CS4681 - Advanced Machine Learning Progress Report

P.P.D. Fernando
210163M

Table Of Contents

1. Project Overview

1.1 Project Title

Enhanced WeNet Speech Recognition Framework with WavLM Self-Supervised Pre-training Integration

1.2 Selected Baseline Model

WeNet Toolkit - A production-oriented, streaming and non-streaming end-to-end speech recognition framework that implements the U2 (Unified Two-Pass) architecture with Transformer/Conformer encoders.

1.3 Enhancement Source Model

WavLM (Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing) - A state-of-the-art self-supervised learning model that achieves superior performance on the SUPERB benchmark and various speech processing tasks including ASR, speaker verification, and diarization.

1.4 Enhancement Objectives

The primary goal is to integrate WavLM's superior self-supervised representations into WeNet's production-ready framework, targeting measurable improvements in:

- Integrate WavLM embeddings into WeNet’s encoder for ASR tasks.
- Evaluate performance improvements on benchmark dataset (LibriSpeech).
- Assess model robustness on noisy speech samples.
- Maintain reasonable inference speed while using WavLM features.

2. Literature Review

2.1 Self-Supervised Speech Learning Evolution

2.1.1 Foundation Models

The evolution of self-supervised learning in speech began with wav2vec (Schneider et al., 2019), which demonstrated that unsupervised pre-training could significantly improve ASR performance. This was followed by wav2vec 2.0 (Baevski et al., 2020), which introduced contrastive learning and achieved remarkable results with minimal labeled data.

2.1.2 HuBERT Framework

HuBERT (Hsu et al., 2021) advanced the field by introducing masked prediction learning similar to BERT in NLP. It uses k-means clustering to create discrete targets for masked audio segments, enabling effective self-supervised learning on continuous speech signals.

2.1.3 WavLM Innovations

WavLM (Chen et al., 2021) represents the current state-of-the-art in self-supervised speech learning with several key innovations:

- Masked Speech Denoising and Prediction: Unlike previous models that focus primarily on clean speech, WavLM jointly learns from noisy/overlapped speech simulation, enabling superior performance on non-ASR tasks
- Gated Relative Position Bias: Enhances the Transformer's ability to capture sequence ordering by adaptively adjusting position bias based on speech content
- Large-Scale Diverse Training: Utilizes 94k hours from LibriLight, GigaSpeech, and VoxPopuli, reducing domain mismatch issues
- Full-Stack Performance: Achieves SOTA results across 15 SUPERB tasks including speaker verification (0.383% EER), speech separation (27.7% WER reduction), and diarization (12.6% DER reduction)

2.2 Production-Oriented Speech Recognition

2.2.1 WeNet Framework Architecture

WeNet (Yao et al., 2021) addresses the critical gap between research and production in E2E speech recognition:

- U2 (Unified Two-Pass) Framework: Unifies streaming and non-streaming models using dynamic chunk-based attention
- Joint CTC/AED Training: Combines CTC and attention-based encoder-decoder for improved stability and performance
- Production-Ready Runtime: Supports both x86 server and Android deployment with quantization support
- PyTorch-Only Ecosystem: Eliminates Kaldi dependencies for simplified installation and deployment

2.2.2 WeNet 2.0 Enhancements

Recent developments include U2++ with bidirectional attention decoders, WFST-based language model integration, contextual biasing, and unified I/O for large-scale training, achieving up to 10% relative improvement over the original U2.

2.3 Integration Opportunities and Challenges

2.3.1 Complementary Strengths

- WavLM: Superior universal representations, multi-task capabilities, noise robustness
- WeNet: Production-ready framework, streaming support, efficient deployment

2.3.2 Technical Challenges

- Feature extraction alignment between WavLM's 20ms stride and WeNet's processing pipeline
- Memory and computational efficiency during fine-tuning and inference
- Maintaining streaming capabilities while leveraging WavLM's full-context benefits
- Balancing universal representations with task-specific optimizations

2.4 Research Gap

While WavLM demonstrates exceptional performance across speech tasks and WeNet provides production-ready deployment, no comprehensive study has explored their systematic integration for enhanced production speech recognition systems.

3. Methodology Outline

3.1 Baseline Establishment

Phase 1: WeNet Baseline Setup

- Install and configure WeNet, verify training/inference pipeline.

Phase 2: WavLM Model Analysis

- Replicate reported results on LibriSpeech (test-clean/test-other)
- Record baseline WER, inference speed, and resource usage

3.2 Integration Design and Implementation

Phase 3: Architecture Design

- Replace WeNet’s raw acoustic features with WavLM pre-trained embeddings.
- Explore a simple **layer selection or averaging** method for representation usage.

Phase 4: Implementation

- Build a WavLM feature extraction module for WeNet
- Add basic configuration support for WavLM models.
- Enable end-to-end training with WavLM features.

3.3 Enhancement Strategies

Strategy 1: Training Strategy

- Start with frozen WavLM layers, fine-tune selectively

Strategy 2: Regularization

- Apply simple data augmentation (e.g., noise addition).

Strategy 3: Loss Function Optimization

- Knowledge distillation from WavLM to WeNet encoder
- Regularization techniques preventing overfitting to specific domains
- CTC/attention loss rebalancing for improved streaming performance

3.4 Experimental Framework

Phase 5: Systematic Evaluation

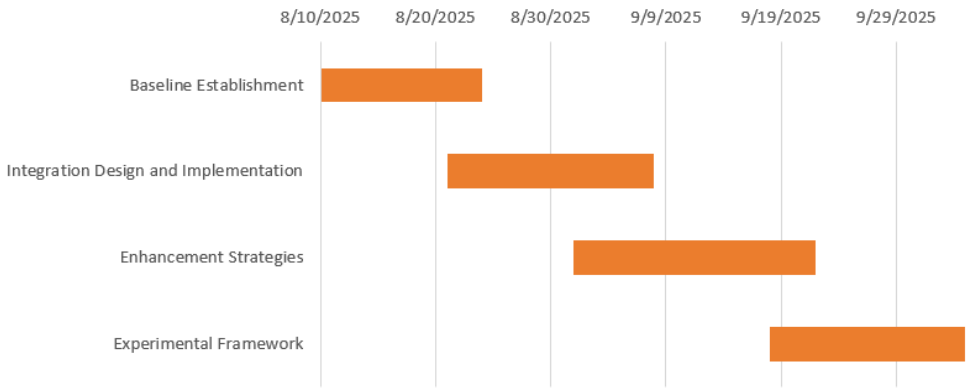
- Datasets: LibriSpeech (English), noisy speech corpora
- Metrics: WER, latency, memory usage, RTF (real-time factor)
- Baselines: Baseline WeNet vs. WavLM-integrated WeNet
- Ablation Studies: Layer combination strategies, training procedures, streaming chunk sizes

Phase 6: Advanced Optimization

- Hyperparameter optimization using grid search or Bayesian optimization
- Model compression techniques (quantization, pruning, knowledge distillation)

4. Project Timeline and Gantt Chart

The project follows a structured 9-week timeline with overlapping phases to ensure continuous progress and iterative improvement. Key milestones align with the assignment's evaluation schedule:



5. Project Planning and Resource Management

5.1 Dataset Preparation

- LibriSpeech: 960h training data already accessible
- Noise Augmentation: DNS Challenge datasets for robustness testing
- Preprocessing Pipeline: Automated scripts for feature extraction and data loading

5.2 Development Environment

- Framework: PyTorch 1.13+, WeNet 2.0, HuggingFace Transformers
- Containerization: Docker environment for reproducible experiments
- Version Control: Git repository with comprehensive documentation
- Monitoring: Weights & Biases for experiment tracking and visualization

5.3 Risk Assessment and Mitigation Strategies

5.3.1 Technical Risks

1. Memory constraints during WavLM integration
    - Mitigation: Implement gradient accumulation, model sharding, and efficient attention mechanisms

2. Performance degradation in streaming scenarios

    - Mitigation: Design chunk-wise processing with look-ahead mechanisms, profile latency extensively
3. Convergence issues during joint training
    - Mitigation: Progressive training strategy, careful learning rate scheduling, extensive hyperparameter tuning

5.3.2 Timeline Risks

1. Extended debugging and integration time
    - Mitigation: Maintain modular design, implement incremental testing, prepare fallback approaches
2. Insufficient computational resources
    - Mitigation: Optimize model sizes, use cloud computing services if needed, focus on most promising approaches

6. Conclusion

This progress evaluation establishes a comprehensive framework for enhancing WeNet's speech recognition capabilities through systematic integration of WavLM's self-supervised representations. The methodology combines rigorous experimental design with practical production constraints, ensuring both scientific validity and real-world applicability.

Expected outcomes include measurable improvements in speech recognition accuracy, robustness, and efficiency, contributing valuable insights to both the research community and practical speech recognition deployments.

7. References

1. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. and Wu, J., 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), pp.1505-1518. Available at: https://arxiv.org/abs/2110.13900
2. Tan, D., Lee, T. (2021) Fine-Grained Style Modeling, Transfer and Prediction in Text-to-Speech Synthesis via Phone-Level Content-Style Disentanglement. Proc. Interspeech 2021, 4683-4687, doi: 10.21437/Interspeech.2021-1129
3. Zhang, B., Wu, D., Peng, Z., Song, X., Yao, Z., Lv, H., Xie, L., Yang, C., Pan, F. and Niu, J., 2022. Wenet 2.0: More productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455*. Available at: https://arxiv.org/abs/2203.15455
4. Pham, N.-Q., Ha, T.-L., Nguyen, T.-N., Nguyen, T.-S., Salesky, E., Stüker, S., Niehues, J., Waibel, A. (2020) Relative Positional Encoding for Speech Recognition and Direct Translation. Proc. Interspeech 2020, 31-35, doi: 10.21437/Interspeech.2020-2526
5. Chi, Z., Huang, S., Dong, L., Ma, S., Zheng, B., Singhal, S., Bajaj, P., Song, X., Mao, X.L., Huang, H. and Wei, F., 2021. XLM-E: Cross-lingual language model pre-training via ELECTRA. *arXiv preprint arXiv:2106.16138*.Available at: https://arxiv.org/abs/2106.16138

6. V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.
7. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R. (2020) Conformer: Convolution-augmented Transformer for Speech Recognition. Proc. Interspeech 2020, 5036-5040, doi: 10.21437/Interspeech.2020-3015
8. 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," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451-3460, 2021, doi: 10.1109/TASLP.2021.3122291.
9. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33, 12449-12460.
10. Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V. (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proc. Interspeech 2019, 2613-2617, doi: 10.21437/Interspeech.2019-2680