Less can be More: Towards a Parameter-Efficient Fine-Tuning of Wav2Vec 2.0 XLSR for Low-Resource Cape Verdean Creole ASR

Mateus N. Andrade[1], **Mouhamadou Lamine Ba**[2], **Idy Diop**[2], **Arlindo O. da Veiga**[1]
[1] University of Cape Verde, Cabo Verde
[2] Université Cheikh Anta Diop, Sénégal
mateus.andrade@docente.unicv.edu.cv, a.veiga@unicv.cv
mouhamadoulamine.ba@esp.sn, idy.diop@esp.sn

Abstract

Automatic Speech Recognition (ASR) for low-resource languages remains challenging due to limited annotated data and high linguistic variability. In this work, we investigate parameter-efficient fine-tuning strategies for Cape Verdean Creole ASR using the Wav2Vec 2.0 XLSR model. We evaluate the impact of structured layer freezing on model performance, training stability, and computational efficiency. Experiments conducted on a newly curated Santiago-dialect dataset show that full fine-tuning achieves the best absolute performance (WER 0.212, CER 0.120). However, several freezing configurations achieve comparable recognition performance while substantially reducing the number of trainable parameters and exhibiting more stable convergence. These results highlight a trade-off between adaptability and efficiency, showing that selective freezing can serve as an effective regularization strategy in low-resource settings. This work provides practical insights into parameter-efficient adaptation for under-resourced Creole languages.

Keywords: ASR, Wav2Vec 2.0, XLSR, Cape Verdean Creole, Low-Resource Languages, Layer Freezing, Computational Efficiency

1. Introduction

Automatic Speech Recognition (ASR) technologies have achieved remarkable progress in recent years, largely driven by deep learning and large-scale self-supervised pre-trained models (Sikasote and Anastasopoulos, 2022; Zhao and Zhang, 2022; Pindoh and Yonta, 2025). However, these advances have disproportionately benefited high-resource languages (Grosman, 2021), while low-resource and under-documented languages continue to face significant performance gaps (Dione, 2021; Yi et al., 2021). This disparity is particularly pronounced for Creole languages (Macaire et al., 2022), which are often characterized by limited annotated data, high dialectal variability, and the absence of standardized orthographic conventions.

Cape Verdean Creole exemplifies these challenges. Spoken across multiple islands, it exhibits substantial phonetic, lexical, and orthographic variation between dialects. Moreover, the lack of large publicly available speech corpora complicates both acoustic and language modeling (Valdman et al., 2015), making direct training of end-to-end ASR systems impractical and often ineffective. In this context, transfer learning from pre-trained multilingual speech models has emerged (Caubrière and Gauthier, 2024) as a promising strategy to enable ASR in such under-resourced settings.

Recent self-supervised models, such as Wav2Vec 2.0 (Anidjar et al., 2024) and its multilingual extensions (e.g., XLSR) (Grosman, 2021; Gulli et al., 2024; Dat et al., 2025), have demonstrated strong cross-lingual transfer capabilities by learning universal acoustic representations from large volumes of unlabeled speech. Nevertheless, effectively adapting these models to low-resource languages remains an open research problem. Excessive fine-tuning can lead to overfitting and unstable convergence, while overly restrictive adaptation can limit the model's ability to capture language-specific phonetic and orthographic characteristics.

An increasingly explored solution is selective layer freezing (Eberhard et al., 2021; Pasula, 2025), in which subsets of pre-trained model layers are kept fixed during fine-tuning. Previous studies suggest that the lower layers encode general acoustic-phonetic representations that can be transferred between languages, whereas the higher layers capture more language-specific information (Yosinski et al., 2014; Peters et al., 2019). Freezing lower layers can therefore act as an implicit regularization mechanism, improving training stability and helping mitigate overfitting in low-resource scenarios. However, the optimal depth and extent of freezing, particularly for Creole languages, remain insufficiently explored.

Given the strong historical and linguistic ties between Portuguese and Cape Verdean Creole, pre-trained models offer a promising foundation for developing ASR systems for this low-resource language. Building on this, we investigate transfer learning strategies for Cape Verdean Creole ASR using a pre-trained Wav2Vec2-large-XLSR model. Our work focuses specifically on structured layer

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 62–71
12 May 2026.

freezing as a mechanism to improve training stability and computational efficiency, analyzing its impact on convergence and final recognition accuracy. To this end, we conducted experiments on a newly curated dataset of 1,787 speech recordings that capture multiple dialectal variations of Cape Verdean Creole. Unlike prior work that often relies solely on Word Error Rate (WER), our evaluation framework incorporates both WER and Character Error Rate (CER). This dual-metric approach enables a finer-grained analysis of orthographic and subword-level errors, a crucial consideration for languages with non-standardized spelling. We systematically vary the depth of frozen layers to identify practical adaptation strategies that balance performance with stability.

This paper makes two primary contributions. First, we introduce a new, curated speech dataset for Cape Verdean Creole, adding to the growing body of resources for African-influenced languages. Second, we provide an empirically validated, parameter-efficient adaptation strategy for large, self-supervised speech models in a low-resource context.

The rest of the paper is structured as follows. Section 2 reviews related work on self-supervised speech models and parameter-efficient adaptation strategies for low-resource ASR. Section 3 presents the Cape Verdean Creole dataset, our baseline ASR model, and the hierarchical layer-freezing strategy adopted in this study. Section 4 presents and discusses the experimental results, covering performance metrics (WER, CER), convergence analysis, error analysis, and computational trade-offs. Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Recent advances in self-supervised learning, particularly Wav2Vec 2.0, have significantly improved ASR performance in resource-constrained scenarios. Self-supervised speech representation learning has fundamentally reshaped the landscape of automatic speech recognition (ASR). In particular, models such as Wav2Vec 2.0 are capable of learning high-level latent acoustic representations from large-scale unlabeled speech corpora, which can subsequently be fine-tuned using relatively limited amounts of labeled data. This self-supervised paradigm has led to substantial improvements in ASR performance by enabling robust transfer learning across diverse languages and application domains (Baevski et al., 2020; Caubrière and Gauthier, 2024). Building upon this foundation, cross-lingual extensions such as XLSR-53 and its successor XLS-R have further demonstrated strong generalization capabilities, especially in low-resource language settings, by leveraging multilingual pre-training to learn language-agnostic acoustic representations (Conneau et al., 2021). These advances provide a strong motivation for exploring targeted adaptation strategies of self-supervised speech models to under-resourced languages, which constitutes the focus of the methodology presented in this work.

Despite these advances, fine-tuning large pre-trained models on limited labeled datasets introduces significant challenges, including overfitting, unstable convergence, and catastrophic forgetting of pre-trained representations, particularly when the target language exhibits high phonetic or orthographic variability. Recent studies have emphasized the importance of controlled fine-tuning strategies—such as selective layer freezing (Pasula, 2025) to mitigate these effects and helps mitigate overfitting in low-resource speech recognition scenarios (Kunze et al., 2022). Eberhard and Zesch (Eberhard et al., 2021) conducted a systematic analysis of layer freezing when transferring ASR models to under-resourced languages, showing that freezing lower layers preserves universal acoustic-phonetic representations and helps mitigate overfitting. Prior work suggests that lower Transformer layers capture general acoustic representations, while higher layers are more task-specific. Similar findings are reported by Pasula (Pasula, 2025), who demonstrates that selective freezing improves convergence speed and reduces error rates in multilingual ASR experiments using the MuST-C dataset.

In parallel, several works, like (Severini, 2023), highlight the importance of character-level evaluation for low-resource languages. While Word Error Rate (WER) remains the dominant metric in ASR, Character Error Rate (CER) provides finer-grained insight into orthographic and subword-level errors, particularly for languages with non-standardized spelling systems (Kumar et al., 2025). This is especially relevant for Creole languages, where spelling variation can inflate WER without necessarily reflecting phonetic recognition errors.

Research on ASR for Creole languages remains limited. Existing studies often focus on Haitian Creole (Havard et al., 2025) or Mauritian and Gwadloupéyen Creole (Macaire et al., 2022), with far fewer contributions addressing Cape Verdean Creole. Reported approaches typically rely on small datasets and do not systematically explore adaptation strategies such as layer freezing or phonetic normalization. Consequently, there is a lack of empirical guidance on how to best adapt large pre-trained ASR models to Creole languages.

In contrast to prior work, the present study combines selective layer freezing, data augmentation, and soft phonetic normalization within a unified ex-

perimental framework. Furthermore, it provides a joint analysis of WER and CER, enabling a more comprehensive evaluation of ASR performance in a linguistically complex, low-resource setting.

3. Methodology

We detail here the methodology used to investigate parameter-efficient adaptation strategies for Cape Verdean Creole ASR. We first describe the speech corpus and the pre-processing steps applied to the audio and text data (Section 3.1). We then present the baseline model architecture and our proposed structured freezing strategy (Section 3.2). Finally, we outline the experimental setup, including training configurations, evaluation metrics, and computational resources (Section 3.3).

3.1. Dataset and Pre-processing

This section describes the speech corpus used in our study and the pre-processing steps applied to the audio and text data to prepare them for model training.

3.1.1. The Cape Verdean Creole Corpus

Cape Verdean Creole exists as a dialectal continuum throughout the Cape Verde archipelago, broadly categorized into the Sotavento (southern) and Barlavento (northern) groups (Veiga, 1995; Baptista, 2002). Although sharing a largely Portuguese-derived lexicon, these varieties differ in phonetic realization, morphosyntactic patterns, and orthographic practices. The absence of a fully unified orthographic standard further contributes to cross-dialectal variation (Veiga, 2004).

The dataset for this study is drawn exclusively from the Santiago dialect, a prominent variety of the Sotavento group. This choice is motivated by the fact that it is one of the most widely spoken and linguistically influential varieties, frequently used in media and education, making it a logical starting point for the development of ASR. The corpus comprises 1,787 speech recordings, totaling approximately 2 hours of audio. Although focused on a single dialect, it presents substantial intra-dialectal variability in speaker background, pronunciation, and transcription practices. Consequently, it provides a challenging and realistic testbed for developing robust ASR models, capturing the phonetic and spelling inconsistencies representative of real-world conditions for Cape Verdean Creole.

3.1.2. Phonetic Normalization

In languages with high orthographic variability, as in our scenario, light normalization strategies are commonly adopted to improve tokenizer consistency and ASR stability, particularly in low resource scenarios (Besacier et al., 2014; Lippmann, 1997). As a result, we implemented a phonetic normalization step to improve data consistency before training. A custom dictionary was created to map common spelling variations and diacritics to a canonical phonetic form (e.g., normalizing words with different accent marks or alternative spellings to a single representation). This light normalization process reduces character-level variance, which is particularly beneficial for the stability of the CTC-based decoder and for obtaining a more meaningful Character Error Rate (CER) by preventing the model from being penalized for inconsistent but phonetically equivalent spellings.

3.1.3. Data Augmentation

To mitigate data scarcity and improve model robustness against acoustic variations, we applied online data augmentation during training, following the methods described by (Huh et al., 2023). The augmentations included temporal perturbations, such as varying the speaking rate, and the addition of background noise to simulate different acoustic environments. This is particularly important for Creole ASR, where available recordings often reflect limited acoustic diversity, and facilitates improved generalization to unseen conditions.

3.2. Model and Adaptation Strategy

This section details the architecture of the baseline ASR model and introduces our structured freezing strategy for parameter-efficient adaptation.

3.2.1. Baseline Model Architecture

Our baseline ASR system is built upon the Wav2Vec 2.0 large-XLSR architecture, a Transformer-based model designed for multilingual self-supervised learning (see Figure 1). The model processes the raw audio waveform through a convolutional feature extractor, which generates latent acoustic representations. These representations are then fed into a 24-layer Transformer encoder that models long-range temporal dependencies using self-attention mechanisms. Finally, a linear layer followed by a Connectionist Temporal Classification (CTC) head projects the contextualized hidden states to the target vocabulary for sequence prediction.

Based on a preliminary study aimed at selecting the most suitable model initialization (see Table 1), all subsequent experiments were initialized from the XLSR-CORAA checkpoint. This model, fine-tuned on Portuguese from the multilingual XLSR-53 backbone, demonstrated the competitive out-of-the-box performance on Cape Verdean Creole.

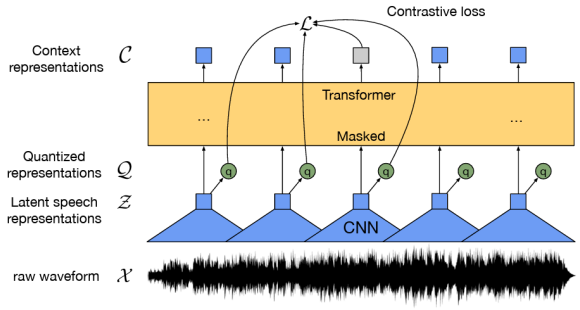

Figure 1: The Wav2Vec 2.0 approach (Baevski et al., 2020)

To ensure a fair comparison across configurations, the core architecture was kept identical in all experiments.

The preliminary evaluation was conducted using two test sets with different sizes and characteristics:

- **Test Configuration (TC) 1:** 132 Creole audio files in `.mp3` format (9 minutes and 38 seconds, 13.1 MB).
- **Test Configuration (TC) 2:** 366 Creole audio files in both `.mp3` and `.wav` formats (22 minutes and 43 seconds, 32.6 MB).

We evaluate four pre-trained Wav2Vec2.0 models: W2V2-PT/EN/FR, a set of Monolingual Models for Portuguese, English, and French (Jonatasgrosman/wav2vec2-large-xlsr-53-*); XLSR-CORAA, a multilingual model fine-tuned on the Portuguese CORAA dataset (Edresson/wav2vec2-large-xlsr-coraa-portuguese); W2V2-960h, a large model pre-trained and fine-tuned on 960 hours of LibriSpeech (facebook/wav2vec2-large-960h); and W2V2-LV60, a large model pre-trained on 60k hours of speech (facebook/wav2vec2-large-lv60).

These experiments provided the basis for selecting the baseline model and establishing the evaluation setup used in the subsequent analysis. Preliminary results exhibited elevated CER values, later attributed to the absence of standardized text normalization and improper decoding of masked labels, as discussed in Section 4.

Language	Model	TC	WER	CER	Time
Portuguese	W2V2-PT	1	39.1	84.5	05:20
	W2V2-PT	2	34.3	84.1	13:22
	XLSR-CORAA	1	43.7	83.5	06:10
	XLSR-CORAA	2	32.9	84.0	16:11
English	W2V2-960h	2	100.0	100.0	12:08
	W2V2-LV60	2	96.9	98.2	13:38
	W2V2-EN	2	36.5	84.4	13:27
French	W2V2-FR	2	37.7	85.3	13:36

Table 1: Performance comparison of pretrained Wav2Vec 2.0 models (monolingual and multilingual) for baseline model selection.

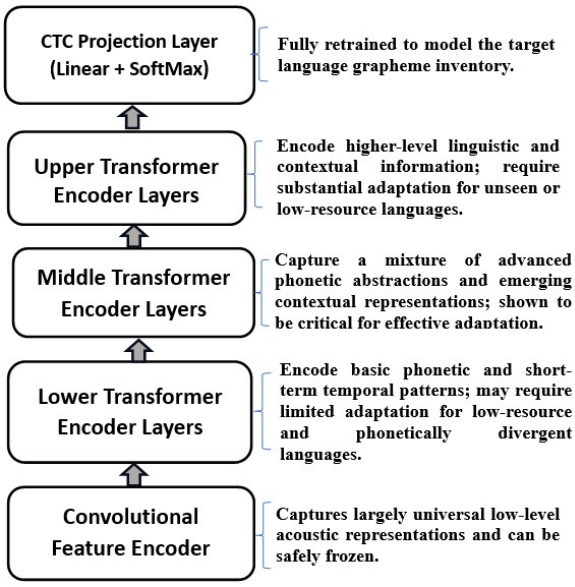

Figure 2: Conceptual representation of linguistic abstraction across Wav2Vec 2.0 layers and its implications for selective layer freezing in Creole ASR

3.2.2. Structured Freezing Strategy

Our adaptation methodology is motivated by the hierarchical nature of the Wav2Vec 2.0 architecture (see Figure 2). Based onon findings from Eberhard et al. (2021) that the lower layers of the Transformer capture universal acoustic-phonetic features, we adopt a hierarchical freezing strategy that prioritizes the progressive freezing of the upper layers. To test this, we employ a structured freezing strategy that systematically varies the adaptation depth to analyze the trade-off between preserving robust pre-trained representations and enabling sufficient adaptation to Cape Verdean Creole.

As depicted in Table 2, we define several experimental configurations:

- **OF (Full Fine-Tuning):** The baseline where all model parameters (feature extractor and Transformer) are trainable.
- **FEF (Feature Extractor Frozen):** Only the convolutional feature extractor is frozen; all 24 Transformer layers are trained.
- **F > N (Selective Freezing):** Transformer layers 0 to N are trainable, while layers N+1 to 23 are frozen.

This structured approach allows for a controlled analysis of how adaptation depth influences not only recognition accuracy but also computational efficiency and convergence stability—critical considerations for low-resource ASR development.

Cfg	TL	Params (%)	Time
OF	0–23	100.0	01:12:51
FEF	0–23	98.7	01:06:50
F>3	0–3	14.8	01:00:28
F>5	0–5	22.8	01:33:21
F>6	0–6	26.8	00:45:16
F>11	0–11	46.8	00:46:43
F>12	0–12	50.8	00:46:50
F>13	0–13	54.8	00:46:48
F>14	0–14	58.7	00:47:21
F>15	0–15	62.7	00:46:41
F>16	0–16	66.7	01:08:27
F>20	0–20	82.7	01:08:27

Table 2: Selective freezing configurations, proportion of trainable parameters, and training time.

OF: No freezing; *FEF*: Feature extractor frozen; *F>N*: Transformer layers above N frozen; *TL*: Trainable Layers.

3.3. Experimental Setup

This section outlines the training configurations, evaluation metrics, and computational resources used to conduct our experiments.

3.3.1. Training Configurations

The dataset was partitioned into training and test sets using an 80/20 split with a fixed random seed (seed=42) to ensure reproducibility. The test set was strictly held out and not used during training or model development, and no overlap exists between training and test samples. Given the limited size of the dataset, no separate validation set was used.

Preprocessing was applied separately to each split. Data augmentation techniques were applied exclusively to the training set to improve model robustness, while the test set was processed without augmentation to ensure a fair and unbiased evaluation. All reported WER and CER results are computed on this unseen test set.

To ensure a fair comparison, all models were fine-tuned under identical experimental conditions. The models were trained for a total of 35 epochs using the AdamW optimizer with a learning rate of 5e-5 and a batch size of 8. Model selection was based on training dynamics due to the limited size of the dataset. All other hyperparameters were kept at their default values as specified in the original Wav2Vec 2.0 implementation.

3.3.2. Evaluation Metrics (WER, CER)

We evaluate model performance using two standard ASR metrics: Word Error Rate (WER) and Character Error Rate (CER). WER measures the number of word-level substitutions, deletions, and insertions required to match the hypothesis to the reference transcription. While WER is the primary metric for ASR, CER provides a complementary, finer-grained analysis by operating at the character level.

Given the high orthographic variability and lack of a standardized writing system for Cape Verdean Creole, CER is particularly important. It allows us to assess the model's ability to learn phonetic and sub-word regularities, even when word-level transcriptions are inconsistent. This dual-metric approach enables a more comprehensive evaluation of how our adaptation strategies influence both lexical accuracy and character-level robustness.

3.3.3. Computational Resources

All experiments were conducted using Google Colab as the computational platform. The hardware environment consisted of 53 GB of system RAM, 22.5 GB of GPU memory, and 235.7 GB of available disk space. This configuration provided sufficient resources to fine-tune the large-scale pr-trained speech models while maintaining consistent experimental conditions across all freezing configurations.

4. Results and Discussion

The experimental results presented in this section directly reflect the architectural hypotheses introduced in Section 3.2.2. By progressively freezing lower Transformer layers while keeping the feature extractor frozen across all configurations, we isolate the effect of Transformer-level adaptation and evaluate its influence on training dynamics and recognition performance.

As illustrated in Figure 3, the experimental pipeline combines data augmentation strategies, similarity analysis, and detailed error inspection. While augmentation is used to improve robustness, it does not introduce domain-level or speaker-level diversity in terms of domain or speaker variability. The dataset remains limited in this regard, as it originates from a single domain and publisher, as discussed in Section 6. This structured approach supports a multi-level evaluation of ASR performance, integrating WER, CER, and error-type distributions, which is particularly relevant for languages with high orthographic and phonetic variability.

The evaluation focuses on training dynamics, recognition accuracy, and stability, using validation loss, Word Error Rate (WER), and Character Error Rate (CER) as primary metrics.

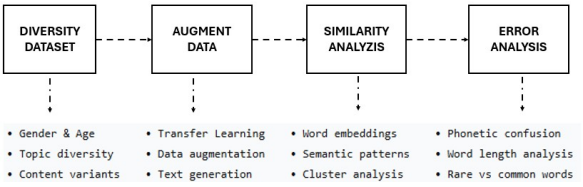

Figure 3: Wav2Vec2-large-XLSR Model Improvement Experiment Strategies

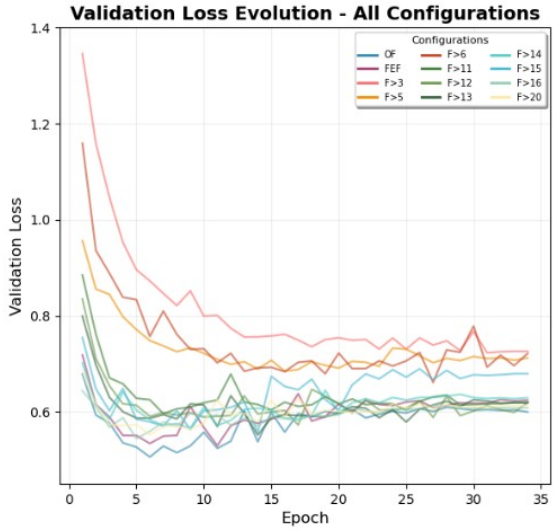

Figure 4: Loss Validation Evolution

4.1. Training Dynamics and Validation Loss

Figure 4 illustrates the evolution of validation loss across all freezing configurations. Models without frozen layers exhibit a rapid decrease in loss during the initial training epochs, reaching the competitive values overall. However, this behavior is accompanied by noticeable oscillations in later epochs, indicating reduced training stability and a higher susceptibility to overfitting.

In contrast, models with selective layer freezing demonstrate smoother loss trajectories and more stable convergence patterns. Configurations with moderate freezing (approximately 5–11 frozen layers) achieve validation loss values comparable to the non-frozen baseline while exhibiting substantially lower variance. Extensive freezing (>12 layers), although beneficial for stable convergence, results in slightly higher validation loss, suggesting a reduced capacity for adaptation.

These findings support the hypothesis that freezing lower layers preserves pretrained acoustic representations while limiting excessive parameter updates that can destabilize training in low-resource settings.

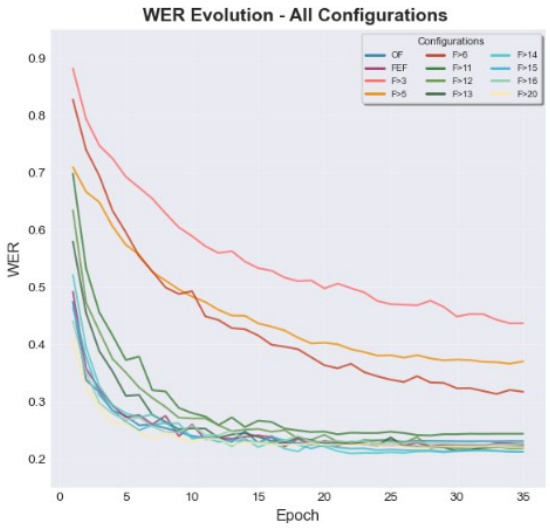

Figure 5: WER Evolution by Configuration

4.2. Word Error Rate (WER) Performance

Figure 5 presents the evolution of WER across training epochs for all freezing configurations. The fully trainable baseline (0F) converges rapidly and achieves the lowest absolute WER (0.212). However, its trajectory exhibits greater oscillation in later epochs compared to several selectively frozen configurations.

Models employing moderate to extended freezing depths (F>14–F>15) achieve nearly identical final WER values (0.213), differing from the baseline by only 0.001. Configurations such as F>12 and F>13 remain competitive (0.219–0.222), while more aggressive freezing (F>3–F>6) leads to clear degradation in performance.

These results indicate that selective hierarchical freezing does not surpass full fine-tuning in absolute WER. Instead, it achieves comparable recognition accuracy while substantially reducing the number of trainable parameters and exhibit more stable convergence. The minimal performance gap between 0F and F>14–F>15 suggests that full adaptation of all Transformer layers is not strictly necessary for competitive word-level recognition in this low-resource setting.

The divergence between validation loss and WER further illustrates that small differences in optimization objective values do not always translate into meaningful improvements in recognition accuracy, particularly in linguistically variable and orthographically non-standardized languages.

4.3. Character Error Rate (CER) Performance

Character-level evaluation provides complementary insight into subword modeling behavior. As

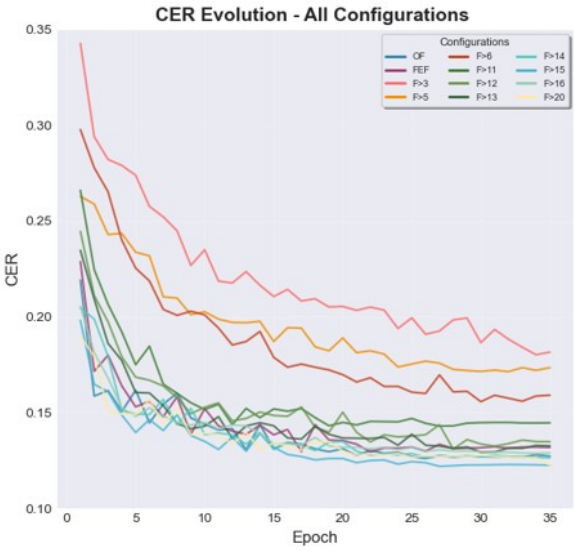

Figure 6: CER Evolution by Configuration

shown in Table 3, the fully trainable baseline (0F) achieves the competitive absolute CER (0.120). Nevertheless, configurations with moderate to extended freezing depths maintain highly competitive character-level performance, with CER values ranging from 0.123 (F>20) to 0.127 (F>15).

While selective freezing does not outperform the baseline in absolute CER, several frozen configurations exhibit more stable convergence. In particular, models with freezing beyond 12 layers demonstrate reduced fluctuation in CER across epochs compared to the fully trainable model.

Aggressive freezing (F>3–F>6) substantially increases CER (0.159–0.182), confirming that excessive constraint of Transformer adaptation limits subword modeling capacity.

Overall, Frozen configurations exhibit more stable convergence compared to full fine-tuning.

4.4. Convergence Speed and Training Stability

The observed stability improvements with moderate freezing depths are consistent with prior observations that partial fine-tuning can mitigate overfitting in low-data regimes (Peters et al., 2019). Models employing layer freezing consistently reach stable WER and CER values faster than the non-frozen model. This effect is most pronounced in configurations with 11-15 frozen layers, where early convergence is achieved with reduced fluctuation.

4.5. Trade-offs Between Adaptability and Generalization

The results highlight a clear trade-off between full adaptability and controlled regularization. While full fine-tuning achieves the best absolute performance, selective layer freezing enables a more efficient adaptation regime with comparable recognition quality.

From a regularization perspective, hierarchical freezing constrains model adaptation in a structured manner, reducing overfitting. Frozen configurations exhibit more stable convergence in low-resource conditions. This suggests that competitive performance can be achieved without fully updating all model parameters, balancing accuracy, efficiency, and robustness.

4.6. Implications for Low-Resource and Creole ASR

The experimental results have several implications for ASR in low-resource and Creole language contexts:

1. **Layer freezing should be considered a standard adaptation strategy** in low-resource settings, particularly when computational efficiency and training stability are critical.

2. While **WER and CER are highly correlated** in our experiments, CER provides additional insight into character-level errors, which is particularly relevant for orthographically variable languages.

3. **Training stability and convergence behavior** are critical evaluation dimensions in low-resource ASR, beyond final accuracy metrics.

By systematically analyzing these factors, this study provides practical guidance for deploying robust ASR systems in linguistically complex and under-resourced environments.

4.7. Error Analysis by Freezing Configuration

Selective freezing depth directly modulates both accuracy and computational efficiency. Configurations with a high proportion of trainable parameters (0F, FEF) show unstable substitution and merge patterns despite lower omission rates, suggesting overfitting in upper Transformer layers. Conversely, aggressive freezing (F>3–F>6), which drastically reduces the percentage of trainable parameters, increases omissions and word merges, leading to the highest WER and CER values.

Intermediate configurations (F>11–F>15), corresponding to a moderate proportion of trainable layers, achieve the competitive trade-off between performance and efficiency. These settings minimize structurally disruptive errors while stabilizing insertion rates, yielding the competitive WER. Deeper freezing (F>20) maintains competitive CER while

significantly reducing the number of trainable parameters.

As illustrated in Figure 5 (WER × % trainable × training time), performance follows a non-linear trend: neither full fine-tuning nor excessive freezing is optimal. Instead, moderate layer freezing maximizes accuracy while substantially reducing computational cost, indicating a regularization effect that mitigates lexical overfitting without sacrificing generalization.

5. Conclusion

This work investigated parameter-efficient fine-tuning strategies for low-resource Cape Verdean Creole ASR using Wav2Vec 2.0 XLSR. The results demonstrate that selective layer freezing can maintain performance comparable to full fine-tuning while reducing computational requirements and improving training stability.

These findings highlight the potential of structured freezing as a practical adaptation strategy for low-resource ASR. However, the conclusions are limited by the size and scope of the dataset, and future work should explore larger and more diverse corpora, as well as alternative parameter-efficient approaches such as adapters and LoRA.

6. Ethical considerations and limitations

6.1. Ethical Considerations

The speech data used in this study were collected from publicly accessible audio materials available on the official website of Jehovah's Witnesses. No private or sensitive data were collected, and no direct interaction with speakers occurred. However, the recordings were not originally produced for ASR research purposes, and explicit consent for computational reuse was not formally documented. Future corpus development should prioritize informed consent procedures tailored to language technology research.

The dataset reflects a single religious and communicative domain, which may introduce lexical and stylistic bias. Sustainable ASR development for Cape Verdian Creole should involve local institutions and community stakeholders to ensure responsible and inclusive technological deployment.

This research does not involve biometric identification, surveillance, or human subject experimentation.

6.2. Limitations

The corpus comprises approximately 2 hours of speech (1,787 recordings), which limits generalization compared to large-scale ASR datasets. Only the Santiago variety (Sotavento group) of Cape Verdian Creole is represented; therefore, findings cannot be generalized to other dialects.

All recordings originate from a single domain, potentially affecting robustness in spontaneous speech contexts. Although phonetic normalization was applied, the absence of a standardized orthography may influence WER evaluation. Finally, the study focuses exclusively on selective layer freezing applied to a pretrained Facebook AI Research Wav2Vec2-large-XLSR model; alternative parameter-efficient adaptation strategies were not explored.

7. Acknowledgements

The authors wish to express their sincere gratitude to the Responsible Artificial Intelligence Lab at Kwame Nkrumah University of Science and Technology (RAIL–KNUST, Kumasi, Ghana), to the International Development Research Centre (IDRC, Ottawa, Canada), and to UK International Development (London, UK), through the AI4PEP Network, for the financial support provided for this research. The authors also acknowledge the support of Université Cheikh Anta Diop (UCAD, Dakar, Senegal) and the DiCentre4AI project, an initiative supported by IDRC and the Foreign, Commonwealth & Development Office (FCDO). Further appreciation is extended to RS2Lab at Uni-CV and to colleagues with whom the laboratory is shared on a daily basis. The authors also thank the linguistics professors at Uni-CV for their continued support throughout this work.

8. Bibliographical References

Anidjar, O. H., Marbel, R., and Yozevitch, R. (2024). Whisper turns stronger: Augmenting Wav2Vec 2.0 for Superior ASR in Low-Resource Languages. *ArXiv*, abs/2501.00425.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*.

Baptista, M. (2002). *The Syntax of Cape Verdean Creole: The Sotavento Varieties*. Linguistik Aktuell/Linguistics Today. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic Speech Recognition for Under-

Category	Metric	0F	FEF	F>3	F>5	F>6	F>11	F>12	F>13	F>14	F>15	F>16	F>20
Global Performance	WER	0.212	0.437	0.226	0.370	0.318	0.244	0.219	0.222	0.213	0.213	0.228	0.221
	CER	0.120	0.132	0.182	0.173	0.159	0.146	0.135	0.133	0.126	0.127	0.129	0.123
Error Types	Phonetic Substitutions	85	92	59	69	75	80	79	85	89	84	88	80
	Omissions	262	305	407	423	368	363	346	322	307	312	296	268
	Insertions	82	65	91	73	89	51	46	55	51	54	63	75
	Word merges	126	132	383	292	255	149	140	135	140	133	139	136
Vocabulary Analysis	Rare Words Errors	139	152	195	211	196	171	152	154	149	140	156	153
	Common Words Errors	377	407	842	700	581	450	398	408	380	390	407	388

Table 3: ASR performance and error analysis across freezing configurations.

Resourced Languages: A Survey. In *Speech Communication*, pages 85–100.

Caubrière, A. and Gauthier, E. (2024). Africa-Centric Self-Supervised Pre-Training for Multilingual Speech Representation in a Sub-Saharan Context. *ArXiv*, abs/2404.02000.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised Cross-lingual Representation Learning for Speech Recognition. In *Proceedings of Interspeech*, pages 2426–2430.

Dat, P. T., Dat, T. H., et al. (2025). Xlsr-Kanformer: A KAN-Intergrated model for Synthetic Speech Detection. In *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, page 1–6. IEEE.

Dione, C. M. B. (2021). Multilingual Dependency Parsing for Low-Resource African Languages: Case Studies on Bambara, Wolof, and Yoruba. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92, Online. Association for Computational Linguistics.

Eberhard, O. et al. (2021). Effects of Layer Freezing on Transferring a Speech Recognition System to Under-resourced Languages. In Evang, K., Kallmeyer, L., Osswald, R., Waszczuk, J., and Zesch, T., editors, *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 208–212, Düsseldorf, Germany. KONVENS 2021 Organizers.

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

Gulli, A., Costantini, F., Sidraschi, D., and Li Destri, E. (2024). Fine-Tuning a Pre-Trained Wav2Vec2 model for Automatic Speech Recognition - Experiments with de Zahrar Sproche. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7336–7342, Torino, Italia. ELRA and ICCL.

Havard, W. N., Govain, R., Lecouteux, B., and Schang, E. (2025). Self-Supervised Models of Speech Processing for Haitian Creole. *BABEL*, 1091(547):544.

Huh, M., Ray, R., and Karnei, C. (2023). A Comparison of Speech Data Augmentation Methods Using S3PRL Toolkit. *arXiv preprint arXiv:2303.00510*.

Kumar, T. D., James, J., Gopinath, D. P., and Krishnan, M. A. (2025). Advocating Character Error Rate for Multilingual ASR Evaluation. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4941–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Kunze, J., Kirsch, L., and Schütze, H. (2022). Transfer Learning for Low-Resource Speech Recognition. In *Proceedings of the International Conference on Speech and Computer*. Springer.

Lippmann, R. P. (1997). Speech Recognition by Machines and Humans. *Speech Communication*, 22(1):1–15.

Macaire, C., Schwab, D., Lecouteux, B., and Schang, E. (2022). Automatic Speech Recognition and Query By Example for Creole Languages Documentation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Pasula, R. (2025). Optimizing Speech Models with Freezing. *International Journal of Innovative Science and Research Technology*, pages 69–73.

Peters, M. E., Ruder, S., and Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 7–14.

Pindoh, P. D. and Yonta, P. M. (2025). Self-supervised and multilingual learning applied to the wolof, swahili and fongbe. *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées*, Volume 42 - Special issue CRI 2023 - 2024/2025.

Severini, S. (2023). *Character-Level and Syntax-Level Models for Low-Resource and Multilingual Natural Language Processing*. PhD thesis, lmu.

Sikasote, C. and Anastasopoulos, A. (2022). BembaSpeech: A Speech Recognition Corpus for the Bemba Language. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Valdman, A., Villeneuve, A.-J., and Siegel, J. (2015). On the influence of the standard norm of Haitian Creole on the Haïtien dialect: Evidence from sociolinguistic variation in the third person singular pronoun. *Journal of Pidgin and Creole Languages*, 30:1–43.

Veiga, M. (1995). *O Crioulo de Cabo Verde: Introdução à Gramática*. Instituto Caboverdiano do Livro.

Veiga, M. (2004). *A Construção do Bilinguismo*. Instituto da Biblioteca Nacional.

Yi, C., Wang, J., Cheng, N., Zhou, S., and Xu, B. (2021). Transfer Ability of Monolingual Wav2Vec2.0 for Low-resource Speech Recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How Transferable Are Features in Deep Neural Networks? *Advances in Neural Information Processing Systems*, pages 3320–3328.

Zhao, J. and Zhang, W.-Q. (2022). Improving Automatic Speech Recognition Performance for Low-Resource Languages with Self-Supervised Models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1227–1241.