BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Ngọc Phương

NGHIÊN CỨU PHÁT TRIỂN HỆ THỐNG THÍCH NGHI GIỌNG NÓI TRONG TỔNG HỢP TIẾNG VIỆT VÀ ỨNG DỤNG

TÓM TẮT LUẬN ÁN TIẾN SỸ NGÀNH HỆ THỐNG THÔNG TIN
Mã số: 9 48 01 04

Hà Nội – 2023

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học: PGS.TS. Lương Chi Mai

Phản biện 1: …
Phản biện 2: …
Phản biện 3: ….

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án tiến sĩ, họp tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi … giờ ..', ngày … tháng … năm 2023

Có thể tìm hiểu luận án tại:

- Thư viện Học viện Khoa học và Công nghệ
- Thư viện Quốc gia Việt Nam

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN
LIÊN QUAN ĐẾN LUẬN ÁN

1. **[**CT1] Pham Ngoc Phuong, Tran Quang Chung, Luong Chi Mai: "Adapt-TTS: High-quality zero-shot multi-speaker text-to-speech adaptive-based for Vietnamese". *Journal of Computer Science and Cybernetics,* V.39, N.2 (2023), pp. 159-173. 1-DOI: 10.15625/1813-9663/18136, VietNam.
2. [CT2] Pham Ngoc Phuong, Tran Quang Chung, Luong Chi Mai: "Improving few-shot multi-speaker text-to-speech adaptive-based with Extracting Mel-vector (EMV) for Vietnamese". *International Journal of Asian Language Processing*, 2023, Vol. 32, No. 02n03, 2350004, pp. 1-15, Singapore.
3. [CT3] Pham Ngoc Phuong, Tran Quang Chung, Do Quoc Truong, Luong Chi Mai: "A study on neural-network-based Text-to-Speech adaptation techniques for Vietnamese", *International Conference on Speech Database and Assessments (Oriental COCOSDA) 2021*, pp. 199-205. IEEE, Singapore.
4. [CT4] Pham Ngoc Phuong, Tran Quang Chung, Nguyen Quang Minh, Do Quoc Truong, Luong Chi Mai: "Improving prosodic phrasing of Vietnamese text-to-speech systems", *Association for Computational Linguistics, 7th International Workshop on Vietnamese Language and Speech Processing*, 12/2020, pp. 19-23, VietNam.
5. [CT5] Nguyen Thai Binh, Nguyen Vu Bao Hung, Nguyen Thi Thu Hien, Pham Ngoc Phuong, Nguyen The Loc, Do Quoc Truong, Luong Chi Mai: "Fast and Accurate Capitalization and Punctuation for Automatic Speech Recognition Using Transformer and Chunk Merging", *International Conference on Speech Database and Assessments (Oriental COCOSDA)* 2019, IEEE, pp. 1-5, Philippines.
6. [CT6] Pham Ngoc Phuong, Do Quoc Truong, Luong Chi Mai: "A high quality and phonetic balanced speech corpus for Vietnamese" *International Conference on Speech Database and Assessments (Oriental COCOSDA)* 2018, pp. 1-5 Japan.
7. [CT7] Tác giả Bảo hộ quyền sở hữu trí tuệ "Phần mềm chuyển đổi văn bản thành giọng nói Adapt-TTS "số 7590/QTG ngày 26/9/2022 tại *Cục Bản quyền tác giả*

MỞ ĐẦU

Thông thường, để xây dựng được tiếng nói tổng hợp với đặc trưng của một người nói cụ thể, cần thu âm một lượng lớn dữ liệu (khoảng 10 giờ trong môi trường phòng thu tiêu chuẩn) của chính giọng nói đó để huấn luyện. Điều này khiến việc tạo ra các giọng nói tổng hợp theo yêu cầu rất tốn kém về chi phí và mất nhiều thời gian, khó thực hiện với các ngôn ngữ nghèo tài nguyên như tiếng Việt. Hơn nữa, hiện nay tổng hợp tiếng nói có các yêu cầu cao hơn so với việc chỉ sử dụng giọng đọc có sẵn, đó là các nhu cầu xây dựng giọng nói riêng, giọng đọc cá nhân hóa, hay nhu cầu phục hồi hoặc nhân bản giọng. Các nghiên cứu điều chỉnh, biến đổi tham số đặc trưng giọng nói và thích nghi người nói chỉ đa phần mới chỉ được áp dụng trong các công trình nghiên cứu của các tác giả nước ngoài trên các ngôn ngữ phổ biến như tiếng Anh, Nhật, Trung. Tại Việt Nam các nghiên cứu này vẫn tiếp cận phương pháp tổng hợp thích nghi dựa trên HMM và cho chất lượng tổng hợp thấp.

Câu hỏi nghiên cứu: Phương pháp nào giúp tổng hợp tiếng nói đảm bảo chất lượng cho ngôn ngữ nghèo tài nguyên như tiếng Việt trong khi chỉ có vài phút mẫu thích nghi? Cần tối thiểu bao nhiêu dữ liệu thích nghi (được huấn luyện cùng hệ thống) để đảm bảo giọng tổng hợp đạt được chất lượng và độ tương đồng cao? Nếu thích nghi bằng mẫu dữ liệu chỉ vài giây và không cần huấn luyện lại mô hình thì hệ thống có thể thực hiện được không và lượng mẫu thích nghi tối thiểu cần bao nhiêu?

Mục tiêu chính của luận án: nghiên cứu và xây dựng được hệ thống tổng hợp tiếng nói tiếng Việt bằng các kỹ thuật huấn luyện thích nghi các đặc trưng âm học của người nói dựa trên DNN nhằm: 1) Nâng cao chất lượng tổng hợp tiếng nói dựa trên thích nghi bằng các đề xuất cải tiến về độ tự nhiên; 2) Tổng hợp giọng nói mới mang các đặc trưng âm học của giọng nói đích với chất lượng và độ tương đồng cao trong khi chỉ cần sử dụng một lượng dữ liệu mẫu nhỏ; 3) Tổng hợp giọng nói tức thì với lượng mẫu nhỏ mà không cần tốn chi phí huấn luyện lại.

Đóng góp của luận án: 1) Đề xuất hai mô hình tổng hợp thích nghi phụ thuộc người nói dựa trên DNN với điều kiện ít dữ liệu mẫu huấn luyện nhưng tạo ra giọng mới tốt nhất có thể *(Few-shot TTS)*: i) Mô hình tổng hợp thích nghi phụ thuộc người nói dựa trên học chuyển đổi (transfer-learning); ii) Mô hình tổng hợp thích nghi phụ thuộc người nói dựa trên vector biểu diễn đặc trưng; 2) Đề xuất mô hình tổng hợp thích nghi độc lập người nói dựa trên DNN với điều kiện chỉ cần một vài câu mẫu mà không cần huấn luyện lại mô hình nhưng vẫn tạo một giọng mới chấp nhận được *(Zero-shot TTS*); 3) Xây dựng được bộ cơ sở dữ liệu (CSDL) tiếng nói tiếng Việt đảm bảo chất lượng làm bộ dữ liệu cơ sở cho nhiệm vụ huấn luyện mô hình tổng hợp và thích nghi. Phương pháp xây dựng bộ CSDL chi phí thấp và các cải cải tiến gán nhãn nhằm tăng cường độ tự nhiên; 4) Xây dựng được ứng dụng thích nghi đa người nói sử dụng được trên các thiết bị đa nền tảng.

Đối tượng và phạm vi nghiên cứu của luận án: *Hệ thống tổng hợp tiếng nói tiếng Việt có thể cá nhân hóa bằng phương pháp thích nghi trong điều kiện số lượng mẫu thích nghi hạn chế có huấn luyện và không phải huấn luyện lại*. Nghiên cứu cũng sẽ xây dựng ứng dụng cho việc bắt chước hoặc phục hồi giọng được tích hợp hoặc chạy trên các nền máy tính đa nền tảng. Dữ liệu huấn luyện và dữ liệu mẫu (giọng đích) được chọn giới hạn ở giọng miền Bắc và giọng miền Nam với phong cách đọc thông tin thời sự chủ đề chính trị, xã hội.

Cấu trúc luận án gồm các phần:

Chương 1: Giới thiệu tổng quan về tổng hợp tiếng nói và tổng hợp tiếng nói với khả năng điều chỉnh đặc trưng đầu ra. Cấu trúc tổng quan của một hệ thống tổng hợp tiếng nói dựa trên thích nghi cơ bản. Tổng quan tình hình nghiên cứu về tổng hợp tiếng nói dựa trên thích nghi nói chung và thích nghi tiếng Việt nói riêng. Giới thiệu các mục tiêu và phạm vi nghiên cứu chính của luận án.

Chương 2: Xây dựng bộ cơ sở dữ liệu (CSDL) tiếng Việt cho hệ thống tổng hợp và thích nghi và các quy trình kèm theo nhằm nâng cao chất lượng, giảm chi phí khi xây dựng bộ CSDL đa người nói cho các hệ thống tổng hợp tiếng Việt. Bên cạnh phương pháp bổ sung thông tin nhãn như chèn điểm dừng lấy hơi và phiên âm từ mượn giúp tăng cường độ tự nhiên của mô hình tổng hợp. Bộ CSDL tiếng và kỹ thuật tăng cường nhãn thông tin này cũng chính là phần cơ sở để xây dựng các mô hình thích nghi ở các chương tiếp theo.

Chương 3: Trình bày phương pháp cải tiến chất lượng mô hình tổng hợp tiếng nói dựa trên thích nghi thông qua hai đề xuất: 1) Cải tiến mô hình tổng hợp thích nghi (Few-shot TTS) bằng Muli-pass fine-tune dựa trên kỹ thuật học chuyển đổi người nói và ngôn ngữ (transfer-learning) với lượng mẫu phải học ít hơn nhiều so với huấn luyện mô hình cơ sở và 2) Cải tiến mô hình tổng hợp thích nghi (Few-shot TTS) bằng vector EMV biểu diễn đặc trưng người nói chỉ với vài câu nói. Cả hai kỹ thuật thích nghi đều yêu cầu dữ liệu mẫu phải có trong tập huấn luyện và với các mô hình đề xuất hướng tới sử dụng lượng dữ liệu thích nghi ít dần.

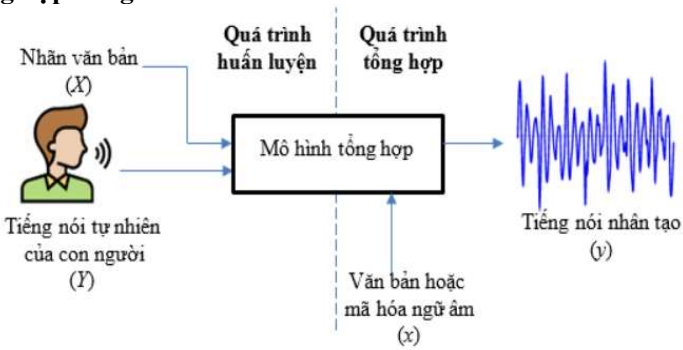
Chương 4: Đề xuất phương pháp nâng cao hiệu năng của mô hình tổng hợp thích nghi chi phí thấp với điều kiện mẫu ít nhất có thể mà không cần huấn luyện lại mô hình (Zero-shot TTS) thông qua hai kỹ thuật: 1) Áp dụng vector biểu diễn đặc trưng người nói hiệu quả; 2) Mô hình khử nhiễu khuếch tán phổ Mel (Mel-spectrogram denoiser) cho phép tổng hợp âm thanh chất lượng cao hơn so với các mô hình cơ sở. Mô hình tổng hợp dựa thích nghi bằng Zero-shot TTS không đòi hỏi dữ liệu thích nghi phải có trong tập huấn luyện và chỉ sử dụng duy nhất một câu mẫu của người nói để thích nghi. Hướng tiếp cận này giúp đơn giản hóa trong việc tổng hợp giọng mới và mở rộng khả năng ứng dụng của các mô hình tổng hợp thích nghi.

Kết luận : Trình bày các đóng góp chính của luận án và chỉ ra các hạn chế và hướng phát triển tiếp theo.

Chương 1. CÁC NGHIÊN CỨU LIÊN QUAN VÀ KIẾN THỨC CƠ SỞ VỀ TỔNG HỢP VÀ THÍCH NGHI TIẾNG NÓI

Trong Chương 1, phần đầu tiên giới thiệu tổng quan các nghiên cứu liên quan về hệ thống tổng hợp tiếng nói và các vấn đề khó khăn cần giải quyết. Tiếp theo, trình bày về nhu cầu tổng hợp tiếng nói với khả năng điều chỉnh đặc trưng đầu ra và các nghiên cứu liên quan về tổng hợp tiếng nói thích nghi và ứng dụng. Sau đó, mô tả các kiến thức sở và các thành phần chính của một hệ thống tổng hợp dựa trên thích nghi, các đánh giá chất lượng tổng hợp dựa trên thích nghi, tổng quan về tình hình nghiên cứu trong và ngoài nước và cuối cùng là xác định các hướng nghiên cứu chính và phạm vi của luận án.

1.1. Tổng hợp tiếng nói



Hình 1 : Mô hình tổng hợp tiếng nói nhân tạo

Khái niệm tổng hợp tiếng nói

Tổng hợp tiếng nói (Speech synthesis) là quá trình tạo ra tiếng nói con người một cách nhân tạo từ đầu vào là văn bản hoặc các mã hóa ngữ âm. Tổng hợp tiếng nói chính là một phần trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Tổng hợp tiếng nói từ văn bản (Text to speech – viết tắt là TTS) là một công nghệ quan trọng trong tổng hợp tiếng, công nghệ này tạo ra sóng âm tiếng nói đầu ra một cách tùy ý từ văn bản bằng đầu vào.

Có thể mô tả hệ thống TTS bằng mô hình tính xác suất phân phối dự đoán:

$$p(y|x, Y, X)$$

trong đó, Y là âm thanh tiếng nói dùng để huấn luyện và X là văn bản gán nhãn tương ứng, x là văn bản đầu vào và y là tiếng nói cần tổng hợp.

1.1.1. Phân loại các phương pháp tổng hợp tiếng nói

Hai kiến trúc TTS phổ biến tiên tiến nhất hiện nay là: 1) *Kiến trúc tự hồi quy* (autoregressive) và; 2) *Kiến trúc không tự động hồi quy* (non-autoregressive). Mỗi một kiến trúc có những ưu nhược điểm khác nhau.

1.1.2. Tổng hợp tiếng nói với khả năng điều chỉnh đặc trưng đầu ra

Hiện nay, các kỹ thuật DNN đã thay thế hoàn toàn mô hình HMM trong việc xây dựng mô hình âm học và mô hình trường độ. Các mô hình DNN chỉ yêu cầu một lần tính toán duy nhất để dự đoán đặc trưng, làm cho nó phù hợp hơn cho việc

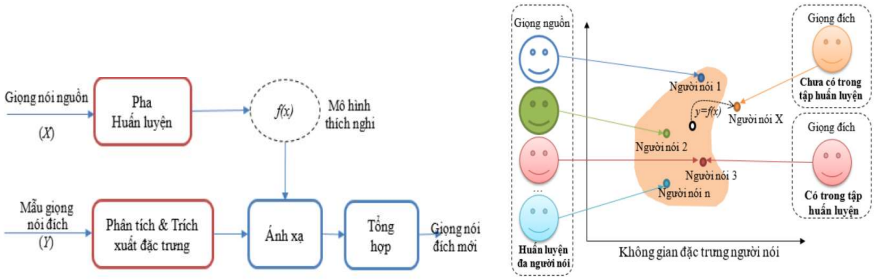
tổng hợp theo thời gian thực. Tuy nhiên, các nghiên cứu hiện tại về mô hình DNN tập trung chủ yếu cho mô hình phụ thuộc người nói, điều này đòi hỏi một lượng dữ liệu đáng kể từ một người nói duy nhất để tạo ra một mô hình âm học ổn định.

1.2. Thích nghi trong tổng hợp tiếng nói

1.2.1. Khái niệm

Thích nghi trong tổng hợp tiếng nói (hay gọi tắt là '*thích nghi TTS'* hoặc *'TTS thích nghi'*) là khả năng tổng hợp giọng nói tùy ý từ bất kỳ người nào với một lượng dữ liệu mẫu thực nhỏ (*reference speech*), giọng nói tổng hợp sẽ mang đặc trưng của giọng nói đích (*target speaker*) với các đặc trưng của giọng nói (*voice characteristics*) và các đặc trưng ngữ điệu (*prosodic features*). Thích nghi TTS được gọi với các thuật ngữ khác nhau trong học thuật và công nghiệp, chẳng hạn như thích nghi giọng nói (*voice adaptation*), nhân bản giọng nói (*voice cloning*), cá nhân hóa giọng nói (*custom voice*).

Biểu diễn tín hiệu giọng nói nguồn và giọng nói đích lần lượt là *X* và *Y,* biểu diễn các đặc trưng tiếng nói nguồn và đích là *x* và *y*, hàm chuyển đổi có thể xây dựng như sau: *y= f(x)* trong đó *f(.)* còn được gọi là hàm ánh xạ theo khung.



Hình 2: Mô hình tổng quát và không gian đặc trưng hệ thống TTS dựa trên thích nghi

Thích nghi TTS được xem xét từ hai khía cạnh: *1) Thiết lập thích nghi chung*, bao gồm các cải tiến về tổng quát hóa mô hình TTS nguồn để hỗ trợ giọng nói mới; *2) Thích nghi hiệu quả*, bao gồm việc giảm dữ liệu thích nghi và các thông số thích nghi

1.2.2. Thích nghi chung

1.2.3. Thích nghi hiệu quả

Mục tiêu thích nghi sử dụng càng ít dữ liệu và tham số nhưng vẫn đạt được chất lượng thích nghi giọng nói cao. Có thể chia các nghiên cứu trong thể loại này thành một số nhóm*: 1) Thích nghi ít dữ liệu; 2) Thích nghi ít tham số; 3) Thích nghi dữ liệu chưa được gán nhãn; 4) Thích nghi không phải huấn luyện lại mô hình (Zero-shot TTS).*

1.3. Tình hình nghiên cứu hiện nay về tổng hợp thích nghi

1.3.1. Một số nghiên cứu gần đây trên một số ngôn ngữ khác

- Thích nghi ít dữ liệu Few-shot TTS. Có hai phương pháp tiếp cận chính trong hướng này, đó là: 1) Thích nghi thông qua tinh chỉnh một phần mô hình hoặc toàn bộ mô hình; 2) Thích nghi dựa trên vector biểu diễn đặc trưng người nói. Một số nghiên cứu đã xây dựng thích nghi few-shot bằng cách chỉ sử dụng một vài cặp dữ liệu văn bản và giọng, thay đổi từ vài phút đến vài giây. Chien và cộng sự khám phá một vài kiểu speaker embedding khác nhau để thích nghi few-shot. Yue và cộng sự dùng tận dụng chuỗi tiếng nói để thích nghi few-shot. Chen và cộng sự, Arık và cộng sự so sánh chất lượng giọng nói với các lượng dữ liệu thích nghi khác nhau và thấy rằng chất lượng giọng nói cải thiện nhanh chóng với sự gia tăng của dữ liệu thích nghi khi kích thước dữ liệu nhỏ (dưới 20 câu) và cải thiện chậm với hàng chục câu thích nghi.

- Thích nghi ít tham số Một số công trình đề xuất giảm thông số thích nghi càng ít càng tốt trong khi vẫn duy trì chất lượng thích nghi. AdaSpeech đề xuất chuẩn hóa lớp có điều kiện để tạo các tham số thang đo và độ lệch trong chuẩn hóa lớp từ các speaker embedding dựa trên việc tạo tham số theo ngữ cảnh và chỉ tinh chỉnh các tham số liên quan đến chuẩn hóa lớp có điều kiện và speaker embedding để đạt được chất lượng thích nghi. Moss và cộng sự đề xuất một phương pháp tinh chỉnh chọn các siêu tham số mô hình khác nhau cho nhiều giọng nói khác nhau dựa trên tối ưu hóa Bayes.

- Thích nghi Zero-shot. Một số nghiên cứu tiến hành thích nghi zero-shot, sử dụng speaker embedding nói để trích xuất các speaker embedding của các âm thanh mẫu. Kịch bản này khá hấp dẫn vì không cần dữ liệu và tham số thích nghi. Tuy nhiên, chất lượng thích nghi không đủ tốt, đặc biệt khi giọng nói đích (target-speaker) rất khác với giọng nói nguồn (source-speaker).

1.3.2. Các nghiên cứu End-to-end cho tổng hợp tiếng Việt

Theo thống kê của VLSP, trong 4 năm tổ chức đánh giá, vào năm 2018 các mô hình DNN đạt ưu thế trong VLSP, vào hai năm 2019-2020 các nhóm nghiên cứu tập trung vào sử dụng các mô hình Tacotron2 chiếm ưu thế để phát triển các hệ thống tổng hợp tiếng Việt (với mô hình âm học sử dụng Tacotron2 kết hợp với Vocoder phổ biến như Waveglow hoặc HifiGAN). Cũng theo nghiên cứu từ tổ chức này, từ những năm 2021 trở đi các nhóm nghiên cứu tổng hợp tiếng Việt đã tập trung sử dụng mô hình Fastspeech2 chiếm ưu thế, một số nhóm đề xuất sử dụng VITS cũng đạt được các kết quả nổi bật. Có thể thấy rằng, các nghiên cứu End-to-end cho tổng hợp tiếng Việt đã khá cập nhật với nghiên cứu quốc tế, tuy nhiên các vấn đề thách thức trong các nghiên cứu các tổng hợp tiếng nói với đặc trưng của tiếng Việt như xây dựng bộ CSDL chuyên biệt cho tổng hợp tiếng Việt.

1.3.3. Một số nghiên cứu hiện nay về tổng hợp thích nghi cho tiếng Việt

Kỹ thuật thích nghi dựa trên HMM: SơnPT và NinhDK sử dụng mô hình lai ghép HMM cho tổng hợp tiếng Việt và để điều chỉnh tham số thích nghi khi tổng hợp, nghiên cứu đã đề xuất áp dụng thuật toán hồi quy tuyến tính khả năng cực đại (MLLR) kết hợp với thuật toán cực đại hậu nghiệm (MAP) hoặc

kết hợp MAP với thuật toán làm mịn trường véc tơ (VFS). SơnPT đã chỉ ra ra rằng chỉ với một lượng dữ liệu "đích" hạn chế (100 câu), kết hợp với tập HMM chung của nhiều giọng nói, nhiều đặc trưng và phong cách nói khác nhau, có thể tổng hợp được tiếng nói có chất lượng được cải thiện. Tuy nhiên, cách tiếp cận tổng hợp thích nghi dựa trên tham số thống kê HMM có những nhược điểm cố hữu đó là: 1) Mô hình dựa trên HMM cho chất lượng tổng hợp thấp hơn rất nhiều so với DNN; 2) Các nghiên cứu hiện nay chỉ ra rằng các kỹ thuật tổng hợp thích nghi dựa trên HMM cho chất lượng thấp hơn so với các kỹ thuật DNN; 3) Mô hình thích nghi dựa trên HMM không thể thực hiện được các nhiệm vụ tổng hợp thích nghi với dữ liệu rất nhỏ (chỉ vài câu) hoặc thích nghi không cần huấn luyện lại.

1.4. Kết luận chương 1 và các nội dung nghiên cứu chính của luận án

Tổng hợp tiếng nói dựa trên thích nghi là một bài toán thuộc lĩnh vực chuyển đổi và thích nghi tiếng nói với mục tiêu biến đổi kết quả tổng hợp giọng mới mang các đặc trưng của giọng nói mẫu. Với tiếng Việt, đây là ngôn ngữ nghèo tài nguyên và là ngôn ngữ phức tạp do ngôn ngữ có chứa thành phần ngữ điệu nên tổng hợp tiếng nói cũng như thích nghi tiếng nói trong tổng hợp vẫn là bài toán khó ít người giải. Do vậy, vẫn tồn tại những khó khăn như các bài toán tổng hợp tiếng nói và chuyển đổi thích nghi về chất lượng tổng hợp và chi phí mẫu huấn luyện cho tiếng Việt, có thể kể đến như sau*: 1) Các nghiên cứu về tổng hợp thích nghi cho tiếng Việt còn rất hạn chế, cần có các nghiên cứu đánh giá ảnh hưởng của mẫu tiếng nói đích hạn chế hoặc không có trong quá trình huấn luyện; 2) Các nghiên cứu về thích nghi cho tiếng Việt mới chỉ sử dụng mô hình HMM cho huấn luyện dữ liệu mẫu hạn chế và chưa có mô hình thích nghi cho tiếng Việt nào sử dụng DNN; 3) Chưa có nghiên cứu nào áp dụng mô hình thích nghi End-to-End cho tiếng Việt với dữ liệu mẫu nhỏ có huấn luyện hoặc không huấn luyện; Cần có thêm các nghiên cứu về nâng cao chất lượng tổng hợp thích nghi tiếng nói tiếng Việt; 4) Cần có ứng dụng để đánh giá tính khả thi của các mô hình thích nghi tiếng nói tiếng Việt.*

Từ các vấn đề thực tế trên dẫn đến luận án sẽ tập trung nghiên cứu một số nội dung chính như sau: *1) Xây dựng bộ CSDL phục vụ cho tổng hợp và thích nghi; 2) Nghiên cứu kỹ thuật thích nghi Few-shot TTS dựa trên DNN cho tiếng Việt và đánh giá; 3)Nghiên cứu kỹ thuật thích nghi Zero-shot TTS dựa trên DNN cho tiếng Việt và đánh giá.*

Phạm vi nghiên cứu: Đối tượng nghiên cứu là tiếng nói tiếng Việt; Mô hình nghiên cứu là tổng hợp thích nghi tiếng tiếng Việt nhằm cá nhân hóa giọng nói tổng hợp, cụ thể là nhân bản giọng nói với dữ liệu của đơn người nói (Single-Speaker) và đa người nói. Ứng dụng nhân bản giọng đánh giá tính khả thi là ứng dụng nhân bản giọng.

Chương 2. XÂY DỰNG CƠ SỞ DỮ LIỆU TIẾNG VIỆT CHI PHÍ THẤP CHO TỔNG HỢP VÀ THÍCH NGHI TIẾNG NÓI

Một trong những nhiệm vụ chính của luận án khi nghiên cứu tổng hợp tiếng nói dựa trên thích nghi là nghiên cứu mô hình tổng hợp tiếng nói tiếng Việt. Tuy nhiên, trong nghiên cứu tổng hợp tiếng nói tiếng Việt có hạn chế lớn nhất chính là *thiếu bộ CSDL đủ lớn, đa dạng, đảm bảo chất lượng và chi phí thấp cho các nghiên cứu tổng hợp; Ngoài ra, một trong những vấn đề còn tồn tại của hệ thống tổng hợp tiếng Việt là độ tự nhiên của câu dài và đọc từ mượn*. Chương 2 phân tích các kỹ thuật xây dựng bộ CSDL tiếng cho tổng hợp thích nghi và các phương pháp gán nhãn và phiên âm bổ sung cải thiện ngữ điệu cho hệ thống tổng hợp tiếng nói tiếng Việt, các nội dung gồm: 1) Trình bày phân tích các bộ CSDL cho tổng hợp tiếng nói hiện nay; 2) Trình bày quy trình xây dựng bộ CSDL tiếng đảm bảo chất lượng cho tổng hợp và thích nghi CT6] [CT4]; 3) Một số phương pháp bổ sung thông tin nhãn nhằm tăng cường độ tự nhiên của hệ thống TTS tiếng Việt thông qua các kỹ thuật như thêm dấu câu, chèn điểm dừng lấy hơi và phiên âm từ mượn [CT5] [CT4]; 4) Kết quả xây dựng bộ CSDL.

2.1. Xây dựng bộ CSDL tổng hợp và thích nghi

Các nghiên cứu End-to-end cho tổng hợp tiếng Việt đã khá cập nhật với nghiên cứu quốc tế. Tuy nhiên, các nghiên cứu cũng chỉ ra các vấn đề thách thức trong các nghiên cứu tổng hợp tiếng nói với đặc trưng của tiếng Việt như xây dựng bộ CSDL chuyên biệt cho tổng hợp tiếng Việt, các vấn đề với tổng hợp câu dài hoặc đọc các từ mượn.

2.1.1. Thống kê các bộ CSDL cho tổng hợp hiện nay

*** Một số bộ CSDL tiếng Việt mở cho TTS được công bố:**

Bộ dữ liệu	Thời gian (giờ)	Số người đọc	Số câu	Phong cách đọc
VLSP 2020	9,5	1	7.770	Đọc truyện với nhiều ngữ điệu
VAIS1000	0,2	1	1.000	Phát thanh viên VOV
INFORE	25	1	14.935	Đọc truyện bằng TTS
VietTTS-v1.1	35,9	1	22.884	Đọc truyện bằng TTS

Ngoài ra, có thể thấy rằng tại Việt Nam có không ít những tập đoàn doanh nghiệp đơn vị nghiên cứu lớn, nhưng những bộ CSDL phục vụ cho nghiên cứu TTS lại rất thiếu và hạn chế về cả chất lượng và độ đa dạng. Từ thực tế đó, luận án xác định tính cấp thiết khi xây dựng một bộ cơ sở dữ liệu tiếng nói tiếng Việt đảm bảo chất lượng, chi phí thấp với giọng nói đa dạng (giới tính, độ tuổi, khu vực sống) và chiến lược rõ ràng, phục vụ cho các nghiên cứu tổng hợp và thích nghi tiếng nói.

Khi nghiên cứu tổng hợp hoặc thích nghi tiếng nói, việc tạo một mô hình cho một người nói cụ thể với lượng dữ liệu hạn chế là một thách thức ngay cả với các ngôn ngữ giàu tài nguyên. Đối với tiếng Việt rất khó để thực hiện một nghiên cứu như vậy do yêu cầu về thiết kế dữ liệu. Đầu tiên, âm thanh phải được ghi lại

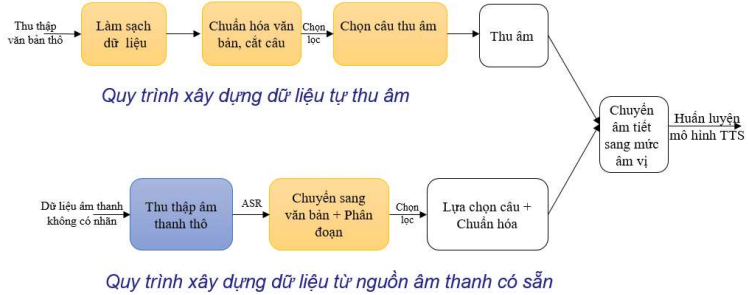
trong một phòng thu cách âm, không nhiễu để đảm bảo giọng thu chất lượng. Thứ hai, người nói phải được chọn để bao hàm nhiều độ tuổi cũng như khu vực sinh sống. Thứ ba, cần chọn lọc để đảm bảo độ cân bằng âm học (đủ các âm vị, từ vựng trong tiếng nói). Như vậy, CSDL đủ tốt để huấn luyện hệ thống tổng hợp và thích nghi tiếng nói phải là sự kết hợp có tính chọn lọc giữa thu thập CSDL từ nguồn âm thanh sẵn có và xây dựng CSDL tự thu âm.

2.1.2. Quy trình xây dựng bộ CSDL cho tổng hợp và thích nghi

Để đảm bảo bộ CSDL cho tổng hợp tiếng đủ lớn và đa dạng cần số cần xác định xây dựng CSDL từ 2 nguồn: Một là tự thu âm; Hai là gán nhãn âm thanh từ nguồn có sẵn. Quy trình mô tả như hình 3 có thể mô tả như sau:

Quy trình xây dựng CSDL tự thu âm : Xây dựng văn bản thu âm (Phát triển các công cụ chọn nội dung, chọn lọc văn bản đảm bảo độ cân bằng âm)→ Chọn giọng đọc (Xác định kiểu giọng đọc, kiểm tra giọng đọc)→ Chuẩn bị thu âm (Quy tắc phát âm, trang tiết bị thu âm, môi trường thu âm, phần mềm thu âm)→ Thu âm (thu âm trên phần mềm, kiểm tra giám sát quá trình thu âm)→Bộ CSDL thu âm chuẩn (rà soát, chọn lọc và tổng hợp).

Quy trình xây dựng CSDL từ nguồn âm thanh có sẵn : Chọn chủ đề audio →Thu thập các nguồn audio/video → Chuyển đổi về audio và chuẩn hóa định dạng → Nhận dạng sơ bộ ra văn bản →Chọn lọc audio tiếng việt dựa vào văn bản → Chia nhỏ audio đã nhận dạng sơ bộ →Phát triển các công cụ gán nhãn →Quy tắc phiên âm→ Gán nhãn → Kiểm tra chéo → Nghiệm thu.



Hình 3: Quy trình xây dựng bộ CSDL tổng hợp và thích nghi

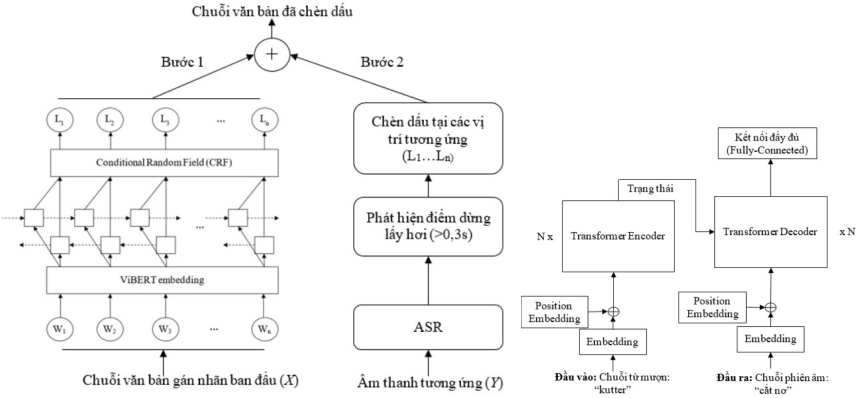
2.1.3. Bổ sung nhãn thông tin cho CSDL

2.1.3.1. Chèn điểm điểm dừng lấy hơi và dấu câu cho văn bản nhãn

Nhãn văn bản được tăng cường thêm thông tin về vị trí chèn dấu câu vào vị trí thích hợp theo hai chiến lược, theo chuẩn tiếng Việt và theo kiểu dừng lấy hơi của người đọc. Để giải quyết bài toán này, luận án đã sử dụng hai giải pháp: Thứ nhất, sử dụng mô hình BERT để khôi phục dấu câu trong câu văn bản gán nhãn; Thứ hai, bắt chước thời gian dừng của người đọc, sử dụng mốc thời gian đo được từ hệ thống ASR có sẵn, nếu thời gian im lặng hơn 0,3 giây sẽ đặt dấu phẩy ở vị trí im lặng này.

Bên cạnh đó, hệ thống cũng thêm một dấu chấm ở cuối văn bản gán nhãn để biểu diễn kết thúc câu nói. Văn bản sau khi xử lý sẽ được đưa đến bước hậu xử

lý phiên âm tiếp theo trước khi dùng huấn luyện mô hình TTS. Kiến trúc mô tả trong Hình 4a.



a) b)

Hình 4: Phương pháp bổ sung nhãn thông tin và phiên âm từ mượn cho CSDL

2.1.3.2. Từ điển phiên âm từ mượn

Để xử lý và giải quyết vấn đề này, cách phát âm tiếng Việt đã được sử dụng để phiên âm các từ tiếng Anh này, ví dụ: "kuttner" sẽ được phát âm bằng "cắt nơ". Để chuyển từ phiên âm nước ngoài sang phiên âm tiếng Việt, luận án đã sử dụng mô hình dịch Transformer cơ sở với kiến trúc mô tả trong Hình 4b. Để huấn luyện mô hình phiên dịch này, phải tạo ra một số lượng lớn các cặp từ Anh-Việt. Kết quả xây dựng được bộ từ điển phiên âm từ mượn phục vụ cho TTS tiếng Việt.

2.1.4. Kết quả

Kết quả đã xây dựng được bộ CSDL 54 đa người nói, bao gồm 26 giọng nam, 28 giọng nữ với phương ngữ Bắc-Nam với độ dài của từng giọng đa dạng từ vài chục phút đến vài tiếng (bao gồm các giọng tự thu âm và thu thập từ nguồn âm thanh có sẵn).

Stt	Bộ dữ liệu	Tổng số câu	Tổng số âm tiết xuất hiện	Số âm tiết trung bình/câu	Số âm tiết khác biệt
I	**Bộ CSDL tự thu âm**				
1.1	Bộ 250 câu chung đa người nói	250	3.268	13,06	1.205
1.2	Bộ 9.600 câu riêng đa người nói	9.600	105.232	10,96	5.516
II	**Bộ CSDL từ nguồn âm thanh có sẵn**				
2.1	Bộ đơn người nói nữ	5.074	100.284	19,76	2.278
2.2	Bộ đơn người nói nam	13.125	280.130	21,34	2.893

2.2. Kết luận chương 2

Chương 2 đã tiến liệt kê các bộ dữ liệu đã công bố cho tổng hợp tiếng nói của các ngôn ngữ giàu tài nguyên và ngôn ngữ tiếng Việt. Phân tích cho thấy, thiếu trầm trọng các bộ CSDL cho nghiên cứu tổng hợp và thích nghi tiếng nói tiếng Việt. Chương 2 cũng trình bày kết quả nghiên cứu xây dựng bộ CSDL chi phí thấp từ hai nguồn dữ liệu tự ghi âm và dữ liệu có sẵn theo các quy trình chặt chẽ để đạt được bộ CSDL đơn người nói và đa người nói đảm bảo chất lượng cho tổng hợp và thích nghi [CT6] [CT3]. Bên cạnh đó, trình bày phương pháp bổ sung nhãn thông tin thông tin văn bản thông qua kỹ thuật chèn dấu kết hợp chèn điểm dừng lấy hơi và phát âm từ mượn để tăng cường độ tự nhiên khi huấn luyện các hệ thống TTS tiếng Việt [CT5][CT4].

Phần này trình bày quy trình và phương pháp. Mục đích của quy trình và phương pháp này để đảm bảo chi phí thấp trong xây dựng CSDL từ dữ liệu không được gắn nhãn sẵn có trên Internet sử dụng ASR chất lượng tốt và một chiến lược thu thập và lọc dữ liệu hiệu quả. Kết quả đã xây dựng được 02 bộ CSDL tiêu chuẩn đơn người nói (1 nam và 1 nữ) và 02 bộ CSDL đa người (250 câu chung và 9.600 câu riêng) và bộ CSDL kiểm thử áp dụng cho tổng hợp và thích nghi. Trong đó, bộ CSDL đa người nói có 54 người nói, gồm 26 giọng nam và 28 giọng nữ với phương ngữ Bắc/Nam với độ dài của từng giọng đa dạng từ vài chục phút đến vài giờ (bao gồm các giọng tự thu âm và thu thập từ nguồn âm thanh có sẵn). Đây là nền tảng quan trọng để xây dựng các hệ thống tổng hợp và thích nghi tiếng Việt trong các Chương tiếp theo.

Chương 3. MÔ HÌNH TỔNG HỢP THÍCH NGHI CÓ HUẤN LUYỆN VỚI MẪU NHỎ (FEW-SHOT TTS)

Chương 3 sẽ trả lời cho câu hỏi: Phương pháp nào giúp tổng hợp tiếng nói tốt cho ngôn ngữ nghèo tài nguyên như tiếng Việt trong khi chỉ có vài phút mẫu thích nghi? Cần tối thiểu bao nhiêu dữ liệu thích nghi (được huấn luyện cùng hệ thống) để đảm bảo giọng tổng hợp đạt được chất lượng và độ tương đồng cao? Chương 3 sẽ trình bày các đề xuất cải tiến mô hình tổng hợp dựa trên thích nghi tổng hợp giọng nói tiếng Việt chất lượng cao, nội dung gồm: 1) Trình bày phương pháp thích nghi tiếng nói tiếng Việt dùng Multi-pass fine-tune [CT3]; 2) Trình bày phương pháp thích nghi tiếng nói tiếng Việt dùng vector đặc trưng người nói (EMV) [CT2].

3.1. Thích nghi cho tổng hợp tiếng và các phương pháp

3.1.1. Phương pháp

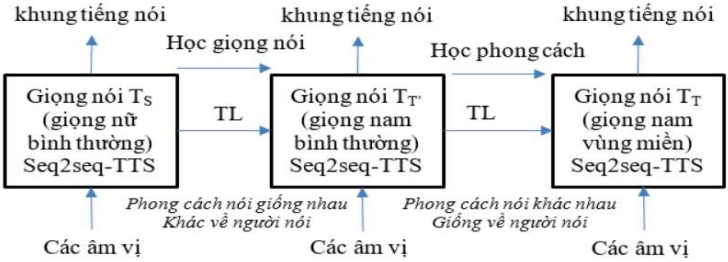
1) Phương pháp thích nghi mô hình dựa trên fine-tune lại mô hình: Phương pháp này sử dụng dữ liệu của đa người nói để huấn luyện một mô hình trung bình. Đối với một người nói mục tiêu cụ thể, mô hình trung bình được điều chỉnh với một lượng nhỏ dữ liệu mục tiêu. Sự thích nghi có thể được đạt được bằng cách điều chỉnh lại tất cả các tham số của mô hình hoặc một phần của chúng.

2) Phương pháp thích nghi mô hình dựa trên mã hóa đặc trưng người nói: Trong phương pháp này, một vector hoặc mạng nhúng được sử dụng để

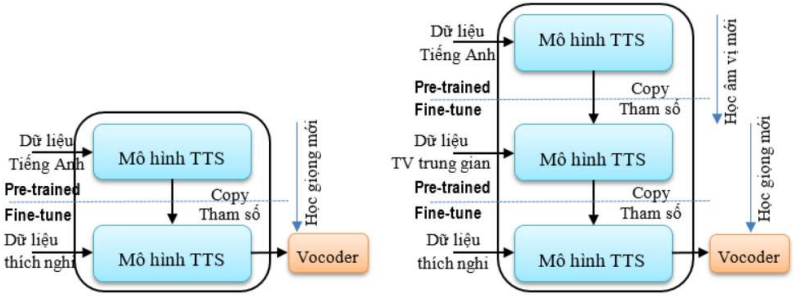
điều kiện danh tính và phong cách nói của giọng nói huấn luyện. Trong quá trình huấn luyện mô hình trung bình, vector hoặc mạng nhúng được sử dụng để phân biệt các đặc trưng âm thanh từ các danh tính người nói và phong cách nói khác nhau.

3.2. Nâng cao chất lượng TTS thích nghi đơn người nói bằng kỹ thuật Multi-pass fine-tune

Với cách tiếp cận tinh chỉnh truyền thống, để tạo ra một tiếng nói mới bằng một ngôn ngữ mới khác với mô hình huấn luyện trước vẫn cần một lượng lớn dữ liệu (≥ 5 giờ và điều này rất khó với các ngôn ngữ ít tài nguyên). Nếu sử dụng lượng dữ liệu quá nhỏ, rất dễ gây ra hiện tượng quá khớp (overffiting) thích nghi trực tiếp trên mô hình âm học End-to-end (Hình 5).



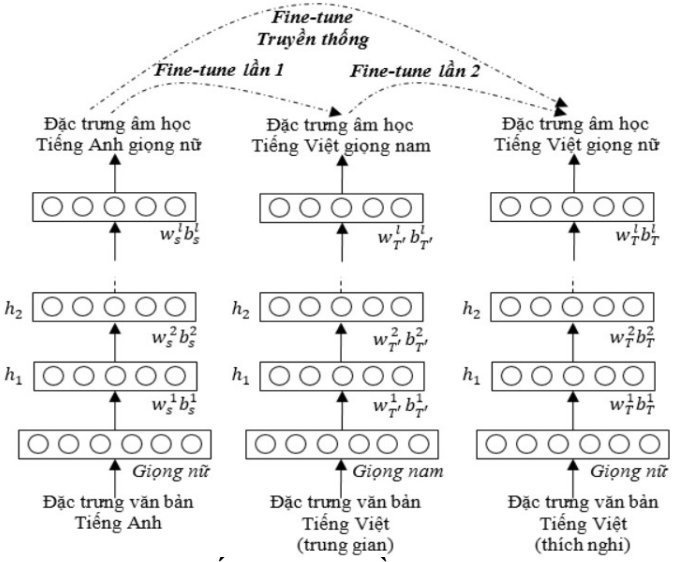
Hình 5 Sơ đồ luồng thích nghi giọng nói bằng tinh chỉnh truyền thống



Hình 6: Thích nghi tiếng nói một giọng nói mới với Multi-pass fine-tune

Để giải quyết những vấn đề này, luận án đề xuất một mô hình mượn một mô hình huấn luyện trước của tiếng Anh và tinh chỉnh lần thứ nhất với một mô hình huấn luyện trước bằng tiếng Việt làm trung gian, sau đó tinh chỉnh lần thứ hai với một mẫu giọng nói thích nghi. Luận án gọi nó là phương pháp gọi là "multi-pass fine-tune" và biển thị sơ đồ ở bên phải của Hình 6, chỉ yêu cầu một mẫu nhỏ để điều chỉnh một giọng nói mới. Vì đặc trưng âm học chính đã được học/chuyển từ tiếng Anh (tập dữ liệu lớn) và tiếng Việt (tập dữ liệu trung bình), để tạo ra giọng nói mới, chỉ cần một lượng nhỏ dữ liệu thích nghi với mô hình học đặc trưng giọng nói đích.

Hình 7 mô tả các bước để thực hiện Multi-pass fine-tune: 1) Đầu tiên huấn luyện mạng với số lượng lớn dữ liệu tiếng Anh để sinh ra bộ tham số của mô hình tiếng Anh; 2) Sau đó dùng mạng này để huấn luyện thích nghi giọng tiếng Việt trung gian, các tham số của mô hình tiếng Anh được cập nhật bằng các tham số của mô hình tiếng Việt trung gian; 3) Cuối cùng dùng mạng này để huấn luyện thích nghi giọng tiếng Việt đích, các tham số của mô hình tiếng Việt trung gian được cập nhật bằng các tham số thích nghi của giọng tiếng Việt đích Với phương pháp tinh chỉnh truyền thống mô hình chỉ huấn luyện mô hình tiếng Anh và cập nhật bộ tham số bởi bộ tham số thích nghi.



Hình 7: Cập nhật tham số thích nghi bằng Multi-pass fine-tune và tinh chỉnh truyền thống

3.2.1. Thử nghiệm đánh giá

Dữ liệu sử dụng : Tập dữ liệu tiếng Anh: Sử dụng kho ngữ liệu LSpeech-1.1 ; Tập dữ liệu tiếng Việt trung gian: Giọng nam đọc tin tức có tổng thời lượng 15 giờ. ; Bộ dữ liệu thích nghi: 4 bộ thích nghi từ 50 đến 800 câu (tương ứng từ 4 đến 60 phút). Sử dụng Tacotron2 + Waveglow cơ sở để đánh giá hệ thống TTS thích nghi.

3.2.2. Kết quả

***Chất lượng mô hình tinh chỉnh truyền thống**

Bảng 1: Bảng thống kê chất lượng thích nghi (MOS) theo mô hình Multi-pass fine-tune và các mô hình khác

Thời gian	Huấn luyện từ đầu (dữ liệu TV)	Mô hình huấn luyện trước bằng tiếng Anh + thích nghi	Mô hình huấn luyện trước bằng TV trung gian + thích nghi
16 phút	1.29	1.33	**3.78**

60 phút	1.31	2.68	3.87
5 giờ	2.66	N/A	N/A

- Nếu huấn luyện từ đầu với bộ dữ liệu tiếng Việt, với 5 giờ dữ liệu chất lượng bài phát biểu rất kém (MOS = 2.66). Huấn luyện với dữ liệu dưới 1 giờ sẽ không nghe được gì.
- Nếu tinh chỉnh từ mô hình huấn luyện trước bằng tiếng Anh với 1 giờ dữ liệu thích nghi tiếng Việt thì chất lượng sẽ ngang như huấn luyện từ đầu dữ liệu tiếng Việt 5 giờ, nhưng chất lượng âm tổng hợp vẫn kém (MOS = 2.68).

*** Chất lượng mô hình multi-pass fine-tune**

Trong các cột 4 của Bảng 1, dựa trên mô hình huấn luyện trước bằngtiếng Anh, nếu tinh chỉnh từ bộ dữ liệu tiếng Việt trung gian sang bộ dữ liệu thích nghi nhỏ thì chỉ cần 16 phút (200 câu) đã cho chất lượng thoại khá tốt với điểm MOS là 3.78/4.69 so với của giọng người nói.

*** Tính tương đồng**

Bảng 2: Bảng đánh giá độ tương đồng của mô hình tinh chỉnh truyền thống và Multi-pass fine-tune khi so sánh với giọng người nói với chỉ 4 phút dữ liệu thích nghi

Mô hình thử nghiệm	MCD	SIM
Âm thanh gốc (Groundtruth)	-	3,99
Tinh chỉnh truyền thống	10.65	1.13
Multi-pass fine-tune	**7.94**	**2.87**

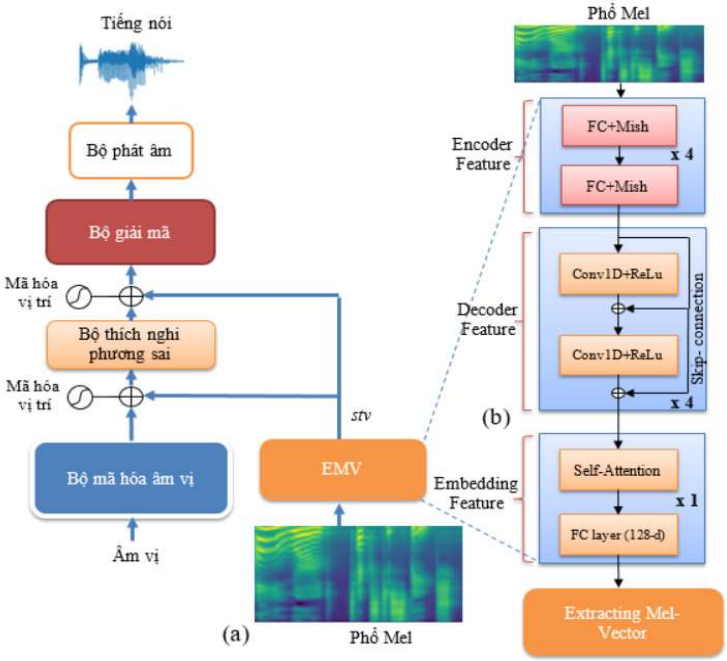
Bảng 2 cho thấy rằng Multi-pass fine-tune cho phép tạo ra giọng nói mới với MCD thấp hơn nhiều so với tinh chỉnh truyền thống (giảm **2.74**). Chỉ với ***4 phút dữ liệu thích nghi***, mô hình Multi-pass fine-tune tạo ra một giọng nói tổng hợp có độ tương đồng cao hơn nhiều so với giọng nói tổng hợp từ phương pháp tinh chỉnh truyền thống (**2.87/3.99** của giọng thật người nói). Tiến hành phân tích đánh giá SIM theo phương pháp của Mirjam Wester và cộng sự [65]. Qua kết quả phân tích ANOVA cho thấy mô hình tinh chỉnh truyền thống có ($F=5,188 > F$ tới hạn, $p<0,05$) và mô hình đề xuất có ($F=12,287 > F$ tới hạn, $p < 0,05$) cho thấy các kết quả thực nghiệm có sự khác biệt và có ý nghĩa thống kê.

3.3. Nâng cao chất lượng tổng hợp thích nghi bằng vector đặc trưng EMV

3.3.1. Đề xuất vector trích xuất đặc trưng Extracting Mel-Vector (EMV)

Hệ thống tổng hợp đa người nói dựa trên thích nghi phải sử dụng đặc trưng của người nói để huấn luyện và điều chỉnh mô hình thích nghi. Phương pháp truyền thống thường sử dụng một module embedding để trích xuất vector thứ tự đại diện. Tuy nhiên, phương pháp cơ bản này không thể nắm bắt được các đặc điểm riêng của từng người nói, chẳng hạn như danh tính, giới tính, tuổi tác và sức khỏe của họ vì nó chỉ dựa vào chỉ số định danh người nói làm đầu vào. Để giải quyết vấn đề này, một số nghiên cứu đề xuất một phương pháp thay thế liên quan đến một vectơ phong cách đại diện cho phong cách nói của người nói. Do đó, luận án đề xuất một mô-đun Mel-Vector Extraction (gọi tắt là mô-đun EMV)

được đề xuất dựa trên kiến trúc ban đầu của bộ mã hóa phong cách nói Mel-style đã sửa đổi có thể trích xuất một vectơ cố định từ phổ Mel, như được mô tả trong Hình 8.



Hình 8: a) Sơ đồ kiến trúc của mô hình dựa trên thích nghi Multi-TTS tiếng Việt với mô-đun Trích xuất Mel-vector (EMV)

Coi $\hat{y}$ là tiếng nói tổng hợp được tạo ra bởi mô hình sinh G với đầu vào là văn bản x và vector đặc trưng stv và các tham số có thể huấn luyện được θ, ta có biểu diễn đầu ra phổ Mel như sau: $\hat{y} = G(x, stv; \theta)$

trong đó, vector biểu diễn đặc trưng người nói *stv* được sinh bởi mô-đun EMV thông qua mã hóa phổ Mel của âm thanh gốc X làm mẫu thích nghi như sau:

$$stv = EMV(Mel\_X)$$

Kiến trúc tổng thể của mô hình đề xuất bao gồm các thành phần chính như sau: Mô-đun mã hóa âm vị (Phoneme Encoder) dùng để biến đổi chuỗi âm vị đầu vào thành chuỗi âm vị ẩn (hidden sequence). Positional Encoding (mã hóa vị trí) để mô hình có khả năng xác định được thông tin về vị trí tương đối của các từ trong câu. Mô-đun EMV được sử dụng để trích xuất các đặc trưng về người nói và phong cách nói từ đầu vào phổ Mel thành một vector đặc trưng người nói. Sau đó, Bộ điều hợp phương sai (Variance adapter) sẽ thêm thông tin về trường độ, cao độ và cường độ vào chuỗi âm vị ẩn này. Và bộ giải mã sẽ sử dụng các thông tin này để dự đoán ra phổ Mel. Cuối cùng, khối Vocoder sẽ chuyển đổi các phổ Mel này thành tín hiệu tiếng nói. Kiến trúc tổng thể được mô tả trong Hình 8.

Chức năng và kiến trúc chi tiết của các mô-đun EMV đề xuất gồm ba thành phần chính là khối mã hóa đặc trưng (Encoder Feature), Khối giải mã đặc trưng (Decoder Feature) và Khối vector nhúng biểu diễn đặc trưng (Embedding Feature). Theo đó: Tại khối Encoder Feature, đầu tiên đầu vào Mel-spectrogram được đưa đến lớp fully-connected (FC) và các hàm kích hoạt Mish (Hàm kích hoạt Mish được lựa chọn do đạt được hiệu năng vượt trội hơn ReLU và Swish trong nhiều thử nghiệm) để chuyển đổi mỗi frames của Mel-spectrogram thành các chuỗi ẩn, sau đó nó sẽ chuyển qua hai lớp FC để chuyển đổi đặc trưng đầu vào thành đặc trưng bộ mã hóa. Tiếp theo, vector này sẽ được đưa qua khối Decoder Feature. Bằng cách sử dụng Conv1D +ReLu với kết nối dư để nắm bắt (capture) được chuỗi thông tin từ tiếng nói mẫu đã cho, mục tiêu của mô-đun này sẽ chuyển đổi Decoder Feature thành đặc trưng bộ giải mã. Ngoài ra, skip connection cũng tích hợp, sẽ sử dụng các đặc trưng có giá trị của các khối trước đó và giải quyết được vấn đề triệt tiêu gradient. Cuối cùng, đầu ra của Decoder Feature sẽ được chuyển sang mô-đun Embedding Feature, mô-đun này có mô-đun self-attention với kết nối dư cộng lớp affine để mã hóa các thông tin một cách tổng quát toàn bộ đặc trưng người nói và phong cách nói. Áp dụng nó ở cấp frames để EMV có thể trích xuất thông tin phong cách nói tốt hơn ngay cả với một mẫu tiếng nói ngắn. Sau đó, tạm thời tính trọng số trung bình đầu ra tự chú ý của các mẫu thích nghi và có được một vectơ kiểu một chiều *stv*. Như vậy mô-đun này sẽ tạo ra một vectơ đại diện cho Mel-spectrogram và vectơ này sẽ thêm vào mô hình chuyển văn bản thành giọng nói.

3.3.2. Hàm mất mát huấn luyện

Hàm mất mát tổng quát của mô hình như sau:

$$L_{final} = L_{mel} + L_{duration} + L_{pich} + L_{energy}$$

1. Giá trị hàm mất mát L_{mel}: Khoảng cách giữa phổ Mel dự đoán và phổ Mel mục tiêu được mô tả như sau: $L_{mel} = \mathbb{E}[\| \hat{y} - y \|_1]$
2. Giá trị hàm mất mát phương sai $L_{duration}$, L_{pich}, L_{energy}: Sai số bình phương trung bình giữa trường độ các âm tiết, cao độ và cường độ của mẫu dự đoán và mục tiêu như sau:

$$L_{duration} = ||d - \hat{d}||_2^2, L_{pich} = ||p - \hat{p}||_2^2, L_{energy} = ||\varepsilon - \hat{\varepsilon}||_2^2,$$

3.3.3. Thử nghiệm đánh giá và kết quả

a) Thử nghiệm đánh giá

Để đánh giá các hệ thống TTS nhiều người nói, phần này sử dụng mô hình tổng hợp giọng nói đa người nói hiện đại như Fastspeech2 và bộ phát âm HifiGAN làm mô hình cơ sở (baseline model). Trong mô hình Multi-TTS dựa trên thích nghi đề xuất như được mô tả trong Hình 8, sẽ thay thế speaker embedding cơ sở bằng mô-đun EMV để mã hóa các đặc trưng của speaker trực tiếp từ Mel-spectrogram; Kỹ thuật data-distributing cũng được sử dụng để giữ các tham số đặc tính tiếng nói thích nghi.

Bộ dữ liệu: Dữ liệu thích nghi được chia thành bốn bộ (lần lượt là 1 phút, 2 phút, 4 phút và 16 phút) để huấn luyện các mô hình Few-shot dùng EMV.

b) Kết quả

Trên bảng 3 chỉ ra rằng, chỉ cần 1 phút dữ liệu tiếng nói đích thì mô hình Multi-speaker TTS dựa trên thích nghi đã có thể tổng hợp được âm thanh có điểm MOS đạt 3.81 so với điểm 4.6 của người nói. Điểm số này cao hơn hẳn so với MOS sinh từ mô hình Multi-TTS cơ sở (sử dụng 16 phút giọng đích). Điểm số WER cũng thể hiện rằng mô hình Multi-TTS dựa trên thích nghi tổng hợp giọng tốt hơn mô hình Multi-TTS cơ sở.

Bảng 3: Bảng đánh giá chất lượng giữa mô hình Multi-TTS cơ sở (sử dụng speaker embedding cơ bản) và Mô hình Multi-TTS dựa trên thích nghi (sử dụng mô-đun EMV và kỹ thuật phân phối dữ liệu)

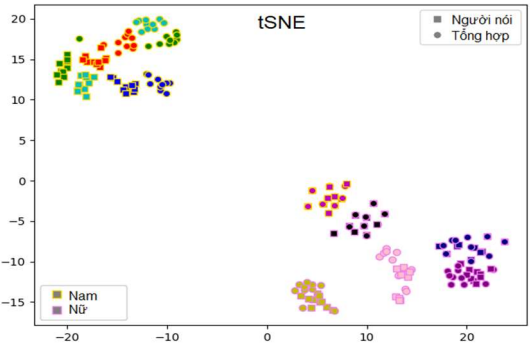
Trường độ/ Mô hình	Mô hình Multi-TTS cơ sở		Mô hình Multi-TTS dựa trên thích nghi	
	MOS (↑)	WER(↓)	MOS(↑)	WER(↓)
Âm thanh gốc	4.60	1.35	4.60	1.35
1 phút	3.39	8.40	**3.81**	**5.00**
2 phút	3.52	7.28	3.87	2.75
4 phút	3.59	6.16	4.00	2.00
16 phút	3.61	5.60	-	1.25

Bảng 4 cho thấy rằng, chỉ với 1 phút dữ liệu giọng nói mẫu thích nghi, mô hình Multi-TTS dựa trên thích nghi có điểm tương đồng SIM là 2.60 so với 4.0 của giọng nói con người. Điểm số này cao hơn nhiều so với 1.96 điểm SIM của mô hình Multi-TTS cơ sở (sử dụng 1 phút mẫu thích nghi). Điểm MCD của mẫu Multi-TTS dựa trên thích nghi cũng giảm hơn 10% so với mẫu TTS cơ sở. Tiến hành phân tích đánh giá SIM theo phương pháp của Mirjam Wester và cộng sự [65]. Qua kết quả phân tích ANOVA ta thấy mô hình cơ sở có ($F=4,636 > F$ tới hạn, $p<0,05$) và mô hình đề xuất có ($F=4,608 > F$ tới hạn, $p < 0,05$) cho thấy các kết quả thực nghiệm có sự khác biệt và có ý nghĩa thống kê.

Bảng 4: Mức độ tương đồng giữa các Mô hình Multi-TTS cơ sở và Mô hình Multi-TTS dựa trên thích nghi so với âm thanh gốc chỉ với 1 phút dữ liệu thích nghi

Mô hình thử nghiệm	MCD	SIM
Âm thanh gốc (Groundtruth)	-	4.0
Mô hình Multi-TTS cơ sở	7.36	1.96
Mô hình Multi-TTS dựa trên thích nghi	**6.54**	**2.60**

Hình 8, minh họa phép chiếu t-SNE của các vectơ phong cách nói giữa giọng người nói và giọng tổng hợp tương ứng. Sử dụng giọng nói của 10 người nói (5 nam và 5 nữ), có thể thấy mô hình người nói của hệ thống sử dụng EMV thể hiện rất tốt đặc trưng của người nói khi thể hiện sự tương đồng giữa giọng nói của con người và giọng nói tổng hợp thông qua các điểm thể hiện được phân cụm rõ ràng giữa từng người nói và điểm tổng hợp lân cận với giọng người nói.



Hình 8: Hình ảnh t-SNE phân bố của giọng nói của người nói và giọng nói tổng hợp (sử dụng EMV)

3.4. Kết luận chương 3

Trong Chương 3 đã trình bày một số kỹ thuật thích nghi tiên tiến hiệu quả cao với lượng dữ liệu thích nghi nhỏ trên thế giới. Từ phân tích đã chỉ ra các nhược điểm của các phương pháp tinh chỉnh truyền thống. Trên cơ sở đó đề xuất xây dựng một hệ thống tổng hợp thích nghi giọng Việt Multi-pas fine-tune sử dụng kỹ thuật học chuyển đổi và thử nghiệm để đánh giá hệ thống thích nghi. Kết quả đánh giá đã chứng minh rằng: 1) Với kỹ thuật Multi-pass fine-tune, chỉ cần một lượng dữ liệu nhỏ (***4 phút***) cho phép hệ thống tổng hợp được tiếng nói có độ tương đồng cao (với điểm số ***SIM đạt 2.87/3.99***) so với 1.13/3.99 của mô hình tinh chỉnh truyền thống và chỉ cần ***16 phút*** dữ liệu thích nghi cho phép hệ thống tổng hợp được tiếng nói với chất lượng cao với điểm MOS đạt ***3.78/4.69*** (tương đương 4.03/5) so với 2.68/3.99 của mô hình tinh chỉnh truyền thống [CT3].

Mô hình được đề xuất Trong Chương 3 đã cho thấy hiệu suất vượt trội so với mô hình Multi-TTS cơ sở sử dụng vector biểu diễn đặc trưng giọng nói truyền thống. Qua thực nghiệm, chỉ với 1 phút, mô hình đề xuất đạt độ tương đồng cao và chất lượng giọng nói tốt so với giọng nói gốc. Chỉ với ***1 phút*** dữ liệu thích nghi, mẫu Multi-TTS trên có khả năng thích nghi đã cho chất lượng ***MOS đạt 3.8/4.6*** và điểm tương đồng ***SIM đạt 2.6/4*** [CT2]. Điểm MOS này tương đương với điểm sử dụng 16 phút dữ liệu thích nghi dựa trên kỹ thuật Multipass-fine-tune mà đã trình bày trong 3.3.1. Điều đó chứng tỏ mô-đun EMV đã biểu diễn hiệu quả các đặc trưng của người nói so với vector biểu diễn đặc trưng giọng nói cơ bản, phù hợp với mô hình huấn luyện Few-shot TTS và có khả năng biểu diễn các đặc trưng ẩn của người nói nhìn thấy trong quá trình huấn luyện. Tuy nhiên việc đòi hỏi huấn luyện lại mô hình là một hạn chế của các kỹ thuật này và lượng dữ liệu thích nghi một vài phút vẫn chưa đủ hấp dẫn. Trong Chương tiếp theo, luận án sẽ đánh giá mô-đun EMV để tăng cường hiệu năng của hệ thống tổng hợp dựa trên thích nghi cho mô hình Zero-shot TTS với dữ liệu chưa từng xuất hiện trong tiến trình huấn luyện. Trên cơ sở đó sẽ tiến hành nghiên cứu và đề xuất các giải pháp và mô hình cải tiến trong Chương 4.

Chương 4. MÔ HÌNH TỔNG HỢP THÍCH NGHI KHÔNG HUẤN LUYỆN VỚI MẪU TỐI THIỂU (ZERO-SHOT TTS)

Chương 3 đã trình bày các kỹ thuật tổng hợp tiếng nói dựa trên thích nghi hiện nay dựa trên 2 luồng chính: một là tinh chỉnh mô hình bằng dữ liệu thích nghi có kích thước nhỏ và hai là huấn luyện toàn bộ mô hình thông qua một vector biểu diễn đặc trưng người nói của giọng đích. Tuy nhiên, cả hai phương pháp này đòi hỏi dữ liệu thích nghi phải xuất hiện trong quá trình huấn luyện, điều này khiến thời gian huấn luyện sinh ra giọng mới khá tốn kém. Ngoài ra, mô hình TTS truyền thống sử dụng hàm loss đơn giản để tái tạo các đặc trưng âm học, tuy nhiên việc tối ưu này dựa trên các giả định phân phối không chính xác dẫn đến kết quả âm thanh tổng hợp bị nhiễu. Chương 4 đề xuất mô hình Adapt-TTS cho phép nâng cao hiệu năng tổng hợp âm thanh ở mức chấp nhận được từ mẫu thích nghi nhỏ không cần huấn luyện. Chương 4 sẽ trình bày các nội dung: Kiến trúc Extracting Mel-vector (EMV) cho phép biểu diễn đặc trưng của người nói và phong cách nói tốt hơn; Mô hình Zero-shot TTS cải tiến với thành phần khuếch tán khử nhiễu phổ Mel (Mel-spectrogram denoiser) sử dụng tính chất của quá trình khuếch tán ngược tích hợp với vector đặc trưng EMV để trích chọn giọng nhằm sinh giọng mới cho mô hình Zero-shot TTS cho phép tổng hợp giọng mới mà không cần phải huấn luyện lại (without training) với chất lượng tốt hơn [CT1]; Thử nghiệm và kết quả đánh giá cho mô hình đề xuất [CT7];

4.1. Đề xuất mô hình Adapt-TTS cải tiến hiệu năng cho tổng hợp thích nghi tiếng Việt

Kiến trúc của Adapt-TTS bao gồm các thành phần chính: Mô-đun EMV để trích xuất các đặc trưng người nói và phong cách nói thành một vector đặc trưng, Mô-đun mã hóa âm vị (Phoneme Encoder) dùng để biến đổi chuỗi âm vị thành phoneme hidden sequence, sau đó Bộ điều hợp phương sai (Variance adapter) sẽ thêm các thông tin về trường độ, cao độ và cường độ vào chuỗi ẩn này. Bộ khử nhiễu khuếch tán phổ Mel (Mel-spectrogram denoiser) sẽ nhận các thông tin ẩn ở các bước trước đó để thực hiện giải mã ra thành các phổ Mel với chất lượng cao dựa trên nhân kiến trúc mô hình khuếch tán. Cuối cùng, khối Vocoder sẽ chuyển đổi các phổ Mel này thành tín hiệu tiếng nói.

4.1.1. Mã hóa đặc trưng với EMV

Chương 4 đề xuất một mô-đun tương đồng với Chương 3 gọi là Vector trích xuất Mel (mô-đun Extracting-Mel vector hay còn gọi là EMV) có thể trích xuất một vectơ cố định từ biểu đồ phổ mel của speaker để biểu chính xác đặc trưng của speaker như đặc trưng người nói và phong cách nói. EMV là sẽ lấy tiếng nói tham chiếu X làm đầu vào, mục đích của khối này lã trích xuất một vector embedding *stv* có chứa phong cách và đặc trưng của người nói X.

4.1.2. Mel-spectrogram denoiser

Khối Decoder nhận đầu vào từ chuỗi âm vị ẩn thông qua Variance adapter để thêm thông tin phương sai (như trường độ, cao độ, năng lượng) sau đó kết hợp với vector EMV (để biểu diễn đặc trưng người) sau đó trong khối decoder khối

Mel-spectrogram denoiser sẽ nhận đầu vào là chuỗi x_t, biến c là đầu ra của variance adapter và bước thời gian t để thực hiện khử nhiễu và tổng hợp âm thanh chất lượng cao dựa trên trên Diffusion model. Quá trình infer của Diffusion model cho Multi-TTS là tối ưu hóa hàm mục tiêu $f_\theta(x_t|t,c)$ để chuyển đổi các phân phối nhiễu thành 1 phân phối Mel-spectrogram tương ứng với văn bản đã cho trước, mô hình này gồm 2 tiến trình chính :

Quá trình khuếch tán xuôi (Diffusion process): Đầu tiên, Mel-spectrogram bị dần làm nhiễu với nhiễu Gauss và biến đổi thành các biến tiềm ẩn. Quá trình này được gọi là quá trình khuếch tán xuôi. Đặt $x_1 \ldots . x_T$ là chuỗi các biến có cùng chiều, trong đó $t= 0,1\ldots.,T$ là chỉ số cho bước thời gian diffusion. Khi đó, quá trình khuếch tán xuôi chuyển đổi Mel-spectrogram x_0 thành một nhiễu Guassian x_T thông qua một chuỗi chuyển đổi Markov.

Quá trình phục hồi (Reverse process): Học cách khôi phục phân bố dữ liệu, là quy trình tạo phổ Mel từ nhiễu Gauss. Quá trình khôi phục được định nghĩa là phân phối có điều kiện $p_\theta(x_{0;T}|x_T,c)$, và nó có thể được factorized thành nhiều chuyển đổi dựa trên tính chất chuỗi Markov. Thông qua các chuyển đổi ngược lại $p_\theta(x_{t-1}|x_t,c)$, các biến tiềm ẩn dần dần được khôi phục thành Mel-spectrogram tương ứng với bước thời gian khuếch tán với điều kiện văn bản. Nói cách khác, Mel-spectrogram denoiser học phân phối mô hình $p_\theta(x_0|c)$ thu được từ quy trình ngược lại. Đặt $\alpha = 1-\beta_t$ và $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ Mục tiêu huấn luyện của bộ khử nhiễu khuếch tán phổ Mel như sau:

$$minL_\theta = E_{x_0,\epsilon,t}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c)||_1\right]$$

4.1.3. Sinh âm thanh có điều kiện

Với nhiệm vụ sinh âm thanh có điều kiện dựa trên nhiều thông tin đầu vào, coi y là các nhãn thông tin điều kiện bổ sung thì chuyển mọi công thức khuếch tán trên với điều kiện y như sau: $p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t,y)$

$$\epsilon_\theta(x_t,t) \rightarrow \epsilon_\theta(x_t,t,y)$$

trong đó y là các điều kiện như biến c là đầu ra của bộ thích nghi phương sai, vector đặc trưng người nói stv sinh bởi EMV, thì biểu diễn các công thức trên như sau:

$$p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t,c,stv) \quad v\grave{a}\ \epsilon_\theta(x_t,t) \rightarrow \epsilon_\theta(x_t,t,c,stv)$$

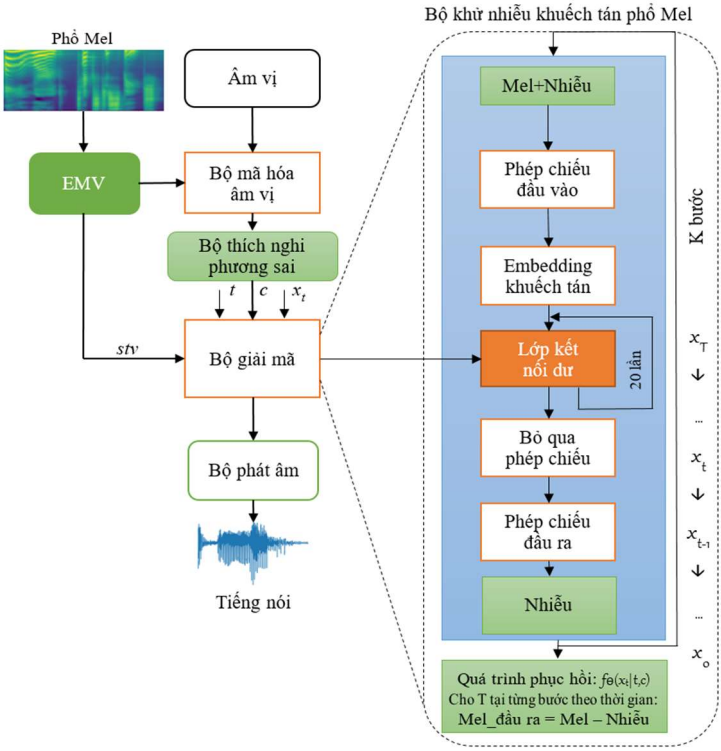
Tóm tắt quá trình huấn luyện và suy luận:

Quá trình huấn luyện: Giá trị hàm mất mát tổng quát trong quá trình huấn luyện bộ khử nhiễu Mel-spectrogram bao gồm các phần sau đây:

$$L_{final} = L_\theta + L_{SSIM} + L_{duration} + L_{pich} + L_{energy}$$

Trong đó : Giá trị hàm mất mát L_θ là sai số bình phương trung bình MSE giữa mẫu phổ mel dự đoán và mục tiêu; Giá trị loss chỉ số tương đồng cấu trúc L_{SSIM}(SSIM) là 1- chỉ số SSIM giữa mẫu phổ mel dự đoán và mục tiêu; Giá trị hàm mất mát phương sai $L_{duration}$, L_{pich}, L_{energy} là sai số bình phương trung bình giữa trường độ các âm tiết, phổ pitch và năng lượng của mẫu dự đoán và mục tiêu.

Quá trình suy diễn: Trong quá trình suy luận, Mel-spectrogram denoiser dự đoán đầu vào x_0 không bị nhiễu và sau đó thêm lại nhiễu bằng cách sử dụng phân phối hậu nghiệm, từ đó tạo ra các mặt phẳng Mel-spectrogram với chi tiết tăng dần. Cụ thể, mô hình khử nhiễu $f_\theta(x_t, t, c)$ trước tiên dự đoán x_t, sau đó x_{t-1} được lấy mẫu bằng cách sử dụng phân phối hậu nghiệm $q(x_{t-1}|x_t, x_0)$ được cho bởi $x_t$và dự đoán x_{t-1}. Cuối cùng, mặt phẳng spectrogram được tạo ra từ x_0 được chuyển đổi thành hình dạng sóng bằng cách sử dụng một vocoder được huấn luyện trước.



Hình 9: Kiến trúc chi tiết của khối Mel-spectrogram Denoiser

4.2. Thử nghiệm đánh giá và kết quả

4.2.1. Thử nghiệm đánh giá

Để đánh giá chất lượng âm thanh tổng hợp tạo từ các mô hình đề xuất, 5 bộ dữ liệu đã được chuẩn bị, trong đó, 4 bộ được tổng hợp từ các audio tham chiếu với khoảng thời gian tương ứng là 1 giây, 3 giây, 5 giây và 1 bộ giọng gốc để đối sánh. Sử dụng hai mô hình để tổng hợp: một là mô hình cơ sở đề xuất bởi Fastspeech2 và mô hình Adapt-TTS. Sử dụng 30 người nghe để đánh giá. Đánh giá hệ thống tổng hợp bằng cách phối hợp đánh giá cả theo phương pháp khách quan (WER) và đánh giá chủ quan (MOS/SIM).

4.2.2. Kết quả

*** Chất lượng tổng hợp**

Bảng5 chỉ ra rằng với chỉ 3 giây âm thanh thích nghi của người nói mới, dù không cần huấn luyện lại thì mô hình Adapt-TTS đã tổng hợp được âm thanh đạt điểm số MOS là 3.29 so với 4.53 điểm của người nói. Điểm số này cao hơn điểm số của mô hình cơ sở là 2.16. Điểm số WER cũng cho thấy chỉ với 1 giây âm thanh thích nghi của người nói mới đã có thể tổng hợp được âm thanh đạt WER 3.38.

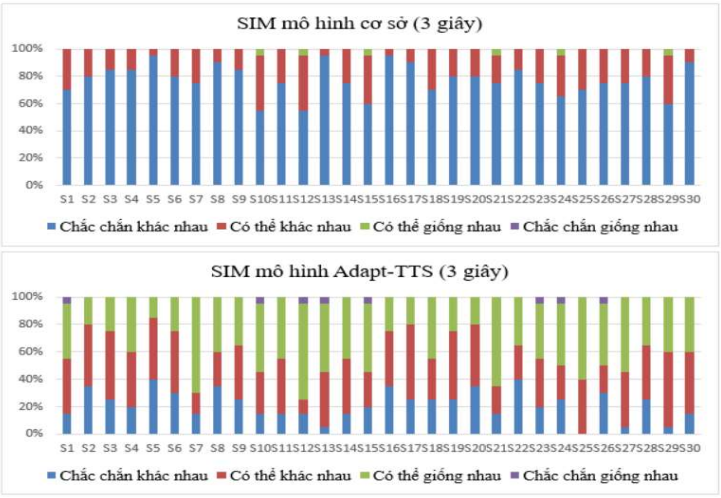
Bảng 5: Kết quả đánh giá chất lượng tổng hợp MOS/WER của các mô hình cơ sở và mô hình đề xuất vố các giọng chưa có trong tập huấn luyện

Trường độ/ Mô hình	Mô hình cơ sở		Mô hình Adapt-TTS	
	MOS (↑)	WER(↓)	MOS(↑)	WER(↓)
Âm thanh gốc	4.53	1.35	4.53	1.35
1 giây	2.05	8.78	2,89	**3.38**
3 giây	2.16	7.77	**3.29**	3.14
5 giây	2.18	6.76	3.31	3.04

*** Độ tương đồng**

Bảng 6: Kết quả đánh giá độ tương đồng SIM của các mô hình cơ bản và mô hình đề xuất với độ tin tưởng 95%

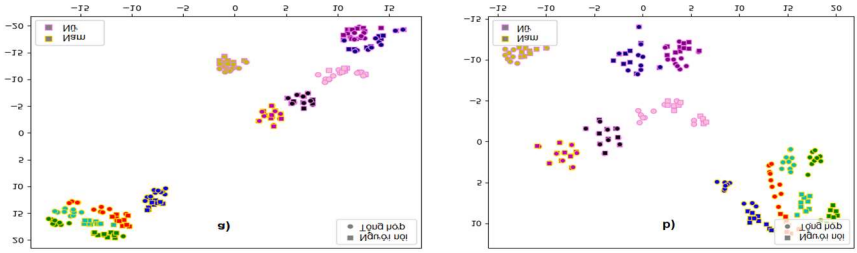
Mô hình/ Trường độ	Mô hình cơ sở	Adapt-TTS
	SIM	SIM
Âm thanh gốc (Groundtruth)	3.90	3.90
1 giây	1.16	1.71
3 giây	1.24	**2.22**
5 giây	1.31	2.6



Hình 10: So sánh sự tương đồng của của mô hình cơ sở (trên) và mô hình Adapt-TTS đề xuất (dưới) trên tất cả các cặp câu đánh giá

Bảng 6 chỉ ra rằng chỉ với 3 giây âm thanh thích nghi của người nói mới, thì mô hình Adapt-TTS đã đạt độ tương đồng với chỉ số SIM là 2.2/3.9 của giọng người nói. Trong khi mô hình cơ sở chỉ đạt chỉ số SIM 1.24/3.9. Tiến hành phân tích đánh giá SIM theo phương pháp của Mirjam Wester và cộng sự [65], tổng hợp điểm tương đồng của người nghe cho toàn bộ các cặp câu được đánh giá (giữa giọng tổng hợp và âm thanh gốc) thể hiện trong Hình 10, trong đó ký hiệu S1, S2, .. biểu diễn thứ tự của người đánh giá. Biểu diễn cho cho thấy độ tự tin về khả năng "chắc chắn giống" và "có thể giống nhau" của hai mô hình Adapt-TTS đề xuất và mô hình cơ sở là rất rõ ràng. Qua kết quả phân tích ANOVA ta thấy mô hình cơ sở có (F=1,675> F tới hạn, p<0,05) và mô hình Adapt-TTS đề xuất có (F=2,099>F tới hạn, p < 0,05) cho thấy các kết quả thực nghiệm có sự khác biệt và có ý nghĩa thống kê.

Phép chiếu t-SNE của vectơ EMV thu được từ những người nói không nhìn thấy trong bộ dữ liệu đa người nói của tiếng Việt, cụ thể, chọn ra 10 người nói (5 nam và 5 nữ). Adapt-TTS cho thấy sự phân tách các vectơ đại diện người nói rõ ràng và gần với âm thanh gốc hơn khi so sánh với mô hình cơ sở. Biểu đồ t-SNE của mô hình Adapt-TTS (Hình 11a) cho thấy âm thanh tổng hợp và âm thanh thật của cùng một người nói co cụm lại gần nhau thể hiện sự tương đồng. Đặc điểm giới tính cũng được phân cụm rõ ràng ở hai vùng khác nhau.



a) b)

Hình 11: Mô hình hóa phân bố không gian t-SNE của a) Mô hình Adapt-TTS và b) Mô hình cơ sở giữa giọng tổng hợp và giọng người nói của 10 người

Như vậy, trong mô hình Adapt-TTS đề xuất, ban đầu không cung cấp bất kỳ thông tin nào về danh tính của người nói cho mô hình encoder, phân phối do bộ encoder dự đoán buộc phải độc lập với danh tính của người nói. Do đó, Adapt-TTS có thể chuyển đổi giọng nói chỉ bằng cách sử dụng mô hình encoder. Bộ mã hóa đặc trưng người nói với EMV và cho phép thích nghi các đặc trưng giọng. Các melspectrogram được dự đoán chính xác hơn nhờ bộ Decoder với thành phần Mel-spectrogram denoiser theo mô hình khuếch tán tạo âm thanh chất lượng cao và ít nhiễu.

4.3. Kết luận chương 4

Chương 4 đã đề xuất một kiến trúc *Adapt-TTS cho phép tổng hợp một giọng mới bằng phương pháp thích nghi giọng nói bằng Zero-shot TTS với chỉ một câu*

nói duy nhất của mẫu âm thanh của người nói mới mà không cần huấn luyện lại mô hình. Các đề xuất sử dụng EMV kết hợp với mô hình khử nhiễu khuếch tán phổ Mel cho phép tổng hợp giọng nói nhân bản với chất lượng chấp nhận được. Thực nghiệm chứng minh rằng chỉ cần một mẫu 1-3 giây tiếng nói của giọng nói mẫu đã có thể tổng hợp được giọng nói có chất lượng ***MOS đạt 3.3/4.5*** và độ tương đồng ***SIM đạt 2.2/3.9*** [CT1]. Chất lượng âm thanh tạo từ mô hình thích nghi Zero-shot TTS tuy không thể đạt được chất lượng hoặc thay thế cho các mô hình tổng hợp thích nghi có huấn luyện lại mô hình như Few-shot TTS nhưng bù lại mô hình đề xuất cho phép học nhanh giọng mới mà không cần phải huấn luyện lại và chất lượng âm thanh tổng hợp vẫn đảm bảo ở mức chấp nhận được và đạt được độ tương đồng tốt so với giọng đích. *Mô hình Adapt-TTS đề xuất đã cho phép tổng hợp giọng nói mới dựa trên duy nhất một mẫu câu thích nghi và không phải huấn luyện lại mô hình, cho phép mở rộng ứng dụng của mô hình tổng hợp và khả năng áp dụng đa dạng trong cuộc sống* [CT7].

KẾT LUẬN

Chương 1 đã trình bày các khảo sát và phân tích chi tiết về các nghiên cứu hiện nay cũng như kiến thức có liên quan về tổng hợp và thích nghi giọng nói. Chương 2 trình bày kết quả xây dựng bộ CSDL bằng các phương pháp hiệu quả với chi phí thấp làm nền tảng xây dựng các mô hình tổng hợp và thích nghi ở các Chương tiếp theo. Các Chương 3, 4 đề đã trình bày các đề xuất và thử nghiệm quan trọng nhất của Luận án '*Nghiên cứu phát triển hệ thống thích nghi giọng nói trong tổng hợp tiếng Việt và ứng dụng'* với các đóng góp chính. Nội dung Chương 3, 4 đã trình bày ba phương pháp tổng hợp tiếng nói đảm bảo chất lượng cho ngôn ngữ nghèo tài nguyên như tiếng Việt trong khi chỉ có vài phút mẫu thích nghi đó là các kỹ thuật thích nghi dựa trên DNN cho mô hình phụ thuộc người nói (Few-shot TTS) và độc lập người nói (Zero-shot TTS). Các kỹ thuật đề xuất chi tiết của các Chương này cũng đã trả lời cho các câu hỏi nghiên cứu về số lượng mẫu tối thiểu đùng để thích nghi (được huấn luyện cùng hệ thống và không huấn luyện cùng hệ thống) kèm theo các thử nghiệm và đánh giá cụ thể:

1) **Đề xuất mô hình Multi-pass fine-tune** để tổng hợp thích nghi Few-shot TTS cho tiếng Việt chất lượng cao bằng kỹ thuật học chuyển đổi. Mô hình phụ thuộc người nói được đề xuất có khả năng nhân bản một giọng mới có qua huấn luyện nhằm giải quyết vấn đề cần ít dữ liệu của giọng nói nhân bản so với phương pháp truyền thống. Chỉ với mẫu câu nói ***4 phút*** cho phép hệ thống tổng hợp được tiếng nói có độ tương đồng cao (với điểm số ***SIM đạt 2.87/3.99***) và chỉ cần ***16 phút*** dữ liệu thích nghi cho phép hệ thống tổng hợp được tiếng nói với chất lượng cao với điểm ***MOS đạt 3.78/4.69*** so với 2.68/3.99 của mô hình tinh chỉnh truyền thống [CT3];

2) **Đề xuất kiến trúc vector EMV (Extracting-Mel vector)** có khả năng trích xuất đặc trưng và biểu diễn người nói hiệu quả và mô hình thích nghi Few-shot TTS cho tiếng Việt giúp tăng cường chất lượng thích nghi. Mô hình phụ thuộc người nói được đề xuất có khả năng nhân bản một giọng mới cần ít dữ liệu hơn

các kỹ thuật tinh chỉnh. Chỉ với **1 phút** dữ liệu thích nghi, mẫu Multi-TTS trên có khả năng thích nghi đã cho chất lượng ***MOS đạt 3.8/4.6*** và đạt điểm tương đồng ***SIM đạt 2.6/4*** [CT2]; Ngoài ra kiến trúc Variance adapter **có khả năng điều chỉnh giọng** (điều khiển các đặc trưng tiếng nói như trường độ, cao độ và cường độ).

3) **Đề xuất mô hình Adapt-TTS** để giải quyết bài toán nhân bản giọng nói không cần huấn luyện lại (Zero-shot TTS). Mô hình độc lập người nói được đề xuất giải quyết bài toán nhân bản một giọng mới với rất ít dữ liệu và không phải huấn luyện lại và có khả năng áp dụng trong thực tế. Mô hình đề xuất có khả năng nhân bản với chỉ một câu mẫu duy nhất ***(1-3 giây)*** thông qua vector biểu diễn đặc trưng **EMV** và kiến trúc khử nhiễu khuếch tán phổ Mel (**Mel-spectrogram denoiser)** mà không cần huấn luyện lại mô hình, cho chất lượng tổng hợp ***MOS đạt 3.3/4.5*** và độ tương đồng ***SIM đạt 2.2/3.9*** [CT1];

4) **Xây dựng bộ CSDL tiếng nói đảm bảo chất lượng và chi phí thấp** cho nhiệm vụ tổng hợp và thích nghi [CT6] [CT3]; Kỹ thuật bổ sung thông tin nhãn nhằm tăng cường độ tự nhiên cho các hệ thống tổng hợp tiếng nói tiếng Việt thông qua (chèn dấu câu, chèn ***điểm*** dừng lấy hơi và phiên âm từ mượn) [CT5][CT4]. Kết quả của phần này chính là bộ CSDL quan trọng cho tổng hợp và thích nghi sử dụng xuyên suốt cho các Chương 3 và 4 của luận án.

5) **Xây dựng được ứng dụng nhân bản giọng sử dụng được trên các thiết đa nền tảng** nhằm bắt chước và tổng hợp giọng nói bất kỳ để chứng minh tính khả thi và hiệu năng của các mô hình đề xuất kèm các minh chứng [CT7]. Với mỗi mô hình thích nghi đề xuất sẽ có ưu nhược điểm riêng và từ đó tính ứng dụng thực tiễn khác nhau: *Mô hình Few-shot TTS sẽ cho chất lượng tổng hợp tốt với chỉ một lượng nhỏ vài phút đến vài chục phút dữ liệu thích nghi cho phép nhân bản giọng hoặc tạo các giọng nói giọng độc quyền phục vụ phát thanh, đọc báo cáo tự động; Mô hình thích nghi Zero-shot TTS với chỉ một câu dữ liệu và không phải huấn luyện phù hợp với học giọng tức thì của người dùng, ứng dụng cho loa thông minh.*

Hướng phát triển

1) Nghiên cứu giải pháp tăng cường chất lượng thích nghi với các mẫu giọng có cảm xúc hoặc giọng mẫu ít dữ liệu.

2) Thực nghiệm các mô hình đề xuất trong nghiên cứu này với các bộ dữ liệu tiếng Anh, tiếng Trung, ... đã công bố để có những đối sánh về tính hiệu quả của mô hình.

3) Áp dụng mô hình đề xuất cho các kỹ thuật thích nghi đa ngôn ngữ (multi-lingual adaptation).

4) Tiếp tục cải tiến mô hình Adapt-TTS và các thuật toán nén cho mô hình huấn luyện/ tổng hợp tương ứng để giảm được chi phí tính toán và có thể chạy trên các thiết bị có tài nguyên nhỏ.