k2SSL: A Faster and Better Framework for Self-Supervised Speech Representation Learning

Yifan Yang*, Jianheng Zhuo*, Zengrui Jin, Ziyang Ma, Xiaoyu Yang, Zengwei Yao, Liyong Guo, Wei Kang, Fangjun Kuang, Long Lin, Daniel Povey, Xie Chen

Next-gen Kaldi Team & X-LANCE Lab

July 2, 2025

The Challenge in Speech-based SSL

- Speech-based SSL models offer generalization ability and can derive universal representations from unlabeled data.
- However, they are incredibly resource-intensive, demanding enormous computational resources.
  - Pre-training HuBERT Base requires 32 GPUs; HuBERT Large needs 128 GPUs, and HuBERT X-Large needs 256 GPUs.
  - WavLM Base needs 32 GPUs, while WavLM Large needs 64 GPUs.
- This high cost in time and memory makes cutting-edge SSL inaccessible for most researchers.
- Furthermore, existing open-source frameworks struggle with excessive memory demands and inefficient data management when handling today's massive datasets, which range up to millions of hours.

Goal

There is a clear need for improving the accessibility, efficiency, and effectiveness of speech-based SSL systems.

Our Solution: The k2SSL Framework

We introduce **k2SSL**, an open-source framework designed to be faster, more memory-efficient, and higher-performing for speech SSL. Our solution is built on two key aspects:

1. Optimized Architecture

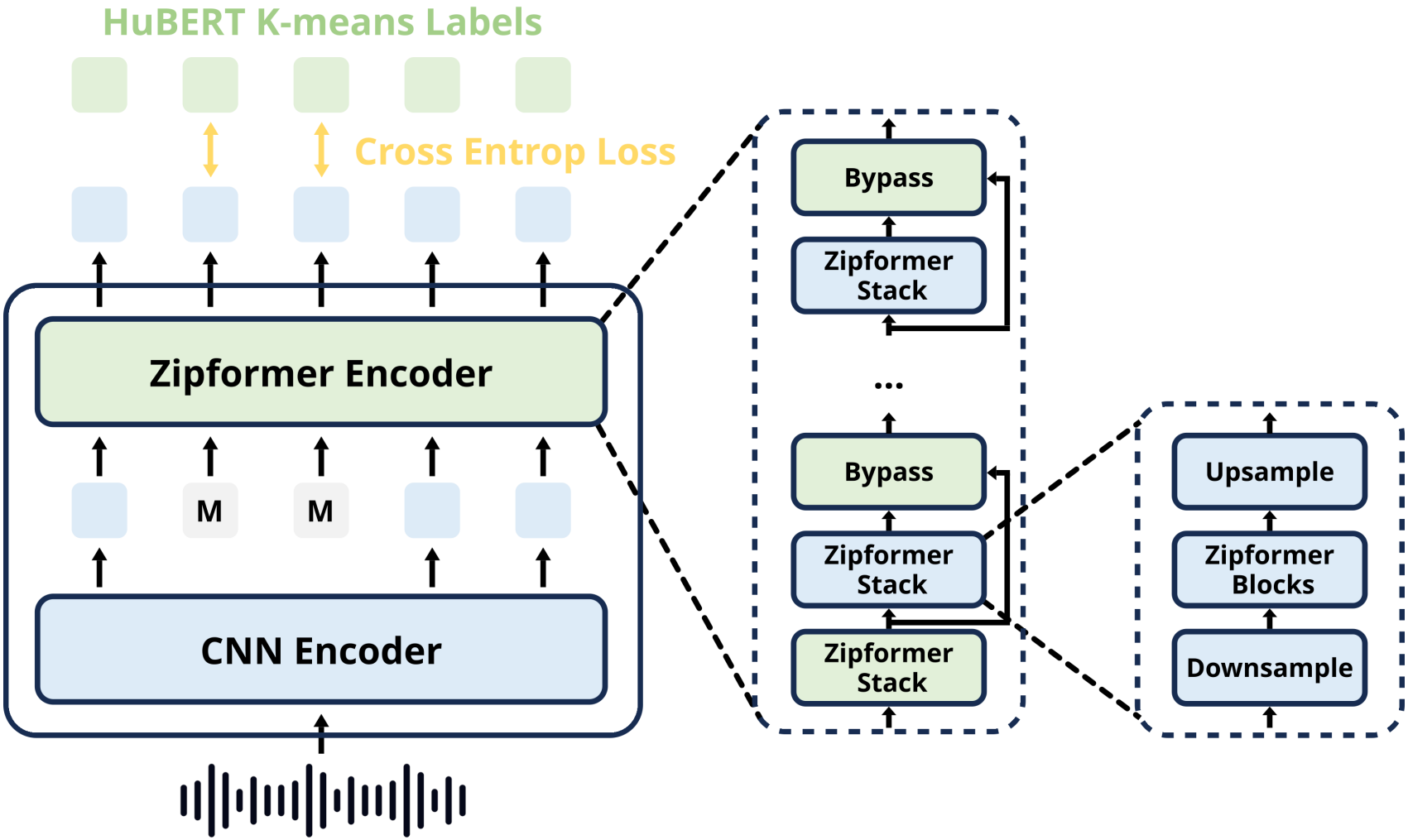
- Memory-lean HuBERT-style Architecture
- Memory-efficient Pre-training Loss
- Effective Fine-tuning Loss
- Advanced Zipformer Backbone & ScaledAdam Optimizer

2. Scalable Framework

- Seamless Scalability of Management and Training
- Constant and Fast Start-up Time
- FP16 Stability
- Compatibility with Fairseq Checkpoints

k2SSL Overview

The flagship model of the k2SSL framework is **Zipformer**. To ensure fair comparisons against HuBERT and WavLM, our experiments use identical parameters, data, targets, and decoding methods.



Results: Superior Efficiency and Performance with Zipformer Base (95M)

Models pre-trained on LibriSpeech-960h, fine-tuned on LibriSpeech-100h using BPE-level pruned RNN-T loss (top) / letter-level CTC loss (bottom)

Model	**Word Error Rate (%) ↓**				**GPU Hours ↓**	**Pre-train Speedup**
	dev-clean	dev-other	test-clean	test-other		
HuBERT Base (fairseq)	4.88	12.06	4.98	11.65	1878	1x
Zipformer Base (k2SSL)	**3.67**	**7.86**	**3.80**	**7.87**	**531**	**3.53x**
Relative WER Reduction	**-24.8%**	**-34.8%**	**-23.7%**	**-32.4%**		

Model	**LM**	**Word Error Rate (%) ↓**			
		dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 Base	None	6.1	13.5	6.1	13.3
HuBERT Base	None	5.3	13.0	5.4	12.6
WavLM Base	None	-	-	5.7	12.0
Zipformer Base	None	**4.7**	**9.6**	**4.4**	**9.8**
wav2vec 2.0 Base	4-gram	2.7	7.9	3.4	8.0
HuBERT Base	4-gram	2.7	7.8	3.4	8.1
WavLM Base	4-gram	-	-	3.4	7.7
Zipformer Base	4-gram	**2.6**	**6.4**	**3.0**	**6.8**

Results: Efficient Scaling with Zipformer Large (306M)

Models pre-trained on 60kh of Libri-Light, fine-tuned on LibriSpeech-960h using BPE-level transducer / pruned RNN-T loss.

Model	Unlabeled Data	Pre-train Steps	Word Error Rate (%) test-clean	test-other
Supervised				
Transformer Transducer	-	-	2.4	5.6
Conformer-L Transducer	-	-	2.1	4.3
Conformer-L Pruned Transducer	-	-	2.5	5.6
Zipformer-L Pruned Transducer	-	-	2.1	4.6
Pre-trained				
Conformer-L	LL-60k	400k	2.0	4.5
HuBERT Large	LL-60k	400k	**1.8**	**3.9**
Zipformer Large	LL-60k	**250k**	**1.8**	4.0

Key Takeaway

Zipformer Large achieves performance comparable to HuBERT Large while requiring only **5/8 of the pre-training steps**. This highlights the remarkable efficiency and scalability.

Follow-up Work: VietASR

Zipformer (68M), FBank frontend, on **50h labeled**, 70kh unlabeled spontaneous Vietnamese speech.

System	# Params (M)	Giga-Speech 2	Common Voice	FLEURS	Avg
Whisper large-v3	1542	17.94	13.74	**8.59**	16.44
Whisper base	72	39.88	44.07	40.41	40.16
MMS L1107	964	46.62	43.88	55.35	47.67
GigaSpeech 2	68	14.72	18.81	13.50	14.75
GigaSpeech 2	152	12.83	14.43	11.59	12.74
Google USM	-	13.28	12.46	11.75	13.03
Azure Speech CLI 1.37.9	-	11.86	**10.21**	11.88	11.78
Zipformer (from scratch)	68	19.7	27.02	23.18	20.54
VietASR Iteration 1	68	9.60	14.75	12.74	10.28
VietASR Iteration 2	68	8.01	12.21	11.40	8.68
VietASR Iteration 3	68	**7.68**	11.46	10.96	**8.31**

Key Takeaway

Through iterative semi-supervised training, 68M Zipformer outperforms Whisper Large-v3 and commercial ASR APIs. VietASR extends the k2SSL framework, demonstrating its practical solution to in-the-wild data.

Conclusion

- We introduced **k2SSL**, a scalable framework for more efficient and effective self-supervised speech representation learning.
- k2SSL drastically reduces memory usage and training time. Our Zipformer-based systems significantly outperform HuBERT and WavLM, demonstrating a 3.5x pre-training speedup and substantial WER reductions up to 35%.
- k2SSL makes high-performance SSL more accessible to the broader research community. VietASR further provides a cost-effective and practical solution to in-the-wild data.

Open Source

The codes, configurations, logs, and pre-trained checkpoints are publicly available to facilitate further research:

`https://github.com/k2-fsa/icefall`

Thank You

If you have any questions, feel free to contact me.

Email: yifanyeung@sjtu.edu.cn

浮槎

北京 海淀

扫一扫上面的二维码图案，加我为朋友。