Context-Driven Dynamic Pruning for Large Speech Foundation Models

Masao Someki[1], Shikhar Bharadwaj[1], Atharva Anand Joshi[1], Chyi-Jiunn Lin[1], Jinchuan Tian[1], Jee-weon Jung[1]
Markus Müeller[2], Nathan Susanj[2], Jing Liu[2], Shinji Watanabe [2]

[1] Language Technologies Institute, Carnegie Mellon University
[2] Neural Efficiency Science, Amazon Artificial General Intelligence

Abstract

- We propose **local Gate Predictor (localGP)**, a layer-wise pruning module that dynamically selects active modules based on **frame-level context** such as speaker and acoustic event embeddings.
- We reduce **56.7 GFLOPs** on the encoder with **+26.1% BLEU improvement on average**, outperforming fully fine-tuned baselines.
- We empirically found a tendency where temporal pruning mimics **VAD-like patterns** in early encoder layers and shows **token-dependent decoder pruning**, revealing structured, context-sensitive computation.

LocalGP architecture

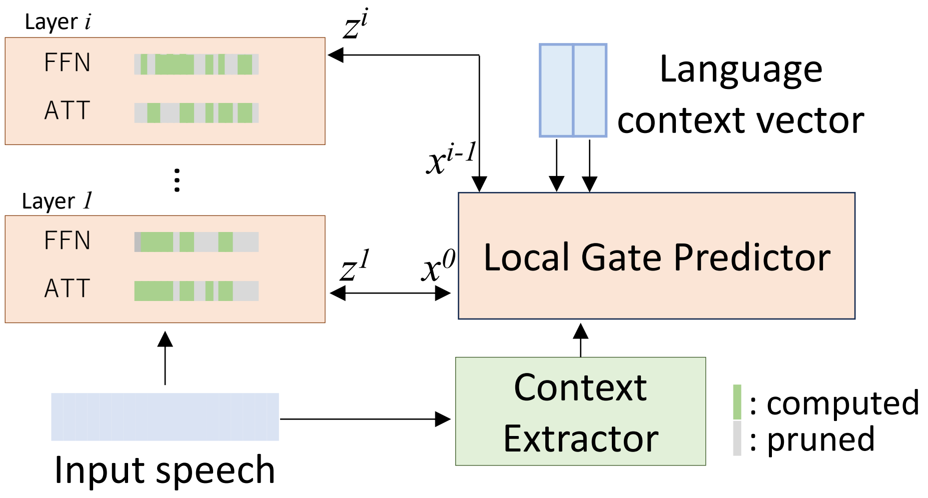


Figure 1. Overview of LocalGP and temporal pruning. At each layer i, LocalGP receives intermediate outputs x^i and computes frame-level probabilities z^i to select or skip modules. Output of layer i-1 guides its own gating via the Local Gate Predictor.

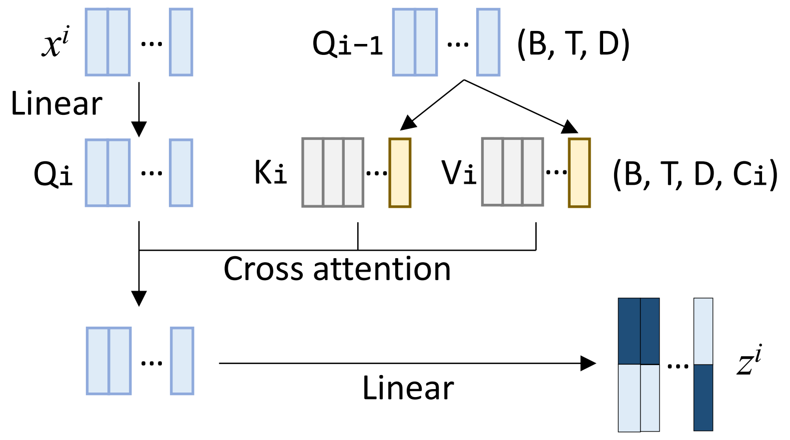


Figure 2. At each layer i, LocalGP computes cross-attention between the x^i and contextual features. The query is used to generate updated context for the next layer. B, T, D, and C denote batch size, frame length, hidden dimension, and the number of contexts, respectively.

Method

- **Local vs Global Decisions:**
  LocalGP makes pruning decisions **independently at each layer**, while globalGP in previous work[1] apply **a single shared mask** across all layers, ignoring layer-specific dynamics.
- **Temporal vs. utterance-wise pruning**
  Temporal pruning (ours) dynamically **skips individual frames**, while utterance-wise pruning removes **entire audio input**, leading to coarse, less flexible behavior.
- **Context-aware vs Fixed Masking:**
  LocalGP with temporal pruning leverages **pretrained context features** to choose efficient computation paths during inference.

Results on Speech Translation

- **LocalGP outperforms full fine-tuning in BLEU:**
  Even without additional context (No. 5), localGP achieves **higher BLEU scores** (12.0 vs 10.7).
  With additional contexts (No. 6–7), BLEU improves to **13.5 and 12.8**.
- GFLOPs drop by **56.7 (spk)** and **58.4 (event)** compared to No.1, showing that localGP achieves **both efficiency and better translation quality**.

No.	Context	de-fr	de-it	fr-de	fr-it	it-de	it-fr	Avg	GFLOPs
1	full fine-tuning (baseline)	8.4	6.4	11.2	13.0	11.8	13.5	10.7	568.5
3	front (baseline)	9.8	8.4	12.2	13.1	11.3	13.4	11.4	–
5	front	10.4	7.8	13.0	14.6	11.2	15.3	12.0	541.6
6	+ spk	**12.0**	**9.2**	**14.4**	**15.6**	**12.7**	**16.9**	**13.5**	511.8
7	+ event	11.4	8.2	13.5	15.1	12.0	16.3	12.8	510.1
8	+ spk + event	10.6	8.1	13.6	15.0	12.0	16.3	12.6	497.0

Table 1. BLEU scores for German (de), French (fr), and Italian (it) speech translation using full fine-tuning (blue), globalGP with utterance pruning (orange), and localGP with temporal pruning (green). "Front" denotes subsampled speech features.

Analysis

Encoder-side

- **VAD-like patterns emerge**:
  The first layer remain active across almost all frames, while deeper layers prune silence more aggressively.
- When speaker or event embeddings are used, the model allocates computation more precisely to speech segments, reducing redundancy in silence regions.

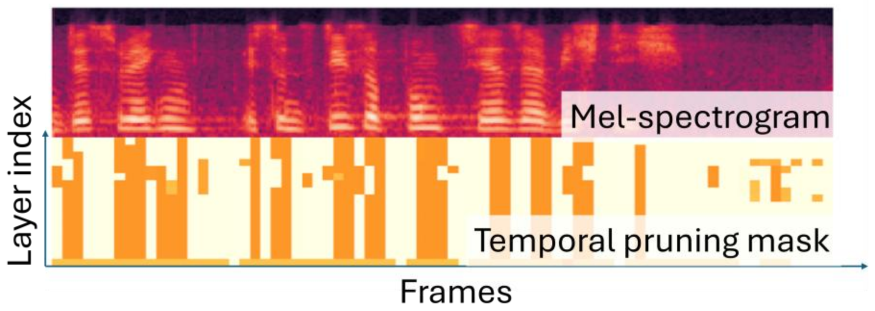
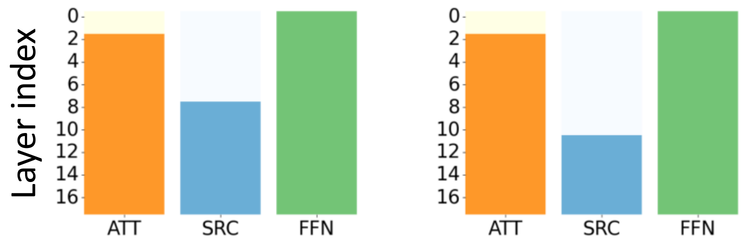


Figure 3. Log-Mel spectrogram (top) and temporal pruning mask for self-attention (bottom). The y-axis shows encoder layers, and the x-axis represents time. Orange regions indicate frames where computation is retained.

Decoder-side

- **Pruning varies by token type:**
  Tokens beginning with a space (e.g., [Space]wollen) activate more source-attention modules, indicating greater audio context is needed at word starts.



(a) *Token:* `[Space]wollen` (b) *Token:* `struktur`

Figure 4. Pruning pattern for tokens `[Space]wollen` and `struktur`. The ATT, SRC, and FFN represent self-attention, source-attention, and the feed-forward network, respectively. The y-axis indicates the layers, with the top representing the first layer. Colored modules indicate activated modules.

Reference

[1] Masao Someki, Yifan Peng, Siddhant Arora, Markus Muller, Athanasios Mouchtaris, Grant Strimel, Jing Liu, & Shinji Watanabe (2025). Context-aware Dynamic Pruning for Speech Foundation Models. In *The Thirteenth International Conference on Learning Representations*.