Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning

Ömer Tarik Özyilmaz[1,2], Matt Coler[3], Matias Valdenegro-Toro[2]

[1]Department of Internal Medicine, Division of Nephrology, University of Groningen, University Medical Center Groningen, The Netherlands; [2]Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen, The Netherlands; [3]Speech Technology Lab, Campus Fryslân, University of Groningen, The Netherlands

`o.t.ozyilmaz@umcg.nl, m.coler@rug.nl, m.a.valdenegro.toro@rug.nl`

Abstract

Although commercial Arabic automatic speech recognition (ASR) systems support Modern Standard Arabic (MSA), they struggle with dialectal speech. We investigate the effect of fine-tuning OpenAI's Whisper on five major Arabic dialects (Gulf, Levantine, Iraqi, Egyptian, Maghrebi) using Mozilla Common Voice for MSA and the MASC dataset for dialectal speech. We evaluate MSA training size effects, benefits of pre-training on MSA data, and dialect-specific versus dialect-pooled models. We find that small amounts of MSA fine-tuning data yield substantial improvements for smaller models, matching larger non-fine-tuned models. While MSA pre-training shows minimal benefit, suggesting limited shared features between MSA and dialects, our dialect-pooled models perform comparably to dialect-specific ones. This indicates that pooling dialectal data, when properly balanced, can help address data scarcity in low-resource ASR without significant performance loss.

Index Terms: Automatic Speech Recognition (ASR), dialectal Arabic, fine-tuning, Modern Standard Arabic

1. Introduction

Automatic Speech Recognition (ASR) systems have achieved remarkable performance for major world languages, yet they often fail to serve speakers of regional dialects and variants [1]. This disparity is particularly evident in Arabic, where commercial ASR systems primarily support Modern Standard Arabic (MSA) — the formal, standardized variant used in broadcast media and official communication — while struggling with the rich diversity of regional dialects spoken in daily life. This gap impacts millions of Arabic speakers who rely on their regional dialects for daily communication [2]. The lack of robust dialectal ASR systems creates significant barriers to technology access to those who primarily use their regional dialects in daily life. Without effective dialectal support, users are often forced to code-switch to MSA when using voice assistants, creating an unnatural and potentially exclusionary interaction. This affects applications from healthcare to business where accurate transcription is essential.

Despite Arabic's large speaker population, its dialects face many of the same challenges as typical low-resource languages in ASR development. In their review, Alsayadi et al. [1] identified several critical barriers: training data for Arabic ASR remains skewed towards MSA, dialectal datasets lack volume and variation, and commercial systems that claim multi-dialect support show significant performance degradation when handling regional variants. These challenges are acute in real-world applications, where ASR systems trained on MSA struggle to generalize across the phonological and lexical variations present in dialectal speech. Current models show inconsistent performance across dialects and datasets [3], further complicated by training data imbalances. This imbalance risks creating ASR systems that exhibit systematic bias, performing well for majority dialects while marginalizing speakers of less-resourced variants.

These challenges highlight two critical requirements for advancing dialectal Arabic ASR: developing methods that perform robustly with limited and homogeneous training data, and creating unified models that maintain accuracy across multiple dialects despite dataset imbalances. While large multilingual models like Whisper [4] and USM [5] show promise for low-resource languages, they trade broad coverage for per-language accuracy. Additionally, the substantial computational requirements of these advanced, multilingual models constrain their deployment on devices with limited resources, as optimizing them for such environments could potentially degrade performance for less commonly spoken languages. We investigate whether small models with fine-tuning could achieve similar performance to larger models. Next to this, this paper will focus on leveraging the advantages of both large ASR models and using the large amount of MSA speech data to develop a novel dialectal Arabic ASR system and investigate its strengths and weaknesses. This includes the linguistic diversity across dialects as well as different combinations and sizes of training data.

This study investigates three key research questions: (1) What is the effect of MSA training data size on ASR performance? (2) Is MSA pre-training beneficial for dialectal Arabic fine-tuning? (3) How do dialect-specific and dialect-pooled models compare in performance?

We hypothesize that MSA pre-training will enhance dialectal ASR performance through shared linguistic features [6], with improvements varying by dialect and training configuration. We expect diminishing returns with increased training data and superior performance from dialect-specific models over pooled models [1].

2. Related Work

Alsharhan and Ramsay [7] demonstrated that dialect-specific training outperformed multi-dialect approaches, finding particular divergence in Maghrebi Arabic. Using Deep Neural Network Hidden Markov Models (DNN-HMM), they showed that increased training data in general models did not necessarily improve performance over specialized ones. However, their findings for traditional hybrid models may not extend to end-to-end approaches. Recent end-to-end models demonstrated strong multilingual performance through general speech representation learning and multi-task training [4, 5]. The fine-tuning of networks such as Whisper by Radford et al. [4] and USM

by Zhang et al. [5] was compared against state-of-the-art systems by Abdelali et al. [3] in a novel Arabic benchmark. They demonstrated that their specialized state-of-the-art system outperformed multilingual zero-shot networks such as USM and Whisper in most Arabic speech datasets, with USM showing the edge in some cases. They mentioned that even though USM outperforms Whisper in all cases, fine-tuning Whisper on just 2 hours of speech data drastically improved this and came near to closing this gap. Unfortunately, Abdelali et al. [3] omitted fine-tuning Whisper on larger amounts of data or dialect-specific data. The advantage of Whisper is the open-source nature compared to the closed-source, more powerful USM, which allows Whisper to be fine-tuned and compared against domain-specific models.

Research on both dialectal diversity using DNNs and fine-tuning multilingual ASR systems on larger amounts of data in the target domain is present but limited to smaller dialects [8, 9]. The current paper investigates the effect and extent of shared feature learning from MSA in the context of dialectal Arabic ASR performance and the level of linguistic similarity between dialects according to a DNN architecture is presented.

3. Methodology

3.1. Data

Mozilla Common Voice is a widely favored option in the ASR industry for accessing both high-quality and extensive datasets [10]. The dataset was collected through volunteers writing, recording, and/or validating samples. We used the Arabic partition of Common Voice 16.1[1]. This partition contained mostly MSA samples, as the contributed sentences are expected to be, which is why we can treat it as our MSA dataset.

The audio was resampled from 48 kHz to 16 kHz to accommodate the feature extractor of the Whisper model architecture. Afterwards, the log-Mel features were extracted from the audio samples. The labels were obtained by tokenizing the transcriptions with the Whisper tokenizer and applying zero-padding such that all labels had the same fixed dimension. A split already existed in the original HuggingFace implementation of the dataset. The original "train" and "validation" partitions were merged, totaling 40 hours of speech, then reshuffled with a fixed random seed, and finally split again with a fixed random seed with a 80:20 ratio into a train- and validation-set, respectively. The "test" partition contained 13 hours of speech and was kept separate until final evaluation.

The **Massive Arabic Speech Corpus (MASC)** dataset is another large-scale Arabic ASR dataset [11]. The data was crawled from YouTube videos, consists of over 1,000 hours of speech and required segmentation from full audio files and matching with corresponding subtitle time-segments. Each processed sample combined four elements: the transcribed text, segment duration, dialect classification (defined in the metadata), and the audio segment itself. The dialects of interest were the five large dialects utilized by Alsharhan and Ramsay [7], namely Gulf, Levantine, Iraqi, Egyptian, and Maghrebi. Since the MASC audio was already at 16 kHz, no resampling was required. Finally, we extracted log-Mel features and applied the Whisper tokenizer with zero-padding similar to the Common Voice dataset.

The MASC dataset also contained a train-, validation-, and test-set. We decided to use the training set only since the size

---

was already sufficient. A large imbalance was present in the dataset, as the Egyptian (385h) and Levantine (148h) datasets were clearly overrepresented, while Iraqi (13h) and Maghrebi (17h) were more limited. Due to this imbalance, a maximum of 20 hours of speech per dialect was used. Dialects with more than 20 hours were randomly shuffled and undersampled, while dialects with less than 20 hours are used in full.

From these resampled datasets, we split the data with a 80:20 ratio into a train- and test-set with a fixed random seed. These test sets are used for final evaluation. The train partition is subsequently split further into a 80:20 partition of train and validation, respectively, again with a fixed random seed. Finally, all datasets are filtered on transcriptions that are longer than 448 tokens, because of the model configuration. This filtering only removes two samples from the Egyptian train set and one sample from the Maghrebi test set. The datasets are ported into the HuggingFace `Dataset` format, making the data-loading identical to the Common Voice data.

3.2. Model

The model of choice, as explained previously, is the state-of-the-art and multilingual ASR system Whisper by OpenAI [4]. The reason for using Whisper is its open-sourced nature compared to USM [5], and its impressive multilingual performance. The best model for this comparison was deemed to be the `whisper-small` configuration, as its size is located amidst two smaller and two larger versions. Furthermore, this configuration still has a competitive inference speed on lower-grade hardware, while containing enough parameters to exhibit model flexibility. For details, the reader is directed to Radford et al. [4]. The weights are obtained through the HuggingFace Transformers library [12], using the `openai/whisper-small` (244 million parameters) checkpoint.

For each of the five dialects and a dialect-pooled scenario (trained on all dialects), three fixed random seeds were used for the train-validation split and each of these were trained with- and without pre-training on MSA. Six scenarios and six models per scenario resulted in 36 models. Furthermore, the `whisper-small` checkpoint was fine-tuned with 20%, 40%, 60%, 80%, and 100% of the MSA data, to obtain a course of the effect of the amount of data. The data subsets are obtained by taking a random subset of the data with a fixed random seed. This results in five additional models being trained. These models were evaluated with three different fixed random seeds (42, 84, and 168) by taking a random 80% of the MSA test set. The reported WER and CER results for each model represent the average performance across these three evaluations.

All models were trained for a maximum of 6,000 steps with a batch size of 8. Early stopping was used to halt training if no decrease in evaluation loss was observed after three evaluation rounds. Thirty-three out of the total 36 models triggered early stopping, at an average of 2,682 steps. For each model, the checkpoint with the lowest validation loss was selected as the final model for evaluation. AdamW was used as an optimizer with a linear learning rate scheduler [13, 14]. The learning rate increased from 5.0×10^{-7} until the maximum learning rate of 1.0×10^{-5} in the first 500 (warm-up) steps, after which it decreased again. The hyperparameters β_1 and β_2 of AdamW were kept at 0.9 and 0.99, respectively. Since this is a fine-tuning task, the learning rate was kept as low as possible to not disturb the original weights of the Whisper model. Evaluation was performed every 250 steps, after which a model checkpoint was saved as well. The loss function used was cross-entropy.

3.3. Tools and Technologies

The fine-tuning and evaluation of the models were performed by an NVIDIA A100 GPU. The implementation was done using Python version 3.10.4, PyTorch version 1.12.1 [15], and HuggingFace Transformers version 4.39.3 [12]. The code is available on GitHub at `https://github.com/O-T-O-Z/finetune-ar-dialects`. All additional requirements and dependencies can be found there as well. The fine-tuned models and pre-processed datasets are available on HuggingFace at `https://huggingface.co/collections/otozz/arabic-dialects-67949653e522c7de7fdddc7a`. The Whisper fine-tuning blog post by Sanchit Gandhi from HuggingFace was used as inspiration [16].

3.4. Metrics

The metrics of choice are the word error rate (WER) and character error rate (CER). WER is defined as the ratio between the number of modifications or edits and the total number of words:

$$\text{WER} = \frac{S_w + I_w + D_w}{N_w}, \tag{1}$$

where S_w denotes the number of substitutions, I_w the number of insertions, D_w the number of deletions, and N_w the number of words in the reference or ground truth. A WER can exceed 100% when more words are inserted, deleted, and/or substituted than there were words in the reference, which is often an indication of "hallucination", or the model generating nonsensical and exceedingly long output. The lowest WERs for English ASR are currently below 10% [17], but many other languages are higher. The CER is defined as the ratio between the number of edits and total number of characters in the ground truth and is also evaluated to detect any inconsistencies in analysis.

4. Results and Discussion

4.1. Effect of MSA training size

The goal of the first experiment is to visualize the effect of the MSA training data size on the performance of Whisper. As observed in Figure 1, the expected downward trend in both WER and CER was recognized. This decrease was most noticeable when moving from `whisper-small` to the model fine-tuned on just 20% of the MSA training dataset. There was an improvement of almost 10% visible in WER. After that, another 80% was required to improve the model by around 6%. This validates our hypothesis and leads us to believe that even small amounts of training data can be extremely beneficial for the performance of Whisper on MSA. Furthermore, our findings are largely in line with Alsharhan and Ramsay [7], who found an increase in model performance with more data up to a certain extent. Even though they performed this similar experiment with multi-dialectal data, the conclusion is equivalent.

The most interesting outcome of this experiment was the comparable performance of the fine-tuned `whisper-small` model checkpoint with the much larger `whisper-large-v3` checkpoint (1550 million parameters) in the ASR performance of MSA. Even though the latter has access to higher model complexity, the former was able to specialize in the specific task of ASR of MSA with fewer weights and a limited amount of fine-tuning. The average inference time per sample was measured as 0.12s for `small` and 0.29s for `large-v3` (for the MSA test set). Inference time and computational cost can thus be reduced heavily through a small drop in WER performance [4].
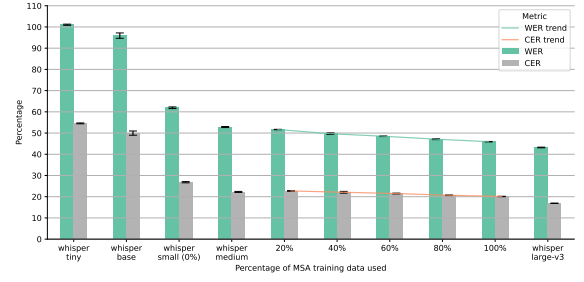

Figure 1: *Barplot shows increasing train size for `whisper-small` results in better performance that nears larger model performance.*

4.2. Impact of Pre-training

The results of the remaining 36 models are found in Figure 2. The models with- and without pre-training are compared on the test sets in Figure 2. No large differences between pre-training and no pre-training was observed, as most results were close to each other in terms of WER. The sole exception was the MSA test set, where a large difference was visible. To validate our findings, multiple Wilcoxon signed-rank tests were performed, each on the differences between test set performances. The Holm-Bonferroni correction was applied to correct for multiple comparisons and no differences were found between the test set performances. For MSA, a slight drop in performance was definitely present, since Figure 1 showed a 45.93% WER on the pre-trained model before fine-tuning, while $\mu = 58.71$ for all pre-trained models after fine-tuning. This is likely due to the model "forgetting" or re-calibrating its weights to try to accommodate both the new information and the previously learnt information.

Furthermore, another Wilcoxon signed-rank test was performed to investigate the difference between the two groups of models as a whole. Pre-trained models ($\mu = 83.87, \sigma = 13.24$) showed better performance ($Z = 120.00, p = 5.1\text{E}-4$) than non pre-trained models ($\mu = 86.77, \sigma = 11.94$). This could indicate that some general information about Arabic is preserved and generally beneficial for performance.
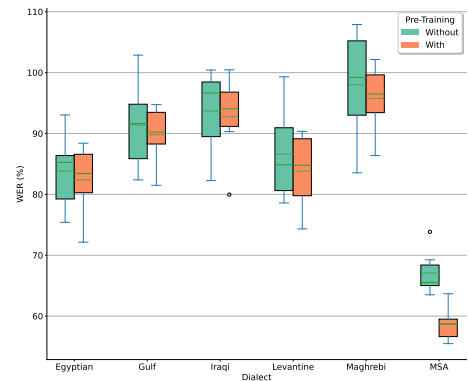

Figure 2: *Differences are small between pre-training and no pre-training, except for MSA.*

4.3. Dialect-Specific vs Pooled Models

For the difference in performance between the dialect-specific and dialect-pooled models, we focused on the performance of the pre-trained models per dialectal training set. Models trained on a specific dialectal dataset generally worked best on the MSA test set (likely due to pre-training), followed by their corresponding test set. The model fine-tuned on the dialect-pooled training set performed close to the models fine-tuned on each dialect specifically (Figure 3). This is contrary to previous literature that mention dialect-pooled performance as a challenge [1]. The average drop in WER performance from dialect-pooled to dialect-specific models is 8.42%, while CER performance reduces by 3.37%, which indicates that only a small amount of performance is sacrificed when opting for a dialect-pooled approach. Better yet, when considering the improved performance on the other dialects that could be considered out-of-distribution for the dialect-specific models, the dialect-pooled models show much potential in overall performance.
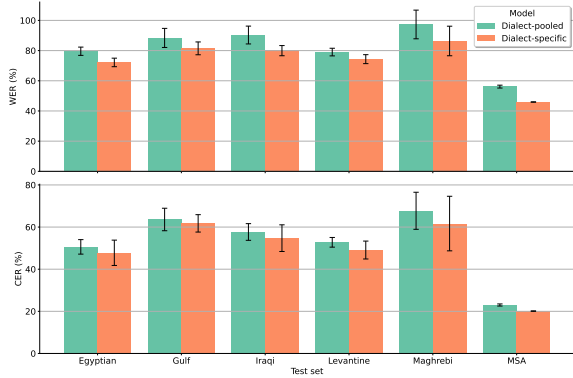

Figure 3: *Dialect-pooled and dialect-specific models perform close to each other on all dialectal test sets (both for WER and CER).*

Another advantage of using the dialectal speech datasets for training, is the ability it provided us to investigate the linguistic differences, similar to Alsharhan and Ramsay [7]. In Figure 4, we find that fine-tuning on Egyptian, Iraqi, and Maghrebi datasets decreases performance on other dialects most (rows). This is contrary to Alsharhan and Ramsay [7], who reported that Egyptian was similar to other Arabic dialects. The authors also note that Maghrebi is farthest from other dialects due to French influence on the language, which was observed in the poor performance on the Maghrebi test set when trained on a different dialect. Further, the large difference between Iraqi and Maghrebi was arguably expected due to the large geographical distance between the two regions. Conversely, Levantine and Gulf are close both geographically and in terms of cross-dialect performance. These results cannot be taken as a truth and a linguistic study would obviously be more suited for this purpose. Yet, the results for end-to-end models are mostly in line with similar comparisons for DNN-HMM models and invite further exploration in linguistic studies [7].

4.4. Limitations

A technical limitation was our reliance on Whisper's built-in tokenizer configured for Arabic transcription, without implementing additional dialect-specific text normalization. Dialectal Arabic benefits from specialized text normalization includ-
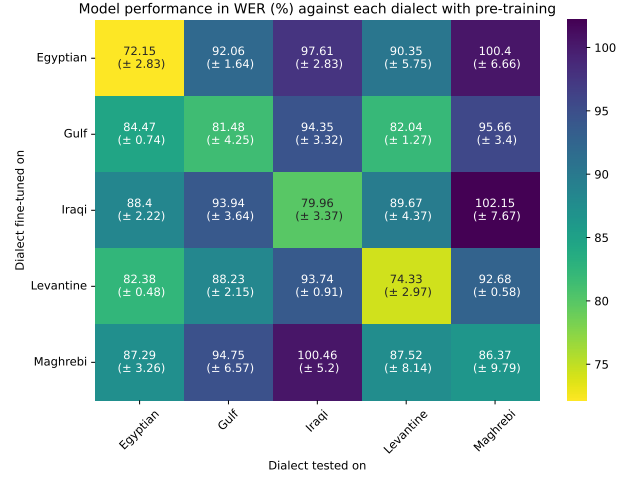

Figure 4: *Confusion matrix of WER performance between the dialectal train and test sets. Iraqi and Maghrebi show clear distance from other sets.*

ing standardization of character variants (for example, different forms of alef and hamza), diacritic handling, and normalization of dialectal spelling variations. This limitation may have affected our WER measurements across dialects, as dialectal variations in spelling and character usage could inflate error rates even when the phonetic content was correctly recognized. Future work should investigate the impact of dialect-aware text normalization on ASR performance.

5. Conclusion

Data scarcity poses a problem in dialectal Arabic ASR, as most commercial applications are focused on MSA performance. All in all, we demonstrated that fine-tuning `whisper-small` with a small amount of MSA data can achieve performance comparable to `whisper-large-v3` at a fraction of the inference time. This could have a profound impact on industry-grade applications of (Arabic) ASR, with the deployment of smaller models being preferred over large, computationally expensive, and slow models. Costs will be reduced on the side of the producer and the response time decreases on the client's side. Pre-training on MSA was also found to be beneficial for average performance in a downstream dialectal fine-tuning task, which means that training time can be reduced as well.

Further optimization showed that it seems beneficial to combine dialectal datasets for a more robust DNN due to the larger size and increased variation of its training dataset. This could be proposed as a solution to the lack of variation and data present in Arabic ASR as well as low-resource language ASR in general. By pooling similar dialects, one could reduce computational burden. When there are more similarities than in dialectal Arabic, generalization is expected to increase even further, such as with cross-continental Spanish or Portuguese.

Finally, large performance differences were present between the Arabic dialect-specific models. Dialects that are geographically further apart showed larger differences, such as the Maghrebi (North-African) and Iraqi dialects, while regions that are closer to each other showed more similar performances, for instance the Levantine (Jordan, Palestine) and Gulf (UAE, Saudi Arabia) dialects.

6. Acknowledgements

7. References

[1] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, B. Alotaibi, and Z. T. Fayed, "Deep investigation of the recent advances in dialectal arabic speech recognition," *IEEE access*, vol. 10, pp. 57 063–57 079, 2022.

[2] D. Vergyri and K. Kirchhoff, "Automatic diacritization of arabic for acoustic modeling in speech recognition," in *Proceedings of the workshop on computational approaches to Arabic script-based languages*, 2004, pp. 66–73.

[3] A. Abdelali, H. Mubarak, S. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, S. Abdaljalil, Y. El Kheir, D. Izham, F. Dalvi *et al.*, "Larabench: Benchmarking arabic ai with large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 487–520.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[5] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[6] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. sn, 2010.

[7] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975–998, 2020.

[8] S. Rijal, S. Adhikari, M. Dahal, M. Awale, and V. Ojha, "Whisper finetuning on nepali language," *arXiv preprint arXiv:2411.12587*, 2024.

[9] M. Torgbi, A. Clayman, J. J. Speight, and H. T. Madabushi, "Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the united kingdom," *arXiv preprint arXiv:2501.08502*, 2025.

[10] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[11] M. Al-Fetyani, M. Al-Barham, G. Abandah, A. Alsharkawi, and M. Dawas, "Masc: Massive arabic speech corpus," 2021. [Online]. Available: https://dx.doi.org/10.21227/e1qb-jv46

[12] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[16] S. Gandhi, "Fine-tune whisper for multilingual asr with transformers," 2022. [Online]. Available: https://huggingface.co/blog/fine-tune-whisper

[17] V. Srivastav, S. Majumdar, N. Koluguri, A. Moumen, S. Gandhi *et al.*, "Open automatic speech recognition leaderboard," https://huggingface.co/spaces/hf-audio/open_asr_leaderboard, 2023.