

Munsit at NADI 2025 Shared Task 2: Pushing the Boundaries of Multidialectal Arabic ASR with Weakly Supervised Pretraining and Continual Supervised Fine-tuning

Mahmoud Salhab
CNTXT AI
Abu Dhabi, UAE
mahmoud.salhab@cntxt.tech

Shameed Sait
CNTXT AI
Abu Dhabi, UAE
shameed.ali@cntxt.tech

Mohammad Abusheikh
CNTXT AI
Abu Dhabi, UAE
mas@cntxt.tech

Hasan Abusheikh
CNTXT AI
Abu Dhabi, UAE
has@cntxt.tech

Abstract

Automatic speech recognition (ASR) plays a vital role in enabling natural human–machine interaction across applications such as virtual assistants, industrial automation, customer support, and real-time transcription. However, developing accurate ASR systems for low-resource languages like Arabic remains a significant challenge due to limited labeled data and the linguistic complexity introduced by diverse dialects. In this work, we present a scalable training pipeline that combines weakly supervised learning with supervised fine-tuning to develop a robust Arabic ASR model. In the first stage, we pretrain the model on 15,000 hours of weakly labeled speech covering both Modern Standard Arabic (MSA) and various Dialectal Arabic (DA) variants. In the subsequent stage, we perform continual supervised fine-tuning using a mixture of filtered weakly labeled data and a small, high-quality annotated dataset. Our approach achieves state-of-the-art results, ranking first in the multi-dialectal Arabic ASR challenge. These findings highlight the effectiveness of weak supervision paired with fine-tuning in overcoming data scarcity and delivering high-quality ASR for low-resource, dialect-rich languages.

1 Introduction

Automatic speech recognition (ASR), or speech-to-text (STT), converts spoken language into text, enabling voice-based interaction with machines (Al-gihab et al., 2019; Kheddar et al., 2024). ASR is widely applied in healthcare, robotics, law enforcement, telecommunications, smart homes, and consumer electronics, among other domains (Vajpai and Bora, 2016). Arabic, the fourth most used language online and one of the UN’s six official

languages, remains underrepresented in ASR research despite serving millions across 22 countries (Alwajeeh et al., 2014). Arabic exists in three forms: Classical Arabic (CA), the language of historical and religious texts; Modern Standard Arabic (MSA), used in formal contexts; and Dialectal Arabic (DA), comprising diverse regional variants (Al-Ayyoub et al., 2018). While some datasets, such as MASC (Al-Fetyani et al., 2021) and SADA (Alharbi et al., 2024), have advanced Arabic ASR, they remain limited in size and linguistic diversity, hindering model generalization. Neural ASR systems require vast transcribed datasets (Lu et al., 2020; Wang et al., 2021), but manual transcription is costly and time-intensive (Gao et al., 2023). We address this by proposing a weakly supervised Arabic ASR system based on the Conformer architecture (Gulati et al., 2020), trained on large-scale weakly labeled MSA and DA speech. In the first stage, we pretrain the model on 15,000 hours of weakly labeled speech covering both Modern Standard Arabic (MSA) and various Dialectal Arabic (DA) variants. In the subsequent stage, we perform continual supervised fine-tuning using a mixture of filtered weakly labeled data and a small, high-quality annotated dataset provided as part of the challenge (Talafha et al., 2025). This approach eliminates the need for extensive manual transcription and attains state-of-the-art results on the challenge’s standard benchmarks, demonstrating the potential of weak supervision for low-resource languages.

2 Background

Arabic Automatic Speech Recognition (ASR) remains challenging due to data scarcity, lexical variation, morphological complexity, and dialect diver-

sity across 22 Arab countries (Ali et al., 2014; Cardinal et al., 2014; Diehl et al., 2012). Traditional systems often used hybrid HMM-DNN pipelines (Cardinal et al., 2014; Bouchakour and Debyeche, 2018). Dialectal variation is a major bottleneck, as most systems focus on Modern Standard Arabic (MSA) and high-resource dialects, performing poorly on low-resource varieties (Djanibekov et al., 2025). To address this, Djanibekov et al. released open-source ASR models covering 17 countries, 11 dialects, and code-switched Arabic-English/French speech. Other efforts integrate dialect identification directly into ASR (Waheed et al., 2023) or build dialect-specific systems, e.g., for Egyptian (Mousa et al., 2013) and Algerian Arabic (Menacer et al., 2017).

End-to-end architectures have advanced Arabic ASR by eliminating the need for intermediate feature extraction (Radford et al., 2023a). Notable examples include large-scale weakly supervised systems such as Whisper (Radford et al., 2023b). Weak supervision has proven particularly effective; for instance, (Salhab et al., 2025) trained a Conformer model from scratch on 15,000 hours of weakly labeled MSA and dialectal speech, achieving state-of-the-art results without relying on manual transcription.

3 Methodology

Our approach consists of two main stages: weakly supervised pretraining followed by continual supervised fine-tuning. In the first stage, we train the model on a large-scale, diverse speech dataset with weak labels—labels that are not guaranteed to be accurate (i.e., not manually verified)—in line with the strategy proposed in (Salhab et al., 2025).

In the second stage, the pretrained model is further fine-tuned using a smaller, high-quality dataset constructed from two main sources: (1) the official training data released for the task (the Casablanca training set (Talafha et al., 2024)), which is expanded through various augmentation techniques, including random Gaussian noise injection, background noise addition, and silence insertion; and (2) a filtered subset derived from the initial 15,000 hours of weakly labeled training data, selected through a rigorous data cleaning and filtering process.

An overview of the complete pipeline is presented in Figure 1. The following subsections provide a detailed explanation of each stage of the

proposed approach.

3.1 Weakly Supervised Learning

Traditional supervised ASR training uses high-quality, human-annotated pairs (x_i, y_i) , where the input x_i is typically a mel-spectrogram and the output y_i consists of a sequence of tokens, each selected from a predefined vocabulary. These accurate labels are assumed to be independently drawn from a clean data distribution, enabling the model to learn a function that performs well on unseen test examples. On the other hand, weakly supervised learning depends on automatically generated or crowd-sourced labels \hat{y}_i , which may contain errors or noise. These weak labels come from a noisier distribution and might not precisely reflect the true transcription. Nonetheless, models trained on such data aim to generalize effectively when evaluated on clean datasets.

Building upon the approach introduced in (Salhab et al., 2025), we adopted the same training pipeline and experimental settings to develop the initial foundation model. Specifically, the model was trained on 15,000 hours of weakly annotated speech data, with automatic labeling performed using the same method described in the aforementioned work.

3.2 Continual supervised finetuning

In neural network-based ASR systems, training typically begins either from scratch—with randomly initialized weights and a large training corpus—or from a pretrained model that has already been exposed to a large-scale dataset. The latter approach enables faster convergence and often better generalization on the target task due to prior knowledge encoded in the pretrained weights.

In this stage, we adopt the second strategy by initializing the model with weights obtained from the first stage, which was trained on weakly labeled data. We then fine-tune this model using a smaller yet higher-quality dataset comprising 3,000 hours of filtered weakly annotated data. The filtering process was designed to exclude news content—largely composed of Modern Standard Arabic (MSA)—and to retain only segments that passed stringent quality thresholds, as outlined in the pipeline of (Salhab et al., 2025). Additionally, we incorporate the Casablanca Challenge training dataset, which is further expanded through various data augmentation techniques. Unlike the first stage that relied on noisy supervision, this fine-

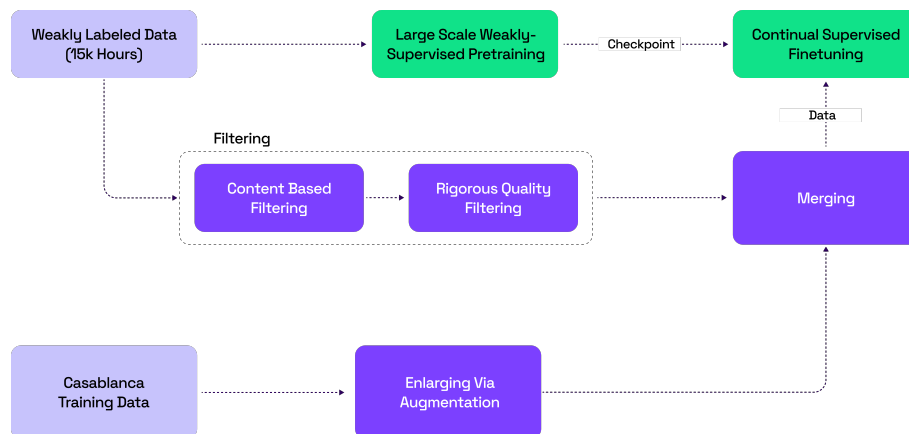


Figure 1: The solution’s full pipeline encompasses large-scale pretraining followed by continual fine-tuning.

tuning phase leverages only high-quality transcriptions.

3.3 Model Architecture

The Conformer architecture (Gulati et al., 2020) effectively models both long- and short-range dependencies in speech through a combination of convolutional modules and multi-head self-attention, making it highly suitable for automatic speech recognition. In this work, we adopt the same architecture as introduced in the original paper, specifically using the large variant of the model.

3.4 Experimental Setup

Our ASR experiments utilized the Conformer architecture trained with the Connectionist Temporal Classification (CTC) objective (Graves et al., 2006). To tokenize the transcripts, we employed a SentencePiece tokenizer (Kudo and Richardson, 2018) trained on the same training corpus, with a vocabulary of 128 tokens.

Model training was carried out in a distributed setting across 8 NVIDIA A100 GPUs using a global batch size of 512. Input features were 80-dimensional mel-spectrograms, extracted using a 25 ms frame length and a 10 ms hop size.

During the weakly supervised pretraining phase, optimization was performed using the AdamW optimizer combined with the Noam learning rate scheduler, incorporating 10,000 warm-up steps and peaking at a learning rate of 2×10^{-3} . For regularization purposes, we applied a dropout rate of 0.1 across all layers and used L2 weight decay. For the fine-tuning stage, the learning rate was reduced by a factor of ten.

To optimize training speed and reduce memory overhead, computations were performed using

bf16 precision. The Conformer model was initialized with random weights and comprised 18 encoder layers. Each layer featured a hidden dimension of 512, 8 attention heads, a convolutional kernel size of 31, and a feedforward expansion factor of 4. The complete model architecture contained approximately 121 million parameters.

3.5 Evaluation Metrics & Datasets

The model’s performance was evaluated using Word Error Rate (WER) and Character Error Rate (CER). Training used a development set with paired speech and transcriptions, while testing involved blind evaluation on speech-only data via CodeBench.

4 Results

We evaluate our proposed system, against all participating teams using both Word Error Rate (WER) and Character Error Rate (CER) metrics, reported across multiple Arabic dialects. The results demonstrate the robustness of our approach across both evaluation and testing phases, as well as its ability to generalize across diverse dialectal variations.

As shown in Table 1, our system achieved the lowest average WER (35.69%), outperforming all other submissions. Notably, our work consistently maintained lower WER in most of the dialects, particularly excelling in Jordanian (20.68%), Egyptian (20.89%), and Emirati (22.67%) dialects. Similarly, Table 2 shows that our model achieved the lowest average CER (12.21%), with the best performance observed in Jordanian (5.64%) and Egyptian (7.33%) dialects. Tables 3 and 4 present a breakdown of WER and CER across evaluation and testing phases/datasets. The average WER decreased

Participant	Avg	JOR	EGY	MOR	ALG	YEM	MAU	UAE	PAL
msalhab96 (Ours)	35.69	20.68	20.89	41.72	53.62	44.62	59.03	22.67	22.28
youssef_saidi	38.54	28.03	26.83	38.27	53.73	46.63	58.11	29.35	27.36
yusser	39.78	28.84	29.50	43.07	55.04	46.42	59.37	28.38	27.66
alhassan10ehab	42.05	32.25	24.73	48.22	60.32	51.77	66.23	28.01	24.87
badr_alabsi	44.15	31.74	37.24	43.31	56.12	46.15	63.32	38.65	36.63
Baseline	93.90	46.10	100.07	100.38	101.03	101.09	100.59	101.15	100.77
rafiulbiswas	104.90	44.97	113.98	104.08	116.60	113.54	111.59	116.79	117.61

Table 1: Dialect-wise WER (%) Comparison Across Participants.

Participant	Avg	JOR	EGY	MOR	ALG	YEM	MAU	UAE	PAL
msalhab96 (Ours)	12.21	5.64	7.33	14.04	18.44	14.30	23.28	6.55	8.06
youssef_saidi	14.53	9.36	11.44	13.66	20.43	16.66	24.53	9.91	10.20
yusser	14.76	9.47	11.91	15.52	20.59	16.05	24.85	9.04	10.59
alhassan10ehab	16.19	9.90	10.21	18.12	23.34	20.41	29.11	8.99	9.41
badr_alabsi	15.59	9.95	12.57	15.07	21.39	15.69	26.70	11.15	12.19
Baseline	72.79	19.29	81.38	80.42	79.59	80.58	82.89	80.28	77.93
rafiulbiswas	84.69	19.19	97.66	87.59	94.27	94.56	92.85	97.01	94.42

Table 2: Dialect-wise CER (%) Comparison Across Participants.

Dialect	Evaluation	Testing	Dialect	Evaluation	Testing
Avg	36.83	35.69	Avg	11.94	12.21
JOR	21.52	20.68	JOR	5.39	5.64
EGY	22.89	20.89	EGY	7.50	7.33
MOR	44.20	41.72	MOR	14.06	14.04
ALG	54.78	53.62	ALG	17.71	18.44
YEM	47.69	44.62	YEM	14.73	14.30
MAU	57.62	59.03	MAU	21.73	23.28
UAE	24.05	22.67	UAE	6.97	6.55
PAL	21.91	22.28	PAL	7.40	8.06

Table 3: Comparison of WER (%) Across Evaluation and Testing Datasets.

Table 4: Comparison of CER (%) Across Evaluation and Testing Datasets.

from 36.83% during evaluation to 35.69% in testing, suggesting that our model generalizes well to unseen data. This trend is consistent across most dialects. For instance, the WER in the Jordanian dialect dropped from 21.52% to 20.68%, and in the Yemeni dialect from 47.69% to 44.62%.

Similarly, the average CER exhibited a slight increase from 11.94% (evaluation) to 12.21% (testing), though the variation across dialects remained minimal, underscoring the model’s stability. These consistent results across both phases affirm the robustness and dialectal adaptability of our ASR system.

5 Conclusion

We present a scalable two-stage pipeline—pretraining on 15,000 hours of

weakly labeled audio, then fine-tuning on a filtered 3,000-hour weak subset plus an augmented official training set—that, with data filtering, augmentation, and a Conformer backbone, achieved state-of-the-art performance and first place in the multi-dialectal Arabic ASR challenge, demonstrating that carefully curated weak supervision combined with targeted fine-tuning can overcome data scarcity and dialectal diversity.

References

- Mahmoud Al-Ayyoub, Aya Nuseir, Kholoud Alsmearat, Yaser Jararweh, and Brij Gupta. 2018. [Deep learning for arabic nlp: A survey](#). *Journal of Computational Science*, 26:522–531.
- Mohammad Al-Fetyani, Muhammad Al-Barham,

- Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. [Masc: Massive arabic speech corpus](#).
- Wajdan Algihab, Noura Alawwad, Anfal Aldawish, and Sarah AlHumoud. 2019. Arabic speech recognition with deep learning: A review. In *Social Computing and Social Media. Design, Human Behavior and Analytics*, pages 15–31, Cham. Springer International Publishing.
- Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghada Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. [Sada: Saudi audio dataset for arabic](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.
- Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014. Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers*, pages 156–162.
- Ahmed Alwajeih, Mahmoud Al-Ayyoub, and Ismail Hmeidi. 2014. [On authorship authentication of arabic articles](#). In *2014 5th International Conference on Information and Communication Systems (ICICS)*, pages 1–6.
- L Bouchakour and M Debyeche. 2018. Improving continuous arabic speech recognition over mobile networks dsr and nsr using mfccs features transformed. *International Journal of Circuits, Systems and Signal Processing*, 12:1–8.
- Paul Cardinal, Ahmed Ali, Najim Dehak, Yifan Zhang, Takahiro A. Hanai, Yu Zhang, James R. Glass, and Stephan Vogel. 2014. [Recent advances in asr applied to an arabic transcription system for al-jazeera](#). In *Proceedings of Interspeech 2014*, pages 2088–2092.
- Frank Diehl, Mark JF Gales, Marcus Tomalin, and Philip C Woodland. 2012. Morphological decomposition in arabic asr systems. *Computer Speech & Language*, 26(4):229–243.
- Amirbek Djanibekov, Hawau Olamide Toyin, Raghada Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. [Dialectal coverage and generalization in Arabic speech recognition](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.
- Dongji Gao, Hainan Xu, Desh Raj, Leibny Paola Garcia Perera, Daniel Povey, and Sanjeev Khudanpur. 2023. Learning from flawed data: Weakly supervised automatic speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. [Conformer: Convolution-augmented transformer for speech recognition](#). *Preprint*, arXiv:2005.08100.
- Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, page 102422.
- Taku Kudo and John Richardson. 2018. [Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). *Preprint*, arXiv:1808.06226.
- Liang Lu, Changliang Liu, Jinyu Li, and Yifan Gong. 2020. Exploring transformers for large-scale speech recognition. *arXiv preprint arXiv:2005.09684*.
- Mohamed Amine Menacer, Odile Mella, Dominique Fohr, Denis Jouviet, David Langlois, and Kamel Smaïli. 2017. [Development of the arabic loria automatic speech recognition system \(alasn\) and its evaluation for algerian dialect](#). *Procedia Computer Science*, 117:81–88. Arabic Computational Linguistics.
- Amr Mousa, Hong-Kwang Kuo, Lidia Mangu, and Hagen Soltau. 2013. [Morpheme-based feature-rich language models using deep neural networks for lvcsr of egyptian arabic](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8435–8439.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023a. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023b. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Mahmoud Salhab, Marwan Elghitany, Shameed Sait, Syed Sibghat Ullah, Mohammad Abusheikh, and Hasan Abusheikh. 2025. [Advancing arabic speech recognition through large-scale weakly supervised learning](#). *Preprint*, arXiv:2504.12254.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou cheikh tourad, Rahaf Alhamouri, Rwa

- Assi, Aisha Alraeesi, Hour Mohamed, Fakhreddin Al-wajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. [Casablanca: Data and models for multidialectal arabic speech recognition](#). *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.
- Jayashri Vajpai and Avnish Bora. 2016. Industrial applications of automatic speech recognition systems. *International Journal of Engineering Research and Applications*, 6(3):88–95.
- Abdul Waheed, Bashar Talafha, Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Voxarabica: A robust dialect-aware arabic speech recognition system. *arXiv preprint arXiv:2310.11069*.
- Yongqiang Wang, Yangyang Shi, Frank Zhang, Chunyang Wu, Julian Chan, Ching-Feng Yeh, and Alex Xiao. 2021. [Transformer in action: A comparative study of transformer-based acoustic models for large scale speech recognition applications](#). In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6778–6782.