|  |  |
| --- | --- |
|  | **TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  **KHOA CÔNG NGHỆ THÔNG TIN** |

**ĐỒ ÁN MÔN XỬ LÝ TIẾNG NÓI – K31**

**HiFi-GAN: Generative Adversarial Networks for  
Efficient and High Fidelity Speech Synthesis**

1. **THÔNG TIN CHUNG**

**Học viên thực hiện:**

21C11036: Nguyễn Thị Ngọc Trâm

***Hướng ứng dụng:***Tìm hiểu hoàn chỉnh một giải pháp cho một vấn đề cụ thể trong lĩnh vực xử lý tiếng nói. Xây dựng 1 ứng dụng, có sử dụng các công nghệ về xử lý/nhận dạng/tổng hợp tiếng nói.

Link Run Code: <https://colab.research.google.com/drive/1Ngm2yC4UcPiGoZZNYtuXyuNZwbYAqCb0?usp=sharing>

Link Data + Source Code:

<https://drive.google.com/drive/folders/16AhH3B77Eh4ICvRFQs5z4r_tHz4gRK_b?usp=sharing>

1. **NỘI DUNG BÁO CÁO**
2. **GIỚI THIỆU**

Giọng nói là một trong những phương thức giao tiếp được con người sử dụng thường xuyên và theo cách tự nhiên nhất. Với sự phát triển gần đây của công nghệ, giọng nói đang được sử dụng làm giao diện chính trong các dịch vụ trợ lý giọng nói trí tuệ nhân tạo (AI) như Amazon Alexa và nó cũng được sử dụng rộng rãi trong xe tự động, nhà thông minh, v.v. Theo đó, trước nhu cầu con người trò chuyện với máy móc ngày càng tăng, công nghệ tổng hợp lời nói tự nhiên giống như lời nói của con người đang được tích cực nghiên cứu.

Gần đây, với sự phát triển của mạng nơ ron, công nghệ tổng hợp giọng nói đã đạt được những bước tiến nhanh chóng. Hầu hết các mô hình nơ ron tổng hợp giọng nói đều sử dụng quy trình hai giai đoạn:

*Giai đoạn 1:* Dự đoán biểu diễn trung gian có độ phân giải thấp như melspectrogram (Shen và cộng sự., 2018, Ping và cộng sự., 2017, Li và cộng sự., 2019) hoặc các đặc điểm ngôn ngữ (Oord và cộng sự, 2016) từ văn bản (mô hình hóa biểu diễn mức độ thấp của tiếng nói con người từ văn bản)

Text 🡪 Melspectrogram (Biểu diễn trung gian)

* *Giai đoạn 2:* Tổng hợp âm thanh dạng sóng thô từ biểu diễn trung gian (Oord và cộng sự, 2016, 2018, Prenger và cộng sự, 2019, Kumar và cộng sự, 2019). Với tối đa 24.000 mẫu mỗi giây và độ trung thực lên tới 16 bit. Trong nghiên cứu này, nhóm tác giả tập trung vào việc thiết kế một mô hình tổng hợp hiệu quả các dạng sóng có độ trung thực cao (fidelity) từ các mel-spectrogram.

Intermediate Representation 🡪 Raw waveform audio

(MelSpec) (high fidelity)

1. **CÁC CÔNG TRÌNH LIÊN QUAN**

Nhiều công trình đã được tiến hành để cải thiện chất lượng tổng hợp âm thanh và hiệu quả ở giai đoạn thứ hai.

* WaveNet (Oord và cộng sự, 2016) là một mạng nơ ron tích chập tự hồi quy (AR) thể hiện khả năng của các phương pháp dựa trên mạng nơ ron vượt qua các phương pháp thông thường về chất lượng. Tuy nhiên, do cấu trúc AR, WaveNet tạo ra một mẫu ở mỗi thao tác chuyển tiếp; nó cực kỳ chậm trong việc tổng hợp âm thanh có độ phân giải cao theo thời gian.
* Các mô hình Flow-based generative được đề xuất để giải quyết vấn đề này. Do khả năng mô hình hóa các raw waveform bằng cách biến đổi song song các chuỗi nhiễu có cùng kích thước, các mô hình tổng quát dựa trên dòng chảy sử dụng đầy đủ các bộ xử lý điện toán song song hiện đại để tăng tốc độ lấy mẫu.
* Parallel WaveNet (Oord et al., 2018) là một luồng tự hồi quy nghịch đảo (IAF) được đào tạo để giảm thiểu sự phân kỳ Kullback-Leibler của nó từ một WaveNet được đào tạo trước được gọi là giáo viên. So với teacher model, nó cải thiện tốc độ tổng hợp lên 1.000 lần hoặc hơn mà không làm giảm chất lượng.
* WaveGlow (Prenger và cộng sự, 2019) loại bỏ nhu cầu chắt lọc teacher model và đơn giản hóa quá trình học tập thông qua ước tính khả năng tối đa bằng cách sử dụng các luồng song ánh hiệu quả dựa trên Glow (Kingma và Dhariwal, 2018). Nó cũng tạo ra âm thanh chất lượng cao so với WaveNet. Tuy nhiên, nó đòi hỏi nhiều tham số cho kiến trúc sâu với hơn 90 lớp.
* Các mạng đối kháng (GAN) (Goodfellow et al., 2014), là một trong những mô hình tạo sâu chiếm ưu thế nhất, cũng đã được áp dụng để tổng hợp giọng nói. Kumar và cộng sự. (2019) đã đề xuất một kiến trúc đa tỷ lệ cho các bộ phân biệt đối xử hoạt động trên nhiều tỷ lệ dạng sóng thô.
* Với việc xem xét kiến trúc phức tạp, trình tạo MelGAN đủ nhỏ gọn để cho phép tổng hợp thời gian thực trên CPU. Yamamoto et al. (2020) đã đề xuất hàm mất STFT đa độ phân giải để cải thiện và ổn định quá trình đào tạo GAN, đồng thời đạt được hiệu quả tham số tốt hơn và thời gian đào tạo ít hơn so với mô hình IAF, ClariNet (Ping et al., 2018).
* Thay vì mel-spectrogram, GAN-TTS (Binkowski và cộng sự, 2019) tạo thành công dạng sóng âm thanh thô chất lượng cao từ các đặc điểm ngôn ngữ thông qua nhiều bộ phân biệt đối xử hoạt động trên các kích thước cửa sổ khác nhau. Mô hình cũng hiển thị ít FLOP hơn so với Parallel WaveNet. Mặc dù có những ưu điểm nhưng vẫn có khoảng cách về chất lượng mẫu giữa các mô hình GAN và AR hoặc các mô hình dựa trên dòng chảy.

=> Nhóm tác giả đề xuất HiFi-GAN, đạt được cả hiệu quả tính toán và chất lượng mẫu cao hơn so với AR hoặc các mô hình flow-based. Vì âm thanh lời nói bao gồm các tín hiệu hình sin với các chu kỳ khác nhau, nên việc lập mô hình các mẫu theo chu kỳ đóng vai trò quan trọng để tạo ra âm thanh lời nói trung thực. Do đó, nhóm tác giả đề xuất một bộ phân biệt bao gồm các bộ phân biệt phụ nhỏ, mỗi bộ phân biệt chỉ thu được một phần định kỳ cụ thể của dạng sóng thô. Kiến trúc này chính là nền tảng giúp mô hình tổng hợp thành công âm thanh giọng nói trung thực. Khi trích xuất các phần khác nhau của âm thanh discriminator, nhóm tác giả cũng thiết kế một mô-đun đặt nhiều khối dư, mỗi khối quan sát song song các mẫu có độ dài khác nhau và áp dụng nó cho generator.

HiFi-GAN đạt MOS score cao hơn so với các mẫu tốt nhất hiện có, WaveNet và WaveGlow. Nó tổng hợp âm thanh lời nói chất lượng của người ở tốc độ 3,7 MHz trên một GPU V100. Các nhà nghiên cứu tiếp tục cho thấy tính tổng quát của HiFi-GAN đối với sự đảo ngược mel-spectrogram của những người nói không nhìn thấy và tổng hợp giọng nói từ đầu đến cuối. Cuối cùng, phiên bản HiFi-GAN nhỏ gọn chỉ yêu cầu các thông số 0,92M trong khi vượt trội so với các mẫu tốt nhất có sẵn công khai và phiên bản nhanh nhất của các mẫu HiFi-GAN nhanh hơn 13,44 lần so với thời gian thực trên CPU và nhanh hơn 1.186 lần so với thời gian thực trên một đĩa đơn. GPU V100 với chất lượng tương đương với đối tác tự hồi quy.

*Mã nguồn:*

<https://github.com/jik876/hifi-gan>

<https://jik876.github.io/hifi-gan-demo/>

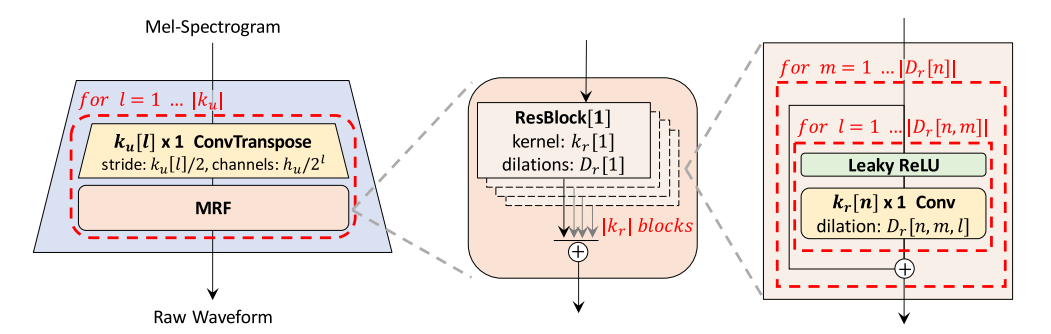
1. **PHƯƠNG PHÁP**

**3.1. Tổng quan**

HiFi-GAN bao gồm một generator và hai discriminators: multi-scale và multi-period discriminators. Generator và discriminator được huấn luyện đối kháng, cùng với hai loss bổ sung để cải thiện độ ổn định đào tạo và hiệu suất mô hình.

**3.2. Generator**

Generator là một mạng fully CNN. Sử dụng một mel-spectrogram làm đầu vào và lấy mẫu thông qua tích chập chuyển vị cho đến khi độ dài của chuỗi đầu ra khớp với độ phân giải tạm thời của dạng sóng thô. Mỗi tích chập chuyển vị được theo sau bởi một mô đun hợp nhất trường (Multi-receptive field fusion - MRF).

****

*Hình 1. Generator lấy mẫu mel-spectrogram lên đến |ku| lần để khớp với độ phân giải tạm thời của dạng sóng thô. Mô-đun MRF thêm các tính năng từ |kr| các khối còn lại kernel size và tốc độ giãn nở khác nhau. Cuối cùng, khối dư n-th với kích thước kernel kr[n] và tốc độ giãn nở Dr[n] trong mô-đun MRF được mô tả.*

Hình 1 thể hiện kiến trúc của generator. Như trong nghiên cứu trước (Mathieu và cộng sự, 2015, Isola và cộng sự, 2017, Kumar và cộng sự, 2019), tiếng ồn không được cung cấp cho generator như một đầu vào bổ sung.

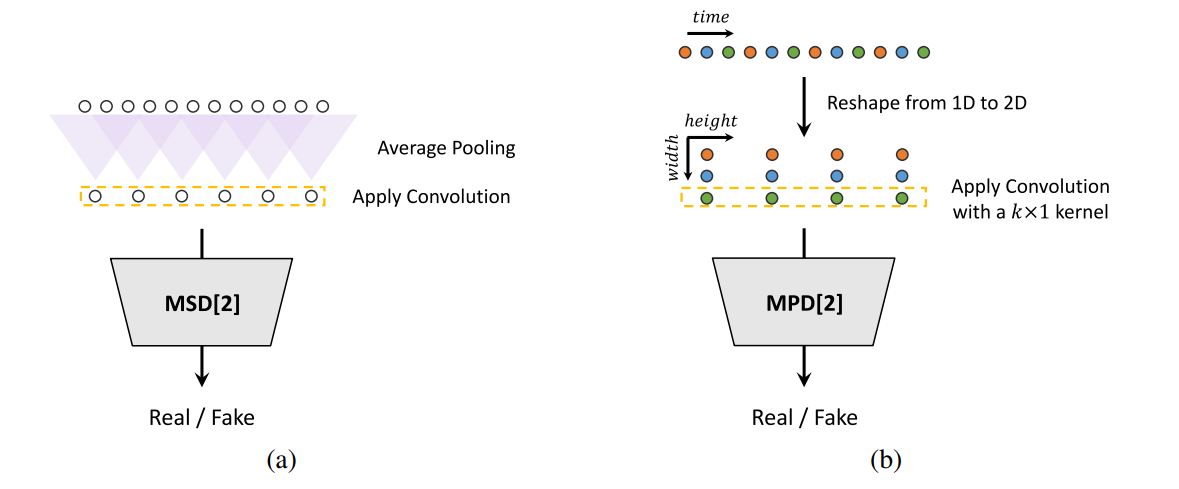
***Multi-Receptive Field Fushion:*** Nhóm tác giả thiết kế mô-đun hợp nhất trường đa tiếp nhận (MRF) cho generator, mô-đun này quan sát song song các mẫu có độ dài khác nhau. Cụ thể, mô-đun MRF trả về tổng đầu ra từ nhiều khối dư. Các kích thước kernel và tốc độ giãn nở khác nhau được chọn cho từng khối còn lại để tạo thành các mẫu trường tiếp nhận đa dạng. Kiến trúc của mô đun MRF và khối dư được thể hiện trong Hình 1. Chúng tôi để lại một số tham số có thể điều chỉnh trong generator; kích thước ẩn hu, kích thước kernel ku của các chập chuyển vị, kích thước hạt nhân kr và tốc độ giãn nở Dr của các mô-đun MRF có thể được điều chỉnh để phù hợp với yêu cầu của mỗi người trong sự đánh đổi giữa hiệu quả tổng hợp và chất lượng mẫu.

**3.3. Discriminator**

Xác định các phụ thuộc dài hạn là chìa khóa cho mô hình âm thanh giọng nói thực tế. Ví dụ: thời lượng của âm vị có thể dài hơn 100ms, dẫn đến mối tương quan cao giữa hơn 2.200 mẫu liền kề ở dạng sóng thô. Vấn đề này đã được giải quyết trong công trình của (Donahue cùng cộng sự., 2018) bằng cách tăng các trường tiếp nhận của generator và discriminator. Nhóm tác giả tập trung vào một vấn đề quan trọng khác vẫn chưa được giải quyết; vì âm thanh giọng nói bao gồm các tín hiệu hình sin với các chu kỳ khác nhau, nên các mẫu định kỳ đa dạng bên dưới dữ liệu âm thanh cần phải được xác định.

Để đạt được mục tiêu này, nhóm tác giả đề xuất multi-period discriminator (MPD) bao gồm một số sub-discriminators, mỗi bộ xử lý một phần tín hiệu định kỳ của âm thanh đầu vào. Ngoài ra, để nắm bắt các mẫu liên tiếp và các phụ thuộc dài hạn, sử dụng multi-scale discriminator (MSD) được đề xuất trong MelGAN (Kumar et al., 2019), đánh giá liên tục các mẫu âm thanh ở các cấp độ khác nhau. Bài báo đã tién hành các thí nghiệm đơn giản để cho thấy khả năng nắm bắt các mẫu định kỳ của MPD và MSD và kết quả có thể được tìm thấy trong Phụ lục B.

**Multi-Period Discriminator**: MPD là sự kết hợp của các sub-discriminators, mỗi bộ chỉ chấp nhận các mẫu âm thanh đầu vào có khoảng cách đều nhau; không gian được đưa ra như là khoảng thời gian p. Các sub-discriminators được thiết kế để bắt các cấu trúc ẩn khác nhau bằng cách xem xét các phần khác nhau của âm thanh đầu vào. Đặt các khoảng thời gian thành [2, 3, 5, 7, 11] để tránh trùng lặp nhiều nhất có thể. Như được hiển thị trong Hình 2b, trước tiên lđịnh hình lại âm thanh thô 1D có độ dài T thành dữ liệu 2D có chiều cao T /p và chiều rộng p , sau đó áp dụng các kết cấu 2D cho dữ liệu được định hình lại. Trong mỗi lớp tích chập của MPD, giới hạn kích thước kernel ở trục chiều rộng là 1 để xử lý các mẫu định kỳ một cách độc lập. Mỗi sub-discriminator là một chồng các lớp tích chập có bước kích hoạt đơn vị tuyến tính được chỉnh lưu (ReLU) bị rò rỉ . Sau đó, chuẩn hóa trọng lượng (Salimans và Kingma, 2016) được áp dụng cho MPD. Bằng cách định hình lại âm thanh đầu vào thành dữ liệu 2D thay vì lấy mẫu tín hiệu định kỳ của âm thanh, độ dốc từ MPD có thể được phân phối tới mọi bước thời gian của âm thanh đầu vào.

****

*Hình 2: (a) sub-discriminator thứ 2 của MSD. (b) sub-discriminator thứ 2 của MPD ở giai đoạn 3*

**Multi-Scale Discriminator:** Bởi vì mỗi bộ phân biệt phụ trong MPD chỉ chấp nhận các mẫu rời rạc, nên chúng tôi thêm MSD để đánh giá chuỗi âm thanh một cách liên tục. Kiến trúc của MSD được rút ra từ kiến trúc của MelGAN (Kumar et al., 2019). MSD là sự kết hợp của ba sub-discriminators hoạt động trên các scale đầu vào khác nhau: âm thanh thô, âm thanh tổng hợp trung bình ×2 và âm thanh tổng hợp trung bình ×4, như thể hiện trong Hình 2a. Mỗi sub-discriminators trong MSD là một chồng các lớp tích chập được sắp xếp và nhóm lại với kích hoạt ReLU bị rò rỉ. Kích thước của bộ phân biệt đối xử được tăng lên bằng cách giảm bước tiến và thêm nhiều lớp hơn. Chuẩn hóa trọng lượng được áp dụng ngoại trừ bộ phân biệt đối xử phụ đầu tiên hoạt động trên âm thanh thô. Thay vào đó, chuẩn hóa phổ (Miyato et al., 2018) được áp dụng và ổn định quá trình đào tạo như đã báo cáo.

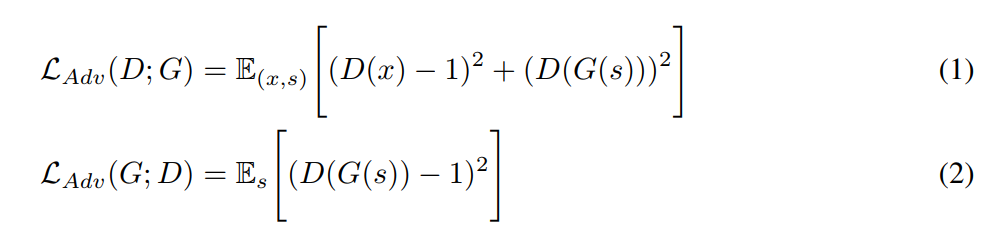
Lưu ý rằng MPD hoạt động trên các mẫu rời rạc của dạng sóng thô, trong khi MSD hoạt động trên dạng sóng được làm mịn.

Đối với công việc trước đây sử dụng kiến trúc đa phân biệt đối xử như MPD và MSD, Binkowski et al. (2019) cũng có thể được giới thiệu. Kiến trúc bộ phân biệt được đề xuất trong tác phẩm này giống với MPD và MSD ở chỗ nó là sự kết hợp của các bộ phân biệt, nhưng MPD và MSD là bộ phân biệt hoàn toàn dựa trên cửa sổ Markovian, trong khi nó tính trung bình đầu ra và có các bộ phân biệt có điều kiện . Ngoài ra, sự giống nhau giữa MPD và RWD (Binkowski và cộng sự, 2019) có thể được xem xét trong phần định hình lại âm thanh đầu vào, nhưng MPD sử dụng các khoảng thời gian được đặt thành số nguyên tố để phân biệt dữ liệu của càng nhiều khoảng thời gian càng tốt, trong khi RWD sử dụng các yếu tố định hình lại của các khoảng thời gian chồng chéo và không xử lý riêng từng kênh của dữ liệu được định hình lại, điều này khác với mục tiêu mà MPD hướng tới. Một biến thể của RWD có thể thực hiện hoạt động tương tự như MPD, nhưng nó cũng không giống MPD về mặt chia sẻ tham số và tích chập có bước sóng với các tín hiệu liền kề. Thông tin chi tiết về sự khác biệt về kiến trúc có thể được tìm thấy trong Phụ lục C.

**3.4. Training Loss Term**

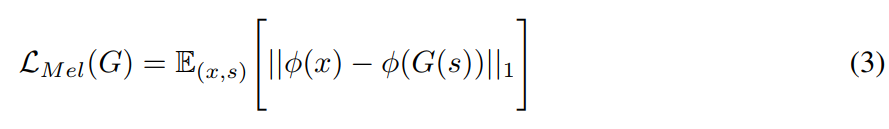
**GAN Loss:** Để cho ngắn gọn, nhóm tác giả mô tả các discriminator, MSD và MPD. Đối với generator và discriminator, việc huấn luyện đối tượng tuân theo LS GAN (Mao và cộng sự, 2017), thay thế cross entropy nhị phân của các mục tiêu GAN ban đầu (Goodfellow và cộng sự, 2014) bằng các hàm mất mát bình phương nhỏ nhất (least squares loss) cho luồng non-vanishing gradient.

Discriminator được đào tạo để phân loại các mẫu grouth truth thành 1 và các mẫu được tổng hợp từ generator thành 0. Generator được huấn luyện để đánh lừa bộ discriminator bằng cách cập nhật chất lượng mẫu để được phân loại thành giá trị gần bằng 1. GAN losses cho generator G và Discriminator D được định nghĩa:



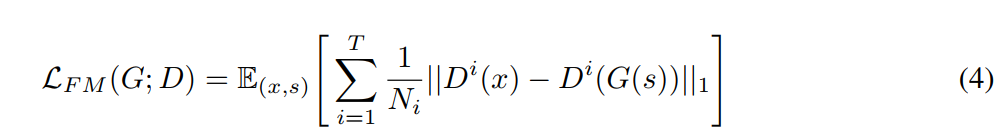
trong đó x biểu thị ground truth audio và s biểu thị điều kiện đầu vào, mel spectrogram của ground truth audio.

**Mel-Spectrogram Loss** Ngoài mục tiêu, đối tượng của GAN, thêm mel-spectrogram loss để cải thiện hiệu quả đào tạo của trình tạo và độ trung thực của âm thanh được tạo. Đề cập đến công trình trước đây (Isola et al., 2017), áp dụng reconstruction loss cho mô hình GAN giúp tạo ra kết quả thực tế và trong Yamamoto et al. (2020), phân phối tần số thời gian được nắm bắt một cách hiệu quả bằng cách cùng nhau tối ưu hóa các hàm tổn thất đối nghịch và biểu đồ quang phổ đa độ phân giải. Chúng tôi đã sử dụng suy hao mel-spectrogram theo các điều kiện đầu vào, điều này cũng có thể có tác dụng tập trung nhiều hơn vào việc cải thiện chất lượng cảm nhận do các đặc điểm của hệ thống thính giác của con người. Suy hao mel spectrogram là khoảng cách L1 giữa mel-spectrogram của dạng sóng do bộ tạo tổng hợp và của dạng sóng thực nền. Nó được định nghĩa là:



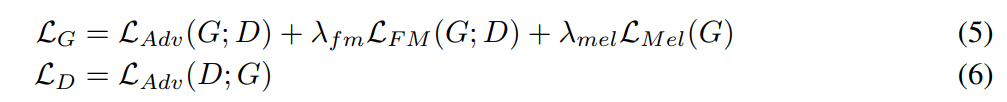
trong đó φ là hàm biến đổi dạng sóng thành mel-spectrogram tương ứng. Mất mel-spectrogram giúp trình tạo tổng hợp dạng sóng thực tế tương ứng với điều kiện đầu vào, đồng thời ổn định quá trình huấn luyện đối thủ ngay từ giai đoạn đầu.

**Feature Matching Loss** Mất kết hợp tính năng là một chỉ số tương tự đã học được đo bằng sự khác biệt về các tính năng của bộ phân biệt giữa mẫu sự thật cơ bản và mẫu được tạo (Larsen và cộng sự, 2016, Kumar và cộng sự, 2019). Vì nó đã được áp dụng thành công cho sis tổng hợp giọng nói (Kumar et al., 2019), chúng tôi sử dụng nó như một tổn thất bổ sung để đào tạo trình tạo. Mọi tính năng trung gian của bộ phân biệt được trích xuất và khoảng cách L1 giữa mẫu sự thật cơ bản và mẫu được tạo có điều kiện trong mỗi không gian tính năng được tính toán. Mất kết hợp tính năng được định nghĩa là:

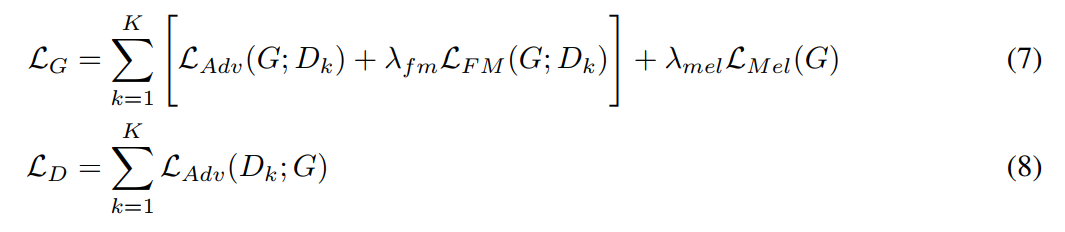
****

trong đó T biểu thị số lượng lớp trong bộ phân biệt đối xử; Di và Ni biểu thị các tính năng và số lượng các tính năng trong lớp thứ i của bộ phân biệt đối xử, tương ứng.

**Final Loss:** Mục tiêu cuối cùng đối với bộ tạo và bộ phân biệt đối xử là

****

Đặt λfm = 2 và λmel = 45. Bởi vì discriminators là một tập hợp các sub-discriminators của MPD và MSD, Phương trình 5 và 6 có thể được chuyển đổi tương ứng các sub-discriminators:

****

trong đó Dk biểu thị bộ phân biệt phụ thứ k trong MPD và MSD.

1. **THỰC NGHIỆM**

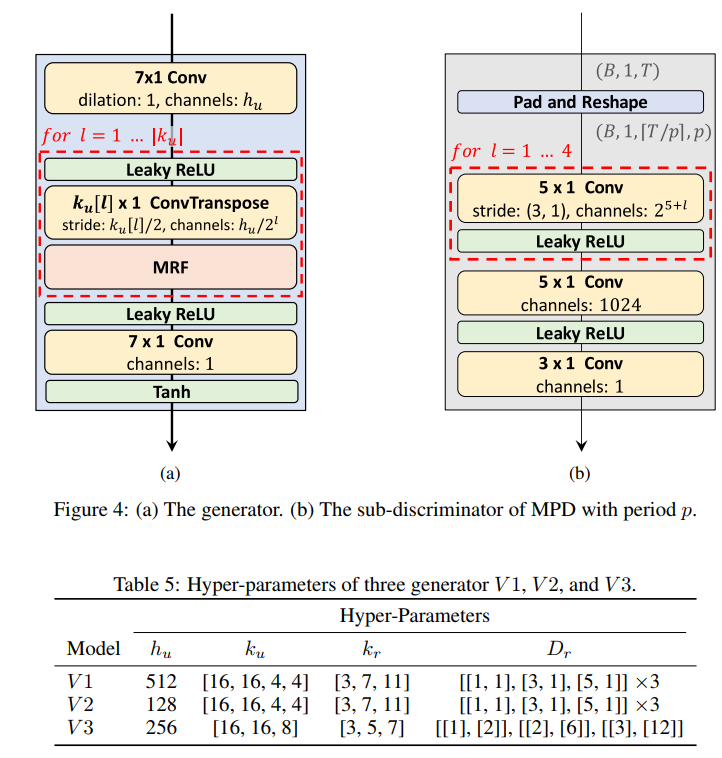
Để so sánh công bằng và có thể lặp lại với các mô hình khác, sử dụng bộ dữ liệu LJSpeech (Ito, 2017) trong đó nhiều mô hình tổng hợp giọng nói được đào tạo. LJSpeech bao gồm 13.100 đoạn âm thanh ngắn của một người nói với tổng thời lượng khoảng 24 giờ. Định dạng âm thanh là PCM 16 bit với tốc độ lấy mẫu là 22 kHz; nó đã được sử dụng mà không có bất kỳ thao tác nào. HiFi-GAN được so sánh với các mô hình tốt nhất có sẵn công khai: hỗn hợp phổ biến của hậu cần (MoL) WaveNet (Oord et al., 2018) triển khai (Yamamoto, 2018) và MelGAN và triển khai chính thức WaveGlow (Valle, 2018b). (Kumar, 2019). Chúng tôi đã sử dụng các trọng số được đào tạo trước được cung cấp cho tất cả các mô hình.

Để đánh giá tính tổng quát của HiFi-GAN đối với sự đảo ngược mel-spectrogram của những người nói không nhìn thấy, chúng tôi đã sử dụng bộ dữ liệu nhiều người nói VCTK (Veaux et al., 2017), bao gồm khoảng 44.200 đoạn âm thanh ngắn được phát ra bởi 109 người nói tiếng Anh bản địa với nhiều điểm nhấn khác nhau. Tổng thời lượng của các đoạn âm thanh là khoảng 44 giờ. Định dạng âm thanh là PCM 16 bit với tốc độ lấy mẫu là 44 kHz. Giảm tốc độ lấy mẫu xuống 22 kHz. Nhóm tác giả đã chọn ngẫu nhiên 9 diễn giả và loại trừ tất cả các đoạn âm thanh của họ khỏi tập huấn luyện. Sau đó, đã đào tạo MoL WaveNet, WaveGlow và MelGAN với cùng cài đặt dữ liệu; tất cả các mô hình đã được đào tạo cho đến 2,5 triệu bước.

Để đánh giá chất lượng âm thanh, chúng tôi đã thu thập các bài kiểm tra MOS 5 thang đo có nguồn gốc từ cộng đồng thông qua Amazon Mechanical Turk. Điểm MOS được ghi lại với khoảng tin cậy (CI) 95%. Những người đánh giá đã nghe các mẫu thử nghiệm một cách ngẫu nhiên, trong đó họ được phép đánh giá từng mẫu âm thanh một lần. Tất cả các đoạn âm thanh đã được chuẩn hóa để tránh ảnh hưởng của sự khác biệt về âm lượng đối với người đánh giá. Tất cả các đánh giá chất lượng trong Phần 4 đều được thực hiện theo cách này và không được lấy từ các bài báo khác.

Tốc độ tổng hợp được đo trên môi trường GPU và CPU theo xu hướng nghiên cứu gần đây về hiệu quả của mạng thần kinh (Kumar và cộng sự, 2019, Zhai và cộng sự, 2020, Tan và cộng sự, 2020). Các thiết bị được sử dụng là một GPU NVIDIA V100 và một máy tính xách tay MacBook Pro (CPU Intel i7 2.6GHz). Ngoài ra, chúng tôi đã sử dụng các phép tính dấu phẩy động 32 bit cho tất cả các mô hình mà không có bất kỳ phương pháp tối ưu hóa nào.

Để xác nhận sự đánh đổi giữa hiệu quả tổng hợp và chất lượng mẫu, chúng tôi đã tiến hành các thử nghiệm dựa trên ba biến thể của generator, V1, V2 và V3 trong khi vẫn duy trì cấu hình bộ phân biệt giống nhau. Đối với V1, chúng tôi đặt hu = 512, ku = [16, 16, 4, 4], kr = [3, 7, 11] và Dr = [[1, 1], [3, 1], [5 , 1]] × 3]. V2 đơn giản là một phiên bản nhỏ hơn của V1, có kích thước ẩn nhỏ hơn hu = 128 nhưng có các trường tiếp nhận hoàn toàn giống nhau. Để tiếp tục giảm số lượng lớp trong khi vẫn duy trì các trường tiếp nhận rộng, kích thước hạt nhân và tốc độ giãn nở của V3 đã được lựa chọn cẩn thận. Cấu hình chi tiết của các mô hình được liệt kê trong.



Sử dụng 80 dải mel-spectrogram làm điều kiện đầu vào. Kích thước FFT, cửa sổ và hop được đặt lần lượt là 1024, 1024 và 256. Các mạng được huấn luyện bằng trình tối ưu hóa AdamW (Loshchilov và Hutter, 2017) với β1 = 0,8, β2 = 0,99 và phân rã trọng số λ = 0,01. Tốc độ học giảm dần được lên lịch theo hệ số 0,999 trong mỗi kỷ nguyên với tốc độ học ban đầu là 2 × 10-4.

Thực hiện train và inference trên colab:

<https://drive.google.com/drive/folders/16AhH3B77Eh4ICvRFQs5z4r_tHz4gRK_b?usp=sharing>

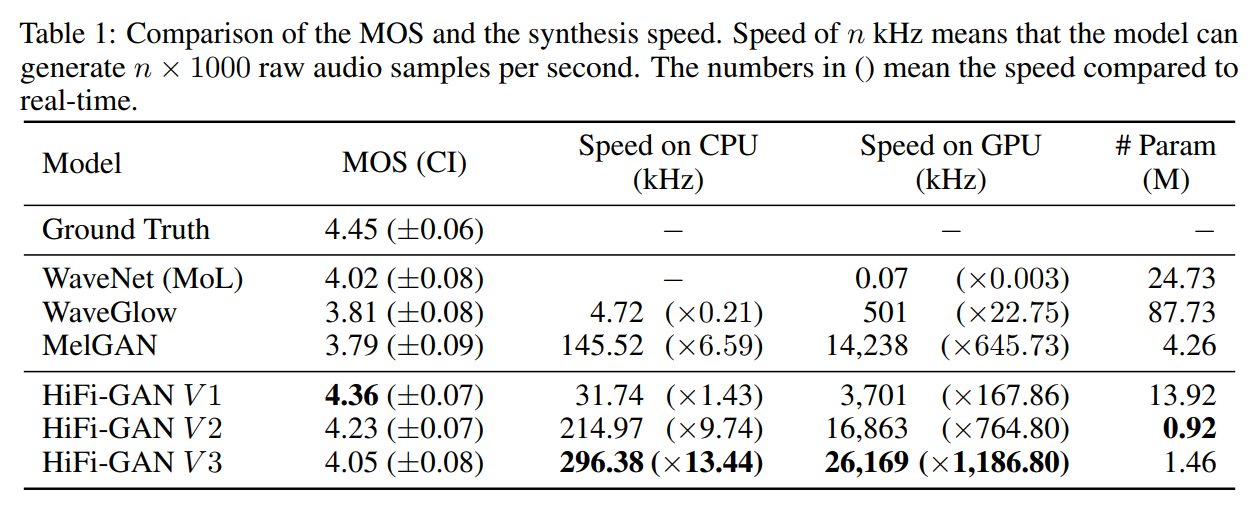
1. **KẾT QUẨ**

**5.1. Audio Quality and Synthesis Speed**

Để đánh giá hiệu suất của các mô hình của chúng tôi về cả chất lượng và tốc độ, chúng tôi đã thực hiện thử nghiệm MOS để đảo ngược phổ và đo tốc độ. Đối với bài kiểm tra MOS, chúng tôi đã chọn ngẫu nhiên 50 cách phát biểu từ bộ dữ liệu LJSpeech và sử dụng biểu đồ phổ chân lý cơ bản của các cách phát biểu đã bị loại khỏi quá trình đào tạo làm điều kiện đầu vào.

Để dễ dàng so sánh chất lượng âm thanh, tốc độ tổng hợp và kích thước mô hình, các kết quả được tổng hợp và trình bày trong Bảng 1. Đáng chú ý, tất cả các biến thể của HiFi-GAN đều đạt điểm cao hơn các mẫu khác. V 1 có thông số 13,92M và đạt MOS cao nhất với khoảng cách 0,09 so với âm thanh trung thực; điều này ngụ ý rằng âm thanh tổng hợp gần như không thể phân biệt được với giọng nói của con người.

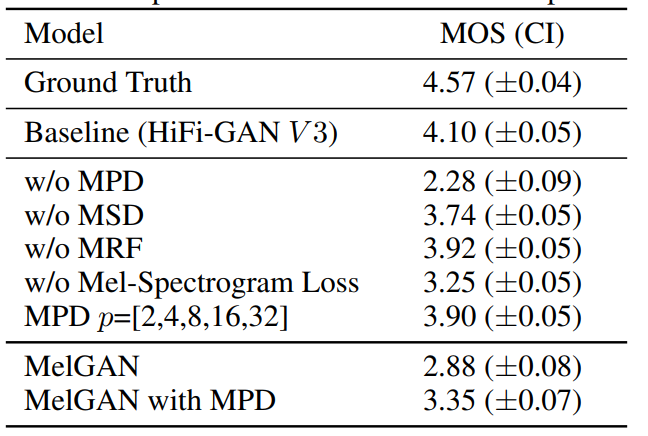
Về tốc độ tổng hợp, V1 nhanh hơn WaveGlow và MoL WaveNet. V2 cũng thể hiện chất lượng tương tự như con người với MOS là 4,23 trong khi giảm đáng kể yêu cầu bộ nhớ và thời gian suy luận so với V 1. Nó chỉ yêu cầu 0,92M tham số. Mặc dù có MOS thấp nhất trong số các kiểu máy của chúng tôi, V3 có thể tổng hợp giọng nói nhanh hơn 13,44 lần so với thời gian thực trên CPU và nhanh hơn 1.186 lần so với thời gian thực trên một GPU V100 trong khi vẫn cho thấy chất lượng cảm nhận tương tự với MoL WaveNet. Vì V3 tổng hợp lời nói trên CPU một cách hiệu quả nên nó có thể rất phù hợp cho các ứng dụng trên thiết bị.



**5.2. Ablation Study**

Nhóm tác giả đã thực hiện một nghiên cứu cắt bỏ MPD, MRF và mel-spectrogram để xác minh ảnh hưởng của từng thành phần HiFi-GAN đối với chất lượng của âm thanh tổng hợp. V3 có công suất biểu đạt nhỏ nhất trong số ba biến thể của bộ tạo đã được sử dụng làm bộ tạo cho nghiên cứu cắt bỏ và các tham số mạng được cập nhật lên tới 500 nghìn bước cho mỗi cấu hình.

Kết quả đánh giá MOS được thể hiện trong Bảng 2, cho thấy cả ba thành phần đều góp phần vào hiệu suất. Việc loại bỏ MPD làm giảm đáng kể chất lượng cảm nhận, trong khi sự vắng mặt của MSD cho thấy sự xuống cấp tương đối nhỏ nhưng đáng chú ý. Để điều tra ảnh hưởng của MRF, một khối còn lại có trường tiếp nhận rộng nhất được giữ lại trong mỗi mô-đun MRF. Kết quả cũng tệ hơn so với đường cơ sở. Thử nghiệm về suy hao mel-spectrogram cho thấy nó giúp cải thiện chất lượng và chúng tôi quan sát thấy rằng chất lượng cải thiện ổn định hơn khi áp dụng suy hao.

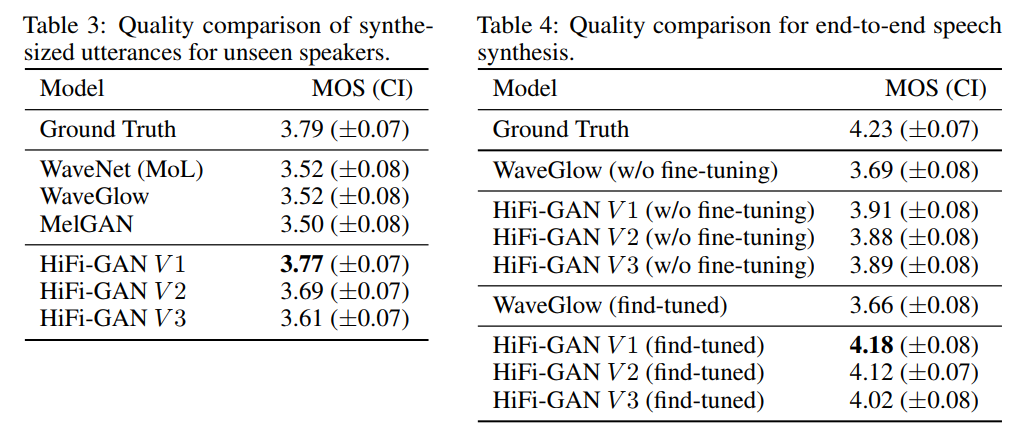


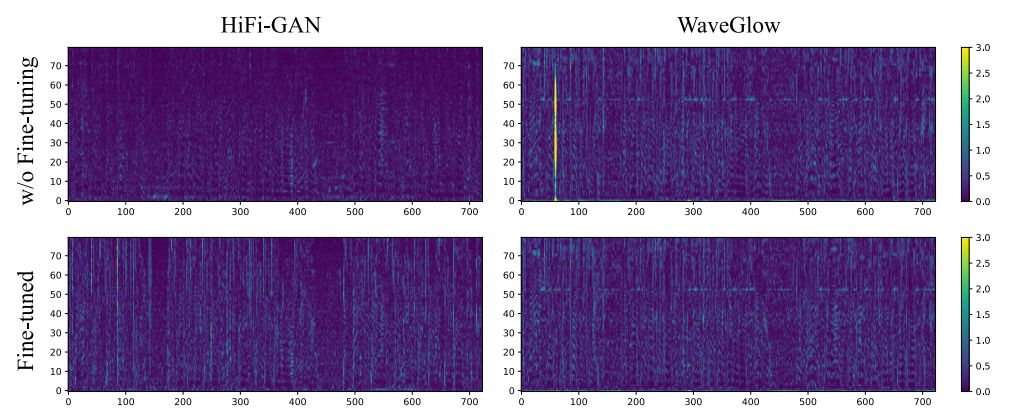
Để xác minh tác dụng của MPD trong cài đặt của các kiểu GAN khác, chúng tôi đã giới thiệu MPD trong MelGAN. MelGAN được đào tạo với MPD vượt trội so với bản gốc với khoảng cách 0,47 MOS, điều này cho thấy sự cải thiện đáng kể về mặt thống kê.

Thực nghiệm với các chu kỳ lũy thừa bằng 2 để xác minh ảnh hưởng của các chu kỳ được đặt thành các số nguyên tố. Mặc dù giai đoạn 2 cho phép các tín hiệu được xử lý chặt chẽ, nhưng nó dẫn đến sự xuống cấp đáng kể về mặt thống kê với mức chênh lệch 0,20 MOS so với baseline

**5.3. Generalization to Unseen Speakers**

Sử dụng 50 cách phát biểu được chọn ngẫu nhiên của chín diễn giả chưa từng thấy trong tập dữ liệu VCTK đã bị loại khỏi tập huấn luyện cho bài kiểm tra MOS. Bảng 3 cho thấy kết quả thử nghiệm đối với mel-spectrogram đảo ngược của người nói không nhìn thấy. Ba biến thể generator đạt điểm 3,77, 3,69 và 3,61. Tất cả chúng đều tốt hơn AR và các mô hình dựa trên dòng chảy, cho thấy rằng các mô hình được đề xuất khái quát tốt cho người nói không nhìn thấy. Ngoài ra, xu hướng khác biệt về điểm số MOS của các mô hình đề xuất cũng tương tự như kết quả trình bày trong Phần 5.1, thể hiện sự tổng quát hóa trên các bộ dữ liệu khác nhau.





*Hình 3: Sự khác biệt về điểm ảnh trong miền mel-spectrogram giữa các dạng sóng được tạo và một melspectrogram từ Tacotron2. Trước khi tinh chỉnh, HiFi GAN tạo ra các dạng sóng tương ứng với các điều kiện đầu vào một cách chính xác. Sau khi tinh chỉnh, sai số của mức mel-spectrogram tăng lên nhưng chất lượng cảm quan tăng lên.*

**5.4. End-to-End Speech Synthesis**

Tiến hành một thử nghiệm bổ sung để kiểm tra tính hiệu quả của các mô hình được đề xuất khi áp dụng cho quy trình tổng hợp giọng nói từ đầu đến cuối, bao gồm văn bản tới mel-spectrogram và mel spectrogram cho các mô đun tổng hợp dạng sóng. Ở đây, sử dụng Tacotron2 (Shen et al., 2018) để tạo mel-spectrogram từ văn bản. Không có bất kỳ sửa đổi nào, nhóm tác giả đã tổng hợp các mel-spectrogram bằng cách triển khai Tacotron2 (Valle, 2018a) phổ biến nhất với các trọng số pretrain được cung cấp. HiFi GAN V1 (không tinh chỉnh) 3,91 (±0,08) HiFi-GAN V3 Thực địa sóng phát sáng Hình 3: Sự khác biệt về điểm ảnh trong miền mel-spectrogram giữa các dạng sóng được tạo và một melspectrogram từ Tacotron2. Trước khi tinh chỉnh, HiFi GAN tạo ra các dạng sóng tương ứng với các điều kiện đầu vào một cách chính xác. Sau khi tinh chỉnh, sai số của mức mel-spectrogram tăng lên nhưng chất lượng cảm quan tăng lên. Sau đó, chúng tôi cung cấp chúng làm điều kiện đầu vào cho các mô hình giai đoạn hai, bao gồm các mô hình của chúng tôi và WaveGlow được sử dụng trong Phần 5.1

Điểm số MOS được liệt kê trong Bảng 4. Kết quả khi chưa tinh chỉnh cho thấy rằng tất cả các mẫu được đề xuất đều hoạt động tốt hơn WaveGlow trong cài đặt tổng thể, trong khi chất lượng âm thanh của tất cả các mẫu đều không đạt yêu cầu so với âm thanh trung thực. Tuy nhiên, khi sự khác biệt về pixel trong miền mel-spectrogram giữa các dạng sóng được tạo và một mel-spectrogram từ Tacotron2 là được điều tra như thể hiện trong Hình 3, chúng tôi thấy rằng sự khác biệt là không đáng kể, điều đó có nghĩa là phổ mel dự đoán từ Tacotron2 đã bị nhiễu. Để cải thiện chất lượng âm thanh trong cài đặt đầu cuối, chúng tôi đã áp dụng tinh chỉnh với mel-spectrogram dự đoán của Tacotron2 ở chế độ bắt buộc của giáo viên (Shen và cộng sự, 2018) cho tất cả các kiểu máy có tối đa 100 nghìn bước. Điểm MOS của tất cả các mẫu được tinh chỉnh đề xuất đều trên 4, trong khi WaveGlow tinh chỉnh không cho thấy sự cải thiện về chất lượng. Chúng tôi xin cảm ơn Bokyung Son, Sungwon Kim, Yongjin Cho và Sungwon Lyu.9 Kết luận rằng HiFi-GAN thích ứng tốt trên cài đặt đầu cuối với tinh chỉnh.

1. **KẾT LUẬN**

Trong công việc này, nhóm tác giả đã giới thiệu HiFi-GAN, có thể tổng hợp hiệu quả âm thanh giọng nói chất lượng cao. Trên hết, mô hình được đề xuất vượt trội so với các mô hình có sẵn công khai hoạt động tốt nhất về chất lượng tổng hợp, thậm chí có thể so sánh với mức độ của con người. Hơn nữa, nó cho thấy một sự cải thiện đáng kể về tốc độ tổng hợp. Lấy cảm hứng từ đặc điểm của âm thanh giọng nói bao gồm các mẫu với các giai đoạn khác nhau và áp dụng nó vào mạng thần kinh, đồng thời xác minh rằng sự tồn tại của bộ phân biệt đối xử được đề xuất ảnh hưởng lớn đến chất lượng tổng hợp giọng nói thông qua nghiên cứu cắt bỏ. Ngoài ra, công việc này trình bày một số thí nghiệm có ý nghĩa trong các ứng dụng tổng hợp giọng nói. HiFi-GAN cho thấy khả năng khái quát hóa những người nói không nhìn thấy và tổng hợp âm thanh giọng nói tương đương với chất lượng của con người từ đầu vào ồn ào trong cài đặt đầu cuối. Ngoài ra, mô hình dấu chân nhỏ của chúng tôi thể hiện chất lượng mẫu tương đương với đối tác tự hồi quy công khai tốt nhất hiện có, đồng thời tạo mẫu theo thứ tự cường độ nhanh hơn thời gian thực trên CPU. Điều này cho thấy sự tiến bộ đối với quá trình tổng hợp giọng nói tự nhiên trên thiết bị, yêu cầu độ trễ thấp và dung lượng bộ nhớ. Cuối cùng, các thử nghiệm của chúng tôi cho thấy rằng các trình tạo có cấu hình khác nhau có thể được đào tạo với cùng một bộ phân biệt đối xử và cơ chế học tập, điều này cho thấy khả năng lựa chọn linh hoạt cấu hình trình tạo theo thông số kỹ thuật đích mà không cần tìm kiếm siêu tham số tốn thời gian cho những người phân biệt đối xử. Phát hành HiFi-GAN dưới dạng mã nguồn mở. Nhóm tác giả dự kiến rằng công trình của họ sẽ là cơ sở cho các nghiên cứu tổng hợp giọng nói trong tương lai.