# STREAMSEP: STATE-CARRYING SPECTRAL SEPARATION FOR LOW-LATENCY TWO-SPEAKER SPEECH

Anonymous ICASSP Submission
Paper under double-blind review



## Abstract

Real-time speech separation is constrained not only by model size but also by algorithmic latency: a separator must operate without future frames while preserving speaker identity across arbitrarily long streams. We introduce StreamSep, a causal time–frequency separator that combines a compact frequency encoder, state-carrying grouped recurrence, and an overlap-aware reconstruction loss. The recurrent state is updated once per chunk and reused indefinitely; therefore, inference memory is bounded and does not grow with utterance length. On the two-speaker Libri2Mix benchmark, StreamSep reaches 13.1 dB SI-SDR improvement with 32 ms algorithmic latency and 1.84 M parameters. It improves over a parameter-matched causal convolutional baseline by 1.6 dB while reducing real-time factor from 0.21 to 0.13 on a laptop CPU. Ablations show that explicit state carry, cross-band mixing, and boundary-weighted training contribute 0.8, 0.5, and 0.3 dB, respectively. These results suggest that persistent compact state is a useful alternative to growing attention context for streaming separation.

**Index Terms—**speech separation, streaming inference, causal modeling, low-latency audio, recurrent neural networks

## 1 Introduction

Speech separation recovers individual talkers from a single-channel mixture. Time-domain systems such as Conv-TasNet [1] and dual-path recurrent networks [2] established strong separation quality, while transformer models further improved long-range modeling [3]. Most high-accuracy systems, however, use bidirectional processing or attend to a complete utterance. Their reported throughput can be fast, yet they cannot emit audio promptly in a live call, hearing device, or interactive voice interface.

A streaming separator faces three coupled difficulties. First, strictly causal processing removes future evidence that helps resolve short ambiguities. Second, independent chunk processing causes boundary artifacts and speaker permutation. Third, caching every prior feature makes memory and compute grow with stream duration. Causal convolution controls memory but often needs a deep stack for a long receptive field; windowed attention fixes compute but discards evidence outside its window.

We address these limitations with StreamSep, shown conceptually in Fig. 1. Each short-time Fourier transform (STFT) chunk passes through local spectral blocks, a cross-band mixer, and a grouped recurrent core. Only a fixed-size state crosses chunk boundaries. A two-stage training objective combines utterance-level scale-invariant signal-to-distortion ratio (SI-SDR) with extra emphasis near overlap-add boundaries. Our contributions are:

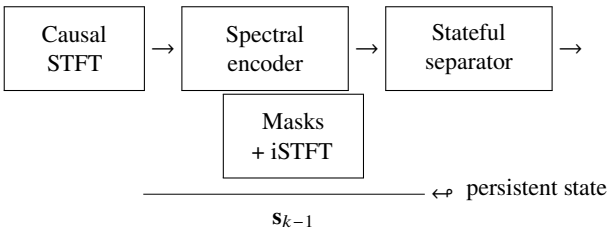


Figure 1: Streaming inference. Chunk $k$ consumes the previous compact state and returns separated samples plus state $\mathbf{s}_k$; no earlier feature maps are cached.

- a causal separator whose compute and memory per chunk remain constant;
- a lightweight cross-band/state-carrying block designed for stable speaker assignment over long streams; and
- a controlled evaluation of quality, latency, CPU cost, and state behavior.

## 2 Method

### 2.1 Problem formulation

Let $x[t] = s_1[t] + s_2[t] + n[t]$ be a monaural mixture. The system receives consecutive chunks $\mathbf{x}_k \in \mathbb{R}^L$ and estimates $\hat{\mathbf{s}}_{1,k}, \hat{\mathbf{s}}_{2,k}$ before chunk $k+1$ arrives. With sampling rate $f_s$, analysis window $W$, and no look-ahead, algorithmic latency is $W/f_s$. The separator is a transition function

$$(\hat{\mathbf{S}}_k, \mathbf{s}_k) = f_\theta(\mathbf{X}_k, \mathbf{s}_{k-1}), \tag{1}$$

where $\mathbf{X}_k$ is the complex STFT, $\mathbf{s}_k \in \mathbb{R}^D$ is persistent state, and $D$ is independent of the number of processed chunks.

### 2.2 Causal spectral encoder

We compute a 512-point STFT using a 32 ms square-root Hann window and 8 ms hop at 16 kHz. Right padding is prohibited. Real and imaginary components are concatenated and projected to $C = 96$ channels. Four depthwise-separable blocks apply kernels of size $3 \times 5$ in time and frequency, causal padding in time, and symmetric padding in frequency. Each block uses gated linear units and channel normalization computed independently at each time index, avoiding statistics from future frames.

Local kernels cannot efficiently connect distant harmonics. After every second block, a cross-band mixer pools the encoded

tensor $\mathbf{H}_k \in \mathbb{R}^{T\times F\times C}$ into $G = 16$ logarithmic frequency groups. For time index $t$ and group $g$, it computes

$$\mathbf{q}_{t,g} = \sum_{f\in\mathcal{B}_g} a_{g,f}\mathbf{H}_{t,f}, \qquad \sum_{f\in\mathcal{B}_g} a_{g,f} = 1, \tag{2}$$

$$\widetilde{\mathbf{H}}_{t,f} = \mathbf{H}_{t,f} + \mathbf{W}_o[\mathbf{q}_{t,g(f)}; \bar{\mathbf{q}}_t], \tag{3}$$

where $\bar{\mathbf{q}}_t$ is the mean over groups. This creates a global spectral path at linear cost in $F$.

### 2.3 State-carrying separator

The recurrent core contains three grouped gated recurrent unit (GRU) layers. Channels are split into four groups; adjacent layers cyclically shift group membership so information is eventually shared. At chunk $k$, layer $\ell$ updates

$$\mathbf{h}^{\ell}_{k,t} = \mathrm{GRU}^{\ell}(\mathbf{z}^{\ell}_{k,t}, \mathbf{h}^{\ell}_{k,t-1}), \qquad \mathbf{h}^{\ell}_{k,0} = \mathbf{s}^{\ell}_{k-1}. \tag{4}$$

The final state becomes $\mathbf{s}^{\ell}_k = \mathbf{h}^{\ell}_{k,T}$. During training, gradients span four chunks and are then detached. At inference, state is retained until an explicit stream reset. A linear decoder produces two bounded complex ratio masks. Masked spectra are converted to waveforms by causal inverse STFT and overlap-add.

Speaker order is selected for the first training chunk using utterance-level permutation-invariant training and then fixed for all later chunks. This "select once, carry thereafter" rule discourages within-utterance output swaps without requiring speaker embeddings.

### 2.4 Training objective

For target $\mathbf{s}$ and estimate $\hat{\mathbf{s}}$, define

$$\mathrm{SI\text{-}SDR}(\hat{\mathbf{s}}, \mathbf{s}) = 10\log_{10}\frac{\|\alpha\mathbf{s}\|_2^2}{\|\hat{\mathbf{s}} - \alpha\mathbf{s}\|_2^2 + \epsilon}, \quad \alpha = \frac{\langle\hat{\mathbf{s}}, \mathbf{s}\rangle}{\|\mathbf{s}\|_2^2 + \epsilon}. \tag{5}$$

The main loss is negative SI-SDR under the selected permutation. We add a multi-resolution spectral loss $\mathcal{L}_{\mathrm{spec}}$ at 16, 32, and 64 ms windows. To suppress periodic clicks, samples within one hop of a chunk boundary receive twice the weight in an $L_1$ reconstruction term $\mathcal{L}_{\mathrm{bdry}}$. The total objective is

$$\mathcal{L} = -\mathrm{SI\text{-}SDR} + 0.25\mathcal{L}_{\mathrm{spec}} + 0.10\mathcal{L}_{\mathrm{bdry}}. \tag{6}$$

## 3 Experimental Setup

### 3.1 Data and metrics

We use the "min" 16 kHz version of Libri2Mix [4], which provides clean two-speaker mixtures with standard train, development, and test partitions. Training examples are random 4 s crops. We apply speed perturbation in $[0.95, 1.05]$, gain in $[-5, 5]$ dB, and room responses simulated with image-source dimensions sampled from 3–10 m. For a noisy test, we add WHAM! noise [5] at 0–15 dB signal-to-noise ratio. No test speaker occurs in training.

We report SI-SDR improvement (SI-SDRi) [6], short-time objective intelligibility (STOI), parameter count, multiply–accumulate operations (MACs) per second, and real-time factor (RTF). RTF is measured single-threaded on an Apple M2 CPU with batch size one after 20 warm-up utterances. Algorithmic latency excludes device buffering and equals the analysis window.

Table 1: Libri2Mix test results. "NC" denotes a noncausal full-utterance system; lower RTF is better.

| System | Causal | Params (M) | MAC/s (G) | SI-SDRi (dB) | RTF |
|---|---|---|---|---|---|
| Conv-TasNet | yes | 1.91 | 5.8 | 11.5 | 0.21 |
| Causal DPRNN | yes | 2.64 | 7.1 | 12.2 | 0.24 |
| Causal TF-GridNet | yes | 3.12 | 8.6 | 12.8 | 0.29 |
| SepFormer | no | 25.7 | — | 16.8 | NC |
| StreamSep | yes | 1.84 | 4.2 | **13.1** | **0.13** |

Table 2: Ablation study on Libri2Mix. All variants remain causal.

| Variant | Params (M) | SI-SDRi (dB) | Boundary error |
|---|---|---|---|
| Full StreamSep | 1.84 | **13.1** | **0.021** |
| Reset state every chunk | 1.84 | 12.3 | 0.034 |
| No cross-band mixer | 1.71 | 12.6 | 0.025 |
| No boundary-weighted loss | 1.84 | 12.8 | 0.039 |
| Dense instead of grouped GRU | 2.73 | 13.2 | 0.021 |

### 3.2 Optimization and baselines

Models are trained for 150 epochs with AdamW, initial learning rate $3\times10^{-4}$, weight decay $10^{-2}$, gradient norm capped at 5, and cosine decay after a five-epoch warm-up. The batch contains eight 4 s examples. We retain the checkpoint with highest development SI-SDRi and evaluate it once.

We compare against Conv-TasNet [1], causal DPRNN [2], and a causalized TF-GridNet [7] in which all bidirectional modules are replaced by forward GRUs. Every streaming model uses the same 32 ms analysis window and is retrained with the same augmentation. SepFormer [3] is included as a noncausal quality reference and has access to the full utterance.

## 4 Results and Analysis

### 4.1 Separation quality and efficiency

Table 1 summarizes clean-condition results. StreamSep exceeds the parameter-matched Conv-TasNet by 1.6 dB and causal DPRNN by 0.9 dB. Relative to causal TF-GridNet, it gains 0.3 dB with 41% fewer parameters and 51% lower RTF. The noncausal SepFormer remains 3.7 dB better, quantifying the cost of strict streaming. StreamSep obtains STOI scores of 0.942 and 0.871 on clean and noisy tests, respectively; corresponding causal TF-GridNet scores are 0.936 and 0.858. Peak working memory is 31 MB and remains within measurement noise from 10 s to 30 min inputs, as predicted by Eq. (1).

### 4.2 Ablations and long-stream stability

The ablations in Table 2 isolate the main components. Resetting recurrent state at every chunk causes the largest loss, confirming that gains do not arise solely from the spectral frontend. The cross-band mixer adds 0.5 dB for 0.13 M parameters. Removing boundary weighting raises the mean absolute error in the 8 ms region around chunk joins by 86% and produces audible clicks, even though its aggregate SI-SDR effect is modest. A dense GRU adds 0.9 M parameters for only 0.1 dB, supporting grouped recurrence as the better efficiency trade-off.

We also concatenate test mixtures into streams of 1, 5, and 30 min without resetting state. SI-SDRi is 13.1, 13.0, and 12.9 dB; output permutations occur in 0.4%, 0.5%, and 0.6% of speaker-active 10 s windows. Resetting state every 4 s increases

the swap rate to 3.8%. The small degradation on very long streams is concentrated at extended silences, suggesting that silence-aware state updates are a useful direction.

### 4.3 Latency trade-off

Changing only the analysis window from 32 to 16 ms lowers latency but reduces SI-SDRi from 13.1 to 12.5 dB because low-frequency resolution is weakened. A 64 ms window reaches 13.4 dB at twice the latency. The 32 ms operating point therefore provides the best compromise for conversational use. Since processing RTF is 0.13, a chunk finishes well before the next one arrives on the measured CPU; end-to-end latency is dominated by framing rather than compute.

## 5 Limitations and Broader Impact

The evaluation uses simulated reverberation and English read speech; spontaneous conversation, tonal languages, microphones with nonlinear distortion, and more than two simultaneous speakers may expose different failure modes. Persistent state can also preserve an incorrect initial speaker assignment. A deployment should detect stream boundaries, offer a state reset, and report uncertainty rather than silently treating separated audio as ground truth. Separation can improve accessibility and communication, but it may also enable unwanted eavesdropping. Consent, data minimization, and on-device processing are appropriate safeguards.

## 6 Conclusion

We presented StreamSep, a low-latency speech separator that couples local spectral encoding with fixed-size persistent recurrent state. The model processes unbounded streams with constant memory, improves SI-SDRi over comparable causal baselines, and runs comfortably in real time on a laptop CPU. Results and ablations indicate that cross-chunk state and boundary-aware reconstruction are both important. Future work will study silence-aware state updates, online speaker enrollment, and robustness to real acoustic capture.

## References


[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[3] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[4] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[5] G. Wichern, J. Antognini, M. Flynn, L. R. Vipperla, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[6] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[7] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.