

# MARCO-ASR: A PRINCIPLED AND METRIC-DRIVEN FRAMEWORK FOR FINE-TUNING LARGE-SCALE ASR MODELS FOR DOMAIN ADAPTATION

Xuanfan Ni, Fei Yang, Fengping Tian, Qingjuan Li, Chenyang Lyu\*, Yichao Du,  
Longyue Wang, Weihua Luo, Kaifu Zhang

Alibaba International Digital Commerce

## ABSTRACT

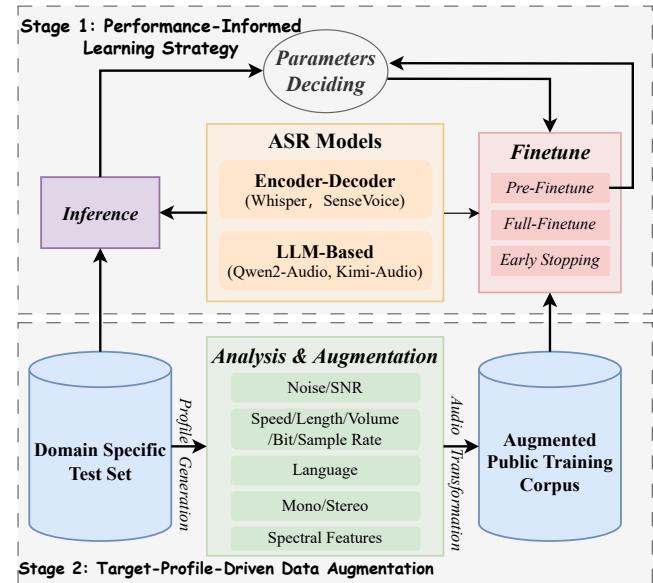
Automatic Speech Recognition (ASR) models have achieved remarkable accuracy in general settings, yet their performance often degrades in domain-specific applications due to data mismatch and linguistic variability. This challenge is amplified for modern Large Language Model (LLM)-based ASR systems, whose massive scale and complex training dynamics make effective fine-tuning non-trivial. To address this gap, this paper proposes a principled and metric-driven fine-tuning framework for adapting both traditional and LLM-based ASR models to specialized domains. The framework emphasizes learning rate optimization based on performance metrics, combined with domain-specific data transformation and augmentation. We empirically evaluate our framework on state-of-the-art models—including Whisper, Whisper-Turbo, and Qwen2-Audio—across multi-domain, multilingual, and multi-length datasets. Our results not only validate the proposed framework but also establish practical protocols for improving domain-specific ASR performance while preventing overfitting.

**Index Terms**— Automatic Speech Recognition, Fine-Tuning, Domain Adaptation

## 1. INTRODUCTION

Automatic Speech Recognition (ASR) systems, powered by large-scale pre-trained models like Whisper [1], have become pervasive in various applications. More recently, the integration of Large Language Models (LLMs) has given rise to a new generation of powerful audio-foundation models, such as Qwen2-Audio [2] and Kimi-Audio [3], which exhibit even stronger capabilities in understanding and transcribing speech. However, despite their impressive performance on general-purpose benchmarks, these off-the-shelf models often falter when deployed in specialized, or out-of-domain (OOD), scenarios like medical transcription, legal dictation, or financial calls. This performance drop is primarily due to distribution shifts in acoustic conditions, speaker characteristics, and, most critically, domain-specific terminology.

While fine-tuning is a conventional approach for domain adaptation, applying it to modern LLM-based ASR models presents unique and significant challenges. Their vast parameter spaces and intricate architectures make them highly sensitive to training configurations. For instance, selecting an appropriate learning rate is not merely a matter of optimization but is critical to prevent catastrophic forgetting or training instability. Furthermore, our empirical findings reveal that traditional indicators of training progress, such as a decreasing loss, do not reliably correlate with improvements in transcription accuracy (i.e., Word Error Rate, WER [4]), making the fine-tuning



**Fig. 1:** An overview of our framework, which consists of: (1) hyper-parameter optimization based on model architecture and performance, and (2) domain-specific data analysis and augmentation.

process opaque and difficult to control. These complexities necessitate a more structured and principled approach to adaptation.

To address these issues, this paper introduces a principled and metric-driven fine-tuning framework designed to be applicable to both traditional encoder-decoder and newer LLM-based ASR models. Our framework provides a systematic guide through the critical stages of domain adaptation: (1) principled hyperparameter selection, with a focus on identifying optimal learning rates through pre-finetuning trials and continuous model performance analysis using validation metrics and (2) rigorous data filtering and augmentation to prepare high-quality, domain-relevant training sets.

Through extensive experiments, we first demonstrate the extent of the OOD problem by showing how leading models underperform on various domain-specific and low-resource language datasets. We then apply our proposed framework to systematically fine-tune these models, presenting detailed results that highlight the impact of learning rate choices and the effectiveness of data augmentation. Our analysis provides practical insights and establishes a clear protocol for practitioners to effectively adapt large-scale ASR models for their specific needs, transforming them from general-purpose tools into highly accurate, domain-specialized solutions.

\* Corresponding Author.

## 2. A PRINCIPLED, METRIC-DRIVEN DOMAIN ADAPTATION FRAMEWORK

We formalize our framework for fine-tuning ASR models on domain adaptation, as shown in Fig 1.

### 2.1. Performance-Informed Learning Strategy

The fine-tuning framework is initiated by quantifying a baseline performance metric on the target domains by calculating the WER on a subset of the target dataset (e.g., 200 instances). This baseline metric, in tandem with the model’s architectural paradigm, is a critical determinant for selecting the optimal initial learning rate ( $\eta$ ).

#### 2.1.1. Encoder-Decoder ASR Models

Fine-tuning these models requires a balanced learning rate. A low rate is stable but slow, while a high rate can destabilize training and cause the WER to fluctuate. To address this trade-off, we employ a dynamic learning rate strategy based on discrete training cycles. Our strategy partitions the fine-tuning process into cycles, where each cycle  $j$  uses a constant learning rate,  $\eta_j$ . Throughout cycle  $j$ , the model is trained for a fixed number of steps, with periodic evaluations on the validation set. This process generates a list of WER measurements within that single cycle. We then quantify the training stability by calculating the standard deviation of this list of WERs, denoted as  $\sigma_{WER}(j)$ .

The learning rate for the subsequent cycle,  $\eta_{j+1}$ , is adapted based on training stability. A high  $\sigma_{WER}(j)$  indicates instability, prompting a smaller learning rate, while a low value signals stable convergence, allowing for a larger learning rate. This is formalized by the update rule:

$$\eta_{j+1} = \eta_{\max} - (\eta_{\max} - \eta_{\min}) \cdot \text{clip}\left(\frac{\sigma_{WER}(j)}{\sigma_{\text{ref}}}, 0, 1\right) \quad (1)$$

where  $[\eta_{\min}, \eta_{\max}]$  is the allowed learning rate range, and  $\sigma_{\text{ref}}$  is a reference standard deviation.

#### 2.1.2. LLM-Based ASR Models

For LLM-based ASR models like *Qwen2-Audio* and *Kimi-Audio*, we conceptualize the initial baseline  $WER_0$ , as a quantitative proxy for the domain mismatch between the source (pre-training) and target datasets. The learning rate,  $\eta_{\text{llm}}$ , is therefore dynamically calibrated as a function of this domain gap. A significant mismatch, manifested as a high  $WER_0$ , necessitates a larger learning rate for accelerated convergence and substantial parameter adaptation. Conversely, a minimal domain gap (low  $WER_0$ ) warrants a smaller rate for fine-grained refinement. This dynamic scaling is formalized as:

$$\eta_{\text{llm}} = \eta'_{\min} + (\eta'_{\max} - \eta'_{\min}) \cdot \text{clip}(WER_0, 0, 100)/100 \quad (2)$$

where  $\eta'_{\min}$  (e.g.,  $1 \times 10^{-6}$ ) and  $\eta'_{\max}$  (e.g.,  $1 \times 10^{-4}$ ) define the learning rate boundaries for this model family. The  $\text{clip}(\cdot)$  function constrains the  $WER_0$  value to the interval  $[0, 1]$ , allowing the learning rate to scale linearly with the perceived domain gap.

We fine-tune the ASR model starting with an empirically derived initial learning rate. During this process, the learning rate is adjusted based on the WER in accordance with the formulas above, and training is halted once the WER converges and shows no significant improvement. The empirical validation and comparative analysis of these learning rate strategies are detailed in Section 4.2.

### 2.2. Target-Profile-Driven Data Augmentation

To bridge the domain gap between public training corpora and domain datasets, we employ a profile-driven data augmentation strategy. This two-phase strategy first analyzes audio from the target domain to build a statistical profile, which then guides the augmentation of clean source data to realistically emulate the target’s characteristics.

In the analysis phase, we generate a quantitative profile of the target audio domain from a representative corpus. This process involves analyzing each audio file to extract a set of technical attributes, namely **sample rate**, **bit depth**, and **channel count**. Concurrently, we compute key acoustic descriptors, including the Signal-to-Noise Ratio (**SNR**), integrated loudness (**LUFS**), the mean **Spectral Centroid** for perceptual brightness, and the **Spectral Rolloff** for spectral shape. The final profile summarizes the dataset by defining a characteristic range of values for each of these metrics.

The synthesis phase transforms a clean source audio signal  $x(t)$  using the generated profile. First, the signal is resampled to a target sample rate and bit depth randomly chosen from the profile’s distributions. Subsequently, a series of augmentations are applied. Reverberation is introduced via convolution with a randomly selected Room Impulse Response (RIR), with the application probability being inversely related to the profile’s mean SNR. Background noise is added, with the target SNR of the resulting mixture being sampled from the range  $[\text{SNR}_{\min}, \text{SNR}_{\max}]$  defined in the profile. The audio’s loudness is then normalized by applying a gain factor to match a target LUFS value drawn from a normal distribution  $\mathcal{N}(\mu_{\text{LUFS}}, \sigma_{\text{LUFS}}^2)$  derived from the profile’s statistics. Finally, to emulate channel effects, spectral shaping is performed using a low-pass or high-pass filter. The choice of filter and its cutoff frequency are guided by the profile’s mean spectral rolloff, thereby simulating conditions such as muffled or telephonic audio.

The entire augmentation process, which transforms a clean signal  $x(t)$  into its augmented version  $x_{\text{aug}}(t)$ , can be formally expressed as a chain of operators:

$$x_{\text{aug}}(t) = (F_{\text{filt}} \circ F_{\text{lufs}} \circ F_{\text{noise}} \circ F_{\text{rev}} \circ R)(x(t)) \quad (3)$$

where  $\circ$  denotes function composition. The operator chain begins with resampling ( $R$ ) to a target sample rate and bit depth. Next,  $F_{\text{rev}}$  probabilistically applies reverberation ( $*h_{\text{rir}}$ ) with a probability  $p_{\text{rev}} \propto 1/\mu_{\text{SNR}}$ . Subsequently,  $F_{\text{noise}}$  adds noise to achieve an SNR drawn from  $U(\text{SNR}_{\min}, \text{SNR}_{\max})$ , and  $F_{\text{lufs}}$  normalizes the loudness to a target LUFS from  $\mathcal{N}(\mu_{\text{LUFS}}, \sigma_{\text{LUFS}}^2)$ . The final step,  $F_{\text{filt}}$ , applies a spectral filter ( $*h_{\text{filt}}$ ) guided by the mean spectral rolloff  $\mu_{\text{rolloff}}$ .

## 3. EXPERIMENTS

### 3.1. Models

We fine-tune and evaluate a selection of representative state-of-the-art (SOTA) ASR models, including Whisper-Large-V3 [1], its light-weight variant Whisper-Large-V3-Turbo, Qwen2-Audio [2], and Kimi-Audio [3]. For comprehensive comparison and to establish strong baselines, we also include SenseVoice [5], SeamlessM4T [6], FireRedASR-AED [7], and Step2-Audio [8]. To focus on the challenges of OOD speech recognition, we have excluded leading ASR models with limited language support, such as Canary [9] and Paraformer [10], from our experiments.

| Models                 | #Param | Zh    | En    | Es    | Ja     | Th    | Tr     | Vi     | Ko    | Id     |
|------------------------|--------|-------|-------|-------|--------|-------|--------|--------|-------|--------|
| <b>Encoder-Decoder</b> |        |       |       |       |        |       |        |        |       |        |
| Whisper-Large-V3       | 1.5B   | 24.77 | 10.26 | 9.14  | 29.35  | 21.92 | 17.20  | 31.46  | 9.50  | 13.67  |
| Whisper-Large-V3-Turbo | 0.8B   | 24.80 | 9.90  | 11.28 | 46.55  | 41.62 | 29.40  | 31.34  | 9.32  | 21.48  |
| SenseVoice             | 0.2B   | 15.11 | 12.73 | --    | 37.13  | --    | --     | --     | 32.41 | --     |
| SeamlessM4T            | 2.3B   | 12.34 | 4.05  | 4.92  | 237.60 | 4.29  | 11.71  | 18.74  | 32.74 | 6.27   |
| FireRedASR-AED         | 1.1B   | 15.11 | 12.73 | --    | 37.13  | --    | --     | --     | 32.41 | --     |
| <b>LLM-Based</b>       |        |       |       |       |        |       |        |        |       |        |
| Qwen2-Audio-Base       | 8.4B   | 20.51 | 5.47  | 11.87 | 37.98  | 97.78 | 102.18 | 114.05 | 26.31 | 73.47  |
| Kimi-Audio-Base        | 9.7B   | 7.12  | 3.45  | 22.35 | 127.07 | 8.55  | 70.30  | 88.11  | 96.11 | 51.26  |
| Step2-Audio-Mini-Base  | 8.3B   | 8.92  | 3.57  | 11.69 | 42.26  | 93.24 | 131.80 | 101.93 | 42.09 | 123.17 |

**Table 1:** Performance of ASR Models on Common Voice21 test set. Columns 3 to 11 in the first row are language codes according to ISO 639.

| Models           | En           |              |              | Zh           |             |
|------------------|--------------|--------------|--------------|--------------|-------------|
|                  | AMI          | Earnings22   | Covo-DA      | Covo-DA      | MDT         |
| SenseVoice       | 23.02        | 32.24        | 49.40        | 21.39        | 15.03       |
| SeamlessM4T      | 1055         | 270.94       | 57.82        | 16.37        | 13.66       |
| Kimi-Audio-Base  | 1100         | 29.98        | 40.50        | 37.07        | <b>7.47</b> |
| Whisper-Large-V3 | 36.24        | 35.92        | 28.47        | 32.60        | 27.37       |
| + FT, 1e-7       | 33.61        | 32.38        | 21.39        | 21.30        | 19.75       |
| + DA             | 21.14        | 28.58        | 20.55        | 16.03        | 17.28       |
| Whisper-Turbo    | 38.59        | 36.90        | 25.12        | 34.56        | 26.38       |
| + FT, 1e-6       | 33.34        | 31.60        | 21.47        | 20.09        | 25.09       |
| + DA             | 19.10        | <b>17.33</b> | <b>18.83</b> | 15.93        | 14.11       |
| Qwen2-Audio-Base | 45.80        | 37.25        | 75.39        | 48.37        | 10.87       |
| + FT, 1e-5       | 34.76        | 43.72        | 56.61        | 17.38        | 10.55       |
| + DA             | <b>15.53</b> | 24.10        | 19.10        | <b>12.73</b> | 7.49        |

**Table 2:** Performance of ASR Models on English and Chinese domain specific datasets. **FT** and **DA** refers to Fine-Tune and Data Augmentation, respectively.

### 3.2. Datasets

We use a diverse suite of public speech datasets across two dimensions in our experiments: (1) **Multilingual** datasets, include Fleurs [11], Common Voice21 [12], and GigaSpeech-2 [13], which cover a wide range of languages and accents. (2) **Multi-Domain** datasets, includes AMI [14] (English meetings), Earnings-22 [15] (English financial calls), WeNet Speech [16] (YouTube), and MDT-AC001<sup>1</sup> (Chinese in-car speech).

Besides public datasets, we perform data augmentation on the CommonVoice corpus to create a new dataset, which we term **Covo-DA**. This process aims to simulate telephony channel conditions by applying a series of audio transformations including reducing the sampling rate and bit depth, applying band-pass filtering, and introducing distortion, saturation, and white noise.

### 3.3. Experimental Setup

For evaluation, we adhere to the official implementations and recommended settings provided by the respective model developers. For fine-tuning, we employ the AdamW optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.995$ . The training is configured with a per-device batch size of 4 and 8 gradient accumulation steps. All fine-tuning experiments are conducted on a server equipped with  $4 \times$ NVIDIA H100 GPUs, resulting in a global batch size of 128. The learning rate bound-

<sup>1</sup><https://www.magicdatatech.cn/datasets/asr/mdt-asr-c001-mandarin-chinese-speech-recognition-corpus>

aries  $[\eta_{\min}, \eta_{\max}]$  are set to  $[10^{-7}, 10^{-5}]$  for Whisper models and  $[10^{-6}, 10^{-4}]$  for LLM-based ASR models, respectively. For our adaptive scheduler, the reference standard deviation  $\sigma_{\text{ref}}$  is fixed at 0.5.

## 4. RESULTS AND ANALYSIS

Our experiments highlight the importance of careful data selection and hyper-parameter tuning in achieving robust domain adaptation. The framework is effective across both traditional and LLM-based ASR models, demonstrating its practical utility in real-world scenarios.

### 4.1. Language and Domain OOD problems

We evaluated the performance of ASR models on the CommonVoice21 test set. The results, presented in Tables 1 and 2, reveal that nearly all models face significant OOD challenge in both language and domain. A clear performance gap is observed between high-resource languages, such as English (**En**) and Spanish (**Es**), and low-resource languages, where models perform poorly on Japanese (**Ja**), Thai (**Th**), and Vietnamese (**Vi**). Furthermore, all models struggle to generalize to domain-specific datasets, especially those with high levels of background noise, such as **phone calls** and **meeting recordings**.

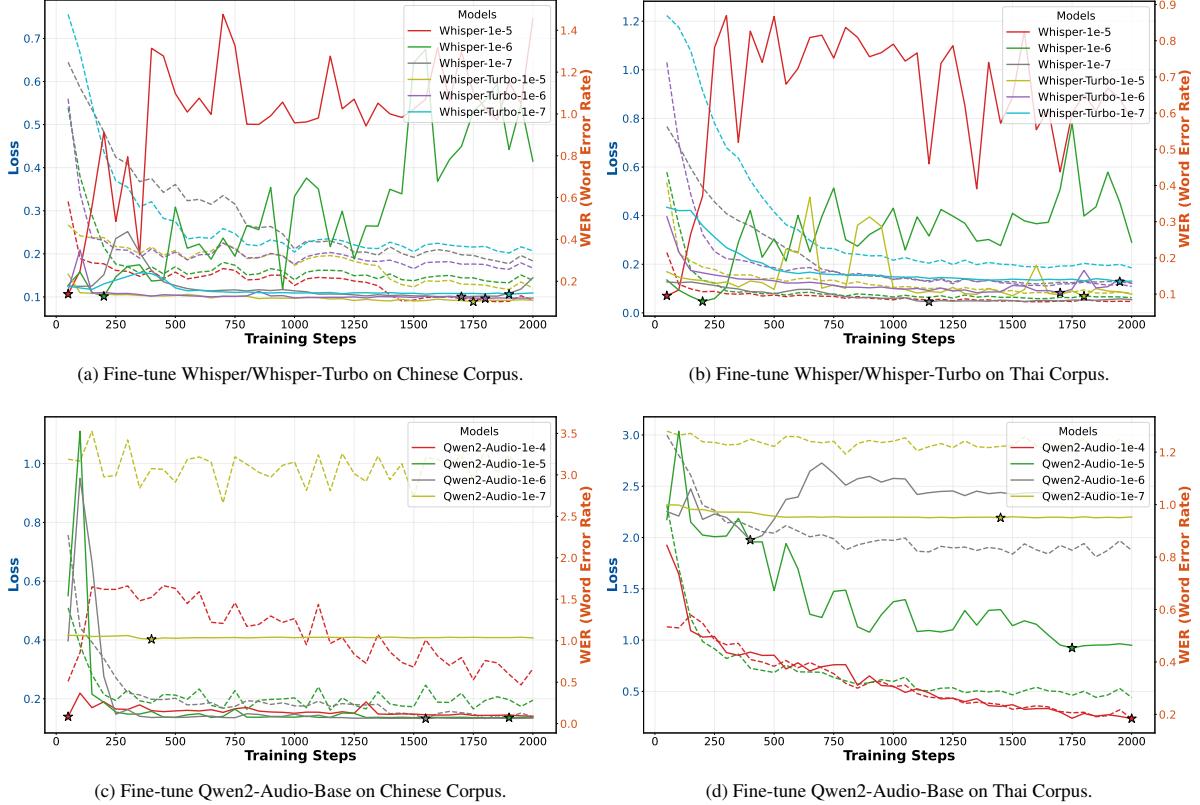
### 4.2. Hyper-Parameter Selection of Fine-tuning ASR Models

We conduct fine-tuning on the three models detailed in Section 3.1, employing a range of hyperparameter combinations for Chinese and Thai. The training corpora consist of a 1000-hour subset of CommonVoice21-Zh combined with WeNetSpeech for Chinese, and a 1000-hour subset of CommonVoice21-Th with GigaSpeech2-Th for Thai. During training, both the loss and WER are recorded at intervals of 50 steps. The comprehensive results are presented in Figure 2.

Our key findings are summarized as follows:

(1) The relationship between training loss and WER is non-linear. As illustrated in Figures 2a and 2b, the Whisper model exhibits a monotonically decreasing loss, whereas its WER score fluctuates significantly. Notably, the optimal WER is often achieved in the early stages of training (e.g., around 1000 steps). This finding highlights the critical need for **well-designed early stopping criteria or pre-finetuning strategies to capture the best-performing checkpoint and prevent overfitting**.

(2) ASR models demonstrate **significant sensitivity to the learning rate, which is tied to both architecture and initial per-**



**Fig. 2:** Comparison of WER and Loss for different models fine-tuned on Chinese and Thai. The metrics are plotted against the number of training steps. In each plot, solid lines denote the WER, while dashed lines represent the training loss. The legend in the top-right corner specifies the model name and the learning rate used for its fine-tuning. The star marker indicates the lowest WER score.

**formance.** For instance, a learning rate of  $10^{-6}$  proves insufficient for Whisper-Large-V3 to reduce WER on the Chinese task. In contrast, Whisper-Large-V3-Turbo, a variant with a shallower decoder, successfully converges with a higher learning rate of  $10^{-5}$  on the same task. Similarly, while Qwen2-Audio achieves its optimal WER on Thai with a learning rate of  $10^{-6}$ , effective fine-tuning on Chinese necessitates an increase to  $10^{-4}$ .

We attribute the learning rate sensitivity in Whisper models primarily to their decoder depth. The 32-layer decoder of Whisper-Large-V3 requires a conservative learning rate to prevent training instability, which can manifest as out-of-vocabulary (OOV) token hallucination or task confusion (i.e., performing translation instead of transcription). In contrast, the much shallower 4-layer decoder in Whisper-Large-V3-Turbo exhibits greater stability, tolerating higher learning rates.

Conversely, LLM-based ASR models often benefit from larger learning rates. Their strong intrinsic language capabilities enable rapid convergence within a few hundred steps, particularly on languages where the model has strong baseline performance. For example, with a  $10^{-4}$  learning rate, Qwen2-Audio-Base achieves a competitive WER on Chinese after only 200 fine-tuning steps.

#### 4.3. Data Augmentation for Domain Specific Fine-tuning

We apply our data augmentation strategy across multiple scenario datasets in both Chinese and English. The Chinese training corpus is the same as described in Section 4.2, while the English training

corpus is entirely sourced from CommonVoice21. The training volume remains at 1,000 hours. We fine-tune ASR models on English and Chinese, using learning rate in Section 4.2 and employing early stopping. As shown in Table 2, the model’s performance on both original and augmented data demonstrates the substantial benefits of our domain-specific augmentation. A clear performance hierarchy emerges: **fine-tuning on augmented data yields the lowest WER**, followed by fine-tuning on original data, with both methods significantly outperforming the baseline pre-trained model. This result validates the effectiveness of our automated augmentation pipeline and our empirical guidelines for selecting fine-tuning learning rates.

## 5. CONCLUSION

This paper presents a unified framework for the domain-specific fine-tuning of ASR models, designed to systematically address key challenges in adapting both traditional architectures and modern LLM-based systems. Our framework integrates two core components: a rigorous hyperparameter optimization process to maximize metric-driven performance, and a principled data augmentation strategy to enhance model generalization and robustness. Through empirical evaluations across a diverse range of models and datasets, we consistently demonstrate the effectiveness of this approach via significant improvements in recognition accuracy. Ultimately, this work provides practitioners with a clear and validated set of guidelines, aiming to transition the field from ad-hoc fine-tuning to a more structured and reliable paradigm for ASR domain adaptation.

## 6. REFERENCES

- [1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” 2022.
- [2] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhi-fang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou, “Qwen2-audio technical report,” 2024.
- [3] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al., “Kimi-audio technical report,” *arXiv preprint arXiv:2504.18425*, 2025.
- [4] Andrew Cameron Morris, Viktoria Maier, and Phil D. Green, “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition,” in *8th International Conference on Spoken Language Processing, INTERSPEECH-ICSLP 2004, Jeju Island, Korea, October 4-8, 2004*. 2004, pp. 2765–2768, ISCA.
- [5] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al., “Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms,” *arXiv preprint arXiv:2407.04051*, 2024.
- [6] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al., “Seamlessm4t: massively multilingual & multimodal machine translation,” *arXiv preprint arXiv:2308.11596*, 2023.
- [7] Kaituo Xu, Feng-Long Xie, Xu Tang, and Yao Hu, “Firedasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to LLM integration,” *CorR*, vol. abs/2501.14350, 2025.
- [8] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al., “Step-audio 2 technical report,” *arXiv preprint arXiv:2507.16632*, 2025.
- [9] Krishna C Puvvada, Piotr Želasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al., “Less is more: Accurate speech recognition & translation without web-scale data,” *arXiv preprint arXiv:2406.19674*, 2024.
- [10] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, “Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition,” *arXiv preprint arXiv:2206.08317*, 2022.
- [11] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna, “FLEURS: few-shot learning evaluation of universal representations of speech,” in *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*. 2022, pp. 798–805, IEEE.
- [12] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
- [13] Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, et al., “Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement,” *arXiv preprint arXiv:2406.11546*, 2024.
- [14] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., “The ami meeting corpus: A pre-announcement,” in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.
- [15] Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra, “Earnings-22: A practical benchmark for accents in the wild,” *CorR*, vol. abs/2203.15591, 2022.
- [16] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng, “WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 2022, pp. 6182–6186, IEEE.