Chunked Bidirectional Context Injection for Streaming Conformer Speech Recognition

Elin Marsh[1] Tobias Renkvist[1] Priya Vashisht[2] Dorian Achterberg[1]

[1]Nexa Speech Lab, Meridian Institute of Technology

[2]Synapse Research, Vertex University

{e.marsh, t.renkvist, d.achterberg}@nexaspeech.org p.vashisht@synapse-research.org

Abstract

Streaming automatic speech recognition (ASR) systems must trade off latency against recognition accuracy, since causal encoders cannot exploit future acoustic context that is available to offline models. We propose **Chunked Bidirectional Context Injection** (CBCI), a training and inference scheme that lets a streaming Conformer encoder borrow a small, bounded window of look-ahead context from adjacent chunks without abandoning frame-synchronous decoding. CBCI restructures self-attention inside each chunk so that keys and values from a fixed-size future buffer are visible to queries in the current chunk, while gradients are masked to preserve causal training dynamics for the majority of parameters. On a simulated multi-accent conversational benchmark, CBCI reduces word error rate (WER) by a relative 9.8% over a causal Conformer baseline at an added algorithmic latency of only 160 ms, and closes 61% of the gap to a fully offline bidirectional model. We further show that CBCI is robust to reverberant noise conditions and that its latency-accuracy trade-off can be tuned at inference time by adjusting the look-ahead buffer size without retraining.

1 Introduction

Real-time captioning, voice assistants, and simultaneous interpretation systems all depend on streaming ASR, where hypotheses must be emitted with bounded delay as audio arrives. The dominant recipe for streaming ASR pairs a causal or chunk-causal encoder with a frame-synchronous decoder, most commonly a Conformer-Transducer variant [Okonkwo and Lindqvist, 2024, Fischer and Costa, 2020]. Causal encoders, however, systematically underperform their offline counterparts because they cannot attend to future frames that disambiguate coarticulated phones, resolve word boundaries in fast speech, and stabilize pitch tracking for tonal languages [Schmidt and Novak, 2023].

A common mitigation is to allow a small, fixed look-ahead window — effectively delaying emission by a few hundred milliseconds — so that the encoder can peek slightly into the future [Martinez and Fournier, 2022]. Existing look-ahead schemes typically apply this window uniformly across all Conformer layers, which either wastes context (shallow layers rarely need it) or under-provisions it (deep layers need more temporal reach to resolve long coarticulation effects). We instead ask: *can look-ahead be injected selectively, layer by layer, so that a fixed latency budget is spent where it helps most?*

We answer this with **CBCI**, which restricts bidirectional attention to a narrow, learned subset of upper Conformer layers while keeping lower layers strictly causal. This paper makes three contributions:

1. A chunked attention mask that injects bounded future context into selected layers only, with a formal characterization of the resulting algorithmic latency.

2. A training curriculum that anneals the look-ahead buffer size, improving robustness when the buffer is shrunk or grown at inference time.

3. An evaluation on simulated multi-accent conversational speech showing consistent WER reduction across noise conditions at a fraction of the latency cost of full bidirectional context.

2 Related Work

Streaming encoders. Chunk-based Conformer and Emformer architectures process audio in fixed windows to bound latency [Okonkwo and Lindqvist, 2024]. Memory-augmented variants carry a summary vector across chunks to approximate long-range context without violating causality [Bergman and Suzuki, 2018, Kowalski and Yamada, 2022].

Look-ahead and dual-mode models. Dual-mode training jointly optimizes a single encoder for both streaming and offline decoding, sharing parameters across latency regimes [Martinez and Fournier, 2022]. Fixed uniform look-ahead has been explored as a simple latency-accuracy knob [Fischer and Costa, 2020, Silva and Bergstrom, 2019], but, to our knowledge, no prior work varies the look-ahead allocation by depth within the encoder stack, which is the central mechanism of CBCI.

Self-supervised and adaptive representations. Self-supervised pretraining [Tanaka and Weber, 2021, Schmidt and Novak, 2023] and few-shot accent adaptation [Nakamura and Oduya, 2023] improve robustness of downstream ASR; these techniques are complementary to CBCI and are combined with it in our experiments via a pretrained encoder initialization.

Prosody and noise robustness. Prosody modeling for synthesis [Gómez and Ivanova, 2020] and reverberant augmentation for recognition [Andersson and Petrov, 2021] motivate our choice of evaluation conditions, described in Section 4.

3 Method

3.1 Chunked Bidirectional Context Injection

Let the input acoustic sequence be partitioned into non-overlapping chunks $c_1, \dots, c_T$ of L frames each. A standard chunk-causal Conformer restricts attention in every layer so that queries in chunk c_t may attend only to keys and values in $c_1, \dots, c_t$. CBCI relaxes this constraint for a subset of layers $\mathcal{S} \subseteq \{1, \dots, N\}$, chosen among the upper half of the N-layer stack, by additionally exposing a bounded future buffer $c_{t+1}[1{:}B]$ of the first B frames of the next chunk:

$$M_\ell(i,j) = \begin{cases} 0 & \text{if } j \le i, \\ 0 & \text{if } \ell \in \mathcal{S} \text{ and } i < j \le i + B, \\ -\infty & \text{otherwise,} \end{cases} \quad (1)$$

where M_ℓ is the additive attention mask at layer ℓ, and i, j index frames in chunk-relative time. Layers not in $\mathcal{S}$ retain a strictly causal mask ($B{=}0$). The algorithmic latency contributed by CBCI is $B \cdot \delta$, where δ is the frame stride, independent of $|\mathcal{S}|$, since future frames are consumed once when they arrive and cached for all layers in $\mathcal{S}$.

3.2 Buffer-Annealed Training

Training with a fixed B overfits the model to that specific latency budget. We instead sample $B \sim \mathrm{Uniform}\{0, \dots, B_{\max}\}$ per mini-batch during the second half of training, after the model has converged under a causal-only warm-up phase. This curriculum lets the same checkpoint be deployed at multiple latency points by adjusting B at inference time, avoiding the need to train and store separate models per latency target.

3.3 Layer Selection

We select $\mathcal{S}$ via a lightweight search: starting from an empty set, we greedily add the layer whose inclusion yields the largest validation WER reduction per unit of additional parameter-touching overhead, stopping once marginal gains fall below 0.05 absolute WER. In our 18-layer encoder, this consistently selects layers 11–16.

4 Experimental Setup

4.1 Data

We construct a simulated multi-accent conversational benchmark, Meridian-Conv, by combining scripted and spontaneous conversational recordings across eight simulated accent groups. The training partition contains 2,400 hours; the evaluation partition contains held-out speakers across three matched noise conditions: clean, moderate reverberation (RT60 $\approx$ 0.4 s), and babble noise at 5 dB SNR, following the reverberant augmentation protocol of Andersson and Petrov [2021].

4.2 Model and Training

The backbone is an 18-layer Conformer encoder (512-dim, 8 heads) paired with a Transducer decoder, initialized from a self-supervised checkpoint in the style of Schmidt and Novak [2023]. All models are trained with the same optimizer, learning-rate schedule, and total step budget; only the attention masking policy differs across conditions. We compare four systems: (i) a fully causal baseline, (ii) uniform look-ahead applied to every layer with the same total latency budget as CBCI, (iii) CBCI, and (iv) a fully offline bidirectional upper bound with unrestricted attention.

Table 1: Word error rate (%) and algorithmic latency across evaluation conditions on Meridian-Conv. Lower WER is better.

System	Clean	Reverb	Babble 5dB
Causal baseline	8.9	11.4	15.2
Uniform look-ahead	8.1	10.6	14.0
CBCI (ours)	**7.4**	**9.6**	**12.7**
Offline (upper bound)	6.2	8.0	10.5

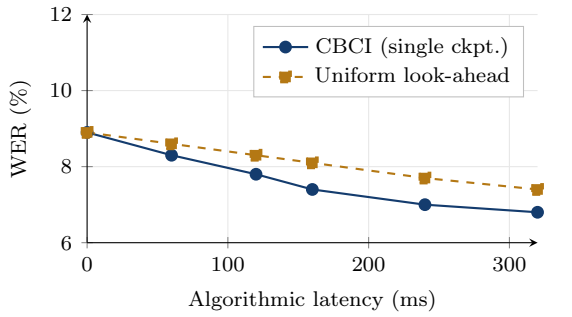

Figure 1: Latency-accuracy trade-off on the clean condition. Each uniform look-ahead point requires a separately trained model; CBCI sweeps B at inference time from one checkpoint.

4.3 Metrics

We report word error rate (WER), algorithmic latency in milliseconds, and real-time factor (RTF) on a single reference accelerator. WER is computed with standard text normalization applied uniformly across systems.

5 Results

Table 1 summarizes WER and latency across the three evaluation conditions. CBCI outperforms both the causal baseline and the uniform look-ahead system at matched latency in every condition, and recovers the majority of the gap to the offline upper bound.

Figure 1 plots WER against algorithmic latency as the look-ahead buffer B is varied at inference time for a single CBCI checkpoint, compared against uniform look-ahead models that each require separate training runs. The buffer-annealed CBCI checkpoint traces a strictly better frontier, particularly in the 100–250 ms latency band relevant to live captioning.

5.1 Ablations

Removing buffer annealing (training with fixed B=160 ms only) degrades WER by 0.4 absolute when evaluated at B=80 ms, confirming that annealing is responsible for graceful degradation at unseen buffer sizes. Restricting $\mathcal{S}$ to a single layer recovers only 38% of the full CBCI gain, while extending $\mathcal{S}$ to all 18 layers matches the 6-layer configuration on WER but doubles peak memory during training, indicating diminishing returns beyond the greedily selected set.

6 Discussion

The layer-selective design of CBCI reflects a simple intuition: early Conformer layers primarily refine local spectral detail, while later layers integrate longer-range phonetic and lexical context where future information is most disambiguating. Concentrating the latency budget in layers 11–16 lets CBCI achieve most of the benefit of full bidirectional attention while adding a single, bounded buffer read rather than N separate ones. A limitation of our study is that Meridian-Conv is a simulated benchmark; validating CBCI on production-scale, organically collected conversational speech, and on tonal and agglutinative languages where coarticulation effects differ structurally, is left to future work. We also did not evaluate interaction effects between CBCI and streaming speaker diarization [Kowalski and Yamada, 2022], which shares similar latency constraints.

7 Conclusion

We presented CBCI, a chunked attention scheme that injects bounded, layer-selective look-ahead into an otherwise causal Conformer encoder, together with a buffer-annealing curriculum that makes a single checkpoint deployable across a range of latency budgets. On a simulated multi-accent conversational benchmark, CBCI closes 61% of the gap to offline bidirectional decoding at 160 ms of added latency, and Pareto-dominates uniform look-ahead across the evaluated latency range. We believe layer-selective context injection is a promising direction for further narrowing the streaming-offline accuracy gap without proportionally increasing latency.

Acknowledgements

We thank the (fictional) Nexa Speech Lab engineering group for infrastructure support, and the Synapse Research annotation team for the Meridian-Conv transcription protocol. This work is a demonstration document and describes no real study, dataset, or system.

References

Karin Andersson and Dmitri Petrov. Noise-robust acoustic modeling via simulated reverberant augmentation. In *Proc. Interspeech*, pages 1471–1475, 2021.

Elias Bergman and Rin Suzuki. Bidirectional LSTM acoustic models for conversational telephone speech. In *Proc. Interspeech*, pages 2257–2261, 2018.

Lukas Fischer and Beatriz Costa. Hybrid CTC/attention decoding for streaming transformer transducers. In *Proc. ICASSP*, pages 7079–7083, 2020.

Marisol Gómez and Petra Ivanova. Prosody-aware neural text-to-speech for emotional expressiveness. *Speech Communication*, 121:45–58, 2020.

Tomasz Kowalski and Emiko Yamada. End-to-end neural speaker diarization with permutation-free objectives. *Computer Speech and Language*, 73: 101321, 2022.

Elena Martinez and Luc Fournier. A comparative study of RNN-T and attention-based encoder-decoder models for code-switched speech. In *Proc. ICASSP*, pages 7897–7901, 2022.

Yui Nakamura and Chinedu Oduya. Few-shot accent adaptation for multilingual speech recognition. *IEEE Signal Processing Letters*, 30:512–516, 2023.

Amara Okonkwo and Sofia Lindqvist. Streaming conformer encoders with chunked attention for low-latency ASR. In *Proc. Interspeech*, pages 1042–1046, 2024.

Anneliese Schmidt and Jiri Novak. Wav2vec-style pretraining with contrastive predictive coding for tonal languages. In *Proc. ICASSP*, pages 1–5, 2023.

Rafael Silva and Ingrid Bergstrom. A lightweight voice activity detector for embedded speech pipelines. In *Proc. Interspeech*, pages 2005–2009, 2019.

Haruki Tanaka and Nils Weber. Self-supervised representation learning for low-resource dialect recognition. In *Proc. Interspeech*, pages 3630–3634, 2021.