Compact Conformer Models for On-Device Speech Recognition: A Reproducibility Study

Sample Author 1, Sample Author 2, Sample Author 3
Brainiall Research, San Francisco, USA

Abstract

We revisit the use of small-footprint Conformer-CTC models for on-device automatic speech recognition. Through a controlled study on the LibriSpeech and TED-LIUM benchmarks, we show that a 13-million-parameter Conformer-CTC encoder, post-training quantized to INT8, delivers a word error rate within 1.4 absolute points of a 600M Parakeet model while running 8.4 times faster on a single CPU core. We further demonstrate that, when paired with a small phone-projection head, the same backbone supports phoneme-level pronunciation assessment with Pearson correlation 0.59 on speechocean762, exceeding inter-annotator agreement.

1. Introduction

End-to-end ASR systems have converged on Conformer-style architectures, which combine self-attention and depth-wise convolutions to capture both global and local acoustic structure. While much recent work focuses on scaling encoders to billions of parameters, compact models remain critical for offline, low-latency, and privacy-sensitive deployments. In this paper we report a reproducibility study of the Conformer-CTC Small variant under post-training INT8 quantization, and quantify the gap to large frontier models on three standard tasks.

2. Equation Layout

The CTC posterior at frame t for token y_k is:

$$p(y_k \mid x_t) = \mathrm{softmax}(W \cdot h_t + b)_k$$

where h_t is the encoder hidden state at frame t and W is the projection matrix. The forward-backward algorithm computes path posteriors in O(T·V) time with V being the vocabulary size.

3. Experimental Setup

Model	Params	Stride	WER (clean)	RTF (CPU)
Conformer-CTC Small (INT8)	13.2M	4	5.8%	0.027
Conformer-CTC Medium (FP32)	30.8M	4	4.9%	0.061
Parakeet-CTC 0.6B	600M	8	4.4%	0.231

4. Results and Discussion

Table 1 summarizes the trade-off between accuracy and on-device latency. The 13M Conformer Small model achieves 5.8 percent WER on LibriSpeech test-clean while sustaining a real-time factor of 0.027 on a single Skylake CPU core. The 600M Parakeet baseline narrows the gap to 4.4 percent but at 8.6 times the compute cost, making it unsuitable for batch deployments on shared infrastructure. We further evaluated phoneme-level scoring on speechocean762: a 9-model ensemble built on top of the Conformer Small encoder reaches a Pearson correlation of 0.59 with median annotator scores, just 0.02 below the published 0.61 SOTA, and 0.035 above the inter-annotator agreement of 0.555.

5. Conclusion

Compact Conformer-CTC encoders, when quantized and paired with task-specific scoring heads, remain a strong default for production speech recognition and assessment. We release the evaluation harness and pretrained weights at github.com/brainiall/research-asr.

References

[1] Gulati, A. et al. Conformer: Convolution-augmented Transformer for Speech Recognition. INTERSPEECH 2020.

[2] Graves, A. et al. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML 2006.

[3] Panayotov, V. et al. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. ICASSP 2015.

[4] Hu, W. et al. Improvements to Speaker Diarization with Pyannote 3.1. arXiv 2024.

[5] Han, K. J. et al. Multistream CNN for Robust Acoustic Modeling. INTERSPEECH 2021.

[6] Chen, X. et al. SpeechOcean762: An Open-Source Non-Native English Pronunciation Dataset. INTERSPEECH 2021.

[7] Reddy, C. K. A. et al. The INTERSPEECH 2020 Deep Noise Suppression Challenge. INTERSPEECH 2020.

[8] Defossez, A. et al. Hybrid Spectrogram and Waveform Source Separation. ISMIR 2021.