# Tăng cường huấn luyện mô hình học tự chú ý cho phân tích và phân đoạn tiếng nói

Hà Minh Tân, Nguyễn Kim Quốc

Khoa Công nghệ Thông tin, Trường Đại học Nguyễn Tất Thành
hmtan@ntt.edu.vn, nkquoc@ntt.edu.vn

## Tóm tắt

Nghiên cứu này đề xuất một phương pháp tăng cường mới sử dụng mô hình tự chú ý để phân tách giọng nói đơn kênh. Đầu tiên, đóng băng tất cả các lớp trong mô hình tự chú ý đã được huấn luyện trước. Tiếp theo, tiến hành huấn luyện lại mô hình qua ba giai đoạn, sử dụng cơ chế lập lịch để điều chỉnh tốc độ học tập và mở khóa các lớp trong mô hình theo lịch trình. Qua quá trình này, mô hình được cập nhật và nâng cấp từ kiến thức trước đó, giúp cải thiện hiệu suất mô hình đồng thời giảm thiểu thời gian và chi phí huấn luyện. Phương pháp này không chỉ giúp tăng hiệu suất của mô hình so với các phương pháp truyền thống mà còn có thể được áp dụng để cải thiện hiệu suất của các mô hình hiện có. Kết quả thử nghiệm choa thấy rằng mô hình được huấn luyện theo phương pháp này vượt trội hơn các phương pháp hiện tại đối với nhiệm vụ tách giọng nói đơn âm trên các tập dữ liệu thông thường.



Nhận 10/03/2024
Được duyệt 17/03/2024
Công bố 29/03/2024

## Từ khóa

Mô hình được huấn luyện trước, bộ chuyển đổi (transfomer), cơ chế tự chú ý, mô hình tinh chỉnh, khung che dấu thời gian, tách giọng nói

## 1 Đặt vấn đề

Học sâu đã chứng tỏ tính hiệu quả trong việc giải quyết các vấn đề mang tính quy tắc không rõ ràng trong lĩnh vực tự nhiên. Trong lĩnh vực thị giác máy tính, nhiều mô hình học sâu đã đạt được thành công đáng kể [1-5]. Tương tự, trong xử lý ngôn ngữ tự nhiên [6, 7] và xử lý âm thanh [8-12], các mô hình học sâu cũng đã mang lại những kết quả tích cực. Trong việc xử lý tiếng nói, một thách thức quan trọng là phải tách biệt các âm thanh. Phương pháp huấn luyện phổ, đặc biệt là phương pháp không gian nhúng sâu, đã đạt được hiệu suất và hiệu quả tiên tiến nhất đối với số lượng người nói chưa biết [13-15]. Gần đây, các nghiên cứu dựa trên tiếng nói miền thời gian đã đạt được thành công đáng kể. Ví dụ, một khung phân tách cơ bản được đưa ra bởi Luo đã đạt được hiệu suất vượt trội [10]. Các tác giả trong một nghiên cứu khác đã chỉ ra rằng phương pháp này vượt trội hơn phương pháp phân tách quang phổ bằng cách sử dụng mặt nạ tần số thời gian (time-frequency mask) [16]. Cấu trúc phân tách giọng nói đặc biệt hiệu quả, bao gồm cấu trúc nội bộ và liên phân đoạn, đã được đề xuất để đơn giản hóa quá trình tính toán và cải thiện khả năng hiểu ngữ cảnh của chuỗi âm thanh dài [9].

Cơ chế chú ý và cơ chế tự chú ý đại diện cho các thành phần then chốt trong cấu trúc bộ chuyển đổi (transformer) [17], tạo thành nền tảng cho các mô hình học sâu hiện đại. Bước đột phá mang tính biến đổi này đã cách mạng hóa các cấu trúc tuần tự, đạt được thành công vượt trội trong xử lý ngôn ngữ [7], đặc biệt là trong dịch máy [18] và mở rộng tác động của nó đến lĩnh vực thị giác máy tính [2, 3]. Bộ chuyển đổi nhận diện bối cảnh phụ thuộc vào các chuỗi dài phù hợp cho các tác vụ xử lý giọng nói và âm thanh. Nó đã thực sự thành công trong việc nhận dạng giọng nói, xác minh người nói, sự tách biệt [6, 19]. Nhận dạng bối cảnh của một chuỗi thời gian cực kỳ dài là động lực trong việc tách nguồn âm thanh trong miền thời gian. Đặc biệt,

https://doi.org/10.55401/8vq83773

việc tách giọng nói đòi hỏi phải xử lý chuỗi thời gian cực dài, đặt ra thách thức cho mô hình bộ chuyển đổi do độ phức tạp bậc hai của cơ chế tự chú ý [17]. Vì vậy, bộ chuyển đổi thường bị tắc nghẽn tính toán với chuỗi thời gian cực kỳ dài. Việc giảm tắc nghẽn bộ nhớ trong cơ chế tự chú ý của bộ chuyển đổi đã được các tác giả nghiên cứu trong nhiều năm qua. Cách phổ biến để giải quyết vấn đề này là tạo các tập hợp con. Để nắm bắt cả ngắn hạn và phụ thuộc lâu dài, bộ chuyển đổi phân đoạn [3] sử dụng cả khung trượt cục bộ và khung trượt giãn nở trên một số lượng phần tử không đổi. Các phiên bản sau này dựa trên [20], LongFormer [21] sử dụng các đầu chú ý tổng thể kết hợp để cải thiện bộ chuyển đổi trong khi Linformers [22] sử dụng phép nhân ma trận bậc thấp để ước tính toàn bộ chú ý phân đoạn. Một phương pháp thành công là chia nhạy cảm cục bộ sử dụng để phân cụm các phần tử [23].

Tiếp nối thành công của cấu trúc bộ chuyển đổi, nó được tích hợp vào cấu trúc nội bộ. Trong nghiên cứu này, dựa trên những thành công của Sepformer [17], đề xuất được sử dụng để huấn luyện lại mô hình này với nhiều thông số kỹ thuật và áp dụng nó cho các bộ dữ liệu khác nhau, cụ thể là khung tự chú ý được huấn luyện trước. Đầu tiên, tất cả các lớp trong khuôn khổ tự chú ý được huấn luyện trước đều bị đóng băng. Sau đó, mô hình được huấn luyện lại qua ba giai đoạn bằng cách sử dụng cơ chế lập lịch cho tốc độ học tập và các lớp của khung được mở khóa theo lịch trình. Bằng cách này, mô hình được học lại, nâng cao và cập nhật từ kiến thức trước đó. Kết quả cho thấy rằng đợt huấn luyện tiếp theo đã mang lại hiệu quả rõ rệt, đạt kết quả tốt hơn so với mô hình ban đầu, đồng thời phương pháp này cũng giúp giảm đáng kể chi phí và thời gian huấn luyện. Đây cũng là phương pháp tiếp cận hiệu quả cho các ứng dụng cụ thể dựa trên phương pháp huấn luyện lại và đạt được kết quả thành công trong việc tách tiếng nói.

## 2 Hệ thống tách giọng nói miền thời gian

### 2.1 Khối mã hóa thông tin đầu vào

Dạng sóng hỗn hợp trong miền thời gian, $y \in R^{l}$, của một số bộ nói độc lập được sử dụng làm tính năng đầu vào của bộ mã hóa. Giống như trong miền tần số thời gian, nhiều nghiên cứu trước đây áp dụng tính năng biến đổi Fourier thời gian ngắn. Trong miền thời gian, tính năng biến đổi Fourier thời gian ngắn được thay thế bằng tính năng trung gian, được tạo bởi khối mã hóa tích chập và được tính như sau:

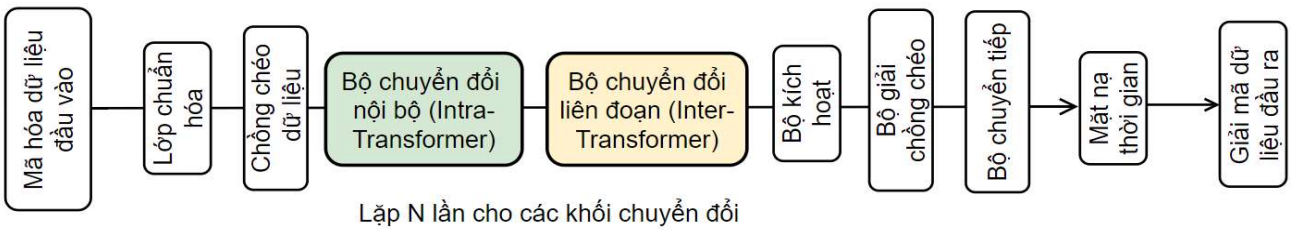
$$e = \mathrm{Re}\, Lu(conv1D(y)) \qquad (1)$$

Trong đó cấu trúc bộ mã hóa có thể được định nghĩa là lớp chập 1D (Conv1D) với kích hoạt đơn vị tuyến tính được chỉnh lưu (ReLU) để đảm bảo kết quả không âm. Cơ chế mã hóa được sử dụng để tạo ra các tính năng trung gian ảnh hưởng đến hiệu suất, kích thước và chi phí tính toán của mô hình.

### 2.2 Khối phân tách

*Cấu trúc chung:* cấu trúc chi tiết của hệ thống phân tách được hiển thị trong Hình 1. Đầu vào của khối mặt nạ là đặc điểm trung gian $e \in R^{l \times d}$ trong khi đầu ra của khối mặt nạ là mặt nạ dự đoán $\hat{m}_1, \hat{m}_2, ..., \hat{m}_c$ của người nói c trong cách nói hỗn hợp.

Tính năng trung gian (trong Phần 2.1) được chuẩn hóa bằng cách chuẩn hóa lớp và được thực hiện bởi một lớp được kết nối đầy đủ với khung tuyến tính. Các phân đoạn chồng chéo được tạo với kích thước sử dụng mức chồng chéo 50 %. Đặc tính tensor, $e^{(1)} \in R^{s \times k \times d}$, trong đó d biểu thị đặc điểm thứ nguyên, s biểu thị độ dài đoạn và k biểu thị số đoạn. Khối chính của khối mặt nạ là bộ chuyển đổi, với thành phần chính là khung tự chú ý. Biểu diễn tensor được cấp cho các bộ chuyển đổi sử dụng hai bộ chuyển đổi (tức là các bộ chuyển đổi nội bộ và liên đoạn) mắc nối tiếp.




**Hình 1** Sơ đồ tổng quát hệ thống tách giọng nói miền thời gian

Cấu trúc này có thể học kiến thức từ các chuỗi phụ thuộc ngắn hạn và dài hạn. Sau đó, đầu ra của mô đun chính, $e^{(2)} \in R^{s\times k\times d}$, được xử lý bởi đơn vị tuyến tính chỉnh lưu tham số (PReLU). Sau đó, lớp tuyến tính được sử dụng và đầu ra là $e^{(3)} \in R^{s\times k\times d\times c}$. Cuối cùng, lớp khử chồng lấp được sử dụng để tạo $e^{(4)} \in R^{l\times d\times c}$ Mô hình chạy biểu diễn này thông qua hai lớp chuyển tiếp nguồn cấp dữ liệu, theo sau là hàm ReLU, để tạo ra mặt nạ mc của mỗi người nói.

Bộ chuyển đổi liên đoạn và nội bộ: các bộ chuyển đổi được chế tạo cho bộ chuyển đổi liên đoạn và nội bộ dựa trên [17]. Đầu vào bộ chuyển đổi biểu thị ϑ và mã hóa vị trí của bộ chuyển đổi là μ. Công thức được định nghĩa như sau: $\vartheta^{(1)} = \vartheta + \mu$. Mã hóa vị trí bổ sung kiến thức về thứ tự của các mục riêng lẻ trong chuỗi, nâng cao hiệu quả phân tách. Sau đó, nhiều mô đun bộ chuyển đổi được sử dụng. Trong mỗi bộ chuyển đổi, lớp định mức và sự chú ý nhiều đầu (MHA) với lớp chuẩn hóa (Normlayer) được áp dụng và được xác định như sau:

$$\vartheta(2) = \text{MHA}(\text{Normlayer}(\vartheta^{(1)})) \quad (2)$$

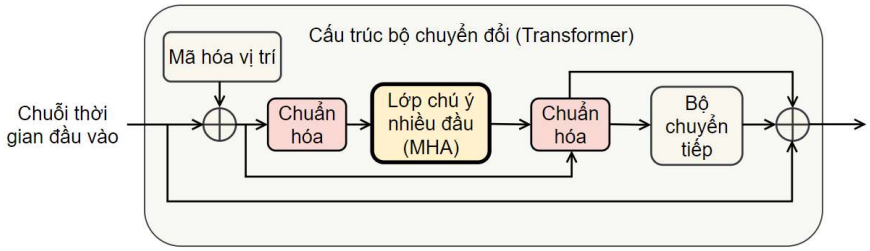
Mỗi phần đầu của sự chú ý sẽ tính toán sự chú ý của sản phẩm chấm theo tỷ lệ cho tất cả các thành phần khác nhau của một chuỗi. Mạng chuyển tiếp nguồn cấp dữ liệu (FFN) được sử dụng ở vị trí độc lập và được xác định như sau:

$$\vartheta(3) = \text{FFN}(\text{Normlayer}(\vartheta^{(2)} + \vartheta^{(1)})) + \vartheta^{(2)} + \vartheta^{(1)} \quad (3)$$

Mô-đun bộ chuyển đổi tổng thể được tính như sau:

$$f(\vartheta) = h^{\kappa}(\vartheta + \mu) + \vartheta, \quad (4)$$

ở đây $h^{\kappa}(\cdot)$ là κ lớp của $h(\cdot)$ (tức là $h(\cdot)$ là mô đun bộ chuyển đổi [8]). Theo [17], κ cũng là số lượng bộ chuyển đổi liên và bộ chuyển đổi nội bộ. Trong Hình 2, các bộ chuyển đổi giữa và trong là mô đun xử lý chính [8]. Kiến trúc này dựa trên DPRNN [9] để mô hình hóa các chuỗi dài hạn và ngắn hạn. Trong [8], các bộ chuyển đổi liên đoạn và nội bộ được sử dụng để thay thế các mạng hồi tiếp giữa và trong. Bộ chuyển đổi liên đoạn được sử dụng để mô hình hóa các trình tự dài hạn trong khi bộ chuyển đổi nội bộ được sử dụng để xử lý các trình tự ngắn hạn. Như tính toán trong phương trình (4), chức năng tổng thể của các bộ chuyển đổi giữa và trong bộ chuyển đổi được xác định như sau: $e^{(2)} = f_{inter}(f_{intra}(e^{(1)}))$, trong đó $f_{inter}$ và $f_{intra}$ lần lượt biểu thị bộ chuyển đổi liên đoạn và bộ chuyển đổi nội bộ.




**Hình 2** Kiến trúc bộ chuyển đổi trong mô hình học sâu.

### 2.3 Khối giải mã tích chập chuyển đổi

Trong khối giải mã, lớp tích chập chuyển đổi được sử dụng. Lớp tích chập chuyển đổi được cấu hình với kích thước lõi và bước tiến bằng khối mã hóa. Các cách phát âm ước tính được xây dựng lại bằng phép nhân theo từng phần tử của mặt nạ ước tính của nguồn tín hiệu phát âm $\hat{m}_c$ và đặc điểm trung gian e và được viết như sau $\hat{x}_c = ConvTr1D(\hat{m}_c \Theta e)$ trong đó $\hat{x}_c$ biểu thị cách phát âm ước tính của câu nói ban đầu thứ c. $ConvTr1D(\cdot)$ là lớp chập 1D được chuyển đổi.

## 3 Khung tự chú ý được huấn luyện tăng cường

Trong nghiên cứu này, khung tự chú ý được huấn luyện trước và được đề xuất thành ba giai đoạn của quá trình học tập. Trong giai đoạn đầu tiên, bộ chuyển đổi nội bộ và liên đoạn chuyển đổi thứ nhất được đóng băng và tinh chỉnh trọng số của bộ chuyển đổi nội bộ và liên đoạn chuyển đổi thứ hai. Trong giai đoạn này, mô hình được huấn luyện lại và cập nhật trọng số cho khối thứ hai bằng cơ chế lập kế hoạch cho tốc độ học tập và các lớp của khung được mở khóa theo lịch trình. Trong giai đoạn thứ hai, bộ chuyển đổi nội bộ và liên đoạn chuyển đổi thứ

https://doi.org/10.55401/8vq83773

hai bị đóng băng và trọng số của bộ chuyển đổi đầu tiên được cập nhật bằng cơ chế lập kế hoạch tốc độ học tập. Ở giai đoạn cuối, tất cả các khối thành phần của khung đều được tinh chỉnh. Kế thừa các trọng số được tạo ban đầu, ở công việc tiếp theo, quá trình huấn luyện sẽ thực hiện từ các khối cuối cùng trước đến các khối tiếp theo. Điều này giúp hệ thống tinh chỉnh và cập nhật những thiếu sót trước đó, như nhiệm vụ rà soát để đưa ra quyết định chính xác hơn. Điểm mạnh của phương pháp này là tăng hiệu suất đồng thời giảm đáng kể thời gian huấn luyện. Theo cách tiếp cận này, hiệu suất được nâng cao của mô hình so với mô hình không được huấn luyện trước bắt nguồn từ việc sử dụng các tính năng ở giai đoạn đầu được huấn luyện trước trong suốt các giai đoạn sau. Chiến lược tái sử dụng các tính năng này không chỉ tận dụng lượng kiến thức phong phú đã được mã hóa trong mô hình được huấn luyện trước mà còn tạo điều kiện cho quá trình học tập hiệu quả hơn. Bằng cách tận dụng các biểu diễn có sẵn này, mô hình này có thể tập trung nỗ lực huấn luyện vào việc tinh chỉnh và tinh chỉnh các tính năng được trích xuất để phù hợp hơn với nhiệm vụ cụ thể trước mắt. Do đó, cách tiếp cận này dẫn đến khả năng khái quát hóa được cải thiện, cho phép mô hình để xử lý hiệu quả các đầu vào và tình huống đa dạng ngoài những tình huống gặp phải trong quá trình huấn luyện trước. Phương pháp này nêu bật tầm quan trọng của việc tái sử dụng tính năng và học hỏi chuyển giao trong việc củng cố hiệu suất mô hình trên một loạt ứng dụng.

## 4 Thực nghiệm

### 4.1 Bộ dữ liệu

Trong bài báo này, kho dữ liệu WSJ0 với sự kết hợp của bộ dữ liệu WSJ0-2mix và WSJ0-3mix được sử dụng để huấn luyện. Tích hợp này được tạo ra bởi sự kết hợp ngẫu nhiên của các giọng nói trong khoảng từ -5 dB đến +5 dB SNR. Tập dữ liệu huấn luyện với 20 000 cách phát âm, 5 000 cách phát biểu của tập dữ liệu xác thực và 3 000 cách phát biểu của tập dữ liệu thử nghiệm được sử dụng cho mô hình huấn luyện. Tất cả các tập dữ liệu được lấy mẫu ở tần số 8 kHz. Sự kết hợp của các dạng sóng phát âm sử dụng độ dài ngắn hơn. Đây là các bộ dữ liệu chuẩn để so sánh hiệu suất của các thuật toán tách âm thanh. Hơn nữa, các thử nghiệm còn mở rộng sang các bộ dữ liệu đầy thách thức, tức là WHAM! và WHAMR!, là những điều kiện ồn ào và âm vọng trong các tính năng hỗn hợp. Môi trường tự nhiên WHAM! và WHAMR! bộ dữ liệu chứa những lời nói ồn ào và âm vọng từ các nhà hàng và quán cà phê. Trộn động (DM) [24] được sử dụng để tăng cường, một kỹ thuật liên quan đến việc tạo ra các cách nói hỗn hợp mới theo thời gian thực từ từng người nói những lời phát biểu. Trong nghiên cứu này, phương pháp hiệu quả được nâng cao bằng cách đưa ra nhiễu loạn tốc độ trước khi trộn các cách phát âm. Tốc độ dao động tùy ý giữa mức giảm tốc độ 95 % và tốc độ tăng tốc 105 %.

### 4.2 Cấu hình mô hình

Cấu hình như [8] với 256 bộ lọc của lớp chập, kích thước nhân 16 mẫu và hệ số bước 8 mẫu được sử dụng. Bộ mã hóa sử dụng kích thước nhân và hệ số bước giống hệt như bộ giải mã. Khối mặt nạ được thiết kế theo từng đoạn có kích thước 250 và độ chồng lên nhau là 50 %. Sepformer với các bộ chuyển đổi giữa và trong được sử dụng. Luồng xử lý đường dẫn kép được lặp lại hai lần. Tám đầu chú ý song song và mỗi bộ chuyển đổi có mô hình cấp nguồn theo vị trí 1024 được sử dụng để huấn luyện.

### 4.3 So sánh và kết quả

*So sánh với các mô hình khác:* phương pháp đề xuất (tức là sử dụng khung Sepfomer [8]) được huấn luyện lại trên các bộ dữ liệu chuẩn WSJ0-2mix như trong Bảng 1. Mô hình này được coi là mô hình hiện đại hàng đầu, tức là kết hợp hai phương pháp hàng đầu là mô hình bộ chuyển đổi [17] và giữa các liên đoạn và nội bộ [9]. Những kết quả thử nghiệm này vượt trội hơn các phương pháp tiên tiến trước đó.

**Bảng 1** So sánh hiệu quả của các mô hình trên tập dữ liệu WSJ0-2mix.

| Mô hình | Kích thước (M) | SI-SNRi | SDRi |
|---|---|---|---|
| TasNet [10] | - | 10,8 | 11,1 |
| Conv-TasNet [16] | 5,1 | 15,3 | 15,6 |
| DPRNN [9] | 2,6 | 18,8 | 19,0 |
| SepFormer [8] | 26 | 20,4 | 20,5 |
| Đào tạo tăng cường | 26 | 20,8 | 21,0 |
| Đào tạo tăng cường + DM | 26 | 22,7 | 22,8 |

*So sánh với khung cơ sở:* trong thử nghiệm này, Sepformer được sử dụng làm đường cơ sở. Mô hình này tiếp tục huấn luyện lại với ba giai đoạn theo những cách khác nhau. Trước tiên, bộ chuyển đổi liên đoạn và nội bộ bị đóng băng và phương pháp đề xuất được huấn luyện lại trong 10 chu kỳ đầu tiên với tốc độ học tập là $10^{-5}$, giảm giá trị 70 % nếu không có sự cải thiện rõ rệt về hiệu suất xác nhận trong 2 chu kỳ liên tiếp. Trong giai đoạn thứ hai, bộ chuyển đổi liên đoạn và nội bộ thứ hai được cố định và huấn luyện trong 10 chu kỳ tiếp theo với tốc độ học là $0.2 \times 10^{-5}$. Tốc độ học giảm tương tự như cài đặt của 10 chu kỳ đầu tiên. Trong giai đoạn cuối cùng, tất cả các lớp của mô hình được huấn luyện lại với 15 chu kỳ tiếp theo với tốc độ học bằng giá trị cuối cùng của giai đoạn trước. Việc đóng băng các lớp có trật tự đư c thực hiện nhằm giúp các lớp cập nhật tốt các trọng số mới mà vẫn ổn định hiệu quả mô hình đư c huấn luyện trước đó. Quan sát quá trình huấn luyện, kết quả hội tụ tốt ở 10 chu kỷ đầu trong 15 chu kỷ của giai đoạn cuối trong khi các chu kỷ còn lại hội tụ rất ít. Như thể hiện trong các kết quả trong Bảng 2, huấn luyện lại đã chứng minh rằng hiệu quả của việc huấn luyện lại đã làm tăng hiệu suất của mô hình so với đường cơ sở.

Dựa trên mô hình huấn luyện trước, khuôn khổ này đư c triển khai để huấn luyện lại WHAM! và WHAMR! các bộ dữ liệu so sánh với các phương pháp tiên tiến trong những năm gần đây, thể hiện trong Bảng 4. Kết quả thực nghiệm cho thấy mô hình tái tạo đạt kết quả tiên tiến, vượt trội so với các mô hình khác và cũng vượt qua mô hình Sepformer ban đầu. Ngoài ra, để tăng hiệu suất của mô hình, phương pháp tăng dữ liệu DM [24] đư c triển khai. Kết quả cho thấy phương pháp này có hiệu quả và đạt được kết quả vượt trội khi dựa trên mô hình tiên tiến sẵn có.

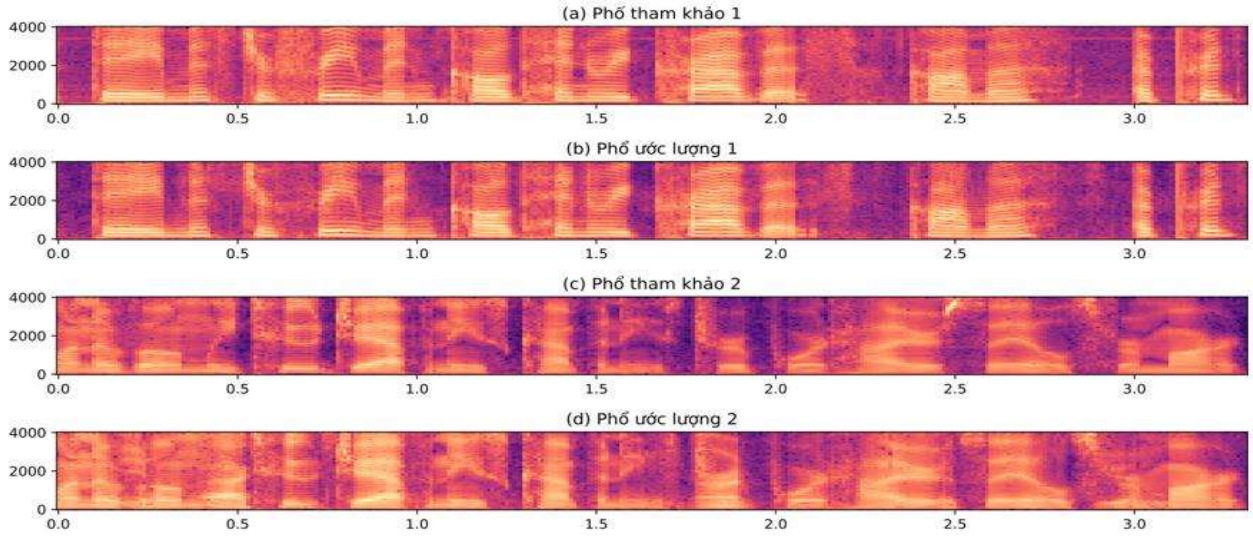
**Bảng 2** So sánh với đường cơ sở trên tập dữ liệu WSJ0-2mix.

| Mô hình | Chu kỳ | Cấu hình | SI-SNRi | SDRi |
|---|---|---|---|---|
| Sepformer [8] | 200 | Đầy đủ | 20,4 dB | 20,5 dB |
| Đào tạo tăng cường + DM | 10 | Đóng băng lớp 1 | 22,5 dB | 22,6 dB |
| Đào tạo tăng cường + DM | 20 | Đóng băng lớp 2 | 22,6 dB | 20,7 dB |
| Đào tạo tăng cường + DM | 35 | Đầy đủ | 22,7 dB | 20,8 dB |

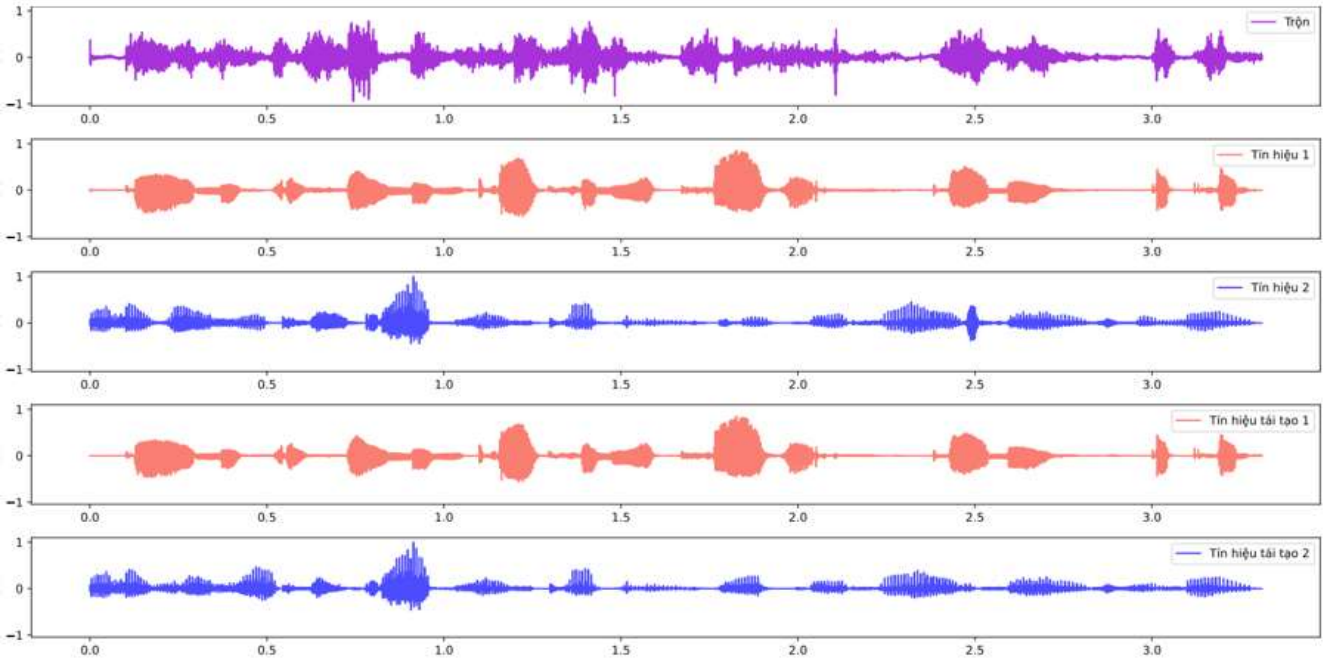
Phổ âm thanh thu được đư c minh họa trên Hình 3. Đánh giá chủ quan khi quan sát, mô hình đề xuất thu đư c phổ phục hồi có độ tương đồng cao với mẫu âm thanh gốc.

**Bảng 3** So sánh trên tập dữ liệu WHAM! và WHAMR!.

| Mô hình | WHAM! | | WHAMR! | |
|---|---|---|---|---|
| | SI-SNRi | SDRi | SI-SNRi | SDRi |
| Conv-TasNet [16] | 12,7 | - | 8,3 | - |
| Wavesplit + DM [24] | 16,0 | 16,5 | 13,2 | 12,2 |
| SepFormer [8] | 14,7 | 15,1 | 11,4 | 10,4 |
| Đào tạo tăng cường | 15,1 | 15,2 | 11,6 | 11,0 |
| Đào tạo tăng cường + DM | 16,5 | 16,7 | 11,9 | 11,4 |




**Hình 3** Phổ tiếng nói ước lượng từ mô hình so với phổ gốc ban đầu (phổ tham khảo) của mô hình tăng cường huấn luyện học tự chú ý.




**Hình 4** Dạng sóng ước lượng so với dạng sóng gốc của mô hình tăng cường huấn luyện học tự chú ý

## 5 Kết luận

Trong nghiên cứu này, cấu trúc tự chú ý được huấn luyện trước có hiệu quả được đề xuất để phân tách giọng nói nhằm mang lại hiệu suất phân tách tiên tiến mới. Theo cách tiếp cận này, hiệu suất mô hình được nâng cao so với mô hình không được huấn luyện trước, xuất phát từ việc sử dụng các tính năng ở giai đoạn đầu được huấn luyện trước trong suốt các giai đoạn sau. Ngoài việc sử dụng tập dữ liệu tiêu chuẩn, các thử nghiệm đầy thử thách cũng được mở rộng, ví dụ: các tập dữ liệu WHAM! và WHAMR! Kết quả cho thấy hiệu quả của cơ chế tự chú ý trong bài toán tách tiếng nói vượt trội so với các phương pháp trước đây và hiệu quả này càng tăng lên khi triển khai mô hình được huấn luyện trước. Điều này cho thấy việc huấn luyện lại của mô hình có hiệu quả trong việc giảm thời gian, tăng hiệu suất và được ứng dụng rộng rãi trong các nhiệm vụ học sâu.

Tài liệu tham khảo

1. Vu, D. Q., & Thu, T. P. T. (2023). Simultaneous context and motion learning in video prediction. *Signal, Image and Video Processing,* 17(8), 3933-3942.
2. Wang, Y., Huang, R., Song, S., Huang, Z., & Huang, G. (2021). Not all images are worth 16 × 16 words: Dynamic transformers for efficient image recognition. *Advances in Neural Information Processing Systems*, 34, 11960-11973.
3. Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12179-12188).
4. Phung, T., Vu, D. Q., Mai-Tan, H., & Nhung, L. T. (2022, November). Deep models for mispronounce prediction for Vietnamese learners of English. In *International Conference on Future Data and Security Engineering* (pp. 682-689). Singapore: Springer Nature Singapore.
5. Vu, D. Q., Le, N., & Wang, J. C. (2021). Teaching yourself: A self-knowledge distillation approach to action recognition. *IEEE Access*, 9, 105711-105723.
6. Dong, L., Xu, S., & Xu, B. (2018, April). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5884-5888). IEEE.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*
8. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 21-25). IEEE.
9. Luo, Y., Chen, Z., & Yoshioka, T. (2020, May). Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 46-50). IEEE.
10. Luo, Y., & Mesgarani, N. (2018, April). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 696-700). IEEE.
11. Tan, H. M., Vu, D. Q., Lee, C. T., Li, Y. H., & Wang, J. C. (2022, May). Selective mutual learning: an efficient approach for single channel speech separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3678-3682). IEEE.
12. Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)* (pp. 31-35). IEEE.
13. Tan, H. M., Liang, K. W., Lee, Y. S., Li, C. T., Li, Y. H., & Wang, J. C. (2022). Speech separation using augmented-discrimination learning on squash-norm embedding vector and node encoder. *IEEE Access,* 10, 102048-102063.
14. Tan, H. M., Liang, K. W., & Wang, J. C. (2023, June). Discriminative vector learning with application to single channel speech separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
15. Tan, H. M., Vu, D. Q., Thi, D. N., & Thu, T. P. T. (2023, December). Voice Separation Using Multi Learning on Squash-Norm Embedding Matrix and Mask. In International Conference on Advances in Information and Communication Technology (pp. 327-333). *Cham: Springer Nature Switzerland.*
16. Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, And Language Processing,* 27(8), 1256-1266.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

18. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8, 726-742.
19. Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.
20. Child, R., Gray, S., Radford, A., Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*
21. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv*:2004.05150.
22. Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv preprint arXiv*:2006.04768.
23. Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*
24. Zeghidour, N., & Grangier, D. (2021). Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2840-2849.

## Enhanced training of self-attention learning models for analysis and segmentation of speech

Ha Minh Tan, Nguyen Kim Quoc
Faculty of Information Technology, Nguyen Tat Thanh University
hmtan@ntt.edu.vn, nkquoc@ntt.edu.vn

**Abstract** This study introduces a novel augmentation approach employing a self-attention model for isolating single-channel speech. Initially, we immobilize all layers within the pre-trained self-attention model. Subsequently, we embark on a three-stage retraining process, incorporating a scheduling mechanism to adapt the learning rate and gradually unlock layers based on a pre-defined schedule. This iterative procedure facilitates the refinement and enhancement of the model's capabilities, leveraging prior knowledge to elevate performance metrics while curtailing training duration and expenses. Notably, this technique not only surpasses conventional methodologies in terms of efficacy but also holds promises for enhancing the performance of pre-existing models. Experimental results underscore the superiority of models trained using this methodology over established techniques in the domain of monosyllabic speech separation across standard datasets.

**Keywords** Pre-trained framework, transformer, self-attention, fine-tuning, temporal masking, voice separation.