A LIGHTWEIGHT ATTENTION MODEL FOR SPEECH SEPARATION PROBLEM

Nguyen Thi Duyen*, Ha Thi Nguyet Anh
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 01/10/2025
Revised: 30/12/2025
Published: 31/12/2025

KEYWORDS

Speech separation
Cocktail party
Attention mechanism
Neural network
Lightweight model

ABSTRACT

This paper focuses on developing a lightweight yet efficient single-channel speech separation model that addresses the challenge of balancing the separation quality and computational complexity. We propose to exploit the SeliNet backbone combined with the CBAM (Convolutional Block Attention Module) attention mechanism, which allows the model to maintain high performance while minimizing the number of parameters and FLOPs. The research inherits recent advances in time-domain deep learning, especially depthwise separable convolutions, bottleneck blocks, and Atrous Temporal Pyramid Pooling to capture multi-scale context. By integrating CBAM, the model is able to emphasize important channels and temporal locations, which significantly improves SI-SNRi, SDRi, and other quality metrics. Experiments on the WSJ0-2mix dataset show that the model achieves competitive performance compared to heavier deep networks while significantly reducing FLOPs and the number of parameters. This result opens new approaches for real-time speech separation systems on mobile and edge devices.

MỘT MÔ HÌNH CHÚ Ý NHẸ CHO BÀI TOÁN TÁCH TIẾNG NÓI

Nguyễn Thị Duyên*, Hà Thị Nguyệt Ánh
Trường Đại học Công nghệ thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 01/10/2025
Ngày hoàn thiện: 30/12/2025
Ngày đăng: 31/12/2025

TỪ KHÓA

Tách tiếng nói
Bữa tiệc cocktail
Cơ chế chú ý
Mạng nơron
Mô hình nhẹ

TÓM TẮT

Bài báo này tập trung vào việc phát triển một mô hình tách tiếng nói kênh đơn nhẹ nhưng hiệu quả, nhằm giải quyết thách thức cân bằng giữa chất lượng tách và độ phức tạp tính toán. Chúng tôi đề xuất khai thác backbone SeliNet kết hợp với cơ chế chú ý CBAM (Convolutional Block Attention Module), cho phép mô hình vừa duy trì hiệu năng cao vừa giảm thiểu số lượng tham số và phép tính FLOPs. Nghiên cứu kế thừa các tiến bộ gần đây trong học sâu miền thời gian, đặc biệt là depthwise separable convolutions, bottleneck blocks và Atrous Temporal Pyramid Pooling để nắm bắt ngữ cảnh đa tỷ lệ. Bằng cách tích hợp CBAM, mô hình có khả năng nhấn mạnh các kênh và vị trí thời gian quan trọng, giúp cải thiện đáng kể SI-SNRi, SDRi và các chỉ số chất lượng khác. Các thí nghiệm trên bộ dữ liệu WSJ0-2mix cho thấy mô hình đạt hiệu năng cạnh tranh so với các mạng sâu nặng nề hơn trong khi giảm đáng kể FLOPs và số tham số. Kết quả này mở ra hướng tiếp cận mới cho các hệ thống tách tiếng nói thời gian thực trên thiết bị di động và biên.

DOI: https://doi.org/10.34238/tnu-jst.13724

* Corresponding author. *Email: duyennt@ictu.edu.vn*

1. Giới thiệu

Tách tiếng nói – bài toán tách các nguồn tín hiệu giọng nói riêng lẻ từ một hỗn hợp tín hiệu chồng lấn – đã thu hút sự quan tâm nghiên cứu mạnh mẽ trong thập kỷ qua. Con người có khả năng nổi bật trong việc tập trung vào một giọng nói duy nhất trong môi trường nhiều người nói – hiệu ứng "bữa tiệc cocktail" – nhưng việc xây dựng các mô hình tính toán đạt được hiệu quả tương đương vẫn là một thách thức. Khó khăn này xuất phát từ tính phi tĩnh cao của tín hiệu tiếng nói, sự chồng lấn phổ tần và yêu cầu xử lý độ trễ thấp trong các ứng dụng thực tế như hội nghị trực tuyến, thiết bị trợ thính hay điện thoại di động. Do đó, tách tiếng nói đã trở thành một thành phần quan trọng trong lĩnh vực tăng cường và nhận dạng tiếng nói, tác động trực tiếp tới các bài toán phía sau như nhận dạng tiếng nói tự động và phân tách người nói [1].

Với sự xuất hiện của học sâu, tách tiếng nói kênh đơn (SCSS) đã có những tiến bộ đáng kể. Các phương pháp che phổ thời gian–tần số (time–frequency masking) tiên phong đã bị vượt qua bởi các kiến trúc miền thời gian end-to-end. Chẳng hạn, Conv-TasNet [2] đã chứng minh rằng việc học tách trực tiếp trong miền thời gian có thể vượt qua che phổ cường độ TF lý tưởng. Các mô hình tiếp theo như Dual-Path RNN (DPRNN) [3], mô hình dựa trên Conformer [4] và SepFormer [5] đã cải thiện hơn nữa khả năng mô hình hóa chuỗi dài và cơ chế attention cho SCSS. Tuy nhiên, những mô hình tiên tiến này thường dựa trên các backbone sâu và nặng, chứa hàng chục triệu tham số, gây khó khăn cho việc triển khai trên thiết bị biên và hệ thống thời gian thực không có GPU chuyên dụng [6]. Hạn chế này đã thúc đẩy hướng nghiên cứu song song về các kiến trúc nhẹ, cân bằng giữa hiệu suất tách và hiệu quả tính toán [7]–[9].

Câu hỏi đặt ra là làm thế nào để thiết kế một mô hình vừa giữ được sức mạnh biểu diễn của các mạng sâu hiện đại vừa giảm đáng kể số tham số và phép tính. Các tích chập tách chiều, vốn được giới thiệu trong các mạng hình ảnh kiểu MobileNet, là một hướng đầy hứa hẹn. Một hướng khác là tích hợp các mô-đun attention nhẹ nhằm làm nổi bật các đặc trưng quan trọng mà không làm tăng nhiều chi phí tính toán. Convolutional Block Attention Module (CBAM) [10] chẳng hạn, thực hiện attention kênh và không gian tuần tự bằng các phép pooling đơn giản và MLP nhỏ. CBAM đã chứng minh hiệu quả trong các bài toán thị giác máy tính, vượt trội so với mô-đun Squeeze-and-Excitation (SE) trước đó trong khi chỉ tăng chưa tới 1% tham số. Tuy nhiên, tiềm năng của CBAM trong tách tiếng nói miền thời gian vẫn chưa được khai thác đầy đủ.

Nghiên cứu này giải quyết khoảng trống trên bằng cách đề xuất một kiến trúc nhẹ mới cho tách tiếng nói kênh đơn, kết hợp SeliNet (một backbone tách nhẹ) với cơ chế attention CBAM. Lấy cảm hứng từ MobileNetV3 và DeepLabV3+, SeliNet sử dụng các tích chập tách chiều 1D, các khối bottleneck và Atrous Temporal Pyramid Pooling (ATPP) để nắm bắt ngữ cảnh thời gian đa tỷ lệ trong khi vẫn duy trì độ phức tạp thấp. Bằng cách tích hợp CBAM vào mỗi khối bottleneck, mô hình có thể thích ứng nhấn mạnh các kênh và vị trí thời gian quan trọng trong đặc trưng được mã hóa, từ đó nâng cao mặt nạ tách mà không gây gánh nặng tính toán.

Mục tiêu chính của nghiên cứu này là (i) khảo sát xem việc kết hợp SeliNet với CBAM có thể đạt chất lượng tách cạnh tranh so với các mô hình tiên tiến nặng nề hơn hay không, và (ii) đánh giá một cách hệ thống sự đánh đổi giữa kích thước mô hình, chi phí tính toán và hiệu suất tách ở các cấu hình tỷ lệ khác nhau. Chúng tôi huấn luyện và đánh giá các mô hình đề xuất trên bộ dữ liệu WSJ0-2mix – một benchmark phổ biến cho SCSS – và so sánh với các mạng tách sâu đại diện. Đánh giá sử dụng nhiều chỉ số như SI-SNRi, SDRi, FLOPs và Params để cung cấp một bức tranh toàn diện về chất lượng tách.

Đóng góp chính của chúng tôi gồm:

- Thứ nhất, giới thiệu SeliNet+CBAM – mô hình tách miền thời gian nhẹ nhưng hiệu quả, tận dụng depthwise separable convolutions, ATPP và attention CBAM để đạt hiệu suất tách cao với chi phí thấp hơn nhiều so với các phương pháp hiện có.

• Thứ hai, tiến hành nghiên cứu thực nghiệm mở rộng với các cấu hình tỷ lệ khác nhau (0,25×–1,0×) để phân tích sự đánh đổi hiệu suất–độ phức tạp và chứng minh rằng mô hình nhỏ nhất vẫn giữ hiệu năng tốt trong khi giảm đáng kể GFLOPs và tham số so với Conv-TasNet.

• Thứ ba, chỉ ra rằng kiến trúc đề xuất đạt SI-SNRi và SDRi hàng đầu trong số các mô hình nhẹ và đặc biệt phù hợp để triển khai trên thiết bị di động và biên yêu cầu xử lý thời gian thực.

Bằng cách khai thác sự cộng hưởng giữa backbone nhẹ được thiết kế riêng và cơ chế attention hiệu quả, nghiên cứu này đóng góp vào dòng nghiên cứu đang phát triển về tách tiếng nói trong điều kiện tài nguyên hạn chế và đưa ra các gợi ý thực tiễn cho việc thiết kế các mô hình thế hệ tiếp theo vừa chính xác vừa dễ triển khai.

2. Phương pháp nghiên cứu

2.1. Mô tả bài toán

Giả sử ta có tập dữ liệu huấn luyện:

$$D = \{(s_{mix}^{(i)}, s_1^{(i)}, s_2^{(i)})\}_{i=1}^{N}$$

trong đó $s_{mix}^{(i)}$ là tín hiệu hỗn hợp (mixture) bao gồm hai nguồn nói đồng thời, và $s_1^{(i)}$, $s_2^{(i)}$ là hai tín hiệu sạch của từng người nói mục tiêu mà ta mong ước tách ra. Mỗi tín hiệu ở miền thời gian, có độ dài T. Bài toán tách tiếng nói đơn kênh có thể được mô tả như sau:

$$s_{mix}(t) = s_1(t) + s_2(t), \quad t = 1, 2, 3, \dots, T$$

Mục tiêu là học một hàm f_θ tham số hóa bởi mô hình (mạng neural) sao cho:

$$(\hat{s}_1, \hat{s}_2) = f_\theta(s_{mix})$$

và các tín hiệu ước lượng $\hat{s}_1, \hat{s}_2$ càng gần với s_1, s_2 theo các chuẩn đo chất lượng như SI-SNRi hoặc SDRi. Trong thực tế, mạng sẽ cho ra các mặt nạ hoặc giá trị tách trong miền tính năng trước khi tái tạo tín hiệu đầu ra.

Ta có thể viết hàm mất mát tổng quát là:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{sep}(\hat{s}_1^{(i)}, \hat{s}_2^{(i)}, s_1^{(i)}, s_2^{(i)})$$

trong đó, $\mathcal{L}_{sep}$ là hàm SI-SNR loss với các thành phần mất mát như sau:

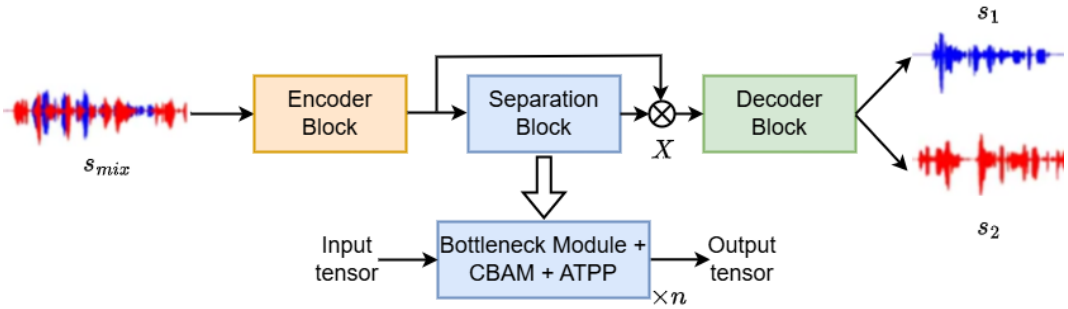
$$\mathcal{L}_{sep} = -SI-SNRi\left(\hat{s}_1^{(i)}, s_1^{(i)}\right) - SI-SNRi(\hat{s}_2^{(i)}, s_2^{(i)})$$

Ngoài ra, trong quá trình thiết kế mô hình, chúng tôi còn sử dụng phương pháp "Permutation Invariant Training (PIT)" để đảm bảo rằng việc ghép cặp ước lượng $\hat{s}$ với các nguồn đích s là đúng cách nhất (không nhầm lẫn giữa các nguồn) – kỹ thuật này được nhiều công trình áp dụng trong các mô hình tách tiếng nói gần đây [11].

2.2. Framework đề xuất: SeliNet + CBAM cho tách tiếng nói

Chúng tôi đề xuất một kiến trúc nhẹ "SeliNet+CBAM" nhằm cân bằng giữa hiệu suất tách và độ phức tạp tính toán. Kiến trúc này gồm phần backbone SeliNet và cơ chế chú ý CBAM được tích hợp vào từng khối bottleneck của SeliNet. Mô hình kiến trúc được mô tả như Hình 1.

Phương pháp của chúng tôi bao gồm 3 thành phần chính: khối Encoder, khối Separation và khối Decoder. Trọng tâm của mô hình nằm ở khối Separation, nó bao gồm một backbone nhẹ được xây dựng dựa trên tích chập tách chiều một chiều, khối bottleneck để giảm số kênh nội bộ, cách kích hoạt Hard-Swish để tăng phi tuyến, và module Atrous Temporal Pyramid Pooling (ATPP) để thu nhận ngữ cảnh thời gian đa tỉ lệ. Ý tưởng là thay vì dùng các lớp convolution thông thường với nhiều tham số, ta sử dụng convolution tách chiều để giảm đáng kể FLOPs và tham số mà vẫn giữ được khả năng trích đặc trưng. ATPP kết hợp các lỗ giãn với các cấp độ khác nhau để mở rộng receptive field mà không tăng nhiều tầng sâu.



Hình 1. *Tổng quan mô hình SeliNet + CBAM*

2.2.1. Khối mã hóa (Encoder Block)

Tín hiệu hỗn hợp dạng sóng trong miền thời gian $x \in R^l$, được tạo ra từ nhiều người nói, được sử dụng làm đầu vào cho bộ mã hóa (encoder). Thay vì sử dụng phép biến đổi Fourier ngắn hạn truyền thống (STFT) trong miền thời gian - tần số (TF), chúng tôi sử dụng bộ mã hóa để thay thế STFT trong quá trình mã hóa tín hiệu. Quá trình này được tính như sau:

$$X = \mathrm{ReLU}\big(\mathrm{Conv1D}(s_{mix})\big) \in R^{l \times d}$$

Trong đó, kiến trúc của bộ mã hóa bao gồm lớp tích chập một chiều (1D convolutional layer) và hàm kích hoạt ReLU nhằm đảm bảo đầu ra không âm. Bộ mã hóa có nhiệm vụ tạo ra các đặc trưng trung gian, và những đặc trưng này sẽ ảnh hưởng trực tiếp đến hiệu suất, kích thước mô hình, cũng như chi phí tính toán.

2.2.2. Khối Separation (Separation Block)

Tích chập phân tách theo chiều sâu 1D: Chúng ta biết rằng lớp tích chập 1D tiêu chuẩn nhận một đặc trưng đầu vào có kích thước $\times d_i$, và áp dụng một kernel $K \in R^{d_i \times k \times d_j}$ để tạo ra đầu ra có kích thước $l \times d_j$. Chi phí tính toán của lớp tích chập 1D có thể được tính bằng: $l \cdot d_i \cdot k \cdot d_j$

Các đặc trưng trung gian được mã hóa từ dạng sóng hỗn hợp và sau đó được đưa vào khối tách. Trong Hình 1 kiến trúc backbone nhẹ bao gồm các mô-đun bottleneck sử dụng tích chập phân tách theo chiều sâu 1D, chia thành tích chập theo chiều sâu và tích chập điểm. Tích chập theo chiều sâu được áp dụng cho mỗi kênh riêng biệt của đặc trưng, trong khi tích chập điểm sẽ kết hợp đầu ra từ tích chập theo chiều sâu trên tất cả các kênh. Tích chập phân tách theo chiều sâu 1D có chi phí tính toán: $l \cdot d_i \cdot (k + d_j)$. Việc sử dụng tích chập phân tách theo chiều sâu 1D giúp mô hình giảm chi phí tính toán xuống:

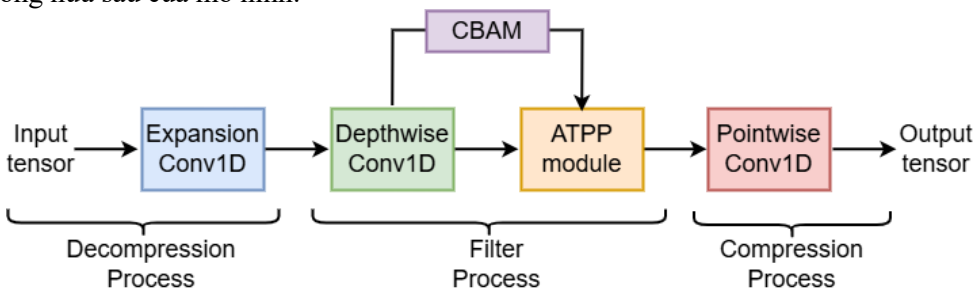
$$\frac{1}{d_j} + \frac{1}{k}.$$

Bottleneck Module: Mạng nơ-ron sâu bao gồm các lớp thành phần, và mỗi lớp có một tensor kích hoạt với kích thước $l_i \times d_i$. Trong đó, tập các kích hoạt lớp tạo thành một đa tạp của thông tin quan trọng. Để giảm tiêu thụ tài nguyên tính toán trong thiết kế mạng, cần thiết kế kích thước mạng nhỏ nhất có thể trong khi vẫn đảm bảo chất lượng hiệu suất. Để giải quyết vấn đề này, như được minh họa trong Hình 1, module bottleneck (BM) được chèn vào trong các mô-đun tích chập nhằm nắm bắt sự đa dạng thông tin cần thiết. BM mở rộng dữ liệu để thu thập thông tin, sau đó nén dữ liệu để giảm kích thước mạng. Các kết nối residual giữa đầu vào và đầu ra của lớp bottleneck là cốt lõi để bảo toàn thông tin.

Lấy cảm hứng từ kiến trúc bottleneck trong MobileNetV3 [10], CBAM được thêm vào mô-đun bottleneck để tạo thành kiến trúc nâng cao — giúp tăng hiệu suất mà chỉ tăng không đáng kể kích thước mô hình và chi phí tính toán, như minh họa trong Hình 2. CBAM là một nhánh phụ sử dụng global average pooling để ghi lại ngữ cảnh toàn cục của lớp trước. Kết quả cuối cùng trong

nhánh này là một vector ngữ cảnh toàn cục, được sử dụng để scale đầu vào. Hàm Hard-swish (HS) [8] được sử dụng trong module bottleneck (như trình bày trong Bảng 1) và HS giúp giảm số lần truy cập bộ nhớ, từ đó giảm đáng kể độ trễ.

Lợi ích của HS được thể hiện rõ nhất khi sử dụng trong các lớp sâu hơn. Do đó, HS được áp dụng trong nửa sau của mô hình.



Hình 2. *Kiến trúc của module bottleneck*

Về attention, chúng tôi tích hợp CBAM cho mỗi khối bottleneck. CBAM gồm hai thành phần: attention kênh (channel attention) và attention không gian (spatial attention). Trong tầng kênh, module này học trọng số cho từng kênh đặc trưng bằng cách sử dụng hai phép pooling (average pooling và max pooling) kết hợp MLP nhỏ, rồi nhân trọng số đó vào các kênh tương ứng. Trong tầng không gian, sau khi hợp kênh, một bản đồ attention không gian được học dựa vào pooling theo chiều kênh và MLP để nhấn mạnh các vị trí thời gian quan trọng trong chuỗi đặc trưng. Việc kết hợp CBAM vào mỗi khối cho phép mô hình tự điều chỉnh nhấn mạnh các kênh và vị trí thời gian quan trọng, giúp lọc đặc trưng nhiễu hoặc không liên quan.

2.2.3. Khối giải mã (Decoder Block)

Trong bộ giải mã, chúng tôi đơn giản sử dụng một lớp tích chập chuyển vị với kích thước kernel và stride giống như ở bộ mã hóa, nhằm tái tạo lại tín hiệu lời nói đã được ước lượng. Đầu vào của decoder là phép nhân từng phần tử (element-wise) giữa mặt nạ $\widehat{m_c}$ của từng nguồn tín hiệu và đầu ra của encoder e, và được biểu diễn như sau:

$$\widehat{s_c} = \mathrm{ConvTr1D}(X \odot \widehat{m_c})$$

Trong đó, $\widehat{s_c} \in \mathrm{R}^1$ là tín hiệu lời nói được ước lượng của nguồn thứ c. ConvTr1D(·) là lớp tích chập chuyển vị một chiều và mặt nạ $\widehat{m_c}$ là kết quả của khối Separation.

Toàn bộ kiến trúc được mô tả như sau: tín hiệu hỗn hợp s_{mix} đi qua lớp encoder 1D convolution để sinh ra biểu diễn trung gian X. Sau đó, biểu diễn này được đi qua chuỗi các khối bottleneck SeliNet (mỗi khối gồm depthwise separable conv + CBAM + ATPP nếu áp dụng) để tạo ra mã feature nội tại $\widehat{m_c}$. Cuối cùng, phần decoder sử dụng các lớp transposed convolution hoặc deconvolution để tái tạo các tín hiệu $\hat{s}_1, \hat{s}_2$. Mỗi khối bottleneck có thể được nhân bản nhiều lần tùy theo hệ số scale (ví dụ 0,25×, 0,5×, 1,0×), để điều chỉnh độ phức tạp mô hình.

Nếu biểu diễn công thức hoạt động của một khối bottleneck i với CBAM như:

$$Y_i = PW(ATPP\left(CBAM\left(DWConv(X_i)\right)\right)),$$

trong đó DWConv là lớp convolution tách chiều, và CBAM gồm attention kênh và spatial như:

$$CBAM(F) = M_{spatial}(M_{channel}(F) \odot F) \odot F,$$

với $M_{channel}: \mathbb{R}^{C\times T} \to \mathbb{R}^{C}$ (trọng số cho từng kênh) và $M_{spatial}: \mathbb{R}^{C\times T} \to \mathbb{R}^{1\times T}$ (trọng số cho từng vị trí thời gian) và PW là khối convolution 1×1 (Pointwise Conv).

Ngoài ra, để cải thiện sự ổn định và hiệu quả hội tụ, chúng tôi áp dụng kỹ thuật giảm learning rate theo epoch, dropout, và normalization giữa các khối bottleneck. Trong các thử nghiệm ablation (mô-đun bỏ CBAM hoặc bỏ ATPP), chúng tôi đánh giá ảnh hưởng của từng thành phần. Chúng tôi cũng so sánh mô hình của mình với các kiến trúc khác như SeliNet, Conv-TasNet, DPRNN,… Toàn bộ framework được thiết kế để dễ mở rộng, cụ thể chúng ta có thể thêm mô-

đun attention khác hoặc thay đổi số tầng bottleneck mà không ảnh hưởng đến kiến trúc tổng thể. Khi đánh giá, các cấu hình scale nhỏ hơn (0,25×, 0,5×) cho phép triển khai trên thiết bị biên với hạn chế tài nguyên.

Bảng 1. *Kiến trúc mô hình SeliNet + CBAM*

Input	Block	EZ	CBAM	NL	s	Output
32000×64	BM, 3	16	✓	RL	2	16
16000×16	BM, 3	72	-	RL	2	24
8000×24	BM, 3	88	-	RL	1	24
8000×24	BM, 3	96	✓	HS	2	40
4000×40	BM, 5	240	✓	HS	1	40
4000×40	BM, 5	240	✓	HS	1	40
4000×40	BM, 5	120	✓	HS	1	48
4000×48	BM, 5	144	✓	HS	1	48
4000×48	BM, 5	288	✓	HS	2	96
2000×96	BM, 5	576	✓	HS	1	96
2000×96	BM, 5	576	✓	HS	1	96
2000×96	BM, 5	576	✓	HS	1	96

3. Thử nghiệm

3.1. Bộ dữ liệu

Một trong những tập dữ liệu nền tảng và được sử dụng rộng rãi nhất là WSJ0-2mix. Tập dữ liệu này được xây dựng từ Wall Street Journal corpus – một kho dữ liệu âm thanh chuẩn về tiếng Anh đọc từ các bài báo. Trong WSJ0-2mix, các đoạn âm thanh từ hai người nói được chọn ngẫu nhiên, chuẩn hóa về độ dài và trộn lẫn với nhau theo một tỷ lệ nhất định để tạo ra hỗn hợp hai giọng nói. Bên cạnh file âm thanh hỗn hợp, tập dữ liệu còn cung cấp các tệp tham chiếu tương ứng cho từng người nói gốc, giúp việc tách và đánh giá trở nên minh bạch, tái lập và có thể so sánh trực tiếp giữa các mô hình.

3.2. Kết quả thử nghiệm

Chúng tôi trình bày kết quả thực nghiệm của mô hình SeliNet + CBAM thông qua hiệu suất tách nguồn, kích thước mô hình và yêu cầu tính toán, so với các phương pháp được công bố gần đây trên tập dữ liệu WSJ0-2mix, như được thể hiện trong Bảng 2. Dễ nhận thấy rằng SeliNet + CBAM đạt được hiệu suất SOTA trong khi sử dụng một mô hình có kích thước nhỏ đáng kể cùng với chi phí tính toán thấp hơn. Cụ thể, SeliNet + CBAM đạt được các chỉ số SDRi: 15,88 dB, SI-SNRi=15,72dB. Kích thước của mô hình là 2,11M và chi phí tính toán là 2,18 GFLOPs.

Điều quan trọng là mô hình đề xuất có khả năng triển khai trên nhiều thiết bị khác nhau, đặc biệt là các thiết bị biên với chi phí tính toán thấp. SeliNet + CBAM được xây dựng dựa trên kiến trúc nhẹ, hoàn toàn sử dụng các lớp tích chập một. Kiến trúc này cho thấy hiệu quả vượt trội so với các mô hình dựa trên BLSTM và Transformer trong môi trường hạn chế tài nguyên.

Mặc dù DPRN có kích thước mô hình nhỏ, chi phí tính toán của nó vẫn lớn hơn gấp 22 lần so với SeliNet + CBAM 1.00x. Điều đáng chú ý là SeliNet + CBAM đạt được kết quả cao hơn nhờ kiến trúc 1D convolution, cụ thể với ΔSDRi = +0,2 dB, ΔSI-SNRi = +0,31 dB. Về mặt kiến trúc, mô hình SeliNet + CBAM có lượng phép tính (GFLOPs) thấp hơn khoảng một nửa so với mô hình Conv-TasNet. Đồng thời, kích thước mô hình của SeliNet + CBAM cũng chỉ bằng khoảng một nửa kích thước của Conv-TasNet. Ngoài ra, SeliNet + CBAM còn có lợi thế rõ rệt về số lượng tham số và chi phí tính toán khi so sánh với các mô hình hiện đại gần đây.

Bảng 2. *So sánh giữa SeliNet + CBAM và các phương pháp khác trên tập dữ liệu WSJ0-2mix*

Method	Params	GFLOPs	SI-SNRi	SDRi
Conv-TasNet [2]	5,10M	5,30	15,3	15,6
DPRNN [3]	2,60M	48,9	18,8	19,0
Demucs [15]	415M	10,34	12,12	-
TDCN [16]	8,63M	7,23	16,1	-
SuDoRM-RF [17]	2,66M	2,56	17,02	-
SeliNet [18]	2,10M	2,18	15,61	15,80
SepFormer [19]	26M	-	20,40	20,50
SeliNet + CBAM 1.00x	**2,11M**	**2,18**	**15,72**	**15,88**

Các điều chỉnh khác nhau trong mô hình SeliNet + CBAM 1.00x có ảnh hưởng đến hiệu suất tách tiếng nói trên tập dữ liệu WSJ0-2mix, như được thể hiện trong Bảng 3. Có thể quan sát thấy rằng thành phần ATPP có ảnh hưởng đáng kể đến hiệu suất tách nguồn, trong khi CBAM điều chỉnh linh hoạt phản hồi đặc trưng theo từng kênh. Ngoài ra, SE cũng giúp cải thiện hiệu suất tách một cách nhẹ nhưng ổn định.

Hàm kích hoạt hard-swish (HS) được sử dụng thay thế cho PReLU. Điều thú vị là mô hình SeliNet đạt kết quả tốt hơn khi sử dụng HS, trong khi hiệu suất giảm mạnh khi dùng PReLU. Trong thí nghiệm này, chúng tôi quan sát thấy hiệu suất cải thiện trong các trường hợp cấu hình khác nhau được trình bày trong Bảng 3. Cuối cùng, chúng tôi ghi nhận rằng hiệu suất tốt nhất đạt được khi kết hợp cả ba thành phần: ATPP, SE, và HS.

Trong Bảng 4, chúng tôi khảo sát hiệu suất của các phiên bản khác nhau của mô hình SeliNet + CBAM trên tập dữ liệu WSJ0-2mix. Có thể thấy rằng mô hình đề xuất có thể được thu gọn để phù hợp với các ứng dụng khác nhau. Khi hệ số thu nhỏ giảm, kích thước mô hình và chi phí tính toán cũng giảm theo. Tuy nhiên, điều này sẽ đánh đổi với việc hiệu suất của mô hình cũng bị giảm đi tương ứng.

Bảng 3. *So sánh các thành phần CBAM, ATPP và hàm phi tuyến trong mô hình SeliNet + CBAM trên tập dữ liệu WSJ0-2mix*

Hàm phi tuyến	CBAM	ATPP	SI-SNRi	SDRi
PReLU	✗	✗	12,70	12,90
PReLU	✓	✗	12,97	13,24
PReLU	✓	✓	13,94	14,24
HS	✗	✗	14,37	14,72
HS	✓	✗	14,56	14,84
HS	✓	✓	15,72	15,88

Kích thước mô hình nhỏ và chi phí tính toán thấp sẽ giảm theo tương ứng. Tuy nhiên, mô hình vẫn đạt được hiệu suất chấp nhận được. Điều này đặc biệt phù hợp cho các ứng dụng trên nhiều thiết bị thời gian thực có tài nguyên hạn chế.

Bảng 4. *Hiệu suất tách tiếng nói của mô hình SeliNet + CBAM với các hệ số thu nhỏ khác nhau trên tập dữ liệu WSJ0-2mix*

Mô hình	#Params	GFLOPs	SI-SNRi	SDRi
SeliNet + CBAM 0.25x	0,7M	0,56	12,76	12,94
SeliNet + CBAM 0.50x	1,0M	1,20	14,68	14,85
SeliNet + CBAM 0.75x	1,5M	1,61	15,24	15,74
SeliNet + CBAM 1.00x	2,1M	2,18	15,72	15,88

4. Kết luận

Nghiên cứu này cho thấy việc kết hợp backbone nhẹ SeliNet với cơ chế attention CBAM mang lại hiệu quả rõ rệt trong bài toán tách tiếng nói kênh đơn, vừa duy trì được chất lượng tách cao vừa giảm đáng kể chi phí tính toán. Kết quả thực nghiệm trên bộ dữ liệu WSJ0-2mix chứng minh mô hình đề xuất đạt hiệu năng SI-SNRi và SDRi cạnh tranh với các mạng sâu hiện đại nhưng có ít tham số hơn đáng kể. Khả năng thích ứng với nhiều cấu hình tỷ lệ cũng cho thấy tính

linh hoạt cao trong việc triển khai trên các nền tảng phần cứng khác nhau. Kiến trúc nhẹ này góp phần thu hẹp khoảng cách giữa các mô hình học sâu mạnh mẽ và yêu cầu xử lý thời gian thực trong môi trường hạn chế tài nguyên. Những kết quả đạt được không chỉ khẳng định tiềm năng của các kỹ thuật depthwise separable convolutions và attention nhẹ mà còn mở ra hướng nghiên cứu tiếp theo về tối ưu kiến trúc cho các bài toán tăng cường và nhận dạng tiếng nói. Mặc dù mô hình được thiết kế nhẹ, việc triển khai trên thiết bị biên vẫn có thể gặp khó khăn do hạn chế về bộ nhớ, tiêu thụ năng lượng và độ trễ trong suy luận. Cơ chế attention, dù được tinh giản, vẫn làm tăng phần nào độ phức tạp tính toán so với các mô hình hoàn toàn dựa trên convolution.

Lời cám ơn

Nghiên cứu này được hỗ trợ bởi đề tài T2025-07-02 được thực hiện tại Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, “A review of deep learning techniques for speech processing,” *Information Fusion*, vol. 99, 2023, Art. no. 101869.

[2] 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.

[3] 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.

[4] A. Gulati *et al.*, “Conformer: Convolution-augmented Transformer for speech recognition,” in *Proc. Interspeech*, 2020, pp. 5036-5040.

[5] 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.

[6] K. Tan, Y. Zhang, and D. Wang, “Deep learning based real-time speech separation for mobile devices,” *IEEE Signal Process. Lett.*, vol. 28, pp. 1–5, 2021.

[7] Y. Xiang and D. Wang, “Lightweight speech separation with depthwise separable convolutions,” in *Proc. ICASSP*, 2022, pp. 126–130.

[8] H. Li, L. Chen, and Z. Huang, “Resource-efficient speech enhancement via mobile architectures,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1234–1247, 2022.

[9] S. Zhang *et al.*, “Designing efficient neural networks for on-device speech separation,” *Neural Networks*, vol. 157, pp. 98–109, 2023.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. ECCV*, 2018, pp. 3–19.

[11] Z. Yao, W. Pei, F. Chen, G. Lu, and D. Zhang, “Stepwise-Refining Speech Separation Network via Fine-Grained Encoding in High-Order Latent Domain,” *arXiv preprint*, 2021.

[12] J. Wang, “An Efficient Speech Separation Network Based on Recurrent Fusion Dilated Convolution and Channel Attention,” *Interspeech*, 2023, pp. 3699-3703.

[13] H. Ma, “A novel end-to-end deep separation network based on the attention mechanism,” *IET Signal Processing*, vol. 17, no. 2, pp. 1-10, 2023.

[14] K. Wang, H. Zhou, J. Cai, and W. Li, “Time-domain adaptive attention network for single-channel speech separation,” *EURASIP J. Audio, Speech, Music Process.*, vol.21, pp. 1-15, 2023.

[15] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” *Proc. Interspeech*, Sep. 2020, pp. 3291–3295, doi: 10.21437/Interspeech.2020-2309.

[16] Y. Luo and N. Mesgarani, “TasNet: Time-domain audio separation network for real-time, single-channel speech separation,” in *Proc. ICASSP*, Apr. 2018, pp. 696–700.

[17] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. R. Hershey, “Unsupervised sound separation using mixture invariant training,” *Proc. NeurIPS*, 2020, pp. 1-12.

[18] H. M. Tan, D.-Q. Vu, and J.-C. Wang, "Selinet: a lightweight model for single channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5.

[19] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5749-5753.