Tachelhiyt-Darija: a parallel speech corpus for two underrepresented languages

Noureddine Atouf
Chouaib Doukkali University
El Jadida, Morocco
atouf.noureddine@ucd.ac.ma

Elsayed Issa
Purdue University
West Lafayette, IN, USA
esissa@purdue.edu

Said Ouzbayr
Chouaib Doukkali University
El Jadida, Morocco
ouzbayrsaid@gmail.com

Abstract

Despite recent advances in speech technology, several languages remain underrepresented. This linguistic disparity exacerbates the existing technological divide, resulting in limited access to speech-driven technologies. A key factor contributing to this challenge is the scarcity of datasets necessary to develop diverse speech recognition systems for low-sourced languages such as Amazigh and Moroccan Arabic. While the Amazigh language is emblematic of cultural identity and is deeply embedded in history, Darija remains the dialect spoken by the majority in Morocco. In this work, we introduce the first Tachelhiyt-Darija speech parallel corpus. A total of 24 Amazigh and 11 Moroccan Darija speakers recorded the parallel textual data, yielding a corpus of 2,772 audio segments. We also conducted benchmarking and fine-tuning of the Whisper ASR model. The results underscored the need for the development of datasets for under-resourced languages.

1 Introduction

According to the latest population census in Morocco (High Commission for Planning, 2024), Moroccan Arabic and Tamazight, both reckoned to be descendants of the Semitic branch of the Afro-Asiatic language family, are the most widely spoken languages in the country (Aissati et al., 2011). In a striking comparison, 92% of Moroccans use Darija (a term often interchangeably used with Moroccan Arabic), while only 25% of the population speaks Amazigh. At the sociolinguistic level, Moroccan Arabic and Amazigh are two dialectical forms spoken in different geographical areas and both have standardized forms: Modern Moroccan Arabic and Tamazight (Sadiqi, 2014; Youssi, 1992).

While Modern Standard Arabic (MSA) remains the country's first official language primarily utilized in formal speeches, administrative correspondences and documentation, news broadcast, and education as the medium of instruction, Moroccan Arabic and Tamazight have long maintained the status of the language of daily social interactions and conversations. In contrast to Darija, and particularly in 2001, Tamazight was recognized as a national heritage for all Moroccans, with the highest authority in the country (King Mohammed VI, 2001) issuing a decree to preserve, promote, and reinforce the Amazigh language and culture throughout the establishment of the Royal Institute of Amazigh Culture (IRCAM - Institut Royal de la Culture Amazighe). Amazigh was introduced into primary school curricula in 2003, marking a shift from its oral tradition to a formalized, codified language. This transition faced major challenges (Aissati et al., 2011; Ennaji, 2014), including standardizing the language across its three main varieties (namely, Tamazight, Tachelhiyt, and Tarifiyt) and selecting an appropriate script—ultimately, the Tifinagh script was adopted, evolving from the ancient Lybico-Berber alphabet.

The present work introduces the *Tachelhiyt-Darija* parallel speech corpus[1]. The corpus includes speech transcriptions in addition to speaker gender and dialect. In addition to describing our dataset, we developed baseline systems for automatic speech recognition (ASR). To summarize, our contributions are as follows. First, we introduce *Tachelhiyt-Darija*, a fully supervised speech dataset for two underrepresented dialects, labeled with transcriptions, dialect, and gender. Second, we evaluate Whisper as the state-of-the-art (SoTA) multilingual ASR model across the two dialects to assess its performance in recognizing underrepresented speech.

2 Related work

There is a limited body of research that develops speech data for underrepresented languages to

[1] https://huggingface.co/datasets/NoureddineMOR/tachelhiyt-darija

serve different purposes such as ASR models, machine translation, language learning, et cetera. Several scholars designed datasets for individual letters and digits for Amazigh spoken digit recognition tasks (Abakarim and Abenaou, 2023; Boulal et al., 2023; Hamidi et al., 2020; Telmem and Ghanou, 2018; El Ghazi et al., 2014; Satori and ElHaoussi, 2014).

On the word and sentence levels, El Ouahabi et al. (2017) developed a corpus for recognizing 520 spoken Amazigh words from 50 native Tarifiyt speakers, while (Oukas et al., 2024) created a database of Arabic vocal data by Tamazight speakers to train ASR models sensitive to Tamazight-accented Arabic. These efforts, along with (Daouad et al., 2023), focus on building word-based corpora to improve speech recognition systems and human-machine vocal interaction. Additionally, Mozilla Common Voice (Ardila et al., 2020) also contributes to this goal, offering a publicly available dataset with 398 Tamazight audio files for training and 159 for testing in version 17.0.

A different mode of database creation is attested in Moroccan Arabic. A variety of Moroccan Arabic databases have been developed, with the Darija Open Dataset (Outchakoucht and Es-Samaali, 2021) standing out for its 10,000 annotated entries featuring English translations and detailed linguistic information. Other studies, such as (Zaidani et al., 2024b) and (Zaidani et al., 2024a), constructed corpora through manual transcription and audio segmentation of YouTube content, while (Talafha et al., 2024) compiled 48 hours of transcribed data from North African and other Arabic dialect speakers—though Amazigh was notably excluded. Additional contributions include (Samih and Maier, 2016), (Ali et al., 2019), and (Labied et al., 2023), each aiming to support NLP, ASR, and speech-to-text translation tasks for Moroccan Arabic.

To our knowledge, there has not been an attempt to consider recording sentence stimuli in a parallel corpus consisting of two underrepresented languages, where the goal is to increase accessibility of the language in terms of the used script. In the context of the present experiment, we expand the Amazigh language speech corpus to include longer sequences instead of isolated words[2]. However, some publicly available datasets on Huggingface provide sentence recordings along with their written transcripts in Tifinagh[3] and phoentic symbols[45]. This study introduces a parallel corpus featuring two underrepresented languages—Tamazight and Darija—with vocal recordings transcribed in Arabic script for greater accessibility. Unlike existing repositories such as HuggingFace, which use Tifinagh or phonetic scripts for Tamazight, this work contributes a balanced, Arabic-script-based speech dataset for both languages.

3 Corpus Design

3.1 Parallel data and speech recoding

The process of designing the corpus was convoluted given the unavailability of written resources in Amazigh. Offline and online materials such as stories and written poetry exist in substantial amounts. Still, they are scripted in Tifinagh and require to be translated by a literate reader in the language. Moreover, visual content in Amazigh on the YouTube platform is abundant but is devoid of automatically generated subtitles. The absence of subtitled episodes with pure backgrounds determined our modus operandi with respect to data collection. The first stage of compiling the data was conducted through listing down numerous random (functional) sentences in Moroccan Arabic. The sentences were both generated by the authors and excerpted from the "Moroccan Arabic Textbook" designed by Peace Corps Morocco [6].

Two native speakers of Tachelhyit (the target Amazigh variety in the extant study) were, then, recruited to either literally translate the Moroccan Darija sentences into Tachelhyit. The rendered sentences were provided in the Arabic script. At the end, the data was further broken into sub-lists, which were sent to the participants who in turn recorded the stimuli in their first language.

The participants were two groups of Amazigh and Moroccan Arabic native speakers. The varying size of the groups, with the Amazigh participants forming the majority (n=24), reflected the study's sampling technique. Both male (n=14, 40%) and female (n=21, 60%) participants, whose age range differed across three major groups were included to

[2] https://huggingface.co/datasets/Tamazight-NLP/tamawalt-n-imZZyann

[3] https://huggingface.co/datasets/Tamazight-NLP/tosd

[4] https://huggingface.co/datasets/TifinLab/moroccan_amazigh_asr

[5] https://huggingface.co/datasets/TifinLab/tamazight_asr

[6] https://www.friendsofmorocco.org/Docs/Darija/Moroccan%20Arabic%20textbook%202011.pdf

Tachelhiyt				Darija			
audio	**transcript**	**speaker**	**gender**	**audio**	**transcript**	**speaker**	**gender**
t_1.wav	أزول فلاون	t_sp14	male	d_1.wav	السلام عليكم	d_sp14	male
t_2.wav	رادفتوخ ناف أوزكا	t_sp14	male	d_2.wav	غادي نمشي بعد غدا	d_sp14	male
t_3.wav	لوقت ن الطوبيس	t_sp12	female	d_3.wav	توقيت الطوبيسات	d_sp12	female

Table 1: Sample of the Tachelhiyt-Darija speech parallel corpus

diversify the spoken form of our data. The majority of participants (57.10%) were under 19 years old, followed by 25.70% aged 20–29, and 17.10% aged 30–40. Table 2 shows the count and percentage rates for such demographic information as gender and participants' native tongue. For the age factor, the table displays the mean, range and standard deviation:

Variables	Values	N	%	M	R	SD
Gender	Male	14	40			
	Female	21	60			
Native L.	Amazigh	24	68			
	Moroccan	11	31			
Age				22	19	6
Education	high school	21	60			
	undergrad.	5	14			
	postgraduate	9	25			

Table 2: Demographic Variables

3.2 Data preparation and segmentation

All participants recorded the data in the wild without any professional equipment, resulting in audio files with different sampling rates, 44.100 and 48.000, for Tachelhiyt and Darija respectively. The participants were asked to make a five-second pause between a recorded string and the following one. This allowed for optimal splitting of the data using *AudioSegment*[7] to extract the speech segments. The resulted data was verified by the experimenters to make sure the audio files were segmented properly and were aligned well with the corresponding transcripts.

Each audio segment was converted to a single channel 16 kHz 16-bit PCM encoded WAV files using the FFmpeg library (Tomar, 2006). For each dialect, there is a comma-separated (CSV) file containing two columns with the audio file name as the first column and the transcript as the second column. The metadata file includes eight columns; four for each dialect as shown in Table 1. The metadata further facilitated the filtering of specific speakers on the basis of their number and gender.

A total of 2772 speech segments made up the developed parallel corpus in Tachelhiyt and Draija.The duration of the Tachelhiyt recordings is 71.68 minutes, with an average audio segment length of 3.11 seconds, while the duration of the Darija recordings is 61.84 minutes, with an average segment length of 2.68 seconds.

4 Experiments

We evaluated Whisper (Radford et al., 2023) performance using zero-shot and full finetuning evaluation. Our primary goal is to report the importance of speech datasets by benchmarking one of the state-of-the-art ASR models.

4.1 Zero-shot evaluation

We evaluated Whisper small (242M) and large-v3 (1.54B) in a zero-shot setting using the test dataset (i.e., 300 speech samples). Then, we reported Word Error Rate (WER) and Character Error Rate (CER).

Model	Tachelhiyt	Darija
	wer/cer	wer/cer
Baseline (small)	127.09/80.85	89.72/60.47
Baseline (large)	145.54/85.86	76.08/34.44

Table 3: Results of Word Error Rate (WER) and Character Error Rate (CER) for the baseline models.

Table 3 shows the performance results for the baseline models on Tachelhiyt and Darija. The results reveal notable discrepancies across both languages and model sizes. For Tachelhiyt, the WER/CER increased from 127.09/80.85 in the

[7] https://github.com/jiaaro/pydub

small model to 145.54/85.86 in the large model, indicating a performance degradation with the larger model. In contrast, Darija showed improvement with model scaling: the WER/CER decreased from 89.72/60.47 in the small model to 76.08/34.44 in the large model. These results suggest that the large model is more effective for Darija but less suited for Tachelhiyt. This is possibly due to differences in linguistic structure, training data distribution, or model overfitting.

4.2 Full Finetuning

For fine-tuning, we sampled audio data at a sampling rate of 16 kHz. All experiments were conducted on a single-node Google Colab instance equipped with an A100 GPU. We fine-tuned two Whisper small models separately for Tachelhiyt and Darija, using identical hyperparameters for both models. The fine-tuning process employed a training batch size of 16 and an evaluation batch size of 8. We set the learning rate to 1e-5 and trained for a maximum of 1000 steps, equivalent to approximately 15 epochs. These hyperparameters were chosen to accommodate the relatively small size of the available training data. For decoding, we used a maximum sequence length of 225 tokens. No additional post-processing steps were applied to the decoded outputs. Although we evaluated both the Whisper small and large-v3 models in a zero-shot setting, we opted to fine-tune the small model due to the limited size of the available dataset.

Model	**Tachelhiyt**	**Darija**
	wer/cer	wer/cer
(small)	7.45/3.25	4.26/1.38

Table 4: Results of Word Error Rate (WER) and Character Error Rate (CER) for the finetuned models.

As Table 4 shows, the fine-tuned small model demonstrates a substantial performance improvement over the baseline, achieving significantly lower error rates across both languages. For Tachelhiyt, WER and CER dropped to 7.45 and 3.25, respectively, while Darija saw even lower error rates of 4.26 (WER) and 1.38 (CER), indicating the effectiveness of fine-tuning in enhancing model accuracy for both language varieties.

The findings have important implications for the development and use of ASR systems in underrepresented languages. First, the baseline performance highlights the challenges that pre-trained ASR models face when deployed in low-resource languages such as Tachelhiyt and Darija. The very high WER and CER across board consistently show that such models are not inherently able to generalize to linguistically diverse contexts without some form of adaptation. This finding further supports the need for the creation of datasets and finetuning to broaden the applicability of ASR technologies to resource-poor languages, which mitigates the existing technological gap.

5 Conclusion

To conclude, this study introduces the first sequential parallel corpus of Tachelhiyt-Darija, comprising 71.68 minutes of Tachelhiyt and 61.84 minutes of Darija. Based on our review of the literature, this is the first parallel dataset of North African speech data from two less represented dialects. It has the potential to be used in Automatic Speech Recognition (ASR), speech-to-speech translation, and machine translation, with further applications for cross-lingual studies between these two languages. We have also applied and validated the Whisper ASR model by means of the utilized dataset for benchmarking and fine-tuning. The findings reveal the importance of creating language-specific datasets for less-documented languages in order to improve the current state of the art in speech technologies as well as enhance linguistic accessibility.

Limitations

This study has certain limitations that should be acknowledged. The first limitation is that the quality and consistency of the audio data may have been affected by variations in participants' recording equipment, eventually leading to noise and possible background interference. This is ascribed to the small size and parallel nature of the dataset, which may have limited the scope of training and evaluation for the fine-tuned models, consequently limiting the generalizability of the results. However, the results are quite insightful. A larger and more diverse corpus could very well enhance the performance and robustness of the system.

In this vein, future work will involve the release of an expanded dataset— another parallel corpus—aimed at increasing both the quantity and quality of data for Tamazight three varieties and Darija, thereby enhancing ASR technology for these linguistic groups. In addition, future research will bring structural and linguistic differences (i.e., phonological and morphological components) that characterize these two languages. This should further improve the performance of the ASR system.

Ethics Statement

In developing Tachelhiyt-Darija corpus, we adhered to ethical principles to ensure responsible and respectful use of data. All speech data used in this study were collected in strict accordance with ethical research standards. Participants were fully informed about the purpose of the study, the nature of the data being collected, and their right to withdraw at any time without consequence. No personally identifiable information was collected, and all audio recordings were anonymized to protect participant privacy. The collected data are used exclusively for academic and research purposes related to language resource development and are stored securely to prevent unauthorized access.

References

Fadwa Abakarim and Abdenbi Abenaou. 2023. Enhancing amazigh speech recognition system with mfdwc-svm. In *International Conference on Computational Science and Its Applications*, pages 471–488. Springer.

Abdelilah El Aissati, Susy Karsmakers, and Jeanne Kurvers. 2011. "we are all beginners": Amazigh in language policy and educational practice in morocco. *Compare: A Journal of Comparative and International Education*, 41(2):211–227.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Hossam Boulal, Mohamed Hamidi, Mustapha Abarkan, and Jamal Barkani. 2023. Amazigh spoken digit recognition using a deep learning approach based on mfcc. *International journal of electrical and computer engineering systems*, 14(7):791–798.

Mohamed Daouad, Fadoua Ataa Allah, and El Wardani Dadi. 2023. An automatic speech recognition system for isolated amazigh word using 1d & 2d cnn-lstm architecture. *International Journal of Speech Technology*, 26(3):775–787.

Ahmed El Ghazi, Cherki Daoui, and Najlae Idrissi. 2014. Automatic speech recognition for tamazight enchained digits. *World Journal Control Science and Engineering*, 2(1):1–5.

Safâa El Ouahabi, Mohamed Atounti, and Mohamed Bellouki. 2017. A database for amazigh speech recognition research: Amzsrd. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, pages 1–5. IEEE.

Moha Ennaji. 2014. *Multiculturalism and Democracy in North Africa: Aftermath of the Arab Spring*. Routledge.

Mohamed Hamidi, Hassan Satori, Ouissam Zealouk, and Khalid Satori. 2020. Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology*, 23(1):101–109.

High Commission for Planning. 2024. Morocco population and housing census.

Maria Labied, Abdessamad Belangour, and Mouad Banane. 2023. Darija-c: towards a moroccan darija speech recognition and speech-to-text translation corpus. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pages 1–4. IEEE.

Nourredine Oukas, Tiziri Chabi, and Tilelli Sari. 2024. A novel dataset for arabic speech recognition recorded by tamazight speakers. *Authorea Preprints*.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Fatima Sadiqi. 2014. Berber and language politics in the moroccan educational system. In Moha Ennaji, editor, *Multiculturalism and Democracy in North Africa: Aftermath of the Arab Spring*. Routledge.

Younes Samih and Wolfgang Maier. 2016. An arabic-moroccan darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4170–4175.

Hassan Satori and Fatima ElHaoussi. 2014. Investigation amazigh speech recognition using cmu tools. *International Journal of Speech Technology*, 17:235–243.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, et al. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

Meryam Telmem and Youssef Ghanou. 2018. Estimation of the optimal hmm parameters for amazigh speech recognition system using cmu-sphinx. *Procedia Computer Science*, 127:92–101.

Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Abderrahim Youssi. 1992. *Grammaire et lexique de l'arabe marocain moderne*. Wallada, Casablanca.

Hajar Zaidani, Abderrahim Maizate, Mohammed Ouzzif, and Rim Koulali. 2024a. Building a corpus for the underexplored moroccan dialect (cfmd) through audio segmentations. *Revue d'Intelligence Artificielle*, 38(3):857.

Hajar Zaidani, Abderrahim Maizate, Mohammed Ouzzif, and Rim Koulali. 2024b. Cfmd: Corpus for moroccan dialect as under researched dialect. In *Future of Information and Communication Conference*, pages 61–69. Springer.