ELYADATA & LIA at NADI 2025: ASR and ADI Subtasks

Haroun Elleuch[1,2,†], Youssef Saidi[1,‡], Salima Mdhaffar[2], Yannick Estève[2], Fethi Bougares[1,2]

[1]ELYADATA, France [2]LIA, Avignon University, France

† Main contributor of the ADI subtask ‡ Main contributor of the ASR subtask

Correspondence: haroun.elleuch@elyadata.com

Abstract

This paper describes Elyadata & LIA's joint submission to the NADI multi-dialectal Arabic Speech Processing 2025. We participated in the Spoken Arabic Dialect Identification (ADI) and multi-dialectal Arabic ASR subtasks. Our submission ranked first for the ADI subtask and second for the multi-dialectal Arabic ASR subtask among all participants. Our ADI system is a fine-tuned Whisper-large-v3 encoder with data augmentation. This system obtained the highest ADI accuracy score of **79.83%** on the official test set. For multi-dialectal Arabic ASR, we fine-tuned SeamlessM4T-v2 Large (Egyptian variant) separately for each of the eight considered dialects. Overall, we obtained an average WER and CER of **38.54%** and **14.53%**, respectively, on the test set. Our results demonstrate the effectiveness of large pre-trained speech models with targeted fine-tuning for Arabic speech processing.

1 Introduction

Arabic is one of the most widely spoken languages in the world, both in terms of number of speakers and geographical spread (Lane, 2025). This wide distribution, coupled with centuries of contact with other languages and cultures, has led to the emergence of numerous colloquial varieties collectively known as Arabic dialects. Although the exact granularity and classification of these dialects remain a matter of debate, a common working assumption in computational processing is to associate a dialect with a country-level variety (Bouamor et al., 2014; Shon et al., 2020), or to a larger area where subdialects are the most similar (Gulf, Levant, North Africa) (Dhouib et al., 2022; Ali et al., 2017).

Dialectal Arabic poses unique challenges for speech and language processing. Unlike Modern Standard Arabic (MSA), dialects are predominantly spoken rather than written (Ferguson, 1959), with significant variation in phonology, lexicon, and syntax. They also lack standardized orthographic conventions, despite recent efforts such as CODA (Conventional Orthography for Dialectal Arabic) (Habash et al., 2012), and later efforts of Habash et al. (2018) and Alhafni et al. (2024). These properties complicate both Automatic Speech Recognition (ASR) and Automatic Dialect Identification (ADI) tasks, where systems must generalize across substantial linguistic variability.

The 2025 Nuanced Arabic Dialect Identification (NADI) Shared Task (Talafha et al., 2025) addresses these challenges through three subtasks aimed at improving the coverage and robustness of speech technologies for Arabic dialects:

- Spoken Arabic Dialect Identification (ADI)
- Multidialectal Arabic ASR using the recently released Casablanca dataset (Talafha et al., 2024)
- Diacritic Restoration focusing on dialectal variations of Arabic

Our team participated in the first two subtasks, achieving first place in ADI and second place in multi-dialectal ASR on the official test sets. In both cases, we leveraged large-scale pre-trained speech models with targeted fine-tuning strategies to address dialectal variability.

Our main contributions are (1) We propose an effective two-stage fine-tuning approach for ADI, using the Whisper-large-v3 encoder to achieve state-of-the-art results. (2) We demonstrate that separately fine-tuning the SeamlessM4T-v2 Large model for each dialect yields competitive ASR performances.

2 Arabic Dialect Identification

The ADI subtask aims to classify speech utterances into their respective country-level dialect categories automatically. Our approach leverages large-scale

Proceedings of The Third Arabic Natural Language Processing Conference, pages 762–766

pre-trained speech representations and a two-stage fine-tuning process to effectively adapt to the dialectal nuances present in the provided dataset. In the following subsections, we describe the datasets used, our ADI model architecture, and the considered training strategy. We also present our experimental results and follow up with an analysis of the ADI system performances.

2.1 Datasets

We utilize several datasets to train and evaluate our ADI system, including established corpora covering multiple Arabic dialects, as well as the official NADI 2025 ADI dataset. The following is a detailed description of each dataset.

2.1.1 ADI-17 and ADI-20

The ADI-17 dataset (Shon et al., 2020) comprises 3,033 hours of dialectal Arabic speech from 17 country-level dialects for training, along with approximately 2 hours per dialect in the development and test splits, respectively.

The ADI-20 dataset (Elleuch et al., 2025) is an expanded and rebalanced version of ADI-17, extending its coverage from 17 to 20 Arabic varieties by including Tunisian and Bahraini dialects as well as Modern Standard Arabic (MSA). It also increases representation for previously underrepresented dialects, such as Jordanian and Sudanese, by incorporating additional speech material. In total, the training partition contains 3,556 hours of speech, while the development and test sets retain the same structure as in ADI-17, supplemented with approximately 2 hours per newly added variety in each split. To enable experiments under resource-constrained conditions and ensure per-dialect balance, ADI-20-53h, a stratified subset containing up to 53 hours of training data for each variety, resulting in a total of 1,060 hours is also available. Our future experiments will use this subset rather than the full ADI-20 dataset for the reasons mentioned earlier.

2.1.2 NADI 2025 ADI Dataset

The official dataset for the ADI subtask covers eight country-level Arabic dialects: Algeria, Egypt, Jordan, Mauritania, Morocco, Palestine, UAE, and Yemen. It includes an *adaptation* split of approximately 15 hours (12,900 utterances) with associated country labels, a validation split of similar size (12,700), and an eleven-hour held-out test set with 6268 utterances.

2.2 Model Architecture

Our system follows the best-performing configuration from Elleuch et al. (2025). The Whisper-large-v3 encoder (Radford et al., 2023) is used as a feature extractor, followed by an attention pooling layer that aggregates frame-level representations into fixed-length utterance embeddings. These are passed through a fully connected layer with a softmax activation for classification over the target dialects.

We freeze the first 16 layers of the Whisper encoder during fine-tuning to preserve general speech representations while adapting the upper layers to the ADI task. To enhance robustness, we apply additive noise, speed perturbation, frequency masking, and chunk-level dropout. Training is performed with SpeechBrain (Ravanelli et al., 2024) using negative log-likelihood loss, the Adam optimizer, and a NewBob learning rate scheduler starting from 1×10^{-5} for frozen encoder layers and 1×10^{-4} for trainable layers. Training runs for up to 100 epochs on NVIDIA H100 80GB GPUs, with early stopping based on validation performance.

2.3 Experiments and Results

We first evaluated the model after fine-tuning only on ADI-17 and ADI-20-53h to assess zero-shot performance on the NADI validation set. As shown in Table 1, fine-tuning on ADI-17 yields an accuracy of 31.84%, while ADI-20-53h substantially improves zero-shot accuracy to 78.33%.

Fine-tuning dataset	Accuracy (%)
ADI-17	31.84
ADI-20-53h	**78.33**

Table 1: Zero-shot evaluation on the NADI 2025 ADI validation set.

Our final submission builds on the ADI-20-53h model, further adapted with the NADI adaptation split. This two-stage fine-tuning yields substantial gains, as shown in Table 2. The system ranked first, achieving 98.08% accuracy on validation and 79.83% on the test set, with corresponding average costs of 0.0171 and 0.1788 using the 2022 NIST LRE formulation.

Analysis of the validation confusion matrix in 1 shows that the Algerian dialect is the most challenging to predict, with only 96% of utterances correctly classified. Misclassifications primarily

involve the geographically adjacent Moroccan dialect, and conversely, 30 Moroccan utterances are labeled as Algerian. Misclassifications between Egyptian and Jordanian are largely reciprocal; despite their geographic proximity, this pattern is unexpected from the perspective of Arabic speakers.

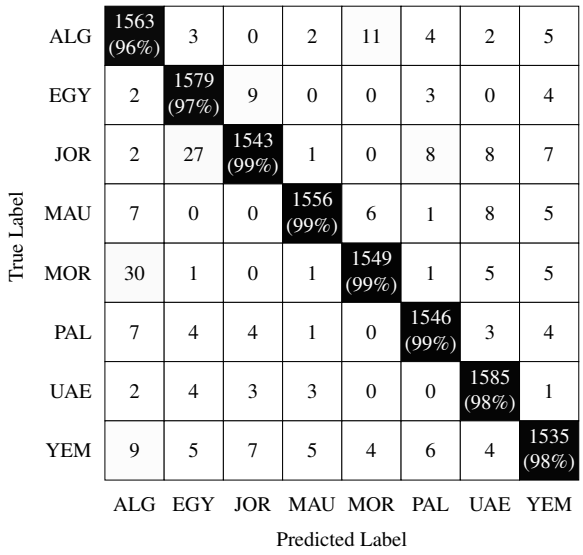

Figure 1: Confusion matrix on the provided development set.

Split	Accuracy (%) ↑	LRE avg. Cost ↓
Validation	98.08	0.0171
Test	79.83	0.1788

Table 2: Final ADI subtask results.

3 Multi-dialectal Arabic ASR

The multi-dialectal Automatic Speech Recognition (ASR) subtask focuses on transcribing spoken Arabic across eight country-level dialects. This subtask aims to highlight the challenges posed by phonetic, lexical, and syntactic diversity of Arabic dialects. In this section, we describe the dataset, model architecture, training methodology, and the obtained results of our approach.

3.1 Dataset

The NADI 2025 ASR dataset includes the same eight dialects as the ADI subtask. Each dialect has 1,600 utterances in both the training and validation splits, with durations ranging from 1 to 30 seconds. The total duration is 30.72 hours (15.44 hours for training, 15.27 for validation). Table 3 shows per-dialect durations.

Dialect	Train (h)	Validation (h)
Algeria	1.91	1.84
Egypt	2.01	1.85
Jordan	1.93	1.89
Mauritania	1.66	1.63
Morocco	1.60	1.67
Palestine	2.43	2.41
UAE	1.87	1.86
Yemen	2.01	2.11
Total	**15.44**	**15.27**

Table 3: Durations per dialect in the NADI 2025 datasets

3.2 Models

We adopted two distinct architectures in our experiments: Whisper and SeamlessM4T-v2 (Barrault et al., 2023). Whisper is an encoder–decoder Transformer model trained on a large-scale multilingual and multitask dataset of speech and text, enabling robust automatic speech recognition (ASR) across a wide range of languages. Its architecture integrates a Transformer-based encoder for speech representation learning and a Transformer decoder for transcription generation. The Whisper-large-v3 model contains approximately 1.55 billion parameters, while Whisper-medium has around 769 million parameters, offering a faster and more memory-efficient alternative.

SeamlessM4T is a multilingual sequence-to-sequence model designed for speech and text translation across more than 100 languages. In its v2 release, it builds upon the UnitY2 architecture, combining a Conformer-based speech encoder with a Transformer-based text decoder. We selected the Egyptian variant due to its demonstrated effectiveness in Arabic transcription tasks. Given the substantial phonetic, lexical, and syntactic divergence between Arabic dialects, we empirically found that fine-tuning a separate model for each dialect outperformed a single unified model for all dialects.

3.3 Experiments and Results

Our experimental process for the multi-dialectal ASR subtask followed three main steps: (i) evaluation of Whisper-based systems, (ii) comparison between per-dialect and unified models, (iii) comparison of the best Whisper model with SeamlessM4T-v2 Large. All results are reported in terms of WER and CER, computed on the NADI 2025 validation and test sets.

For Whisper-based experiments, we fine-tuned different variants (Medium and Large) under two hardware configurations: (i) on an NVIDIA P100 16GB GPU with the AdamW optimizer, a fixed learning rate of 1×10^{-5}, a batch size of 1, and gradient accumulation over 4 steps; and (ii) on an NVIDIA A100 80GB GPU with a batch size of 8 using the same optimizer and learning rate.

For SeamlessM4T, we fine-tuned the v2 Large Egyptian variant for six epochs on an NVIDIA A100 40GB GPU. Training employed the AdamW optimizer with a learning rate warmed up over 100 steps from 1×10^{-9} to 5×10^{-5}. We used label-smoothed negative log-likelihood loss with a smoothing factor of 0.2 and a batch size of 2.

3.3.1 Whisper Large vs Whisper Medium

We first fine-tuned both Whisper-large-v3 and Whisper-medium on the full multi-dialectal dataset (all eight dialects combined). On average, Whisper-large-v3 achieved a WER of 72.20% and a CER of 58.51% while Whisper-medium, despite its smaller size, outperformed it with a WER of 48.21% and a CER of 17.94%. Given these substantial improvements, Whisper-medium was chosen for all subsequent experiments.

3.3.2 Multi vs Mono-dialectal Models

We evaluated two training strategies using Whisper-medium: a multi-dialectal model trained jointly on all dialects, and mono-dialectal models obtained via dedicated fine-tuning for each dialect, yielding eight specialized models. The mono-dialectal approach achieved a lower average Word Error Rate (WER) of 46.71%, compared to 48.21% for the multi-dialectal model. In terms of Character Error Rate (CER), both approaches performed similarly, with the multi-dialectal model scoring 17.97% and the specialized models 17.94% on average. These results suggest that training separate mono-dialectal models is the best approach.

3.3.3 Whisper vs SeamlessM4T-v2 Large

We also compared the best Whisper setup (one specialized whisper-medium system per-dialect) with the SeamlessM4T-v2 Large Egyptian model (Barrault et al., 2023) also fine-tuned separately for each dialect. Due to time constraints, only the large variant was considered for our experiments. Table 4 shows that the Seamless-based system consistently outperforms our best Whisper system for all dialects in both WER and CER.

	Seamless	**Whisper-med.**
Dialect	**WER / CER (%)**	**WER / CER (%)**
Jordan	**25.26 / 7.68**	32.53 / 9.93
Egypt	**30.05 / 12.52**	39.38 / 15.97
Morocco	**39.24 / 13.48**	49.22 / 18.34
Algeria	**54.13 / 19.34**	60.61 / 22.41
Yemen	**50.49 / 16.85**	61.28 / 25.54
Mauritania	**56.93 / 23.91**	62.79 / 26.97
UAE	**30.90 / 10.59**	35.38 / 12.48
Palestine	**26.35 / 9.64**	32.51 / 12.11
Average	**39.17 / 14.25**	46.71 / 17.97

Table 4: SeamlessM4T-v2 Large vs. Whisper-medium WER and CER on the validation sets of each NADI 2025 dialect using one fine-tuned model per dialect.

3.3.4 Official Submission

Based on the validation results presented in table 4, we selected the SeamlessM4T-v2 Large per-dialect models for submission. It ranked second overall, with an average WER 38.54% and CER 14.53% on the test set. Table 5 shows the results per dialect. As it can be seen, performance varied notably across dialects, with Levantine and Egyptian achieving the lowest WERs, while Maghrebi dialects remained the most challenging.

Dialect	**WER (%)**	**CER (%)**
Jordan	28.03	9.36
Egypt	26.83	11.44
Morocco	38.27	13.66
Algeria	53.73	20.43
Yemen	46.63	16.66
Mauritania	58.11	24.53
UAE	29.35	9.91
Palestine	27.36	10.20

Table 5: WER and CER on the NADI 25 test sets.

4 Conclusion

This paper presented the ELYADATA–LIA submissions to the NADI 2025 shared task, addressing both the Arabic Dialect Identification and Multi-dialectal Automatic Speech Recognition subtasks. For ADI, we demonstrated the effectiveness of a two-stage fine-tuning approach using the Whisper-large-v3 encoder, achieving first place with 79.83% accuracy on the test set. For ASR, fine-tuning the SeamlessM4T-v2 Large model separately for each dialect resulted in a strong performance, ranking second on the leaderboard with an average WER of 38.54%.

Acknowledgments

This work was partially funded by the ESPERANTO project. The ESPERANTO project has received funding from the European Union's Horizon 2020 (H2020) research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666. This work was granted access to the HPC resources of IDRIS under the allocations AD011015051R1 and AD011012551R3 made by GENCI.

References

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. Arabic automatic speech recognition: A systematic literature review. *Applied Sciences*, 12(17).

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. Adi-20: Arabic dialect identification dataset and models. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Rotterdam Ahoy Convention Centre, Rotterdam, The Netherlands. To appear.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

James Lane. 2025. The 10 most spoken languages in the world in 2025. https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world. Accessed: 2025-08-12.

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.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research*.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. Nadi 2025: The first multidialectal arabic speech processing shared task. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.