# Parameter-Efficient Adaptation of Self-Supervised Models for Arabic Speech Recognition

Wafa Alshehri Wasfi G. Al-Khatib Mohammad Amro
Information and Computer Science Department
Interdisciplinary Research Center for Intelligent Secure Systems
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia
{g202427480@kfupm.edu.sa, wasfi@kfupm.edu.sa, mamro@kfupm.edu.sa}

## Abstract

Arabic speech recognition systems face distinct challenges due to the language's complex morphology and dialectal variations. Self-supervised models (SSL) like XLS-R have shown promising results, but their size with over than 300 million of parameters, makes fine-tuning computationally expensive. In this work, we present the first comparative study of parameter-efficient fine-tuning (PEFT), specifically LoRA and DoRA, applied to XLS-R for Arabic ASR. We evaluate on the newly released Common Voice Arabic V24.0 dataset, establishing new benchmarks. Our full fine-tuning achieves state-of-the-art results among XLS-R-based models with 23.03% Word Error Rate (WER). In our experiments, LoRA achieved a 36.10% word error rate (WER) while training just 2% of the model's parameters. DoRA reached 45.20% WER in initial experiments. We analyze the trade-offs between accuracy and efficiency, offering practical guidance for developing Arabic ASR systems when computational resources are limited. The models and code are publicly available.

## 1 Introduction

Arabic, spoken by over 420 million people across 22 countries (Dhouib et al., 2022), poses distinct challenges for automatic speech recognition (ASR). First, Arabic orthography typically omits diacritical marks indicating short vowels, creating ambiguity where identical written forms have multiple pronunciations and meanings. Second, substantial dialectal variation exists across Arabic regions, yet labeled speech datasets predominantly cover Modern Standard Arabic (MSA). Third, Arabic's complex morphology characterized by intricate affixation patterns, produces extensive vocabularies that increase out-of-vocabulary rates and complicate language modeling. These factors, combined with limited large-scale labeled speech corpora compared to English, have constrained Arabic ASR development. Self-supervised learning (SSL) addresses data scarcity in ASR through models like wav2vec 2.0 (Baevski et al., 2020) and its multilingual extension XLS-R (Conneau et al., 2020), which learn speech representations from unlabeled audio and enable effective fine-tuning with limited labeled data. XLS-R, pretrained on 53 languages including Arabic, demonstrates strong cross-lingual transfer and achieves state-of-the-art ASR results. However, these models like XLS-R contains approximately 317 million parameters, makes full fine-tuning computationally expensive and often infeasible for resource-constrained researches.

Parameter-efficient fine-tuning (PEFT) methods address this computational challenge. Low-Rank Adaptation (LoRA) (Hu et al.) freezes pretrained weights and introduces small trainable low-rank matrices, drastically reducing trainable parameters while maintaining performance. Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) extends LoRA by decomposing updates into magnitude and direction components, potentially improving adaptation quality. While PEFT methods are increasingly explored for ASR (Song et al., 2024; Omar et al.), a significant gap exists at the intersection of PEFT and Arabic ASR. Existing PEFT studies for ASR focus predominantly on Whisper (Omar et al.), an encoder-decoder architecture, while Arabic ASR research using SSL models (Younis et al.; Talafha et al., 2023) relies exclusively on full fine-tuning. To our knowledge, no prior work has applied LoRA to CTC-based self-supervised models like XLS-R for Arabic ASR, nor has DoRA been explored for Arabic speech recog-



*Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026)*, pages 322–328

nition on any model architecture. This paper addresses these gaps through a comparative study of PEFT methods for Arabic ASR using XLS-R. Our contributions include:

- First application of LoRA and DoRA to XLS-R for Arabic ASR, achieving competitive performance while training only 2.2% of parameters.
- State-of-the-art results among XLS-R Arabic models with 23.03% WER on Arabic Common Voice (CV) V24.0, the first evaluation on this dataset.
- Accuracy-efficiency trade-off analysis, demonstrating LoRA achieves 36.10% WER with 47× smaller adapter storage than full fine-tuning.
- Release of trained models and code for reproducibility and future research (Alshehri et al., 2026).

The remainder of this paper is organized as follows. Section 2 reviews related work on PEFT methods for ASR and Arabic speech recognition using SSL models. Section 3 describes our methodology, including the dataset, model architecture, and fine-tuning approaches. Section 4 presents experimental results, Section 5 discusses the findings, Section 6 concludes the paper, and Section 7 presents a dedicated Limitations section summarizing the main constraints of the study and directions for future work.

## 2 Related Work

PEFT methods have been increasingly explored for speech recognition, with LoRA being applied to Whisper for multilingual settings (Song et al., 2024; Kwok et al., 2025), Turkish (Polat et al., 2024), and Japanese (Bajo et al.). However, previous studies have focused predominantly on encoder-decoder architectures or high-resource languages. For Arabic, only Omar et al. (Omar et al.) have applied a PEFT method, using LoRA on Whisper for multi-dialectal ASR.

Beyond LoRA and DoRA, other PEFT approaches such as adapters and prefix-tuning have been proposed. Adapter-based methods introduce additional bottleneck layers between Transformer blocks, while prefix-tuning prepends learnable virtual tokens to attention mechanisms. Although effective, these approaches require architectural modifications or additional inference-time components. In contrast, LoRA and DoRA directly modify existing weight matrices with minimal overhead, making them particularly suitable for large CTC-based self-supervised speech models such as XLS-R.

Arabic ASR using self-supervised learning remains significantly underexplored compared to other languages. Table 1 summarizes the limited studies in this area. Younis and Mohammad (Younis et al.) compared SSL models for Arabic, with XLS-R achieving 40% WER on Common Voice (CV) dataset. Talafha et al. (Talafha et al., 2023) evaluated Whisper and XLS-R under various settings, reaching 31.16% WER on CV 11.0. Toyin et al. (Toyin et al.) developed ArTST, an Arabic-specific speech transformer based on the SpeechT5 architecture, achieving 12.8% WER on the Multi-Genre Broadcast (MGB-2) dataset outperforming multilingual models like Whisper. Alkanhal et al. (Alkanhal et al., 2023) introduced the Aswat dataset (732 hours) with wav2vec and data2vec pretraining, achieving state-of-the-art WERs of 10.3% on MGB-2 and 11.7% on CV. Notably, no prior work has explored PEFT methods on CTC-based self-supervised models for Arabic, nor has DoRA been evaluated for Arabic speech recognition on any architecture. All existing Arabic SSL studies rely exclusively on full fine-tuning, leaving parameter-efficient approaches unexplored for this morphologically complex language.

This gap is particularly significant given Arabic's unique challenges, which make efficient fine-tuning methods especially valuable. Our work addresses this gap by: applying LoRA to XLS-R, a CTC-based self-supervised model, for Arabic ASR the first such study to our knowledge; evaluating DoRA for Arabic speech recognition, which has not been explored on any model architecture; and providing the first results on the newly released CV Arabic V24.0 dataset. These contributions provide a more complete understanding of PEFT applicability for Arabic ASR.

Table 1: Prior Work on SSL-Based Arabic ASR

| Study | Model | Method | Dataset |
|---|---|---|---|
| Younis & Mohammad (2023) (Younis et al.) | HuBERT, XLS-R, MMS | Full FT | CV |
| Talafha et al. (2023) (Talafha et al., 2023) | Whisper, XLS-R | Full FT | CV, MGB-2/3/5, FLEURS |
| Toyin et al. (2023) (Toyin et al.) | ArTST | Full FT | MGB-2 |
| Alkanhal et al. (2023) (Alkanhal et al., 2023) | wav2vec2, data2vec | Full FT | CV, MGB-2, Aswat |
| Alharbi et al. (2024) (Alharbi et al., 2024) | XLS-R, Whisper, MMS | Full FT | SADA |
| Alrashoudi et al. (2024) (Alrashoudi et al., 2024) | Wav2Vec2, HuBERT, Whisper | Full FT | Shehri (Jibbali) |
| Omar et al. (2024) (Omar et al.) | Whisper-Small | LoRA | CV 16.1, MASC |

## 3 METHODOLOGY

This section describes the experimental setup for evaluating PEFT methods for Arabic ASR, including dataset, preprocessing, model architecture, fine-tuning approaches, and evaluation metrics.

### 3.1 Dataset

We use Mozilla CV Arabic version 24.0 (2025 release) (Com), the first evaluation on this version. The dataset contains crowdsourced recordings from various speakers and environments, capturing real-world diversity in Arabic pronunciation and regional accents. Inconsistent audio quality with varying background noise makes the dataset challenging yet representative of real-world scenarios. The dataset includes approximately 92 hours of validated speech data using official train, development, and test splits with 28,881, 10,181, and 10,508 samples, respectively. For preprocessing, audio recordings are resampled to 16 kHz mono format for XLS-R compatibility. Text normalization includes: removing punctuation, Arabic diacritical marks (tashkeel), and non-Arabic characters; and unifying letter forms (e.g., أ, إ, آ to ا). This reduces vocabulary complexity and focuses the model on phonetic-to-character mappings. We employ a character-level tokenizer with 39 tokens: 34 Arabic characters (ء, ؤ, ئ, ا, ب, ة, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ك, ل, م, ن, ه, و, ي, ى) and special tokens for padding [PAD], unknown [UNK], word boundary |, and sequence markers <s>, </s>.

### 3.2 Base Model

We adopt wav2vec2-XLSR-53 (Conneau et al., 2020) as our base model. Wav2vec 2.0 (Baevski et al., 2020) is a self-supervised learning framework with a CNN feature encoder processing raw audio and a Transformer context network capturing long-range dependencies. During pretraining, it learns representations through contrastive learning by masking latent speech representations and identifying correct segments from distractors, enabling learning from unlabeled audio. XLSR-53 (Conneau et al., 2020) extends wav2vec 2.0 to multilingual settings, pretrained on 53 languages including Arabic. This enables cross-lingual transfer where high-resource language knowledge benefits low-resource languages. The model consists of 24 Transformer layers with 1024 model dimension, 4096 feed-forward dimension, 16 attention heads, and approximately 317M total parameters. For ASR fine-tuning, we add a linear classification head projecting contextualized representations to character vocabulary. Following standard practice, we freeze the CNN feature extractor and update only Transformer layers and classification head. The model is trained using CTC loss, enabling alignment-free training by marginalizing over all possible alignments between input audio and output character sequence.

### 3.3 Fine-Tuning Approaches

We compare three approaches: full fine-tuning (baseline) and two PEFT methods, LoRA and DoRA. We focus on LoRA and DoRA as they provide parameter-efficient adaptation without modifying the base architecture or inference pipeline, which is desirable for large-scale CTC-based SSL speech models.

#### 3.3.1 Full Fine-Tuning

serves as our baseline approach, updates all pretrained XLS-R Transformer encoder parameters and the classification head during training (CNN feature extractor remains frozen). While achieving best performance through full adaptation, it requires substantial computational

resources and storage.

#### 3.3.2 LoRA (Low-Rank Adaptation)(Hu et al.)

freezes pretrained weights and injects trainable low-rank matrices. For weight matrix $W \in \mathbb{R}^{d\times k}$ where d is the input dimension and k is the output dimension, LoRA represents update as:

$$W' = W + \Delta W = W + BA \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll \min(d, k)$. Here, r denotes the low-rank bottleneck dimension, which controls the adapter capacity and parameter count. Only $A$ and $B$ are updated, while $W$ remain frozen, dramatically reducing trainable parameters. Scaling factor $\alpha$ controls update magnitude with effective learning rate $\alpha/r$. Dropout is applied for regularization.

#### 3.3.3 DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al., 2024)

extends LoRA by decomposing weights into magnitude and direction:

$$W' = m \cdot \frac{W + BA}{\|W + BA\|} \tag{2}$$

where $m$ is a learnable magnitude vector, enabling independent adjustment of magnitude and direction for potentially improved stability and performance. In our implementation, DoRA uses the same as LoRA configuration but enables the weight decomposition through the use_dora=True flag in the PEFT library.

### 3.4 Experimental Configuration

Table 2 presents the experimental configuration for all fine-tuning approaches. We apply LoRA and DoRA adapters to attention projection matrices and feed-forward layers to maximize adaptation capacity. To ensure fair comparison, we use identical hyperparameters across all experiments, isolating the effect of fine-tuning methods. All experiments use early stopping based on validation WER with patience of 3 epochs, resulting in 50 epochs for full fine-tuning, 41 for LoRA, and 33 for DoRA

### 3.5 Evaluation Metrics

We evaluate performance using Word Error Rate (WER) and character Error Rate (CER).

Table 2: Training Configuration

| Category | Parameter | Value |
|---|---|---|
| Environment | | |
| | Platform | Google Colab Pro+ |
| | GPU | NVIDIA A100 (80 GB) |
| | System RAM | 167 GB |
| Training | | |
| | Batch size | 32 |
| | Grad. accum. | 2 (effective $\approx$ 64) |
| | Learning rate | $3 \times 10^{-4}$ |
| | Max epochs | 50 |
| | Early stopping | Val. WER (patience 3) |
| | Warmup ratio | 0.1 |
| | Optimizer | AdamW |
| LoRA/DoRA | | |
| | Rank ($r$) | 16 |
| | Alpha ($\alpha$) | 32 |
| | Dropout | 0.05 |
| | Bias | None |
| | Target modules | q_proj, k_proj, v_proj, out_proj, intermediate_dense, output_dense |

WER measures word-level errors:

$$\text{WER} = \frac{S + I + D}{N} \tag{3}$$

where $S$, $I$, $D$ denote substitutions, insertions, deletions, and $N$ is total words in reference. CER applies the same formulation at character level, providing fine-grained assessment particularly relevant for morphologically rich Arabic, where word-level errors may reflect minor character mistakes. We also report trainable parameters and model size to analyze accuracy-efficiency trade-offs.

## 4 Results

Table 3 and Figure 1 present results for the three approaches. Full fine-tuning achieves 23.03% WER and 6.7% CER, the lowest among all reported XLS-R Arabic models (Table 4). LoRA obtains 36.10% WER and 9.6% CER training only 2.2% of parameters, competitive with fully fine-tuned models including mohammed/xlsr-arabic (36.70%) and jonatasgrosman/xlsr-arabic (39.59%). DoRA achieves 45.20% WER and 12.54% CER with similar parameter efficiency. Table 5 shows LoRA and DoRA require approximately 2% of full fine-tuning parameters with small adapter sizes

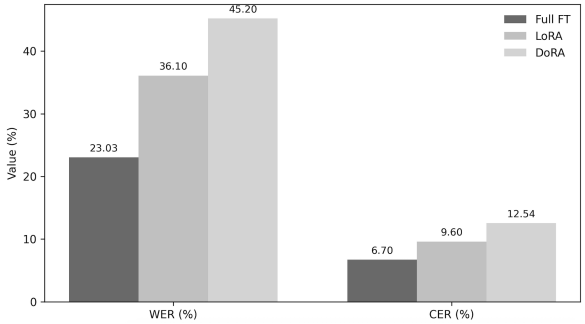


Figure 1: WER and CER comparison between full fine-tuning and PEFT methods on the CV Arabic V24.0 test set.

Table 3: Main Results on CV Arabic V24.0 Test Set

| Method | WER (%) | CER (%) |
|---|---|---|
| Full Fine-Tuning | 23.03 | 6.7 |
| LoRA | 36.10 | 9.6 |
| DoRA | 45.20 | 12.54 |

(27-28 MB), enabling efficient deployment of multiple adapters sharing a single base model. Direct comparison with previous work is limited by dataset version and training data differences

## 5 Discussion

Table 4 and Figure 2 shows that our full fine-tuning achieves state-of-the-art results among XLS-R Arabic models (23.03% WER), a 3.52 percentage point improvement over elgeish/xlsr-53-arabic (26.55%). This is notable given training exclusively on CV Arabic V24.0 without additional corpora, while competing models used supplementary datasets, validating XLS-R's multilingual pretraining effectiveness and demonstrating that single high-quality dataset fine-tuning achieves competitive performance. LoRA demonstrates parameter-efficient fine-tuning viability with 36.10% WER using only 2.2% trainable parameters, competitive with several fully fine-tuned models. This enables: Arabic ASR development in resource-constrained environments; efficient deployment of multiple specialized models via small adapters (27 MB) sharing one base model; and rapid experimentation. DoRA achieved 45.20% WER, underperforming LoRA and full fine-tuning, contrasting with improvements in other domains (Bhattacharjee et al.). DoRA's weight decomposition requires speech-specific hyperparameter tuning; our identical hyperparameters for fair comparison may not optimize DoRA. DoRA converged faster (33 vs. 41 epochs for LoRA), suggesting premature stopping. Despite lowest performance, DoRA outperforms several fully fine-tuned models including kmfoda/xlsr-arabic (52.53%), demonstrating parameter-efficient methods can exceed some full fine-tuning using only 2.3% of parameters. Future work should explore DoRA-specific optimization.

Our results reveal clear accuracy-efficiency trade-offs (Table 5, Figure 1). The 13.07 percentage point gap between full fine-tuning and LoRA represents parameter efficiency cost. This performance gap is particularly pronounced for Arabic ASR due to several

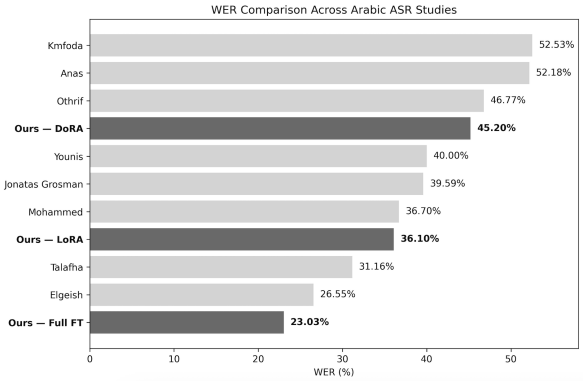


Figure 2: WER comparison across Arabic ASR studies. Our Full Fine-Tuning achieves the lowest WER.

Table 4: Comparison with Previous XLS-R Arabic Models

| Model | Dataset | WER | CER |
|---|---|---|---|
| Models trained on CV + SC | | | |
| jonatasgrosman/xlsr-arabic | CV 6.0 + SC | 39.59 | 18.18 |
| mohammed/xlsr-arabic | CV 13.0 + SC | 36.70 | – |
| elgeish/xlsr-53-arabic | CV + SC | 26.55 | – |
| Models trained on CV only | | | |
| kmfoda/xlsr-arabic | CV 13.0 | 52.53 | – |
| anas/xlsr-arabic | CV | 52.18 | – |
| othrif/xlsr-arabic | CV 13.0 | 46.77 | – |
| (Younis et al.) | CV | 40.0 | – |
| (Talafha et al., 2023) | CV 11.0 | 31.16 | 9.35 |
| Ours (Full FT) | CV 24.0 | 23.03 | 6.7 |
| Ours (LoRA) | CV 24.0 | 36.10 | 9.6 |
| Ours (DoRA) | CV 24.0 | 45.20 | 12.54 |

Table 5: Efficiency Comparison

| Metric | Full FT | LoRA | DoRA |
|---|---|---|---|
| Trainable params | 317M (100%) | 7.1M (2.2%) | 7.3M (2.3%) |
| Param. reduction | – | 97.8% | 97.7% |
| Adapter size | – | 27 MB | 27.9 MB |
| Storage reduction | – | 47× | 45× |

language and data-specific factors. Arabic's rich morphology and large character-level vocabulary require substantial representational adaptation, which may not be fully captured by low-rank updates alone. In addition, dialectal variation and pronunciation diversity introduce acoustic variability that benefits from updating a larger portion of model parameters, as in full fine-tuning. The crowd-sourced nature of Common Voice further increases heterogeneity in speaker traits, recording conditions, and background noise, amplifying the need for more expressive adaptation. While PEFT methods such as LoRA significantly reduce trainable parameters, their constrained update capacity can limit performance in linguistically and acoustically complex settings such as Arabic ASR. Moreover, to ensure a fair and controlled comparison, identical hyperparameters were used across fine-tuning methods, which may not optimally exploit the full capacity of PEFT techniques and could further contribute to the observed performance gap.

The observed performance levels reflect inherent challenges of Arabic ASR, including morphological richness that creates large and complex vocabularies, substantial speaker and dialectal variation, and crowd-sourced recordings with background noise and variable quality. These factors help explain the performance gap between PEFT methods and full fine-tuning, as low-rank adaptations may be insufficient to fully capture Arabic speech complexity in heterogeneous datasets. Based on these findings, we recommend full fine-tuning when maximum accuracy is required and computational resources permit, LoRA for resource-constrained or rapid prototyping scenarios (97.8% parameter reduction), leveraging lightweight adapters for deploying multiple specialized models, and further hyperparameter optimization for DoRA prior to deployment.

## 6 Conclusion

This paper presented the first application of PEFT methods to CTC-based self-supervised models for Arabic ASR using XLS-R. We evaluated full fine-tuning, LoRA, and DoRA on Common Voice Arabic v24.0. Full fine-tuning achieved state-of-the-art performance among XLS-R Arabic models (23.03% WER), while LoRA attained competitive results (36.10% WER) using only 2.2% trainable parameters and 47× smaller adapters. DoRA achieved 45.20% WER. These findings confirm the viability of parameter-efficient fine-tuning for Arabic ASR in resource-constrained environments.

## 7 Limitations

While using a single dataset for evaluation and identical hyperparameters across the different approaches ensure a controlled comparison, they also introduce limitations. Evaluation on Common Voice Arabic v24.0 alone may not fully capture the diversity of Arabic speech across domains and dialects. In particular, future evaluations on additional Arabic corpora such as MGB-2, Aswat, and other dialectal datasets would strengthen the generalizability of our conclusions. In addition, using identical hyperparameters for all fine-tuning approaches may not optimally reflect the full potential of individual PEFT techniques, particularly DoRA. Furthermore, our experiments are limited to a single self-supervised architecture (XLS-R). Future work should consider hyperparameter optimization tailored to each PEFT method, exploration of alternative PEFT approaches such as adapter-based methods, prefix-tuning, and adaptive low-rank variants (e.g., AdaLoRA), and assessment of other SSL architectures (e.g., HuBERT, WavLM, larger XLS-R variants).

## Acknowledgment

The authors would like to acknowledge the funding support provided by the Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS) at King Fahd University of Petroleum & Minerals under Project No. INSS2522.

## References

Common Voice Scripted Speech 24.0 - Arabic | Mozilla Data Collective.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. SADA: SAUDI AUDIO DATASET FOR ARABIC. ICASSP, IEEE International

Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 10286–10290.

L Alkanhal, A Alessa, E Almahmoud … of Arabic-NLP 2023, and undefined 2023. 2023. Aswat: Arabic audio dataset for automatic speech recognition using speech-representation learning. aclanthology.org, pages 120–127.

NA Alrashoudi, OS Alshahri Proceedings of the 6th …, and undefined 2024. 2024. Arabic Speech Recognition of zero-resourced Languages: A Case of Shehri (Jibbali) Language. aclanthology.orgNA Alrashoudi, OS Alshahri, H Al-KhalifaProceedings of the 6th Workshop on Open-Source Arabic Corpora and, 2024 • aclanthology.org, pages 84–92.

Wafa Alshehri, Wasfi Al-Khatib, and Mohammad Amro. 2026. Parameter-efficient fine-tuning of xls-r for arabic speech recognition.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 33:12449–12460.

M Bajo, H Fukukawa, R Morita arXiv preprint arXiv …, and undefined 2024. Efficient adaptation of multilingual models for japanese asr. arxiv.orgM Bajo, H Fukukawa, R Morita, Y OgasawaraarXiv preprint arXiv:2412.10705, 2024 • arxiv.org.

S Bhattacharjee, J Mishra, HS Shekhawat arXiv preprint arXiv …, and undefined 2025. Parameter-Efficient Fine-Tuning of Foundation Models for CLP Speech Classification. arxiv.org.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 1:346–350.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, Aisha Al Sinani, 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 2022, Vol. 12,, 12(17).

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arxiv.org.

Chin Yuen Kwok, Hexin Liu, Jia Qi Yip, Sheng Li, and Eng Siong Chng. 2025. A Two-Stage LoRA Strategy for Expanding Language Capabilities in Multilingual ASR Models. IEEE Transactions on Audio, Speech and Language Processing, 33:2576–2590.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. openreview.net.

Z Omar, A Abdelazim, M Gomaa, K Ali, and A Jamal. Parameter-Efficient Fine-Tuning of Whisper for Multi-Dialectal Arabic ASR. researchgate.netZ Omar, A Abdelazim, M Gomaa, K Ali, A Jamal, A Faresresearchgate.net.

Hüseyin Polat, Alp Kaan Turan, Cemal Koçak, Hasan Basri Ulaş, Hüseyin Polat, Alp Kaan Turan, Cemal Koçak, and Hasan Basri Ulaş. 2024. Implementation of a Whisper Architecture-Based Turkish Automatic Speech Recognition (ASR) System and Evaluation of the Effect of Fine-Tuning with a Low-Rank Adaptation (LoRA) Adapter on Its Performance. Electronics 2024, Vol. 13,, 13(21).

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 3934–3938.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-August:5092–5096.

HO Toyin, A Djanibekov, A Kulkarni arXiv preprint arXiv …, and undefined 2023. ArTST: Arabic text and speech transformer. arxiv.org.

HA Younis, YF Mohammad 2023 16th International, and undefined 2023. Arabic speech recognition based on self supervised learning. ieeexplore.ieee.orgHA Younis, YF Mohammad2023 16th International Conference on Developments in eSystems, 2023 • ieeexplore.ieee.org.