Báo Cáo Khoa Học: Phân Biệt Giọng Hát Do AI Tạo Ra Hay Con Người Tạo Ra

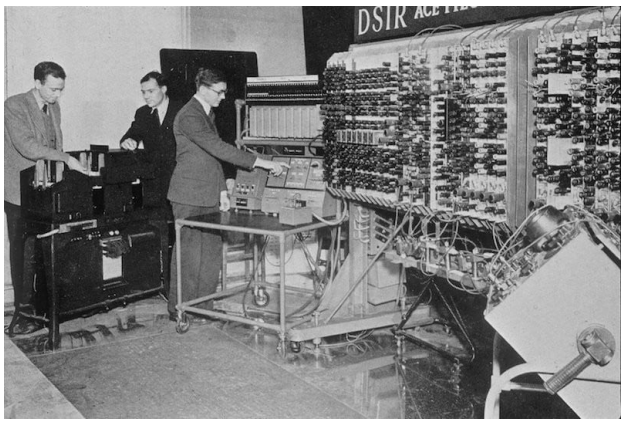
Tống Duy Tân - 22022538 Nguyễn Quốc Tuấn - 22022553
Viện trí tuệ nhân tạo, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Tóm tắt - Trí tuệ nhân tạo (AI) đang được ứng dụng rộng rãi trong lĩnh vực âm nhạc, từ sáng tác, phân phối và tiêu thụ cho đến việc hoàn thành vòng lặp phản hồi để cung cấp thông tin cho việc sáng tác nhạc. AI đang dần được coi như một cộng sự mạnh mẽ của nghệ sĩ khi có thể giúp nghệ sĩ vượt qua những ranh giới sáng tạo của chính mình, nhưng cũng đang dần trở thành mối đe dọa đối với chính nghệ sĩ đó cũng như sự phát triển của nghệ thuật âm nhạc. Bài báo này sẽ đưa ra một vài giải pháp để có thể nhận dạng giọng hát do AI tạo ra để ngăn chặn những vấn đề nhức nhối khi AI có thể mạo danh giọng hát của những người nổi tiếng, rồi có thể dẫn đến các vấn đề xã hội khác.

Keyword: Voice Fingerprint, Singer Matching, AI Detection.

1. ĐẶT VẤN ĐỀ

Ý tưởng sử dụng trí tuệ nhân tạo để tạo ra âm nhạc đã có từ nhiều thập kỷ trước, với những thử nghiệm ban đầu về các thuật toán có từ những năm 1950 và 60[2].



Hình 1: Ý tưởng sử dụng AI để tạo ra âm nhạc từ những năm 1950 và 60.

Âm nhạc AI đã trải qua một cuộc lột xác ngoạn mục, từ những bước chập chững với hệ thống quy tắc đơn giản đến giai đoạn bùng nổ nhờ mạng lưới thần kinh và học sâu. Giờ đây, AI có thể phân tích và học hỏi từ lượng dữ liệu âm nhạc và giọng nói khổng lồ, từ đó tạo ra những bản nhạc phức tạp, tinh tế và mang đậm dấu ấn sáng tạo. Sự phát triển của AI trong lĩnh vực âm nhạc đã dẫn đến sự ra đời của nhiều công ty và công cụ AI khiến bất kỳ ai, bất kể trình độ âm nhạc hay kỹ năng kỹ thuật nào, cũng có thể dễ dàng tạo ra âm nhạc với giọng hát theo ý muốn. Điều này góp phần thúc đẩy sự sáng tạo và đa dạng trong thế giới âm nhạc, và mở ra những tiềm năng to lớn cho tương lai.

Tuy vậy, việc AI có thể tạo ra âm nhạc và ai cũng có thể sử dụng nó tạo ra 1 cuộc tranh cãi quy mô toàn cầu. Những bài hát huyền thoại như “Song 2” của Blur với giọng hát được thay thế bằng Kurt Cobain AI tạo ra , "Piano Man" của Billy Joel được hát bởi Paul McCartney do AI tạo nên và bài "Rickroll" với giọng hát của Rick Astley được thay thế bằng Michael Jackson AI,... [4] Gần đây, vào tháng 4/2023, một người dùng có tên Ghostwriter977 đã phát hành bài hát "Heart on My Sleeve" của Drake và The Weeknd được AI tạo ra và trình bày trên TikTok, bài hát này đã thu được hàng triệu lượt xem trước khi bị xóa. Rõ ràng không chỉ chúng tôi mà cả thế giới đều coi rằng công nghệ AI đã đạt đến mức chất lượng và khả năng tiếp cận vừa thú vị vừa đáng sợ đối với nhiều người sáng tạo và người hâm mộ âm nhạc.

Khi những bản nhạc này được tải lên các nền tảng, ngành công nghiệp âm nhạc đang phải vật lộn với những câu hỏi lớn mang tính hiện sinh, bao gồm âm nhạc sẽ phát triển như thế nào khi khả năng của AI phát triển, liệu sự chuyển đổi này sẽ là điều tốt hay xấu khi cách người sáng tạo có thể kiếm tiền trong đó khiến một ngành công

nghiệp đang thay đổi. [4] Nhiều người phân vân rằng việc sử dụng AI để tạo ra âm nhạc sử dụng chất giọng nghệ sĩ nổi tiếng có bị dính bản quyền.

VẤN ĐỀ BẢN QUYỀN [8]: Bằng việc tận dụng các thuật toán học máy như GAN, AI có thể tạo ra các tác phẩm âm nhạc, khiến cho ranh giới giữa trí tuệ nhân tạo và trí tuệ con người trở nên gần hơn. Điều này dẫn đến hai vấn đề lớn về bản quyền. Thứ nhất: Việc sao chép, lưu trữ và chỉnh sửa các tác phẩm âm nhạc có bản quyền đặt ra các vấn đề về bản quyền. Tuy vậy tác tác phẩm âm nhạc thuộc phạm vi công cộng thì không ảnh hưởng đến quá trình sáng tạo của AI. Thứ hai: Bản quyền của tác phẩm âm nhạc cuối cùng do AI tạo ra trở nên rất phức tạp. Liệu tác phẩm này được coi là một "tác phẩm" được bảo hộ bản quyền hay rơi vào phạm vi công cộng? Hiện tại, AI chưa thực sự tự chủ, vẫn cần sự tham gia của con người trong việc thiết lập cấu trúc. Mặc dù hiện tại không có quy định cụ thể về việc sử dụng AI cho mục đích âm nhạc, nhưng có những ngoại lệ có thể áp dụng, đặc biệt là ngoại lệ đối với hoạt động khai thác dữ liệu lớn (TDM). [7]

VẤN ĐỀ ĐẠO ĐỨC [11]: Việc sử dụng AI như một công cụ không chỉ gây ra những vấn đề về bản quyền mà còn tạo ra những vấn đề về đạo đức khi sử dụng nó. Đầu tiên, vấn đề thay thế việc làm trong lĩnh vực âm nhạc: Sự phát triển của AI có thể dẫn đến việc máy móc thay thế con người trong sáng tác nhạc. Điều này đặt ra lo ngại về thất nghiệp cho các nhạc sĩ. Tiếp theo, vấn đề quyền tác giả của nhạc do AI sáng tác: AI có thể tạo ra những tác phẩm âm nhạc sáng tạo nhưng trong sản phẩm lại có đóng góp bằng giọng nói hoặc là giọng văn của các nghệ sĩ ngoài đời. Khi nhiều người có thể sử dụng AI để sao chép thì ca sĩ sẽ dần mất đi bản quyền và có thể ảnh hưởng đến nhiều yếu tố khác của nghệ sĩ (tâm trạng, sự nghiệp, công sức,...) Cuối cùng, vấn đề gây rối trên mạng: Như đã đề cập ở trên, bài hát "Heart on My Sleeve" được một người dùng sử dụng công cụ AI tạo ra đã gây một cuộc náo loạn trên Internet, gây ra một cuộc tranh cãi kéo dài hàng tháng. Chúng tôi đưa ra kết luận rằng máy móc có khả năng sáng tác nhạc sáng tạo ngang bằng con người. Tuy nhiên, để được chấp nhận rộng rãi như một nhạc sĩ thực thụ, AI cần vượt qua rào cản về đạo đức và chính trị. Con người cần đưa ra quyết định về vai trò của AI trong sáng tác nhạc, tránh để AI hoàn toàn thay thế con người hoặc sở hữu toàn bộ quyền tác giả.

2. NGHIÊN CỨU LIÊN QUAN

NHẬN DIỆN LỖI CÁC BỘ BIẾN ĐỔI ÂM THANH ĐỂ LẠI TRONG GIỌNG NÓI TỔNG HỢP BẰNG AI[6]

Nghiên cứu này sử dụng thuật toán dựa trên việc nhận diện các hiện vật (artifacts) do các bộ biến đổi âm thanh (neural vocoder) để lại trong giọng nói tổng hợp. Thuật toán chính được sử dụng là mô hình RawNet2, xử lý trực tiếp các dạng sóng âm thanh để phát hiện các hiện vật đặc trưng của vocoder. Nhóm nghiên cứu đã cải tiến RawNet2 bằng cách tích hợp phương pháp học đa nhiệm, kết hợp giữa phân loại nhị phân (để phân biệt giọng thật và giọng tổng hợp) và nhiệm vụ nhận diện vocoder. Thiết lập hai nhiệm vụ này giúp mô hình nắm bắt tốt hơn các đặc điểm riêng biệt của âm thanh được tạo ra bởi vocoder, dẫn đến việc phát hiện chính xác hơn.Tỷ lệ EER của thuật toán là 4.54% cho thấy độ chính xác của thuật toán cao.

SINGFAKE: PHÁT HIỆN GIỌNG HÁT TẠO[10]

Nghiên cứu này trình bày thuật toán phát hiện giọng hát giả mạo dựa trên phương pháp học sâu. Thuật toán này sử dụng mạng nơ-ron tích chập (CNN) để trích xuất các đặc trưng từ phổ âm thanh, giúp nhận diện các đặc điểm âm thanh quan trọng từ giọng hát. Sau đó, mạng nơ-ron tuần hoàn (RNN) được sử dụng để phân tích các đặc điểm này theo thời gian, giúp theo dõi sự thay đổi và tính liên tục trong giọng hát. Đặc biệt, mô hình Transformer được áp dụng để tăng cường khả năng nhận diện và phân loại giọng hát giả mạo, nhờ vào cơ chế tự chú ý (self-attention), giúp cải thiện đáng kể độ chính xác và khả năng tổng quát của mô hình. Tuy nhiên nghiên cứu cũng chỉ ra một số hạn chế của thuật toán như: hiệu suất không ổn định bởi vì các mô hình có hiệu suất khác nhau đáng kể tùy thuộc vào điều kiện thử nghiệm. Khi có nhạc nền, EER của các mô hình tăng lên đáng kể, giảm hiệu quả phát hiện, nhạc nền có thể khiến các hệ thống gặp khó khăn trong việc phân biệt giọng hát thật và giả mạo khi có nhạc nền, các mô

hình phức tạp như vậy đòi hỏi lượng tài nguyên tính toán lớn và thời gian xử lý đáng kể, gây khó khăn cho việc ứng dụng thực tế.

HỆ THỐNG CẢI THIỆN GIỌNG HÁT CÁ NHÂN DƯA TRÊN AI [1] Có thể thấy nhìn qua mặt chữ "Hệ Thống Cải Thiện Giọng Hát Cá Nhân Dựa Trên AI " khác với chủ đề của chúng tôi "Phân Biệt Giọng Hát Do AI Tạo Ra Hay Con Người Tạo Ra". Tuy vậy phương pháp để 2 vấn đề này được giải quyết bằng phương pháp được cho là tương tự nhau. Hệ thống này là một ứng dụng công nghệ nhằm cải thiện kỹ năng hát của người dùng thông qua việc phân tích và đánh giá giọng hát cá nhân của họ. Hệ thống này sử dụng trí tuệ nhân tạo để nhận diện và đánh giá các đặc điểm của giọng hát và phong cách biểu diễn của các nghệ sĩ mà người dùng chọn. Sau đó, nó cung cấp phản hồi và điểm số cá nhân hóa, cho phép người dùng theo dõi tiến triển và cải thiện kỹ năng hát của mình. Từ đó khích lệ và động viên họ tiếp tục cải thiện kỹ năng hát.

Điểm đặc biệt của phương pháp này đó là sử dụng cùng 1 phương pháp nhận dạng giống giải pháp của chúng tôi, đó là VOICE ID(Singer Matching). Chúng tôi sẽ giải thích kỹ phương pháp này ở phần phương pháp

THỬ THÁCH NHẬN DẠNG [9]: Khó khăn khi nhận biết giọng hát do AI tạo ra.

Phân biệt dựa trên giọng hát: giọng hát do AI tạo ra có thể ngày càng tinh vi và khó phân biệt với giọn hát của con người.Các phương pháp phân biệt dựa trên giọng hát thường tập trung vào các đặc điểm như cao độ, nhịp điệu, âm sắc và cấu trúc. Tuy nhiên, AI có thể học cách tạo ra giọng hát có các đặc điểm tương tự như giọng hát của con người. Phân biệt qua lời bài hát: thông thường, con người sẽ hát những bài hát có lời hay, ý đẹp và có cảm xúc, việc nhận ra giọng hát AI qua lời bài hát có thể gặp dễ nhận ra trong việc phân biệt các sắc thái và ý nghĩa tinh tế trong lời bài hát. Phân biệt dựa trên siêu dữ liệu: Siêu dữ liệu, chẳng hạn như thông tin về tác giả, thể loại và ngày tạo, có thể được sử dụng để phân biệt âm nhạc do AI tạo ra. Tuy nhiên, siêu dữ liệu có thể dễ dàng bị thao tác hoặc giả mạo.

Khó khăn chung: Việc phân biệt âm nhạc do AI tạo ra là một lĩnh vực nghiên cứu mới và đang ngày càng phát triển. Các phương pháp hiện tại được các chuyên gia cho là thường không chính xác và có thể bị đánh lừa bởi các kỹ thuật AI mới. Việc thiếu dữ liệu đào tạo có chất lượng cao cũng là một thách thức lớn. Ngoài ra: Việc phân biệt âm nhạc do AI tạo ra có thể dẫn đến những vấn đề đạo đức và pháp lý. Ví dụ, việc sử dụng AI để tạo ra âm nhạc giả mạo hoặc đạo nhạc có thể gây ra những hậu quả nghiêm trọng.

3. PHƯƠNG PHÁP VOICE ID

Đối với phương pháp truyền thống để có thể phân biệt AI đó là Watermarking (Đánh dấu bản quyền kỹ thuật số bằng hình mờ) và Artifact Detection (Phát hiện lỗi) đều có những hạn chế khi sử dụng để phân biệt âm nhạc do AI tạo ra [5].

Đánh dấu bản quyền kỹ thuật số bằng hình mờ: Cách tiếp cận đầu tiên dựa vào việc phát hiện các hình mờ được thêm vào. Trong quá trình tạo nội dung AI, mô hình AI sẽ nhúng các hình mờ không thể nhận ra vào chính nội dung được tạo. Lỗ hổng chính của thủy vân là không có kỹ thuật thủy vân nào được biết đến có thể tồn tại thành công sau các nỗ lực loại bỏ hoặc sửa đổi. Vì hình mờ được thiết kế để có thể phát hiện được bằng các thuật toán phát hiện hình mờ nên các thuật toán như vậy cũng có thể được sử dụng để sửa đổi hình mờ. Tuy vậy, âm nhạc là một dạng nghệ thuật trừu tượng vì vậy việc phát hiện hình mờ không thể được coi là phương pháp đáng tin cậy để xác định nội dung do AI tạo ra, đặc biệt khi người tạo hoặc nhà xuất bản nội dung đó có ý định xấu hoặc thông tin sai lệch

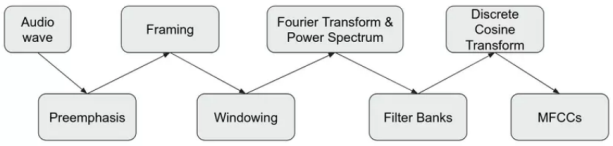
Phát hiện lỗi: Cách tiếp cận thứ hai dựa vào việc phát hiện các hiện vật dành riêng cho AI trong nội dung được tạo. Ví dụ: trong các giải pháp tạo hình ảnh ban đầu, bàn tay con người thường bị biến dạng hoặc khuyên tai không khớp. Những đồ tạo tác như vậy rất dễ được phát hiện nhưng sau đó đã được rèn luyện nhờ những tiến bộ trong AI. Các công cụ phát hiện AI trực tiếp dựa vào việc tìm kiếm các thành phần và tính năng cụ thể của nội dung do AI tạo ra cũng là các mô hình dựa trên mạng thần kinh sâu. Vì vậy, việc phát hiện hiện vật sẽ luôn tụt hậu so với nội dung do AI tạo

ra và do đó không phải là giải pháp phù hợp để nhận dạng.

Vì vậy, chúng tôi quan tâm và nghiên cứu các phương pháp thực tiễn, tiên tiến hơn và vẫn đang trong quá trình phát triển, đó là Voice ID và ACR. Đầu vào kiểm tra (INPUT) sẽ là giọng hát của ca sĩ qua nhiều bài hát(Đã trải qua quá trình lọc nhạc và giọng hát của chúng tôi)

Hình 2: Tổng quan về quá trình phân biệt.

Automated Content Recognition (ACR): Phương pháp nhận dạng nội dung này là một công nghệ cho phép nhận diện và nhận biết nội dung đa phương tiện một cách tự động. Phương pháp này thường được áp dụng trong các hệ thống gợi ý nội dung, quảng cáo tương tác, phân tích dữ liệu phát sóng, và nhiều ứng dụng khác liên quan đến nội dung đa phương tiện. Trong lĩnh vực âm nhạc này thì phương pháp này sử dụng các thuật toán xử lý tín hiệu âm thanh để so sánh mẫu âm thanh (INPUT) với các mẫu đã biết trước để xác định xem chúng có tương tự nhau hay không.



Hình 3: Tổng quan về quá trình tính MFCCs.

Sử dụng MFCC là viết tắt của "Mel-Frequency Cepstral Coefficients". Nó là một cách để trích xuất các đặc trưng (feature extraction) giọng nói (speech) thường được sử dụng trong các model nhận dạng giọng nói (Automatic Speech Recognition) hay phân loại giọng nói (Speech Classification). Đây là một phương pháp phổ biến trong xử lý tín hiệu âm thanh để biểu diễn thông tin âm thanh dưới dạng các vectơ đặc trưng

Hình 4: Singer identification.

Voice Fingerprint (Singer Identification) để trích xuất vân tay giọng nói từ bản cover, chúng tôi sử dụng kỹ thuật nhận dạng giọng nói MFCC (Mel-Frequency Cepstral Coefficients). Nhận dạng ca sĩ: phương pháp được phát triển từ **(ACR)** - xác định danh tính (tên) của ca sĩ trong một bản ghi âm thanh.

Sử dụng Chroma để biểu diễn tần số cho thấy cường độ của 12 cao độ âm nhạc (C, C, D, D, E, F, F, G, G, A, A, B) trong một đoạn thời gian và Spectral Contrast để tính toán độ tương phản phổ đo lường sự khác biệt giữa các đỉnh và đáy trong phổ tần số Phương pháp cần có dữ liệu giọng hát mẫu của ca sĩ trong một cơ sở dữ liệu tham chiếu để so sánh. Tạo dấu vân tay giọng hát (digital fingerprint) của ca sĩ để lưu trữ. Sử dụng dấu vân tay giọng hát để xác định ca sĩ trong các bản ghi khác, kể cả bản ghi có giọng hát do AI tạo ra.

Sử dụng công thức Short-Time Fourier Transform (STFT): Chia tín hiệu âm thanh thành các đoạn nhỏ (cửa sổ) và thực hiện biến đổi Fourier để chuyển đổi từ miền thời gian sang miền tần số:

$$X(w,t) = \sum_{n=-\infty}^{\infty} x[n] * w[n-t] * e^{-jwn},$$

trong đó x[n] là tín hiệu âm thanh, ω là tần số góc.

Lấy log của năng lượng ở mỗi dải tần số. Sau đó dùng công thức Discrete Cosine Transform (DCT) để nén các hệ số và giữ lại các hệ số quan trọng nhất:

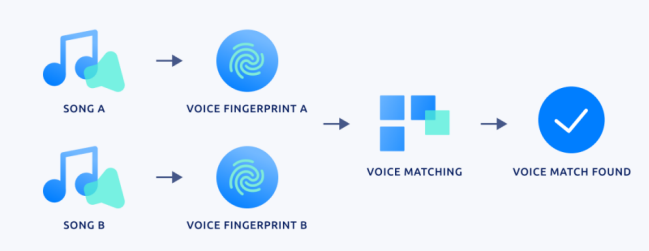
$$c_k = \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right],$$

trong đó c_k là hệ số DCT, x_n là giá trị sau khi log, và N là số lượng điểm.

Trung bình của các giá trị MFCCs:

$$avg\_mfccs = \frac{1}{T}\sum_{t=1}^{T} mfccs[t],$$

mfccs[t] là vector MFCC ở khung thời gian thứ t, T là tổng số khung thời gian trong đoạn âm thanh.



Hình 5: Singer Matching.

Hình 6: Singer matching.

Voice ID: Singer matching (Đối sánh giọng hát). Đây là phương pháp đặc biệt vẫn còn đang phát triển, tốn nhiều tài nguyên nhưng thực sự hiệu quả. Mục tiêu là xác định xem hai bản ghi (hoặc nhiều hơn) có cùng một ca sĩ hát hay không, bất kể phong cách nhạc hay giọng hát. Tìm kiếm sự trùng khớp giữa các đoạn của các bài hát khác nhau, ngay cả khi giọng hát chỉ trùng khớp trong 10 giây. Điểm đặc biệt: Không cần biết danh tính ca sĩ trước, không cần huấn luyện mô hình trên bất kỳ mẫu giọng hát nào. Giả thuyết: Giọng hát của một người có những đặc điểm riêng biệt, ngay cả khi hát theo phong cách khác nhau. Phương pháp này tập trung vào việc phân tích những đặc điểm đó.

Các bước tiềm năng như: Trích xuất đặc điểm giọng hát bao gồm Hệ thống có thể trích xuất các đặc điểm giọng hát từ bản ghi âm thanh, chẳng hạn như: Tần số cơ bản (pitch): độ cao thấp của giọng hát. Âm sắc (timbre): chất lượng âm thanh độc đáo của giọng hát. Phiên âm (formant): đặc trưng cộng hưởng của vòm miệng khi phát ra âm thanh. Các đặc điểm này được chuyển đổi thành dạng dữ liệu máy tính để phân tích.

Hệ thống so sánh các đặc điểm giọng hát được trích xuất từ các bản ghi khác nhau. Các thuật toán học máy có thể được sử dụng để tìm kiếm các mẫu trùng khớp hoặc các mô hình tương tự trong các đặc điểm. Khi xác định 2 bài hát do cùng 1 ca sĩ hát, thì ta sẽ so sánh giọng nói gốc với giọng nói do AI tạo ra và xác định xem chúng có giống nhau hay không. Một ví dụ tuyệt vời về điều này là ca khúc " Heart on my sleeve ", sử dụng giọng hát do AI tạo ra của Drake và The Weeknd. Voice ID khớp giọng của Drake AI trong " Hotline Bling " , cũng như giọng của The Weeknd AI trong " Save Your Tears "

Sau khi có vân tay giọng nói từ bản cover, chúng ta có thể so sánh nó với một cơ sở dữ liệu các vân tay giọng nói đã biết để xác định xem có sự khớp nào không. Các phương pháp như Dynamic Time Warping (DTW) hoặc cosine similarity thường được sử dụng cho mục đích này.

Hệ số tương quan Pearson giữa hai mảng đặc trưng của giọng hát được tính bằng công thức:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

x_i, y_i là các giá trị trong hai mảng đặc trưng đang được so sánh, $\bar{x}, \bar{y}$ là giá trị trung bình của mỗi mảng đặc trưng,

các tổng được tính trên tất cả các giá trị i trong mảng đặc trưng.

AI Detection [4]là bước cuối cùng để phát hiện xem giọng nói của bản cover có được tạo ra bởi trí tuệ nhân tạo (AI) hay không, chúng ta có thể kiểm tra xem người hát của bản cover có phải là một nghệ sĩ ghi âm đã được biết đến hay không. Nếu người hát là một nghệ sĩ ghi âm đã biết và chưa từng ghi âm bài hát đó, có thể nói rằng có khả năng bản cover được tạo ra bằng AI. Chúng tôi chọn ra mô hình tốt nhất trong quá trình training(Random Forest, SVMM, Gradient Boostinga) ,xong đó là mô hình SVM (Support

Vector Machine) với kernel tuyến tính. SVM là một phương pháp học máy giám sát, được sử dụng cho cả phân loại và hồi quy, trong trường hợp này sử dụng cho việc phân loại để có thể dán nhãn giọng hát của bản ghi.

Điều đặc biệt trong cả 3 chức năng trên đó là đều sử dụng phương pháp so sánh dựa trên hệ số tương quan. Điều này có nghĩa là chúng tính toán mức độ tương quan giữa các vector đặc trưng. Hệ số tương quan càng gần với 1 thì các vector đặc trưng càng giống nhau. Hệ số tương quan (Correlation coefficient): Cho hai vector X và Y có n chiều, hệ số tương quan Pearson được tính bằng công thức sau: Tóm gọn lại:

Audio Matching sẽ là bước đầu tiên trong quá trình xử lý phân biệt giọng hát mà con người tạo ra hay do AI tạo ra. Sử dụng thuật toán nhận dạng nội dung Automated Content Recognition(ACR) để phát hiện và so sánh âm thanh của bản cover với bản gốc, chúng ta có thể sử dụng các kỹ thuật xử lý tín hiệu âm thanh như cross-correlation, spectral analysis, hoặc sử dụng các thư viện nhận dạng âm thanh như Librosa trong Python. Sau đó là 3 phương pháp VOICE ID để có thể phân biệt được giọng hát do con người tạo ra hay do AI tạo ra.

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

(X,Y) là hiệp phương sai (covariance) giữa X và Y,

σ_X và σ_Y là độ lệch chuẩn (standard deviation) của X và Y, tương ứng.

4. THỰC NGHIỆM

4.1 Cài đặt

Cài đặt thư viện dành riêng cho việc phân tích và xử lý âm thanh Librosa. Nó cung cấp một bộ công cụ phong phú để làm việc với tín hiệu âm thanh, bao gồm các chức năng để tải, hiển thị, và trích xuất các đặc trưng từ các tệp âm thanh. Librosa được sử dụng rộng rãi trong nghiên cứu và ứng dụng thực tế liên quan đến xử lý âm thanh và nhận dạng giọng nói.

Xây dựng và huấn luyện mô hình: sử dụng ba mô hình khác nhau: Random Forest, SVM (Support Vector Machine), và Gradient Boosting. Mỗi mô hình được huấn luyện và đánh giá bằng cách sử dụng kỹ thuật k-fold cross-validation để đảm bảo tính tổng quát và độ tin cậy của mô hình. Sử dụng GridSearchCV để tối ưu hóa các siêu tham số của mô hình SVM nhằm cải thiện hiệu suất.

4.2 Dữ liệu

4.2.1 Thu thập dữ liệu

Dữ liệu có vai trò quan trọng bậc nhất trong việc phát triển và ứng dụng phát hiện giọng hát sử dụng AI. Do dữ liệu thu thập rất khó khăn, hầu hết dữ liệu đáng tin cậy và đầy đủ đều phải trả phí cho các công ty tự thu thập nên chúng tôi quyết định chỉ thu thập dữ liệu giọng hát của 1 ca sĩ (Justin Bieber [3]) và tập trung vào phần nhận biết 2 bản gốc và bản cover của ca sĩ đó là do ca sĩ đó làm hay do AI tạo ra.

Âm thanh: bản gốc - bài hát gốc được sử dụng để so sánh với các bản cover, bản cover - bài hát được hát lại bởi một ca sĩ khác hoặc do AI tạo ra.

Nguồn dữ liệu: nguồn dữ liệu tự thu thập trên một số nền tảng như Spotify, Youtube, ...

Dữ liệu thu thập sẽ ở dưới dạng file mp3 và có độ dài không quá 6 phút

4.2.2 Tiền xử lý dữ liệu

Dữ liệu sau khi thu thập sẽ trải qua các công đoạn tiền xử lý trước khi đưa vào sử dụng.

Dầu tiên sẽ là: trích xuất ra các đặc trưng của âm thanh: sử dụng thư viện librosa để trích xuất các đặc trưng âm thanh từ mỗi tệp âm thanh. Các đặc trưng bao gồm MFCC (Mel-Frequency Cepstral Coefficients), Chroma, Spectral Contrast. Các đặc trưng này sau đó được tính trung bình theo thời gian để tạo ra một vector đặc trưng duy nhất cho mỗi tệp âm thanh.

Tách dữ liệu thành 2 tập riêng biết để đảm bảo cho việc huấn luyện: giọng hát và nhạc

Xử lý mất cân bằng dữ liệu: sử dụng phương pháp SMOTE (Synthetic Minority Over-sampling Technique) để tạo thêm dữ liệu từ lớp thiểu số (minority class) nhằm xử lý vấn đề mất cân bằng dữ liệu giữa giọng hát thật và giọng hát giả mạo.

Chuẩn hóa dữ liệu: dữ liệu đặc trưng được chuẩn hóa bằng StandardScaler để đảm bảo các đặc trưng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa này giúp cải thiện hiệu suất của các mô hình học máy.

Dữ liệu sau khi tiền xử lý sẽ được chia thành 2 tập: huấn luyện và kiểm thử.

Đối với tập huấn luyện: cặp dữ liệu âm thanh (bản gốc, bản cover) được sử dụng để huấn luyện hệ thống. Dữ liệu sẽ được chuyển sang dạng wav để dễ phân tích và lấy ra được các đặc trưng của giọng hát hơn. Sau đó được chia ra làm hai thư mục: "Music-Human" và "Music-AI". Các tệp âm thanh từ thư mục "Music-Human" được gán nhãn là 0 (giọng hát thật), và các tệp từ thư mục "Music-AI" được gán nhãn là 1 (giọng hát do AI tạo).

Đối với tập kiểm thử: dữ liệu âm thanh (bản gốc, bản cover) không được sử dụng trong quá trình huấn luyện. Và cũng sẽ đuuọc chuyển sang dạng wav.

4.3 Kết quả và đánh giá

Mô hình dự đoán độ chính xác là 93.3%.

Trong quá trình thực nghiệm, khi kiểm tra các tập test, mô hình đưa ra kết quả chính xác, dán nhãn đúng hơn mong đợi. Chúng tôi nhận ra rằng thực chất, khi tách các đặc trưng của giọng hát ra và nhìn vào bảng so sánh các đặc trưng, chúng tôi thấy rằng MFCCs, Spectral Contrast được AI mô phỏng con người rất tốt nhưng với Chroma thì khác. Chroma tập trung vào cường độ của các cao độ âm nhạc, hữu ích trong phân tích và nhận dạng các thuộc tính âm nhạc, giọng của Justin Bieber được đánh giá là thanh và cao nhưng có vẻ AI bắt chước với tỷ lệ không cao.

Mô hình dự đoán chỉ mới thực hiện phân biệt giọng hát 1 ca sĩ (Justin Bieber) là người hay do AI làm nên kết quả mới có thể lên đến trên 90%. Mô hình cần sự bổ sung số lượng lớn dữ liệu để có thể gán nhãn do con người làm hay AI làm và hoàn thiện tập dữ liệu training và testing. Mô hình đã có sự tối ưu khi lựa chọn mô hình tốt nhất trong 3 mô hình (Random Forest, SVM và GradientBoosting) rồi sử dụng GridSearchCV.

5. KẾT LUẬN

Nghiên cứu đã trình bày một giải pháp để phân biệt giọng hát do AI tạo ra và giọng hát do con người thể hiện. Bằng cách sử dụng các đặc trưng âm thanh như MFCCs, Chroma, Spectral Contrast và các kỹ thuật nhận dạng giọng hát tiên tiến, chúng tôi đã đạt được độ chính xác cao trong việc phân biệt giọng hát của ca sĩ Justin Bieber giữa bản gốc và bản do AI tạo ra.

Kết quả thực nghiệm cho thấy mô hình dự đoán đạt độ chính xác 93.3% , chứng minh tính hiệu quả của phương pháp đề xuất. Điều này mở ra triển vọng lớn cho việc áp dụng các kỹ thuật xử lý và phân tích âm thanh trong việc bảo vệ quyền lợi của nghệ sĩ và ngăn chặn việc lạm dụng công nghệ AI để tạo ra các sản phẩm âm nhạc giả mạo.

Tuy nhiên, nghiên cứu này cũng gặp phải một số hạn chế, như việc chỉ tập trung vào một ca sĩ duy nhất và dữ liệu huấn luyện chưa đủ phong phú để bao quát đa dạng giọng hát của nhiều nghệ sĩ khác nhau. Trong tương lai, khi có đủ tài nguyên và dữ liệu, chúng tôi sẽ mở rộng phạm vi nghiên cứu để bao gồm nhiều ca sĩ hơn, từ đó cải thiện độ chính xác và khả năng ứng dụng thực tế của mô hình qua 2 bước lớn chặt chẽ : ACR và Voice ID.

Nhìn chung, nghiên cứu đã góp phần trong việc phát triển các công cụ nhận dạng và phân biệt giọng hát, bảo vệ quyền sở hữu trí tuệ và quyền lợi của các nghệ sĩ trong bối cảnh công nghệ AI ngày càng phát triển mạnh mẽ. Việc tiếp tục phát triển và hoàn thiện các phương pháp này là cần thiết để đối phó với những thách thức đạo đức và pháp lý mà công nghệ AI mang lại trong ngành công nghiệp âm nhạc.

TÀI LIỆU

[1] HTS Caldera and BVKI Vidanage. Ai-driven voice training and singing improvement: A comprehensive literature.

[2] Math Duffin. How does ai music work? from machine learning to viral hits. *Machine Learning to Viral Hits*, page 1, 2024.

[3] Chas Newkey-Burden. *Justin Bieber: the unauthorized biography*. Michael O'Mara, 2010.

[4] Pex. Real or fake: Identifying ai-generated music and voices. *AI-generated music*, page 1, 2023.

[5] Pex. Rrd spotlight: How pex identifies singers and voices in ai-generated content. *AI-generated music*, page 1, 2024.

[6] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 904–912, 2023.

[7] Pham Thanh Tra. The challenge of test data management (tdm) lies in collecting the right data effectively. *TEST DATA MANAGEMENT IN SOFTWARE TESTING*.

[8] Eleftheria Trygousi. Copyright issues pertaining to musical works created by artificial intelligence. 2022.

[9] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H Luan. A survey on chatgpt: Ai-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 2023.

[10] Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. Singfake: Singing voice deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12156–12160. IEEE, 2024.

[11] Ivano Zanzarella. The problem of musical creativity and its relevance for ethical and legal decisions towards musical ai, 2020.