Boosting wav2vec2 Performance in Low Resource Speech Recognition via Enhanced Pretraining and Fine-Tuning

Fonseka W.A.R.T
Department of Computer Science Engineering
Faculty of Engineering, University of Moratuwa
Colombo, Sri Lanka
thathsarana.21@cse.mrt.ac.lk

Uthayasankar Thayasivam
Department of Computer Science Engineering
Faculty of Engineering, University of Moratuwa
Colombo, Sri Lanka
ruthaya@cse.mrt.ac.lk

***Abstract* — Pretrained speech representation models such as Wav2Vec2 have achieved remarkable success in speech recognition. However, many downstream tasks such as low-resource language recognition, child voice recognition, and animal sound detection differ significantly from the pretraining objectives. In such cases, relying solely on fine-tuning pretrained models is often insufficient, while training from scratch is computationally expensive and data intensive. These challenges make it particularly difficult for researchers working with low-resource settings.**

In this work, we explore approaches to improve the efficiency of wav2vec2 for both pretraining and fine-tuning. We experimented with an inter-codebook similarity loss to enhance pretraining efficiency. We also investigated the use of residual quantization during fine-tuning to improve model adaptability with limited data. While our results are preliminary, this study provides insights into the challenges of adapting wav2vec2 to low-resource conditions and highlights potential directions for future research on efficient pretraining and fine-tuning strategies.

***Index Terms* — Wav2Vec2, Low-Resource Speech Recognition, Residual Quantization**

I. INTRODUCTION

Speech recognition has seen remarkable advances in recent years, driven largely by self-supervised learning techniques that leverage large amounts of unlabeled audio data. Among these methods, wav2vec2 [1] has emerged as a state-of-the-art framework, learning rich speech representations during pretraining, and enabling effective fine-tuning on various downstream tasks. These pretrained models have been widely adopted for standard speech recognition tasks in high-resource languages.

However, many practical applications such as low-resource language speech recognition, child voice recognition, and animal sound detection differ significantly from the domains used during pretraining. In these scenarios, relying solely on fine-tuning pretrained models may not fully capture the unique characteristics of the target data [2]. Moreover, training models from scratch is often required to achieve optimal performance, but this approach demands large-scale datasets, substantial computational resources, and extended training times, which can be prohibitive for researchers with limited resources [3].

Addressing these challenges is crucial for democratizing speech recognition technology and extending its benefits to low-resource and specialized domains. In this work, we focus on improving the efficiency of wav2vec2 for low-resource settings, both during pretraining and fine-tuning. We propose methods to enhance the learning of speech representations with limited data and improve model adaptability during downstream tasks. Our study aims to provide practical strategies for efficient pretraining and fine-tuning of large speech models, reducing computational and data requirements while maintaining competitive performance across diverse speech recognition tasks.

II. RELATED WORK

Self-supervised learning (SSL) has become the foundation for modern speech representation models,

allowing the use of vast amounts of unlabeled audio for pretraining. Wav2Vec2.0 [1] is one of the most influential models in this space, extending the earlier Wav2Vec [4] and CPC (Contrastive Predictive Coding) [5] approaches. It consists of a feature encoder, a Transformer-based context network, and a quantization module, enabling robust context-aware speech representations. By masking timesteps and training with contrastive objectives against discrete codebook targets, Wav2Vec2.0 learns representations that generalize across diverse acoustic conditions.

Building on its success, several variants and extensions have been proposed. HuBERT [6] introduced an iterative pseudo-labeling strategy where clustering-based labels guide pretraining, progressively refining both acoustic and linguistic structure capture. Data2Vec [7] generalized the SSL framework to multiple modalities, including speech, text, and vision, while maintaining a unified architecture. Similarly, WavLM [8] specialized in speech by integrating masked prediction with additional training strategies to enhance speaker and acoustic information encoding. These approaches highlight the rapid evolution of SSL models for speech, but they also expose challenges when adapting to low-resource or specialized domains.

Several studies have explored low-resource speech recognition. Techniques such as transfer learning and data augmentation have been widely applied to address the scarcity of labeled data in low-resource languages [9]. These approaches often leverage pretrained models trained on high-resource languages, but performance can degrade when the target domain significantly differs from the pretraining domain, such as in child speech recognition [10] or animal vocalizations.

Other works have focused on improving pretraining efficiency and reducing the computational burden of large speech models. For instance, Luis Lugo et al. [3] investigated strategies for efficient self-supervised learning of speech representations, highlighting the challenges of applying large models to resource-constrained environments. Techniques such as quantization and knowledge distillation have also been employed to reduce model size and accelerate fine-tuning while maintaining performance [11], [12].

Despite these advances, efficiently adapting large pretrained models like wav2vec2 to specialized or low-resource speech tasks remains challenging. Our work builds upon these approaches by exploring strategies that improve pretraining and fine-tuning efficiency, aiming to reduce computational requirements and enable practical deployment in low-resource settings.

III. METHODOLOGY

This section outlines the dataset used in our study, the experimental setup, and the two key approaches we employed to enhance the efficiency of Wav2Vec 2.0: improvements during the pretraining stage and the fine-tuning stage. We also describe the model configurations adopted in our experiments, along with the evaluation metrics used to assess performance. In addition, we highlight specific design choices made to adapt the framework for low-resource scenarios, ensuring a fair and meaningful evaluation.

A. Dataset

For our experiments, we used the LibriSpeech dataset [13], which was also used in the original Wav2Vec 2.0 research. To mimic a low-resource setting, we selected a 10-hour subset from the full 960-hour corpus. We specifically used the cleaned subset to ensure higher data quality.

When constructing the 10-hour dataset, we only included utterances longer than 5 seconds, since very short speech segments often caused issues during the masking step in pretraining. After filtering, our training set contained 2,850 utterances.

For evaluation, we used the original LibriSpeech validation set. Applying the same filtering rule (removing utterances shorter than 5 seconds), the final validation set consisted of 2,703 utterances.

B. Proposed Approaches

To improve the efficiency of Wav2Vec 2.0 in low-resource scenarios, we explored two key strategies focusing on both pretraining and fine-tuning stages.

1. Pretraining Enhancement:

In self-supervised learning models that rely on codebook-based vector quantization, there is a potential issue where all vectors within a single codebook may collapse into a single representation. The original Wav2Vec 2.0 addresses this by introducing a diversity loss term, which encourages uniform usage of codebook entries and helps prevent intra-codebook collapse.

However, when multiple codebooks are used, another issue can arise called inter-codebook similarity, where

distinct codebooks learn overlapping or redundant representations. This reduces the diversity and effectiveness of quantized vectors, limiting the model's representational capacity. The original diversity loss does not explicitly address this form of redundancy, which may lead to inefficient feature learning.

To mitigate this issue, we propose an additional loss term called the Inter-Codebook Similarity Loss (ICSL). This term measures the cosine similarity between embeddings across multiple codebooks and penalizes high similarity, thereby encouraging each codebook to capture distinct aspects of the input signal. By promoting greater inter-codebook diversity, ICSL aims to stabilize training and enhance the quality of learned representations.

Previous research on Wav2Vec 2.0 noted that using a large number of codebooks can degrade model performance, possibly due to such inter-codebook collapse. Through this work, we empirically explore this phenomenon and investigate whether incorporating ICSL can counteract this effect and improve training efficiency.

This approach allows us to examine the impact of inter-codebook regularization under different codebook configurations. The empirical results and comparative analysis of these experiments are presented in Section IV.

2. Fine-Tuning Enhancement

In the original Wav2Vec 2.0 framework, fine-tuning is typically performed by attaching a task-specific head, such as a linear projection layer, on top of the pretrained model and optimizing it using a supervised loss such as Connectionist Temporal Classification (CTC) Loss. While this approach performs well for large and diverse datasets, it may be less effective in low-resource scenarios, where data scarcity can limit the model's ability to adapt effectively to the downstream task.

To address this limitation, we introduce a Residual Quantization Vector (RVQ) representation between the Wav2Vec 2.0 embedding layer and the task-specific head during fine-tuning. The RVQ module refines the latent speech representations by applying a series of residual quantizers, where each subsequent quantizer encodes the residual error left by the previous one. This hierarchical quantization helps capture finer-grained variations in the latent space and reduces redundancy, allowing the model to represent speech features more compactly and adapt more efficiently with limited data.

By integrating RVQ into the fine-tuning process, we aim to improve the convergence rate and overall adaptability of the model under data-constrained conditions. In this study, we used the Hugging Face pretrained Wav2Vec 2.0 model trained on the 960-hour LibriSpeech corpus as the base model. Fine-tuning was conducted using our filtered 10-hour LibriSpeech subset, employing the CTC loss for optimization.

Although we did not perform full-scale fine-tuning to final convergence, preliminary experiments demonstrated that the RVQ-enhanced fine-tuning achieved faster convergence compared to the standard fine-tuning approach. These findings suggest that RVQ may provide a more efficient adaptation pathway for low-resource speech recognition tasks. Further experiments and evaluations are planned to investigate the stability and final accuracy of this method in future work.

C. Experimental Setup

All experiments were conducted using a single NVIDIA P100 GPU with 16 GB memory provided by Kaggle.

	Pretraining Configuration	**Fine-Tuning Configuration**
Model	base (randomly initialized)	base (pretrained weights)
Processor	facebook/wav2vec2-base	facebook/wav2vec2-base
Batch Size	16	8
Gradient Accumulation Steps	4	8
Learning Rate	1×10^{-4}	1×10^{-4}
Scheduler	Linear	Linear
Total Steps	2200	1500
LR Warm-up Ratio	0.1	0.2
LR Hold Ratio	-	0.2
LR Decay Ratio	0.9	0.6

Table 1: Experimental configurations for pretraining and fine-tuning stages

For pretraining, the Inter-Codebook Similarity Loss (ICSL) was added to the total loss function with a weight of 0.1 to encourage diversity across multiple codebooks.

D. Evaluation Metrics

We used different metrics for pretraining and fine-tuning to reflect their distinct objectives.

Pretraining: The model was evaluated using Contrastive Loss, which measures how well it distinguishes between similar and dissimilar audio representations. Lower contrastive loss indicates better feature learning and alignment between latent and quantized representations.

Fine-tuning: We used Connectionist Temporal Classification (CTC) Loss and Word Error Rate (WER). CTC Loss measures how accurately the model predicts sequences without explicit alignment, while WER evaluates end-to-end transcription accuracy by calculating the proportion of insertions, deletions, and substitutions. Lower values indicate better downstream performance.

IV. EXPERIMENTS

In this section, we present the experimental evaluation of our proposed enhancements for Wav2Vec 2.0 in low-resource scenarios. We assess both the pretraining and fine-tuning stages using the 10-hour subset of the LibriSpeech dataset described in Section III.A.

A. Pretraining Experiments

We evaluated the impact of Inter-Codebook Similarity Loss (ICSL) on Wav2Vec 2.0 pretraining using different codebook configurations. Specifically, we compared three setups:

1. 2 codebooks without ICSL (baseline)
2. 2 codebooks with ICSL
3. 8 codebooks with ICSL

The contrastive loss was recorded for both training and validation sets. Figures 1 and 2 show the contrastive loss curves for training and validation, respectively.

Results indicate that using ICSL with 8 codebooks leads to faster convergence compared to both baseline setups. In contrast, the two-codebook configurations, with or without ICSL, produced similar convergence behavior. These findings suggest that increasing the number of codebooks can improve representation learning if inter-codebook redundancy is addressed. In prior work, additional codebooks did not yield better performance, likely due to codebook collapse. By incorporating ICSL, we mitigate this issue, allowing multiple codebooks to contribute effectively.

Although we did not perform full training to final convergence, preliminary results imply that the 8-codebook configuration with ICSL may achieve better final performance than configurations with fewer codebooks. These observations highlight that in low-resource scenarios, using more codebooks with ICSL can provide faster and potentially more effective pretraining, reducing both data and computational requirements. Complete experiments and extended results will be presented in future work.

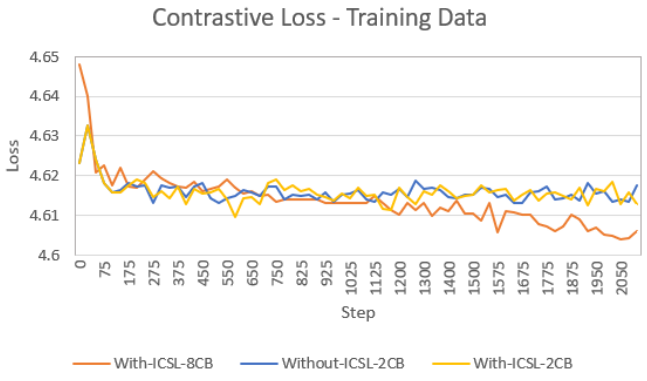

Figure 1: Training contrastive loss with and without ICSL using 2 and 8 codebooks. Models with ICSL and 8 codebooks converge faster.

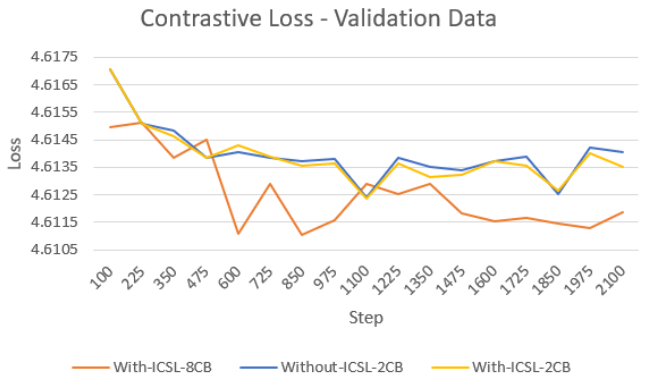

Figure 2: Validation contrastive loss with and without ICSL using 2 and 8 codebooks, showing faster and more stable convergence with ICSL.

B. Fine-Tuning Experiments

To evaluate the effectiveness of Residual Quantization Vectors (RVQ) in low-resource scenarios, we fine-tuned the pretrained Wav2Vec 2.0 base model on the same 10-hour LibriSpeech subset. Two configurations were tested:

1. **Baseline**: Pretrained Wav2Vec 2.0 with a standard classification head.
2. **RVQ-enhanced**: Pretrained Wav2Vec 2.0 with RVQ inserted between the embedding layer and the classification head.

Training was performed using CTC loss as the objective function, and both models were evaluated on training and validation subsets. Figures 3 and 4 illustrate the CTC loss curves for training and validation, respectively.

Results show that the RVQ-enhanced model converges faster than the baseline in both training and validation sets. This suggests that in low-resource conditions, residual quantized representations provide more distinct and effective features than directly using the continuous embeddings. Faster convergence is particularly beneficial for low-resource settings where computational resources are limited, as it reduces training time while maintaining performance potential.

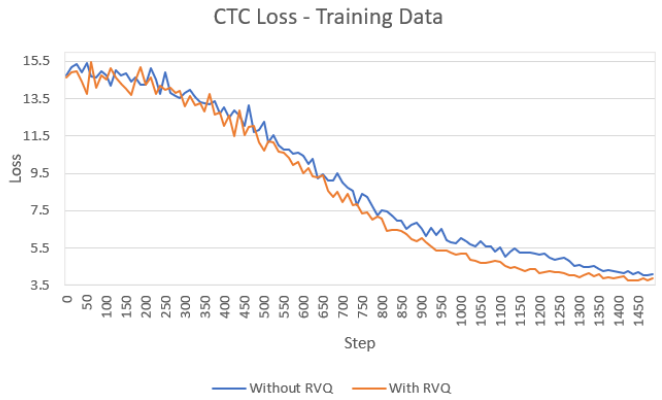

Figure 3: Training CTC loss for models with and without RVQ, showing faster convergence with RVQ integration.

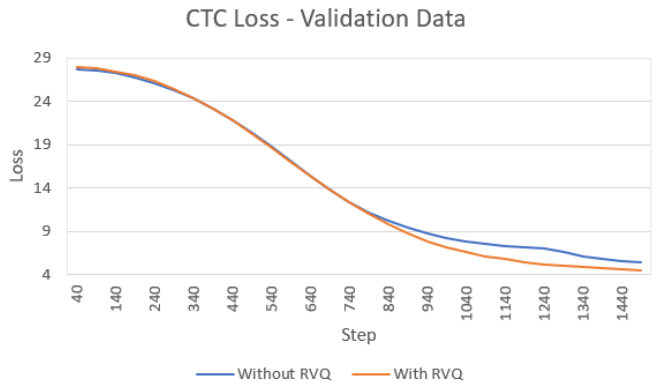

Figure 4: Validation CTC loss for models with and without RVQ, indicating improved convergence with RVQ.

Although full fine-tuning to convergence was not conducted, preliminary observations indicate that RVQ may lead to better final results compared to standard fine-tuning. These findings demonstrate the potential of RVQ for improving model adaptability and efficiency in low-resource speech recognition tasks.

V. FUTURE WORKS

In future work, we plan to conduct full-scale experiments to provide comprehensive results and detailed comparisons across different configurations, including varying numbers of codebooks and training durations. We also aim to explore the use of Residual Quantization Vectors (RVQ) during pretraining as a replacement or complement to existing codebooks, which may further improve representation learning efficiency and stability in low-resource scenarios. Additionally, we intend to systematically study the effect of RVQ hyperparameters, such as the number of quantization levels, residual vector dimensions, and stacking strategies, to identify optimal configurations for both pretraining and fine-tuning. Finally, we plan to evaluate our methods on other low-resource languages and specialized speech domains, such as child speech or animal sounds, and investigate potential computational efficiency gains, including reduced training time and memory usage, to support resource-constrained research environments.

VI. CONCLUSION

In this work, we explored methods to improve the efficiency of Wav2Vec 2.0 for low-resource speech recognition tasks. Specifically, we introduced Inter-Codebook Similarity Loss (ICSL) during pretraining to encourage diversity across multiple codebooks and Residual Quantization Vectors (RVQ) during fine-tuning to enhance convergence speed and representation adaptability. Our preliminary experiments on a 10-hour subset of LibriSpeech demonstrate that ICSL with more codebooks leads to faster convergence, while RVQ improves fine-tuning efficiency in low-resource conditions. Although full-scale training was not conducted, these findings suggest that both methods have the potential to reduce data and computational requirements while maintaining or improving performance. Future work will focus on complete training experiments, exploring RVQ during pretraining, tuning hyperparameters, and evaluating our approaches on other low-resource languages and specialized speech domains. Overall, this study provides insights and practical strategies for efficient pretraining and fine-tuning of large speech models in resource-constrained settings.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 22, 2020, *arXiv*: arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477.

[2] T. Reitmaier *et al.*, "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers," in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–17. doi: 10.1145/3491102.3517639.

[3] L. Lugo and V. Vielzeuf, "Towards efficient self-supervised representation learning in speech processing".

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," Sept. 11, 2019, *arXiv*: arXiv:1904.05862. doi: 10.48550/arXiv.1904.05862.

[5] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 22, 2019, *arXiv*: arXiv:1807.03748. doi: 10.48550/arXiv.1807.03748.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," June 14, 2021, *arXiv*: arXiv:2106.07447. doi: 10.48550/arXiv.2106.07447.

[7] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," Oct. 25, 2022, *arXiv*: arXiv:2202.03555. doi: 10.48550/arXiv.2202.03555.

[8] S. Chen *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.

[9] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview," *IEEE Access*, vol. 8, pp. 163829–163843, 2020, doi: 10.1109/ACCESS.2020.3020421.

[10] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, ISCA, Sept. 1997, pp. 2371–2374. doi: 10.21437/Eurospeech.1997-623.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Mar. 09, 2015, *arXiv*: arXiv:1503.02531. doi: 10.48550/arXiv.1503.02531.

[12] Y. Guo, "A Survey on Methods and Theories of Quantized Neural Networks," Dec. 16, 2018, *arXiv*: arXiv:1808.04752. doi: 10.48550/arXiv.1808.04752.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.