MarsadLab at NADI Shared Task: Arabic Dialect Identification and Speech Recognition using ECAPA-TDNN and Whisper

Md. Rafiul Biswas[1], **Kais Attia**[2], **Shimaa Ibrahim**[3],
Mabrouka Bessghaier[3], **Wajdi Zaghouani**[3]
[1]Hamad Bin Khalifa University, Qatar, [2]Independent Researcher, Tunisia
[3]Northwestern University in Qatar, Qatar
mbiswas@hbku.edu.qa,wajdi.zaghouani@northwestern.edu

Abstract

We participated in NADI 2025 shared tasks on Arabic Dialect Identification (ADI) and Automatic Speech Recognition (ASR) across eight Arabic dialects. For ADI, we employ an enhanced ECAPA-TDNN with VoxLingua107 initialization, featuring self-attention classification head, progressive unfreezing, advanced augmentation, and test-time augmentation. This approach ranked third with 61.6% accuracy and 0.3068 macro cost. For ASR, we implement a zero-shot cascaded system using Whisper Large-v3 and MARBERT with extreme parameter efficiency (0.0004% trainable), ranking seventh with 104.895 WER and 84.693 CER. Our results validate complementary paradigms: direct audio processing for competitive dialect classification versus foundation model robustness for cross-dialectal transcription.

1 Introduction

Arabic is a pluricentric language with a rich continuum of regional and social varieties. This diversity—spanning Egyptian, Levantine, Gulf, Maghrebi, and other dialect groupings alongside Modern Standard Arabic (MSA)—poses unique challenges for speech technologies (Rahman et al., 2024). Despite steady progress in speech processing, reliable recognition and identification of Arabic dialects from speech remains difficult due to limited labeled resources, frequent code-switching with MSA and other languages, and substantial phonetic and lexical variation (Biadsy et al., 2009). Earlier shared tasks on spoken dialect identification helped define the problem space and catalyze benchmarking (Ali et al., 2017, 2019) while recent large-scale models that jointly learn ASR and language identification—such as Whisper (Tang et al., 2022; Radford et al., 2022) and MMS (Pratap et al., 2023) have reset expectations for zero-/few-shot performance. Still, their effectiveness on multidialectal Arabic, especially under domain shift and fine-grained dialect labels, is far from settled (Aboelela and Mansour, 2025).

The NADI 2025 shared task (Talafha et al., 2025) addresses two complementary problems: fine-grained dialect identification from single utterances and robust ASR across dialects using the Casablanca dataset. Building on prior Arabic shared tasks and benchmarks (e.g., MGB-3 (Ali et al., 2017), MGB-2 (Ali et al., 2019)), we adopt two complementary system designs: (1) an adaptation-heavy ECAPA-TDNN (Desplanques et al., 2020a) pipeline for dialect classification and (2) a zero-shot Whisper Large baseline for ASR. Our design choices emphasize reproducibility and computational practicality while exploring methods that improve dialect discrimination and transcription robustness. Our proposed system model using Whisper and MMS dataset demonstrates the power of large-scale multilingual models. community-driven effort to advance multidialectal Arabic speech recognition, while Speechbrain (Ravanelli et al., 2021a), VoxLingua107 (Valk and Alumäe, 2021) and ECAPA-TDNN (Desplanques et al., 2020a) provide crucial multilingual and architectural foundations.

Our contributions are threefold: (i) a practical and reproducible Arabic ASR that is based on ECAPA TDNN that features the self-attention mechanism; (ii) an empirical study of the use of OpenAI Whisper Large v3 in Casablanca for dialect-specific transcription; and (iii) a transparent analysis of errors and per-dialect behavior to inform future multidialectal modeling.

2 Background

NADI subtasks uses Casablanca audio corpus covering eight target dialects (Algerian, Egyptian, Jordanian, Mauritanian, Moroccan, Palestinian, Emirati, Yemeni)(Talafha et al., 2024). Each input is a single-channel WAV file carrying one utter-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 752–756
November 8-9, 2025 ©2025 Association for Computational Linguistics

ance; ADI expects a single dialect label output and ASR expects a text transcription (MSA or dialectal Arabic depending on the speaker). Table 1 presents the NADI 2025 Arabic dialect dataset comprising 25,600 audio samples across 8 Arabic dialects. The dataset is well-balanced with each dialect containing exactly 3,200 samples, split nearly evenly between training (12,900) and validation (12,700) sets. Audio recordings are sampled at 16 kHz with durations ranging from 1.04 to 15.12 seconds (mean: 4.25s, median: 3.56s, std: 2.79s). An additional 6,268 unlabeled test samples are provided for evaluation.

Figure 1 illustrates the audio characteristics analysis of the dataset. The left panel shows the distribution of audio durations, revealing a right-skewed distribution with most samples concentrated between 2-4 seconds, and the mean (4.3s) slightly higher than the median (3.6s) due to longer outliers. Dialects exhibit similar interquartile ranges and median values around 3-4 seconds. Both visualizations confirm the dataset's consistency and balance, making it suitable for robust Arabic dialect identification model training and evaluation.

Dialect	**Train**	**Val**
Algeria	1,610	1,590
Egypt	1,603	1,597
Jordan	1,604	1,596
Mauritania	1,617	1,583
Morocco	1,608	1,592
Palestine	1,631	1,569
UAE	1,602	1,598
Yemen	1,625	1,575
Total	**12,900**	**12,700**
Dataset Overview		
Dialects: 8	Total: 25,600	Test: 6,268
Sampling rate: 16 kHz		
Audio Duration Statistics (seconds)		
Mean: 4.25	Median: 3.56	Std: 2.79
Range: 1.04 – 15.12		

Table 1: Dialectal distribution of NADI dataset

Task Challenges: Prior Arabic speech work demonstrates recurring challenges: dialectal variation, scarcity of labeled data for many dialects, and domain mismatch between broadcast and in-the-wild audio (Ali et al., 2017; Althobaiti, 2020). Recent multilingual foundation models (Whisper (Radford et al., 2022), MMS (Pratap et al., 2023)) show strong zero-shot generalization, while architectures such as ECAPA-TDNN have been effective for representation extraction in speaker and language tasks (Desplanques et al., 2020b). For implementation and tooling we relied on the SpeechBrain toolkit (Ravanelli et al., 2021b).

3 System Overview

We implemented two systems consistent with the memorized process described earlier. Below we summarize the main design choices and components for each subtask.

3.1 Subtask 1: Dialect Identification (ECAPA-TDNN pipeline)

Base architecture: ECAPA-TDNN pre-trained and described in prior work (Desplanques et al., 2020b). We adapt ECAPA as a robust embedding extractor and add a classification pathway on top.

Classification head: Custom multi-layer MLP with Swish activation, BatchNorm, dropout, and a feature-wise self-attention module. The attention reweights ECAPA feature vectors:

$$\begin{aligned} \mathbf{a} &= \sigma(\mathbf{W}_2 \cdot \mathrm{Swish}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2)\,, \\ \hat{\mathbf{h}} &= \mathbf{a} \odot \mathbf{h}, \end{aligned} \tag{1}$$

Training schedule:

- Phase 1: Freeze ECAPA backbone; train classifier head (2,500 steps).
- Phase 2: Unfreeze top ECAPA layers; fine-tune with discriminative learning rates ($\eta_{\text{encoder}} = 1 \times 10^{-6}, \eta_{\text{classifier}} = 5 \times 10^{-5}$).

Loss & regularization: Combined loss $\mathcal{L} = 0.3\mathcal{L}_{\text{focal}} + 0.7\mathcal{L}_{\text{CE}}$ (focal $\gamma = 2.5$), label smoothing, gradient clipping, and cosine-annealing LR with warmup.

Augmentation & inference: Advanced augmentation pipeline (noise, pitch/time perturbations, reverb, volume, frequency/time masking) during training. At inference we applied Test-Time Augmentation (TTA) with 5–10 variants per utterance and averaged softmax outputs; temperature scaling was used for calibration.

3.2 Subtask 2: Automatic Speech Recognition (MARBERT-Whisper pipeline)

We employ OpenAI Whisper Large-v3 (via Hugging Face `pipeline("automatic-speech-recognition")`) as our baseline (Radford et al., 2022). Our approach implements a cascaded architecture for Arabic Dialect Identification (ADI), combining ASR with text classification through parameter-efficient transfer learning.

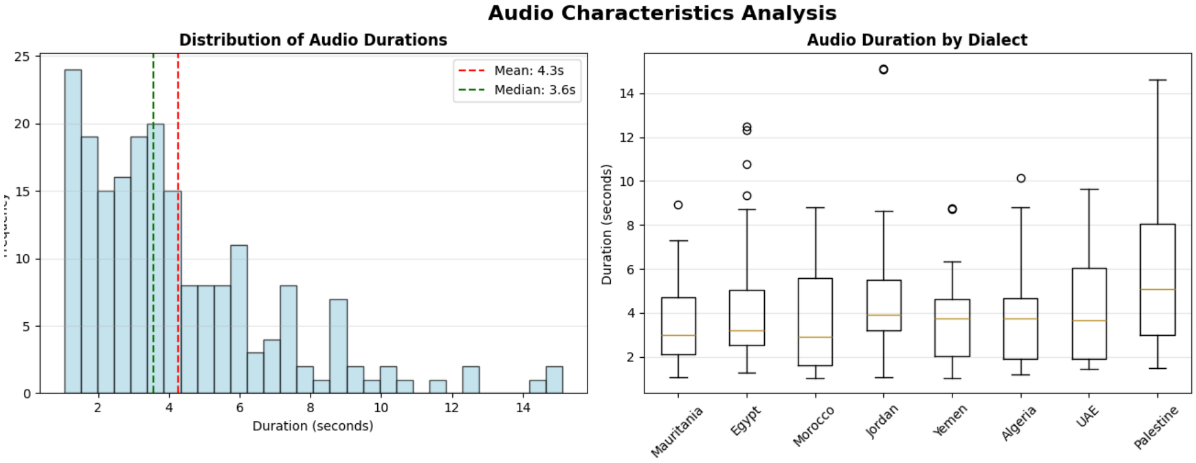

Figure 1: Statistical distribution of audio duration in NADI dataset

Given an input audio signal $\mathbf{x} \in \mathbb{R}^T$ of length T, the system performs sequential transformations:

1. **ASR:** Whisper Large-v3 for speech-to-text.
2. **Encoding:** MARBERT for contextual text embeddings.
3. **Classification:** Trainable linear layer for dialect prediction.

The speech-to-text step uses a frozen Whisper model Φ_{whisper}:

$$t = \Phi_{\text{whisper}}(\mathbf{x}; \boldsymbol{\theta}_{\text{whisper}}), \tag{2}$$

where t is the transcript and $\boldsymbol{\theta}_{\text{whisper}}$ are frozen pretrained parameters.

The transcript t is processed by the frozen MARBERT encoder Φ_{MARBERT}:

$$\mathbf{h} = \Phi_{\text{MARBERT}}(t; \boldsymbol{\theta}_{\text{MARBERT}}), \tag{3}$$

where $\mathbf{h} \in \mathbb{R}^{768}$ is the [CLS] token embedding.

A trainable classifier maps $\mathbf{h}$ to dialect probabilities:

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}), \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{8\times768}$ and $\mathbf{b} \in \mathbb{R}^8$ are the only trainable parameters.

Audio is resampled to 16 kHz mono, truncated at 30 s, and zero-padded. Text is tokenized with MARBERT (max length 512, dynamic padding). Training uses batch-mode transcript processing; inference is sequential with error handling.

This cascaded design achieves $\mathcal{O}(T \log T)$ complexity for ASR and $\mathcal{O}(L^2)$ for encoding, with minimal overhead due to selective parameter updates.

Component	Task 1: ADI	Task 2: ASR
Framework	SpeechBrain	HF Transformers
Pretrained	ECAPA-TDNN	Whisper Large + MARBERT
Optimizer	AdamW	AdamW
Batch size	32	8 (train), 4 (val)
Precision	FP16	FP16
Augmentation	Audio perturb.	None
Learning rate	5e-5	2e-5
Trainable params	Enhanced classifier	6,152 (0.0004%)
Max steps	25,000	3,000
Hardware	8 GB+ GPU	8 GB+ GPU

Table 2: Training configurations for ADI and ASR tasks

4 Experimental Setup

All experiments used the organizer-provided splits (Table 1). Implementations used SpeechBrain for ECAPA-based pipelines and Hugging Face Transformers for Whisper. Important implementation details are summarized in Table 2.

Metrics and evaluation. For ADI we report accuracy and the macro-averaged cost metric provided by the organizers. For ASR we report average WER and CER using the Codabench evaluation script. Recent large-scale approaches and multilingual systems motivate the use of zero-shot baselines for comparison (Pratap et al., 2023; Radford et al., 2022).

5 Results

Table 3 summarizes official results submitted to the organizers and used for official ranking. The enhanced ECAPA-TDNN system achieved a competition score of 0.616 (cost: 0.3068) in Task 1, demonstrating competitive performance against the best system which scored 0.7983 (cost: 0.1788),

Task	Metric 1	Metric 2	Rank
ADI	Acc. 0.616	Macro Cost 0.3068	3
ASR	Avg. WER 104.90	Avg. CER 84.69	7

Table 3: Performance metrics of our proposed system

validating the effectiveness of direct audio processing for Arabic dialect identification.

For Task2, the novel Whisper + MARBERT cascaded approach, while achieving more modest accuracy, offers significant advantages in computational efficiency and interpretability, requiring only minimal parameter training while leveraging the power of large pre-trained models.

5.1 Ablation and analysis (validation splits)

We performed ablations during development on the validation set. Removing the feature-wise attention layer reduced validation discrimination between similar dialect classes and led to decreased stability in low-resource dialects (consistent with our informal validation runs). Progressive unfreezing and discriminative learning rates helped preserve pretrained representations and improved final validation cost.

5.2 Error analysis

We analyzed common confusions on validation and test samples (explicitly noting which split is used where):

- **Dialect confusions:** Moroccan and Algerian Arabic sound very similar in how they're spoken (rhythm/melody) and use similar words/expressions. The same applies to Levantine and Palestinian Arabic. When these linguistic features "overlapped" (were very similar between the pairs), the AI system couldn't reliably distinguish between them.

- **ASR errors:** Whisper zero-shot produced frequent errors in colloquial and code-switched segments (e.g., mixing Arabic and French terms), and often omitted short function words or mis-transcribed named entities.

Example (validation): a Moroccan utterance containing dialectal lexical items was misclassified as Algerian due to shared lexical forms and similar rhythm; manual inspection revealed low SNR and overlapped background speech.

6 Discussion

Our NADI 2025 participation reveals several critical limitations and areas for improvement across both tasks. For Task 1 (ADI), our enhanced ECAPA-TDNN system achieved an accuracy of 0.616 with macro cost of 0.3068, ranking 3rd among participants, compared to the best performing system at 0.7983, indicating substantial room for optimization in fine-tuning strategies and feature extraction despite our sophisticated enhancement techniques including self-attention mechanisms, progressive unfreezing, and advanced data augmentation.

The cascaded approach in Task 2 (ASR) exposed fundamental limitations of speech-to-text pipelines, achieving an average WER of 104.90 and CER of 84.69, ranking 7th in the competition. These high error rates reflect domain mismatch between Whisper's training data and the competition dataset, as well as differences in transcription conventions and dialectal variations that the pre-trained model was not optimized for. Error propagation from the ASR component directly impacts downstream classification performance, as dialectal acoustic features crucial for identification are lost during transcription. This suggests that preserving prosodic and phonetic information through direct audio processing remains superior for dialect-specific tasks.

The limited training data for certain dialect classes exacerbated class imbalance issues in Task 1, despite employing focal loss and data augmentation techniques, while the extremely high error rates in Task 2 suggest fundamental challenges in adapting general-purpose ASR models to dialectal Arabic. Future improvements should focus on dialectal data augmentation strategies, cross-lingual transfer learning from related Arabic varieties, hybrid architectures that combine acoustic and linguistic features for ADI, and specialized ASR models trained specifically on dialectal Arabic corpora.

7 Conclusion

In summary, our experiments presents the complementary strengths of two paradigms: fine-tuned ECAPA-TDNN, augmented with diverse perturbations and targeted architectural refinements, delivers strong dialect classification, whereas Whisper Large serves as a capable zero-shot transcription baseline across dialects without any task-specific adaptation. This contrast suggests a promising avenue in combining the adaptability of tailored acoustic models with the broad coverage of large, general-purpose ASR systems.

Code Reproducibility

To ensure reproducibility of our results, all source code, model implementations, and experimental configurations are made publicly available at `https://github.com/rafiulbiswas/NADI`. The repository includes complete implementations for both tasks with detailed documentation and setup instructions.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Eman Aboelela and Omar Mansour. 2025. A review of speech recognition and application to arabic speech recognition. In *Future of Information and Communication Conference*, pages 13–31. Springer.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. *Preprint*, arXiv:1609.05625.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. *Preprint*, arXiv:1709.07276.

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *Preprint*, arXiv:2009.12622.

Fadi Biadsy, Julia Bell Hirschberg, and Nizar Y Habash. 2009. Spoken arabic dialect identification using phonotactic modeling.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020a. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020b. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Ashifur Rahman, Md Mohsin Kabir, Muhammad Firoz Mridha, Mohammed Alatiyyah, Haifa F Alhasson, and Shuaa S Alharbi. 2024. Arabic speech recognition: Advancement and challenges. *IEEE Access*, 12:39689–39716.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021a. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021b. Speechbrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624.

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.

Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture, and Jimmy Lin. 2022. Speechnet: Weakly supervised, end-to-end speech recognition at industrial scale. *arXiv preprint arXiv:2211.11740*.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.