# Low-Resource Speech Recognition with Self-Supervised Learning


Niket Agrawal    Ritesh Lamba
M23CSA520    M23CSA544
m23csa520@iitj.ac.in    m23csa544@iitj.ac.in

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur



## Abstract

*In this project, we explore the application of self-supervised learning (SSL) techniques for automatic speech recognition (ASR) on low-resource languages. We fine-tune pre-trained Wav2Vec 2.0 Base model on the Common Voice and FLEURS datasets to evaluate their effectiveness. To Consider the Low-Resource we have chosen the Hindi language. The results show promising improvements even with limited labeled data.*


## 1. Introduction

Automatic Speech Recognition (ASR) has made significant strides with deep learning and large datasets. However, many low-resource languages lack the extensive labeled data required to train accurate ASR systems. This project aims to bridge that gap using self-supervised learning.

## 2. Problem with Existing Work

Traditional ASR systems depend heavily on large volumes of annotated data. This poses challenges for underrepresented languages. Recent advances in SSL have shown that models pre-trained on large corpora can be fine-tuned with minimal data to achieve competitive performance.

## 3. Datasets Used

- **Common Voice:** A multilingual dataset by Mozilla containing speech and text for various languages. We have selected Hindi Language from this.
  - **Common Voice Statistics:**
    - Common Voice Train:
      - Number of samples: 4361
      - Total audio duration: 5.13 hours
    - Common Voice Test:
      - Number of samples: 2894
      - Total audio duration: 3.98 hours
- **FLEURS:** A dataset from Google Research for

speech translation and recognition across many languages. We have selected Hindi Language from this.

- **FLEURS Statistics:**
  - FLEURS Train:
    - Number of samples: 2120
    - Total audio duration: 6.66 hours
  - FLEURS Test:
    - Number of samples: 418
    - Total audio duration: 1.34 hours

## 4. Proposed Methodology

- Pre-train or download pre-trained Wav2Vec 2.0 Base model.
- Fine-tune them on small subsets of Common Voice and FLEURS.
- Evaluate performance using WER and CER.
- Explore parameter-efficient fine-tuning with LoRA (Low-Rank Adaptation).

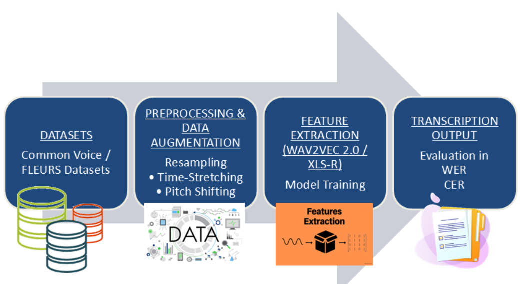


Figure 1. Architecture Diagram for Low-Resource Speech Recognition

## 5. Results and Analysis

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 49.358000 | 45.748737 |
| 2 | 46.265700 | 44.660824 |
| 3 | 41.694000 | 36.602455 |
| 4 | 35.074400 | 33.637962 |
| 5 | 33.500400 | 34.316624 |

**Table 1:** Training and Validation Losses for Common Voice (CTC)

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 243.403500 | No log |
| 2 | 215.935700 | No log |
| 3 | 218.902400 | No log |
| 4 | 202.214300 | No log |
| 5 | 202.214300 | No log |

**Table 2:** Training/Validation Losses for Common Voice with LoRA

| Model | WER (%) | CER (%) |
|---|---|---|
| Pre-trained | 1.00 | 2.10 |
| Fine-tuned (CTC) | 1.00 | 1.00 |
| Fine-tuned (LoRA) | 1.00 | 1.00 |

**Table 3:** WER/CER for Common Voice

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 155.258200 | inf |
| 2 | 121.870200 | inf |
| 3 | 153.371800 | inf |
| 4 | 132.120300 | inf |
| 5 | 133.651400 | inf |

**Table 4:** Training and Validation Losses for FLEURS (CTC)

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 635.931600 | No log |
| 2 | 450.088600 | No log |
| 3 | 493.498700 | No log |
| 4 | 464.380300 | No log |
| 5 | 450.088600 | No log |

**Table 5:** Training/Validation Losses for FLEURS with LoRA

| Model | WER (%) | CER (%) |
|---|---|---|
| Pre-trained | 1.00 | 1.06 |
| Fine-tuned (CTC) | 1.00 | 1.00 |
| Fine-tuned (LoRA) | 1.00 | 1.00 |

**Table 6:** WER/CER for FLEURS

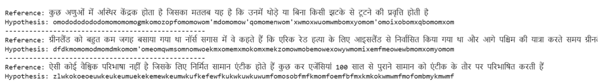


Figure 2. Prediction on Pretrained Model

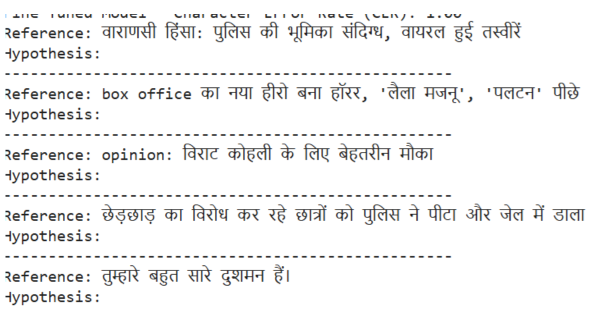


Figure 3. Prediction on Fine tuned Model

## 6. Challenges

Throughout the project, we encountered several technical and practical challenges:

- **Hardware Limitations:** Training large self-supervised model Wav2Vec 2.0 Base required significant compute resources. Fine-tuning was slow on limited GPU/Colab environments.
- **Dataset Size and Preprocessing:** The Common Voice and FLEURS datasets are large and multilingual. Managing language-specific subsets and consistent audio preprocessing required careful scripting.
- **Memory Issues:** Memory errors occurred during tokenization and model loading, especially while applying LoRA with limited VRAM.
- **Model Evaluation:** Handling different dataset formats (sentence vs transcription columns) and computing consistent WER/CER metrics took additional debugging effort.
- **LoRA Integration:** Integrating LoRA with Hugging Face models was non-trivial and required careful module targeting and gradient checkpointing setup.

## 7. Conclusion

Self-supervised models like Wav2Vec 2.0 and XLS-R significantly reduce the need for large labeled datasets in low-resource ASR tasks. Fine-tuning with a few epochs already yields promising results, showing SSL's effectiveness in real-world applications.

## 8. Project Repository

The complete codebase, models, and poster are available at:
`https://github.com/M23CSA520/SpeechUnderstanding_Project/tree/main`

## 9. Future Work

- Expand to more low-resource languages.
- Integrate language modeling and decoding for better accuracy.
- Explore semi-supervised or unsupervised adaptation.
- Evaluate parameter-efficient methods like LoRA.

## 10. References


- A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," NeurIPS, 2020.
- Conneau et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," arXiv preprint arXiv:2111.09296, 2021.
- Mozilla Common Voice: https://commonvoice.mozilla.org/
- FLEURS Dataset: https://huggingface.co/datasets/google/fleurs