

# MDVC corpus: empowering Moroccan Darija speech recognition

Boumehdi Ahmed<sup>1</sup>, Yousfi Abdellah<sup>2</sup>

<sup>1</sup>Information, Communication and Embedded Systems (ICES)-ENSIAS, Morocco Mohammed V University, Temara, Morocco

<sup>2</sup>Faculty of Law, Economics and Social Sciences Souissi-Rabat, Morocco Mohammed V University, Temara, Morocco

## Article Info

### Article history:

Received Nov 25, 2023

Revised Jan 12, 2024

Accepted Jan 14, 2024

### Keywords:

Automatic speech recognition

Low resource language

Moroccan Darija voice corpus

Wav2Vec2

Word error rate

## ABSTRACT

Automatic speech recognition (ASR) technology has significantly transformed human-machine interactions, but it remains limited in its representation of diverse languages and dialects. Moroccan Darija, the lively Moroccan dialect, has long been underrepresented in the realm of language technology. To address this gap, we present a novel corpus of audio files accompanied by meticulously transcribed Moroccan Darija speech. The corpus comprises 1,000 hours of diverse content, featuring multiple Moroccan accents, extracted from 80 YouTube channels. To standardize the representation of Moroccan Darija in our corpus, we made efforts to establish consistent writing norms and conventions. In addition to the dataset creation, we applied fine-tuning using the Wav2Vec2 model on the Moroccan Darija voice corpus (MDVC) dataset achieving a remarkable word error rate (WER) of 9%. This article discusses the current state of Moroccan Darija research, highlighting the scarcity of resources and the need for robust ASR systems. Our contribution offers a valuable resource for researchers and developers, and by standardizing the Darija language, we strive to improve ASR system for this low resource language.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Boumehdi Ahmed

Information, Communication and Embedded Systems (ICES)-ENSIAS, Morocco Mohammed V University  
12000 Chateau 2 Temara, Rabat, Morocco

Email: souregh@gmail.com

## 1. INTRODUCTION

In previous works, we have delved into Arabic speech recognition [1]–[3], and as automatic speech recognition (ASR) technology continues to revolutionize human-machine interactions, it remains essential to ensure inclusivity and representation for diverse languages and dialects. It's worth noting that Ethnologue lists 6,909 [4] known living languages, and Moroccan Darija is among them. This underscores the importance of recognizing and preserving linguistic diversity, especially for languages and dialects like Moroccan Darija that play a significant role in Morocco's cultural and communicative tapestry.

Aligning with this perspective, the Moroccan dialect known as Darija (Moroccan Darija) has garnered attention as a vibrant linguistic entity that often lacks the resources and recognition it deserves in the realm of language technology. In response to this gap, we present an innovative and comprehensive database of audio files, each meticulously accompanied by Moroccan Darija speech transcriptions. This extensive database comprises an impressive 1,000 hours of diverse spoken content, featuring wide range of Moroccan accents that authentically capture the language's multifaceted richness and nuances. The existing research landscape for Darija is explored, highlighting the limited availability of resources and the pressing need for robust ASR systems tailored to this unique dialect.

The core of this article centers on the description of the newly created database, meticulously curated to offer a valuable resource for researchers and developers in the field of Moroccan Darija speech processing. In addition to the database, we aim in this article to build a Wav2Vec2 model for the Moroccan dialect using the Moroccan Darija voice corpus (MDVC) database, something that is achieved for the first time for the Moroccan dialect. Thanks to the size of our training corpus (1,000 hours for 80 speakers), we can say that we have achieved a powerful model for the recognition of the Moroccan dialect.

By combining this innovative database containing standardized Moroccan Darija and with the fine-tuned Wav2Vec2 model, we aim to provide a comprehensive and accessible toolkit for advancing speech recognition technology tailored to the Moroccan Darija dialect. This integrated approach not only addresses the scarcity of resources for Darija but also fosters inclusivity and equal representation in the realm of language technology. Researchers and developers now have a robust foundation to build upon, opening up new possibilities for applications, research, and improved human-machine interactions in Moroccan Darija.

Similarly, we all know that the most spoken language in Morocco in everyday life is the dialect, hence the importance of developing such a speech recognition system for this dialect. This system can be used in several domains, such as:

- In the marketing domain, analyzing audio related to the opinions of Moroccan customers often expressed in dialect on social media can help companies improve the quality of their products.
- In the security domain, this type of system is used to combat crime and acts of terrorism.

## 2. DARIJA AND ITS DIVERSE ACCENTS

The Moroccan dialect, also known as “Darija” or “Moroccan Darija” is the most widely spoken language in Morocco, serving as a common means of communication for the majority of the population. While modern standard Arabic (MSA) is the official language of Morocco and used in formal contexts, Moroccan Darija holds significant importance in everyday interactions, reflecting the nation’s cultural diversity and linguistic heritage. Moroccan Darija differs from MSA in several aspects. First and foremost, Moroccan Darija exhibits a pronounced influence from the Berber languages, which were spoken in Morocco long before the arrival of Arabic. This influence is evident in the vocabulary, pronunciation, and grammar, making Darija distinct from the more formal MSA. Moreover, Darija includes borrowings from French, Spanish, and other languages due to Morocco’s historical interactions with various cultures.

One fascinating feature of the Moroccan dialect is the diverse array of accents found across different regions of the country. Each region, and even specific cities, has its own unique accent, reflecting the local cultural and historical context. For example, the “Tanjawi” accent is associated with the northern city of Tangier, characterized by specific phonetic nuances and colloquial expressions particular to the area. Similarly, the “Fassi” accent is linked to the city of Fes and carries its distinctive linguistic traits.

Other prominent accents include “Marrakshi” from Marrakech, “Casawi” from Casablanca, “Rabati” from Rabat, “Soussi” from the Sous region, and “Sahrawi” from areas bordering Western Sahara. These accents showcase the rich linguistic tapestry of Morocco and reveal the country’s diverse cultural and geographical influences. As the country continues to evolve, Darija remains a testament to Morocco’s collective history and its people’s ability to cherish and preserve their linguistic roots amidst the ever-changing world. In constructing a comprehensive database for Moroccan Darija speech recognition, it was imperative to gather audio samples from these diverse regions across Morocco. By incorporating a wide range of regional accents, our database seeks to represent the authentic linguistic landscape of Morocco, fostering the development of inclusive and accurate speech recognition systems tailored to the intricacies of Darija.

## 3. RELATED WORKS

Previous research on Moroccan Darija language and related works in the field of speech recognition have been relatively limited, showcasing a clear need for further exploration and resource development. While Moroccan Darija serves as a vital communication medium for millions of Moroccans, there has been a scarcity of comprehensive databases and research initiatives dedicated to this distinct dialect. Existing works have mostly centered on broader Arabic language models, neglecting the nuanced variations and regional dialects present in every Arabic country. Presented here are the related works and the database created for Moroccan Darija:

- Belgacem [5] built a speech corpus for 8 dialects representing 8 Arab countries. The Moroccan dialect is included in this corpus, with a speech duration of 90 minutes.
- El Ghazi *et al.* [6] used a database containing 2800 pronunciations of Arabic numerals from 0 to 9 to construct their automatic speech recognition system for the Moroccan Darija in 2011. The authors

employed the hidden Markov model (HMM) to model the phonetic units corresponding to these words from the training database. The obtained results exhibited promising performance, achieving a recognition rate of 91% on a test database.

- Hassine *et al.* [7] developed a classifier for both Tunisian and Moroccan spoken dialects. The test was conducted on numbers from 0 to 9 for speakers. The achieved classification rate is around 98.3%.
- Bezoui [8] developed a compact database comprising twenty speakers, with eleven males and nine females. These speakers pronounced four words in the Moroccan dialect: “سلام” (hi), “كيدايير” (how are you), “لاباس” (there is nothing wrong), and “بيخير” (fine). The database contained a total of 80 words. Utilizing HMM and mel frequency cepstral coefficients (MFCC), they constructed a speech recognition system, achieving an impressive accuracy of approximately 90%. This work showcases a promising step towards advancing speech recognition technology for the captivating Moroccan dialect.
- In 2020, Abderrahim *et al.* [9] embarked on creating a small yet instrumental database for constructing an ASR system tailored to the Darija Moroccan dialect. The corpus was meticulously formed using ten Darija digits, representing specific words in the vocabulary. To compile this database, 20 Moroccan speakers, comprising 10 males and 10 females aged between 14 and 50 years old, were invited to enunciate each digit ten times, generating a total of 2,000 words. During the recording session, each utterance was played back to ensure the inclusion of complete digit representations in the dataset. This database serves as a crucial resource for furthering advancements in ASR technology for the captivating Moroccan Dialect.
- In 2021, Labied and Belangour [10] presented a comprehensive literature review on Moroccan Darija ASR, highlighting its growing importance in human-machine interaction. They outlined the specific constraints of the Moroccan Darija dialect and discussed the limited research on this under-resourced language compared to other international languages in the past. The paper delved into the different works conducted in the field of Moroccan Darija speech recognition and examined the progress made in recent years. Various aspects, including speech analysis techniques, feature extraction, and modeling approaches, were explored, providing insights into the challenges and advancements in ASR for Moroccan Darija in the past. The authors also explored works related to Moroccan Darija resources, such as lexicons, corpora, and dialect identification, emphasizing the importance of annotated corpora for further research in this area.
- In 2021, the Dvoice dataset 1.0, developed by SI2M labs in collaboration with AIOX Lab [11], presents the groundbreaking Dvoice 1.0 database (~1.85GB). This comprehensive compilation comprises 2,576 audio files of Darija speech, each accompanied by its respective transcription. Representing a significant milestone in speech recognition, the dataset features audio files with durations ranging from 1.4 seconds to 10.8 seconds, collectively amounting to an impressive 3 hours, 10 minutes, and 32 seconds. The database upholds high-quality standards with a 44,100 Hz sampling rate, 1,411 kbps bitrate, and 16-bit stereo encoding, ensuring precise and faithful representation of the captivating Darija speech.
- The Dvoice dataset 2.0, also developed later on by the same SI2M labs in collaboration with AIOX Lab, represents a significant advancement in speech recognition. With a size of approximately 2.65GB, the Dvoice 2.0 database boasts an extensive collection of 13,686 audio files. These audio files showcase durations ranging from 0.5 seconds to 13.0 seconds, capturing a diverse range of spoken content and intricacies in the Moroccan dialect. The cumulative duration of 13 hours, 30 minutes, and 14 seconds highlights the substantial size and scope of this valuable resource.  
Notably, Dvoice 2.0 maintains lower audio quality in comparison to Dvoice 1.0. Each audio file in Dvoice 2.0 is meticulously recorded at a sampling rate of 16,000 Hz, with a bitrate of 256 kbps, 16-bit depth, and Mono channel configuration. Despite the quality difference, Dvoice 2.0 stands as a remarkable collection, providing a valuable pool of speech data for further research and development in speech recognition technologies tailored to the captivating Moroccan dialect.
- In 2022, Aitoulghazi *et al.* [12] proposed a solution based on the state-of-the-art architecture named deep speech 2 by Baidu. They conducted tests on 24 hours of speech data (DarSpeech) and achieved promising results, with a word error rate (WER) of 22.7% and a character error rate of 6.03%. This research represents a significant step towards enabling efficient and accurate speech recognition for the widely spoken Moroccan dialect, benefiting both public, and private organizations in the country.

The resources used in the previous works, though significant for their specific research objectives, are relatively small compared to the vast amount of speech data available for languages like English and French. Similarly, the Dvoice 1.0 and Dvoice 2.0 databases, with 3 hours, 10 minutes, and 32 seconds, and 13 hours, 30 minutes, and 14 seconds, respectively, provide valuable datasets, but they still fall short in comparison to the requirements for building robust and accurate speech recognition systems. Even

DarSpeech, with 24 hours of speech, while a considerable advancement, remains limited in the context of the vast linguistic complexity and variations of the Moroccan dialect.

To address this need for more extensive and representative data, we recognized the necessity to create a new database that we called MDVC containing 1,000 hours of pure speech. This sizable database would play a crucial role in constructing a more comprehensive and effective ASR system for the Moroccan dialect. By having access to a larger and more diverse dataset, researchers and developers can overcome the challenges posed by the complexity and richness of the Moroccan dialect, leading to improved ASR models capable of handling a broader range of speech variations and dialectal nuances.

**4. MAJOR CORPORA COMPARISON**

Table 1 that provides a detailed comparison between MDVC, LibriSpeech [13] (for English), common voice corpus (for English, Arabic, French, Spanish, Italian, German, Russian, Turkish, and Chinese) [14] in terms of speech duration. The MDVC stands out in terms of its extensive collection of speech data, offering a remarkable 1,000 hours of spoken content, which reflects its dedication to capturing the diverse linguistic landscape of Moroccan Darija. This dedication to linguistic diversity becomes evident when comparing MDVC to other major corpora.

Table 1. MDVC vs. LibriSpeech and common voice corpora

Corpus	Speech duration (hr)
MDVC (Moroccan Darija)	1,000
LibriSpeech (English)	1,000
Common voice corpus (English)	3,333
Common voice corpus (Arabic)	155
Common voice corpus (French)	1,081
Common voice corpus (Spanish)	2,290
Common voice corpus (Italian)	490
Common voice corpus (German)	1,376
Common voice corpus (Russian)	261
Common voice corpus (Turkish)	118
Common voice corpus (Chinese)	1,223

**5. MOROCCAN DARIJA VOICE CORPUS**

MDVC represents a comprehensive and groundbreaking database specifically tailored for the Moroccan dialect. Developed with great attention to detail, this valuable resource aims to provide researchers and developers with an extensive collection of audio speech data and their corresponding transcriptions, catering to the unique linguistic characteristics and regional variations present in the Darija language see in Figure 1. In Morocco, there are four main classes of dialects:

- The dialect of the North (Tetouan, and chefchaouan).
- The eastern dialect (Oujda, Jerada, and guerssive).
- The West dialect (Rabat, Fes, and Casablanca) (سند: أغلق).
- The dialect of the south or saharian or hassania (laayoun, tantan, and dakhla) (الإغماء).



Figure 1. An extract from the linguistic map of Morocco showing the linguistic distribution in Morocco

For the remainder of the population, they speak either exclusively Tamazight or a combination of Tamazight and Darija see in Figure 1. Within these four classes, there are frequently sub-dialects present. This presents an additional challenge both in terms of collecting all Moroccan dialects and in processing them. The Table 2 shows a difference of Moroccan Darija in different regions of Morocco.

Table 2. Regional variations in Moroccan Darija words

Dialect word	Type	Word in Arabic standar
بدك	North	خمسة
بُلغ	Eastern	أغلق
سند	West	أغلق
الدُّومَة	South	الإغماء
الدُّوْحَة	West-Eastern	الإغماء
النُّور	Eastern	المطر
الشُّتَا	West	المطر

### 5.1. Database creation

To create the MDVC, a meticulous and multi-step process was undertaken. The following outlines the steps involved in its creation. Initially, the process commenced by identifying YouTube channels featuring diverse dialects, as previously specified, with the aim of curating content representing Moroccan Darija speech. After extensive research, we successfully identified approximately 80 YouTube channels where users demonstrated commendable proficiency in pronouncing Darija. These channels were deemed suitable sources for extracting audio content. To automate data collection, a custom script was developed to scrape all video URLs from each identified YouTube channel. The subsequent step involved downloading the identified videos and converting them to the WAV format. For this purpose, the yt\_dlp Python library was employed, ensuring a consistent 16,000 Hz sampling rate, crucial for speech recognition applications.

With the WAV files obtained, the “split\_on\_silence” feature from the pydub Python library was utilized. This technique enabled the segmentation of audio files based on silence detection, setting a threshold of -45 dB and a minimum silence duration of 0.5 seconds to ensure accurate splitting. Audio chunks exceeding 6 seconds in duration were excluded. This step aimed to optimize the dataset for compatibility with models like Wav2Vec2, which we will discuss later on, as they require smaller batch sizes for efficient training on GPUs. For accurate transcriptions, the “AIMP” software was employed for manual transcription of each audio chunk. Transcriptions were then saved into the title tag metadata of the respective chunks, ensuring alignment between the audio and its corresponding text.

Once all transcriptions were completed and saved as metadata, a script was developed to generate a comprehensive and final transcription file by reading the title tags of all audio chunks. This process consolidated the transcriptions into a single, coherent document. To further organize the data, the final transcription file was split into multiple TXT files, and folders were created to correspond with each TXT file. Each TXT file represented a folder containing the audio chunks associated with its transcriptions, effectively organizing the MDVC for efficient access and utilization.

During the creation of the MDVC, we encountered some challenges that required careful handling to ensure the database’s quality and consistency. One of the key issues was the presence of chunks containing non-Arabic words or overlapping voices. Those chunks were excluded from the database.

The final database comprises 1140 TXT files and 1140 folders. For instance, the audios corresponding to the file 113.txt are located within the 113 folder. The contents of “113.txt” begin with the following lines, where the first number points to the name of the chunk (for instance 0.wav) inside the “113” folder, and the Arabic sentence is its transcription:

- 1 اي حاجة (meaning = anything)
- 2 كتبغني تقول (meaning = you want to say)
- 3 نقدر نحتاجها واحد النهار (meaning = I may need it/her one day)
- 4 تتشوفو اللبسة ديال سمية (meaning = You see Samia’s dress)
- 5 تقريبا ساليما القراية (meaning = We almost finished studying)
- 6 المهم حنارا عندنا واحد الإمتحان (meaning = The important thing is that we have an exam)

The choice of 6 seconds was not arbitrary; indeed, it was a deliberate decision made based on the intended use of this database for any ASR system, particularly considering popular models like Facebook’s Wav2Vec2. To achieve optimal results, Wav2Vec2 requires a relatively high batch\_size parameter, such as 32 or 64. This model predominantly relies on GPUs due to their speed advantages compared to CPUs, as training on CPUs could take months to complete. However, when using GPUs with a batch\_size of 32 or 64, there is a risk of running out of memory, particularly with common graphics cards having 8 VRAM

(even when employing the gradient checkpointing technique). To address this limitation, it was necessary to reduce the batch size, which corresponds to reducing the duration of each audio recording. After thorough testing, it was determined that the optimal duration that ensures successful execution and training of the model without errors is approximately 6 seconds.

**5.2. MDVC-problems faced**

**5.2.1. Foreign words**

The creation of the database faced a significant challenge due to dialectal Arabic’s amalgamation of classical Arabic and foreign words, including French, Spanish, and English [15]. This linguistic mix led to complexities in transcribing and segmenting audio content, requiring careful handling. To address these intricacies, we chose to exclude audio segments containing foreign words, maintaining the authenticity of the MDVC. This approach ensures the database accurately represents the nuances of the Moroccan Darija dialect for focused research in speech recognition and language technology. However, in the future, we may enhance this new database by incorporating transcriptions that include foreign words.

**5.2.2. Written representation of Moroccan Darija**

Another challenge encountered was the variability in the written representation of spoken Moroccan Darija words, which can lead to different spellings for the same word. For instance, the sentence “Just go where he goes” in Darija can be written as:

معهمORمعاه غير سير + مشى OR+ مشى + غي سير OR فين ما + مشا ORفينما

Another example, the sentence “and I did not go in the train” in Darija can be written as:

فالتران OR ما مشيتش + في التران OR وانا + مامشيتش OR وانا OR وانا OR وأنا

Another example, just to highlight the difficulty in transcribing Darija, is the word “doctors,” which can be written in multiple ways:

الأطباء، الأطباء، الاطباء، الاطبا، الأطب، لأطبا، لأطباء، لاطباء، لاطبا، لأطب،

These words may exhibit slight pronunciation differences, but they may lead to transcription errors if the transcriber does not pay close attention while listening and transcribing.

**5.2.3. Distinctive Phonemes in Moroccan Darija**

One last challenge faced is that Moroccan Darija exhibits several distinct phonological features including the presence of phonemes that are not found in MSA. These unique phonemes contribute to the distinctiveness and richness of Moroccan Darija:

ك (g): one of the most notable phonemes in Moroccan Darija that is absent in standard Arabic is the sound represented by the letter “ك”. This sound is a voiced velar nasal, similar to the “g” sound in English words like “garage” or “gate”. In Moroccan Darija, this sound appears in words like “كّال” (meaning “he said”) and “كّادير” (referring to the city of Agadir). To simplify the task, ك was transcribed as with ك, ق or ج see in Table 3.

Table 3. New representation in MDVC of words containing the sound of “ك”

Word	New representation in MDVC
أكادير	أكادير
كاراج	كاراج
كاميلة	كاميلة
كاطو	كاطو
كّاع	كّاع
مدكّك	مدكّك
زكا	زكا
كّالس	كّالس
كّاموس	كّاموس
الكنّازة	الكنّازة
كّول	كّول
سايكّ	سايكّ

ف (v): Moroccan Darija also incorporates the phoneme represented by “ف” which corresponds to the “v” sound. This sound is not found in standard Arabic but is prevalent in many Moroccan Darija words.

For instance, “فیراج” (meaning “a turn”) and “فِیلا” (meaning “villa”) is an example where this phoneme is used, “فازلین” (meaning Vaseline-a brand of petroleum jelly) and “فیدیو” (meaning video-a recording of moving visual images). To simplify the task, ف was transcribed as with ف.

پ (p): the letter “پ” represents the “p” sound, which is another phoneme not present in Standard Arabic but used in Moroccan Darija. Words like “پاکستان” (Pakistan-a country in South Asia) and “پودکاست” (podcast refers to a digital audio or video program that is available for streaming or downloading from the internet). To simplify the task, پ was transcribed as with ب.

The presence of these phonemes sets Moroccan Darija apart from standard Arabic and highlights the dialect's historical and linguistic evolution. These phonological distinctions are often accompanied by differences in vocabulary, grammar, and pronunciation, making Moroccan Darija a unique and distinct variety of Arabic. It's important to note that Moroccan Darija has been influenced by various languages and cultures over the centuries, which has contributed to its distinctive phonological features.

While we have mentioned only a few challenges encountered, it's important to note that there were many other complexities in transcribing the Darija language. This variability introduces ambiguity and inconsistency in transcriptions. To address this issue, we established specific transcription norms to ensure uniformity and accuracy across the database. These norms helped maintain a standardized representation of spoken words and promoted clarity in transcriptions, enhancing the usability and reliability of the MDVC for subsequent research and applications in the field of speech recognition and language technology. Let's delve into the specific transcription norms that were established to address the challenges posed by the variability in transcribing the Darija language. The first norm specifically addresses some words ending with the suffix “وا” see in Table 4. In accordance with this norm, a decision has been made to exclude the final letter “ا” in these instances, as it is never pronounced in natural speech.

Table 4. New representation in MDVC of some words ending with the suffix “وا”

Word	New representation in MDVC
کانوا	کانو
فکروا	فکرو
وقفوا	وقفو
قرروا	قررو
جربوا	جربو
قالوا	قالو

The second norm specifically addresses some words ending with the suffix “ا” see Table 5. In accordance with this norm, a decision has been made to replace the final letter “ا” in these instances with the letter “ة”. items, or combine clauses to create a more cohesive and nuanced expression. In the MDVC, the “واو العطف” is not treated as an independent word but is instead written as part of the next word to which it is attached. For instance, in the phrase “کلیت و شربت” (I ate and drank), the conjunction “و” (waaw) is seamlessly attached to the following word, resulting in the written form “کلیت و شربت.”

Table 5. New representation in MDVC of some words ending with the suffix “ا”

Word	New representation in MDVC
باغا	باغة
کاینا	کاینة
غادا	غادة
باقا	باقة
الشهورا	الشهوره
دايرا	دايرة
جایا	جایة
شادا	شادة

The third norm is about the conjunction واو العطف. In the context of Moroccan Darija and Arabic grammar, “واو العطف” (waaw al-‘atf) refers to the conjunction “واو” (waaw), commonly known as the “waaw of coordination” or “waaw of conjunction”. Its versatile nature allows it to connect contrasting ideas, list. The fourth norm addresses words that end with the suffix “ة” in Moroccan Darija. Recognizing the variable pronunciation of the letter “ة,” where it is sometimes pronounced and sometimes silent, the decision has been made to retain the letter “ة” in the representation see in Table 6.

Table 6. New representation in MDVC of some words ending with the suffix “ا”

Word	New representation in MDVC
الصحرا	الصحراء
حمرا خضرا خضرا بيضا	حمراء خضراء خضراء بيضاء
الرجا	الرجاء

In addition to the specific norms outlined above, there are general transcription guidelines that govern the representation of various words in the MDVC. While not explicitly listed, these guidelines ensure consistency and accuracy in the representation of words. Here are examples of words and their new representations in MDVC see in Table 7. It is essential to note that while transcription norms, such as those outlined for MDVC, aim to standardize the representation of words, the choice between a certain representation or its alternative may not necessarily improve the overall accuracy of transcription, but having multiple representations for the same words will potentially lead to an increased WER.

Table 7. New representation in MDVC of some common Moroccan words

Word	New representation in MDVC
هدر نهذرو هدرات الهدرة كيهدر كنهذرو كنهذري ...	هضر نهضرو هضرات الهضرة كيهضر كنهضرو كنهضري ...
اللي للي اللي	لي
وخا	والخا
هداك لهداك فهداك بهداك	هاداك لهاداك فهاداك بهاداك
السي	سي
الحداش حداش لحداش وحداش حداش	الحضاش حضاش لحضاش وحضاش حضاش
عدك عدنا عدناش	عندك عندنا عندناش
كفئش	كيفئش
بصح	بالصح
كولشي	كلشي
لول لولاني لولين لخر لور	اللول اللول اللولاني اللولين اللخر اللور
ايمتا	امتا
لاكارت	لاكارط
كلو	كولو
هداكشي	هاداكشي
نتا نتي انتيا	انت انتي نتيا

### 5.3. Statistics of MDVC

Statistics of the MDVC reveal a vast and diverse collection of linguistic content. The database contains 7 million (7,439,666) words, with 203,756 unique words and a size of ~1.09TB uncompressed. MDVC consists of 1,140 TXT files, and on average, each TXT file consists of approximately 1233.96 lines, indicating that each folder contains an average of 1,234 audio files. Additionally, the MDVC has an average of 6.28 words per transcription. Table 8 listing the 20 most common words and their respective counts.

Table 8. 20 most common words in MDVC

Word	Count
ما	164182
لي	104989
واحد	90486
من	90412
ليا	87594
ديال	87184
انا	65234
على	60503
شي	58301
هاد	54026
ليه	50967
قلت	48551
هو	47090
حتى	42485
اه	42069
غادي	40886
هي	35924
لكن	33630
باش	33106
كان	32669



#### 5.4. Data augmentation

It is worth noting that no data augmentation techniques were used in the construction of the MDVC. The database solely comprises authentic and unaltered audio recordings, ensuring the integrity of the linguistic content collected. However, it is important to acknowledge that data augmentation techniques can be applied to enhance the versatility and robustness of such corpora for various machine-learning tasks [16], [17]. Are five commonly used data augmentation techniques:

- Time stretching: time stretching involves altering the speed of an audio recording without changing its pitch. This technique can be useful for generating variations in speech tempo, which is beneficial for training models to handle different speaking rates.
- Pitch shifting: pitch shifting modifies the pitch of an audio signal while maintaining its duration. It can simulate variations in vocal pitch, which is particularly valuable for training models to recognize speakers with different vocal ranges.
- Noise injection: noise injection involves adding background noise to audio recordings. It helps models become more robust to noisy environments and can simulate real-world conditions where speech may be degraded by various sources of noise.
- Speed perturbation: speed perturbation randomly alters the speaking rate of audio samples by a small factor. This technique can be used to create a more diverse training dataset, enabling models to better adapt to variations in speaking pace.
- Reverberation: reverberation simulates the acoustic effects of sound reflections in different environments. It is typically represented by convolution of the audio signals with a room impulse response [18]. By adding various levels of reverberation to audio recordings, models can become more resilient to changes in recording conditions and room acoustics.

These data augmentation techniques, when applied judiciously, can help improve the performance and generalization of speech and audio processing models trained on limited data [19], without compromising the authenticity of the linguistic content.

### 6. SPEECH RECOGNITION USING MDVC AND WAV2VEC2

In the development of our speech recognition system, we harnessed the power of the MDVC dataset in conjunction with the state-of-the-art Wav2Vec2 model. This combination allowed us to achieve remarkable accuracy in transcribing Moroccan Darija speech, surpassing previous benchmarks in accuracy and efficiency. In the following paragraphs, we delve into the intricacies of the Wav2Vec2 model and detail our fine-tuning process tailored specifically to the MDVC corpus.

#### 6.1. Presentation of Wav2Vec2 model

Wav2Vec2 is an advanced ASR model that has gained significant attention in the field of speech processing. Developed by researchers at Facebook AI, Wav2Vec2 represents a major advancement in ASR technology, achieving state-of-the-art results on various ASR benchmarks [20]. Wav2Vec2's groundbreaking features have redefined the landscape of ASR technology. Central to its success is the concept of self-supervised learning, a paradigm shift in ASR research. By training on vast amounts of unlabeled speech data [21], Wav2Vec2 transcends the limitations of traditional supervised methods. This innovation not only significantly boosts data efficiency but also empowers the model to grasp the intricacies of diverse accents and languages, making it a versatile tool for multilingual ASR applications.

#### 6.2. Fine-tuning Wav2Vec2 using MDVC and Wav2Vec2-large-XLSR-53

The adaptability and versatility of Wav2Vec2 are further exemplified by its fine-tuning capability. A model can be fine-tuned on specific ASR tasks using labeled data. This fine-tuning process fine-tunes its performance, allowing excel in various speech recognition tasks across different low resource languages [22].

Wav2Vec2-large-XLSR-53 is a cutting-edge multilingual speech recognition model introduced by Facebook. It excels at cross-lingual speech representation learning, leveraging raw waveforms from multiple languages. Building on Wav2Vec 2.0, it jointly learns shared latent speech representations, surpassing monolingual training. This model offers a competitive alternative to individual language models, with shared representations benefiting related languages. With 53 languages pretraining, XLSR-53 is poised to advance low-resource speech understanding research [23]. This model was trained on public training data and comprises about 300M parameters [24].

In our research, we undertook several important steps to fine-tune the Wav2Vec2-large-XLSR-53 model for speech recognition tasks using the MDVC. These steps included data preprocessing, model fine-tuning, and the selection of specific hyperparameters and evaluation metrics.

To train this model, we utilized 70% of the MDVC corpus, which corresponds to 700 hours of speech. And to facilitate the integration of the text and audio data, we created a tokenizer. Each token generated by the tokenizer was assigned a unique code or index. This step is essential as it enables the model to represent the textual information in a numerical format that can be processed by machine learning algorithms. In evaluating the performance of our models, we utilized the WER as a key metric as it is the most popular metric for ASR evaluation [25]. WER measures the accuracy of the model’s transcriptions by comparing them to the ground truth text. A lower WER value indicates a more accurate transcription, making it a commonly used metric in speech recognition tasks.

Regarding the hyperparameters, we selected specific values to fine-tune our models effectively. For example, we set attention\_dropout, hidden\_dropout, and feat\_proj\_dropout to 0.0, which means we didn’t employ dropout regularization in these layers. We also used values like mask\_time\_prob and layerdrop to control model behavior during training, ensuring that it learned effectively from the data.

To accommodate our hardware limitations, we leveraged gradient\_checkpointing\_enable. This technique allows us to trade off computation for memory by computing gradients only for a subset of model parameters, helping us avoid memory exceptions while training. So we fine-tuned training-related parameters like per\_device\_train\_batch\_size=32 and gradient\_accumulation\_steps=4 to optimize memory usage and training efficiency.

**6.3. The challenge**

Our carefully configured setup, which included selecting appropriate hyperparameters and fine-tuning the Wav2Vec2 model using the MDVC dataset, proved to be instrumental in achieving the best WER for our speech recognition system. It enabled us to harness the full potential of the models and optimize their performance to achieve highly accurate transcriptions. However, during our evaluation process, we encountered a notable challenge inherent to Moroccan Darija, a dialect of Arabic spoken in Morocco. One of the significant issues we identified while checking the errors generated by the system was the flexibility in writing Moroccan Darija. This dialect can be transcribed in various ways, leading to multiple possible representations for the same spoken content. This variability introduced challenges in the ground truth transcriptions provided by the MDVC dataset, as there can be different valid ways to write the same spoken sentence.

To address this challenge and improve the quality and consistency of the dataset, we undertook a critical step in our research. We re-adjusted the MDVC corpus to standardize the transcriptions of Moroccan Darija with more rules. This involved refining and normalizing the text representations to ensure a more consistent and accurate alignment between the spoken audio and its textual counterpart. By doing so, we were able to create a more reliable and consistent dataset, which in turn contributed significantly to the improved performance and lowered WER rates of our speech recognition system. This critical step underscores the importance of data preprocessing and standardization in the development of robust speech recognition models, particularly when dealing with dialects or languages with varying orthographic conventions.

**6.4. The evaluation**

Following the establishment of standardized rules for Moroccan Arabic transcription, we embarked on a meticulous journey to rectify all potential discrepancies in the data. The model is evaluated on three sets: the global corpus (1,000 hours), the training set (700 hours), and the test set (300 hours). The result of this dedicated effort was a substantial improvement in our speech recognition system, epitomized by a remarkable WER of 9% see in Table 9. To get an idea of the evaluation of our work compared to what already exists, we have compared our system with two Moroccan dialect speech recognition systems. The Table 10 shows this comparison.

Table 9. Performance metrics for fine-tuned speech recognition model on MDVC corpus

Test set	Training set
9%	3%

Table 10. Comparison of our system using MDVC to other systems

	Type of system	Database	WER
Bezoui [8]	Isolated words	80 words for 4 speaker	10%
Allak <i>et al.</i> [11]	Continuous speech	DVoice (13.5 hours)	30%
Aitoulghazi <i>et al.</i> [12]	Continuous speech	DarSpeech (24 hours speech)	22%
Our system using MDVC Corpus, 2024	Continuous speech	MDVC (1,000 hours) for 80 speaker	9%

Additionally, the WER on the training set reached an impressive low of 3%. However, despite the remarkable performance achieved during training, we encountered a situation where further training epochs did not result in noticeable improvement. The test set remained at a low value of 7%. This prompted us to stop the training process, as the model had reached a point of diminishing returns in terms of improved WER on the validation set. Nonetheless, the consistently low validation loss suggests that the model had successfully learned to generalize its knowledge to new and unseen examples, highlighting its robustness and effectiveness in recognizing Moroccan Darija speech.

Finally, we are pleased to announce that the final model, which boasts these impressive metrics, is now accessible on the HuggingFace platform [26]. This Moroccan Darija model represents a valuable resource for future endeavors in speech recognition for Moroccan Darija. Importantly, it offers an efficient alternative to the time-consuming process of fine-tuning from scratch or using pre-existing models like Wav2Vec2-large-XLSR-53. As a result, our model stands as a testament to the power of data standardization and the potential for further advancements in speech recognition for Moroccan Darija and other dialects, making it readily available for utilization in various applications.

## 7. CONCLUSION

In conclusion, the creation of the MDVC marks a significant milestone in advancing speech recognition technologies for the Moroccan dialect. This comprehensive database, comprising 1,000 hours of pure speech extracted from various YouTube channels, captures the rich linguistic diversity and regional accents present in Darija. The meticulous selection process, along with the utilization of state-of-the-art techniques and tools, ensures the quality and accuracy of the dataset. Furthermore, our efforts to fine-tune the Wav2Vec2 model on the MDVC dataset yielded remarkable results, with a WER as low as 9%. This achievement underscores the efficacy of our approach in enhancing speech recognition accuracy for Moroccan Darija. By overcoming the challenges posed by limited resources in existing databases, the MDVC offers a valuable resource for researchers and developers interested in building automatic speech recognition systems and furthering their understanding of the Moroccan dialect. This dataset, in conjunction with fine-tuned models, paves the way for more accurate and efficient speech recognition solutions and opens new possibilities for the development of applications tailored to the Moroccan Darija dialect.





## REFERENCES

- [1] A. Boumejdi and A. Yousfi, "Arabic speech recognition independent of vocabulary for isolated words," in *Lecture Notes in Networks and Systems*, vol. 216, 2022, pp. 585–595.
- [2] A. Boumejdi and A. Yousfi, "Comparison of new approaches of semi-syllable units for speech recognition of any Arabic word," *Journal of Physics: Conference Series*, vol. 2337, no. 1, p. 012002, Sep. 2022, doi: 10.1088/1742-6596/2337/1/012002.
- [3] A. Boumejdi and A. Yousfi, "Construction of a database for speech recognition of isolated Arabic words," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, Sep. 2020, pp. 1–4, doi: 10.1145/3419604.3419752.
- [4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, no. 1, pp. 85–100, Jan. 2014, doi: 10.1016/j.specom.2013.07.008.
- [5] M. Belgacem, "Construction d'un corpus robuste de différents dialectes arabes," *Actes des 8emes Rencontres Jeunes Chercheurs en Parole*, vol. 33, no. 0, 2009.
- [6] A. El Ghazi, C. Daoui, N. Idrissi, M. Fakir, and B. Bouikhalene, "Speech recognition system based on hidden markov model concerning the Moroccan Dialect DARIJA Speech recognition system based on hidden markov model concerning the Moroccan Dialect DARIJA," *Global Journal of Computer Science and Technology*, vol. 11, no. September, 2011.
- [7] M. Hassine, L. Boussaid, and H. Messaoud, "Maghrebian dialect recognition based on support vector machines and neural network classifiers," *International Journal of Speech Technology*, vol. 19, no. 4, pp. 687–695, 2016, doi: 10.1007/s10772-016-9360-6.
- [8] M. Bezoui, "Speech recognition of Moroccan Dialect using hidden markov models," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 1, p. 7, Mar. 2019, doi: 10.11591/ijai.v8.i1.pp7-13.
- [9] A. Ezzine, H. Satori, M. Hamidi, and K. Satori, "Moroccan dialect speech recognition system based on CMU SphinxTools," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Jun. 2020, pp. 1–5, doi: 10.1109/ISCV49265.2020.9204250.
- [10] M. Labied and A. Belangour, "Moroccan dialect 'Darija' automatic speech recognition: a survey," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, Jul. 2021, pp. 208–213, doi: 10.1109/PRML52754.2021.9520690.
- [11] A. Allak, A. M. Naira, I. Benelallam, and K. Gaanoun, "Dialectal Voice : an open-source voice dataset and automatic speech recognition model for Moroccan Arabic dialect," *NeurIPS Data-Centric AI Workshop*, 2021.
- [12] O. Aitoulghazi, A. Jaafari, and A. Mourhir, "DarSpeech: an automatic speech recognition system for the Moroccan Dialect," in *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, May 2022, pp. 1–6, doi: 10.1109/ISCV54655.2022.9806105.
- [13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, vol. 2015-Augus, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
- [14] R. Ardila et al., "Common voice: a massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.





- [15] M. Rafiq, "La situation linguistique au Maroc entre hier et aujourd'hui," *Université Hassan II de Casablanca*, pp. 291–310, 2017.
- [16] I. Rebai, Y. Benayed, W. Mahdi, and J. P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Computer Science*, vol. 112, pp. 316–322, 2017, doi: 10.1016/j.procs.2017.08.003.
- [17] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 810–814, 2014, doi: 10.21437/interspeech.2014-207.
- [18] D. S. Park *et al.*, "SpecAugment: a simple data augmentation method for automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Sept, pp. 2613–2617, 2019, doi: 10.21437/Interspeech.2019-2680.
- [19] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Interspeech 2016*, Sep. 2016, vol. 08-12-Sept, pp. 2378–2382, doi: 10.21437/Interspeech.2016-1386.
- [20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: a framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, Jun. 2020, <http://arxiv.org/abs/2006.11477>.
- [21] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Wav2Vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46938–46948, 2023, doi: 10.1109/ACCESS.2023.3275106.
- [22] Y. Cheng, *et al.*, "Applying Wav2Vec2. 0 to speech recognition in various low-resource languages," arXiv preprint arXiv:2012.12121, 2020.
- [23] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, pp. 346–350, 2021, doi: 10.21437/Interspeech.2021-329.
- [24] A. Babu *et al.*, "XLS-R: self-supervised cross-lingual speech representation learning at scale," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September, pp. 2278–2282, 2022, doi: 10.21437/Interspeech.2022-143.
- [25] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic speech recognition errors detection and correction: a review," *Procedia Computer Science*, vol. 128, pp. 32–37, 2018, doi: 10.1016/j.procs.2018.03.005.
- [26] B. Ahmed "Wav2Vec2-Large-XLSR-53-Moroccan-Darija," <https://huggingface.co/boumehdi/wav2vec2-large-xlsr-moroccan-darija>.

## BIOGRAPHIES OF AUTHORS



**Boumehdi Ahmed**     Mr. Boumehdi Ahmed is a dedicated engineer currently pursuing a Ph.D. at National School of Computer Science and Systems Analysis (ENSIAS) in his sixth year of study. He holds an engineering degree from INPT (National Institute of Posts and Telecommunications). His research interests revolve around automatic speech recognition for Arabic and Moroccan dialects, alongside a focus on natural language processing for the same languages. Balancing his academic pursuits with professional responsibilities, he actively contributes to the practical application of his knowledge. For inquiries and collaboration opportunities, Mr. Boumehdi Ahmed showcasing his dedication to advancing technology in linguistics. He can be contacted at email: [souregh@gmail.com](mailto:souregh@gmail.com).



**Abdellah Yousfi**     was a professor at the Institute of Study and Research for Arabization (2003-2007). Professor at the Faculty of Law, Economics, and Social Sciences of Souissi at Mohamed V University in Rabat since 2007. He is member of the ICES Team in the ENSIAS, at Mohamed V University in Rabat, Morocco. His research interests include creation of corpora for the Arabic language, Arabic speech recognition, Arabic handwriting recognition and correction of Arabic spelling errors. He can be contacted at email: [a.yousfi@um5r.ac.ma](mailto:a.yousfi@um5r.ac.ma).