TRƯỜNG ĐẠI HỌC THỦY LỢI

ĐỀ TÀI:

PHÁT HIỆN GIỌNG NÓI NGƯỜI VỚI GIỌNG NÓI GIẢ MẠO

Khoa: Công nghệ thông tin

Giáo viên hướng dẫn: Ths. Nguyễn Đắc Phương Thảo

Sinh viên thực hiện: Đinh Quốc Việt, 2251262659

Nguyễn Trường An, 2251262568

Lê Đoàn Dương, 2251262592

MỤC LỤC

MỞ ĐẦU

1. Lý do chọn đề tài

Sự phát triển mạnh mẽ của trí tuệ nhân tạo trong những năm gần đây, đặc biệt là các mô hình tổng hợp và chuyển đổi giọng nói, đã giúp tạo ra tiếng nói nhân tạo ngày càng tự nhiên và khó phân biệt với tiếng nói thật. Công nghệ này mang lại nhiều lợi ích trong các ứng dụng như trợ lý ảo, tổng đài tự động hay giáo dục trực tuyến. Tuy nhiên, bên cạnh những lợi ích đó, tiếng nói do AI sinh ra cũng đang bị lợi dụng cho các mục đích lừa đảo và giả mạo.

Trên thực tế, nhiều báo cáo cho thấy thiệt hại do các hình thức lừa đảo sử dụng công nghệ giả mạo, trong đó có giả mạo giọng nói, đã gây tổn thất hàng chục tỷ đô la mỗi năm trên toàn thế giới. Đã xuất hiện những vụ việc sử dụng giọng nói tổng hợp để giả danh lãnh đạo doanh nghiệp hoặc người thân, lừa đảo chuyển tiền với giá trị lên tới hàng trăm nghìn hoặc hàng triệu đô la cho mỗi vụ. Điều này cho thấy mức độ nguy hiểm ngày càng gia tăng của tiếng nói giả mạo và sự cần thiết của các biện pháp phát hiện hiệu quả.

Trước bối cảnh đó, bài toán phát hiện tiếng nói giả mạo đã thu hút nhiều sự quan tâm trong cộng đồng nghiên cứu, với các mô hình học sâu hiện đại. Tuy nhiên, phần lớn các nghiên cứu hiện nay tập trung vào tiếng Anh và các ngôn ngữ phổ biến khác, trong khi các nghiên cứu dành riêng cho tiếng Việt còn khá hạn chế. Đặc điểm ngữ âm và thanh điệu của tiếng Việt khiến việc áp dụng trực tiếp các mô hình sẵn có chưa đạt hiệu quả cao.

Vì vậy, việc nghiên cứu phát hiện tiếng nói giả mạo cho tiếng Việt là cần thiết cả về mặt khoa học lẫn thực tiễn. Xuất phát từ lý do đó, đề tài "Phát hiện tiếng nói giả mạo do AI sinh ra cho tiếng Việt nói" được lựa chọn nhằm góp phần nâng cao độ an toàn và độ tin cậy của các hệ thống xử lý tiếng nói tại Việt Nam.

2. Mục tiêu nghiên cứu

Mục tiêu chính là nghiên cứu và xây dựng mô hình phát hiện tiếng nói giả mạo do trí tuệ nhân tạo sinh ra đối với tiếng Việt nói dựa trên phương diện xử lý tín hiệu tiếng nói và học sâu. Cụ thể hướng tới việc phân tích, đánh giá các đặc trưng tín hiệu tiếng nói và khảo sát hiệu quả của một số mô hình học máy, học sâu hiện đại trong bài toán phân biệt tiếng nói thật và tiếng nói giả mạo.

3. Nhiệm vụ nghiên cứu

Để đạt được các mục tiêu đã đề ra, đề tài tập trung thực hiện các nhiệm vụ nghiên cứu chính sau:

Nghiên cứu tổng quan về tiếng nói giả mạo do trí tuệ nhân tạo sinh ra và các hướng tiếp cận trong bài toán phát hiện tiếng nói giả mạo.

Thu thập, tìm kiếm dữ liệu giọng nói cho bài toán

Thử nghiệm và đánh giá hiệu quả của mô hình học sâu trong phát hiện tiếng nói giả mạo cho tiếng Việt.

Phân tích, so sánh kết quả thực nghiệm và rút ra nhận xét về ưu, nhược điểm của mô hình trong điều kiện dữ liệu tiếng Việt.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là bài toán phát hiện tiếng nói giả mạo do trí tuệ nhân tạo sinh ra, dựa trên xử lý tín hiệu tiếng nói. Nghiên cứu tập trung phân biệt tiếng nói thật và tiếng nói giả mạo đối với tiếng Việt. Phạm vi nghiên cứu sử dụng các dữ liệu giọng nói liệu tiếng Việt sẵn có và tự thu thập tự tạo, bao gồm cả giọng nam và giọng nữ.

5. Ý nghĩa khoa học và thực tiễn

Về mặt khoa học góp phần làm rõ bài toán phát hiện tiếng nói giả mạo do trí tuệ nhân tạo sinh ra, đồng thời cung cấp các kết quả đánh giá thực nghiệm về hiệu quả của các mô hình học sâu trong điều kiện dữ liệu tiếng Việt.

Về mặt thực tiễn kết quả nghiên cứu có thể được sử dụng làm cơ sở tham khảo cho việc triển khai các hệ thống phát hiện tiếng nói giả mạo trong các ứng dụng sử dụng giọng nói tiếng Việt như xác thực người dùng và các hệ thống an ninh dựa trên giọng nói.

I DỮ LIỆU VÀ TIỀN XỬ LÝ

I.1 Thu thập dữ liệu

Dữ liệu được thu thập từ nhiều nguồn khác nhau nhằm phục vụ cho bài toán phát hiện tiếng nói giả mạo do trí tuệ nhân tạo sinh ra đối với tiếng Việt. Do dữ liệu giọng nói tiếng Việt giả mạo không có nhiều nên dữ liệu giả mạo sẽ được tạo ra từ các mô hình giả mạo giọng nói mạnh có sẵn.

Dữ liệu được lựa chọn đảm bảo đa dạng về giới tính người nói, nội dung phát ngôn và hình thức sinh tiếng nói (giọng nói thật và giọng nói do mô hình trí tuệ nhân tạo tạo ra). Quá trình thu thập tập trung vào việc đảm bảo tính đại diện và khả năng sử dụng cho các thử nghiệm phát hiện tiếng nói giả mạo, đồng thời hạn chế các mẫu có chất lượng âm thanh quá thấp hoặc không phù hợp.

Dữ liệu tiếng nói người thật được thu thập từ nhiều nguồn khác nhau như các bộ dữ liệu công khai trên Kaggle, các video hội thoại và phát biểu trên nền tảng YouTube, cũng như các mẫu giọng nói được thu thập trực tiếp ngoài thực tế.

Nguồn thu thập	Mô tả	Thời lượng
Kaggle	Các bộ dữ liệu tiếng Việt công khai, đa dạng nội dung và người nói	2 tiếng 10 phút
YouTube	Video hội thoại, phát biểu tự nhiên bằng tiếng Việt	26 phút
Thu thập thực tế	Ghi âm trực tiếp giọng nói ngoài đời	1 tiếng

Bảng thống kê thu thập giọng nói người thật

Do nguồn dữ liệu tiếng nói giả mạo cho tiếng Việt còn hạn chế, dữ liệu giả mạo chủ yếu được thu thập từ các video có giọng nói do trí tuệ nhân tạo tạo sinh và nhân bản giọng nói từ dữ liệu tiếng nói người thật đã thu thập bằng các mô hình nền tảng giả giọng nói mạnh mẽ.

Nguồn thu thập	Mô tả	Thời lượng
Video tạo sinh	Giọng nói do trí tuệ nhân tạo sinh ra trong các video trực tuyến	1 tiếng 47 phút
F5TTS	Mô hình tạo sinh	40 phút
Minimax	Nền tảng nhân bản giọng nói người thật được xếp hạng cao	1 tiếng

Bảng thống kê thu thập giọng nói giả mạo

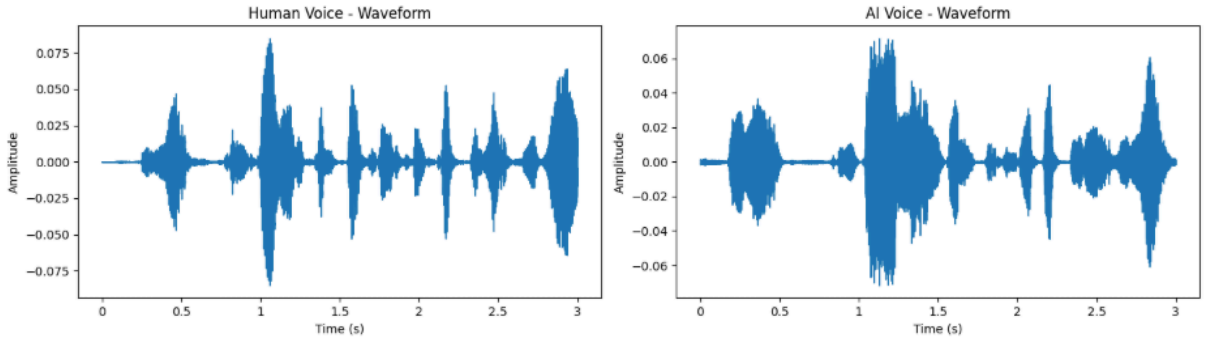
Sau khi thu thập từ nhiều nguồn khác nhau, bộ dữ liệu tiếng nói người thật và tiếng nói giả mạo có thời lượng xấp xỉ 3,5 giờ cho mỗi loại. Với quy mô này, dữ liệu được xem là tương đối đầy đủ và phù hợp để phục vụ trong bài toán phát hiện tiếng nói giả mạo.

I.2 Trực quan hóa dữ liệu

Trực quan hóa dữ liệu âm thanh giúp biểu diễn tín hiệu tiếng nói dưới dạng dạng sóng và phổ tần số, qua đó quan sát đặc điểm năng lượng và cấu trúc của tín hiệu. Bước này hỗ trợ đánh giá chất lượng dữ liệu và nhận diện sự khác biệt giữa giọng nói người thật và giọng nói giả mạo.

Để nhìn trực quan hơn với hai nhãn và sự khác biệt giữa tiếng nói người thật và tiếng nói giả mạo, phần trực quan sẽ sử dụng dữ liệu người thật tự thu thập với giọng nói giả mạo của cùng một người và nội dung nói thông qua Minimax.

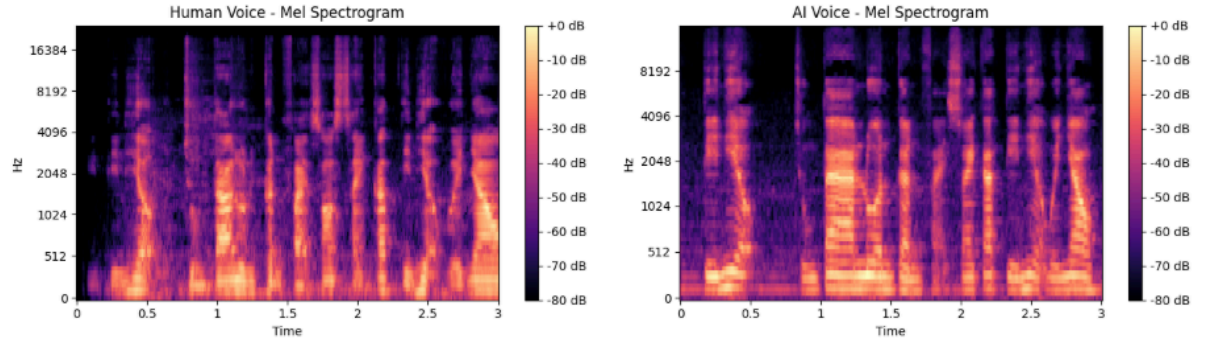
Waveform (dạng sóng) biểu diễn biên độ tín hiệu tiếng nói theo thời gian, cho phép quan sát cấu trúc tổng thể, mức năng lượng và khoảng lặng trong phát ngôn. Qua waveform, có thể thấy tiếng nói người thật thường có dao động tự nhiên và không đều, trong khi tiếng nói giả mạo có xu hướng ổn định và đều hơn. Đây là bước trực quan ban đầu giúp nhận diện sự khác biệt giữa hai loại tiếng nói.



Ảnh: Biểu đồ Waveform trực quan

Qua biểu đồ waveform, có thể thấy giọng nói giả mạo tạo ra có năng lượng khá ổn định, trong khi giọng nói con người thể hiện nhiều dao động tự nhiên và biến thiên nhỏ theo thời gian. Sự khác biệt này là cơ sở để các mô hình học sâu khai thác và phân biệt tiếng nói người thật với tiếng nói giả mạo.

Mel Spectrogram thể hiện phân bố năng lượng của tiếng nói theo thời gian và tần số. Qua trực quan, tiếng nói người thật thường có phổ đa dạng và biến thiên tự nhiên, trong khi tiếng nói giả mạo có xu hướng mượt và đồng đều hơn. Đây là đặc trưng quan trọng để mô hình học sâu phân biệt tiếng nói thật và giả.



Ảnh: Biểu đồ Mel Spectrogram trực quan

Trong khi Waveform chỉ phản ánh mức độ biên độ của tín hiệu, còn Mel Spectrogram cho thấy cấu trúc tần số của âm thanh. Qua đó có thể thấy giọng nói giả mạo thường thiếu các biến thiên tần số tự nhiên và có các dải hài âm rõ, đều hơn so với giọng nói người thật.

Trực quan hóa dữ liệu âm thanh bằng Waveform và Mel Spectrogram giúp quan sát đặc điểm năng lượng và cấu trúc tần số của tiếng nói. Kết quả cho thấy giọng nói người thật có dao động và biến thiên tự nhiên, trong khi giọng nói giả mạo thường ổn định, mượt và đồng đều hơn. Những khác biệt này là cơ sở quan

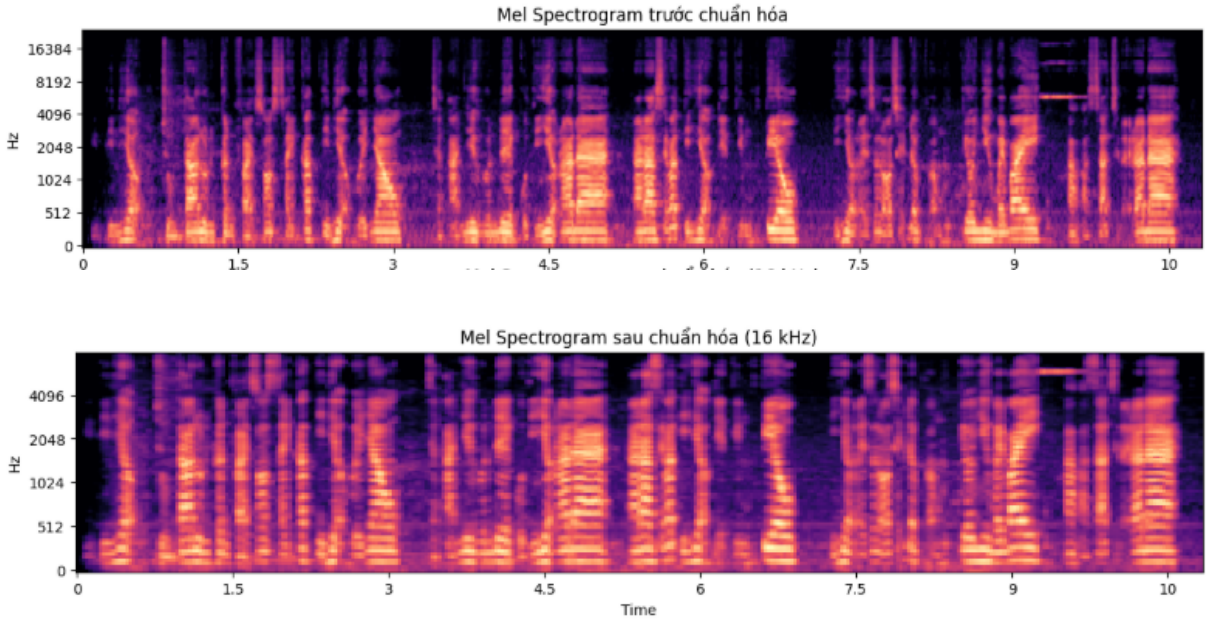
trọng để các mô hình học sâu khai thác đặc trưng và phân biệt tiếng nói người thật với tiếng nói giả mạo.

I.3 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu tiếng nói nhằm chuẩn hóa và làm sạch tín hiệu trước khi đưa vào mô hình học sâu. Các bước thực hiện giúp đồng nhất tần số lấy mẫu, ổn định mức năng lượng, cắt đoạn tiếng nói phù hợp và loại bỏ các phần kém chất lượng. Nhờ đó, dữ liệu đầu vào trở nên ổn định hơn, hỗ trợ mô hình học hiệu quả và phân biệt rõ hơn giữa giọng nói người thật và giọng nói giả mạo.

I.3.1 Chuẩn hóa định dạng và tần số lấy mẫu

Dữ liệu tiếng nói được thu thập từ nhiều nguồn khác nhau nên có sự khác biệt về định dạng, số kênh và tần số lấy mẫu. Trong bước này, tất cả các tệp âm thanh được chuyển về dạng đơn kênh (mono) và chuẩn hóa về cùng tần số lấy mẫu 16 kHz. Việc chuẩn hóa giúp đảm bảo tính đồng nhất của dữ liệu, giảm sai lệch do khác biệt kỹ thuật và phù hợp với yêu cầu đầu vào của các mô hình học sâu trong bài toán phát hiện tiếng nói giả mạo.



Ảnh Mel Spectrogram trước và sau khi chuẩn hóa 16khz

Qua so sánh trước và sau chuẩn hóa, dữ liệu được đưa về tần số lấy mẫu 16kHz, nên phổ tần chỉ còn thể hiện đến khoảng 8kHz. Đây là dải tần quan trọng nhất của tiếng nói con người. Việc chuẩn hóa này giúp dữ liệu đồng nhất,

giảm độ phức tạp cho mô hình và loại bỏ các thành phần tần số cao không cần thiết. Phổ sau chuẩn hóa có năng lượng rõ ràng và ổn định hơn, hỗ trợ mô hình học hiệu quả hơn.

Chuẩn hóa giúp dữ liệu nhất quán, giảm ảnh hưởng thiết bị thu âm và làm nổi bật các đặc trưng hữu ích, tạo điều kiện để mô hình học sâu phân biệt hiệu quả giữa giọng nói người thật và giọng nói giả mạo.

I.3.2 Chuẩn hóa năng lượng tín hiệu

Chuẩn hóa năng lượng (RMS normalization) giúp đưa các tín hiệu tiếng nói về cùng mức cường độ, hạn chế sự khác biệt do thiết bị thu âm hoặc âm lượng phát âm. Qua trực quan waveform, có thể thấy biên độ tín hiệu sau chuẩn hóa trở nên đồng đều hơn trong khi cấu trúc thời gian vẫn được giữ nguyên. Trên Mel Spectrogram, năng lượng được phân bố cân bằng hơn, giúp mô hình học ổn định và hội tụ nhanh hơn.

Trạng thái dữ liệu	**RMS**
Trước chuẩn hóa	0.01038
Sau chuẩn hóa	0.03113

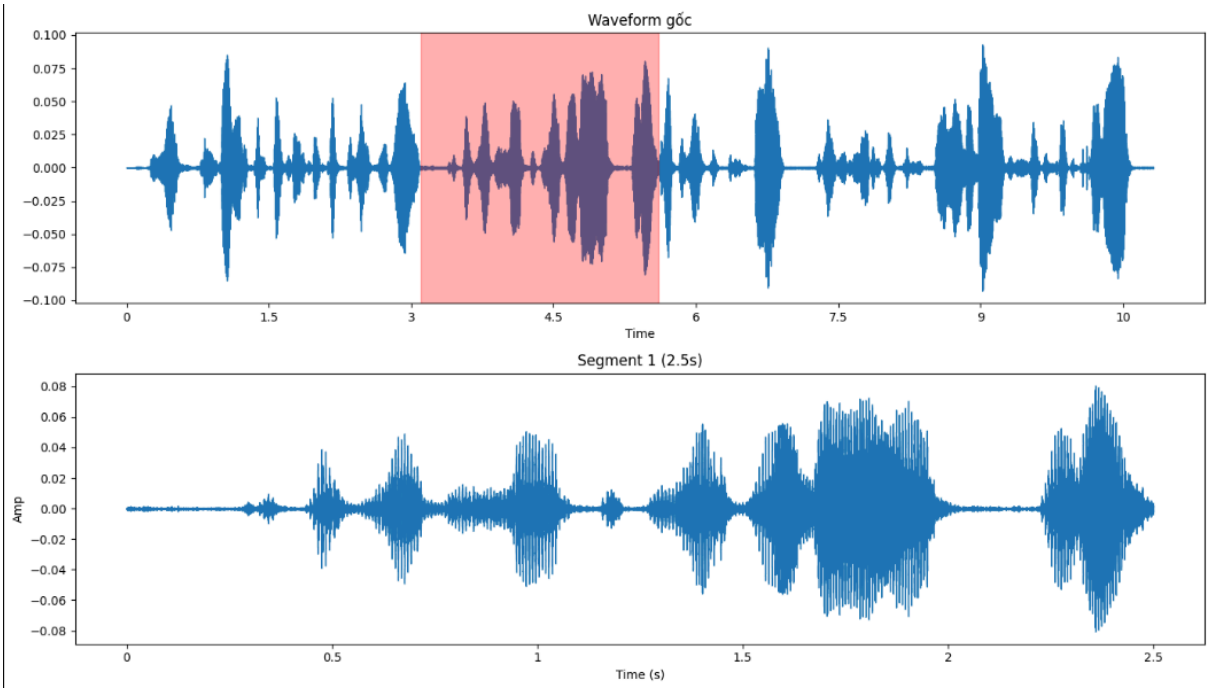
Bảng minh họa RMS

Kết quả cho thấy RMS của tín hiệu sau chuẩn hóa tăng đáng kể so với ban đầu, tuy nhiên được giới hạn bởi hệ số khuếch đại tối đa nhằm tránh làm nổi nhiễu nền. Điều này giúp cân bằng năng lượng giữa các mẫu trong khi vẫn giữ được đặc trưng tự nhiên của tín hiệu tiếng nói.

I.3.3 Cắt đoạn tín hiệu (Segmentation)

Do độ dài các tệp âm thanh không đồng nhất, dữ liệu được chia thành các đoạn ngắn có độ dài cố định để phù hợp với đầu vào của mô hình học sâu. Mỗi đoạn được cắt ngẫu nhiên từ tín hiệu gốc với thời lượng xác định, đồng thời loại bỏ các đoạn có năng lượng quá thấp nhằm tránh khoảng lặng hoặc nhiễu. Cách

tiếp cận này giúp tăng số lượng mẫu huấn luyện, đảm bảo tính đồng nhất về độ dài và cải thiện khả năng học đặc trưng của mô hình.

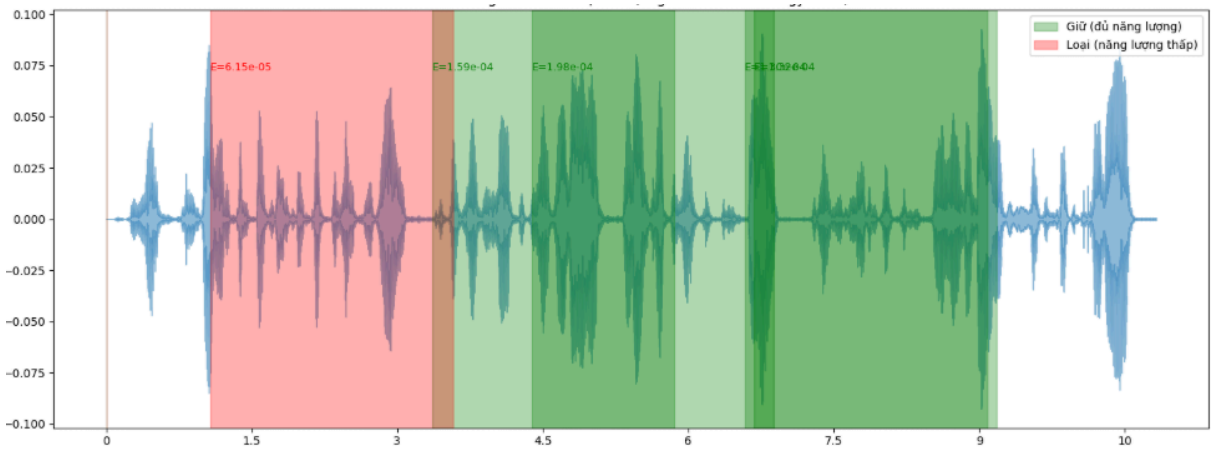


Ảnh cắt đoạn tín hiệu

Waveform gốc thể hiện toàn bộ phát ngôn với độ dài không cố định, bao gồm cả các đoạn có năng lượng thấp và khoảng lặng. Trong khi đó, Segment 1 (2.5s) là một phần tín hiệu ngắn được cắt ra từ waveform gốc, có độ dài cố định và chứa chủ yếu thông tin tiếng nói. Đoạn này giữ được cấu trúc và năng lượng cần thiết cho việc trích xuất đặc trưng, đồng thời loại bỏ các phần dư thừa, giúp dữ liệu phù hợp hơn cho quá trình huấn luyện mô hình.

I.3.4 Lọc đoạn có năng lượng thấp

Sau khi cắt đoạn tín hiệu, các đoạn có mức năng lượng quá thấp được loại bỏ nhằm tránh đưa vào mô hình những khoảng lặng hoặc tín hiệu không chứa thông tin tiếng nói. Việc lọc này giúp giảm nhiễu, đảm bảo các mẫu huấn luyện đều mang nội dung hữu ích, từ đó cải thiện độ ổn định và hiệu quả học của mô hình trong bài toán phân biệt giọng nói người thật và giọng nói giả mạo.



Ảnh lọc đoạn có năng lượng thấp

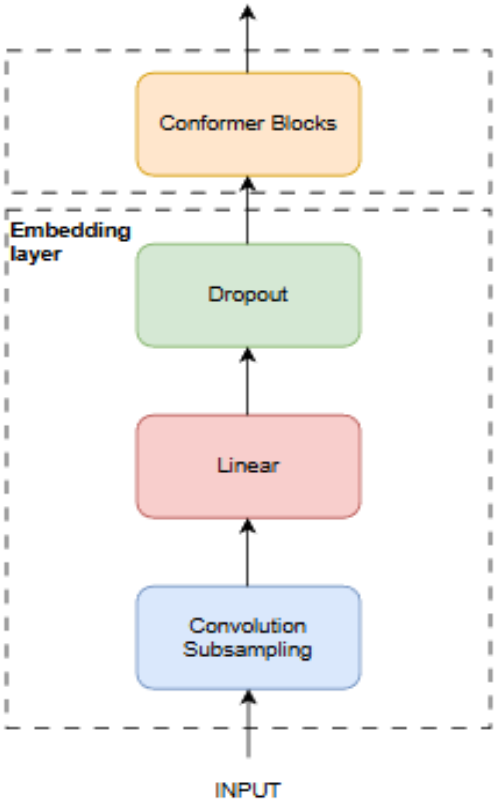
Sau khi lọc đoạn năng lượng thấp, dữ liệu chỉ còn các phần chứa tiếng nói rõ ràng, loại bỏ khoảng lặng và nhiễu nền. Điều này giúp giữ đặc trưng giọng nói, đồng nhất năng lượng các segment và tăng hiệu quả huấn luyện mô hình.

II MÔ HÌNH VÀ TRIỂN KHAI

II.1 Mô hình Conformer

Đầu vào là phổ(dạng vecto) Mel Spectrogram.

Conformer là một mô hình học sâu kết hợp giữa CNN và Transformer, được thiết kế đặc biệt cho các tác vụ xử lý chuỗi tín hiệu âm thanh như nhận dạng tiếng nói. Phần CNN giúp trích xuất các đặc trưng cục bộ từ phổ âm thanh, nắm bắt các mẫu tần số chi tiết, trong khi phần Transformer dùng cơ chế self-attention để mô hình hóa các mối quan hệ dài hạn và cấu trúc tổng thể của chuỗi âm thanh. Sự kết hợp này cho phép Conformer vừa hiểu được chi tiết âm sắc, vừa nắm bối cảnh dài, nhờ đó cải thiện hiệu quả nhận dạng tiếng nói và phân biệt giọng thật với giọng giả mạo so với các mô hình Transformer thuần túy.

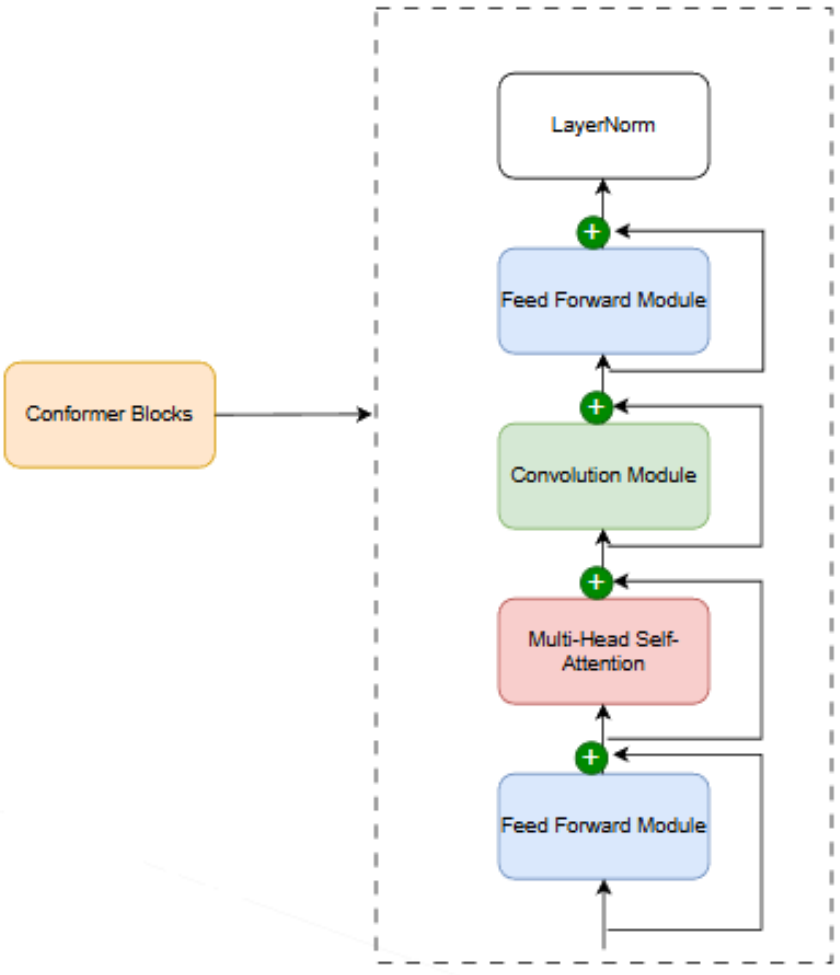


Ảnh cấu trúc Conformer tổng quan

Trước tiên Embedding layer là khối biến đổi đặc trưng đầu vào (log-Mel spectrogram) thành dạng biểu diễn phù hợp để đưa vào Conformer Encoder. Trong bài toán xử lý tiếng nói, embedding không phải là word embedding, mà

là feature embedding theo thời gian. Cụ thể, embedding layer gồm Convolution Subsampling + Linear + Dropout. Convolution Subsampling giúp giảm độ dài chuỗi theo thời gian, mở rộng receptive field ban đầu và giảm chi phí tính toán cho self-attention. Lớp Linear chiếu đặc trưng sau convolution về không gian ẩn cố định mà Conformer block yêu cầu. Dropout được sử dụng để regularization, giúp mô hình tổng quát tốt hơn và tránh overfitting. Nhờ embedding layer, tín hiệu speech thô được chuẩn hóa về kích thước và tốc độ khung, tạo nền tảng ổn định để Conformer Encoder học các đặc trưng ngữ cảnh dài và cục bộ hiệu quả.

Đi sâu vào trọng tâm của mô hình Conformer Blocks là thành phần trung tâm của kiến trúc Conformer, nhằm khắc phục hạn chế của Transformer truyền thống trong xử lý tín hiệu tiếng nói. Trong khi self-attention có khả năng mô hình hóa quan hệ dài hạn hiệu quả, nó lại thiếu khả năng khai thác đặc trưng cục bộ quan trọng trong speech. Conformer Blocks được thiết kế để kết hợp self-attention và convolution trong cùng một khối, giúp mô hình học được cả ngữ cảnh dài hạn và đặc trưng ngắn hạn của tín hiệu âm thanh.



Ảnh cấu trúc Conformer Blocks

Trong bài toán phân biệt giọng nói người thật và giọng giả, mô hình không cần sinh chuỗi đầu ra theo thời gian mà chỉ yêu cầu trích xuất biểu diễn đặc trưng toàn cục của tín hiệu để phục vụ phân loại. Do đó, Conformer được sử dụng dưới dạng encoder-only, tập trung vào việc học các quan hệ ngắn hạn và dài hạn trong tín hiệu speech thông qua cơ chế self-attention kết hợp với convolution. Việc không sử dụng decoder giúp mô hình đơn giản hơn, giảm chi phí tính toán nhưng vẫn đảm bảo hiệu quả cho bài toán chống giả mạo giọng nói.

Conformer Block là khối xử lý cốt lõi của mô hình Conformer, kết hợp Self-Attention: học ngữ cảnh dài, Convolution: học đặc trưng cục bộ, Feed Forward Network (FFN): tăng khả năng biểu diễn phi tuyến

Feed Forward Module trong Conformer thực hiện biến đổi phi tuyến để tăng khả năng biểu diễn đặc trưng. Cấu trúc gồm hai lớp tuyến tính và hàm kích hoạt Swish (SiLU):

$$FFN(x) = W2 \cdot Swish(W1x)$$

Khác với Transformer truyền thống, Conformer chỉ cộng một nửa đầu ra FFN vào kết nối dư. Việc chia FFN thành hai nửa giúp cân bằng ảnh hưởng của FFN với các mô-đun attention và convolution, đồng thời cải thiện sự ổn định trong quá trình huấn luyện.

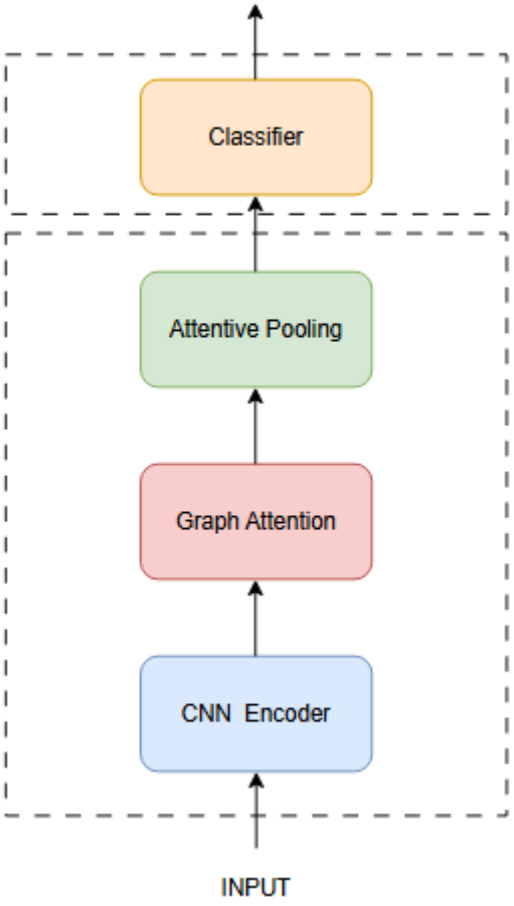
Mô-đun Multi-Head Self-Attention cho phép mỗi frame âm thanh học quan hệ với toàn bộ các frame còn lại trong chuỗi. Trong Conformer, attention thường kết hợp relative positional encoding, giúp mô hình nhận biết vị trí tương đối giữa các frame, phù hợp hơn với tín hiệu giọng nói liên tục.

Convolution Module giúp mô hình học các đặc trưng cục bộ theo thời gian, như phoneme và các artefact ngắn hạn. Module này thường bao gồm convolution 1×1, GLU, depthwise convolution với kernel lớn, batch normalization và hàm kích hoạt Swish.

Cuối mỗi Conformer Block, Layer Normalization được sử dụng để chuẩn hóa đầu ra. Layer Normalization giúp ổn định gradient, giảm hiện tượng bùng nổ hoặc tiêu biến gradient và cải thiện khả năng hội tụ của mô hình khi xử lý chuỗi dài.

II.2 Mô hình ASSIST

AASIST là một mô hình học sâu được thiết kế chuyên biệt cho bài toán phát hiện giả mạo giọng nói. Mô hình kết hợp CNN và Graph Attention Network (GAT) nhằm khai thác hiệu quả cả đặc trưng cục bộ và quan hệ toàn cục trong tín hiệu âm thanh.



Ảnh cấu trúc mô hình AASIST

Phần CNN đóng vai trò encoder ở đầu vào, giúp trích xuất các đặc trưng cục bộ từ phổ thời gian–tần số của tín hiệu giọng nói, qua đó nắm bắt các mẫu tần số chi tiết và các đặc trưng do giọng nói tổng hợp tạo ra.

Tiếp theo, Graph Attention được sử dụng để mô hình hóa mối quan hệ giữa các vùng phổ khác nhau, cho phép phát hiện các mẫu lặp bất thường và sự thiếu biến thiên tự nhiên – những dấu hiệu điển hình của giọng nói giả mạo. Sự kết hợp này giúp AASIST học được biểu diễn giàu thông tin hơn, từ đó nâng cao

hiệu quả phân biệt giữa giọng nói người thật và giọng nói do AI tạo ra so với các mô hình chỉ dựa trên CNN hoặc Transformer thuần túy.

Sau giai đoạn học quan hệ toàn cục bằng Graph Attention Network, mô hình AASIST sử dụng Attentive Pooling để tổng hợp các đặc trưng quan trọng của toàn bộ tín hiệu giọng nói. Khác với các phương pháp pooling thông thường như average hoặc max pooling, Attentive Pooling cho phép mô hình tự động gán trọng số lớn hơn cho những vùng phổ chứa nhiều thông tin phân biệt, đặc biệt là các đặc trưng của giọng nói giả mạo. Nhờ đó, biểu diễn cuối cùng của mỗi mẫu âm thanh tập trung vào các đặc trưng quan trọng nhất cho bài toán phát hiện giả mạo.

Biểu diễn sau khi được tổng hợp sẽ được đưa vào Classifier, thường bao gồm một hoặc nhiều lớp fully connected, để thực hiện phân loại. Bộ phân loại này học cách ánh xạ embedding của toàn bộ tín hiệu giọng nói thành nhãn đầu ra, tương ứng với giọng nói người thật hoặc giọng nói giả mạo.

II.3 Tối ưu mô hình

Quá trình tối ưu mô hình Conformer cho bài toán phân biệt giọng nói người thật và giọng nói giả mạo được thực hiện thông qua việc kết hợp nhiều kỹ thuật huấn luyện và điều chỉnh siêu tham số nhằm nâng cao khả năng tổng quát hóa và hạn chế hiện tượng quá khớp.

Thành phần	Thiết lập	Vai trò
Hàm mất mát	BCEWithLogitsLoss	Phân loại nhị phân người thật / giọng giả
Bộ tối ưu	Adam (lr = 1e-4)	Huấn luyện ổn định, hội tụ nhanh

Điều chỉnh LR	ReduceLROnPlateau	Giảm learning rate khi mô hình không cải thiện
Early Stopping	Patience = 6	Tránh overfitting
Dropout	0.3	Giảm học thuộc dữ liệu
Tiêu chí chọn mô hình	EER thấp nhất	Phù hợp bài toán chống giả mạo

Trong quá trình huấn luyện, mô hình sử dụng BCEWithLogitsLoss làm hàm mất mát để xử lý bài toán phân loại nhị phân giữa giọng nói người thật và giọng giả mạo. Thuật toán tối ưu Adam với learning rate $1\times10-4$ được lựa chọn nhằm đảm bảo quá trình huấn luyện ổn định và hội tụ nhanh. Để thích nghi với quá trình học, ReduceLROnPlateau được áp dụng để tự động giảm learning rate khi hiệu năng trên tập validation không còn cải thiện.

Bên cạnh đó, kỹ thuật Early Stopping với ngưỡng patience bằng 6 giúp dừng huấn luyện sớm, hạn chế hiện tượng overfitting. Dropout với tỷ lệ 0.3 được sử dụng trong các lớp fully-connected và Conformer block nhằm tăng khả năng tổng quát hóa của mô hình. Cuối cùng, mô hình được lựa chọn dựa trên giá trị EER thấp nhất, phù hợp với yêu cầu đánh giá trong bài toán phát hiện giọng nói giả mạo.

II.4 Huấn luyện mô hình

II.4.1 Hệ số đánh giá mô hình

Trong bài toán phân biệt giọng nói người thật và giọng nói giả mạo, đầu ra của mô hình thường là xác suất hoặc điểm tin cậy. Vì vậy, các chỉ số đánh giá dựa trên ngưỡng quyết định như AUC và EER được sử dụng nhằm phản ánh chính xác năng lực phân biệt của mô hình.

1) AUC (Area Under the ROC Curve) mô tả mối quan hệ giữa: TPR (True Positive Rate) – tỷ lệ phát hiện đúng giọng giả mạo, FPR (False Positive Rate) – tỷ lệ chấp nhận nhầm giọng giả mạo.

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

AUC là diện tích dưới đường cong ROC: $AUC = \int_0^1 TPR(FPR)d(FPR)$

TP: số mẫu giả mạo được phát hiện đúng,

FP: số mẫu người thật bị nhận nhầm là giả mạo,

TN: số mẫu người thật được nhận diện đúng,

FN: số mẫu giả mạo bị bỏ sót.

AUC phản ánh xác suất mà mô hình gán điểm cao hơn cho một mẫu giả mạo so với một mẫu người thật được chọn ngẫu nhiên. AUC càng lớn thì khả năng phân biệt hai lớp của mô hình càng tốt.

2) EER là điểm tại đó tỷ lệ chấp nhận sai (FAR) bằng tỷ lệ từ chối sai (FRR):

FAR (False Acceptance Rate) – Tỷ lệ chấp nhận nhầm giọng giả mạo:

$$FAR = \frac{FP}{FP + TN}$$

FRR (False Rejection Rate) – Tỷ lệ từ chối nhầm giọng nói người thật:

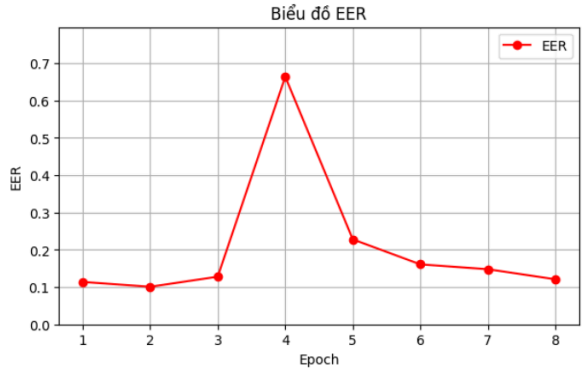
$$FRR = \frac{FN}{FN + TP}$$

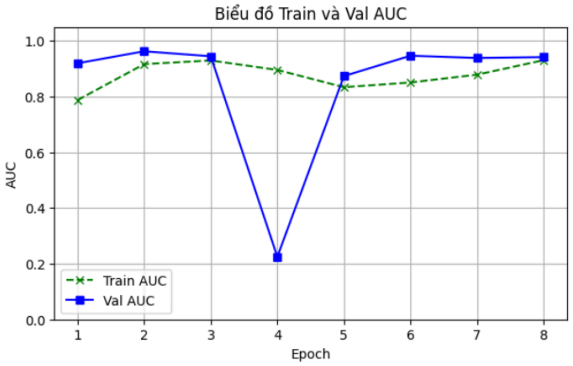
EER được xác định tại ngưỡng một ngưỡng sao cho: $EER = FAR = FRR$

EER là điểm cân bằng sai số giữa chấp nhận nhầm và từ chối nhầm. EER càng thấp chứng tỏ mô hình càng chính xác và ổn định.

II.4.2 Kết quả huấn luyện Conformer

Quá trình huấn luyện mô hình Conformer được đánh giá thông qua ba chỉ số chính gồm AUC trên tập huấn luyện (Train AUC), AUC trên tập kiểm tra (Validation AUC) và Equal Error Rate (EER)





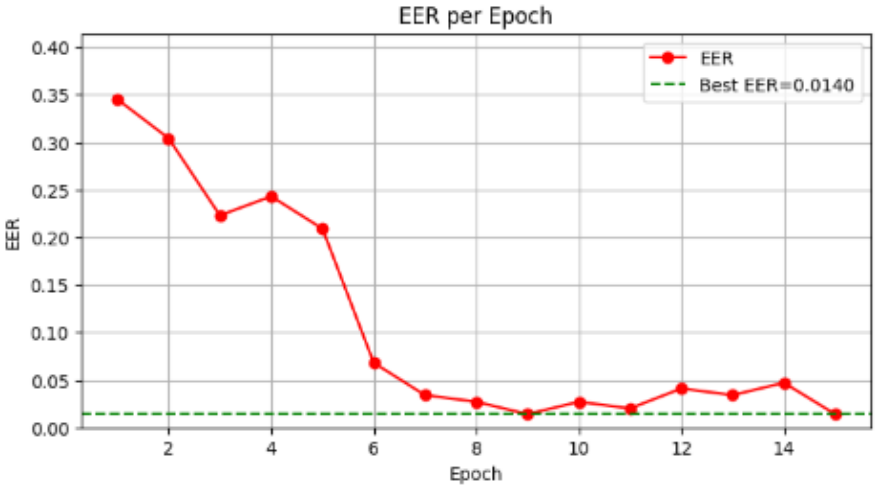
Trong các epoch đầu, mô hình cho thấy khả năng học đặc trưng khá tốt khi Train AUC tăng nhanh từ 0.788 lên 0.930, cho thấy Conformer có khả năng khai thác hiệu quả các đặc trưng phổ Mel trong bài toán phân biệt giọng nói người thật và giọng nói giả mạo. Trên tập validation, AUC đạt giá trị cao ngay từ sớm (0.920 – 0.963), chứng tỏ mô hình có khả năng tổng quát hóa tốt trong giai đoạn đầu huấn luyện.

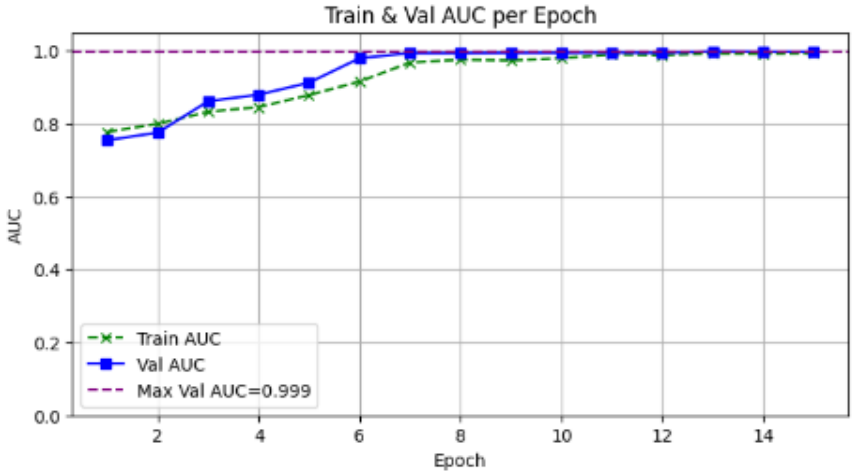
Chỉ số EER phản ánh trực tiếp hiệu năng của hệ thống trong bối cảnh bài toán an ninh sinh trắc học. Kết quả cho thấy: EER thấp nhất đạt 0.101, tương ứng với epoch có Validation AUC cao nhất (0.963), Trong phần lớn các epoch ổn định, EER dao động trong khoảng 0.11 – 0.16, cho thấy khả năng phân biệt tương đối tốt giữa hai lớp.

Nhìn chung, mô hình Conformer đạt được hiệu năng cao và ổn định sau giai đoạn đầu huấn luyện. Dù đã xảy ra overfitting tạm thời sau đó đã ổn định dần nhưng do Earlystopping dừng sau 6 epoch không cải thiện nên nếu chạy tiếp có thể sẽ cho mô hình tốt hơn. Nhưng tổng quan kết quả là chấp nhận được.

II.4.2 Kết quả huấn luyện AASIST

Quá trình huấn luyện mô hình AASIST được đánh giá thông qua ba chỉ số chính gồm AUC trên tập huấn luyện (Train AUC), AUC trên tập validation (Validation AUC) và Equal Error Rate (EER). Các chỉ số này phản ánh lần lượt khả năng học đặc trưng của mô hình, mức độ tổng quát hóa và hiệu năng phân biệt trong bối cảnh bài toán phát hiện giả mạo giọng nói.





Trong các epoch đầu, mô hình AASIST cho thấy khả năng học đặc trưng hiệu quả khi Train AUC tăng từ 0.778 lên 0.916 trong 6 epoch đầu, cho thấy mô hình nhanh chóng khai thác được các đặc trưng phổ Mel phục vụ phân biệt giọng nói người thật và giọng nói giả mạo. Từ epoch 7 trở đi, Train AUC duy trì ở mức rất cao (trên 0.96), phản ánh năng lực biểu diễn mạnh mẽ của kiến trúc CNN kết hợp Graph Attention.

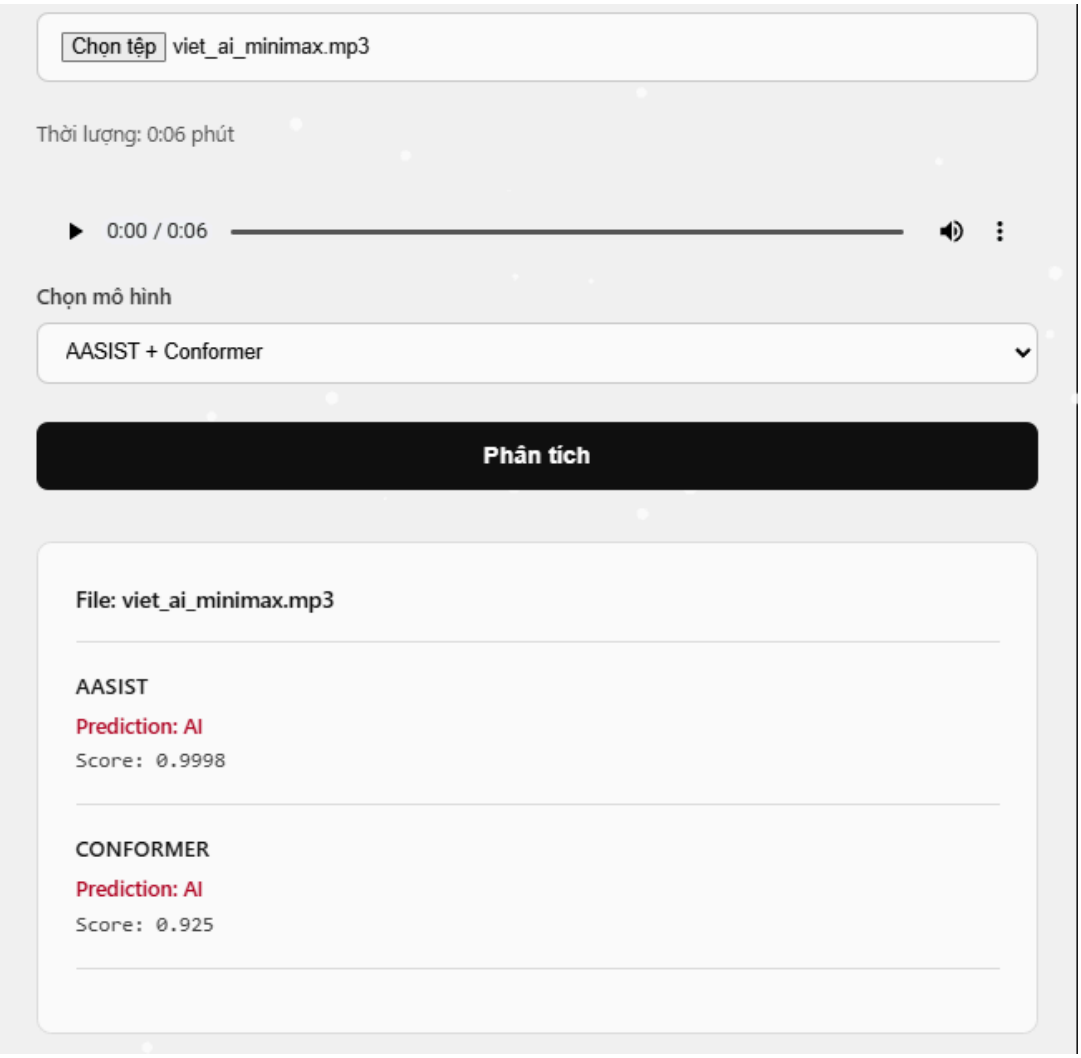
Trên tập validation, Validation AUC tăng nhanh và ổn định, từ 0.755 ở epoch đầu lên 0.980 tại epoch 6, và tiếp tục đạt các giá trị rất cao trong các epoch sau (0.994–0.999). Kết quả này cho thấy mô hình AASIST có khả năng tổng quát hóa tốt và hạn chế được hiện tượng overfitting so với các mô hình CNN hoặc Transformer thuần túy.

Chỉ số EER giảm mạnh trong quá trình huấn luyện, từ 0.345 ban đầu xuống 0.068 tại epoch 6 và đạt mức thấp nhất 0.014 tại các epoch 9 và 15. Trong các epoch ổn định, EER duy trì ở mức thấp (0.014–0.047), cho thấy mô hình

AASIST phân biệt hiệu quả giữa hai lớp giọng nói và đáp ứng tốt yêu cầu của bài toán phát hiện giả mạo giọng nói.

II.4.3 Chạy thử nghiệm

Cuối cùng để kiểm tra độ hiệu quả của mô hình, chúng ta sẽ thực nghiệm trên thực tế với các dữ liệu, file âm thanh mới tạo ra. Chúng ta sẽ đánh giá và đưa ra kết quả theo Score: là xác suất dự đoán giọng nói là giả mạo. Tức là score càng cao thì sẽ là giọng nói giả mạo, còn score càng thấp sẽ là người thật nói.

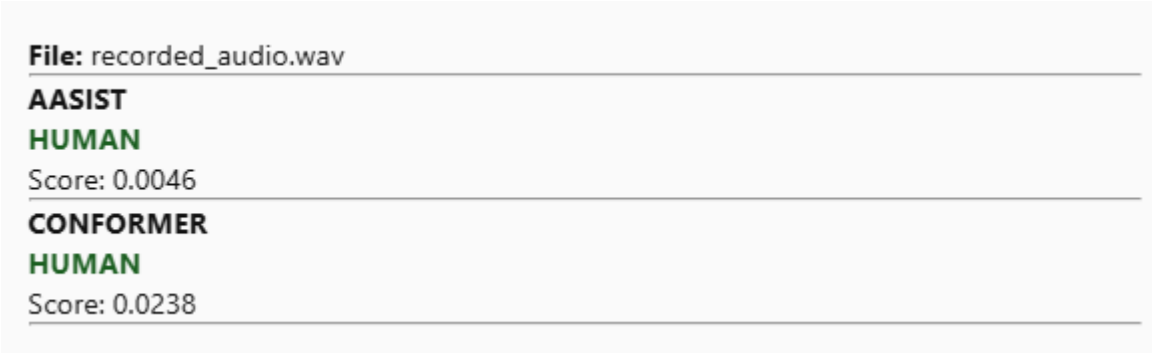


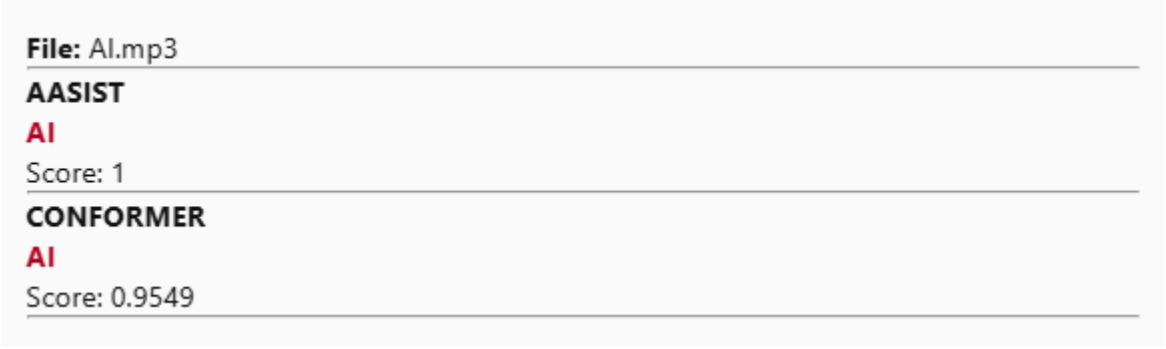
Ảnh kết quả chạy thử nghiệm giọng nói giả mạo

Đối với file viet_ai_minimax.mp3, cả hai mô hình AASIST và Conformer đều dự đoán là giọng nói AI, cho thấy sự thống nhất cao giữa các phương pháp. Trong đó, AASIST đạt điểm rất cao (0.9998), phản ánh mức độ tự tin gần như tuyệt đối trong việc phát hiện dấu hiệu giả mạo. Conformer cũng cho kết quả

nhất quán với score 0.925, tuy thấp hơn AASIST nhưng vẫn vượt xa ngưỡng phân loại, khẳng định khả năng nhận diện giọng AI hiệu quả. Nhìn chung, kết quả cho thấy mẫu âm thanh mang đặc trưng giả mạo rõ rệt và AASIST tỏ ra nhạy hơn trong việc phát hiện các artefact của giọng tổng hợp.

Sau đây là một số kết quả khác được chạy ra:





Kết quả cho thấy hệ thống phân loại hoạt động ổn định và nhất quán: các file ghi âm thật đều được cả hai mô hình (AASIST và Conformer) nhận diện là giọng người với điểm số thấp, thể hiện độ tin cậy cao, trong khi file giọng tổng hợp được xác định rõ ràng là AI với điểm số rất cao. Điều này cho thấy mô hình có khả năng phân biệt tốt giữa giọng người thật và giọng nhân tạo trong các mẫu thử nghiệm.

III ĐÁNH GIÁ, KẾT LUẬN

Hai mô hình học sâu tiêu biểu cho bài toán phát hiện giả mạo giọng nói là Conformer và AASIST đã được trình bày, phân tích và triển khai một cách chi tiết từ kiến trúc, cơ chế hoạt động cho đến quá trình huấn luyện và đánh giá.

Quá trình tối ưu và huấn luyện được thiết kế hợp lý với các kỹ thuật phổ biến nhưng hiệu quả như Adam optimizer, ReduceLROnPlateau, Early Stopping và tiêu chí chọn mô hình dựa trên EER – chỉ số đặc trưng cho các hệ thống an ninh sinh trắc học. Việc sử dụng các chỉ số AUC và EER giúp đánh giá toàn diện cả khả năng phân biệt và độ tin cậy của mô hình trong các ngưỡng quyết định khác nhau.

Các kết quả thử nghiệm thực tế trên các file âm thanh mới cho thấy hai mô hình hoạt động nhất quán và ổn định. Các mẫu giọng nói người thật đều được nhận diện chính xác với score thấp, trong khi các mẫu giọng nói tổng hợp được phát hiện rõ ràng với score rất cao. Đặc biệt, AASIST thể hiện mức độ tự tin cao hơn trong việc phát hiện giọng AI, phù hợp với kết quả huấn luyện và thiết kế kiến trúc của mô hình.

Conformer là một lựa chọn mạnh và linh hoạt, có khả năng tổng quát hóa tốt nhờ kết hợp CNN và self-attention, phù hợp cho các hệ thống cần cân bằng giữa hiệu năng và độ phức tạp. Tuy nhiên, AASIST tỏ ra vượt trội hơn về độ chính xác và độ ổn định, đặc biệt trong việc phát hiện các artefact tinh vi của giọng nói tổng hợp, nhờ tận dụng Graph Attention để mô hình hóa quan hệ toàn cục trong phổ âm thanh.

Nhìn chung, kết quả của Chương II khẳng định tính khả thi và hiệu quả của các mô hình học sâu trong bài toán chống giả mạo giọng nói. Đây là cơ sở vững chắc để tiếp tục mở rộng nghiên cứu, cải tiến mô hình hoặc triển khai hệ thống phát hiện giọng nói giả mạo trong các ứng dụng thực tế như xác thực sinh trắc học, an ninh thông tin và phòng chống gian lận.

TÀI LIỆU THAM KHẢO

[1] Sooftware, “Conformer: Convolution-augmented Transformer for Speech Recognition” : github.com/sooftware/conformer

[2] Hugging Face,“Wav2Vec2-Conformer Documentation”,Transformers library: huggingface.co/docs/transformers/model_doc/wav2vec2-conformer

[3] W. Gulati, J. Qin, C. Chiu, N. Gao, M. Duarte, Y. Wang, Z. Zhang, J. Han, S. A. Raffel, N. González, et al., “Conformer: Convolution-augmented Transformer for Speech Recognition,”. Interspeech 2020, pp. 5036–5040: isca-archive.org/interspeech_2020/gulati20_interspeech.pdf

[4] Clova AI Research, “AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks”: github.com/clovaai/aasist