Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing under Data Constraints

Badr M. Abdullah[1,5] **Yusser Al Ghussin**[1,2] **Zena Al-Khalili**[1,5] **Ömer Tarik Özyilmaz**[3,4]
Matias Valdenegro-Toro[4] **Simon Ostermann**[1,2] **Dietrich Klakow**[1,5]

[1]Saarland University, Germany [2]German Research Center for AI (DFKI), Germany
[3]University Medical Center Groningen, University of Groningen, The Netherlands
[4]University of Groningen, The Netherlands
[5]SFB 1102 - Information Density and Linguistic Encoding (IDeaL)

Abstract

We present our systems for the NADI 2025 shared task on multidialectal Arabic speech processing, participating in both spoken dialect identification (ADI) and automatic speech recognition (ASR) subtasks. Working under data constraints by using only the provided shared task resources for dialect adaptation, we explore effective model adaptation strategies for dialectal Arabic speech. For ADI, we fine-tune w2v-BERT 2.0 and employ voice conversion as data augmentation, improving accuracy from 68.71% to 76.40% on a blind cross-domain test set. For ASR, we develop two complementary approaches: (1) a CTC-based model pre-trained on public Arabic speech data, and (2) Whisper-based models using two-stage fine-tuning. Our experiments show that while dialect-centric CTC models exhibit better zero-shot dialectal performance (58.89 vs 93.90 WER), Whisper achieves better performance after dialect-specific adaptation, which reduces WER from 93.89 to 39.78 WER. We also demonstrate that using character error rate (CER) as a validation criterion provides practical benefits with minimal performance trade-offs. Despite using no external resources for dialect adaptation beyond the shared task data, our systems ranked second in ADI and third in ASR, demonstrating that careful adaptation strategies can overcome data constraints in dialectal speech processing.

1 Introduction

The Arabic language exhibits a rich linguistic variation landscape. While Modern Standard Arabic (MSA) serves as the official language and codified variety across all Arabic-speaking countries, it primarily exists in formal situations such as scripted news broadcasts and official documents. Daily spoken communication occurs exclusively in regional dialects that differ from MSA and each other at every linguistic level: prosody, phonology, lexicon, and syntax. Although spoken dialects still lack a standardized orthography and are not formally taught in schools, they maintain a strong cultural presence through songs, folktales, and cinema (Holes, 2004; Habash, 2010).

Despite recent advances in language technology, MSA remains the only Arabic variety that is well-supported by AI-powered speech technology. For example, while state-of-the-art ASR systems (e.g., Radford et al. (2023)'s Whisper model) work well on MSA speech, they fail to adequately transcribe and translate dialectal speech. To address this gap, recent community efforts have focused on building speech resources for Arabic dialects. Notable among these is the Casablanca corpus (Talafha et al., 2024), the largest fully supervised Arabic speech dataset covering eight regional dialects. The NADI 2025 shared task builds on this resource to advance speech technologies for Arabic dialects across three speech processing subtasks.

In the NADI 2025 shared task, we participated in two subtasks: spoken Arabic dialect identification (ADI) and multidialectal Arabic ASR. Working exclusively with the provided datasets by the organizers, we explored which model adaptation techniques are most effective under resource constraints. For ADI, we adapted the multilingual pretrained w2v-BERT 2.0 model using supervised fine-tuning and voice conversion as audio augmentation. We found that this approach improves robustness to domain mismatch, which is consistent with our prior work (Abdullah et al., 2025). For ASR, we developed two systems: (1) a dialect-centric model based on connectionist temporal classification (CTC) loss and (2) fine-tuned Whisper models. While the dialect-centric approach performed better in zero-shot settings, dialect-specific Whisper-based models achieved superior performance after fine-tuning. Overall, our best ADI system ranked second while our best ASR system ranked third in their respective subtasks, despite our data constrained setup.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 745–751

2 Shared Task Description

The NADI shared task series has evolved significantly over the years, with previous iterations (2020-2024) focusing primarily on text-based dialect identification at various granularities (Abdul-Mageed et al., 2020, 2021; Abdul-Mageed et al., 2022, 2023, 2024). NADI 2025 represents a major shift to speech processing, recognizing that dialectal variation is most naturally expressed in spoken form and that speech technology lags behind text processing for Arabic dialects.

The NADI 2025 shared task focuses on advancing multidialectal Arabic speech processing through three complementary subtasks that address critical challenges in dialect-aware speech technology (Talafha et al., 2025). Building on the Casablanca corpus (Talafha et al., 2024), the task provides participants with resources for eight Arabic dialects throughout the Middle East and North Africa. The dataset covers eight country-level dialects with the following abbreviations used throughout this paper: Algerian (ALG), Egyptian (EGY), Emirati (UAE), Jordanian (JOR), Mauritanian (MAU), Moroccan (MOR), Palestinian (PAL), and Yemeni (YEM).

2.1 Subtask 1: Spoken Arabic Dialect Identification (ADI)

This subtask requires systems to predict the spoken Arabic dialect from short audio clips. Given the rich linguistic diversity of Arabic and the limited availability of labeled dialectal speech data, accurate dialect identification remains challenging, especially in domain mismatch settings (Sullivan et al., 2023; Abdullah et al., 2025). This subtask aims to evaluate how well modern multilingual speech models and embedding techniques can distinguish between dialectal variations using acoustic-phonetic features. The provided dataset for this subtask consists of dialect-annotated speech samples for three splits: adaptation, validation, and test, where each split is 8 hours of speech.

2.2 Subtask 2: Multidialectal Arabic ASR

In this subtask, participants are required to develop ASR systems capable of adequately transcribing speech across multiple Arabic dialects. The primary challenge lies in handling the substantial phonological, lexical, and syntactic variations between dialects while maintaining high-quality transcriptions across all varieties. Systems are evaluated using both Word Error Rate (WER) and Character Error Rate (CER) metrics, which measure the extent to which ASR generated transcripts match gold human transcriptions. The provided dataset for this subtask consists of transcribed speech samples for three splits: adaptation (12,800 utterances), validation (12,800 utterances), and test (10,298 utterances).

3 System Overview

In this section, we describe our systems for the shared task. We refer to all our systems under the name **BYZÖ**, an acronym formed from the first letters of each core team member's first name.

3.1 Spoken Arabic Dialect Identification

We fine-tuned the multilingual pre-trained speech model w2v-BERT-2.0 for ADI using only the provided shared task data. We add an 8-way classification head that is randomly initialized on top of the pre-trained model for this task. To improve the model's robustness against unpredictable recording variations, we used k-nearest neighbor (k-NN) voice conversion (Baas et al., 2023) to create resynthesized samples from the training data using target voices from LibriVox audiobook recordings. We used four target voices from LibriVox who spoke standard Arabic. Using this approach, we created synthesized data that is four times larger than the original dataset. Our results show that using a combined dataset (natural + resynthesized) significantly improves performance without adding any natural samples or requiring architectural modifications.

3.2 Multidialectal Arabic ASR

3.2.1 System 1: `BYZÖ-whisper`

Similar to prior research in dialectal Arabic ASR using Whisper (Özyilmaz et al., 2025), we fine-tune the Whisper-large-v3 model for multidialectal Arabic ASR and examine how different training strategies affect its performance. Our Whisper-based approach consists of three aspects:

1. Two-stage fine-tuning procedure. First, we perform domain adaptation by fine-tuning all model layers on the combined dataset from all dialects, creating a domain-adapted multidialect baseline. Second, we conduct dialect adaptation by fine-tuning eight dialect-specific models, each trained exclusively on its respective dialect data using the same configuration. This approach combines the

benefits of shared dialectal knowledge with dialect-specific optimization.

2. Alternative validation criterion. We experiment with CER as an alternative validation metric to stop early during dialect adaptation. While domain adaptation uses WER for validation, we compare WER versus CER as stopping criteria for dialect-specific fine-tuning. Using CER for early stopping may prevent overfitting to frequent word patterns and yield better character-level performance.

3. Parameter-efficient fine-tuning via LoRA. We also experiment with Low-Rank Adaptation, or LoRA (Liu et al., 2024), as an efficient alternative to full fine-tuning. LoRA inserts trainable rank-decomposition matrices into the model's weight layers while keeping original weights frozen, reducing computational costs and potential overfitting on limited dialect data.

3.2.2 System 2: BYZÖ-ctc

As an alternative to Whisper-based models, we developed our own dialect-centric ASR model by fine-tuning w2v-BERT-2.0 (580M parameters) with CTC loss. The model underwent two-stage training: (1) supervised fine-tuning on public Arabic ASR datasets including Arabic Common Voice (Ardila et al., 2020), SADA (Alharbi et al., 2024), Linto (Abdallah et al., 2024; Naouara et al., 2025), D-Voice 2.0 (Allak et al., 2021), and the Egyptian Arabic ASR dataset on Kaggle, and (2) dialect-specific fine-tuning using only the shared task data. This encoder-only architecture is more efficient than Whisper-based models and we show that it outperforms Whisper-large in zero-shot settings.

To enhance the dialectal fidelity of ASR output, we trained dialect-specific n-gram language models with Kneser-Ney smoothing (with $n = 3$) using curated text corpora for each dialect. These LMs were integrated into BYZÖ-ctc's decoding to constrain acoustically plausible but linguistically unlikely word sequences, reducing grammatical and lexical errors in the final transcriptions. The LMs training corpora are detailed in Appendix B.

4 Experimental Setup

We used the Hugging Face Transformers library and the Trainer module to fine-tune our ASR and ADI systems. For our Whisper-based systems, we used the AdamW optimizer with a linear learning rate warmup for 500 steps to a peak of 1×10^{-5}, followed by cosine decay. Each model was trained for up to 2000 steps. For our w2vBERT 2.0-based systems, we used the AdamW optimizer with a linear learning rate warmup for 10% of the adaptation samples to a peak of 1×10^{-5}, followed by linear decay. We applied minimal text processing to the text transcripts for the ASR systems. We share our code and models for reproducibility[1].

System	Accuracy (%)	Avg. Cost
Baseline	61.09	0.342
BYZÖ-ADI	68.71	1.136
BYZÖ-ADI + VC	**76.40**	**0.227**

Table 1: Dialect identification performance metrics. Our approach with voice conversion (VC) achieves optimal performance with 76.4% accuracy (higher is better) as well as the lowest cost value (lower is better).

5 Experimental Results

5.1 Spoken Arabic Dialect Identification

Table 1 presents the ADI results on the NADI 2025 test set with two evaluation metrics: accuracy and average cost as define by the NIST Language Recognition Evaluation campaign. The baseline system, which is based on a Pretrained ECAPA-TDNN VoxLingua107 system fine-tuned on adaptation split, achieves 61.09% accuracy with a cost of 0.342. Our initial BYZÖ-ADI model improves accuracy to 68.71%, though at a higher cost of 1.136, indicating increased confusion between dialects. However, incorporating voice conversion (VC) as a data augmentation strategy yields substantial improvements on both metrics. The BYZÖ-ADI + VC system achieves the best performance with 76.40% accuracy while simultaneously reducing the cost to 0.227. This 7.69 percentage point improvement in accuracy over the base model demonstrates that voice conversion effectively enhances the model's robustness to acoustic variations while improving its discriminative ability across dialects.

5.2 Multidialectal Arabic ASR

Table 2 shows WER results across eight dialects. The zero-shot Whisper baseline fails completely on dialectal speech with an average WER of 93.90, except for Jordanian (46.10). Our BYZÖ-ctc model performs outperforms Whisper in zero-shot set-

[1] https://github.com/Yusser95/NADI-NLP-2025-Whisper

System	ALG	EGY	JOR	MAU	MOR	PAL	UAE	YEM	AVG
Whisper (Zero-shot)	101.0	100.1	46.09	100.6	100.4	100.8	101.6	101.1	93.89
BYZÖ-ctc (Zero-shot)	75.17	48.40	40.67	81.25	72.21	52.24	46.91	54.23	58.89
BYZÖ-ctc + SFT	60.82	40.59	44.52	67.00	50.74	45.45	42.31	49.24	50.08
BYZÖ-ctc + SFT + LM	57.12	35.23	32.62	62.81	45.46	37.32	38.20	46.42	44.40
BYZÖ-whisper + SFT I	65.10	32.88	31.49	69.80	57.80	31.31	35.69	53.14	47.15
BYZÖ-whisper + SFT II	**55.04**	**29.50**	**28.84**	**59.37**	**43.07**	**27.66**	**28.38**	**46.42**	**39.78**

Table 2: Word Error Rate (WER) performance across eight Arabic dialects on the NADI 2025 test set. All our systems used only shared task data for dialect adaptation. The baseline Whisper zero-shot results demonstrate the challenge of dialectal ASR, while our BYZÖ systems show progressive improvements. Best results (in bold) are achieved by two-stage fine-tuning with CER as criterion. Lower values indicate better performance.

tings (WER of 58.89), showing that dialect-centric pre-training is effective for dialectal speech. After dialect-specific fine-tuning, both our systems improve significantly. The CTC model reduces average WER from 58.89 to 50.08 with supervised fine-tuning, and further to 44.40 when adding language models. The Whisper-based models achieve better final results despite worse zero-shot performance. Whisper fine-tuning gives a WER of 47.15, while two-stage fine-tuning with CER as a validation criterion achieves the best performance at 39.78. This 4.62 point gap suggests the encoder-decoder architecture handles dialectal variations better than CTC when properly fine-tuned. Both models show the largest gains on low-resource dialects like Mauritanian and Moroccan, reducing WER by over 40 points from baseline.

On the other hand, Table 3 shows the performance measured by CER for the eight dialects. Interestingly, the model that yields the lowest WER for a dialect does not necessarily yield the lowest CER. This finding suggests that WER and CER might capture different model competences and therefore should be combined when evaluating ASR models.

6 Discussion

Our results reveal several important insights about adapting ASR systems for dialectal Arabic speech. The dramatic failure of zero-shot Whisper (93.9 WER average) highlights a fundamental challenge: models trained primarily on MSA and high-resource languages cannot generalize to Arabic dialects, despite Whisper's multilingual capabilities. This performance gap shows how the distinct phonological and lexical features, which separate dialectal Arabic varieties from MSA, affect the performance of ASR systems. The success of our adaptation strategies raises interesting questions about model architecture choices. While our CTC-based model shows better zero-shot dialectal speech-to-text transcription (58.89 vs 93.90 WER), the Whisper architecture ultimately achieves superior performance after fine-tuning (39.78 WER). This suggests that encoder-decoder models may have greater capacity for dialectal adaptation when provided with adequate supervision for each dialect, possibly due to their ability to model longer-range dependencies and contextual information during decoding.

7 Conclusion

We presented data-constrained approaches for the NADI 2025 shared task, achieving competitive results in both dialect identification and ASR subtasks. Our key findings include: (1) voice conversion improves ADI accuracy by 7.69 percentage points while reducing classification uncertainty, (2) dialect-centric pre-training provides better zero-shot performance than general multilingual models, and (3) two-stage fine-tuning with character-level optimization yields the best ASR results. Our experiments reveal important architectural trade-offs. CTC models offer better initial dialectal understanding and efficiency, while encoder-decoder architectures show superior adaptation capacity after fine-tuning. Future work should address the persistent performance disparities across dialects (27.7-59.4 WER range), which cannot be resolved through equal data distribution alone. Promising directions including cross-dialectal transfer learning and extending voice conversion techniques to ASR tasks. Our competitive rankings despite using only shared task data demonstrate that advancing dialec-

System	ALG	EGY	JOR	MAU	MOR	PAL	UAE	YEM	AVG
Whisper (Zero-shot)	79.58	81.37	19.28	82.89	80.42	77.92	80.27	80.58	84.69
BYZÖ-ctc (Zero-shot)	32.65	16.68	11.23	39.47	28.52	16.07	12.76	18.23	21.95
BYZÖ-ctc + SFT	**20.17**	13.06	12.25	**24.64**	16.20	13.91	11.68	**15.32**	15.90
BYZÖ-ctc + SFT + LM	22.03	12.02	10.17	26.25	15.89	12.30	11.00	15.85	15.69
BYZÖ-whisper + SFT I	26.69	13.41	10.36	30.12	21.21	12.23	11.91	24.79	18.84
BYZÖ-whisper + SFT II	20.59	**11.91**	**9.47**	24.85	**15.52**	**10.59**	**9.04**	16.05	**14.76**

Table 3: Character Error Rate (CER) performance across eight Arabic dialects on the NADI 2025 test set. All our systems used only shared task data for dialect adaptation. The baseline Whisper zero-shot results demonstrate the challenge of dialectal ASR, while our BYZÖ systems show progressive improvements. Best result for a dialect is shown in bold. Lower values indicate better performance.

tal Arabic speech technology requires not massive resources, but careful adaptation strategies tailored to the unique characteristics of Arabic dialects.

Acknowledgments

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102 and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2024. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. In *ICASSP 2024-2024 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pages 12607–12611. IEEE.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Badr M Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. Voice conversion improves cross-domain robustness for spoken arabic dialect identification. *arXiv e-prints*, pages arXiv–2505.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013.

Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a Levantine corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 769–778, Marseille, France. European Language Resources Association.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, and 1 others. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290. IEEE.

Anass Allak, Naira Abdou Mohamed, Imade Benelallam, and Kamel Gaanoun. 2021. Dialectal voice : An open-source voice dataset and automatic speech recognition model for moroccan arabic dialect. In *Proceedings of the Data Centric AI (NeurIPS 2021)*.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. In *Interspeech 2023*, pages 2053–2057.

Omar A Essameldin, Ali O Elbeih, Wael H Gomaa, and Wael F Elsersy. 2025. Arabic dialect classification using rnns, transformers, and large language models: A comparative analysis. *arXiv preprint arXiv:2506.19753*.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Mustafa Jarrar, Fadi A Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2022. Lisan: Yemeni, iraqi, libyan, and sudanese arabic dialect copora with morphological annotations.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Batista-Navarro. 2023. UniManc at NADI 2023 shared task: A comparison of various t5-based models for translating Arabic dialectical text to Modern Standard Arabic. In *Proceedings of ArabicNLP 2023*, pages 658–664, Singapore (Hybrid). Association for Computational Linguistics.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Hedi Naouara, Jean-Pierre Lorré, and Jérôme Louradour. 2025. Linto audio and textual datasets to train and evaluate automatic speech recognition in tunisian arabic dialect. *arXiv preprint arXiv:2504.02604*.

Ömer Tarik Özyilmaz, Matt Coler, and Matias Valdenegro-Toro. 2025. Overcoming data scarcity in multi-dialectal arabic asr via whisper fine-tuning. *arXiv preprint arXiv:2506.02627*.

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.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the robustness of arabic speech dialect identification. In *Interspeech 2023*, pages 5326–5330.

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.

A Training parameters for Whisper

A.1 System Configurations and Training Setup

We train and evaluate three Whisper-based ASR systems, summarized as follows:

1. **Whisper + SFT**: A two-stage fine-tuning system with WER loss in first and second stage. This configuration trains all model weights (no LoRA). The maximum generation length used in training is 225 tokens.

2. **Whisper + SFT + 2OPT**: A full fine-tuning system with WER loss in first stage (same shared across all systems) and CER loss for the second stage. This configuration trains all model weights (no LoRA) to directly compare against System **Whisper + SFT** and show the effect of CER-based training. The maximum generation length is 225 tokens.

3. **Whisper + SFT + LORA + 2OPt**: We use the same first stage model trained using full fine-tuning system with WER loss and for the scond stage we train a parameter-efficient system using LoRA and a CER loss. because it showed that it was effective in System **Whisper + SFT + 2OPT**. We freeze Whisper's original weights and fine-tune only LoRA adapter parameters inserted in each layer (rank $r = 32$). The CER loss term ($\lambda = 0.5$) is added to the training objective to directly optimize character accuracy. We impose a stricter maximum generation length of 125 tokens to

simulate potential truncation and evaluate its effect, especially in conjunction with LoRA. This setup updates only $\sim$1% of parameters, significantly reducing training memory and time, making it appealing for low-resource or deployment scenarios if accuracy trade-offs are acceptable.

B Training Corpora of n-gram Language Models

The training corpora for the n-gram language models were compiled from several existing, dialect-annotated datasets. These primary sources include the Palestinian Curas corpus (Al-Haff et al., 2022), the Yemeni Lisan corpus (Jarrar et al., 2022), the Emirati Emi-NADI (Khered et al., 2023), the Moroccan Darija-LID dataset[2], and the multi-dialect QADI corpus (Abdelali et al., 2021).

To augment these resources, we expanded the training data by automatically annotating a subset of the Arabic-tweets dataset (Al-Fetyani et al., 2023). This dialect identification task was performed using the MARBERTv2 model (Essameldin et al., 2025).

C Correlation between Different Models

Figure 1 shows the correlation between different models in their dialect performance. One can observe a strong correlation between the models, which indicates that the different systems behave similarly for the dialectal Arabic ASR task.

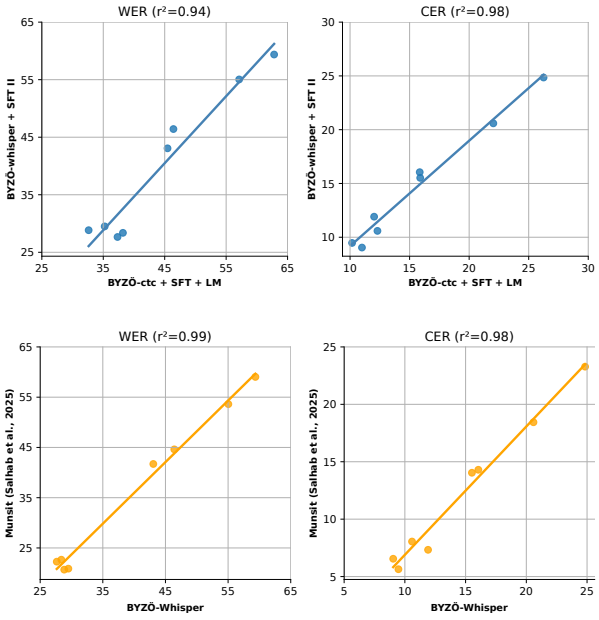

Figure 1: Performance correlation between different models: Our CTC- and Whisper-based systems (top), and the top performing system in the shared task vs. our best system. Each data point in the figure corresponds to a dialect.

[2] https://huggingface.co/datasets/atlasia/Darija-LID