# Improving on the Limitations of the ASR Model in Low-Resourced Environments Using Parameter-Efficient Fine-Tuning

**Rupak Raj Ghimire, Prakash Poudyal, Bal Krishna Bal**
rughimire@gmail.com, { prakash, bal }@ku.edu.np
Information and Language Processing Research Lab (ILPRL)
Kathmandu University, Nepal

## Abstract

Modern general-purpose speech recognition systems are more robust in languages with high resources. In contrast, achieving state-of-the-art accuracy for low-resource languages is still challenging. The fine-tuning of the pre-trained model is one of the highly popular practices which utilizes the existing information while efficiently learning from a small amount of data to enhance the precision and robustness of speech recognition tasks.

This work attempts to diagnose the performance of a pre-trained model when transcribing the audio from the low-resource language. In this work, we apply an adapter-based iterative parameter-efficient fine-tuning strategy on a limited dataset aiming to improve the quality of the transcription of a previously fine-tuned model. For the experiment, we used Whisper's multilingual pre-trained speech model and Nepali as a test language. Using this approach we achieved Word Error Rate of 27.9%, which is more than 19% improvement over pre-trained $Whisper\ Large - V2$.

**Keywords** - Nepali ASR, Low-Resourced ASR, PEFT, LoRA

## 1 Introduction

Automatic Speech Recognition (ASR) is a subset of speech technology that uses machine learning techniques and neural networks to analyze and transcribe audio recordings or convert real-time speech into text. With the emergence of deep learning methods in recent years, Speech-based technology has made significant advancements. Machine learning-based ASR systems can be trained using supervised, semi-supervised, or unsupervised techniques. Supervised ASR systems acquire knowledge via precise alignment of spoken and transcribed text, requiring a substantial amount of highly curated data from manual alignment. A lot of time and work has to go into manually aligning spoken words with their written versions to make sure they match up correctly.

The initial proposal for unsupervised ASR implementation was presented by Liu et al. (2018). Following this, unsupervised methods also have gained widespread popularity. A recent study by Baevski et al. (2022) revealed that the performance of the unsupervised model is equally comparable to that of supervised models. The availability of the larger pre-trained speech models are increasing in number with access of the computing resources and advancement in deep learning techniques.

Large pre-trained speech models are typically trained on thousands of hours of diverse speech datasets. Recently released $Wav2Vec2 - BERT2.0$ (Chung et al., 2021) was pre-trained on $4.5M$ hours of unlabeled audio data covering more than 143 languages. In line with this, the $whisper - large$ model (Radford et al., 2023) is trained on $680,000$ hours of labeled audio data and has $1550M$ parameters. These models excel at capturing complex acoustic and linguistic patterns, enabling them to generalize effectively across multiple languages, accents, and noises. Having said that their size also poses challenges, such as the need for significant computational resources and the risk of over-fitting particularly in low-resource environments.

The large speech model can be fine-tuned for any speech-related downstream tasks. The full parameter fine-tuning paradigm requires multiple Graphical Processing Unit (GPU) working in parallel which is very inefficient and non-sustainable. Parameter-Efficient Fine-Tuning (PEFT) based approaches has gained popularity due to its efficiency and effectiveness in adapting pre-trained models to specific tasks, especially in resource-constrained environments. Unlike full parameter fine-tuning, the PEFT-based methods introduce the concept of a small adapter that can be trained leaving the majority of the pre-trained parameters untouched. This

approach significantly reduces the number of trainable parameters, making it computationally efficient and reducing the risk of over-fitting, particularly in low-resource settings.

Hu et al. (2022) proposed the fine-tuning technique by freezing the pre-trained model weights and injecting the trainable rank decomposition matrix to each transformer layer called Low-Rank Adaptation (LoRA). This approach is effective in the context of the Large Language Model (LLM). For example in $GPT - 3175B$ model, LoRA reduced the trainable parameters by $10,000$ times and GPU requirements by 3 times (Hu et al., 2022).

In this paper, we propose the fine-tuning architecture for the low-resourced language, Nepali. We first introduce the language to the model by full-weight fine-tuning. The limitations of the resulting module are identified and later adapter-based fine-tuning is applied to enhance the overall quality of the ASR. Additionally, we investigated the use of LoRA (Hu et al., 2022) for fine-tuning the $Whisper - Large$ model (Radford et al., 2023).

The rest of the paper is organized as follows: in section 2 the related works are explained followed by the methodology in section 3. Section 4 and 5 presents the conducted experiments and discussion of results. Finally, the paper concludes with section 6 where summary of findings, future plans, and potential extensions to the work are explained.

## 2 Related Works

Pre-trained large speech models have revolutionized the speech-related downstream task such as ASR. There are various types of the pre-trained model that we can use for the fine-tuning task. Multilingual supervised, semi-supervised, and unsupervised - all type of the model can be effectively fine-tuned. Wav2Vec2-Conformer (Wang et al., 2020), Whisper (Radford et al., 2023), MMS-1B (Pratap et al., 2024), HuBERT (Hsu et al., 2021), Wav2Vec2-BERT 2.0 (Chung et al., 2021), Wav2Vec2-Phoneme (Xu et al., 2022), Wav2Vec2.0 (Baevski et al., 2020a; Baevski et al., 2020b), $Wav2Vec$ (Schneider et al., 2019) are some of the example of pre-trained speech models trained on massive amounts of multilingual speech datasets. $Whisper$, for example, is a powerful encoder-decoder model that can transcribe speech into text in multiple languages. $Wav2Vec$ and its successors (Chung et al., 2021; Xu et al., 2022; Wang et al., 2020) use contrastive learning to learn robust speech representations.

These pre-trained models have significantly improved the accuracy and robustness of ASR systems, making them more accessible and useful in a variety of applications. Various published research (Arunkumar et al., 2022; Khare et al., 2021; Luo et al., 2021; Singh et al., 2023; Zheng et al., 2023; Ghimire et al., 2023a) shows that the accuracy of the ASR in low-resource languages including Nepali can be improved by fine-tuning the pre-trained models(Ghimire et al., 2023a). As per these researches, the fine-tuning approach requires less computing and also reduces the model training time significantly compared to full parameter training. However due to higher number of the parameters involved in the network the full parameter fine-tuning is still challenging.

The lightweight adapter tuning for speech-related task has been addressed by Le et al. (2021). As per the author this type of the adapter tuning can be used to - (1) fine-tune the large and generic model for downstream task, and (2) glue the two adapters to find solutions to new downstream tasks. This type of the adapter can be merged into the pre-trained model either in serial or parallel fashion to produce the output.

Use of the parameter-efficient fine-tuning approach LoRA and its variants are becoming very common on speech model fine-tuning. Liu et al. (2024a) used LoRA for fine-tuning the $Whisper$ model on child speech dataset. They found that LoRA-based techniques are very effective in fine-tuning the $Whisper$ model on a low-resourced child dataset. Song et al. (2024) proposed the parameter-efficient and extensible model for $Whisper$ fine-tuned with LoRA called LoRA-Whisper. The LoRA-Whisper yields a relative gain of $18.5\%$ over baseline system for multilingual ASR model. The LoRA based model adaptation mechanism naturally allows multiple LoRA modules can be formed and merged. Loading multiple LoRA and merging them with main model for inference purpose still requires higher memory capacity. To address this situation, Sheng et al. (2024) has proposed a method in which we can serve thousands of concurrent LoRA adapters called S-LoRA. Liu et al. (2024b) demonstrate that fine-tuning using LoRA is much more effective on the pre-trained models for low-resource ASR.

The Nepali ASR is still in its early stages of research and development. However, there are some

promising results as reported in previous works ( Ghimire et al., 2023a; Shrestha et al., 2021; Regmi and Bal, 2021; Ghimire et al., 2023b; Ghimire and Bal, 2017). Among them, the work reported by Ghimire et al. (2023a) is the only work related to fine-tuning for building Nepali ASR system. The author proposed semi-supervised fine tuning of the pre-trained model using an active learning approach. This research uses the SLR54 (Kjartansson et al., 2018) dataset for the full parameter fine-tuning resulting $6.77\%$ CER on the Massively Multilingual Speech (MMS)-1B ($mms1b$) (Pratap et al., 2024) model.

The use of LoRA and similar parameter-efficient fine-tuning approaches for speech-related downstream tasks is increasing, even in the case of low-resourced languages. These approaches are designed to solve downstream tasks. Unfortunately, it should be noted that the use of multiple LoRA adapters to enhance the output is the least explored. We can train and adopt multiple adapters, which reduces the limitations of each other. This is the primary motivation for this work and we have experimented with this approach in low-resource Nepali ASR.

## 3 Methodology

### 3.1 Nepali Speech Corpus

The large Nepali ASR training dataset (Kjartansson et al., 2018) is available in Open Speech Language Resources[1]. This is the only speech corpus publicly available that is suitable for ASR task. This dataset has $157K$ utterances. There are other Nepali speech corpus (Sodimana et al., 2018[2]; Khadka et al., 2023[3]) also available, but these are single speaker datasets and are only suitable for Text-to-Speech (TTS) task.

The whole dataset (say $D$) is further divided into the following sub datasets:

- $D_{train}$ : Dataset used for the full parameter tuning
- $D_{val}$ : Dataset used for the validation of the models
- $D_{stock}$ : They are stock dataset which will be used for the LoRA based fine-tuning

[1][SLR54] - https://www.openslr.org/54/
[2][SLR43] - https://www.openslr.org/43/
[3][SLR143] - https://www.openslr.org/143/

The overall methodology can be further divided into the three stages: 1) full parameter fine-tuning, 2) identifying the limitation of the fine-tuned model, building an adapter to address those limitations, and 3) merging the selected adapters and building the final fine-tuned model. These steps are outlined in Figure 1 and are explained in more detail in Subsections 3.2, 3.3, 3.4 and 3.5.

### 3.2 Fine-tuning of Pre-trained Model

A crucial initial step is to (re)introduce language into the large speech model by full-parameter fine-tuning.

Suppose $W$ represents the weight matrix of the pre-trained model, the goal of the fine-tuning is to identify the weight changes $\Delta W$. The weight changes $\Delta W$ is computed as a negative gradient of the loos times learning rate i.e. $\Delta W = \alpha \times (-\nabla L_W)$. Now the updated original weight ($W'$) is $W' = W + \Delta W$.

The dataset $D_{train}$ is used, which is a subset of the entire data set $D$. The computational resources and time required for training are dependent upon the number of parameters of the pre-trained model and the size of the speech dataset.

### 3.3 Identification of the transcription errors

After full parameter or LoRA fine-tuning we have to analyze the output of the model to identify the patterns of the recognition errors. Recognition errors can be quantified using Character Error Rate. The whole process of determining the error or limitation is explained in Algorithm 1. The validation dataset ($D_{val}$) are used for error identification purposes. The CER ($CER_{d_i} \forall D_{val}$) is calculated. The $CER_{d_i}$ is then compared with the threshold CER value $CER_{TH}$. Those data whose CER is greater than $CER_{TH}$, are marked as not acceptable transcripts, deciding these are areas for improvement. To address these problems, we have to train the model on more datasets containing the problematic tokens. This dataset can be generated from $D_{stock}$. Newly generated subset of dataset are called $\Delta DS$.

Getting the appropriate value of the threshold CER ($CER_{TH}$) is very important. For this experiment, we chose the CER of the previously merged adapter while evaluating using the $D_{val}$ dataset. The identification of shortcomings and the dataset generation technique is explained in Algorithm 1.

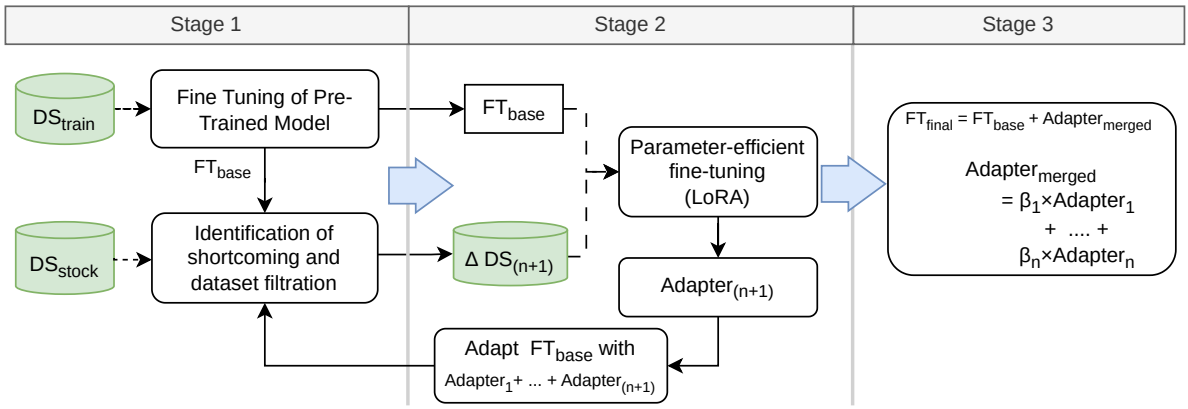


Figure 1: Proposed Architecture for Multi-Stage LoRA Fine-Tuning of Base model $FT_{base}$ (**Stage 1**: (re)introduce the language, **Stage 2**: multiple adapter building by fine tuning the base model guided by Algorithm 1, and **Stage 3**: merging the adapters)

**Algorithm 1:** Identification of shortcomings of $(n)^{th}$ Adapter and dataset generation for $(n+1)^{th}$

**Input:**
$Adapter_n$ : LoRA adapter of $n^{th} iteration$
**Output:**
$\Delta DS_{(n+1)}$ : Dataset to be used for $(n+1)^{th}$ Adapter
**Data:**
$FT_{base}$ : fine-tuned base model
$DS_{train}$ : training dataset for full-parameter fine tuning
$DS_{val}$ : validation dataset
$DS_{stock}$ : stock dataset
$CER_{TH}$ : threshold value of CER
$A$ : list of adapters

$Label_{weak}$ = [ ]
$FT_{modal}$ = $FT_{base} \cup Adapter_n \cup \{A_i\}_{i \epsilon N}$
**forall** $d_i \in DS_{val}$ **do**
    $CER_i$ = $FT_{modal}.Evaluate(d_i.Audio)$
    **if** $CER_i > CER_{TH}$ **then**
        $Label_{weak}.Append(d_i.Label)$
**forall** $d_i \in DS_{stock}$ **do**
    **if** $d_i.Label \in Label_{weak}$ **then**
        $\Delta DS_{n+1}.Append(d_i)$
**return** $\Delta DS_{(n+1)}$

### 3.4 Parameter-Efficient Fine-Tuning

Pre-trained models have a lower intrinsic dimension when they are adjusted to new challenges (Hu et al., 2022). A low intrinsic dimension suggests that the data can be represented or approximated by a lower-dimensional space while most of its structure is preserved. This allows us to split the new weight matrix for the adapted task into smaller matrices without compromising important details. Suppose $\Delta W$ is the weight update for an $A \times B$ weight matrix. Then we can decompose the weight update matrix into two smaller matrices: $\Delta W = W_A W_B$, where $W_A$ is an $A \times r$-dimensional matrix and $W_B$ is an $r \times B$-dimensional matrix. We keep the original weight $W$ frozen and only train the new matrices $W_A$ and $W_B$. This will be considered as a new fine-tuned adapter.

The LoRA is used as the parameter-efficient fine-tuning. We started by loading the fine-tuned base ($FT_{base}$) model which is the output of Section 3.2. In this stage only the parameters of the LoRA adapters are updated, while the remaining weights ($W$) of $FT_{base}$ were kept frozen.

### 3.5 Merging the LoRA Adapters

Building the model by correcting for each weakness in a single training can be costly, time consuming, and takes up storage space. Multiadapter training can overcome some of these limitations by training a model to solve multiple weaknesses. Loading multiple adapters and doing inference on those require more memory. This can be solved by merging multiple adapters together to form the single adapter. There are various techniques to merge

Table 1: Evaluation of models in terms of Word Error Rate

| **Model** | **Details** | **WER**(%) |
|---|---|---|
| Whisper large-v2 | Model proposed by (Radford et al., 2023) | 47.1 |
| $FT_{base}$ | Full parameter fined tuned base model | 36.2 |
| $Adapter_1$ | LoRA fine-tuned from the base model $FT_{base}$ | 31.1 |
| $Adapter_2$ | LoRA fine-tuned from the base model $FT_{base}$ | 33.1 |
| $Adapter_3$ | LoRA fine-tuned from the base model $FT_{base}$ | 30.1 |
| $Adapter_{merged}$ | All adapter merged into $FT_{base}$ | **27.9** |

the adapters together; among them, the weighted average method is computationally friendly and the easiest technique. We merged multiple adapters with adapter weight $\beta_i$ as in Equation(1).

$$Adapter_{merged} = \sum_{i}^{n} \beta_i \times Adapter_i \quad (1)$$

Where, $\beta_i$ is the weight of $i^{th}$ adapter and holds $\sum_i^n \beta_i = 1$. The values for $\beta_i$ are generated from the hyperparameter search. The actual values used are reported in Section 4.

## 4 Experiments

We used the SLR54 dataset (Kjartansson et al., 2018) which has a Nepali speech corpus of $157K$ utterances and has 165 hours of speech recording featuring 527 unique speakers. Among these 4 hours of speech dataset that are used as $DS_{train}$, 20 minutes are used as the validation dataset $DS_{val}$ and the remaining are used for the pool or stock dataset that can be used for further adapter training ($DS_{stock}$).

The $Whisper$ (Radford et al., 2023) model family is used as the pre-trained speech model. They have $tiny$, $base$, $small$, $medium$, and $large$ models ranging from 39M parameters to 1550M parameters. We used a pre-trained $Whisper - Large - V2$ model. This pre-trained model also receives Nepali language training. So, there is no need to explicitly introduce the vocabulary.

For this experiment, we used the Hugging Face Transformer library[4]. To make our experiment comparable, we trained each adapter with a total of five (5) epochs with the following parameters:

- mixed precision training using float16 (*fp16*) data type
- 8-bit Adam optimizer(*adamw_bnb_8bit*)
- learning rate of $1e-3$ and
- training batch size to $4$

Parameter-efficient fine-tuning is implemented using the PEFT library (Mangrulkar et al., 2022). The best combination of LoRA configurations is estimated by using a hyperparameter tuning technique. The best estimated parameters are as follows:

- $r = 32$
- $alpha = 64$
- $dropout = 10\%$

To check the effectiveness of the module, a total of 3 adapters are fine-tuned, namely $Adapter_1$, $Adapter_2$, and $Adapter_3$. All of these adapters are linearly merged to form $Adapter_{merged}$ with the weights $\beta$ as in Equation(1). The value of $\beta$ obtained from the hyperparamerter tuning is as follows:

$$\beta = \{0.7, 0.2, 0.1\}$$

Fine-tuning may be performed by training any number of adapters until the model starts to overfit. We observed overfitting after three iterations of fine-tuning because of the limited size of the dataset employed in our trials. This discovery highlights the necessity of carefully overseeing model performance during fine-tuning to avoid overfitting

[4]Hugging Face: https://huggingface.co/docs/transformer

and guarantee effective generalization to unseen data.

The newly produced adapter, termed $Adapter_{merged}$ is later integrated with the fine-tuned basis model ($FT_{base}$). Upon completion of the integration, an assessment is performed to evaluate the performance of the merged model.

## 5 Results and Discussions

We used $Whisper - Large - V2$ pre-trained model. This is a multilingual large speech model also trained in Nepali speech as well. The reported Word Error Rate (WER) of this model is $47.1\%$. While inspecting the output, we saw that the transcript of the module is more toward Hindi language. Hence, we performed full-parameter fine-tuning using our dataset and ended up with a WER of $36.2\%$. This language (re)introduction by full-parameter fine-tuning improved the performance of the base model. Now, our base model is $FT_{base}$ with **36.2%** WER. All experiments are listed in Table 1.

Three different Low-Rank Adaptation (LoRA) based adapters are trained as per Section 3.3 and Algorithm 1. The individual WERs of the adapter range from 33.1% to 30.1%. The best WER is achieved by combining the three adapters together, which is **27.9%**[5]. This is more than **8%** improvement over full-parameter fine-tuned base model ($FT_{base}$) and more than **20%** improvement over the pre-trained model $Whisper - Large - V2$.

Our work can also be compared with existing fully supervised End-to-End Nepali ASR models. The Hidden Markov Mode (HMM) based model proposed by Baral and Shrestha (2020) has achieved 29.45% WER. The CNN-GRU-based model proposed by Joshi et al. (2023) has reported a WER of 37. 50%. In all of these cases, our model performs better. The fine-tuned model based on active learning proposed by Ghimire et al. (2023a) has reported the 6.77% CER. Because of the different evaluation matrices, we could not compare the result.

The results of our investigation indicate that combining numerous LoRA adapters, each trained on different subsets of data, allows them to complement each other's strengths and minimize shortcomings. An adaptor may capture only specific patterns or features, resulting in a higher WER. However, when integrated, these adapters can offer a more thorough adaptation by addressing the shortcomings of the others. This collective effect improves the model's capacity to generalize and reliably recognize speech, resulting in a lower overall WER.

[5]The final models are available through Information and Language Processing Research Lab's website, https://ilprl.ku.edu.np

## 6 Conclusion

Fine-tuning of Large Language Models for low-resource languages is a prevalent and growing practice, particularly in the context of speech-related tasks. Due to Nepali being a low-resourced language, the fine-tuning task has been comparatively less investigated. Our study focused on evaluating the usefulness of adapter-based fine-tuning through experiments conducted with LoRA and Whisper. Our strategy involves initially (re)introducing the language into the bigger model through full parameter tuning. Subsequently, the shortcomings of the model are recognized, and by addressing those drawbacks, we refine certain aspects of the model using LoRA, thereby improving the overall quality of the model. Using this methodology, we achieved a Word Error Rate (WER) of $27.9\%$, which is an improvement over $19.2\%$ as reported in previous work.

Although we observed a significant improvement in implementing the suggested strategy, there is still plenty of space to improve the precision of the model. The algorithm for generating $\Delta DS$ uses the CER compared to some threshold. However, to achieve further improvements, this area can be analyzed using language-specific similarity metrics. Additionally, at this stage, our study focused only on Whisper. The study could be extended to other larger models to compare the corresponding results.

## References


A Arunkumar, Vrunda Nileshkumar Sukhadia, and Srinivasan Umesh. 2022. Investigation of Ensemble Features of Self-Supervised Pretrained Models for Automatic Speech Recognition. In *INTERSPEECH 2022*, pages 5145–5149. ISCA.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2022. Unsupervised speech recognition. In *Advances in Neural Information Processing Systems 34*, pages 27826–27839.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. Vq-Wav2vec: Self-Supervised Learning of Discrete Speech Representations. In *ICLR 2020*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in neural information processing systems 33*, pages 12449–12460.

Elina Baral and Sagar Shrestha. 2020. Large Vocabulary Continuous Speech Recognition for Nepali Language. *International Journal of Signal Processing Systems*, 8(4):68–73.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

Rupak Raj Ghimire and Bal Krishna Bal. 2017. Enhancing the Quality of Nepali Text-to-Speech Systems. In Alla Kravets, Maxim Shcherbakov, Marina Kultsova, and Peter Groumpos, editors, *Creativity in Intelligent Technologies and Data Science*, volume 754, pages 187–197. Springer International Publishing.

Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. 2023a. Active learning approach for fine-tuning pre-trained ASR model for a low-resourced language: A case study of nepali. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 82–89. NLP Association of India (NLPAI).

Rupak Raj Ghimire, Bal Krishna Bal, Balaram Prasain, and Prakash Poudyal. 2023b. Pronunciation-aware syllable tokenizer for nepali automatic speech recognition system. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 36–43. NLP Association of India (NLPAI).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 3451–3460.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Basanta Joshi, Bharat Bhatta, and Ram Krishna Maharjan. 2023. End to End based Nepali Speech Recognition System. 17(102–109).

Supriya Khadka, Ranju G.C., Prabin Paudel, Rahul Shah, and Basanta Joshi. 2023. Nepali text-to-speech synthesis using tacotron2 for melspectrogram generation. In *SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages*.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *INTERSPEECH 2021*, pages 1529–1533. ISCA.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 52–55. ISCA.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824. Association for Computational Linguistics.

Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, and Lin-shan Lee. 2018. Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings. In *INTERSPEECH 2018*, pages 3748–3752. ISCA.

Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2024a. Sparsely shared lora on whisper for child speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11751–11755.

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

Jian Luo, Jianzong Wang, Ning Cheng, and Jing Xiao. 2021. Loss Prediction: End-to-End Active Learning Approach For Speech Recognition. In *2021 International Joint Conference on Neural Networks (IJCNN). IEEE*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Sunil Regmi and Bal Krishna Bal. 2021. An end-to-end speech recognition for the Nepali language. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 180–185, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised Pre-training for Speech Recognition. In *INTERSPEECH 2019*. ISCA.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. 2024. S-lora: Serving thousands of concurrent lora adapters.

Rupesh Shrestha, Basanta Joshi, and Suman Sharma. 2021. Nepali Speech Recognition using LSTM-CTC. In *Proceedings of 10th IOE Graduate Conference*.

Satwinder Singh, Feng Hou, and Ruili Wang. 2023. A Novel Self-training Approach for Low-resource Speech Recognition. In *NTERSPEECH 2023*, pages 1588–1592. ISCA.

Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. 2018. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-whisper: Parameter-efficient and extensible multilingual ASR. In *INTERSPEECH 2024*. ISCA.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *INTERSPEECH 2022*. ISCA.

Zhisheng Zheng, Ziyang Ma, Yu Wang, and Xie Chen. 2023. Unsupervised Active Learning: Optimizing Labeling Cost-Effectiveness for Automatic Speech Recognition. In *INTERSPEECH 2023*. ISCA.