Lahjati at NADI 2025: A ECAPA-WavLM Fusion with Multi-Stage Optimization

Sanad Albawwab[1] **Omar Qawasmeh**[2]

[1]Knowledge Technologies Department, Applied AI Division,
Royal Scientific Society, Amman, Jordan
[2]Data Science Department, Princess Sumaya University for Technology, Amman, Jordan
Correspondence: sanad.bawwab@rss.jo, o.alqawasmeh@psut.edu.jo

Abstract

This paper presents Lahjati (ECAPA-WavLM Fusion with Multi-Stage Optimization) system for the spoken Arabic Dialect Identification (ADI) subtask at Nadi 2025 (Talafha et al., 2025), The task aims to automatically identify the dialect of spoken Arabic utterances, a challenging problem due to the rich linguistic diversity of Arabic and the scarcity of labeled speech resources. Our approach combines ECAPA-TDNN embeddings from SpeechBrain with WavLM-base representations. The proposed system achieved 94.08% accuracy on the validation set and ~51.0% on the test set. Challenges included differentiating acoustically similar dialect pairs and mitigating the effects of varied recording conditions, which likely contributed to performance degradation on unseen data. These findings highlight both the potential and limitations of fusing complementary speech representations for robust dialect identification.

1 Introduction

Arabic dialect identification from speech presents a significant challenge due to the extensive linguistic diversity across the Arab world and the scarcity of large, high-quality labeled datasets. These factors considerably hinder the development and optimization of speech technology applications. The task addressed here involves recognizing speech from eight distinct dialectal varieties: Palestinian, Yemeni, Mauritanian, Algerian, Moroccan, Jordanian, Egyptian, and Emirati Arabic. Achieving accurate identification of these dialects is essential for enhancing a wide range of downstream speech-processing technologies, including automatic speech recognition, machine translation, and conversational systems that can adapt effectively to regional linguistic variations.

This work builds upon the foundation laid by the Nuanced Arabic Dialect Identification (NADI) shared task, first introduced in 2020 (Abdul-Mageed et al., 2020), which provided standardized benchmark datasets and evaluation protocols for country-level Arabic dialect classification . The NADI initiative not only unified fragmented research efforts in this domain but also established a baseline for systematic comparison of approaches, paving the way for more fine-grained and context-aware dialect identification systems (Abdul-Mageed et al., 2021; Abdul-Mageed et al., 2022, 2023, 2024). By leveraging such frameworks and addressing current data limitations, this task aims to push the boundaries of robust, real-world Arabic speech dialect identification.

Main System Strategy

Our proposed system, **ECAPA_WavLM_Fusion**, integrates two complementary pretrained speech encoders to jointly capture *speaker-level* and *contextual acoustic* features, enabling robust Arabic dialect identification across eight dialect classes.

- **Speaker-Level Encoder:** The first component is **ECAPA-TDNN**, initialized from SpeechBrain's VoxLingua107 model, which generates 256-dimensional speaker embeddings directly from raw waveforms. This encoder excels at modeling speaker-specific timbre and prosodic characteristics, which are often correlated with dialectal traits.

- **Contextual Acoustic Encoder:** The second component is **WavLM-base** from Microsoft, a transformer-based model that produces 768-dimensional frame-level contextual embeddings. These embeddings are mean-pooled over time to obtain utterance-level representations, then linearly projected to 256 dimensions to match the ECAPA-TDNN feature space.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 740–744

The outputs from both encoders are concatenated to form a 512-dimensional fused representation, which is `LayerNorm`-normalized and passed through a two-layer feedforward classifier with dropout regularization for final dialect prediction.

Training follows a two-stage fine-tuning strategy:

1. **Stabilization phase:** Encoder weights are frozen for the first 8,000 steps, allowing the classifier layers to learn stable decision boundaries from fixed embeddings.

2. **Full fine-tuning phase:** All parameters are unfrozen, and training continues with a multi-phase learning rate scheduler consisting of linear warmup, constant hold, and cosine annealing.

We optimize with **AdamW**, incorporating weight decay for regularization and gradient clipping to mitigate exploding gradients.

By combining *speaker-discriminative* and *context-aware* representations, the **ECAPA_WavLM_Fusion** architecture effectively captures subtle phonetic and prosodic cues that differentiate closely related Arabic dialects, mitigating challenges posed by intra-dialect similarities. Code and pretrained models are available at our GitHub repository [1].

2 Background

This task addresses **Arabic dialect identification** directly from raw speech waveforms. The input consists of 16 kHz audio clips containing spoken Arabic from **eight dialects**: Palestinian, Yemeni, Mauritanian, Algerian, Moroccan, Jordanian, Egyptian, and Emirati. The system outputs a predicted dialect label corresponding to one of these classes. For example, given a short audio excerpt from a television program, the model must determine whether the speech is Egyptian, Moroccan, or one of the other target dialects.

The dataset used in this work comprises **high-quality multidialectal Arabic speech recordings** sampled at 16 kHz. It contains approximately **12,900 training samples** (∼8 hours of speech) and **12,700 validation samples** (∼8 hours), totaling around **16 hours of labeled audio**. An additional **8-hour blind test set** is provided for final evaluation.

A qualitative examination of the audio reveals that many clips are drawn from diverse media sources such as television dramas, movies, and talk shows, similar to the Casablanca dataset.(Talafha et al., 2024). This diversity introduces *natural conversational speech* with a wide range of acoustic conditions—including variations in background noise, recording quality, and speaker expressiveness—thereby creating a realistic and challenging benchmark for dialect classification.

Each audio sample is annotated with one of the eight target dialect labels, covering a spectrum of speech genres and speaker demographics. This diversity helps improve the robustness and generalization capabilities of trained models, making them more applicable to real-world settings.

Our system was developed for the **Spoken Arabic Dialect Identification (ADI)** track of the **Nuanced Arabic Dialect Identification (NADI)** shared task, which offers standardized datasets, clear evaluation protocols, and a competitive benchmarking platform for advancing research in fine-grained Arabic dialect recognition.

3 System Overview

Our system, `Lahjati (ECAPA_WavLM_Fusion)`, leverages two complementary pretrained speech encoders to jointly capture *speaker-discriminative* and *context-aware* acoustic representations for Arabic dialect identification. This design aims to exploit both timbre/prosody cues (often linked to speaker identity and dialect) and broader contextual speech patterns for robust classification.

Key Architecture and Algorithms: The architecture integrates:

- **ECAPA-TDNN encoder** (Desplanques et al., 2020), initialized from the SpeechBrain VoxLingua107 model, which extracts 256-dimensional speaker embeddings from raw audio waveforms.

- **WavLM-base encoder** (Chen et al., 2022), a transformer-based model producing 768-dimensional contextual embeddings from frame-level speech representations, subsequently mean-pooled to form utterance-level features.

Both embeddings are linearly projected to 256 dimensions, concatenated into a unified 512-dimensional vector, normalized with `LayerNorm`, and passed through a two-layer feedforward neural

[1] https://github.com/sanadbawab0/nadi2025/

network with dropout regularization to predict one of eight target dialect classes.

The core forward computation of `Lahjati` can be formulated as:

$$\begin{aligned}
\mathbf{e}_{\text{ecapa}} &= \text{ECAPA-TDNN}(x), \\
\mathbf{e}_{\text{wavlm}} &= \text{meanpool}(\text{WavLM}(x)), \\
\mathbf{h}_{\text{ecapa}} &= W_{\text{ecapa}}\,\mathbf{e}_{\text{ecapa}} + b_{\text{ecapa}}, \\
\mathbf{h}_{\text{wavlm}} &= W_{\text{wavlm}}\,\mathbf{e}_{\text{wavlm}} + b_{\text{wavlm}}, \\
\mathbf{h} &= \text{LayerNorm}\big([\mathbf{h}_{\text{ecapa}}; \mathbf{h}_{\text{wavlm}}]\big), \\
\hat{y} &= \text{Classifier}(\mathbf{h}),
\end{aligned} \tag{1}$$

where x is the input audio waveform and $\hat{y}$ is the predicted dialect label. The intermediate representations are defined as follows: $\mathbf{e}_{\text{ecapa}} \in \mathbb{R}^{256}$ is the ECAPA-TDNN speaker embedding, $\mathbf{e}_{\text{wavlm}} \in \mathbb{R}^{768}$ is the pooled WavLM contextual embedding, $\mathbf{h}_{\text{ecapa}}, \mathbf{h}_{\text{wavlm}} \in \mathbb{R}^{256}$ are the respective projected embeddings, and $\mathbf{h} \in \mathbb{R}^{512}$ is the fused representation after concatenation and normalization.

We employed pretrained `ECAPA-TDNN` and `WavLM-base` encoders, both initially frozen to exploit their rich acoustic representations while mitigating the risk of overfitting on the limited dialectal dataset. All experiments were conducted on the NADI 2025 dataset `UBC-NLP/NADI2025_subtask1_SLID` (UBC-NLP, 2025), available via Hugging Face.

Staged Fine-tuning Strategy: To address data scarcity and substantial dialectal overlap, we adopted a two-phase training procedure: (i) for the first 8,000 steps, encoder weights were frozen to allow the classifier to adapt to fixed embeddings; (ii) all parameters were then unfrozen, enabling joint fine-tuning with a multi-phase learning rate schedule (linear warmup, constant hold, cosine annealing). This approach balances early training stability with later model adaptability.

Training Pipeline: The training process follows five sequential steps: (1) raw audio is processed in parallel by ECAPA-TDNN and WavLM encoders; (2) embeddings are projected to a common 256-dimensional space, concatenated, and normalized; (3) a two-layer feedforward classifier produces logits for the eight target dialect classes; (4) cross-entropy loss is used for optimization; and (5) learning rate scheduling and gradient clipping are applied to ensure stable convergence.

Experimental Configurations: We compared training durations of 50,000 and 100,000 steps. Extending training to 100,000 steps improved validation accuracy by approximately $+2\%$, underscoring the benefits of prolonged fine-tuning for this task.

3.1 Experimental Setup

Data Splits We used the official splits released by the NADI 2025 organizers without modification. For Subtask 1 (SLID), we employed the `UBC-NLP/NADI2025_subtask1_SLID` dataset, comprising

- **Training:** 12,900 samples
- **Validation:** 12,700 samples

For the Subtask 1 (ADI) test phase, we used the `UBC-NLP/NADI2025_subtask1_ADI_Test` set containing 6,268 samples. No external data was incorporated.

Data Format Each instance consists of an audio recording and its label. In the SLID dataset, fields include `id`, `audio`, and `country`. In the ADI test set, fields include `id` and `audio`, where `audio` is stored as an array of float values along with the sampling rate.

Preprocessing Audio waveforms were loaded at their original sampling rate and resampled to 16 kHz using the `AutoFeatureExtractor` from the Hugging Face `microsoft/wavlm-base` model. Dialect labels were mapped to integer IDs via:

$$\text{labels2id} = \{\text{country} : \text{index}\}.$$

For batch preparation:

- Raw waveforms were padded to the longest sequence in the batch for ECAPA-TDNN input.
- WavLM inputs were prepared using `AutoFeatureExtractor` (`return_tensors="pt"`, `padding=True`) with a 16 kHz sampling rate.

No data augmentation was applied.

Batching Training batches comprised 4 audio samples each, randomly shuffled for training and kept in sequential order for validation.

Training Hyperparameters Models were trained for up to 100,000 steps using `AdamW` with a weight decay of 10^{-2}. Learning rates were set to 1×10^{-5} for ECAPA-TDNN and WavLM encoder parameters, and 1×10^{-4} for the projection layers and classifier. Gradient clipping (max norm = 1.0) was applied to mitigate exploding gradients.

Freezing Strategy & Learning Rate Schedule Pretrained encoders were frozen for the first 8,000 steps, followed by full-network fine-tuning. The learning rate schedule, implemented via PyTorch, consisted of:

1. **Frozen phase (0–8,000 steps):**
   - *Warmup* (0–3,000): `LinearLR`, start factor $= \frac{1}{3}$.
   - *Constant* (3,000–8,000): `ConstantLR`.
2. **Unfrozen phase (8,000–100,000 steps):**
   - *Warmup* (8,000–12,000): `LinearLR`, start factor $= \frac{1}{10}$.
   - *Constant* (12,000–52,000): `ConstantLR`.
   - *Cosine decay* (52,000–100,000): `CosineAnnealingLR`.

Evaluation Metrics System performance was assessed using two metrics:

- **Accuracy:** Proportion of correctly classified samples over the total number of evaluated samples.
- **Average Cost:** Following the NIST LRE 2022 formulation (Lee et al., 2022), log-likelihood ratios were computed from model logits via pairwise class comparisons to estimate prediction confidence. The cost combines false positive rate (FPR) and false negative rate (FNR) as:

$$\text{Cost} = \text{FPR} + \text{FNR}.$$

  This metric balances penalties for missed detections and false alarms across varying decision thresholds.

Accuracy served as the primary metric, with average cost providing a complementary error-sensitive measure.

4 Results

Experiments were conducted on the official NADI 2025 validation and blind test splits.

Validation: Our system achieved **94.08%** accuracy with an average NIST cost of **6.37%**, ranking 3rd on the validation leaderboard.

Test: Performance dropped to **~51.0%** accuracy with an average NIST cost of **~49.0%**, likely due to domain mismatch and data distribution shifts between validation and test sets.

No additional ablation or error analysis was performed; results focus on the primary leaderboard metrics.

5 Conclusion

We presented `Lahjati` (`ECAPA_WavLM_Fusion`), a dual-encoder fusion model combining ECAPA-TDNN and WavLM to jointly capture speaker-level and contextual acoustic representations for Arabic dialect identification. A staged training regime—initial encoder freezing followed by fine-tuning—yielded competitive results, with a validation accuracy of 94.08% on the NADI 2025 dataset.

Limitations include the absence of data augmentation and challenges from dialectal overlap, both of which may hinder generalization to unseen data. Future work will investigate advanced augmentation, alternative fusion architectures, and hyperparameter optimization to improve robustness.

This study offers a competitive, reproducible benchmark for Arabic dialect identification, contributing toward improved speech processing for underrepresented language varieties.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2310.16117*.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2407.04910*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, WANLP@COLING 2020, Barcelona, Spain (Online), December 12, 2020*, pages 97–110. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim A. Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop, WANLP 2021, Kyiv, Ukraine (Virtual),*

April 9, 2021, pages 244–259. Association for Computational Linguistics.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Yooyoung Lee, Craig Greenberg, Lisa Mason, and Elliot Singer. 2022. Nist 2022 language recognition evaluation plan.

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.

UBC-NLP. 2025. Nadi 2025 subtask 1: Spoken language identification (slid) dataset. https://huggingface.co/datasets/UBC-NLP/NADI2025_subtask1_SLID. Accessed 2025-08-11.