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

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội - 2023

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

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

Mã số: 9 48 01 04

Xác nhận của Học viện
Khoa học và Công nghệ

Người hướng dẫn
(Ký, ghi rõ họ tên)

PGS.TS. Lương Chi Mai

Hà Nội - 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận án này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận án là trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm trước phát luật.

Hà Nội, ngày tháng năm 2023

Tác giả luận án

Phạm Ngọc Phương

LỜI CẢM ƠN

Luận án của tác giả được thực hiện tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, dưới sự hướng dẫn tận tình của PGS.TS. Lương Chi Mai. Tôi xin được bày tỏ lòng biết ơn sâu sắc đến Cô về định hướng nghiên cứu, sự động viên và hướng dẫn tận tình giúp tôi vượt qua những khó khăn để hoàn thành luận án này. Tôi cũng xin gửi lời cảm ơn chân thành đến các nhà khoa học, các đồng tác giả của các công trình nghiên cứu đã được trích dẫn trong luận án. Đây là những tư liệu quý báu có liên quan giúp tôi hoàn thành luận án.

Tôi xin chân thành cảm ơn đến Ban lãnh đạo Học viện Khoa học và Công nghệ, Viện Công nghệ Thông tin đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập, nghiên cứu.

Tôi xin chân thành cảm ơn Ban lãnh đạo Trung tâm Số - Đại học Thái Nguyên và các đồng nghiệp đã giúp đỡ và tạo điều kiện thuận lợi để tôi có thể thực hiện kế hoạch nghiên cứu, hoàn thành luận án.

Tôi xin chân thành cảm ơn TS. Đỗ Quốc Trường, NCS. Trần Quang Chung và các thành viên tại công ty VAIS cũng như công ty AIMed đã giúp đỡ và tạo điều kiện thuận lợi để tôi có thể thực hiện nghiên cứu.

Tôi xin được bày tỏ tình cảm và lòng biết ơn vô hạn tới những người thân trong Gia đình, những người luôn dành cho tôi sự động viên, khích lệ, sẻ chia, giúp đỡ trong những lúc khó khăn.

Hà Nội, ngày tháng năm 2023

Người thực hiện

Phạm Ngọc Phương

MỤC LỤC

DANH MỤC THUẬT NGỮ

Thuật ngữ	Diễn giải
Anova	Kiểm định Anova hay còn gọi là phân tích phương sai
Attention	Cơ chế tự chú ý
Baseline	Mô hình hoặc kiến trúc cơ bản, làm cơ sở so sánh
Cepstrum	Phổ trên thang logarit với trục hoành là nghịch đảo tần số tín hiệu, trục tung là biên độ logarit
Decoder	Bộ giải mã
Distillation	Quá trình chưng cất/lọc thông tin
Duration	Trường độ thể hiện độ dài thời gian của âm thanh
Embedding	Kỹ thuật đưa vector có số chiều lớn về không gian có chiều nhỏ hơn mang tính đại diện , còn gọi là vector nhúng
Encoder	Bộ mã hóa
End-to-end	Mô hình từ một luồng vào ra
F0	Tần số cơ bản
F1	Độ đo F1
Few-shot	Mô hình hóa bằng cách học một lượng nhỏ dữ liệu
Fine-tune	Kỹ thuật tinh chỉnh các tham số học từ mô hình huấn luyện trước (pre-trained model)
Groundtruth	Âm thanh gốc, thường là âm thanh của người nói
Loss	Hàm mất mát
Mel-Spectrogram	Phổ Mel âm thanh (viết tắt là phổ Mel)
One-shot	Mô hình hóa bằng cách học duy nhất một mẫu dữ liệu
Overfit	Mô hình xây dựng quá khớp với dữ liệu huấn luyện
Pitch	Pitch là cảm nhận âm thanh của tần số cơ bản F0
Pre-trained model	Mô hình đã được huấn luyện từ trước
Sequence-to-Sequence	Chuỗi từ chuỗi (hay còn viết là Seq2seq)
Speaker	Người nói, người phát biểu

Speaker Adaptation	Thích nghi người nói
Speaker-embedding	Vector mã hóa biểu diễn đặc trưng giọng nói
Spectrogram	Phổ âm thanh
Text to speech	Văn bản thành tiếng nói
t-SNE	Biểu diễn giảm chiều phân phối ngẫu nhiên các vector liền kề
Variance Adaptor hoặc Variance Adapter	Bộ thích nghi phương sai
Vocoder	Bộ phát âm
Zero-shot	Mô hình hóa mà không cần dữ liệu huấn luyện

DANH MỤC CÁC KÝ HIỆU VÀ TỪ VIẾT TẮT

Từ viết tắt	Diễn giải	Ý nghĩa
ASR	Automatic Speech Recognition	Nhận dạng tiếng nói
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
CRF	Conditional Random Field	Trường ngẫu nhiên có điều kiện
DBF	Deep Belief Networks	Mạng niềm tin sâu
DCT	Discrete Cosine transform	Biến đổi cosin rời rạc
DDPM	Denoise Diffusion Probabilistic Model	Mô hình xác suất khuếch tán khử nhiễu
DFT	Discrete Fourier Transform	Biến đổi Fourier rời rạc
DNN	Deep Neural Network	Mạng nơ-ron học sâu
EER	Equal Error Rate	Tỷ lệ câu bị lỗi
EMV	Extracting Mel-spectrogram Vector	Vector trích xuất đặc trưng từ phổ Mel
FFT	Feed-Forward Transformer	Transformer chuyển tiếp
G2P	Graph to Phone	Hình vị thành âm vị
GAN	Generative Adversarial Network	Mạng sinh đối nghịch
GMM	Gaussian Mixture Model	Mô hình phân phối trộn Gauss
GPU	Graphical Processing Unit	Bộ xử lý đồ họa
GT	Ground Truth	Âm thanh gốc làm đối sánh
HMM	Hidden Markov Model	Mô hình Markov ẩn
IPA	International Phonetic Alphabet	Bản phiên âm quốc tế
MAE	Mean Absolute Error	Sai số tuyệ đối trung bình (hàm mất mát L1)
MAP	Maximum A Posteriori	Thuật toán cực đại hậu nghiệm
MCD	Mel-Cepstral Distortion	Đo sự biến dạng phổ mel
MFA	Montreal Forced Align	Công cụ trích xuất trường độ dựa trên căn chỉnh thời gian bằng cách sử dụng từ điển phát âm

MLLR	Maximum Likelihood Linear Regression	Thuật toán hồi quy tuyến tính ước lượng khả năng cực đại
MFCC	Mel Frequency Cepstral Coefficients	Hệ số phổ quang tần số Mel
MOS	Mean Opinion Score	Điểm ý kiến trung bình
MSE	Mean Squared Error	Sai số bình phương trung bình (hàm mất mát L2)
MSD	Multi-Space Distribution	Phân phối đa không gian
LSTM	Long Short Term Memory	Bộ nhớ ngắn dài hạn
L1	Loss 1	Hàm mất mát MAE
L2	Loss 2	Hàm mất mát MSE
OOV	Out Of Vocabulary	Các từ ngoài từ điển
PCA	Principal Component Analysis	Phép phân tích thành phần chính
PLDA	Probabilistic Linear Discriminant	Phân tích biệt thức tuyến tính xác suất
ReLU	Rectified Linear Unit	Hàm kích hoạt sửa chữa tuyến tính
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
SIM	Similarity score	Điểm đo độ tương đồng
SPSS	Statistical Parametric Speech Synthesis	Tổng hợp dựa trên tham số thống kê
t-SNE	t-Distributed Stochastic Neighbor Embedding	Biểu diễn ngẫu nhiên các embedding phân tán
TTS	Text to speech	Văn bản thành tiếng nói
UBM	Universal Background Model	Mô hình UBM
VAE	Variational Autoencoder	Bộ mã hóa tự động biến đổi
VLSP	Vietnamese Language and Speech Processing	Hiệp hội Xử lý tiếng nói và văn bản tiếng Việt
VPS	Vector Field Smoothing	Thuật toán làm mịn trường vector
WER	Word Error Rate	Tỷ lệ lỗi từ

DANH MỤC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Tiếng nói nhân tạo hay còn gọi là tiếng nói tổng hợp đã có lịch sử trên 200 năm. Đến nay, tiếng nói tổng hợp đã phát triển vượt bậc khi có chất lượng gần giống con người (độ dễ nghe, dễ hiểu) lẫn khả năng ứng dụng rộng rãi trong đời sống xã hội. Hiện nay, có thể dễ dàng bắt gặp sản phẩm tổng hợp tiếng nói ở nhiều nơi trên internet, trên các ứng dụng di động, hệ thống hỏi đáp tự động ... Khi nghiên cứu về tổng hợp tiếng nói, một trong những chủ đề được quan tâm nhất hiện nay là phương pháp điều khiển và thích nghi các đặc trưng tiếng nói để tạo ra tiếng nói tổng hợp theo phong cách và ngữ điệu tùy ý. 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 [1]. Điều này khiến việc tạo ra các giọng nói tổng hợp mới theo yêu cầu rất tốn kém về chi phí, mất nhiều thời gian và khó thực hiện với các ngôn ngữ nghèo tài nguyên. 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 [2] [3]. Do vậy, với lượng dữ liệu mẫu thích nghi nhỏ (từ vài câu đến vài chục câu) thì việc nâng cao chất lượng tổng hợp vẫn còn là một thách thức.

Các đặc trưng riêng biệt của người nói (gồm đặc trưng giọng nói và đặc trưng ngữ điệu) đều bao hàm trong phổ tín hiệu, tần số cơ bản và trường độ. Do đó, để thực hiện kỹ thuật chuyển đổi và thích nghi giọng nói, cần phải chuyển đổi tất cả các tham số đặc trưng giọng nói nguồn thành các tham số đặc trưng giọng nói đích. Các nghiên cứu điều chỉnh, biến đổi tham số đặc trưng giọng nói và thích nghi giọng nói 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, tiếng Nhật, tiếng Trung và vẫn đang còn là thách thức [4] [5].

Với tiếng Việt, đây là ngôn ngữ nghèo tài nguyên và là ngôn ngữ phức tạp do có chứa thành phần ngữ điệu và nhiều từ mượn, ngay cả các kỹ thuật tổng hợp tiên tiến nhất áp dụng cho tổng hợp tiếng Việt cũng chưa giải quyết được triệt để các vấn đề như đọc câu dài và từ mượn [6]. Đã có một số nghiên cứu về chuyển

đổi đặc trưng giọng nói và thích nghi giọng nói áp dụng đối với tiếng Việt [7] [8], tuy nhiên, các nghiên cứu này vẫn tiếp cận phương pháp tổng hợp thích nghi dựa trên HMM và cho chất lượng tổng hợp thấp. Vì vậy, việc nghiên cứu một giải pháp tổng hợp tiếng nói tiếng Việt dựa trên thích nghi là một vấn đề cấp thiết cả về tính khoa học và tính kinh tế.

Luận án cần trả lời được các 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?
- Kích thước mẫu sẽ ảnh hưởng như thế nào đến chất lượng tổng hợp và ưu nhược điểm của các phương pháp này?

Từ các lý do cấp thiết này tôi đã chọn luận án "**Nghiên cứu phát triển hệ thống thích nghi giọng nói trong tổng hợp tiếng Việt và ứng dụng"**. Với mục tiêu chính là 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 là đề xuất phương pháp tổng hợp giọng nói dựa trên kỹ thuật thích nghi bằng mạng nơ-ron sâu (DNN) để cải thiện chất lượng tổng hợp. Và quan trọng nhất là khả năng bắt chước hoặc tạo một giọng nói mới bất kỳ với ngữ liệu huấn luyện từ đa người nói và đa phong cách với chỉ một lượng mẫu nhỏ, cụ thể:

- 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 (*từ giờ trở đi luận án gọi tắt khái niệm này bằng thuật ngữ Few-shot TTS*): 1) Mô hình tổng hợp thích nghi phụ thuộc người nói dựa trên học chuyển đổi (transfer-learning); và 2) Mô hình tổng hợp thích nghi phụ thuộc người nói dựa trên trích xuất vector biểu diễn đặc trưng.
- 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 (*từ giờ trở đi luận án gọi tắt khái niệm này bằng thuật ngữ Zero-shot TTS*).
- 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 kỹ thuật bổ sung thông tin nhãn thông qua phương pháp chèn điểm dừng lấy hơi, chèn dấu câu và phiên âm từ mượn.
- Xây dựng được ứng dụng thích nghi đa người nói sử dụng được trên các thiết bị đa nền tảng.

Luận án có ý nghĩa thực tiễn lớn bởi việc tăng cường chất lượng tổng hợp dựa trên thích nghi giúp giảm thiểu chi phí để xây dựng một giọng mới (chi phí tính toán, chi phí xây dựng dữ liệu cũng như thời gian tổng hợp), cho phép tạo giọng đọc có tính cá nhân hóa cao phục vụ đa mục đích. Hơn nữa, thích nghi giọng nói sẽ giúp tăng hiệu quả và tính thân thiện của giao tiếp người – máy bằng tiếng nói (ví dụ: các hệ thống chỉ dẫn bằng tiếng nói trong giao thông, các ki-ốt bán hàng tự động, hệ thống đọc sách báo tự động, hỗ trợ các hệ thống phiên dịch có thích nghi lời dịch, các hệ thống biến đổi/phục hồi giọng/nhân bản giọng v.v.). Tất cả các ưu điểm này sẽ mở rộng khả năng đưa công nghệ tổng hợp tiếng nói dễ dàng ứng dụng vào thực tế.

Đối tượng và phạm vi nghiên cứu của luận án là *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.

Phương pháp luận sử dụng:

- Khảo sát, phân tích các phương pháp tổng hợp tiếng nói dựa trên thích nghi mới nhất, hiệu quả nhất đã được dùng trên thế giới, lựa chọn phương pháp hiệu quả và phù hợp với tiếng Việt;
- Kế thừa các nghiên cứu đã có của cộng đồng nghiên cứu, tiếp tục nghiên cứu phát triển các phương pháp tổng hợp tiếng nói dựa trên thích nghi phù hợp với tiếng Việt;
- Dựa trên các phương pháp được nghiên cứu, phát triển thử nghiệm ứng dụng tái tạo/nhân bản giọng nói nhằm đánh giá kỹ lưỡng chất lượng tổng hợp của các mô hình.

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 cũng chính là phần cơ sở để xây dựng các mô hình thích nghi ở các chương tiếp theo. Các thử nghiệm và đánh giá cũng cho thấy rằng với các phương pháp tổng hợp thông thường bắt buộc phải sử dụng một tài nguyên lớn (hàng chục giờ cho mỗi giọng mới và hàng chục giờ huấn luyện) để thực hiện tổng hợp giọng mới và điều này là không khả thi trong thực tế. Từ đó chỉ ra nhu cầu ứng dụng các kỹ thuật thích nghi trong tổng hợp tiếng nói để giải quyết các tồn tại trên.

- **Chương 3**: Trình bày phương pháp cải tiến mô hình tổng hợp dựa trên thích nghi nhằm nâng cao chất lượng tổng hợp thông qua hai đề xuất: 1) Cải tiến mô hình tổng hợp thích nghi Few-shot TTS bằng phương pháp tinh chỉnh nhiều lần (Multi-pass fine-tune) dựa trên kỹ thuật học chuyển đổi (Transfer-learning) người nói và ngôn ngữ với lượng mẫu phải học ít hơn nhiều mô hình chinh chỉnh truyền thống (Fine-tune) hoặc so với huấn luyện mô hình cơ sở từ đầu và 2) Cải tiến mô hình tổng hợp thích nghi Few-shot TTS bằng vector EMV để biểu diễn đặc trưng giọng nói chỉ với vài câu nói. Cả hai kỹ thuật thích nghi đều yêu cầu dữ liệu mẫu phải xuất hiện trong quá trình huấn luyện và với các mô hình đề xuất hướng tới sử dụng lượng dữ liệu thích nghi ít dần. Nội dung trình bày cũng bao gồm các nghiên cứu hiện nay, mô hình đề xuất và đánh giá. Hướng tiếp cận thích nghi chỉ với vài phút dữ liệu giúp giảm độ phức tạp khi muốn xây dựng một giọng tổng hợp mới, điều này giúp tăng khả năng ứng dụng các mô hình tổng hợp trong thực tế. Từ đó, chỉ ra nhu cầu xây dựng mô hình thích nghi với lượng dữ liệu chỉ một câu duy nhất mà không cần huấn luyện lại mô hình.
- **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 giọng 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. Chương 4 trình bày một hướng tiếp cận khác để thích nghi giọng nói trong điều kiện chỉ có một câu thích nghi duy nhất và không phải huấn luyện lại mô hình. 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. Cuối luận án cũng đề xuất một ứng dụng thử nghiệm nhân bản giọng có thể chạy trên các thiết bị tính toán đa nền tảng để đánh giá tính khả thi và mô tả các minh chứng liên quan.

Chương 1. CÁC NGHIÊN CỨU LIÊN QUAN VÀ KIẾN THỨC CƠ SỞ VỀ TỔNG HỢP VÀ THÍCH NGHI GIỌNG 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, cách 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. Đặt vấn đề

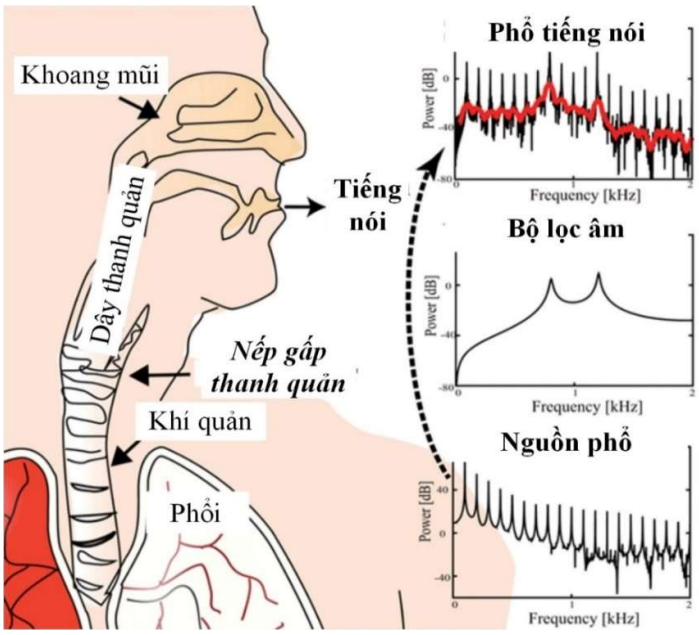
Tổng hợp tiếng nói nhân tạo đã được ứng dụng rộng rãi trong cuộc sống và từ lâu đã trở thành một chủ đề lớn trong nghiên cứu về trí tuệ nhân tạo, ngôn ngữ tự nhiên và xử lý giọng nói. Khi có sự phát triển của học sâu, tổng hợp tiếng nói dựa trên mạng nơ-ron đã phát triển mạnh, một lượng lớn công trình nghiên cứu tập trung vào các khía cạnh khác nhau của tổng hợp tiếng nói dựa trên mạng nơ-ron [9] [10] [11]. Do đó, chất lượng của tiếng nói tổng hợp đã được cải thiện đáng kể trong những năm gần đây.

Mục tiêu quan trọng nhất của một hệ thống TTS là tổng hợp được tiếng nói chất lượng cao. Chất lượng của tiếng nói được xác định bởi nhiều khía cạnh liên quan đến nhận thức lời nói, bao gồm tính dễ hiểu, tính tự nhiên, biểu cảm, ngữ điệu, cảm xúc, phong cách, độ mạnh mẽ, khả năng điều khiển giọng, v.v. Trong khi các phương pháp tiếp cận nơ-ron đã cải thiện đáng kể chất lượng của giọng nói tổng hợp đạt được như con người thì vẫn còn nhiều hướng đi nhằm cải tiến các tồn tại của hệ thống tổng hợp tiếng nói, trong đó thu nhỏ kích thước dữ liệu huấn luyện bằng các kỹ thuật thích nghi là cách tiếp cận phổ biến và chiếm ưu thế nhất [12] [13] [14].

1.2. Tổng quan về tổng hợp tiếng nói và tổng hợp thích nghi

1.2.1. Tổng hợp tiếng nói

Trước khi nói về tổng hợp tiếng nói, cần phân tích bản chất cách con người tạo ra âm thanh và tiếng nói. Âm thanh và tiếng nói của con người được tạo ra bởi sự tương tác phức tạp của các thành phần trong cơ thể người. Hầu hết các âm thanh và tiếng nói đều bắt đầu bằng hệ thống hô hấp, hệ thống này sẽ đẩy không khí ra khỏi phổi theo mô tả trong Hình 1.



Hình 1: Cấu tạo bộ lọc nguồn tạo ra âm thanh và lời nói của con người [15]

Theo lý thuyết bộ lọc nguồn [15], luồng không khí từ phổi gây ra rung động ở nếp gấp thanh quản, nơi tạo ra âm thanh nguồn. Bộ lọc âm định hình cấu trúc phổ của âm thanh nguồn. Âm thanh và lời nói được lọc cuối cùng được phát ra từ miệng.

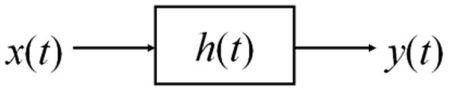
Có thể mô tả chi tiết như sau: Không khí đi qua khí quản và đi vào thanh quản, nơi có hai nếp gấp cơ nhỏ, được gọi là "Nếp gấp thanh quản". Khi các nếp gấp thanh âm được tập trung lại với nhau để tạo thành một đường dẫn khí hẹp, luồng khí làm cho chúng dao động theo chu kỳ. Các rung động nếp gấp thanh quản điều chỉnh áp suất không khí và tạo ra âm thanh tuần hoàn. Những âm thanh được tạo ra, khi các nếp gấp thanh quản rung động, được gọi là "âm hữu thanh", trong khi âm thanh mà các nếp gấp thanh quản không rung động được gọi là "âm

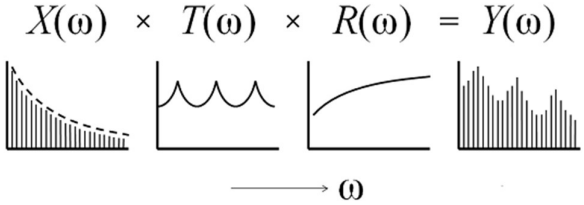
vô thanh". Các đường dẫn khí phía trên thanh quản được gọi là "Dây thanh quản". Các luồng không khí hỗn loạn được tạo ra tại các phần bị co thắt của thanh môn hoặc thanh quản cũng góp phần tạo ra âm thanh nguồn không theo chu kỳ được phân phối trên một dải tần số rộng. Hình dạng của đường thanh âm và vị trí của các khớp nối (tức là hàm, lưỡi, vật đệm, môi, miệng, răng và vòm miệng cứng) là yếu tố quan trọng để xác định đặc trưng âm học của tiếng nói. Trạng thái của các nếp gấp thanh âm cũng như vị trí, hình dạng và kích thước của các khớp nối thay đổi theo thời gian để tạo ra các âm thanh khác nhau một cách tuần tự.

Con người có thể kiểm soát quá trình phát âm (tạo nguồn) và đọc âm (lọc) một cách độc lập. Do đó, âm thanh và lời nói được coi là phản ứng của bộ lọc đường âm, nơi cung cấp nguồn âm thanh. Để mô hình hóa các hệ thống lọc nguồn như vậy cho việc tạo ra tiếng nói, nguồn âm thanh hoặc tín hiệu kích thích $x(t)$ thường được thực hiện như một hệ thống phát xung định kỳ cho giọng nói, trong khi tiếng ồn trắng được sử dụng làm nguồn cho lời nói không chỉnh âm. Nếu cấu hình đường âm không thay đổi theo thời gian, bộ lọc đường âm sẽ trở thành một hệ thống tuyến tính bất biến theo thời gian (LTI) và tín hiệu đầu ra $y(t)$ có thể được biểu thị bằng tích chập của tín hiệu đầu vào $x(t)$ và đáp ứng xung của hệ thống $h(t)$ như công thức sau:

$$y(t) = h(t) * x(t), \tag{1.1}$$

trong đó dấu hoa thị * biểu thị tích chập.





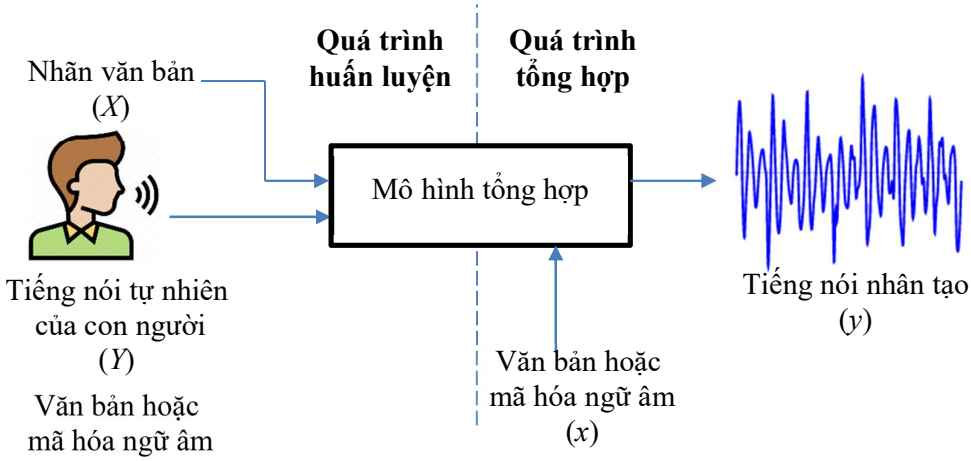
Phương trình trên được mô tả trong miền thời gian, cũng có thể được biểu diễn trong miền tần số như sau:

$$Y(\omega) = H(\omega)\, X(\omega). \tag{1.2}$$

Công thức miền tần số cho biết rằng phổ giọng nói *Y(ω)* được mô hình hóa như một sản phẩm của phổ âm thanh nguồn *X(ω)* và phổ âm thanh của bộ lọc đường âm *H(ω)*. Phổ của bộ lọc đường âm *H(ω)* được biểu thị bằng tích của hàm truyền qua đường thanh âm *T(ω)* và các đặc tính bức xạ từ miệng và mũi *R(ω)*, nghĩa là:

$$Y(\omega)=[T(\omega)\ R(\omega)]X(\omega). \quad (1.3)$$

Dựa trên nguyên tắc tổng hợp tiếng nói tự nhiên, việc tổng hợp tiếng nói nhân tạo cũng theo nguyên tắc với đầu vào dựa trên tiếng nói tự nhiên của con người nhằm tạo ra một mô hình tổng hợp, sau cùng có thể dùng mô hình tổng hợp này để sinh tiếng nói ngẫu nhiên bằng đầu vào văn bản hoặc các mã hóa ngữ âm (âm vị, mã âm thanh, mã phần cứng). Hình 2 mô tả trực quan quá trình tổng hợp tiếng nói nhân tạo.



Hình 2: 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.

Chuyển đổi văn bản thành tiếng nói (Text to Speech – viết tắt là TTS) là một công nghệ quan trọng trong tổng hợp tiếng nói, công nghệ này tạo ra sóng âm tiếng nói đầu ra một cách tùy ý từ văn bản bằng đầu vào. Nghiên cứu của luận án là một nhánh nằm trong tổng hợp tiếng nói từ văn bản (TTS).

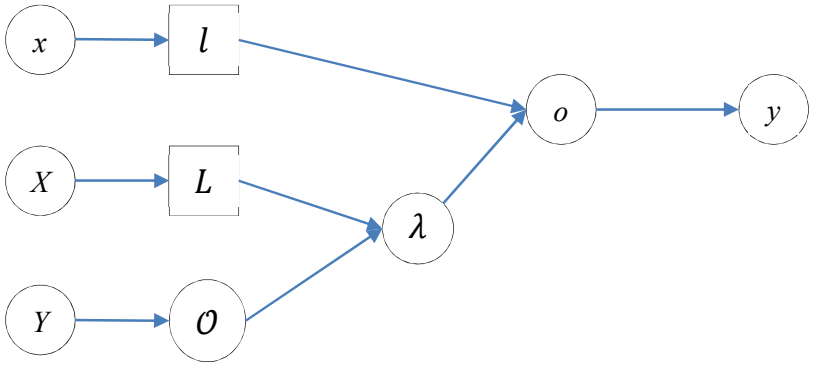
Có thể mô tả hệ thống TTS bằng mô hình tính xác suất phân phối dự đoán:

$$p(y|x,Y,X) \quad (1.4)$$

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.

Nếu coi o và $\mathcal{O}$ lần lượt là đặc trưng âm học của y và Y, l và L lần lượt là đặc trưng ngôn ngữ của x và X, λ là mô hình, thì ta có thể biểu diễn biểu thức trên dưới dạng các biến đại diện và phục thuộc như sau:

$$p(y|x,Y,X) = \iiint \sum_{\forall l} \sum_{\forall L} \left\{ \frac{p(y|o)p(o|l,\lambda)p(l|\lambda)p(l|x)p(Y|\mathcal{O})p(\mathcal{O}|L,\lambda)p(\lambda)p(L|X)}{p(Y)} \right\} dod\mathcal{O}d\lambda \quad (1.5)$$



1.2.2. Phân loại các phương pháp tổng hợp tiếng nói

Có thể phân loại các nghiên cứu về tổng hợp tiếng nói theo các nhóm sau:

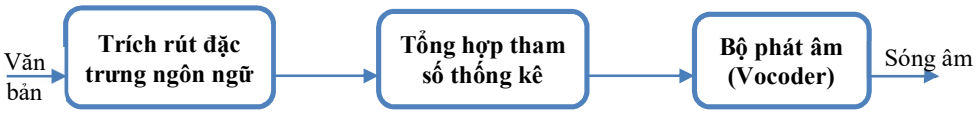
***Tổng hợp dựa trên khớp nối (Articulatory Synthesis):** Tổng hợp khớp nối [16] tạo ra tiếng nói bằng cách mô phỏng hành vi của bộ phận khớp nối của con người như môi, lưỡi, thanh quản và đường thanh âm chuyển động. Tuy nhiên, rất khó để mô hình hóa các hành vi của khớp nối này hoạt động trong thực tế. Do đó, chất lượng giọng nói do tổng hợp khớp nối thường kém hơn so với chất lượng giọng nói bằng các kỹ thuật tổng hợp sau này.

***Tổng hợp formant (Formant Synthesis):** Tổng hợp formant [17] tạo ra lời nói dựa trên một tập hợp các quy tắc điều khiển mô hình bộ lọc nguồn đơn giản hóa. Các lời nói được tổng hợp bởi một mô-đun tổng hợp phụ và một mô hình âm học với các thông số khác nhau như tần số cơ bản, giọng nói và mức tiếng ồn. Tổng hợp formant có thể tạo ra giọng nói dễ hiểu với tài nguyên tính toán vừa phải, phù hợp với các hệ thống nhúng và không dựa vào kho ngữ liệu giọng nói của con người quy mô lớn như trong tổng hợp ghép nối. Tuy nhiên, giọng nói

tổng hợp nghe kém tự nhiên hơn và hơi máy. Hơn nữa, rất khó để xác định các quy tắc tổng hợp.

***Tổng hợp dựa trên ghép nối (Concatenative Synthesis):** Tổng hợp dựa trên ghép nối [18] dựa vào việc nối các đoạn lời nói được lưu trữ trong cơ sở dữ liệu. Cách này chỉ thực sự hiệu quả khi bộ dữ liệu âm thanh đủ lớn cả về kích thước, độ đa dạng phát âm và các đặc trưng phổ âm thanh. Có ba kiểu tổng hợp ghép nối: Tổng hợp chọn đơn vị âm; Tổng hợp âm kép (dipphone) và; Tổng hợp chuyên ngành. Tuy nhiên, TTS dựa trên ghép nối yêu cầu cơ sở dữ liệu ghi âm khổng lồ, bao gồm tất cả các xác suất kết hợp có thể có của các đơn vị tiếng nói. Một nhược điểm nữa là giọng nói được tạo ra kém tự nhiên và kém tính truyền cảm do điểm ghép nối kém mượt do sự căng thẳng, cảm xúc, ngữ điệu, khác nhau tại từng thời điểm thu âm.

***Tổng hợp dựa trên tham số thống kê (Statistical Parametric Synthesis-SPSS):** Để giải quyết những hạn chế của TTS ghép nối, tổng hợp tiếng nói tham số thống kê (SPSS) được đề xuất [19]. Ý tưởng cơ bản là thay vì tạo trực tiếp dạng sóng thông qua ghép nối, trước tiên có thể tạo ra các tham số âm thanh [20] [21] cần thiết để tạo ra giọng nói, sau đó khôi phục giọng nói từ các tham số âm thanh đã tạo bằng một số thuật toán [22] [23] [24].



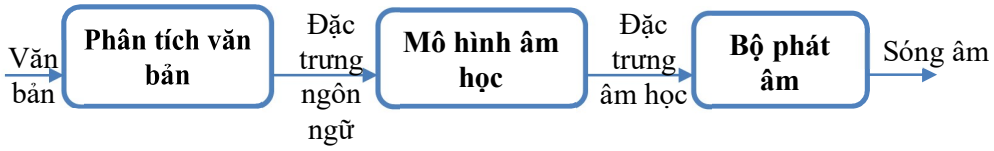
Hình 3: Tổng hợp tiếng nói dựa trên tham số thống kê

SPSS thường bao gồm ba thành phần: một mô-đun phân tích văn bản nhằm trích rút đặc trưng ngôn ngữ ; một mô-đun dự đoán tham số thống kê (mô hình âm học) và một mô-đun phân tích, tổng hợp thành sóng âm. Có thể mô tả các mô-đun trong Hình 3. SPSS có một số ưu điểm so với các hệ thống TTS trước đây: 1) Tính tự nhiên, âm thanh tự nhiên hơn; 2) Tính linh hoạt, để sửa đổi các tham số một cách thuận tiện để kiểm soát lời nói tạo ra; 3) Chi phí dữ liệu thấp, yêu cầu ít bản ghi hơn so với tổng hợp ghép nối. Tuy nhiên, SPSS cũng có những hạn chế

của nó: 1) Giọng nói được tạo ra có độ rõ ràng kém hơn so với thực tế như: âm thanh bị bóp nghẹt, ù hoặc nhiễu; 2) Giọng nói được tạo ra vẫn mang đặc trưng giọng máy và có thể dễ dàng phân biệt với giọng nói ghi âm của con người.

Trong những năm 2010, khi mạng nơ-ron và học sâu đã đạt được tiến bộ nhanh chóng, một số công trình lần đầu tiên đưa mạng nơ-ron sâu (Deep Neural Network viết tắt là DNN) vào SPSS, ví dụ như dựa trên DNN [9] và dựa trên mạng nơ-ron hồi quy (RNN). Tuy nhiên, các mô hình này thay thế HMM bằng mạng nơ-ron và vẫn dự đoán các đặc trưng âm học từ các đặc trưng ngôn ngữ, tuân theo mô hình của SPSS.

***Tổng hợp giọng dựa trên mạng nơ-ron**: Với sự phát triển của học sâu, TTS dựa trên mạng nơ-ron được đề xuất, sử dụng DNN làm mô hình xương sống để tổng hợp giọng nói. Một mô hình tổng hợp dựa trên mạng nơ-ron có thành phần như mô tả trong Hình 4.

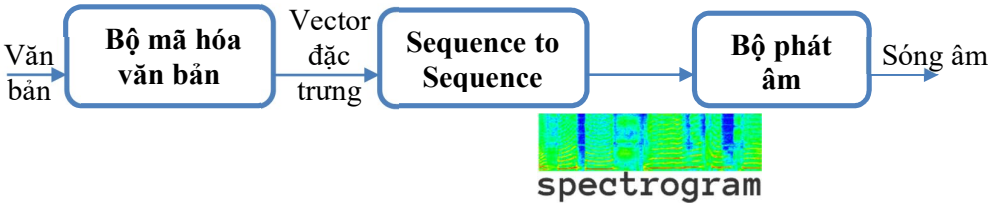


Hình 4: Ba thành phần cơ bản của mạng nơ-ron TTS

Một số mô hình nơ-ron ban đầu được áp dụng trong SPSS để thay thế HMM cho mô hình âm học, sau đó, WaveNet [10] được đề xuất để trực tiếp tạo ra dạng sóng từ các đặc trưng ngôn ngữ. Đây có thể được coi là mô hình TTS dựa trên mạng nơ-ron hiện đại đầu tiên. Các mô hình khác như DeepVoice1/2 [25] vẫn tuân theo ba thành phần trong tổng hợp tham số thống kê, nhưng nâng cấp chúng với các mô hình dựa trên mạng nơ-ron tương ứng. So với các hệ thống TTS trước đây dựa trên tổng hợp ghép nối và tổng hợp tham số thống kê, ưu điểm của tổng hợp giọng nói dựa trên mạng nơ-ron bao gồm chất lượng giọng nói cao cả về độ dễ hiểu và tính tự nhiên, đồng thời ít yêu cầu quá trình tiền xử lý phức tạp. Sau đó, nghiên cứu [26] đề xuất tạo ra trực tiếp các đặc trưng âm thanh từ chuỗi âm vị thay vì các đặc trưng ngôn ngữ. Đây có thể được coi là nghiên cứu đầu tiên về tổng

hợp giọng nói dựa trên mô hình một luồng vào ra (trong luận án sẽ giữ nguyên thuật ngữ tiếng Anh End-to-end)[1].

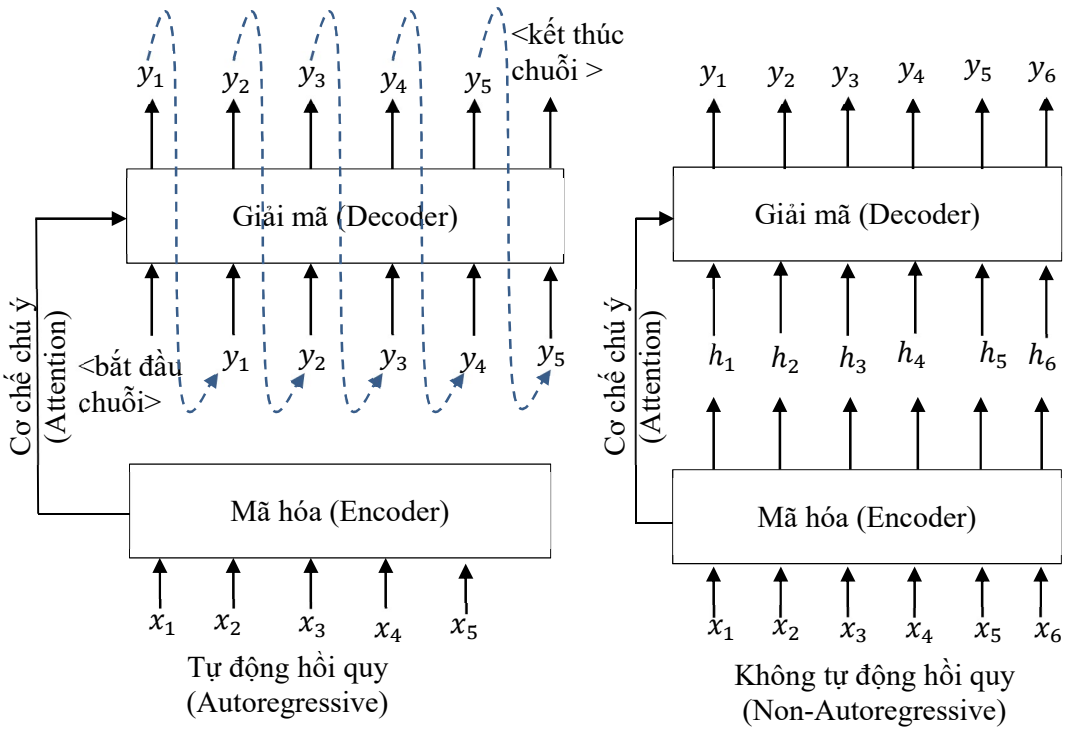
*** Tổng hợp dựa trên End-to-end:** Hình 5 mô tả kiến trúc một hệ thống tổng hợp End-to-end, trong đó phần tạo phổ Mel gồm hai mô đun: Đầu tiên là mô đun mã hóa văn bản để chuyển hóa chuỗi ký tự thành chuỗi các véc tơ biểu diễn, sau đó mô đun Sequence-to-sequence (*từ giờ trở đi luận án gọi tắt là Seq2seq*) sẽ ước lượng phổ Mel từ chuỗi véc tơ này. Cuối cùng, từ phổ Mel được chuyển đổi thành tín hiệu tiếng nói nhờ bộ phát âm. Một số mô hình End-to-end tiêu biểu (Ví dụ: Tacotron 1/2 [27] [28], Deep Voice 3 [29] và FastSpeech1/2 [11] [30]) được đề xuất để đơn giản hóa các mô-đun phân tích văn bản và trực tiếp lấy chuỗi ký tự hoặc âm vị làm đầu vào, và đơn giản hóa các đặc trưng âm thanh với phổ Mel. Sau đó, các hệ thống TTS End-to-end hoàn chỉnh được phát triển để tạo trực tiếp dạng sóng âm từ văn bản, chẳng hạn như ClariNet [31], FastSpeech2 [30] và EATS [32].



Hình 5: Tổng hợp End-to-end TTS

Hiện nay, có hai kiến trúc TTS phổ biến tiên tiến nhất là: 1) Kiến trúc tự hồi quy (autoregressive) và; 2) Kiến trúc không tự động hồi quy (non-autoregressive). So sánh hai kiến trúc được mô tả trong Hình 6.

[1] Thuật ngữ "End-to-end" trong TTS có một ý nghĩa chưa nhất quán. Trong các nghiên cứu ban đầu, "End-to-end" đề cập đến mô hình chuyển văn bản thành phổ mel là End-to-end, nhưng vẫn sử dụng một bộ tổng hợp dạng sóng riêng biệt (vocoder). Nó cũng có thể đề cập đến các mô hình TTS dựa trên mạng nơ-ron không sử dụng các đặc trưng ngôn ngữ hoặc âm học phức tạp. Tuy nhiên, mô hình End-to-end nghiêm ngặt đề cập đến việc tạo trực tiếp dạng sóng từ văn bản.



Hình 6: So sánh mô hình tự động hồi quy và không tự động hồi quy

i) Kiến trúc TTS tự hồi quy (autoregressive TTS)

Cho $x = \{x_1, x_2, \ldots, x_m\}$ biểu diễn cho chuỗi văn bản đầu vào, và $y_{<t}$ biểu diễn cho các khung tiếng nói được tạo ra trước $y_t, y = \{y_1, y_2, \ldots, y_n\}$ đại diện cho chuỗi mục tiêu của các đặc trưng âm học.

Quá trình giải mã khung tiếng nói thứ *t* là y_t tuân theo phân phối có điều kiện $p(y_t|y_{<t}, \boldsymbol{x}, \theta)$, trong đó khung tiếng nói y_t phụ thuộc vào toàn bộ chuỗi văn bản đầu vào *x*, lịch sử giải mã một phần $y_{<t}$ và các tham số mạng θ. Xác suất chung của một bộ giải mã tự động có thể được xác định như sau:

$$p(\boldsymbol{y}|\boldsymbol{x}, \theta) = \prod_{t=1}^{n} p(y_t|y_{<t}, \boldsymbol{x}, \theta) \quad (1.6)$$

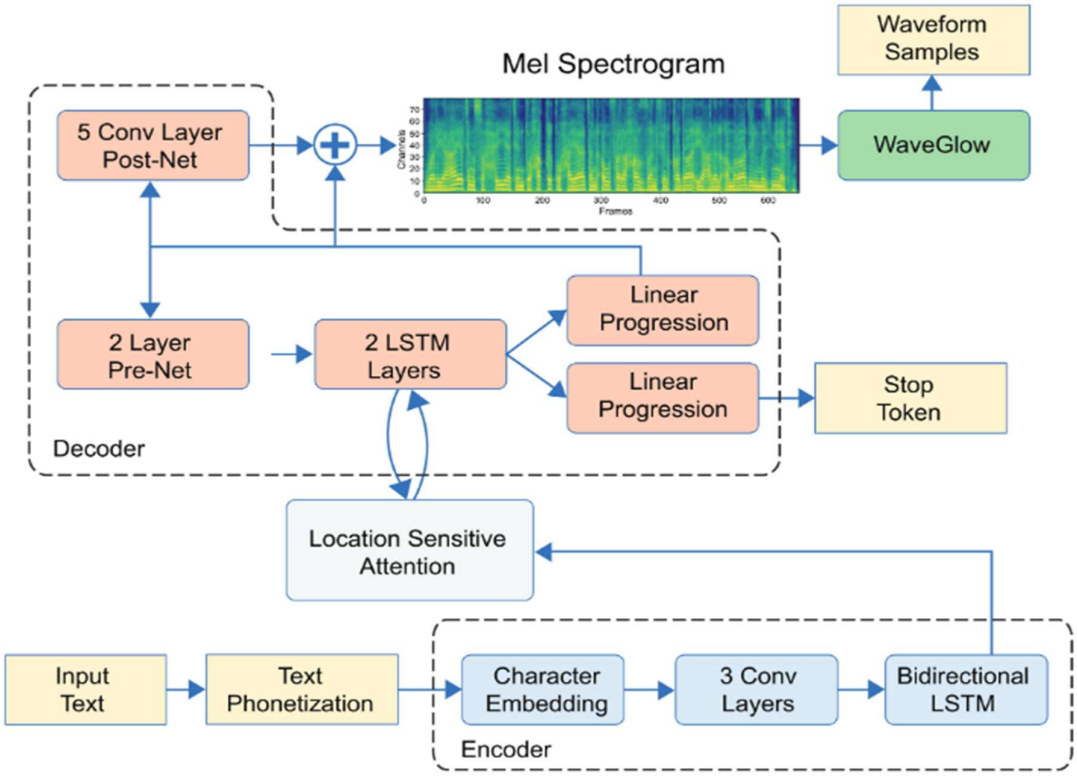
Điều kiện trên *x* đảm bảo tính phù hợp về ngữ âm và điều kiện dựa trên lịch sử giải mã một phần $y_{<t}$ đảm bảo sự phụ thuộc thời gian. Vì bộ giải mã có tác động trực tiếp đến sự phụ thuộc thời gian của tín hiệu âm học đầu ra, phép giải mã tự hồi quy thường được thực hiện trên bộ giải mã. Mặc dù phép giải mã tự hồi quy này đảm bảo sự phụ thuộc thời gian của tín hiệu âm học, nhưng nó không cho phép xử lý song song và liên quan đến tính toán lặp nhiều. Ngoài ra, các vấn đề cố hữu của hệ thống TTS tự hồi quy là khó điều khiển tốc độ đọc và ngữ điệu cũng như lỗi bỏ qua hoặc lặp từ [11] [33].

- ***Kiến trúc chi tiết của Tacotron2***

Kiến trúc TTS tự hồi quy nổi bật như Tacotron2 tạo ra phổ Mel từ văn bản và sau đó tổng hợp giọng nói từ các phổ Mel đã được tạo bằng cách sử dụng một bộ phát âm đã được huấn luyện riêng. Chúng thường mắc phải vấn đề là tốc độ suy diễn chậm và gặp các vấn đề tồn tại (bỏ qua và lặp lại từ). Xương sống của Tacotron là mô hình Seq2seq với cơ chế chú ý (attention) [34]. Hình 7 mô tả mô hình tổng hợp tiếng nói bằng Tacotron2, bao gồm một bộ mã hóa (*luận án sẽ giữ nguyên thuật ngữ tiếng Anh là Encoder*), một bộ giải mã (*luận án sẽ giữ nguyên thuật ngữ tiếng Anh là Decoder*) dựa trên cơ chế chú ý và một mạng xử lý sau. Ở mức cao, mô hình Tacotron lấy các ký tự làm đầu vào và tạo ra các phổ Mel, sau đó được chuyển đổi thành dạng sóng. Kiến trúc của Tacotron và Tacotron2 khá giống nhau, đều chia thành hai phần riêng biệt:

– Phần 1: Mạng dự đoán phổ âm thanh được dùng để chuyển đổi chuỗi ký tự (text) sang dạng phổ Mel ở miền tần số.

– Phần 2: Bộ phát âm biến đổi âm thanh từ phổ Mel (miền tần số) sang sóng âm (miền thời gian).



Hình 7: Sơ đồ khối kiến trúc hệ thống của Tacotron2 [28]

ii) Kiến trúc TTS không tự hồi quy (non-autoregressive TTS)

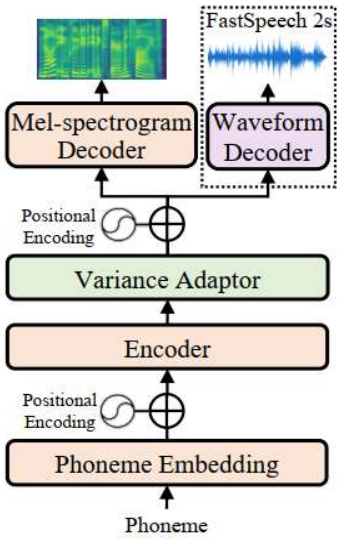
Trong những năm gần đây, các mô hình TTS không tự động hồi quy (non-autoregressive) được thiết kế cho phép tạo ra các phổ Mel với tốc độ cực nhanh và tránh các vấn đề tồn tại của hệ thống TTS tự động hồi quy, đồng thời đạt được chất lượng âm thanh gần tương đương với các mô hình tự hồi quy trước đó dựa trên kiến trúc mạng chuyển đổi chuyển tiếp (Feed-Forward Transformer –viết tắt là FFT) mới cho phép xử lý song song với cơ chế tự chú ý. Một bộ mã hóa FFT mã hóa một chuỗi từ đầu vào văn bản thành các chuỗi ẩn và một bộ giải mã FFT tạo ra các đặc trưng âm thanh đầu ra từ các vector ngữ cảnh một cách song song. Cùng với mô hình bộ phát âm song song, việc triển khai toàn bộ mô hình song song cho độ trễ thấp. Phương pháp không tự hồi quy không phụ thuộc vào lịch sử giải mã, nó tạo ra các đặc trưng âm thanh đầu ra dưới dạng tích của xác suất độc lập có điều kiện của từng khung tiếng nói:

$$p(\boldsymbol{y}|\boldsymbol{x},\theta) = \prod_{t=1}^{n} p(y_t|\boldsymbol{x},\theta) \quad (1.7)$$

Sự độc lập có điều kiện cho phép mô hình tạo ra nhiều khung tiếng nói mục tiêu đồng thời, dẫn đến giảm đáng kể việc tính toán so với phần tử tự hồi quy. Mô hình không tự hồi quy chỉ phụ thuộc vào cơ chế tự chú ý của nó để giải mã các đặc trưng âm thanh từ các đặc trưng văn bản đã được mã hóa, mà không có bất kỳ đầu vào tự hồi quy nào.

- Kiến trúc chi tiết của FastSpeech2

Trong số các phương pháp TTS không tự hồi quy, FastSpeech là một trong những mô hình thành công nhất. FastSpeech2 được ra đời thay thế FastSpeech để giải quyết vấn đề ánh xạ một-nhiều với cơ chế huấn luyện theo cơ chế đường ống (pipeline) đơn giản và cho chất lượng tổng hợp âm thanh cao [11].



Hình 8: Kiến trúc tổng quan của FastSpeech2 [30]

Kiến trúc mô hình tổng thể của FastSpeech2 được thể hiện trong Hình 8. Bộ mã hóa chuyển đổi chuỗi nhúng âm vị (phoneme embedding) thành chuỗi âm vị ẩn, sau đó bộ thích nghi phương sai (variance adaptor) thêm thông tin phương sai khác nhau như trường độ, cao độ và cường độ vào chuỗi ẩn, cuối cùng là bộ giải mã phổ Mel sẽ chuyển đổi chuỗi ẩn đã thích nghi thành chuỗi phổ Mel một cách song song. Kiến trúc sử dụng khối FFT, đây là một lớp ngăn xếp của lớp tự chú ý (self-attention) [35] và tích chập 1D như trong FastSpeech, làm cấu trúc cơ bản cho bộ mã hóa và bộ giải mã phổ Mel, kiến trúc này giúp sinh phổ Mel song song thay vì một các tuần tự như các mô hình TTS tự động hồi quy. Khác với FastSpeech huấn luyện dựa trên hệ thống chưng cất (distillation) giữa bộ dạy và bộ học (teacher-student) và trường độ âm vị (phoneme duration), FastSpeech2 thực hiện được một số cải tiến. Đầu tiên, nó loại bỏ đường ống chưng cất giữa bộ dạy và bộ học và sử dụng trực tiếp các phổ Mel của âm thanh gốc (groundtruth) làm mục tiêu huấn luyện mô hình, điều này có thể tránh mất thông tin trong các phổ Mel được chưng cất và tăng chất lượng giọng nói. Thứ hai, bộ variance adaptor không chỉ bao gồm bộ dự đoán trường độ mà còn cả bộ dự đoán cao độ và cường độ. Trong đó: 1) Bộ dự đoán trường độ sử dụng trường độ âm vị thu được bằng cách căn chỉnh cưỡng bức [36] làm mục tiêu huấn luyện chính xác hơn thay vì trích xuất từ ánh xạ bộ chú ý của mô hình dạy tự phục hồi và 2) Các công

cụ dự đoán cao độ và cường độ bổ sung có thể cung cấp thêm thông tin về phương sai, điều này rất quan trọng để giải quyết vấn đề ánh xạ một-nhiều trong TTS.

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

Trước đây, mô hình của tổng hợp tiếng nói dựa trên tham số thống kê đã thay thế hoàn toàn tổng hợp tiếng nói dựa trên lựa chọn đơn vị bởi khả năng thích nghi và điều khiển các đặc trưng của người nói và phong cách nói. Tổng hợp tiếng nói dựa trên HMM có thể áp dụng thành công cho nhiều nghiên cứu mở rộng bằng các kỹ thuật thích nghi giọng nói và đã được chứng minh là cải thiện đáng kể chất lượng tiếng nói tổng hợp [37]. Bởi vì, tổng hợp thống kê dựa trên HMM có thể sử dụng các phương pháp nội suy [38], hồi quy đa vector cảm xúc [39] và kỹ thuật thích nghi [40] để dễ dàng chuyển đổi hoặc điều chỉnh phong cách và cảm xúc nói, phương pháp này đã trở thành phương pháp chính trong tổng hợp tiếng nói có cảm xúc trong các giai đoạn trước.

Mặc dù tổng hợp dựa trên thống kê bằng HMM đã cho chất lượng tốt nhưng nó vẫn còn các hạn chế. Đầu tiên phải kể đến là ánh xạ đầu vào đến phân cụm dựa trên cây quyết định trong tổng hợp giọng nói dựa trên HMM không hiệu quả để diễn đạt các phụ thuộc ngữ cảnh phức tạp và vấn đề XOR (không tính toán được perceptron đơn), điều này có thể dẫn đến quá khớp (overfit) dữ liệu huấn luyện. Thứ hai, ánh xạ đặc trưng thành cụm sử dụng các phân bố Gauss đơn lẻ với ma trận hiệp phương được thiết lập dựa trên hai giả định về tính độc lập: 1) Sự độc lập có điều kiện giữa các khung trạng thái và 2) Sự độc lập của các đặc trưng âm thanh trong một khung. Điều này dẫn đến các đường bao phổ (envelopes spectral) được tái tạo bị làm mịn quá mức và chất lượng của giọng nói tổng hợp bị giảm sút.

Từ năm 2013, các kỹ thuật thống kê sử dụng DNN chiếm ưu thế [9], đặc biệt trong vài năm gần đây, các kỹ thuật áp dụng mạng đối nghịch GAN cho tổng hợp tiếng [41] và kiến trúc End-to-end [42] càng trở nên phổ biến. 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. Mặt khác, DNN lập mô hình xác suất có điều kiện thay vì xác suất chung như trong mô hình DBN (Deep Belief Networks), nó hiệu quả với nhiệm vụ ánh xạ đặc trưng. 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. Do đó, các phương pháp thích nghi đa người nói dựa trên DNN đã được đề xuất.

1.2.4. Tổng hợp tiếng nói hiệu quả

Khi có thể tổng hợp giọng nói chất lượng cao và có khả năng điều khiển được thì nhiệm vụ quan trọng tiếp theo là tổng hợp giọng nói hiệu quả (Efficient Speech Synthesis), tức là làm thế nào để giảm chi phí tổng hợp giọng nói bao gồm chi phí thu thập và gán nhãn dữ liệu, huấn luyện và duy diễn cho các mô hình TTS. Các cách tiếp cận chính khi xây dựng các mô hình tổng hợp giọng hiệu quả có thể nhóm thành các mục tiêu:

- **TTS ít dữ liệu**. Nhiều ngôn ngữ nghèo tài nguyên và thiếu dữ liệu huấn luyện. Làm thế nào để tận dụng việc học không giám sát/bán giám sát và học chuyển đổi đa ngôn ngữ để giúp các ngôn ngữ có nguồn tài nguyên thấp là một hướng đi nhận được nhiều sự quan tâm. Ví dụ như 'ZeroSpeech Challenge' [43] là một sáng kiến hay để khám phá các kỹ thuật cho phép học mô hình ngôn ngữ trực tiếp từ âm thanh mà không cần bất kỳ văn bản hoặc kiến thức ngôn ngữ nào. Kỹ thuật sử dụng là sự kết hợp của ba thành phần không giám sát: Bộ mã hóa dự đoán tương phản (CPC), bộ phân cụm (k-mean) và mô hình ngôn ngữ (LSTM hoặc BERT). Bên cạnh đó 'Multi-speaker multi-style voice cloning challenge' [5] cũng mở ra các thách thức trong nhân bản giọng, điều chỉnh giọng nói của một người nói đích có ít dữ liệu để thích nghi thông qua hai nhiệm vụ: 1) Nhân bản giọng từ 100 mẫu/người nói cho trước kèm một bộ dữ liệu đa người nói (Few-shot track) và 2) Nhân bản giọng từ 5 mẫu/người nói cho trước (One-shot track), các kỹ thuật chính được sử dụng cho các nhiệm vụ này là thích nghi dựa trên mã hóa đặc trưng (Embedding based) và thích nghi dựa trên tinh chỉnh mô hình (Fine-tuning based).
- **TTS ít tham số**. Các hệ thống TTS bằng nơ-ron ngày nay thường sử dụng mạng nơ-ron lớn với hàng chục triệu tham số để tổng hợp giọng nói chất lượng cao. Hệ thống này hạn chế khả năng ứng dụng trên các thiết bị di động, IoT và các thiết bị tài nguyên nhỏ khác do bộ nhớ và mức tiêu thụ điện năng lớn. Việc thiết kế các mô hình nhỏ, gọn, nhẹ với bộ nhớ nhỏ, tiêu thụ điện năng thấp và ít độ trễ là rất quan trọng.

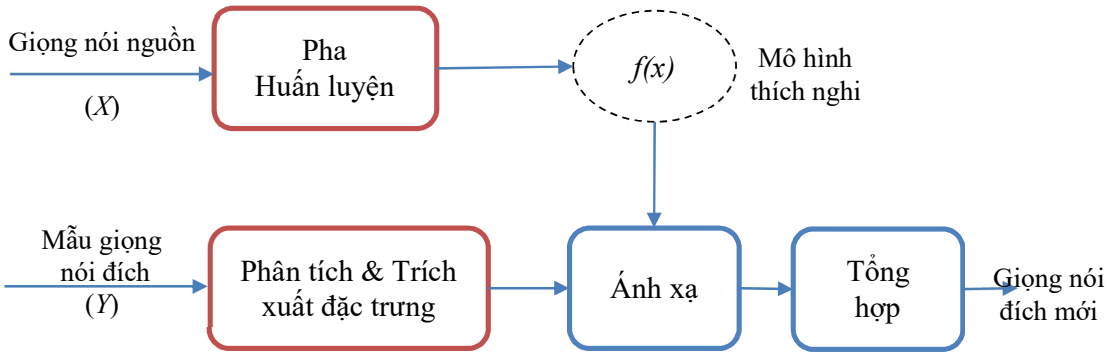
- **TTS ít năng lượng**. Huấn luyện và phục vụ một mô hình TTS chất lượng cao đòi hỏi tiêu tốn nhiều năng lượng và thải ra nhiều carbon. Việc cải thiện hiệu quả sử dụng năng lượng, ví dụ như giảm FLOP trong huấn luyện và suy diễn TTS là giúp nhiều người được hưởng lợi từ các kỹ thuật TTS tiên tiến, đồng thời giảm lượng khí thải carbon ra môi trường.

1.2.5. Thích nghi trong tổng hợp tiếng nói

1.2.5.1. 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 mang phong cách tùy ý từ bất kỳ người nào với một lượng dữ liệu mẫu thực của giọng nói đó (*reference speech*), giọng nói tổng hợp sẽ mang đặc trưng của giọng nói đích hoặc giọng tham chiếu (*target speaker* hoặc *reference speaker*) với các đặc trưng của giọng nói (*voice characteristics*) và các đặc trưng ngữ điệu (*prosodic features*) [4]. Thích nghi TTS tùy theo ngữ cảnh được gọi bằng các thuật ngữ khác nhau trong học thuật và công nghiệp, ví dụ như thích nghi giọng nói (*voice adaptation)* hoặc thích nghi người nói (*speaker adaptation*) [44], nhân bản giọng nói (*voice cloning*) [45], cá nhân hóa giọng nói (*custom voice*) [13]. Thích nghi TTS là một chủ đề nghiên cứu nóng, rất nhiều các công trình trong lĩnh vực tổng hợp tiếng nói gần đây đã nghiên cứu sự thích nghi giọng nói [13] và thử thách nhân bản giọng nói cũng thu hút rất nhiều nhóm nghiên cứu tham gia [5] [46]. Trong luận án này khái niệm thích nghi giọng nói (hoặc thích nghi người nói) được định nghĩa giới hạn là khả năng nhân bản giọng nói của người nói tùy ý với lượng dữ liệu mẫu hạn chế.

1.2.5.2. Mô hình



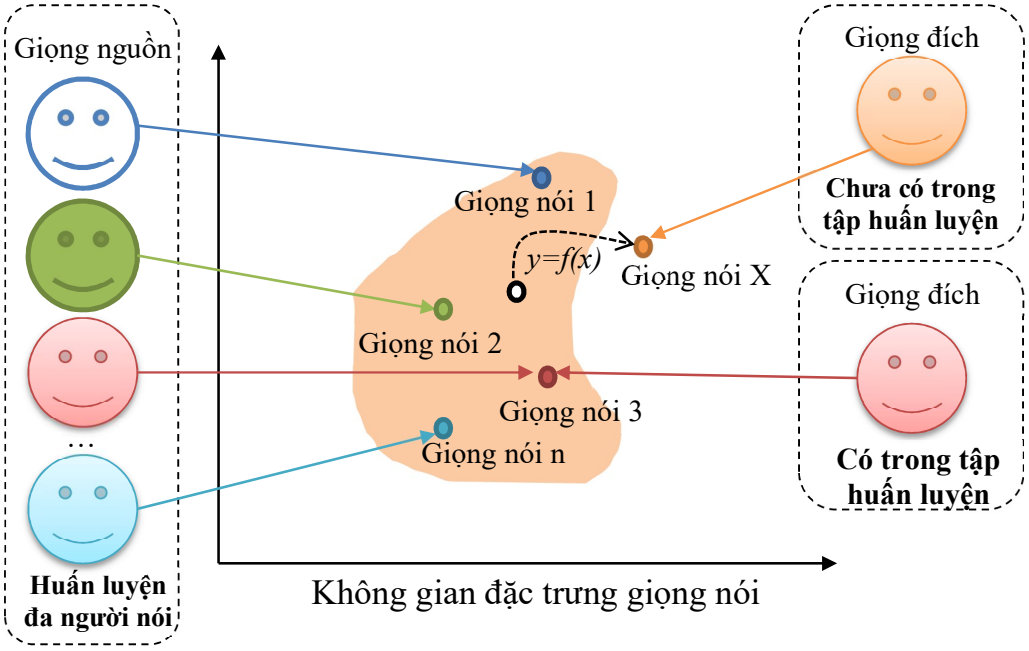
Hình 9: Mô hình tổng quát của hệ thống tổng hợp tiếng nói dựa trên thích nghi

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:

$$y = f(x) \tag{1.8}$$

trong đó $f(.)$ còn được gọi là hàm ánh xạ theo khung. Có thể biểu diễn tổng quát các khối trong hệ thống tổng hợp thích nghi như Hình 9 và và biểu diễn không gian đặc trưng tổng quát với mục tiêu dịch tâm thích nghi trong Hình 10.

Trong kịch bản thích nghi TTS, mô hình nguồn TTS (thường được huấn luyện trên tập dữ liệu đa người nói (Multi-speaker)) thường được thích nghi với ít dữ liệu thích nghi của từng giọng nói đích. 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 và sự thích nghi với các miền khác nhau; 2) Cài đặt thích nghi hiệu quả, bao gồm việc giảm dữ liệu thích nghi và các thông số thích nghi cho từng giọng nói đích.



Hình 10: Không gian đặc của hệ thống tổng hợp tiếng nói dựa trên thích nghi

1.2.5.3. Thích nghi chung

- **Tổng quát hóa mô hình nguồn**. Các công trình nghiên cứu trong nhóm này hướng đến cải thiện tính tổng quát hóa của mô hình TTS nguồn. Trong huấn

luyện mô hình nguồn, văn bản nguồn không chứa đủ thông tin âm học như âm thanh, âm sắc của giọng nói và môi trường ghi âm để tạo ra giọng nói đích. Kết quả là, mô hình TTS có xu hướng được trang bị quá nhiều dữ liệu huấn luyện và có khả năng tổng quát hóa kém đối với những giọng nói mới (thích nghi kém). Nghiên cứu [13] đề xuất mô hình điều kiện âm học, cung cấp thông tin âm học cần thiết làm đầu vào mô hình để học cách ánh xạ văn bản thành giọng nói với khả năng tổng quát hóa tốt hơn thay vì ghi nhớ. Một cách khác để cải thiện tính tổng quát của mô hình TTS nguồn là tăng số lượng và tính đa dạng của dữ liệu huấn luyện. Nghiên cứu [12] nâng số người nói để tăng số lượng giọng khi huấn luyện mô hình TTS nguồn, có thể khái quát tốt giọng nói chưa từng xuất hiện trong quá trình huấn luyện cho mục tiêu thích nghi.

- **Thích nghi đa miền.** Trong thích nghi TTS, một yếu tố quan trọng là tiếng nói dùng thích nghi có các điều kiện hoặc phong cách âm học khác với dữ liệu giọng nói được sử dụng để huấn luyện mô hình TTS nguồn. Theo cách này, các thiết kế đặc biệt cần được xem xét để cải thiện tính tổng quát hóa của mô hình TTS nguồn và hỗ trợ các phong cách của giọng nói đích. AdaSpeech [13] thiết kế mô hình điều kiện âm học để mô hình hóa tốt hơn trong các điều kiện âm học đa dạng như thiết bị ghi âm, tiếng ồn môi trường, trọng âm, tốc độ nói, âm sắc của giọng nói, v.v. Bằng cách này, mô hình có xu hướng tổng quát hóa thay vì ghi nhớ các điều kiện âm học và có thể thích nghi tốt với dữ liệu giọng nói với các điều kiện âm học khác nhau. AdaSpeech3 [3] điều chỉnh mô hình TTS theo phong cách đọc sang phong cách tự do, bằng cách thiết kế thích nghi chèn khoảng dừng, thích nghi nhịp điệu và âm sắc. Một số nghiên cứu khác đề cập kỹ thuật chuyển đổi giữa các phong cách nói một cách tự nhiên, chẳng hạn như tiếng Lombard [47] hoặc tiếng thì thầm [48]. Một số công trình [49], [50] đề xuất thích nghi giọng nói giữa các ngôn ngữ, ví dụ: tổng hợp giọng nói tiếng Trung bằng cách sử dụng giọng nói tiếng Anh để huấn luyện mà dữ liệu huấn luyện không có bất kỳ giọng nói tiếng Trung nào.

1.2.5.4. Thích nghi hiệu quả

Nhiều dữ liệu thích nghi hơn sẽ dẫn đến chất lượng tổng hợp tốt hơn, nhưng chi phí thu thập dữ liệu cao. Đối với thích nghi tham số, toàn bộ mô hình TTS [44] [51] hoặc một phần của mô hình (Ví dụ: bộ giải mã) hoặc chỉ vector biểu diễn đặc trưng giọng nói speaker-embedding (*luận án này sẽ giữ nguyên thuật ngữ tiếng Anh speaker-embedding cho một số trường hợp*) [45] [44] [13] có thể được tinh chỉnh. Tương tự, việc tinh chỉnh nhiều thông số hơn sẽ cho chất lượng âm thanh tốt, nhưng lại làm tăng chi phí triển khai và bộ nhớ. Trong thực tế, mục tiêu thích nghi là sử dụng ít dữ liệu ít tham số nhưng vẫn đạt được chất lượng thích nghi giọng nói cao. Có thể chia các nghiên cứu 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. Các kiến thức cơ sở

1.3.1. Cơ sở vật lý

Cơ sở vật lý của âm thanh đặc trưng bởi ba yếu tố chính: trường độ, cao độ và cường độ (năng lượng).

a) Trường độ (duration): Trường độ liên quan đến độ dài của âm thanh, phụ thuộc vào thời gian lâu hay mau mà âm phát ra.

b) Cao độ (pitch): Cao độ của âm thanh phụ thuộc vào tần số dao động của dây thanh trong một đơn vị thời gian. Âm cao có tần số dao động nhanh hơn, trong khi âm thấp có tần số dao động chậm hơn. Đơn vị đo cao độ là Hz. Độ cao của giọng có thể cung cấp thông tin về giới tính, tuổi tác và cảm xúc.

Một số ngôn ngữ như tiếng Việt và tiếng Trung sử dụng cao độ để tạo ra những đơn vị ngữ âm như thanh điệu, ngữ điệu và trọng âm.

c) Cường độ (energy): Cường độ liên quan đến độ mạnh của âm thanh, và phụ thuộc vào biên độ dao động của vật thể. Khi biên độ dao động càng rộng, âm thanh càng mạnh và vang lớn. Đơn vị đo cường độ là dB. Cường độ là yếu tố quan trọng tạo nên trọng âm trong ngôn ngữ.

1.3.2. Cấu tạo tiếng Việt

Tiếng nói tiếng Việt là một ngôn ngữ phức tạp so với các ngôn ngữ khác bởi nó là ngôn ngữ đơn âm tiết có thanh điệu, mọi âm tiết luôn mang một thanh điệu nào đó (Bảng 1) [52]:

Bảng 1: Sơ đồ cấu tạo âm tiếng Việt

THANH ĐIỆU			
ÂM ĐẦU (phụ âm đầu)	VẦN		
	Âm đệm	Âm chính	Âm cuối

1.3.2.1. Đơn vị ngữ âm

***Âm tiết (syllable):** Chuỗi từ được con người tạo thành từ các hợp âm tĩnh từ lớn đến nhỏ. Đơn vị từ nhỏ nhất là âm tiết. Một từ như "tổng hợp" được phát âm thành "tổng" và "hợp", gồm 2 âm tiết. Có thể chia các âm tiết thành bốn loại [53]: Âm tiết mở; Âm tiết nửa mở; Âm tiết nửa kín; Âm tiết đóng. Tổng số âm tiết không trùng lặp phát âm bằng tiếng Việt là 18.958 nhưng số âm tiết đã sử dụng trong thực tế chỉ khoảng 7000 âm tiết [54].

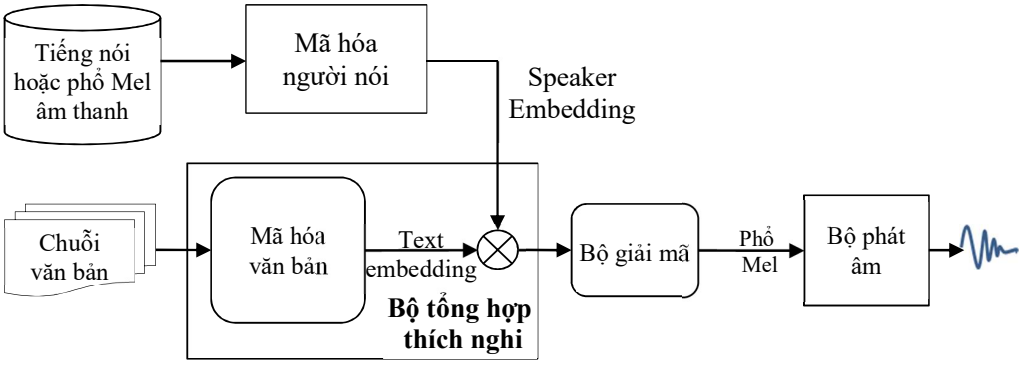
*** Âm thanh (sound):** Âm thanh (hay còn gọi là âm thanh của ngôn ngữ) là đơn vị tự nhiên nhỏ nhất của lời nói. Tiêu chí thường được sử dụng để phân biệt giữa các âm thanh là đặc trưng âm học và đặc trưng nguyên âm. Theo đó, âm thanh được chia thành hai loại chính là nguyên âm (vocal) và phụ âm (consonant).

*** Âm vị (phoneme):** Âm vị là phân đoạn nhỏ nhất của âm thanh, dùng để cấu tạo và phân biệt vỏ âm thanh của các đơn vị có nghĩa của ngôn ngữ hoặc phân biệt giữa các cách phát âm. Dưới âm vị không thể phân chia được nữa, nếu chia nhỏ, sẽ không thu được gì [55]. Tiếng Việt có 22 âm vị phụ âm, 14 âm vị nguyên âm và 2 âm vị bán nguyên âm.

***Hệ thống thanh điệu (tone):** Thanh điệu là thuật ngữ dùng để chỉ độ cao của âm tiết được sinh ra do sự rung động của dây thanh âm. Tùy vào độ rung nhanh hay chậm, mạnh hay yếu của dây thanh âm sẽ cho các âm sắc khác nhau. Tiếng Việt gồm có 6 thanh điệu. Trên chữ viết, năm thanh được ghi lại bằng năm dấu trong chữ viết: 1. Thanh huyền, 2. Thanh ngã, 3. Thanh hỏi, 4. Thanh sắc, 5. Thanh nặng, còn một thanh không được ghi lại bằng một dấu nào cả là thanh ngang.

1.3.3. Các thành phần chính của hệ thống tổng hợp thích nghi

Mô hình tổng hợp thích chung được mô tả trong Hình 11 gồm: 1) Một mạng mã hóa người nói (còn gọi là Speaker Encoder) mã hóa đặc trưng giọng nói thành vector biểu diễn đặc trưng giọng nói (speaker-embedding) dựa trên tính toán một vector chiều cố định từ tín hiệu tiếng nói; 2) Bộ tổng hợp chuỗi từ chuỗi (Seq2seq) dựa trên việc dự đoán phổ Mel từ một chuỗi hình vị hoặc âm vị đầu vào kết hợp với vector biểu diễn đặc trưng nói có điều kiện; 3) Bộ giải mã tái tạo các biểu diễn đặc trưng ẩn đầu vào thành thành phổ Mel; và 4) Bộ phát âm cho phép chuyển đổi phổ Mel thành sóng âm theo miền thời gian.



Hình 11: Sơ đồ khối hệ thống tổng hợp thích nghi cơ sở dựa trên DNN

1.3.3.1. Bộ mã hóa (Encoder)

Bộ mã hóa là một thành phần cho phép mã hóa các chuỗi có chiều dài không cố định thành các vector biểu diễn có chiều cố định, cụ thể là mã hóa chuỗi văn bản hoặc âm vị đầu vào thành các vector ẩn biểu diễn thông tin ngữ âm ở mức khung gọi là *Text embedding*. Ngoài ra, bộ mã hóa người nói cho phép mã hóa các biểu diễn đặc trưng độc lập người nói đó thành các biểu diễn ẩn là vector đặc trưng giọng nói.

1.3.3.2. Vector biểu diễn đặc trưng giọng nói (Speaker-embedding)

Cách tiếp cận khi xây dựng hệ thống tổng hợp đa người nói dựa trên thích nghi hiện nay sử dụng các biểu diễn đặc trưng của người nói để huấn luyện và điều chỉnh mô hình thích nghi. Để thực hiện được việc đó, các hệ thống xử lý tiếng nói trước hết phải chuyển đổi từng đoạn âm thanh có độ dài thay đổi thành một vectơ có độ dài cố định đại diện cho danh tính của người nói, bao gồm các thông tin như cao độ, tông giọng, ngữ điệu, trọng âm của một người nói, được gọi là vector biểu diễn đặc trưng giọng nói, sau đó thực hiện phân cụm dựa trên các

vectơ này. Vector biểu diễn đặc trưng nói cũng được sử dụng rộng rãi trong các tác vụ xử lý ngôn ngữ tự nhiên như: Định danh người nói *(Speaker identity)*, phân đoạn người nói (*Speaker diarization*), thích nghi người nói (*Speaker adaptation*), tổng hợp tiếng nói (*Speech synthesis*).

1.3.3.3. Bộ giải mã (Decoder)

Khi bộ mã hóa mã hóa các chuỗi tiếng nói có chiều dài không cố định thành các vector có chiều dài cố định thì bộ giải mã *Decoder* sẽ giải mã các biểu diễn vector ẩn đó và dự đoán thành các biểu diễn âm học. Cụ thể là bộ giải mã sẽ dự đoán thành đặc trưng âm học tương ứng hoặc thành tiếng nói theo miền thời gian với đặc trưng độc lập người nói. Khi tổng hợp, đặc trưng độc lập của cùng một người nói được mã hóa bởi bộ mã hóa từ chuỗi tiếng nói đầu vào và được dùng để điều khiển giải mã sinh ra các đặc trưng hoặc tiếng nói. Bộ giải mã được huấn luyện để thực hiện ánh xạ giữa các đặc trưng độc lập người nói và các đặc trưng âm học hoặc giọng nói tương ứng của cùng một người nói.

1.3.3.4. Bộ phát âm (Vocoder)

Bộ phát âm chuyển các biểu diễn đặc trưng âm thanh như phổ Mel thành sóng âm thanh có thể nghe thấy được. Sự phát triển của bộ phát âm giọng nói có thể được chia thành hai giai đoạn: Bộ mã hóa giọng nói được sử dụng trong tổng hợp giọng nói tham số thống kê (SPSS) [23] [24] và bộ mã hóa dựa trên mạng nơ-ron [10] [56]. Một số bộ phát âm phổ biến trong SPSS bao gồm STRAIGHT [23] và WORLD [24]. Lấy bộ mã hóa WORLD làm ví dụ, nó bao gồm các bước phân tích phát âm và tổng hợp bộ phát âm. Trong phân tích phát âm giọng nói, nó phân tích giọng nói và nhận các đặc trưng âm thanh như hệ số mel-cepstral, tần số dải và F0. Trong quá trình tổng hợp bộ phát âm, nó tạo ra dạng sóng giọng nói từ các đặc trưng âm thanh này.

Các bộ phát âm dựa trên nơ-ron ban đầu như WaveNet [10], WaveRNN [57] trực tiếp nhận đầu vào là một chuỗi các đặc trưng ngôn ngữ và ngữ âm (chứa thông tin về âm vị, âm tiết, từ, v.v.), sau đó đưa qua mô hình xác suất và tự hồi quy để dự đoán ra mỗi mẫu âm thanh có điều kiện dựa trên các dự đoán trước đó để tạo dạng sóng. WaveNet nguyên bản là loại mạng non-conditional, tức là không nhận đầu vào nào mà chỉ sinh ngẫu nhiên các đoạn âm thanh vô nghĩa. Tuy nhiên trong Tacotron2, nhóm nghiên cứu sử dụng một phiên bản sửa đổi của WaveNet

để tạo ra các mẫu dạng sóng theo miền thời gian có điều kiện trên phổ Mel. Tức là WaveNet được thiết kế để nhận đầu vào là dạng phổ Mel, đầu ra là sóng âm tương ứng. Sau đó, Prenger [56] và Kim [58] lấy phổ Mel làm đầu vào và tạo dạng sóng thay vì đầu vào là các đặc trưng ngữ âm. Vì dạng sóng tiếng nói rất dài nên việc tạo dạng sóng tự hồi quy mất nhiều thời gian suy diễn (do kiến trúc tự hồi quy đóng vai trò nút thắt cổ chai chính khi tạo sóng thời gian thực), do vậy các kiến trúc tự hồi quy như WaveNet và Char2Wav được thay thế hoặc cải tiến để tăng tốc độ quy diễn. Các mô hình tổng quát như mô hình sinh dựa trên dòng chảy (Flow) [59], mạng đối nghịch (GAN) [60], bộ tự mã hóa (VAE) [61] và mô hình xác suất khuếch tán khử nhiễu (DDPM) [62] được sử dụng trong quá trình tạo dạng sóng. Như vậy, có thể chia bộ phát âm mạng nơ-ron thành các loại khác nhau: 1) Bộ phát âm tự hồi quy, 2) Bộ phát âm dựa trên dòng chảy, 3) Bộ phát âm dựa trên GAN, 4) Bộ phát âm dựa trên VAE và 5) Bộ phát âm dựa trên mô hình khuếch tán (Diffusion model).

1.3.4. Đánh giá chất lượng hệ thống tổng hợp thích nghi

Để đánh giá hiệu năng của hệ thống tổng hợp tiếng nói hoặc hệ thống thích nghi giọng nói cần sử dụng kết hợp giữa cách đánh giá chủ quan dựa vào cảm tính (MOS, SIM) và đánh giá khách quan thông qua đo các đại lượng vật lý (MCD, WER) như sau:

1.3.4.1. MOS (Mean Opinion Score)

Sử dụng thang đo ý kiến trung bình (MOS) được sử dụng để đo chất lượng của hệ thống TTS thông qua đánh giá một cách định tính trên cơ sở các bài kiểm tra khả năng nghe hợp lệ và đáng tin cậy. MOS là thước đo được khuyến nghị sử dụng để đánh giá chất lượng giọng nói tổng hợp (Khuyến nghị ITU-T P.85) [63]. Bản thân MOS được tính trung bình từ một số thành phần riêng lẻ của từng phiên đo. Thang đánh giá được sử dụng phổ biến nhất là thang đo Xếp hạng danh mục tuyệt đối (ACR), gồm năm cấp độ: *5. Xuất sắc (như con người); 4. Tốt (gần giống con người); 3: Trung bình (hơi máy); 2. Kém (như máy); 1. Tệ*.

1.3.4.2. MCD (Mel-Cepstral Distortion)

Thang đo sự biến dạng mel cepstral (MCD) [64] là một phép đo lường mức độ khác nhau của hai chuỗi mel cepstral. MCD giữa chuỗi mel cepstral tổng hợp và tự nhiên càng nhỏ thì giọng nói tổng hợp càng gần với việc tái tạo giọng nói tự

nhiên. Để tính toán MCD, trích xuất hệ số cepstral từ tín hiệu âm thanh và tính toán khoảng cách giữa các hệ số này trong hai tín hiệu âm thanh. MCD được tính bằng cách lấy trung bình cộng của khoảng cách Euclide giữa các hệ số cepstral trong hai tín hiệu giọng nói trên một khoảng thời gian nhất định. Nó hoàn toàn không phải là một thước đo hoàn hảo để đánh giá chất lượng của bài phát biểu tổng hợp nhưng thường là một chỉ số hữu ích khi kết hợp với các thước đo khác.

1.3.4.3. SIM (Similarity)

Thang đo sự tương đồng của các giọng nói (SIM) được sử dụng để đánh giá sự giống nhau giữa các hệ thống thích nghi giọng nói được đề xuất bởi Voice Conversion Challenge (VCC) [65]. Người nghe sẽ được nghe các cặp câu âm thanh được tạo bởi một hệ thống tổng hợp tiếng nói và âm thanh gốc để đánh giá xem các câu âm thanh này có phải của cùng một người nói hay không. Người nghe có bốn (4) lựa chọn để cho điểm*: 4. Chắc chắn là giống nhau; 3. Có thể giống nhau; 2. Có thể khác nhau; 1. Chắc chắn là khác nhau*. Điểm càng cao, có ý nghĩa càng có nhiều điểm tương đồng giữa giọng tổng hợp và giọng người nói.

1.3.4.4. WER (Word-error-rate)

Chỉ số lỗi từ (WER) được sử dụng để đo lường tỷ lệ phần trăm các từ bị nhận diện sai trong văn bản nhận dạng từ âm thanh tổng hợp so với văn bản nhận dạng từ âm thanh gốc. Chỉ số này cung cấp thêm thông tin đánh giá chất lượng tổng hợp thông qua khả năng nhận dạng tiếng nói thành văn bản của các hệ thống ASR sẵn có [66].

1.3.4.5. Phân tích phương sai (ANOVA)

Phân tích phương sai (ANOVA) trong tổng hợp tiếng nói là một phương pháp thống kê được sử dụng để kiểm tra sự khác biệt giữa các nhóm trong việc đánh giá chất lượng hoặc hiệu suất của các hệ thống tổng hợp tiếng nói. Trong ngữ cảnh này, các nhóm có thể là các phương pháp hoặc mô hình khác nhau của tổng hợp tiếng nói. Mục tiêu chính của ANOVA là xác định xem có sự khác biệt ý nghĩa nào đó giữa các phương pháp hay mô hình tổng hợp tiếng nói không (ví dụ phân tích ANOVA trong đánh giá MOS hoặc SIM), và nếu có, thì nhóm nào hoặc phương pháp nào gây ra sự khác biệt này [65]. Quy trình thực hiện kiểm định ANOVA trong tổng hợp tiếng nói bao gồm: Chuẩn bị dữ liệu ; Xác định giả thuyết; Thực hiện phân tích ANOVA; Kiểm định hậu nghiệm; Phân tích Kết Quả.

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

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

Các kỹ thuật tổng hợp tiếng nói và tổng hợp dựa trên thích nghi đã là kỹ thuật chính tồn tại trong cả thập niên từ những năm 1990. Tiếp đó, tổng hợp tiếng nói sử dụng DNN trở lên chiếm ưu thế với việc ứng dụng cho mô hình âm học, mô hình trường độ, trích xuất đặc trưng và phân tích văn bản đã có nhiều phân tích theo nhiều nhóm khác nhau. Các nghiên cứu đã chỉ ra rằng, hệ thống DNN cơ sở đã vượt qua các hệ thống HMM tiêu chuẩn về chất lượng tiếng nói tổng hợp thông qua hướng tiếp cận học sâu, vì DNN có thể học các ánh xạ phức tạp từ các đặc trưng ngôn ngữ đến các đặc trưng âm học. Một số nghiên cứu độc lập đã chỉ ra rằng, DNN có thể tạo ra giọng nói tổng hợp tự nhiên hơn so với tổng hợp giọng nói dựa trên HMM thông thường cho một người nói trong các điều kiện huấn luyện khác nhau [9]. Tuy nhiên, chỉ một số các nghiên cứu giải quyết câu hỏi liệu tổng hợp tiếng nói dựa trên DNN có thể cung cấp các kỹ thuật thích nghi có tính linh hoạt tương tự như tổng hợp tiếng nói dựa trên HMM hay không, mặc dù đã có công trình thành công trong lĩnh vực này trong bối cảnh nhận dạng giọng nói dựa trên DNN [67]. Một nghiên cứu sơ bộ đã được thực hiện về sự thích nghi của người nói trong tổng hợp dựa trên DNN [68], nhưng chỉ có một sự chuyển đổi đặc trưng được sử dụng để sửa đổi đầu ra của DNN. Các giai đoạn tiếp sau là sự phát triển của các kỹ thuật tổng hợp tiếng nói từ văn bản dựa trên thích nghi (hay còn gọi là Thích nghi TTS hoặc tổng hợp thích nghi – TTS Adaptation).

Các hướng nghiên cứu chính có thể mô tả như sau:

- **Thích nghi ít dữ liệu Few-shot TTS**. Một số nghiên cứu đã xây dựng mô hình tổng hợp thích nghi ít dữ liệu hay còn gọi là thích nghi Few-shot TTS bằng cách chỉ sử dụng một vài cặp dữ liệu văn bản và âm thanh trong khoảng từ vài giây đến vài phút để huấn luyện mô hình. 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 giọng nói. Nghiên cứu [69] đưa ra phương pháp thích nghi bằng cách huấn luyện trước mô hình trên một bộ dữ liệu lớn và sau đó tinh chỉnh mô hình với một ít dữ liệu nhỏ. Nghiên

cứu [70] chỉ ra một vài kiểu vector biểu diễn đặc trưng giọng nói khác nhau để thích nghi Few-shot TTS: 1) Biểu diễn đặc trưng giọng nói từ mô hình huấn luyện trước dùng d-vector, x-vector và biểu diễn chuyển đổi giọng nói (voice conversion-VC); 2) Biểu diễn đặc trưng giọng nói thông qua tối ưu hóa chung sử dụng vector nhúng mã hóa vị trí để gán ID cho mỗi người nói khác nhau hoặc vector mã hóa thông tin phong cách nói toàn cục (Global style token - GST). Tuy nhiên, bởi vì cả d-vector và x-vector đều được huấn luyện trước cho nhiệm vụ nhận dạng, các mô hình chỉ nhằm mục đích phân biệt cách nói của những người nói khác nhau, vì vậy chúng có thể bỏ qua những thông tin không giúp phân biệt những người nói khác nhau. Điều này có thể không tốt cho mô hình TTS. Một vector đại diện giọng nói lý tưởng cho các mô hình TTS nên mô hình hóa mọi đặc trưng của người nói, nhưng không nhất thiết phải có ranh giới rõ ràng giữa từng người nói. Biểu diễn VC chứa thông tin cần thiết cho việc tạo tiếng nói, đây có thể là lý do khiến nó vượt trội so với các vector biểu diễn huấn luyện trước (d-vector, x-vector) trong các đánh giá khách quan và chủ quan. Thực nghiệm đã chứng minh sự kết hợp các biểu diễn của người nói được huấn luyện trước với các biểu diễn của người nói được tối ưu hóa chung mang lại kết quả tốt hơn so với sử dụng bất kỳ biểu diễn đơn lẻ nào. Nghiên cứu [13] [45] so sánh chất lượng giọng nói với các lượng dữ liệu thích nghi khác nhau và thấy rằng chất lượng giọng nói được cải thiện nhanh chóng với việc tăng dần dữ liệu thích nghi khi kích thước dữ liệu nhỏ (dưới 20 câu) nhưng cải thiện chậm với hàng chục câu thích nghi.

- **Thích nghi ít tham số**. Để hỗ trợ nhiều người dùng, các tham số thích nghi của mỗi giọng nói đích cần đủ nhỏ để giảm mức sử dụng bộ nhớ trong khi vẫn duy trì chất lượng tổng hợp cao. Ví dụ, nếu mỗi người dùng/giọng nói sử dụng tham số 100MB thì tổng dung lượng bộ nhớ sẽ lên đến 100PB cho một triệu người dùng. Một số nghiên cứu đề xuất giảm tham số thích nghi xuống mức thấp nhất trong khi vẫn duy trì chất lượng thích nghi.

Các hệ thống tổng hợp đa người nói thông thường yêu cầu một lượng dữ liệu đáng kể để mô hình hóa những người nói đã được quan sát trong quá trình huấn

luyện. Tuy nhiên, nhiều ứng dụng được cá nhân hóa chỉ có thể cung cấp một lượng rất nhỏ dữ liệu mẫu (Ví dụ: khôi phục khả năng giao tiếp cho những người bị mất giọng nói, khôi phục giọng của các vĩ nhân, khôi phục giọng của những người quá cố …). Những nhu cầu như vậy đòi hỏi một kỹ thuật thích nghi cho phép sao chép giọng nói chỉ cần một vài mẫu tham khảo thông qua huấn luyện, được gọi là thích nghi Few-shot TTS. Như vậy, thích nghi Few-shot TTS là cách tổng hợp một giọng mới sao cho giống giọng đích với một lượng dữ liệu nhỏ có huấn luyện. AdaSpeech [13] đề xuất chuẩn hóa lớp có điều kiện để tạo các tham số, thang đo và độ lệch trong chuẩn hóa lớp từ các vector biểu diễn đặc trưng giọng nói dựa trên việc tạo tham số theo ngữ cảnh [71] và chỉ tinh chỉnh các tham số liên quan đến chuẩn hóa lớp có điều kiện và vector biểu diễn đặc trưng giọng nói để đạt được chất lượng thích nghi. Moss và cộng sự [72] 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. Phương pháp này đạt được mục tiêu tổng hợp giọng nói của một giọng nói cụ thể chỉ với một số lượng nhỏ các mẫu giọng nói.

- **Thích nghi Zero-shot TTS**. Zero-shot TTS là một hướng tiếp cận mới để thích nghi giọng mới với chỉ một câu hoặc vài giây mẫu mà không cần huấn luyện bổ sung. Kỹ thuật này cho phép thích nghi giọng mới mà không cần huấn luyện lại, hơn nữa lượng dữ liệu cần để huấn luyện cực nhỏ (chỉ cần một câu hoặc vài giây dữ liệu giọng đích) [73]. Zero-shot TTS trong tổng hợp tiếng nói là các kỹ thuật hướng đến huấn luyện một mô hình cho phép tạo các giọng mới dưới điều kiện các giọng này chưa từng xuất hiện trong quá trình huấn luyện hoặc không xuất hiện trong suốt quá trình học giám sát [74]. Các nghiên cứu này mở ra một số ứng dụng hữu ích như hệ thống speaker thông minh (có tài nguyên nhỏ) có thể kể chuyện hoặc giao tiếp bằng giọng nói riêng; học giọng mới tại chỗ mà không cần huấn luyện lại; các hệ thống thuyết minh linh hoạt theo giọng của speaker được cung cấp ngay tại chỗ. Các mẫu Zero-shot TTS đa người nói thường sử dụng vector biểu diễn đặc trưng giọng nói có thể dễ dàng thích nghi với giọng nói mới, cho phép chúng tạo giọng nói bằng giọng nói của người nói mới với lượng rất ít dữ liệu hơn nhiều so với các phương pháp thích nghi bằng tinh chỉnh: Hiệu quả

của những mô hình này đầy hứa hẹn về chất lượng của giọng tổng hợp và khả năng khái quát cho những người nói mới. Nhìn chung, Zero-shot TTS là một lĩnh vực thú vị và đang phát triển nhanh chóng, đồng thời có khả năng tác động đáng kể đến cách các hệ thống TTS được xây dựng và sử dụng trong tương lai.

Một số nghiên cứu [75] [12] [76] tiến hành thích nghi Zero-shot TTS, sử dụng vector biểu diễn đặc trưng giọng nói để trích xuất các biểu diễn đặc trưng giọng nói (ví dụ như d-vector, x-vector và vector LDE) từ các âm thanh mẫu. Kịch bản này khá hấp dẫn vì không cần dữ liệu và huấn luyện lại các tham số thích nghi. Tuy nhiên, thử nghiệm cho thấy, chất lượng thích nghi chưa đủ tốt do các vector biểu diễn đặc trưng giọng nói được đề xuất biểu diễn đặc trưng giọng nói chưa hiệu quả, điều này càng thể hiện rõ khi giọng nói đích (target-speaker) rất khác với giọng nói nguồn (source-speaker).

1.4.2. Một số nghiên cứu hiện này về tổng hợp tiếng Việt

1.4.2.1. Các hệ thống tổng hợp tiếng nói tiếng Việt sử dụng kỹ thuật ghép nối và lựa chọn đơn vị âm

"Hoa súng" là một hệ thống TTS Tiếng Việt sử dụng đơn vị âm dip-phone và bán âm tiết. Mô hình CART được chọn để xây dựng mẫu trường độ và thuật toán miền thời gian PSOLA (TD-PSOLA) được sử dụng để điều chỉnh cao độ và trường độ của các đơn vị tiếng nói. Tuy nhiên, hạn chế của Hoa Súng là không bao hàm được hết các âm tiết mới nhất hoặc các từ mượn, kho ngữ liệu nhỏ không thể tổng hợp được các âm tiết bao gồm các bán âm tiết, không tái tạo được những thay đổi quan trọng về F0 do hiện tượng đóng âm… [77]

Nghiên cứu của MICA (Đại học Bách Khoa Hà Nội) về tổng hợp tiếng nói dựa trên kỹ thuật ghép nối các đơn vị âm thanh không đồng nhất [78]. Hệ thống tổng hợp tiếng nói phương Nam (VoS) phát triển bởi Vũ Hải Quân và nhóm phát triển tại AILab thuộc Đại học Công nghệ - Đại học Quốc gia Hồ Chí Minh [79] sử dụng kỹ thuật tổng hợp lựa chọn đơn vị không đồng nhất với các đơn vị tiếng nói là âm tiết trở lên, giọng tổng hợp là của phương ngữ Nam Bộ. Chất lượng tổng hợp tốt hơn trong miền giới hạn (Ví dụ: bình luận bóng đá). Tuy nhiên, quá trình chuyển tiếp giữa các phần nối của sóng âm vẫn không mượt mà, trơn tru.

1.4.2.2. Các hệ thống tổng hợp tiếng nói tiếng Việt dựa trên tham số thống kê (SPSS)

Đối với phần tổng hợp tiếng nói tiếng Việt dựa trên tham số thống kê HMM, chỉ có hai nhóm chính sau: 1) Từ Viện Công nghệ Thông tin (IoIT, thuộc Viện Khoa học và Công nghệ Việt Nam) [80] và 2) Từ trường Đại học Vân Nam, Trung Quốc [81]. Cả hai nhóm đều dùng kiến trúc lõi của HTS để phát triển hệ thống TTS cho tiếng Việt giọng Hà Nội. Theo đánh giá chủ quan, công bố đầu tiên về tổng hợp giọng nói tiếng Việt dựa trên HMM là công trình của Viện công nghệ thông tin [80]. Hệ thống này chỉ áp dụng HTS cho tiếng Việt với kho ngữ liệu huấn luyện bao gồm 3000 câu giàu ngữ âm, được gán nhãn bán tự động ở cấp độ âm. Công trình của trường Đại học Vân Nam [81] cũng chỉ đơn giản là áp dụng kỹ thuật tổng hợp dựa trên HMM cho tiếng Việt bằng cách sử dụng bộ tổng hợp STRAIGHT[2]. Khoảng 600 câu có nhãn được sử dụng để huấn luyện mô hình HMM. Một đánh giá sơ bộ (10 người) đã được thực hiện với cùng một kết luận về giọng nói tổng hợp như công trình của [80]. Các nghiên cứu tiếp theo của Viện CNTT [82] tập trung vào tầm quan trọng của các đặc trưng ngữ điệu. Các đặc trưng ngữ điệu bổ sung từ tiếng Anh sử dụng mô hình ToBI đã được sử dụng trong các nghiên cứu này, bao gồm: 1) Ngữ điệu cuối cụm từ và 2) Trọng âm và cao độ. Từ nửa cuối năm 2013, Viện nghiên cứu Quốc tế MICA (Đại học Bách khoa Hà Nội) và Phòng thí nghiệm Trí tuệ nhân tạo AILab (Đại học Khoa học tự nhiên TP HCM) cũng bắt đầu có những nghiên cứu, phát triển hệ thống tổng hợp tiếng Việt tham số thống kê dựa trên HMM.

1.4.2.3. Các hệ thống TTS Tiếng Việt dựa trên DNN

Mạng nơ-ron sâu (Deep neural network - DNN), với khả năng giải quyết vấn đề HMM, đã trở nên phổ biến không chỉ trong việc tổng hợp giọng nói mà còn trong nhiều vấn đề phụ thuộc vào ngữ cảnh khác như nhận dạng giọng nói tự động [83]. Với sự phát triển của học sâu, TTS dựa trên mạng nơ-ron được đề xuất với việc sử dụng mạng nơ-ron học sâu làm mô hình xương sống để tổng hợp giọng nói. Một số mô hình nơ-ron ban đầu được áp dụng trong tổng hợp tiếng nói dựa

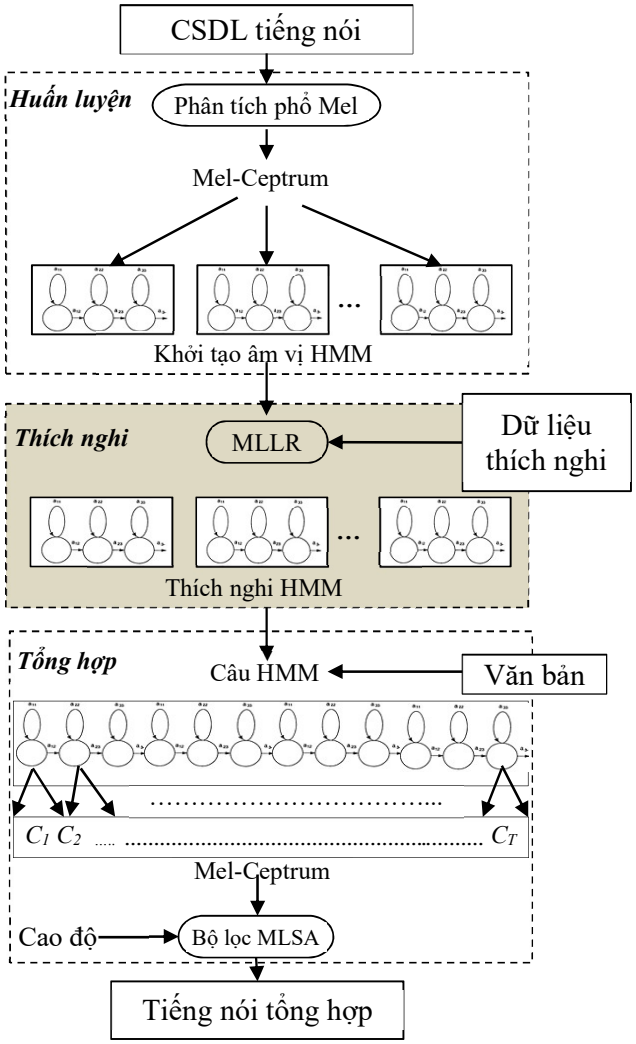
[2] http://web.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html

trên tham số thống kê SPSS để thay thế HMM trong mô hình âm học. Sau đó, WaveNet được đề xuất để trực tiếp tạo ra dạng sóng âm từ các đặc trưng ngôn ngữ và nó có thể được coi mô hình TTS nơ-ron hiện đại đầu tiên [10].

Viettel đã giới thiệu hệ thống TTS tiếng Việt dựa trên DNN đầu tiên đạt MOS (Mean Opinion Score) vượt trội so với các hệ thống tổng hợp tiếng nói tiếng Việt dựa trên HMM về độ dễ hiểu và tự nhiên (so với các hệ thống TTS tiếng Việt khác lúc đó như MICA và VAIS và kết quả được đánh giá trong hội thảo quốc tế về Xử lý tiếng nói và văn bản tiếng Việt - VLSP 2018) [84]. Các hệ thống TTS cho tiếng Việt sau đó tiếp tục được cải tiến và công bố với các công nghệ cập nhật [85].

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

Một số nghiên cứu thích nghi cho tiếng Việt trước đây dựa trên hướng tiếp cận chính là tham số thống kê HMM [7] [8].



Hình 12: Sơ đồ khối tổng hợp tiếng nói dựa trên thích nghi bằng HMM [7]

Để chuyển đổi các đặc trưng giọng nói của lời nói tổng hợp thành các đặc trưng của người nói mục tiêu cần điều chỉnh các HMM cho các âm vị ban đầu thành các HMM của người nói mục tiêu. Với cách như vậy, các tham số xác suất đầu ra của các mô hình Markov ẩn được sửa đổi để phản ánh các đặc trưng giọng nói của người nói mục tiêu. Kết quả là các thông số giọng nói được tạo ra trở nên gần hơn với các thông số của người nói mục tiêu và các đặc trưng giọng nói của lời nói tổng hợp cũng trở nên gần hơn với người nói mục tiêu. Các kỹ thuật thích nghi chính thường được sử dụng như MAP, VFS, MLLR hoặc kết hợp chúng với nhau.

Thuật toán cực đại hậu nghiệm (Maximum A Posteriori - MAP) được sử dụng để cập nhật các tham số của từng phân bố xác suất đầu ra trạng thái tùy thuộc vào dữ liệu thích nghi [86]. Tuy nhiên, để thuật toán MAP thích hợp thì cần có đầy đủ thông tin tiên nghiệm về các tham số mô hình. Trong cách tiếp cận này, thông tin tiên nghiệm thường được sử dụng là các HMM độc lập người nói. Nếu lượng dữ liệu thích nghi nhỏ hoặc rải rác thì các ước lượng MAP không đáng tin cậy. Để khắc phục vấn đề này, thuật toán VFS được đề xuất kết hợp với ước lượng MAP để nội suy các tham số mới cho tập phân bố không có trong tập huấn luyện và làm mịn các véctơ chuyển đổi của các phân bố ước lượng MAP đã huấn luyện [87]. Các nghiên cứu thích nghi sử dụng thuật toán MAP/VFS (Vector Field Smoothing) yêu cầu nhiều tham số để kiểm soát sự thích nghi, ví dụ như trọng số cho hệ số làm mịn của mật độ tiên nghiệm cho VFS và số PDF để tính các vector chuyển đổi. Do đó, không dễ để xác định một tập các tham số cho phép cung cấp hiệu suất tốt nhất trong tổng hợp giọng nói dùng cho HMM.

Thuật toán hồi quy tuyến tính ước lượng khả năng cực đại (MLLR) dựa trên hồi quy tuyến tính ước lượng khả năng cực đại (maximum likelihood) và chỉ yêu cầu một thông số duy nhất đại diện cho số ma trận hồi quy. Thuật toán MLLR được sử dụng để tìm phương thức biến đổi các tham số đặc trưng, đây là thuật toán đầu tiên được áp dụng cho hệ thống TTS dựa trên HMM [88]. MLLR tìm kiếm phép ánh xạ tuyến tính của HMM để cực đại hóa xác suất của dữ liệu thích nghi từ người nói đích. Để buộc các mô hình tương tự nhau cho

quá trình thích nghi, hồi quy hoặc phân cụm dựa trên cây quyết định được sử dụng để chia sẻ phép biến đổi giữa các phân bố của mỗi cụm ngữ cảnh. Chia sẻ các phép biến đổi giữa nhiều phân bố giúp giảm khối lượng dữ liệu cần thiết cho quá trình thích nghi. Do đó, thuật toán thích nghi dựa trên MLLR được ưa chuộng và sử dụng phổ biến hơn thuật toán MAP-VFS trong trường hợp thiếu dữ liệu đích.

Nghiên cứu [7] đề xuất 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 MLLR kết hợp với thuật toán cực đại hậu nghiệm (MAP) thay thế cho các thuật toán đơn thuần chỉ sử dụng MAP hoặc MLLR; hoặc thay thế cho các thuật toán kết hợp MAP với thuật toán làm mịn trường véc tơ (VFS). Nghiên cứu trên cũng đã chỉ 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 [89]. Nghiên cứu [8] đề xuất mô hình lai ghép giữa mô hình Markov ẩn và phối xác suất đa không gian (MSD-HMM). Chi tiết các mô đun của hệ thống trong Hình 12 như sau:

- **Huấn luyện**: Đầu tiên, trong giai đoạn huấn luyện, hệ số mel-cepstral được thu thập từ cơ sở dữ liệu giọng nói bằng cách phân tích mel-cepstral [90] [91]. Các đặc trưng động như hệ số delta và delta-delta cũng được tính toán từ hệ số mel-cepstral. Sau đó, đối với mỗi âm vị, HMM độc lập với người nói còn gọi là HMM âm vị khởi tạo, được huấn luyện bằng cách sử dụng hệ số mel-cepstral, delta và delta-delta của chúng.
- **Thích nghi**: Trong giai đoạn thích nghi, các vectơ đặc trưng được tính toán từ mẫu dữ liệu thích nghi đã cho. Sau đó, các HMM khởi tạo được chuyển thành HMM của giọng nói đích bằng cách áp dụng kỹ thuật thích nghi giọng nói. Ở mô hình này, có thể biến đổi các thông số phổ âm thanh bằng cách sử dụng một số kỹ thuật thích nghi giọng nói như MLLR [92] hoặc MAP-VFS. Đây là một thuật toán kết hợp tối đa thích nghi posteriori (MAP) và làm mịn trường vectơ (VFS). Sau đó, để mô hình hóa và thích nghi đồng thời các thông số kích thích

của tiếng nói cũng như các thông số phổ âm thanh, phân bố xác suất đa không gian (MSD) HMM [93] và thuật toán thích nghi MLLR [88] được sử dụng.

- **Tổng hợp**: Trong giai đoạn tổng hợp, một văn bản cho trước tùy ý cần tổng hợp sẽ được biến đổi thành một chuỗi âm vị. Một câu HMM đại diện cho toàn bộ văn bản được tổng hợp bằng cách nối các âm vị HMM theo trình tự âm vị. Từ câu HMM thu được, chuỗi tham số mel-cepstral được tạo ra bằng cách sử dụng thuật toán [94] được mô tả ngắn gọn trong phần tiếp theo. Bằng cách sử dụng bộ lọc MLSA (Mel Log Spectrum Approximation) [22], lời nói được tổng hợp từ các hệ số mel-cepstral được tạo ra.

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 [84] [95]; 2) Các nghiên cứu hiện nay [96] chỉ ra rằng các kỹ thuật tổng hợp thích nghi dựa trên HMM cho chất lượng thấp hơn so với các kỹ thuật DNN và đã được chỉ rõ trong mục 1.2.3. 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.

1.4.4. Hướng nghiên cứu chính của luận án

Luận án tập trung vào nghiên cứu phương pháp tổng hợp tiếng nói tiếng Việt dựa trên thích nghi với hai hướng tiếp cận chính là: 1) Tiếp cận theo hướng thích nghi ít dữ liệu có huấn luyện lại mô hình (Few-shot TTS) và 2) Tiếp cận theo hướng giảm tối đa dữ liệu mẫu và không phải huấn luyện lại mô hình (Zero-shot TTS). Mỗi hướng tiếp cận này sẽ có ưu, nhược điểm khác nhau, được phân tích chi tiết theo bảng dưới đây:

Bảng 2: So sánh ưu nhược điểm của hai phương pháp tiếp cận tổng hợp dựa trên thích nghi

	Tiếp cận theo hướng Few-shot TTS	**Tiếp cận theo hướng Zero-shot TTS**
Định nghĩa	• Thích nghi với ít dữ liệu và huấn luyện lại mô hình để tạo giọng mới	• Thích nghi với ít dữ liệu và không phải huấn luyện lại mô hình để tạo giọng mới

Phương pháp	• Dựa vào tinh chỉnh, dựa vào vector đặc trưng và DNN	• Dựa vector đặc trưng, DNN và mô hình khuếch tán khử nhiễu
Ưu điểm	• Chất lượng và độ tương đồng cao • Phù hợp các ứng dụng đòi hỏi giọng mới đảm bảo chất lượng nhưng không cần quá nhiều mẫu	• Chỉ cần lượng mẫu rất nhỏ để thích nghi (vài giây) • Không cần huấn luyện lại nên thời gian tổng hợp giọng mới tức thì (vài giây) • Phù hợp cho ứng dụng tạo giọng mới tức thì
Nhược điểm	• Cần số lượng mẫu lớn hơn zero-shot để thích nghi (vài phút) • Cần thời gian lớn để huấn luyện và tốn chi phí triển khai một giọng mới (nhiều chục giờ)	• Chất lượng và độ tương đồng kém hơn (ở mức chấp nhận được)

1.5. 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 giọng 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 có chứa thành phần ngữ điệu nên tổng hợp tiếng nói cũng như thích nghi giọng 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;

4) Cần có ứng dụng để đánh giá tính khả thi của các mô hình thích nghi giọng nói tiếng Việt.

Từ các vấn đề thực tế trên, luận án tập trung nghiên cứu một số nội dung chính như sau và các nội dung đó nhằm trả lời cho các câu hỏi nghiên cứu được đặt ra trong trang 2 phần Mở đầu:

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 giọng tiếng Việt nhằm cá nhân hóa giọng nói tổng hợp, cụ thể là nhân bản giọng nói với dữ liệu của đơn người nói 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 GIỌNG NÓI

Mục đích của hệ thống thích nghi cho tổng hợp tiếng đó là tổng hợp được tiếng nói với mẫu thích nghi nhỏ hoặc ít tài nguyên. Một trong những hợp phần quan trọng nhất trong hệ thống tổng hợp dựa trên thích nghi là hệ thống tổng hợp tiếng nói. Ở một khía cạnh khác, có thể coi thích nghi giọng nói chính là bài toán con trong lĩnh vực tổng hợp tiếng nói. Do vậy, 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* [97].

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

Tổng hợp tiếng nói tiếng Việt đang được chú ý nhiều hơn trong thời gian gần đây do sự cần thiết trong các ứng dụng thực tế. Các giải pháp về báo nói, sách nói, phát thanh viên ảo, tổng đài trả lời tự động ngày càng phổ biến. Thống kê đến thời điểm hiện tại hầu hết các báo điện tử lớn tại Việt Nam đều tích hợp TTS để tự động đọc nội dung các bài viết với ngữ điệu và chất lượng khá tự nhiên (Ví dụ: dantri.vn, tienphong.vn, laodong.vn, ...).

Theo thống kê của VLSP, trong bốn năm tổ chức đánh giá, năm 2018 các mô hình DNN đạt ưu thế trong VLSP [98], 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 bộ phát âm phổ biến như Waveglow hoặc HifiGAN [97] [99]. Cũng theo báo cáo 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 [85].

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 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 [97].

Tiếng Việt là ngôn ngữ đơn âm tiết và có thanh điệu. Để tổng hợp tiếng Việt chất lượng tốt, việc đảm bảo chất lượng của thanh điệu tổng hợp sao cho càng gần với thanh điệu tự nhiên là rất quan trọng. Trong xử lý tiếng nói, việc xây dựng một bộ cơ sở dữ liệu chất lượng phục vụ nghiên cứu tổng hợp và huấn luyện luôn là yếu tố quan trọng. Với các ngôn ngữ giàu tài nguyên như tiếng Anh, tiếng Nhật qua quá trình phát triển lâu đời đã có những bộ cơ sở dữ liệu lớn, đa dạng. Ở Việt Nam, các công trình nghiên cứu xử lý tiếng nói đã có từ những năm đầu thế kỷ 20. Đến nay, nhiều bộ dữ liệu tiếng nói đã được công bố rộng rãi như VOV (nguồn phát thanh vô tuyến), MICA VNSpeechCorpus, AIlab VOS (Voice of Southern Vietnam), AIlab VIVOS, VAIS-1000, VLSP. Tuy nhiên, các bộ cơ sở dữ liệu tiếng Việt phục vụ nghiên cứu tổng hợp tiếng nói vẫn còn nhiều hạn chế, chưa có bộ dữ liệu đảm bảo chất lượng có kích thước trên 10 giờ dành cho TTS được công bố [97]. Ví dụ kho văn bản đề xuất trong MICA VNSpeechCorpus [100] được thiết kế tốt và chứa bài phát biểu chất lượng tốt. Tuy nhiên, mặc dù tổng kích thước của ngữ liệu lớn, nhưng số lượng câu ngắn thích hợp cho sự thích nghi trong tổng hợp tiếng nói là khá nhỏ. Những hạn chế có thể chỉ ra là bộ cơ sở dữ liệu thu âm nhỏ, chất lượng chưa cao, không đa dạng (một số bộ thu thập từ các nguồn dữ

liệu đa phương tiện sẵn có, một số bộ dữ liệu xây dựng từ một vài người nói hoặc giọng nói theo một phương ngữ của một vùng miền nào đó), một số bộ dữ liệu có tiêu chí xây dựng chưa rõ ràng như vừa sử dụng để nhận dạng vừa sử dụng để tổng hợp.

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

*** Một số bộ CSDL của các ngôn ngữ giàu tài nguyên mở cho TTS:**

Bộ dữ liệu	Thời gian (giờ)	Số người đọc	Số câu	Ngôn ngữ
LJSpeech	24	1	13.100	Tiếng Anh
VCTK	44	109	400	Tiếng Anh
AISHELL-3	85	218	88.035	Trung Quốc
DiDiSpeech-1	572	4500	225.000	Trung Quốc
DiDiSpeech-2	227	1500	171.361	Trung Quốc
J-MAC	31,5	39	-	Nhật Bản
JVS	30	100	-	Nhật Bản

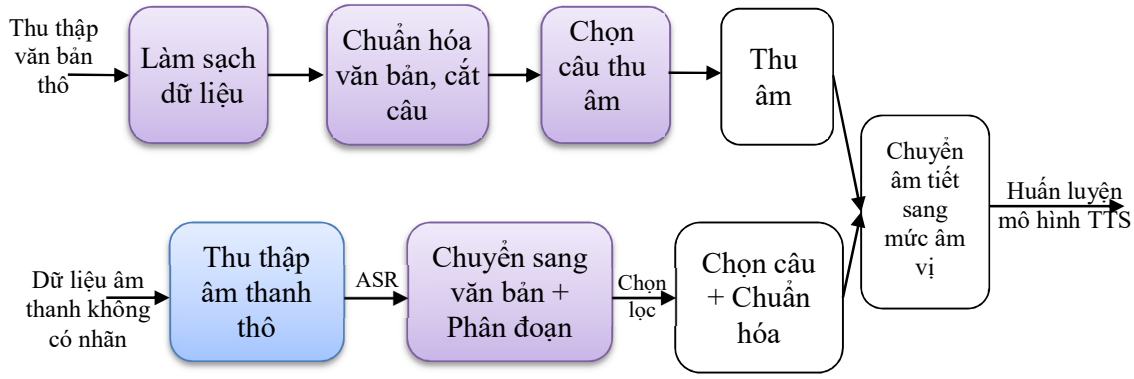
*** Một số bộ CSDL tiếng Việt mở cho TTS được công bố:**

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, khó tiếp cận [97]. 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 giọng nói.

Khi nghiên cứu tổng hợp hoặc thích nghi giọng 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 giọng 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 (Hình 13).



Hình 13: Quy trình xây dựng dữ liệu từ nguồn âm thanh có sẵn và tự thu âm

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

Sự khác biệt giữa bộ xây dựng CSDL cho tổng hợp và thích nghi đó là:

- CSDL cho tổng hợp đơn người nói cần số giờ dữ liệu huấn luyện cho mỗi giọng đọc đủ lớn (trên 10 giờ), trong khi CSDL cho huấn luyện đa người nói thì mỗi giọng đọc không cần quá lớn (khoảng 450-500 câu) nhưng số giọng đọc cần nhiều và đa dạng (từ vài chục đến hàng trăm người). Tuy nhiên cả hai bộ dữ liệu này cần đảm độ cân bằng âm vị và bao hàm số lượng từ vựng nhiều nhất có thể.

-Cơ sở dữ liệu dùng cho thích nghi là các bộ có CSDL đơn người nói có kích thước nhỏ sử dụng cho từng mục đích đánh giá khác nhau (có thể là vài giây hoặc vài phút, có chọn lọc).

Do vậy, để đảm bảo bộ CSDL cho tổng hợp tiếng đủ lớn và đa dạng cần số cần xác định xây dựng CSDL từ hai nguồn: 1) Tự thu âm; 2) Gán nhãn âm thanh từ nguồn có sẵn. Mỗi một nguồn đều có những ưu nhược điểm riêng. Xây dựng CSDL tự thu âm có ưu điểm là chủ động và giám sát được nội dung và chất lượng thu âm. Tuy nhiên, nhược điểm của phương pháp này là đòi hỏi chi phí cao (tốn thời gian, kinh phí, cơ sở vật chất và nguồn lực) và khó thu thập số giọng nói đa dạng; Xây dựng CSDL từ nguồn có sẵn có ưu điểm sẽ tiết kiệm được chi phí thu âm và đa dạng giọng đọc cũng như có thể tạo bộ CSDL có thời lượng dài nhiều giờ, tuy nhiên có nhược điểm của nó là sẽ khó kiểm soát chất lượng và sự đồng đều của dữ liệu, giọng đọc ngẫu nhiên thu thập được cũng sẽ khó đảm bảo sự cân bằng âm.

2.1.2.1. 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). Cách thực hiện:

+ Phát triển hệ thống hỗ trợ thu dữ liệu và tuyển giọng đọc:

- Ứng dụng mobile, web cho người đọc và quản trị;
- Ứng dụng web.

+ Thiết kế văn bản đọc:

- Thiết kế bộ văn bản thu âm bao gồm ~10.000 câu (bao gồm đủ các âm vị trong tiếng Việt, nội dung trong 7.000 từ vựng thường dùng trong tiếng Việt).

+ Thu âm:

- Chọn giọng đọc;
- Thống nhất quy tắc phát âm (tên nước ngoài, tên riêng);
- Người đọc tiến hành đọc văn bản có sẵn trên phần mềm;
- Người đọc đọc lại các câu chưa đạt dựa trên đồ thị sóng âm.

+ Kiểm tra âm thanh:

- Người quản trị kiểm tra chất lượng mỗi âm thanh đã thu âm;

- Đánh dấu các âm thanh chưa đạt.

2.1.2.2. Quy trình xây dựng CSDL từ nguồn âm thanh có sẵn

Chọn chủ đề âm thanh →Thu thập các nguồn âm thanh/video → Chuyển đổi về âm thanh và chuẩn hóa định dạng → Nhận dạng ASR sơ bộ ra văn bản →Chọn lọc âm thanh tiếng Việt dựa vào văn bản → Chia nhỏ âm thanh đã nhận dạng sơ bộ →Phát triển các công cụ gán nhãn →Quy tắc phiên âm→ Gán nhãn → Kiểm tra chéo → Nghiệm thu. Các thực hiện:

+ Phát triển hệ thống hỗ trợ thu dữ liệu và tuyển chọn người gán nhãn:

- Ứng dụng mobile, web cho cộng tác viên và quản trị;
- Ứng dụng web.

+ Lọc dữ liệu thô:

- Thu thập toàn bộ âm thanh/video thu thập từ các nguồn khác nhau (Youtube, VOV, VTV, phim truyện, kênh truyền thanh truyền hình khác v.v.);
- Tiền xử lý âm thanh/video về âm thanh chuẩn;
- Chuyển từ âm thanh sang văn bản cho toàn bộ dữ liệu đã thu thập;
- Chọn lọc âm thanh (âm thanh tiếng Việt, tỷ lệ âm thanh có tiếng nói, loại bỏ âm thanh có nội dung trùng lặp).

+ Gán nhãn dữ liệu:

- Thống nhất quy tắc phiên âm (tên nước ngoài, tên riêng);
- Người gán nhãn tiến hành gán nhãn văn bản cho âm thanh.

+ Kiểm tra chéo:

- Giám sát viên kiểm tra chéo chất lượng, mỗi cặp âm thanh/văn bản đã gán nhãn được kiểm tra bởi ít nhất 1 người, tối đa 3 người. Sử dụng tỉ lệ bình chọn chiếm đa số để lựa chọn kết quả.

2.1.2.3. Chọn lọc và chuẩn hóa văn bản

Nguyên tắc chung khi xây dựng một dữ liệu tiếng nói là dữ liệu sẽ chứa đựng các từ và câu thường xuyên sử dụng nhất. Ngoài ra, dữ liệu bao gồm đủ lớn các biến thể để có thể hỗ trợ tạo ra ngôn ngữ nói tự nhiên và linh hoạt. Do đó, mỗi dữ liệu tiếng nói bao hàm một trong những mục tiêu sau: bao quát về mặt ngữ nghĩa, bao quát về mặt cú pháp, bao quát về mặt ngôn điệu và bao quát về mặt từ. Tuy nhiên để đảm bảo được đồng thời những mục tiêu trên là rất khó. Thông thường, các từ, các câu và các đoạn cố gắng được lựa chọn sao cho chứa đựng tất cả các

phụ âm, vần, phong phú về ngữ cảnh (context) và đa dạng về từ vựng, ngữ nghĩa và cú pháp, đầy đủ về thanh điệu.

Do cơ sở dữ liệu giọng nói cần được xây dựng đảm bảo các tiêu chí là số lượng không quá lớn nhưng đảm bảo về sự cân bằng âm vị (chứa đủ các âm vị) và phủ đa số ngữ cảnh cần có của một ngôn ngữ. Nên việc đầu tiên của xây dựng là tiến hành thu thập các dữ liệu văn bản đa dạng. Đầu tiên, thu thập các bài báo điện tử tiếng Việt từ một số tờ báo lớn ở mọi lĩnh vực. Sau đó, chia quá trình xử lý văn bản thành các pha độc lập với phân bổ tham số lựa chọn và dữ liệu xử lý cho nhiều mục đích. Kết quả cuối cùng là chọn ra trong toàn bộ dữ liệu các câu văn bản cho người thu âm.

Dưới đây là nguyên lý chung của giải thuật Tham lam khi lựa chọn văn bản đã thu thập được. Tuy nhiên trên thực tế, có một số các vấn đề quan trọng khác ảnh hưởng đến quá trình chọn văn bản. Một trong những vấn đề đó là, trong bước chọn câu tiếp theo, nếu có nhiều câu có số lượng bao phủ âm tiết bằng nhau, thì sẽ chọn câu nào? Và hơn nữa mục tiêu là để chọn lựa ra tập câu S nhỏ nhất có thể, đồng thời có phân bố các âm tiết cân bằng. Do vậy ngoài tiêu chí mỗi câu được chọn bao phủ nhiều âm tiết trong từ điển âm tiết nhất, cần phải đưa thêm vào các tiêu chí k khác, ví dụ như độ dài các câu hay phân bố các âm tiết trong câu. Việc chọn lựa và bố trí ảnh hưởng các tiêu chí này với nhau như thế nào thì phụ thuộc vào tiêu chí nào có ưu tiên đáp ứng được yêu cầu của tập câu lựa chọn, nhưng nói chung việc lựa chọn này thường được mô hình hoá thành một hàm trọng số W, thể hiện sự "hữu ích" của mỗi câu so với yêu cầu của tập câu được chọn.

---

* Đầu vào : Tập các câu S* thu thập được ở bước bước trên
* Đầu ra : Bộ câu S đã lọc với các tiêu chí

Bắt đầu :

- S = Ø // S: tập câu được chọn
- Khởi tạo D=(C,f_i = n) // D: Từ điển âm tiết, số đếm âm tiết được gán bằng n.
- Lặp cho đến khi f_i=0 với mọi i
  - // Chọn câu
  - Xây dựng tập ứng viên S* = {s_j}, s_j là các câu chứa nhiều âm tiết trong D nhất.
  - Chọn một câu s_j thuộc S*:
    - Với mỗi tiêu chí k, gán trọng số tiêu chí k cho s_j.
    - Tổng hợp trọng số w_j của câu.
  - Chọn câu s_{max} có w_j lớn nhất.
  - // Cập nhật cho bước lặp tiếp
  - Loại bỏ s_{max} trong C.
  - Cập nhật s_{max} vào S.
  - Cập nhật số đếm của các âm tiết trong D.

Kết thúc.

Bên cạnh dó, khi xây dựng hệ thống TTS tiếng Việt nó vẫn tồn tại những nhược điểm: 1) Ngữ điệu phát âm không tự nhiên khi tổng hợp câu dài (câu phát âm liên tục không ngắt nghỉ ở câu dài), 2) Không tốt khi đọc các từ mượn tiếng Anh. Nguyên nhân bởi vì khi nói các câu dài hoặc các đoạn dài người nói thường ngắt các cụm từ bằng chèn các chuyển từ (các đoạn nghỉ) thay vì ngắt nghỉ theo dấu câu điều đó khiến câu nói dễ để diễn đạt, dễ hiểu hơn và cũng dễ dàng lấy hơi. Ngoài ra, có nhiều từ nước ngoài (từ mượn) trong các câu mà không có trong từ điển phát âm tiếng Việt. Nếu chỉ thay thế các từ nước ngoài bằng bảng phiên âm quốc tế (International Phonetic Alphabet - IPA) thì câu tổng hợp sẽ không phát âm được giống chuẩn tiếng Việt. Do vậy, để giải quyết vấn đề này cần hai mô đun tiền xử lý bộ CSDL có thể cung cấp thêm các thông tin về 1) Hệ thống phát hiện điểm dừng để dự đoán và chèn dấu câu vào các câu dài nhằm tăng tính tự nhiên của câu nói khi tổng hợp. 2) Hệ thống biên dịch các từ mượn thành các từ phát âm theo chuẩn tiếng Việt.

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

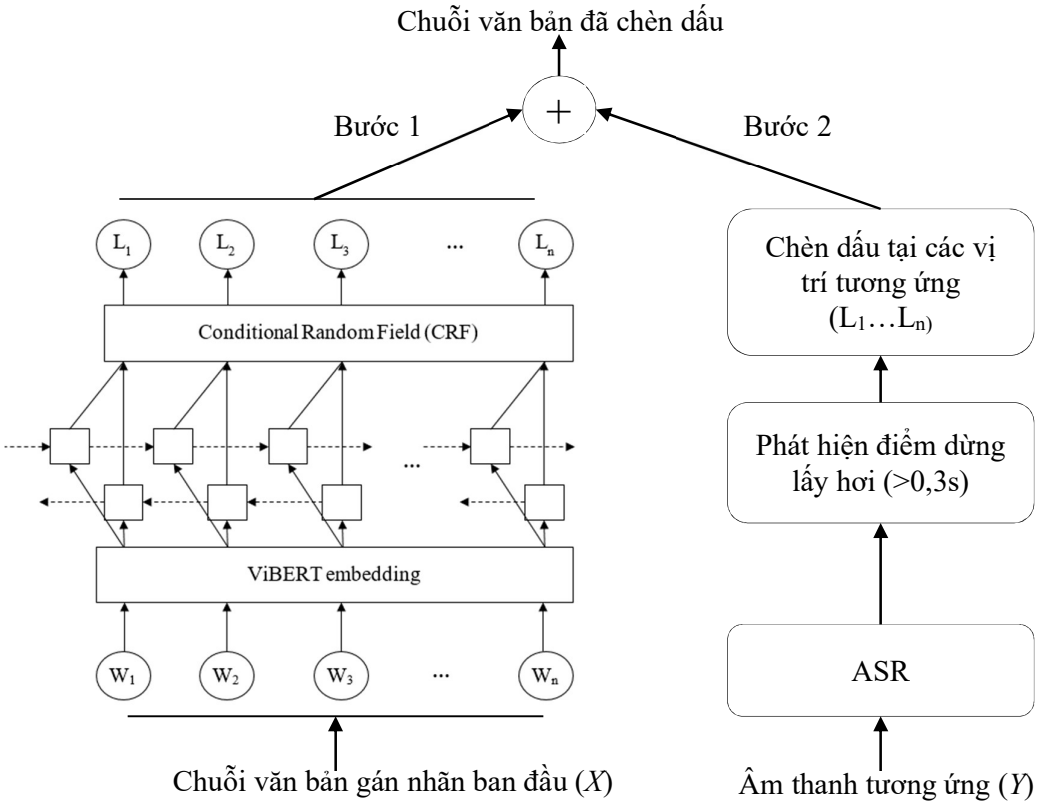
Khi đọc các câu dài, người đọc luôn có xu hướng dừng nghỉ lấy hơi sau một vài từ hoặc sau một khoảng thời gian hoặc ở vị trí của hai hay nhiều từ có tầm quan trọng như nhau về mặt cú pháp (chẳng hạn như danh từ, động từ, v.v.) ngoài các điểm dừng tại vị trí có dấu câu. Do vậy, các điểm dừng lấy hơi trong quá trình phát âm thực tế không hoàn toàn giống các vị trí có dấu câu trong văn bản gán nhãn hoặc văn bản đầu vào. Ngoài ra, dữ liệu nhãn được cung cấp từ kết quả của hệ thống ASR thông thường chỉ có văn bản mà không có dấu câu. Chất lượng âm thanh tổng hợp của mạng nơ-ron sâu phụ thuộc vào dữ liệu đầu vào. Do đó, việc thêm dấu câu ở vị trí thích hợp có thể bổ sung thông tin nhãn giúp nâng cao độ tự nhiên của hệ thống tổng hợp tiếng nói. 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. Hình 14 mô tả phương pháp gán nhãn tăng cường độ tự nhiên thông qua kết hợp giữa tự động chèn dấu câu và chèn điểm dừng lấy hơi vào văn bản. Để huấn luyện mô hình TTS, dữ liệu huấn luyện được sử dụng bao gồm âm thanh gốc và văn bản tương ứng. Các văn bản gán nhãn này

được đưa qua mô hình BERT kết hợp ASR không chỉ tự động chèn dấu câu để sửa định dạng văn bản mà còn đặt dấu câu ở vị trí liên quan đến nhịp thở lấy hơi. Nhờ đó, 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. 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.

Kết quả của mô hình chèn dấu và chèn điểm dừng lấy hơi:

Nhãn thô: *cảm giác đó đến một cách đột ngột nhưng mụ xua đuổi nó đi không cho nó chạm tới mụ cũng như không để cho nó chạm tới nền cộng hòa.*

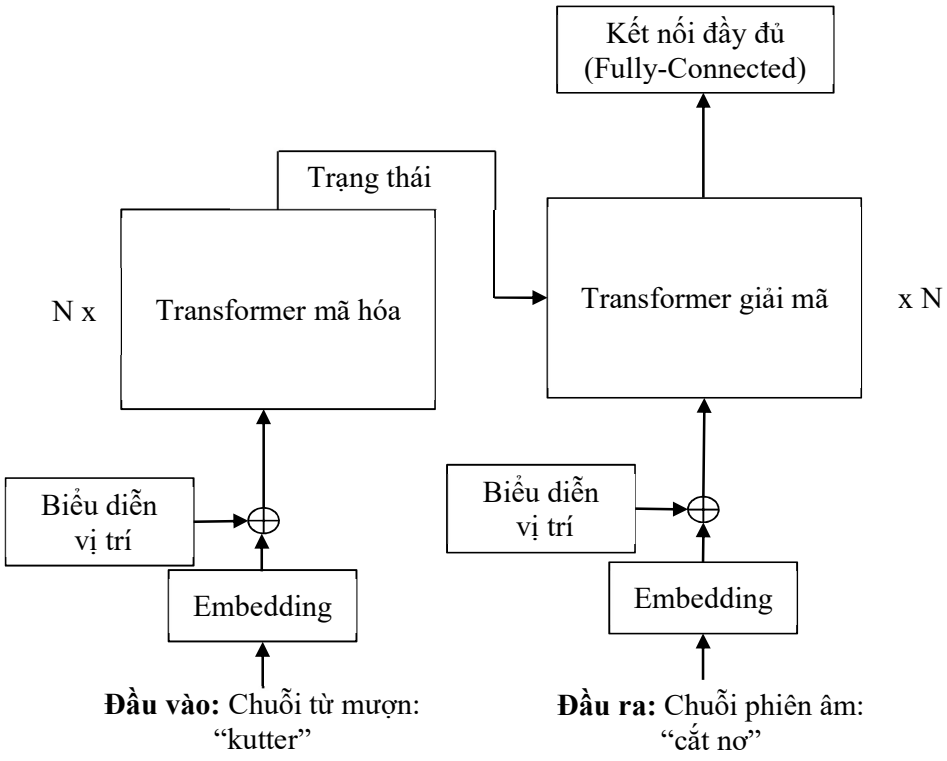
Nhãn sau khi chèn dấu: *cảm giác đó , đến một cách đột ngột , nhưng mụ xua đuổi nó đi , không cho nó chạm tới mụ , cũng như không để cho nó chạm tới nền cộng hòa.*



Hình 14: Phương pháp chèn dấu câu và chèn điểm dừng lấy hơi bổ sung nhãn thông tin cho bộ CSDL

2.1.2.5. Từ điển phiên âm từ mượn

Một trong những thách thức lớn nhất đối với nhiệm vụ VLSP Text-To-Speech [6] là văn bản phiên âm có nhiều từ nước ngoài. Bởi vì các từ nước ngoài nằm ngoài từ vựng tiếng Việt và không thể trực tiếp chuyển sang âm vị, điều này dẫn đến khó khăn cho người tham gia và xây dựng hệ thống TTS tiếng Việt. 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" (xem thêm ví dụ trong Bảng 3). 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 15 [35].



Hình 15: Kiến trúc Transformer cho mô hình phiên âm từ mượn [35]

Kiến trúc Transformer có hai mô-đun là mã hóa và giải mã và hai thành phần được kết nối thông qua một cơ chế chú ý. Mô hình Transformer sử dụng cho phần này bao gồm một chồng (stack) N = 6 lớp giống nhau cho cả khối mã hóa và giải mã.

Khối Transformer: Trong kiến trúc seq2seq, một tầng hồi tiếp được thay thế bằng một khối Transformer. Với bộ mã hóa, khối này bao gồm một tầng chú

ý đa đầu (multi-head attention) và một mạng truyền xuôi theo vị trí (position-wise feed-forward network) với hai tầng kết nối đầy đủ (fully connected layers). Đối với bộ giải mã, khối này cũng bao gồm một tầng chú ý đa đầu khác để nhận trạng thái từ bộ mã hóa.

Biểu diễn vị trí (Position Embedding): Vì tầng chú ý đa đầu không phân biệt thứ tự các phần tử trong một chuỗi, tầng biểu diễn vị trí được sử dụng để bổ sung thông tin về vị trí cho từng phần tử trong chuỗi.

Để huấn luyện mô hình phiên dịch này, phải tạo ra một số lượng lớn các cặp từ Anh-Việt như ví dụ mô tả trong Bảng 3. Kết quả xây dựng được bộ từ điển phiên âm từ mượn phục vụ cho TTS tiếng Việt.

Bảng 3: Phiên âm từ tiếng Anh sang tiếng Việt

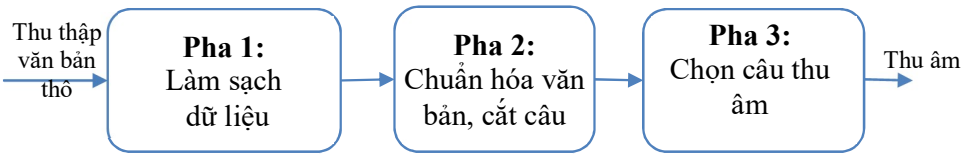
Từ tiếng Anh	Từ tiếng Việt
kuttner	cắt nơ
Anderson	an đơ sơn
vera	vê ra
reme	rê mi

2.1.2.6. Xây dựng bộ dữ liệu bằng phương pháp ghi âm

Quá trình sẽ được thực hiện bằng cách sử dụng các thiết bị và môi trường thu âm chuyên dụng để thu nhận và lưu trữ tiếng nói từ các dữ liệu tiếng nói phục vụ cho tổng hợp và thích nghi giọng nói.

i) Tiền xử lý văn bản

Do cơ sở dữ liệu giọng nói cần được xây dựng đảm bảo các tiêu chí là số lượng không quá lớn nhưng đảm bảo về sự cân bằng âm vị (chứa đủ các âm vị) và phủ đa số ngữ cảnh cần có của một ngôn ngữ. Nên việc đầu tiên của xây dựng là tiến hành thu thập các dữ liệu văn bản đa dạng. Để thực hiện phần này, cần thu thập các bài báo điện tử tiếng Việt lớn ở mọi lĩnh vực. Phần này có hai nhiệm vụ chính là chuẩn hóa văn bản đầu vào và chọn lựa câu thu âm. Để mô tả chi tiết chia quá trình xử lý thành các pha độc lập với khả năng tùy biến tham số đầu vào và xuất dữ liệu đầu ra để sử dụng cho nhiều mục đích. Kết quả cuối cùng là chọn ra trong toàn bộ dữ liệu được câu chung và câu riêng cho người thu âm. Quá trình có thể tóm tắt ở sơ đồ tại Hình 16 sau:



Hình 16: Quá trình lọc và xử lý văn bản thu âm

Pha 1: Ở pha này tiến hành lọc lấy tất cả dữ liệu có nghĩa từ các thẻ nội dung <content> trong tập tin *.txt của các trang báo điện tử đã thu thập được. Tiếp theo cắt thành từng câu, mỗi câu đặt thành một dòng dựa vào các chuỗi ký tự kết thúc câu (“. ”, “?”,“!”). Trong quá trình này cần phải loại bỏ các dòng ít ý nghĩa(thời gian đăng bài, đường dẫn tắt, địa chỉ v.v); xóa các chuỗi ký tự vô nghĩa (các dấu hoa thị, ký tự đặc biệt, các dấu ngắt câu, tên tác giả, chú thích, tên nguồn trích dẫn v.v.); xóa các câu bị trùng lặp.

Pha 2: Ở pha này nhiệm vụ chính là chuẩn hóa văn bản theo các tiêu chuẩn tiếng Việt. Một trong các công việc quan trọng là tiến hành xử lý phiên dịch các chữ số hoặc các đơn vị đo. Cụ thể là phân tách làm hai phần:

Phần 1. Xử lý cách đọc số và thời gian bằng mã nguồn riêng theo cách phiên âm thông dụng:

- Định dạng số:

Phiên âm chữ số thông qua khai báo mảng số và mảng thập phân. Một số trường hợp ngoại lệ bổ sung cách đọc riêng cho mảng số (ví dụ: không mươi → không→ lẻ, mười năm →mười lăm, mươi một → mươi mốt).

- Định dạng ngày, tháng, thời gian:
- Ký tự gắn kèm chữ số:

o Nếu định dạng dd/mm/yyyy tự động dịch thành ngày, tháng, năm.

o Nếu định dạng (dd/mm, dd-mm-yyyy và dd-mm) nhưng trước đó là chữ "ngày" hoặc chữ "hôm" đứng trước (tiền tố) thì mới phiên âm sang dạng ngày, tháng.

o Nếu định dạng hh:mm:ss thì sẽ dịch thành giờ, phút, giây.

o Nếu định dạng hh:mm thì phải có tiền tố "hồi, khoảng, lúc" ở trước thì mới phiên âm thành giờ, phút.

o Tách ký tự gắn kèm chữ số bằng dấu khoảng cách (ví dụ: 10kg→10 kg, 10m→10 m, 11h→11 h, 8/10→ 8 / 10, 90%→ 90 %).

o Sau đó thay thế ký tự phiên âm cho ký hiệu hoặc đơn vị đo bằng từ điển thay thế mức từ (ví dụ: 10 kg → mười ki lô gam, 10 m→ mười mét, 11 hz → mười một hec, 8 / 10 → tám trên/phần mười, 90 % → chín mươi phần trăm).

Phần 2. Phiên âm các từ viết tắt hoặc tên riêng bằng từ điển tự định nghĩa (ví dụ: TP →thành phố, HCM → Hồ Chí Minh, VNĐ → Việt Nam đồng, Paris → Pa ri, Samsung →Sam Sung).

Sau khi chuẩn hóa văn bản , mỗi bài báo điện tử sau khi được xử lý ở pha 2 cần được nối lại với nhau thành các tệp lớn hơn phục vụ cho tổng hợp và xử lý dữ liệu. Sau đó tiến hành cắt chỉ giữ lại các câu có độ dài từ 40 đến 90 âm tiết. Ở cuối công đoạn này cũng thực hiện xóa các dòng trùng lặp và sắp xếp các câu theo thứ tự bảng chữ cái.

Phần 3. Ở pha này tiến hành kiểm tra các câu bằng từ điển tiếng Việt, chỉ giữ lại các câu có các từ nằm trong từ điển. Công đoạn cuối là chuẩn hóa các chữ hoa sang chữ thường.

Pha 3: Ở pha này sử dụng một công cụ chuyên dụng được thiết kế hiệu quả đã được công bố để lựa chọn tập câu tối thiểu để thu thập cơ sở dữ liệu giọng nói để nghiên cứu tiếng nói. Công cụ này thiết kế dựa trên thuật toán tìm kiếm tham lam đã trình này ở Mục 2.3.1.3, thuật toán có đặc tính chọn số câu văn bản nhỏ nhất từ tập con nhưng bao phủ được các đơn vị ngữ âm tập trung để giảm chi phí thu thập. Đây là pha quan trọng quyết định ý nghĩa của dữ liệu. Ở công đoạn này đã chọn ra 2 bộ dữ liệu là tập câu văn bản thu âm chung và tập câu văn bản thu âm riêng.

Bảng 4 tổng hợp dữ liệu văn bản đã thu thập được trong các bước xử lý xây dựng CSDL tự thu âm ở phần trên:

Bảng 4: Thống kê các bước xử lý dữ liệu văn bản tự thu âm

Stt	Tên	Kích thước	Đơn vị tính	Ghi chú
1	Dữ liệu gốc	1.978/9.6	Tệp/GB	Dữ liệu ở định dạng .txt
2	Pha 1	1.978/8.1	Tệp/GB	Lọc nội dung, xử lý thô
3	Pha 2	4.036.312	Câu	Chuẩn hóa văn bản, Cắt câu và ghép tệp
4	Pha 3	250/200	Câu riêng/ Câu chung	Chọn câu thu âm : mỗi người thu 250 câu chung và 200 câu riêng

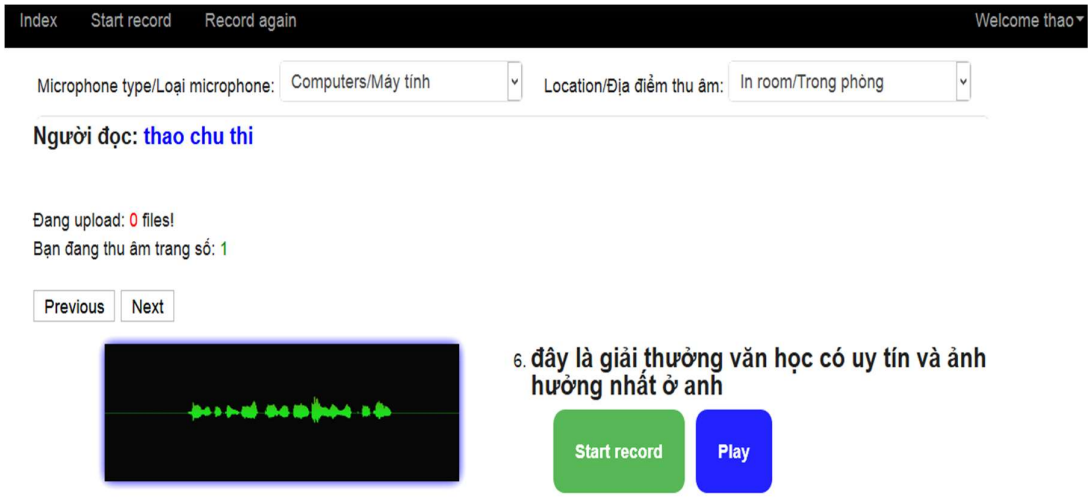
Để đảm bảo đa dạng trong tiếng nói, người đọc đã được chọn đến từ các tỉnh miền Bắc và miền Nam Việt Nam gồm 26 nam và 28 nữ từ 20 đến 35 tuổi. Các câu ghi âm chung mục tiêu là 250 câu có thuộc tính cân bằng ngữ âm. Thử

nghiệm về thuật toán lựa chọn dữ liệu được mô tả trong phần trên cho thấy rằng chỉ cần 250 cách phát biểu để bao trọn bộ monophone và 99% bi-phone. Đây là dấu hiệu tốt rằng không phải chọn nhiều câu để đáp ứng các yêu cầu về tổng hợp hay thích nghi mà vẫn bao hàm đủ các âm vị. Sử dụng 250 câu này cho tất cả người đọc.

Tổng cộng có 54 người mỗi người đọc 200 câu riêng được chọn lọc văn bản cân bằng ngữ âm bên cạnh bộ 250 câu chung. Kết quả đã ghi âm và thu thập được 9.600 câu đọc với thời lượng khoảng 20 giờ.

ii) Thu âm

Cụ thể, phần này sử dụng một công cụ thu âm riêng viết trên nền web dựa trên Docker. Công cụ này cho phép tạo tài khoản để quản lý từng user trong toàn bộ quá trình thu âm (bao gồm việc quản lý thông tin cá nhân của người đọc, quản lý các câu văn bản thu âm và kiểm tra các câu đã thu âm của từng user bằng cửa sổ hiển thị sóng âm kèm chức năng phát lại âm thanh đã thu). Quá trình thu âm bao gồm: Bước đầu tiên sử dụng một microphone TakStar PC-K600 kết nối đến một máy tính có cấu hình trung bình qua đường line in, máy trạm chạy hệ điều hành Windows 10 và trình duyệt Firefox. Quá trình thu được thực hiện trong phòng thu chuyên dụng có cách âm và khử vọng nhằm đảm bảo triệt tiêu tối đa các tạp âm; Tiếp theo người quản trị import toàn bộ dữ liệu văn bản thu âm theo cấu trúc vào hệ thống, sau đó thiết lập số câu thu âm cho mỗi tài khoản, thông số thu âm gồm: Tốc độ lấy mẫu là 48.000Hz, độ phân giải 16 bit, âm thanh thu đơn kênh(mono); Cuối cùng từng người đọc sẽ đăng nhập và đọc theo từng câu văn bản đã được phân bổ. Người đọc sẽ bấm nút “Start recording” và đợi 3s để bắt đầu đọc câu văn bản thu âm hiển thị sẵn bên cạnh, sau khi đọc xong sẽ đợi 1s rồi ấn nút “Stop recording” câu thu âm sẽ được lưu lại, một cửa sổ hiển thị sóng âm đã thu ngay sau khi người đọc chọn dừng thu cho phép giám sát chất lượng sóng âm, nút “Play” cho phép nghe lại câu vừa thu (Hình 17).



Hình 17 : Giao diện thu âm trên nền web

Trong quá trình thu âm nếu câu nào không đảm bảo thì người đọc chỉ cần thực hiện lại bước này để ghi đè bằng câu thu âm mới. Quá trình thu âm được giám sát bởi người quản trị để đảm bảo các câu thu âm đạt chất lượng theo yêu cầu. Với các câu không đạt yêu cầu người quản trị sẽ đánh giá là "Bad" trong mục "verify", khi đó người thu âm sẽ mở cửa sổ "Recording again" để thu âm lại các câu chưa đạt đã được lọc. Quá trình thu âm hoàn tất, dữ liệu thu âm được đóng gói mỗi câu thu âm thành các file từng file *.wav và file văn bản *.info chứa thông tin tương ứng. Sau quá trình thu âm, một nhóm rà soát sẽ làm tiếp nhiệm vụ nghe lại và lọc dữ liệu lần 2 để đảm bảo nội dung thu âm và chất lượng thu âm được kiểm soát.

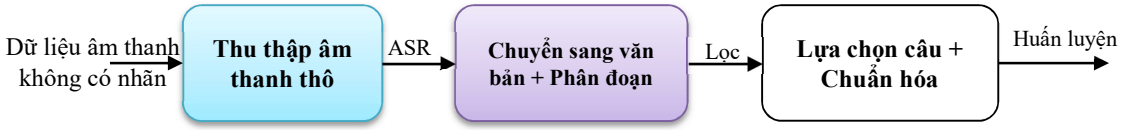
2.1.2.7. Xây dựng bộ dữ liệu chi phí thấp từ nguồn có sẵn

Việc thu thập lượng dữ liệu đó rất tốn kém, mất thời gian và không thể nếu muốn tổng hợp thêm giọng mới. Vì vậy, cần có chiến lược tự xây dựng bộ dữ liệu TTS từ các nguồn ghi chất lượng miễn phí. Tuy nhiên, các nguồn này thường không được phiên âm đầy đủ hoặc được dán nhãn thích hợp, đây là một thách thức lớn trong quá trình phát triển kho ngữ liệu TTS. Do đó, một khía cạnh khác của thích nghi hiệu quả TTS là thích nghi dữ liệu trên dữ liệu chưa được phiên âm. Luận án tiến hành xây dựng hai bộ dữ liệu từ nguồn có sẵn: 1) Dữ liệu cho mô hình cơ sở TTS, đây là bộ dữ liệu tiếng Việt lớn của người nói đơn lẻ; 2) Bộ dữ liệu thích nghi được chia thành nhiều bộ dữ liệu con với nhiều kích cỡ khác nhau để đánh giá chất lượng của hệ thống thích nghi.

i) Xây dựng bộ CSDL từ nguồn âm thanh có sẵn

Thu thập dữ liệu theo phương pháp truyền thống tốn nhiều thời gian và tốn kém (chọn văn bản, chọn người nói, ghi âm ...) [CT7]. Tuy nhiên, đây là một

nhiệm vụ bất khả thi khi bạn muốn xây dựng giọng mới theo yêu cầu một cách nhanh chóng (do người nói không có đủ thời gian để thu âm). Chiến lược thu thập dữ liệu được mô tả trong Hình 18 như sau:



Hình 18: Quy trình xây dựng dữ liệu từ nguồn âm thanh có sẵn

- Thu thập dữ liệu thô của âm thanh: Từ các nguồn âm thanh được ghi âm sẵn với chất lượng tiêu chuẩn và âm thanh rõ ràng đã được ghi lại trong phòng thu với thời lượng ghi âm tối thiểu 20 giờ. Nguồn âm thanh thu được có thể từ một đoạn truyện đọc dài kỳ hoặc giọng của một phát thanh viên hoặc MC trên TV, radio.

- Nhận dạng thô và phân đoạn: Cho bộ âm thanh thô thông qua hệ thống nhận dạng ASR để chuyển sang văn bản, hệ thống ASR chuyển đổi âm thanh thành văn bản và cắt các tệp âm thanh trong tập dữ liệu thành các đoạn âm thanh có độ dài từ 3-10 giây (tùy theo sự tạm dừng hoặc khoảng lặng). Các bản nhạc có tạp âm hoặc phát âm không rõ ràng sẽ tự động bị loại bỏ. Chỉ những đoạn âm thanh có văn bản rõ ràng mới được giữ lại (chọn độ tin cậy ASR trên 95%).

- Chọn các câu văn bản và chuẩn hóa âm thanh: Lọc các câu văn bản được nhận dạng để giữ lại tập hợp các câu ít nhất nhưng bao hàm nhiều âm tiết trong tiếng Việt nhất bằng thuật toán tham lam đã trình bày tại *Mục 2.1.2.3*. Chỉ giữ bộ âm thanh có kích thước khoảng 13 giờ. Kích thước này được coi là đã bao phủ khá nhiều âm vị trong tiếng Việt. Trước khi huấn luyện, cần chuẩn hóa âm thanh sang định dạng tệp .wav, tốc độ mẫu là 20.500Hz, kênh đơn âm. Bên cạnh đó, đã giảm 50% âm lượng của mỗi tệp âm thanh và giảm tiếng ồn bằng cách cắt bớt khoảng lặng ở đầu và cuối mỗi âm thanh trước khi huấn luyện. Cuối cùng các cặp âm thanh – văn bản đã chọn lọc được đưa qua các mô đun làm giàu thông tin nhãn để bổ sung điểm dừng lấy hơi và phiên âm từ mượn.

ii) Xây dựng bộ dữ liệu thử nghiệm thích nghi TTS tiếng Việt

Để xây dựng các tập dữ liệu nhỏ để đánh giá chất lượng thích nghi, trước tiên, chia tập dữ liệu thành các tập nhỏ và tiến hành đánh giá sơ bộ, sau đó lọc và chỉ giữ lại các tập dữ liệu đại diện. Chỉ giữ lại và chia những người nói mục tiêu (đích) thành các tập nhỏ với nhiều kích cỡ khác nhau: 1, 1, 1, 50, 200, 800 và 4500 câu (tương ứng với 1, 3, 5 giây và 4, 16, 60 phút và 5 giờ). Từ tập dữ liệu

của người nói mục tiêu, bằng cách áp dụng lựa chọn văn bản dựa trên tìm kiếm tham lam để tìm các câu tối ưu có phạm vi ngữ âm tốt nhất [101].

2.2. Đánh giá kết quả xây dựng bộ CSDL cho tổng hợp và thích nghi

Kết quả hai bộ CSDL tự thu âm và từ nguồn am thanh có sẵn được liệt phân tích chi tiết như Bảng 5.

Bảng 5: Thống kê dữ liệu đã xây dựng

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

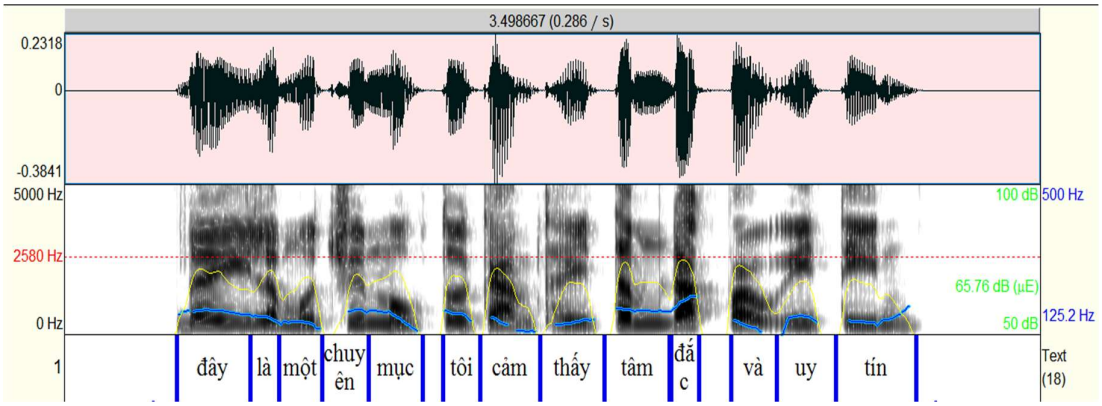
Luận án sử dụng một số mô-đun để đếm hai tập dữ liệu văn bản dựa trên tần suất xuất hiện và sự khác biệt của âm vị, âm tiết và từ trên bộ CSDL đa người nói. Kết quả như trong Bảng 6:

Bảng 6: Thống kê 20 âm vị phổ biến nhất của 2 bộ dữ liệu (bỏ silence)

Bộ dữ liệu 250 câu chung					**Bộ dữ liệu 9.600 câu riêng**			
Stt	**Bi phone**	**Số lần xuất hiện**	**Mono phone**	**Số lần xuất hiện**	**Bi phone**	**Số lần xuất hiện**	**Mono phone**	**Số lần xuất hiện**
1	ea-ngz	89	ngz	526	a-iz	3236	a	17928
2	a-iz	84	a	510	oo-ngz	2576	ngz	16940
3	oo-ngz	78	iz	347	ea-ngz	2544	iz	12532
4	l-a	76	nz	340	aa-nz	2028	nz	11484
5	oa-ngz	59	k	296	ie-nz	2016	k	9728
6	aa-nz	58	i	286	u-ngz	1936	oo	9420

7	ngz-k	56	oo	265	l-a	1928	i	9144
8	u-ngz	54	dd	236	a-nz	1888	dd	7924
9	k-o	53	tr	232	aw-iz	1876	tr	7452
10	aw-iz	52	aa	227	oo-iz	1856	aa	7296
11	a-nz	51	wa	218	i-ngz	1788	kc	6896
12	ie-uz	51	kc	216	w-a	1744	aw	6632
13	wa-ngz	50	ie	211	oa-ngz	1676	ie	6588
14	aa-tc	49	aw	202	k-o	1644	wa	6392
15	ow-iz	49	ee	190	ie-uz	1604	uz	6316
16	ie-nz	48	uz	188	ngz-k	1600	o	5856
17	uw-ngz	48	o	183	k-uo	1556	t	5772
18	w-a	45	uw	182	ow-iz	1556	mz	5768
19	wa-kc	45	th	180	b-a	1544	ee	5640
20	oo-iz	44	tc	175	uw-ngz	1516	m	5544

Chất lượng âm thanh thu được phân tích bằng phần mềm Praat phiên bản 6.0 để đánh giá các đặc trưng về sóng âm, phổ âm thanh, cao độ và cường độ âm thanh. Ở Hình 19 sau sẽ là phân tích sóng âm của hai giọng nam và giọng nữ khi đọc câu văn bản có nội dung *"Đây là một chuyên mục tôi cảm thấy tâm đắc và uy tín"*.



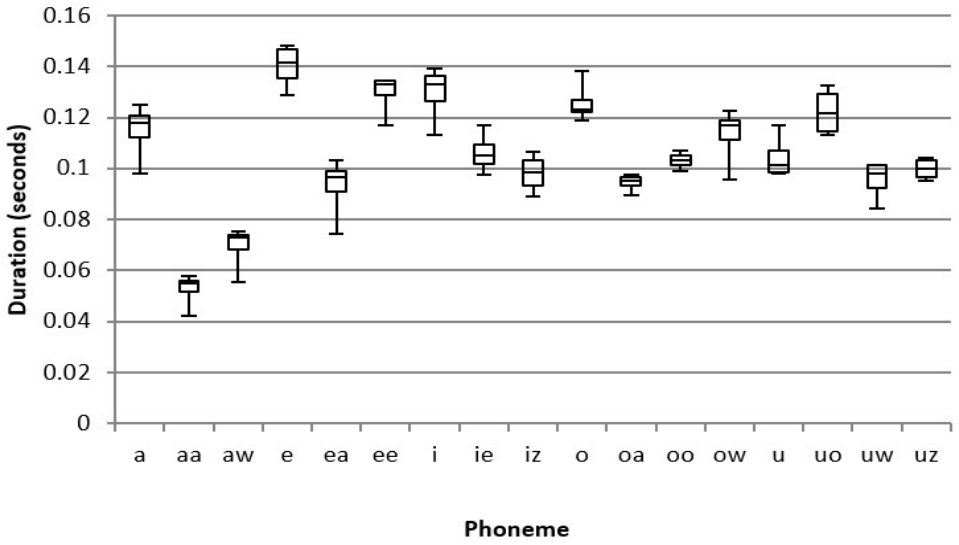
Hình 19: Ảnh sóng âm và ảnh phổ của một câu nói đã thu âm

Việc đánh giá dữ liệu được thực hiện thông qua việc đánh giá môi trường thu âm, tỷ số tính hiệu nhiễu [80]. Qua phân tích và đánh giá toàn bộ dữ liệu có thể đánh giá âm thanh thu ít nhiễu, âm thanh thu rõ ràng và chất lượng tốt.

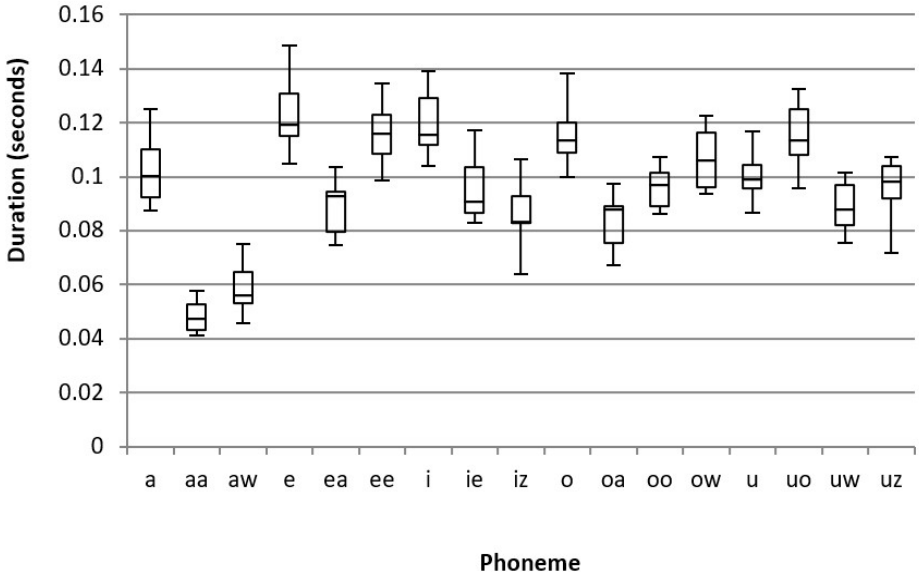
Sử dụng một công cụ phân tích và gán nhãn tự động dựa trên kho dữ liệu thu âm (tên là Audio alignment), từ đó tiến hành phân tích và thống kê sự khác biệt của âm vị giữa các đối tượng thu âm để chỉ ra các đặc trưng của đối tượng. Tự

động dựa gán nhãn dựa trên phương pháp phân khúc mô hình Markov ẩn (HMM). Quá đó, các dạng sóng giọng nói được sắp xếp thẳng hàng với các phiên âm tương ứng bằng cách sử dụng thuật toán Viterbi. Với việc sử dụng một từ điển phát âm, đã chuyển đổi từ mức phiên âm sang các mức âm vị tương ứng, thống kê thời gian âm vị được thu thập từ tệp MLF.

Khoảng thời gian trung bình của các âm vị (mean duration of the phonemes) là yếu tố giúp nâng cao chất lượng tổng hợp tiếng nói thông qua việc lựa chọn âm vị dựa trên cây quyết định với trọng số độ dài âm vị. Qua thống kê các giọng nữ cùng độ tuổi thì các đơn âm vị nguyên âm có lượng thời gian tương đương nhau.



Hình 20: Biểu đồ phân bố trường độ âm vị của các giọng nữ với cùng lứa tuổi



Hình 21: Biểu đồ của các phân bố trường độ âm vị ở nhiều độ tuổi, giới tính

Hình 20 mô tả độ dài âm vị của các giọng nữ với cùng độ tuổi, có thể thấy dải phân bổ khá nhỏ, cho thấy rằng các giọng nữ ở cùng một độ tuổi có tốc độ nói tương đồng nhau. Ở hình 21 chỉ ra rằng ở phân bổ đa dạng đa người nói ở nhiều giới tính độ tuổi khác nhau chúng có thể thấy rõ độ dài âm vị khá rộng.

Như mục đích của bộ dữ liệu này để xây dựng hệ thống tổng hợp dựa trên thích nghi. Kết quả chỉ ra một điều rất quan trọng rằng có thể không cần dùng quá nhiều dữ liệu thay vào đó chỉ cần sử dụng dữ liệu có độ tương đồng như độ tuổi hoặc giới tính để đạt được kết quả tối ưu trong thích nghi.

2.3. 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 giọng 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 giọng Việt trong các Chương tiếp theo.

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

Như đã trình bày ở Chương 2, trình bày phương pháp xây dựng bộ CSDL tiếng nói cho tổng hợp và thích nghi và một số phương pháp tăng cường nhãn thông tin để tăng cường độ tự nhiên của các hệ thống tổng hợp. Thông thường, để tạo ra một giọng tổng hợp mới thì việc huấn luyện một giọng nói từ đầu thường yêu cầu một tập dữ liệu lớn, đòi hỏi thu âm trong môi trường khắt khe về độ ồn. Tuy nhiên, với những hệ thống dựa trên DNN hoặc End-to-end hoàn chỉnh thì lượng dữ liệu cần có thể lên tới nhiều chục giờ dữ liệu, dẫn tới tốn thời gian, công sức và chi phí. Vì vậy, các nghiên cứu về thích nghi giọng nói là rất quan trọng trong việc tạo ra các giọng mới một cách nhanh chóng nhưng yêu cầu ít dữ liệu.

Nội dung của 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 giọng nói tiếng Việt dùng Multi-pass fine-tune [CT3]; 2) Trình bày phương pháp thích nghi giọng nói tiếng Việt dùng vector đặc trưng giọng nói (EMV) [CT2]; 3) Thử nghiệm và đánh giá đã chứng minh rằng, chỉ cần từ 1 đến 4 phút dữ liệu thích nghi đã cho chất lượng tổng hợp tốt và độ tương đồng cao.

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

Có hai phương pháp thích nghi chính được sử dụng trong các hệ thống tổng hợp tiếng nói dựa trên mô hình đa người nói trung bình về đặc trưng giọng nói và phong cách nói:

1) Phương pháp thích nghi mô hình dựa trên tinh chỉnh 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 đạt được bằng cách điều chỉnh lại tất hoặc một phần các tham số của mô hình. Một số đề xuất chọn giữ nguyên một số tham số mô hình và điều chỉnh lại các tham số khác của mô hình, trong khi một số đề xuất chọn điều chỉnh lại tất cả các tham số của mô hình.

2) Phương pháp thích nghi mô hình dựa trên mã hóa đặc trưng giọng nói: Trong phương pháp này, một vector hoặc mạng nhúng được sử dụng để biểu diễn người nói 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 người nói và phong cách nói khác nhau. Vào pha dự đoán, các biểu diễn của người nói mục tiêu được sử dụng như một phần của đặc trưng đầu vào để tạo ra giọng nói của người nói mục tiêu.

Một số mô hình TTS dựa trên mạng nơ-ron bao gồm các mô-đun mã hóa người nói nhằm mục đích mã hóa biểu diễn ẩn để điều khiển các đặc trưng mong muốn khi tổng hợp tiếng nói như giọng nói, ngữ điệu, phong cách nói, hoặc nhiễu [29] [102], ngoài việc mã hóa văn bản. Trong số các mô hình đó, mô hình TTS đa người nói (*từ giờ luận án gọi tắt khái niệm này bằng thuật ngữ Multi-TTS*) có thể tổng hợp bắt chước (cloning) giọng nói của đa người nói. Một thành phần quan trọng trong hệ thống Multi-TTS là bộ mã hóa người nói thể trích xuất không gian vector của người nói từ một hoặc một vài lời nói của người nói mong muốn. Vector nhúng này được sử dụng để tùy chỉnh đầu ra TTS và tạo ra các lời nói mới từ người nói đích. Bộ mã hóa người nói có thể được huấn luyện chung với các mô-đun còn lại của Multi-TTS. Bộ mã hóa người nói cũng có thể được huấn luyện trước và được sử dụng để tạo ra các vector nhúng nhằm huấn luyện Multi-TTS trên tập dữ liệu đa người nói [75].

Để triển khai Few-shot TTS, các nghiên cứu trước đây đã đề xuất quá trình thích nghi giọng của người nói huấn luyện trước các mô hình trên một tập dữ liệu lớn gồm đa người nói, sau đó tiếp tục huấn luyện trên một tập dữ liệu nhỏ của các giọng nói đích [69]. Tuy nhiên, các phương pháp như vậy yêu cầu ít nhất một vài phút mẫu âm thanh cùng với quá trình tinh chỉnh bổ sung. Do đó, nó kém hấp dẫn

hơn so với yêu cầu thực tế với nhu cầu nhân bản giọng nói ngay lập tức các giọng mới một cách tùy ý.

Một số phương pháp tiếp cận khác là dự đoán một vector biểu diễn đặc trưng giọng nói từ giọng nói để nhân bản giọng nói chưa có trong quá trình huấn luyện mà không cần tinh chỉnh, sử dụng bộ mã hóa người nói được huấn luyện chung với mô hình TTS [45] hoặc một mô hình được huấn luyện riêng cho nhiệm vụ xác minh người nói [75]. Tuy nhiên, khó để tạo ra vector biểu diễn vector nhúng đại diện cho mọi đặc trưng của giọng nói, bao gồm cả định danh người nói và phong cách nói. Trên thực tế, các nghiên cứu chỉ ra rằng việc nhúng hoạt động kém hiệu quả nếu câu nói mẫu ngắn hơn câu nói đích [102]. Để giải quyết vấn đề như vậy, một số nghiên cứu đề xuất một số phương pháp mã hóa người nói chuyên biệt, mỗi bộ khác nhau tương ứng để đại diện cho các đặc trưng giọng nói đa dạng (ví dụ: phong cách nói và nhiễu) hoặc chuyển sang vector nhúng có độ dài thay đổi để duy trì thông tin tạm thời [103]. Tuy nhiên các nghiên cứu này chưa áp dụng cho ngôn ngữ tiếng Việt.

Sau đây, luận án sẽ trình bày hai đề xuất nhằm nâng cao chất lượng hệ thống TTS thích nghi cho tiếng Việt với mẫu thích nghi nhỏ có huấn luyện lại mô hình (Few-shot TTS): 1) Nâng cao chất lượng TTS thích nghi đơn người nói bằng kỹ thuật Multi-pass fine-tune; 2) Nâng cao chất lượng tổng hợp thích nghi bằng vector đặc trưng EMV (Extracting Mel-spectogram Vector).

3.1.1. Mô hình tổng hợp thích nghi cơ sở

Xem xét một mô hình tổng hợp đa người nói $f(x_i, s_i; \theta, e_{s_i})$ lấy một văn bản x_i và người nói s_i. Mô hình được tham số hóa bởi θ, có thể huấn luyện các tham số bộ mã hóa và bộ giải mã, và e_{s_i} có thể huấn luyện vector biểu diễn đặc trưng giọng nói tương ứng với s_i. Cả θ và e_{s_i} được tối ưu hóa bằng cách tối thiểu hóa hàm mất mát L thông qua đối sánh giữa âm thanh tổng hợp và âm thanh gốc (sử dụng L1 hoặc L2 trên spectrogram):

$$\min_{\theta,e} \mathbb{E}_{(x_i,y_i)\sim\mathcal{T}_{s_i}}^{s\sim S,} \{L\big(f\big(x_i, s_i; \theta, e_{s_i}\big), y_i\big)\} \tag{3.1}$$

trong đó S là tập của đa người nói, $\mathcal{T}_{s_i}$ là một các cặp văn bản - âm thanh huấn luyện của người nói s_i và y_i là cặp âm thanh gốc tương ứng với văn bản x_i của

người nói s_i. Kỳ vọng là thông qua huấn luyện cặp âm thanh/văn bản của tất cả người nói. Trong thực tế, toán tử $\mathbb{E}$ cho hàm mất mát được xấp xỉ bằng minibatch. Sử dụng $\hat{\theta}$ và $\hat{e}$ để biểu diễn các tham số và các vector nhúng đã được huấn luyện.

Vector biểu diễn đặc trưng giọng nói đã được chứng minh là nắm bắt hiệu quả đặc trưng người nói để tổng hợp đa người nói. Chúng là các biểu diễn liên tục chiều thấp của các đặc trưng giọng nói (Embedding bản chất là vectơ mã hóa dữ liệu đầu vào có số chiều biểu diễn đặc trưng nhỏ hơn nhiều so với dữ liệu thô cần được xử lý).

Trong nhân bản giọng nói few-shot, mục tiêu của mô hình nhằm mục đích trích xuất các đặc trưng của người nói nhìn thấy s_k (có trong tập S) từ một tập âm thanh mẫu để thích nghi Y_{s_k} và tạo ra một âm thanh khác dựa trên một văn bản nhất định cho người nói đó.

3.1.2. Mô hình thích nghi dựa trên tinh chỉnh

Nếu thích nghi chỉ dựa trên vector nhúng, chúng ta có mục tiêu:

$$\min_{e_{s_k}} \mathbb{E}_{(x_k,y_k)\sim\mathcal{T}_{s_k}} \{L\big(f\big(x_k, s_k; \hat{\theta}, e_{s_k}\big), y_k\big)\} \tag{3.2}$$

trong đó $\mathcal{T}_{s_k}$ là các cặp âm thanh – văn bản của giọng nói đích s_k.

Để thích nghi toàn bộ mô hình, ta có mục tiêu sau:

$$\min_{\theta,e_{s_k}} \mathbb{E}_{(x_k,y_k)\sim\mathcal{T}_{s_k}} \{L\big(f\big(x_k, s_k; \theta, e_{s_k}\big), y_k\big)\} \tag{3.3}$$

Mặc dù thích nghi toàn bộ các tham số của mô hình cung cấp nhiều thông tin hơn cho việc thích nghi giọng nói, nhưng việc tối ưu hóa mô hình này là một thách thức, đặc biệt là đối với một số lượng nhỏ mẫu thích nghi. Trong khi chạy tối ưu hóa, việc lựa chọn cẩn thận số lần lặp lại là rất quan trọng để tránh mô hình trang bị thiếu hoặc thừa thông tin.

3.1.3. Mô hình thích nghi dựa trên mã hóa vector đặc trưng

Phương pháp mã hóa người nói để ước tính trực tiếp vector biểu diễn đặc trưng giọng nói từ các mẫu âm thanh mà không cần đưa âm thanh vào huấn luyện. Một mô hình như vậy không yêu cầu tinh chỉnh trong quá trình thích nghi nhân bản giọng nói, do đó, cùng một mô hình có thể được sử dụng cho tất cả người nói không xuất hiện trong quá trình huấn luyện.

Cụ thể, chức năng mã hóa người nói, $Encoder(Y_{s_k};\Theta)$, nhận một tập hợp các mẫu âm thanh thích nghi Y_{s_k} và ước tính e_{s_k}. Hàm tính toán được tham số hóa bởi Θ. Lý tưởng nhất là bộ mã hóa người nói có thể được huấn luyện cùng với mô hình sinh đa người nói từ đầu, với hàm mất mát được xác định cho chất lượng âm thanh được tạo ra:

$$\min_{W,\Theta} \mathbb{E}_{(x_i,y_i)\sim\mathcal{T}_{s_i}^{s\sim S,}} \{L\big(f\big(x_i, s_i; \Theta, Encoder(Y_{s_i};\Theta), y_i\big)\} \tag{3.4}$$

Lưu ý rằng bộ mã hóa người nói được huấn luyện với các người nói hiện có. Trong quá trình huấn luyện, một tập hợp các mẫu âm thanh thích nghi Y_{s_i} được lấy mẫu ngẫu nhiên để huấn luyện người nói s_i. Trong quá trình suy luận, Y_{s_i}, mẫu âm thanh từ người nói mục tiêu s_k, được sử dụng để tính $Encoder(Y_{s_i};\Theta)$. Tuy nhiên, tồn tại các thách thức về tối ưu hóa khi quy trình huấn luyện được xây dựng trong phương trình trên được bắt đầu lại từ đầu (from scratch). Một vấn đề tiềm ẩn là khi mô hình khớp với giọng nói trung bình để giảm thiểu suy hao tổng quát thường gặp phải vấn đề 'suy giảm chế độ' (mode collapse) do một mô hình sinh dữ liệu không đủ đa dạng để phân phối các điểm dữ liệu đầu vào, do vậy mô hình không thể tổng quát hóa đúng cách. Một ý tưởng để giải quyết tình trạng *mode collapse* là dùng các hàm mất mát để phân biệt cho các vector nhúng trung gian, hoặc âm thanh được tổng hợp. Tuy nhiên, trong trường hợp này, những cách tiếp cận như vậy chỉ cải thiện một chút sự khác biệt giữa các người nói. Thay vào đó, nhiều đề xuất huấn luyện riêng bộ mã hóa người nói. Đây cũng là cách tiếp cận của luận án.

Phần vector biểu diễn đặc trưng giọng nói $\hat{e}_{s_i}$ được trích xuất từ mô hình tổng hợp đa người nói đã được huấn luyện trước $f(x_i, s_i; \theta, e_{s_i})$. Sau đó, mô hình mã hóa người nói $Encoder(Y_{s_k};\Theta)$ được huấn luyện để dự đoán các vector nhúng từ mẫu âm thanh thích nghi. Hàm mục tiêu cho bài toán hồi quy tương ứng thu được kết quả tốt nhất bằng cách sử dụng hàm mất mát L1 giữa vector nhúng ước lượng và vector nhúng mục tiêu:

$$\min_{\Theta} \mathbb{E}_{s_i\sim S} \{L|Encoder\big(Y_{s_i};\Theta\big) - \hat{e}_{s_i}|\} \tag{3.5}$$

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

3.2.1. Kỹ thuật học chuyển đổi trong tổng hợp tiếng nói

Có nhiều nghiên cứu về tổng hợp giọng nói với các ngôn ngữ giàu tài nguyên như tiếng Anh và tiếng Trung cho phép tổng hợp giọng nói mới với lượng dữ liệu nhỏ dựa trên các kỹ thuật thích nghi. Một trong những cách tiếp cận để khắc phục hạn chế dữ liệu là tinh chỉnh mô hình. Để thực hiện được điều đó, cần phải một mô hình huấn luyện trước (pre-trained model) học từ một lượng lớn dữ liệu. Tuy nhiên, chưa có nghiên cứu nào đánh giá chính xác số lượng mẫu dữ liệu tối thiểu để tổng hợp một giọng nói mới cho tiếng Việt. Các công trình trước đây [104] đã nghiên cứu khả năng thích nghi để có TTS cảm xúc bằng cách tinh chỉnh mô hình TTS trung tính với một tập dữ liệu cảm xúc nhỏ. Mặc dù sử dụng cùng một ngôn ngữ tiếng Anh nhưng tập dữ liệu thích nghi cần có kích thước khá lớn (40 phút). Hoặc nghiên cứu [105] đã chứng minh rằng việc sử dụng Tacotron2 để thích nghi dữ liệu hiệu quả với giọng nói đa ngôn ngữ và có thể chuyển đổi mô hình TTS hiện có sang một giọng nói mới với cùng một ngôn ngữ hoặc khác ngôn ngữ bằng cách sử dụng 20 phút dữ liệu. Tuy nhiên, với cách áp dụng tinh chỉnh mô hình truyền thống, kết quả có chất lượng thấp và độ tương đồng thấp.

Học chuyển đổi (Transfer-learning, viết tắt là TL) là một kỹ thuật học máy phổ biến cho phép mô hình đã được huấn luyện trên một tác vụ cụ thể có thể được sử dụng lại để huấn luyện cho một mô hình trên một tác vụ khác có liên quan. Thay vì huấn luyện mô hình từ đầu trên dữ liệu mới, TL cho phép sử dụng các tri thức đã học được từ tác vụ trước đó áp dụng vào tác vụ mới [106]. Có thể mô tả cụ thể kỹ thuật học chuyển đổi là việc ứng dụng kiến thức đã học được từ vấn đề này (source domain – D_s), với nhiệm vụ nguồn (source task - T_S) sang vấn đề khác (target domain - D_T) với nhiệm vụ đích (target task - T_T) có liên quan. Một miền D được định nghĩa là:

$$D = \{X, P(X)\} \tag{3.6}$$

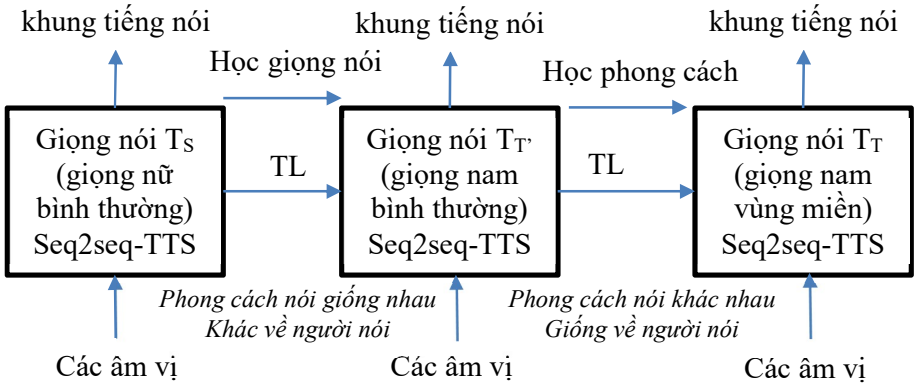
trong đó X là một không gian đặc trưng và $P(X)$ là phân phối xác suất của $X = \{x_1, \ldots, x_n\} \in X$. Cho một miền D, một tác vụ được định nghĩa là $T = \{Y, f(x)\}$, trong đó Y là một không gian nhãn và $f: X \rightarrow Y$ là một hàm dự đoán khách quan.

Kiến thức về D_S và T_S có thể giúp huấn luyện hàm đích $f_T(x)$ của nhiệm vụ đích T_T trong miền đích D_T. Học chuyển đổi nhằm cải thiện việc học hàm $f_T(\cdot)$ cho nhiệm vụ đích T_T trên miền D_T. Đặc biệt, học chuyển đổi giúp cho việc học máy nhiệm vụ nguồn T_S được khắc phục đáng kể trong điều kiện khan hiếm dữ liệu của nhiệm vụ đích T_T. Trong điều kiện dữ liệu hạn chế, việc huấn luyện một mô hình mới với lượng dữ liệu nhỏ có thể dẫn đến mô hình khó có khả năng tổng quát hóa tốt. Tuy nhiên, kiến thức từ nhiệm vụ nguồn T_S được huấn luyện trên một tập dữ liệu lớn có thể rất hữu ích. Cách tiếp cận này đã được sử dụng rộng rãi trong các bài toán vụ học máy. Có hai cách tiếp cận phổ biến của học chuyển đổi: 1) để tinh chỉnh tham số mô hình nguồn cho nhiệm vụ đích [1]; và 2) để học các biểu diễn đặc trưng sử dụng mạng nguồn cho nhiệm vụ đích [107].

Học chuyển đổi áp dụng trong tổng hợp tiếng nói, cho phép mô hình học một giọng nói mới nhanh chóng thông qua học chuyển đổi các tham số huấn luyện từ các mô hình được huấn luyện trước. Học chuyển đổi thường được thực hiện cho các nhiệm vụ mà tập dữ liệu có quá ít dữ liệu để huấn luyện mô hình quy mô đầy đủ từ đầu. Trong tổng hợp giọng nói, bằng cách tinh chỉnh trên hai không gian con nhỏ hơn, có thể chứa ít dữ liệu hơn và giọng nói tổng hợp tự nhiên và giống nhau hơn. Đối với TTS, các công trình trước đây đã tập trung vào phương pháp học chuyển đổi và meta-learning tổng hợp để thích nghi với giọng nói mới [75] [44] [1]. Một cách tiếp cận truyền thống là tinh chỉnh mô hình âm học của giọng nói đã huấn luyện trước với tập dữ liệu của giọng nói đích [104] [105]. Khi tinh chỉnh, mô hình chỉ cập nhật tham số của mô-đun giải mã. Sau khi tinh chỉnh, chúng ta có thể học được toàn bộ đặc trưng của người nói từ một lượng nhỏ dữ liệu.

Trong nghiên cứu của [69] đã đề xuất một phương pháp học chuyển đổi hiệu quả từ một hệ thống tổng hợp tiếng nói từ văn bản dựa trên Seq2seq-TTS đã được huấn luyện bằng số lượng lớn dữ liệu tiếng nói có phong cách nói bình thường của cùng một ngôn ngữ để học chuyển đổi phong cách nói và giới tính. Trong Hình 22, phương pháp học chuyển đổi của nghiên cứu [69] chia làm hai bước:

Đầu tiên huấn luyện một hệ thống Seq2seq-TTS với một giọng nữ bình thường với dữ liệu lớn khoảng 16 giờ (gọi là T_S). Sau đó, hệ thống tinh chỉnh mô hình đã học với giọng nói nam bình thường với dữ liệu hạn chế khoảng 2 giờ (gọi là $T_{T'}$). Cuối cùng, sử dụng một ít mẫu giọng đích D_T với dữ liệu khoảng 30 phút, hệ thống tinh chỉnh lại mô hình một lần nữa để tạo ra tiếng nói tổng hợp có đặc trưng giống giọng đích T_T (giọng nam vùng miền). Vì sự thành công của kỹ thuật học chuyển đổi phụ thuộc vào sự giống nhau giữa các nhiệm vụ nguồn và nhiệm vụ đích, nghiên cứu đã sử dụng cách tiếp cận tinh chỉnh được trình bày trong hình dưới thay vì thích nghi giọng T_S (giọng nữ bình thường) trực tiếp thành giọng T_T (giọng nam vùng miền). Tuy nhiên có thể thấy rằng đề, xuất này mới áp dụng học chuyển đổi trên cùng một ngôn ngữ là tiếng Anh mà chưa áp dụng cho ngôn ngữ nghèo tài nguyên với dữ liệu nguồn hạn chế.



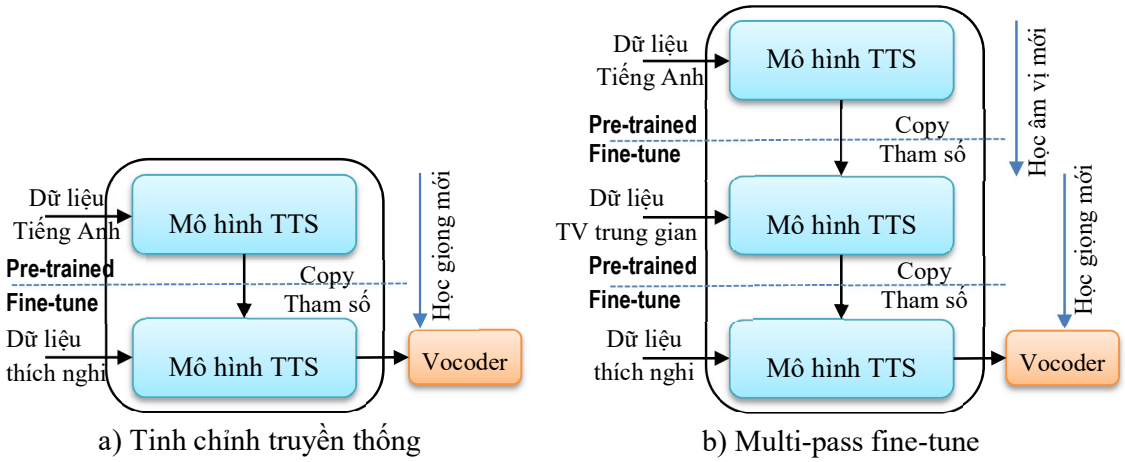
Hình 22: Sơ đồ luồng thích nghi giọng nói bằng tinh chỉnh truyền thống [69]

3.2.2. Đề xuất kỹ thuật Multi-pass fine-tune cho tổng hợp tiếng nói tiếng Việt

Mô hình âm học cơ bản của hệ thống TTS End-to-end có thể coi là một hàm chuyển thông tin văn bản thành các đặc trưng âm thanh và nó cho thấy rằng một cách tiếp cận thông thường là tinh chỉnh mô hình âm học của người nói đã được tinh chỉnh trước với mẫu dữ liệu của giọng nói đích dựa trên Tacotron2 (được biểu thị ở bên trái của Hình 23 [CT3]).

Tuy nhiên, 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 (overfitting do) thích nghi trực tiếp trên mô hình âm học End-to-end.



Hình 23: Thích nghi một giọng nói mới với Multi-pass fine-tune

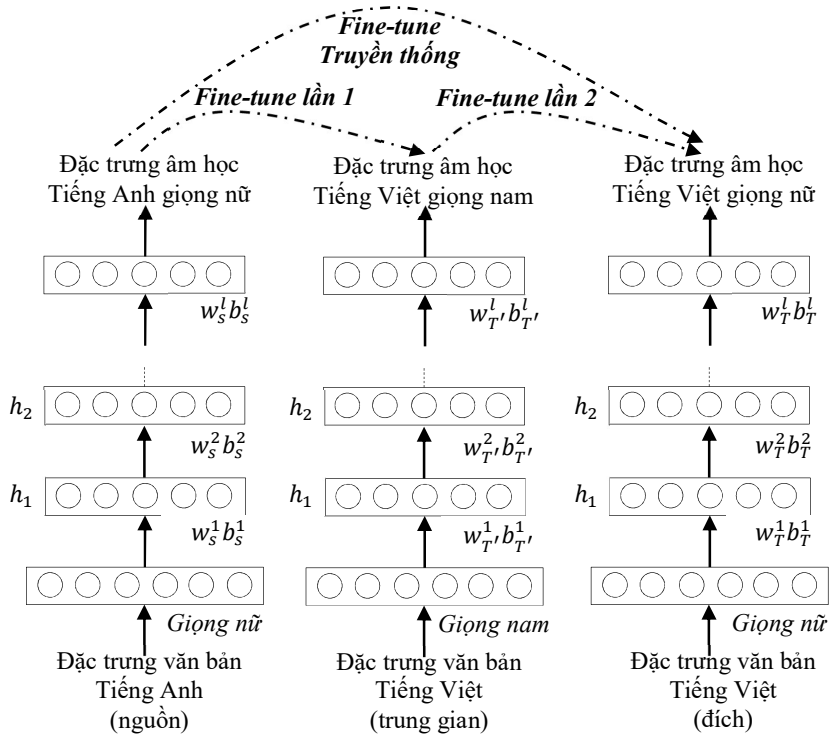
Để giải quyết những vấn đề này, luận án đề xuất ban đầu mượn một mô hình huấn luyện trước của tiếng Anh là ngôn ngữ giàu tài nguyên và có các âm vị tương đương với tiếng Việt. Mô hình huấn luyện trước này đã được huấn luyện bằng giọng đọc nữ có thời lượng hơn 23 giờ, mô hình đã bao hàm đầy đủ âm vị và âm của một ngôn ngữ là tiếng Anh và tiếp theo được tinh chỉnh lần thứ nhất với một mô hình huấn luyện trước bằng tiếng Việt trung gian (huấn luyện bằng giọng nam phong cách đọc có thời lượng khoảng 5 giờ) để mô hình học chuyển đổi các vector nhúng âm vị chia sẻ từ tiếng Anh sang tiếng Việt (âm vị, âm), sau đó tinh chỉnh lần thứ hai với một mẫu giọng nói thích nghi (giọng nữ phong cách đọc với thời lượng vài phút) để học chuyển đổi đặc trưng âm học của người nói giữa hai giọng tiếng Việt với chỉ một lượng mẫu nhỏ. Luận án gọi nó là phương pháp "Tinh chỉnh nhiều lần" hay còn gọi là "Multi-pass fine-tune" và biển thị sơ đồ ở bên phải của Hình 23, mô hình đề xuất chỉ cầu một mẫu nhỏ để điều chỉnh một giọng nói mới. Vì đặc trưng âm học chính đã được học chuyển đổi từ tiếng Anh (tập dữ liệu lớn) và tiếng Việt (tập dữ liệu trung bình), để tạo ra giọng nói mới, do vậy chỉ cần một lượng nhỏ dữ liệu thích nghi cũng cho phép học đặc trưng giọng nói đích.

Hình 24 mô tả các bước để thực hiện Multi-pass fine-tune, trong đó *s* biểu diễn tham số nguồn, *T'* biểu diễn tham số giọng trung gian và *T* biểu diễn tham số giọng đích:

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 $w_s^1, w_s^2, .. w_s^l, b_s^1, b_s^2, .. b_s^l$;

2) Sau đó dùng mạng này để huấn luyện thích nghi giọng tiếng Việt trung gian, các tham số của mô hình tiếng Anh $w_s^1, w_s^2, .. w_s^l, b_s^1, b_s^2, .. b_s^l$ được cập nhật bằng các tham số của mô hình tiếng Việt trung gian $w_{T'}^1, w_{T'}^2, .. w_{T'}^l, b_{T'}^1, b_{T'}^2, .. b_{T'}^l$;

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 $w_{T'}^1, w_{T'}^2, .. w_{T'}^l, b_{T'}^1, b_{T'}^2, .. b_{T'}^l$ được cập nhật bằng các tham số thích nghi của giọng tiếng Việt đích $w_T^1, w_T^2, .. w_T^l, b_T^1, b_T^2, .. b_T^l$. Với phương pháp tinh chỉnh truyền thống mô hình chỉ huấn luyện mô hình tiếng Anh và cập nhật bộ tham số $w_s^1, w_s^2, .. w_s^l, b_s^1, b_s^2, .. b_s^l$ bởi bộ tham số thích nghi $w_T^1, w_T^2, .. w_T^l, b_T^1, b_T^2, .. b_T^l$



Hình 24: Cập nhật tham số thích nghi bằng Multi-pass fine-tune và tinh chỉnh truyền thống

Trong học chuyển đổi đa ngôn ngữ, giọng nói đích được học chuyển đổi từ một ngôn ngữ khác với giọng nói nguồn. Vì ngôn ngữ đích có thể sử dụng kết hợp các âm vị khác nhau để phát âm từ, nên việc đóng băng bộ mã hóa văn bản sẽ gây ra các vấn đề về phát âm trong ngôn ngữ đích. Do vậy, cần thực hiện thành ba bước: Đầu tiên, cần đóng băng bộ giải mã và bộ mã hóa sau (posterior encoder) trong suốt quá trình tinh chỉnh. Tiếp theo, vì vai trò của bộ giải mã và bộ mã hóa sau là tái tạo âm thanh chất lượng cao từ dữ liệu gốc, khi những mô-đun này đã được huấn luyện đủ với một tập dữ liệu lớn (ở đây là bộ dữ liệu tiếng Anh và tiếng Việt trung gian) và không cần phải tinh chỉnh trên một tập dữ liệu nhỏ. Đóng băng bộ giải mã mang lại nhiều lợi ích vì không cần phải thực hiện truyền xuôi qua bộ giải mã đã đóng băng, nên có thể tiết kiệm bộ nhớ và tài nguyên tính toán để sinh âm thanh có độ phân giải cao. Cuối cùng, bộ mã hóa văn bản và bộ dự đoán trường độ (duration predictor) được huấn luyện từ đầu. Để có khả năng tổng quát tốt hơn, không đưa điều kiện vector biểu diễn đặc trưng giọng nói vào bộ dự đoán trường độ.

Mô hình ánh xạ âm vị giữa các ngôn ngữ nguồn và ngôn ngữ đích dựa theo cách phát âm của từ, vectơ biểu diễn đặc trưng giọng nói của người nói cũ bị đóng băng và mô hình TTS chỉ cho phép người nói mới thích nghi vector nhúng. Do đó, kỹ thuật tinh chỉnh nhiều lần này cũng cho phép học chuyển đối với một ngôn ngữ mới.

3.2.3. Thử nghiệm đánh giá và kết quả

3.2.3.1. Thử nghiệm đánh giá

*** Bộ dữ liệu**

Để đánh giá kích thước tối thiểu và số lượng lớp tinh chỉnh nhằm tạo ra một giọng nói mới một cách hiệu quả, luận án sử dụng 3 loại tập dữ liệu:

- Tập dữ liệu tiếng Anh: Sử dụng kho ngữ liệu LJSpeech-1.1, đây là tập dữ liệu giọng nói đã công bố bao gồm 13.100 đoạn âm thanh ngắn của một giọng nữ đọc 7 cuốn sách. Mỗi clip được cung cấp một phiên âm (văn bản) đi kèm. Các đoạn âm thanh có độ dài khác nhau từ 1 đến 10 giây và có tổng thời lượng khoảng 24 giờ.

- Tập dữ liệu tiếng Việt trung gian: Sử dụng bộ dữ liệu tiếng Việt trung gian, bao gồm 13.125 đoạn âm thanh của giọng nam đọc tin tức thời sự được xây dựng

từ Chương 2. Các đoạn âm thanh có độ dài khác nhau từ 1 đến 10 giây và có tổng thời lượng khoảng 15 giờ.

- Bộ dữ liệu thích nghi: 4 bộ thích nghi từ 50 đến 800 câu (tổng thời lượng từ 4 đến 60 phút).

*** Cài đặt thử nghiệm**

Luận án đã sử dụng Tacotron2 + Waveglow cơ sở để đánh giá hệ thống TTS thích nghi. Thông qua tiến hành nhiều thử nghiệm liên quan đến xử lý dữ liệu. Mạng Tacotron2 có hai thành phần: bộ mã hóa và bộ giải mã. một thay đổi nhỏ so với mô hình ban đầu. Để thích nghi với đặc trưng của tiếng Việt, đầu vào mô hình là cấp độ âm vị thay vì cấp độ ký tự. Âm vị được chuyển đến lớp vector nhúng được biểu diễn bằng 512 chiều. Sau đó, các vectơ này đi qua một chồng 3 lớp chập, tiếp theo là các lớp LSTM hai hướng đơn lẻ để tạo ra các đặc trưng được mã hóa [CT4]. Đầu ra của bộ mã hóa được sử dụng bởi một mạng chú ý mang lại một vectơ có chiều cố định. Cuối cùng, bộ giải mã có nhiệm vụ chuyển đổi vectơ này thành một phổ Mel sau đó phổ Mel sẽ được chuyển đổi thành dạng sóng âm nhờ bộ phát âm WaveGlow. Để huấn luyện mô hình Tacotron2 cần giảm thiểu sai số đầu ra của mô hình với tham số gốc bằng cách sử dụng sai số trung bình bình phương (MSE) và huấn luyện mô hình với khởi tạo từ các tham số mô hình huấn luyện trước hoặc khởi tạo từ đầu tương ứng với các thử nghiệm sau:

Thử nghiệm 1. Xác định lượng dữ liệu tiếng Việt tối thiểu để huấn luyện mô hình Tacotron2 từ đầu (from scratch) và tinh chỉnh truyền thống. Kết quả đánh giá được thể hiện trong Bảng 7:

- Cột 2: Huấn luyện mô hình trực tiếp từ đầu với chỉ dữ liệu thích nghi (không sử dụng mô hình huấn luyện trước).

- Cột 3: Huấn luyện mô hình bằng cách tỉnh chỉnh mô hình huấn luyện trước bằng tiếng Anh kết hợp dữ liệu thích nghi (Mô hình huấn luyện trước lần 1 sử dụng tập dữ liệu LJspeech 1.1).

Thử nghiệm 2. Đánh giá mô hình thích nghi giọng nói Multi-pass fine-tune theo phương pháp huấn luyện dữ liệu thích nghi kết hợp với mô hình huấn luyện trước bằng tiếng Việt trung gian (*đã được huấn luyện tỉnh chỉnh từ mô hình huấn luyện trước bằng tiếng Anh*). Đây là huấn luyện thích nghi để tạo ra một giọng nói mới từ dữ liệu thích nghi với kỹ thuật Multi-pass fine-tune. Cỡ mẫu thích nghi

thay đổi từ 4 phút đến 1 giờ để đánh giá chất lượng thích nghi. Kết quả đánh giá thể hiện ở cột 4 của Bảng 7.

Thử nghiệm 3. Đánh giá sự giống nhau của giọng gốc và giọng thích nghi của mô hình tinh chỉnh truyền thống và mô hình Multi-pass fine-tune. Sử dụng thước đo SIM để đo độ tương đồng của giọng người và giọng tổng hợp bằng cách cho 11 người nghe chuyên nghiệp nghe 90 cặp câu được tạo ra bởi mô hình tinh chỉnh truyền thống và mô hình Multi-pass fine-tune (mô hình đề xuất) để đánh giá mức độ tương đồng (giống nhau) giữa các cặp câu, âm thanh gốc cũng được đưa vào đánh giá để đảm bảo sự khách quan. Kết quả đánh giá thể hiện ở cột 2 và cột 3 của Bảng 8.

3.2.3.2. Kết quả

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

Sử dụng thang đo MOS để đánh giá chất lượng giọng nói do hệ thống tạo ra. Phần tổng hợp tiếng nói được đánh giá bởi hai nhóm gồm 23 người nghe là chuyên gia và tình nguyện viên. Trong cột 2 và 3 của Bảng 7, so sánh chất lượng âm thanh (MOS) giữa bộ dữ liệu huấn luyện từ đầu và tinh chỉnh truyền thống:

- 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 giọng 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).

Bảng 7: 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 tiếng Việt)	Mô hình huấn luyện trước bằng tiếng Anh + Dữ liệu thích nghi	Mô hình huấn luyện trước bằng tiếng Việt trung gian + Dữ liệu 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

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

Trong các cột 4 của Bảng 7, dựa trên mô hình huấn luyện trước bằng tiếng 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 khá tốt với điểm MOS là **3.78/4.69** so với của giọng thật của người nói.

*** Tính tương đồng**

Bảng 8: 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**

Cũng trong cùng một bảng, 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 đạt **2.87/3.99** so với giọng gốc của người nói.

Hình 25: So sánh sự tương đồng của của mô hình tinh chỉnh truyền thống (trên) và mô hình đề xuất (dưới) trên tất cả các cặp câu đánh giá

Tiến hành phân tích đánh giá SIM theo phương pháp [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 25, 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 đề xuất và mô hình cơ sở là rất rõ ràng.

Bảng 9: Bảng phân tích ANOVA về điểm đánh giá tương đồng giữa mô hình tinh chỉnh truyền thống và mô hình đề xuất

Mô hình	**Nguồn phương sai**	*Tổng bình phương (SS)*	*Bậc tự do (df)*	*Bình phương trung bình (MS)*	*Giá trị thống kê (F)*	*Giá trị p*	*F tới hạn*
Tinh chỉnh truyền thống	Giữa các nhóm	6.630	10	0.663	5.188	4.69E-07	1.860
	Trong các nhóm	40.767	319	0.128			
Multi-pass fine-tune	Giữa các nhóm	49.121	10	4.912	12.287	4.94E-18	1.860
	Trong các nhóm	127.533	319	0.399			

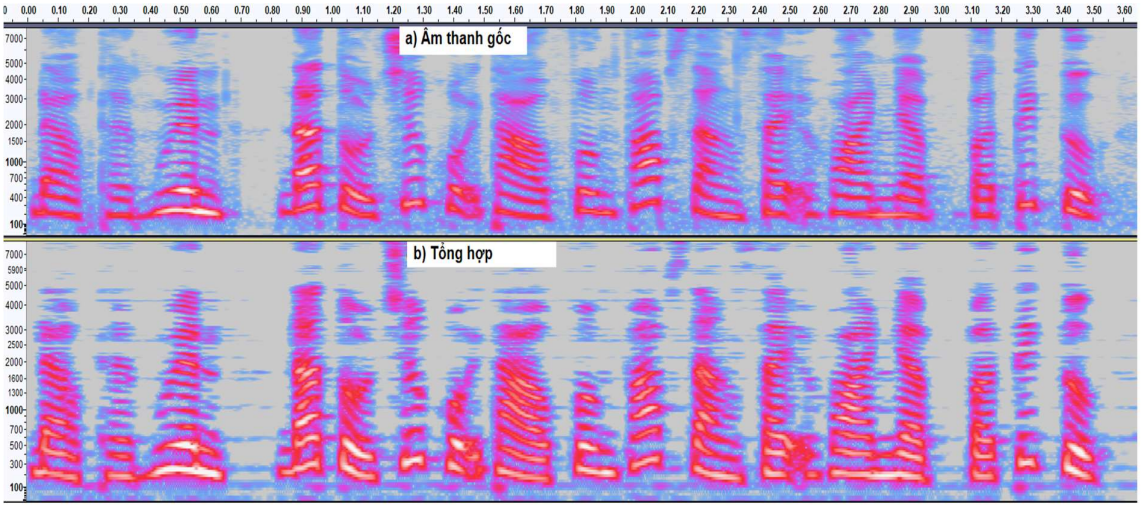
Phân tích ANOVA một chiều (phụ thuộc một biến duy nhất là mô hình TTS) trong đánh giá độ tương đồng SIM giữa các mô hình cho kết quả ở Bảng 9. Các giả thuyết cho phân tích ANOVA ban đầu như sau :

- **Giả thuyết không(H_0):** Tất cả các phương pháp có chất lượng tương đồng bằng nhau (không có sự khác biệt).
- **Giả thuyết thay thế(H_1):** Có sự khác biệt ít nhất một cặp phương pháp.

Qua kết quả phân tích ta 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.

Sử dụng độ đo méo phổ Mel (MCD) để đo lường mức độ khác nhau của hai chuỗi phổ Mel để đánh giá hiệu suất chuyển đổi giọng nói [64]. Bảng 8 cho thấy

rằng Multi-pass fine-tune cho phép tạo ra giọng nói mới với MCD thấp hơn nhiều so với phương pháp tinh chỉnh truyền thống (giảm **2.74**). Hình 26 mô tả phổ Mel của giọng nói tổng hợp và giọng người nói. Chỉ với 4 phút dữ liệu thích nghi (khoảng 50 câu), giọng nói tổng hợp thích nghi cho thấy sự tương đồng về các trường độ âm vị và sự tương đồng cao về vân phổ âm thanh.



Hình 26: Sự tương đồng giữa giọng tổng hợp và giọng người nói chỉ với 4 phút dữ liệu thích nghi

Phần này đã chứng minh rằng nếu sử dụng các kỹ thuật tinh chỉnh truyền thống thì 1 giờ dữ liệu thích nghi giọng Việt không đủ để tổng hợp các giọng mới mà cần tối thiểu 3 giờ. Luận án cũng đề xuất một số thay đổi đơn giản đối với mô hình TTS Seq2seq cơ sở với một số kỹ thuật Multi-pass fine-tune để thích nghi giọng nói mới sang các ngôn ngữ ít tài nguyên như tiếng Việt. Luận án đã chứng minh rằng chỉ mất 4 phút dữ liệu thích nghi để tạo ra giọng nói mới với độ tương đồng cao và chỉ mất 16 phút để tạo ra giọng nói chất lượng tốt.

Có thể thấy rằng sự thích nghi này dẫn đến việc phát âm các từ của giọng nói đích tốt hơn so với giọng nói nguồn trong điều kiện ngôn ngữ đích nghèo tài nguyên có thể mượn các ngôn ngữ nguồn giàu tài nguyên. Tuy nhiên việc giảm dữ liệu thích nghi nhỏ hơn so với 4 phút dữ liệu của phương pháp Multi-pass fine-tune vẫn còn là thách thức khi áp dụng kỹ thuật thích nghi. Trong phần tiếp theo luận án sẽ tiếp tục nghiên cứu các hướng thích nghi khác với dữ liệu ít dần so với phương pháp đã trình bày ở phần này.

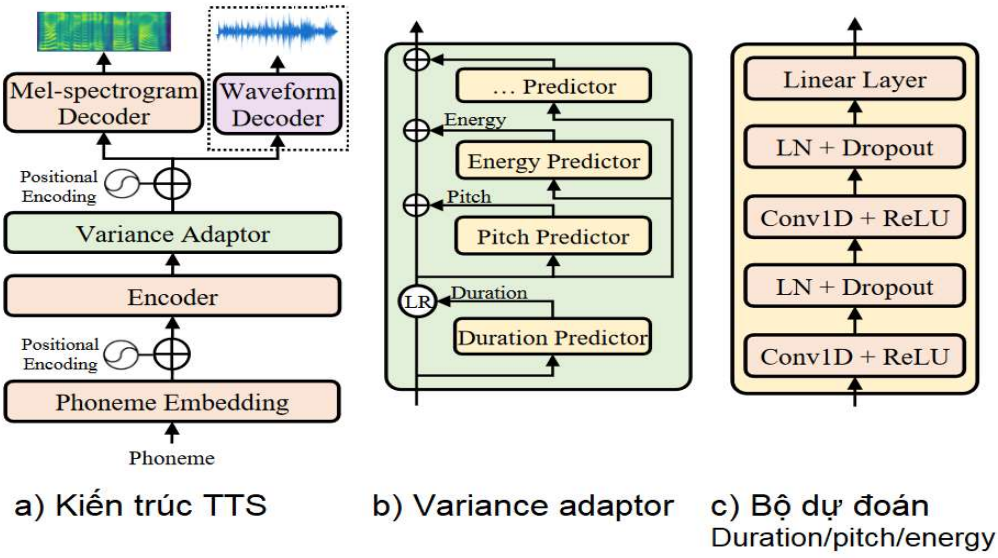
3.3. Nâng cao chất lượng tổng hợp thích nghi bằng vector đặc trưng EMV

Huấn luyện mô hình Multi-TTS yêu cầu nhiều giọng nói khác nhau để tạo ra một mô hình tổng quát. Tuy nhiên, mô hình tổng hợp giọng nói trung bình sẽ bị méo hoặc trung bình hóa, dẫn đến chất lượng thấp nếu giọng nói của người nói mới có quá ít dữ liệu để huấn luyện. Một trong các phương pháp thực hiện là tinh chỉnh mô hình, nếu không, mô hình sẽ có chất lượng thích nghi thấp. Tuy nhiên, để đạt được chất lượng thích nghi cao, ít nhất phải có hàng ngàn bước tinh chỉnh mô hình sử dụng kỹ thuật tinh chỉnh truyền thống hoặc sử dụng kỹ thuật Multi-pass fine-tune (thông qua học chuyển đổi từ các ngôn ngữ giàu tài nguyên) đã đề xuất ở Mục 3.2. Một phương pháp tiếp cận khác để huấn luyện mô hình Multi-TTS dựa trên DNN là sử dụng các mô-đun mã hóa người nói để trích xuất các biểu diễn ẩn đặc trưng của người nói và phong cách nói [108] [109]. Các phương pháp này đã cho thấy khả năng học được các đặc trưng giọng nói mới với thời gian lấy mẫu và thời gian tổng hợp giảm đáng kể. Vì vậy, phần này đề xuất giải pháp thích nghi đa người nói Few-shot TTS dựa trên kiến trúc trích xuất đặc trưng giọng nói với ưu thế là giảm thời gian tính toán huấn luyện so với các phương pháp dựa trên tinh chỉnh tham số.

3.3.1. Dự đoán và điều khiển các đặc trưng tiếng nói

Để giảm khoảng cách thông tin (do đầu vào không chứa tất cả thông tin để dự đoán mục tiêu) giữa đầu vào (chuỗi văn bản) và đầu ra mục tiêu (phổ Mel) và giảm bớt vấn đề ánh xạ một-nhiều đối với huấn luyện mô hình TTS không tự động hồi quy, một số thông tin đặc trưng của giọng nói bao gồm cao độ (pitch), cường độ (energy) và trường độ (duration) được dự đoán chính xác hơn dựa trên nền tảng đã giới thiệu trong mô hình [30] (Hình 27): Trong huấn luyện, trích xuất trường độ, cao độ và cường độ từ dạng sóng giọng nói đích và trực tiếp lấy chúng làm đầu vào có điều kiện; Trong suy diễn, sử dụng các giá trị được dự đoán bởi những bộ dự đoán được huấn luyện chung với mô hình để suy diễn ra phổ Mel hoặc suy diễn trực tiếp sang sóng tiếng nói.

3.3.1.1. Bộ thích nghi phương sai (Variance adaptor)



Hình 27: Kiến trúc Variance adaptor [30]

Bộ thích nghi phương sai (Variance adaptor) nhằm mục đích thêm thông tin phương sai (như trường độ, cao độ, cường độ) vào chuỗi âm vị ẩn, có thể cung cấp đủ thông tin để dự đoán giọng nói biến đổi và giảm vấn đề ánh xạ một-nhiều trong TTS. Mô tả ngắn gọn thông tin về phương sai như sau: 1) Trường độ âm vị, biểu thị trường độ giọng nói phát ra; 2) Cao độ, là một đặc trưng chính để truyền tải cảm xúc và ảnh hưởng lớn đến cảm xúc của giọng nói; 3) Cường độ, cho biết cường độ ở mức khung phổ Mel của tiếng nói và ảnh hưởng trực tiếp đến âm lượng và ngữ điệu của giọng nói. Thông tin về phương sai khác có thể được thêm vào trong bộ thích nghi phương sai, ví dụ như cảm xúc, phong cách nói và người nói. Tương ứng, bộ thích nghi phương sai bao gồm: 1) Bộ dự đoán trường độ (*duration predictor)*; 2) Bộ dự đoán cao độ (*pitch predictor*); và 3) Bộ dự đoán cường độ (*energy predictor*), như được biểu diễn trong Hình 27b. Trong huấn luyện, lấy giá trị của trường độ, cao độ và cường độ trích xuất từ các bản ghi âm người nói làm đầu vào cho chuỗi ẩn để dự đoán các giọng nói mục tiêu. Đồng thời, sử dụng trường độ, cao độ và cường độ dựa trên âm thanh gốc làm mục tiêu để huấn luyện các yếu tố để dự đoán trường độ, cao độ và cường độ, được sử dụng trong suy diễn để tổng hợp giọng nói đích.

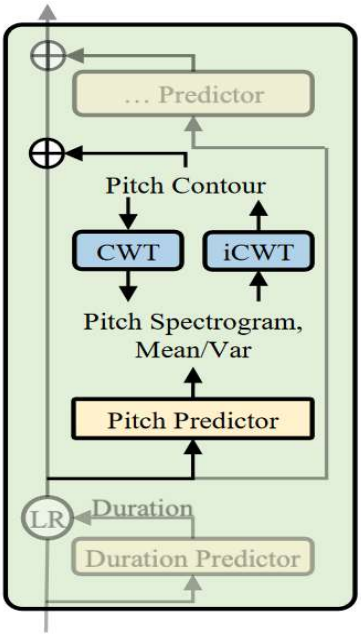
Trong Hình 27c, các yếu tố dự báo trường độ, cao độ và cường độ chia sẻ cấu trúc mô hình tương tự (nhưng các thông số mô hình khác nhau), bao gồm mạng chập 2 lớp 1D với hàm kích hoạt ReLU, mỗi yếu tố dự đoán tiếp theo sẽ đưa đến lớp chuẩn hóa và lớp bỏ học (dropout), và một lớp tuyến tính bổ sung để chiếu các trạng thái ẩn vào chuỗi đầu ra. Các đoạn sau sẽ mô tả chi tiết về ba yếu tố dự đoán tương ứng:

*** Dự đoán trường độ (duration predictor)**

Bộ dự đoán trường độ lấy chuỗi ẩn âm vị làm đầu vào và dự đoán trường độ của mỗi âm vị, biểu thị số khung phổ Mel tương ứng với âm vị này và được chuyển đổi thành miền logarit để dễ dự đoán. Bộ dự báo trường độ được tối ưu hóa với tổn thất sai số bình phương trung bình (MSE), lấy trường độ trích xuất làm mục tiêu huấn luyện. Thay vì trích xuất thời lượng âm vị bằng mô hình TTS tự động hồi quy được huấn luyện trước trong FastSpeech, sử dụng công cụ căn chỉnh cưỡng bức (MFA) của Montreal [36] để trích xuất trường độ âm vị, nhằm cải thiện độ chính xác của căn chỉnh và do đó giảm khoảng cách thông tin giữa đầu vào và đầu ra của mô hình.

*** Dự đoán cao độ (pitch predictor)**

Các hệ thống TTS dựa trên mạng nơ-ron trước đây với dự đoán cao độ [110]; [25] thường dự đoán trực tiếp đường bao cao độ. Tuy nhiên, do sự biến đổi lớn của cao độ của âm thanh gốc, phân phối của các giá trị cao độ dự đoán rất khác với phân phối âm thanh gốc. Để dự đoán tốt hơn các biến thể trong đường bao cao độ, sử dụng biến đổi Wavelet liên tục (CWT) để phân tách chuỗi cao độ liên tục thành biểu đồ cao độ và lấy biểu đồ cao độ làm mục tiêu huấn luyện công cụ dự báo cao độ được tối ưu hóa với tổn thất MSE. Trong suy diễn, bộ dự đoán cao độ dự đoán biểu đồ cao độ, tiếp tục được chuyển đổi trở lại thành đường bao cao độ bằng cách sử dụng biến đổi Wavelet liên tục nghịch đảo (iCWT). Để lấy đường bao cao độ làm đầu vào trong cả huấn luyện và suy diễn, lượng hóa cao độ F0 (giá trị âm thanh gốc/âm thanh dự đoán tương ứng cho huấn luyện/suy diễn) của mỗi khung tiếng nói thành 256 giá trị có thể có trong log-scale và tiếp tục chuyển nó thành vector nhúng cao độ và thêm nó vào chuỗi ẩn đã mở rộng (Hình 28).



Hình 28: Chi tiết trong công cụ dự đoán cao độ. CWT và iCWT lần lượt biểu thị biến đổi wavelet liên tục và biến đổi wavelet nghịch đảo [30]

*** Dự đoán cường độ (energy predictor)**

Tính toán định mức L2 của biên độ của mỗi khung biến đổi Fourier thời gian ngắn (STFT) làm cường độ. Sau đó, lượng tử hóa cường độ của mỗi khung tiếng nói thành 256 giá trị có thể đơn nhất, mã hóa nó thành vector nhúng cường độ *e* và thêm nó vào chuỗi ẩn mở rộng tương tự như cao độ. Sử dụng công cụ dự báo cường độ để dự đoán các giá trị ban đầu của cường độ thay vì các giá trị lượng tử hóa và tối ưu hóa công cụ dự báo cường độ với hàm mất mát MSE.

*** Kiến trúc tổng hợp tiếng nói từ văn bản không tự hồi quy cơ sở (Non-autoregressive TTS)**

Luận án xây dựng kiến trúc bộ TTS không hồi quy cơ sở dựa trên FastSpeech2 [30], là một trong những mô hình tổng hợp đơn người nói phổ biến nhất trong TTS không tự động hồi quy. Mô hình bao gồm ba phần; một bộ mã hóa âm vị, một bộ giải mã phổ Mel và một bộ thích nghi phương sai (Variance adaptor). Bộ mã hóa âm vị chuyển một chuỗi vector biểu diễn đặc trưng giọng nói thành chuỗi âm vị ẩn. Sau đó, bộ thích nghi phương sai dự đoán các thông tin khác nhau trong tiếng nói, chẳng hạn như cao độ và cường độ ở cấp âm vị.

Bộ mã hóa văn bản $encoder^{txt}(\cdot)$ mã hóa chuỗi văn bản có độ dài n là $W = \{w_i \in V | i = 1, \dots, n\}$ với một từ vựng V thành một dãy các vectơ ẩn $\boldsymbol{h} = \{h_i \in \mathbb{R}^D | i = 1, \dots, n\}$ với D chiều như sau:

$$\boldsymbol{h} = Encoder^{txt}(W) \tag{3.7}$$

Để cho phép mô hình tạo âm thanh người nói, một bộ mã hóa người nói $Encoder^{spk}(\cdot)$ được thêm vào để mã hóa đặc trưng âm học của một vector biểu diễn đặc trưng giọng nói toàn cục $\boldsymbol{e} \in \mathbb{R}^{D_e}$ với chiều D_e như sau:

$$\boldsymbol{e} = Encoder^{spk}(\boldsymbol{O}) \tag{3.8}$$

Bộ mã hóa $Decoder(\cdot)$ suy diễn một chuỗi ẩn có độ dài T mang đặc trưng âm học $\boldsymbol{O}$ của chuỗi đích $\boldsymbol{O} = \{\boldsymbol{o}_t \in \mathbb{R}^{D_a} | \, t = 1, \dots, T\}$ với các đặc trưng có chiều D_a:

$$\boldsymbol{o}_t = Decoder(\boldsymbol{o}_{t-1}, \boldsymbol{h}) \tag{3.9}$$

Vectơ biểu diễn đặc trưng giọng nói của người nói $\boldsymbol{e}$ được nối với mọi vectơ văn bản được mã hóa $\boldsymbol{H}$ trên toàn bộ chuỗi. Nghĩa là, chức năng bộ Decoder đa người nói $Decoder^{mlt}$ được mở rộng từ công thức trên như sau:

$$\boldsymbol{o}_t = Decoder^{mlt}(\boldsymbol{o}_{t-1}, \mathrm{Cat}(\boldsymbol{h}, \boldsymbol{e})) \tag{3.10}$$

trong đó $\mathrm{Cat}(\cdot)$ là hàm nối giữa $\boldsymbol{h}$ và $\boldsymbol{e}$. Bỏ qua giải thích về bộ phát âm, thường được huấn luyện riêng để tạo dạng sóng từ biểu đồ phổ Mel.

Hơn nữa, bộ thích nghi phương sai dự đoán trường độ của mỗi âm vị để điều chỉnh độ dài của chuỗi âm vị ẩn thành độ dài của khung tiếng nói. Cuối cùng, bộ giải mã phổ Mel chuyển đổi chuỗi ẩn âm vị được quy định về độ dài thành chuỗi phổ Mel. Cả bộ mã hóa âm vị và bộ giải mã phổ Mel đều bao gồm các khối Feed-Forward Transformer (khối FFT) dựa trên kiến trúc Transformer [35].

Như vậy có thể thấy vai trò của bộ thích nghi phương sai trong việc huấn luyện và trong suy diễn các đặc trưng tiếng nói như cường độ (energy), cao độ (pitch), trường độ (duration). Đây là hợp phần quan trọng trong mô hình tổng hợp thích nghi giúp hệ thống học và điều chỉnh được các đặc trưng giọng, tổng hợp được các tham số đầu ra một cách tùy ý.

3.3.1.2. Biểu diễn đặc trưng giọng nói và phong cách nói

Các nghiên cứu trước đây về vector biểu diễn đặc trưng giọng nói đã đề xuất nhiều dựa trên các phương pháp tiếp cận GMM và DNN, để trích xuất vectơ có chiều cố định. Nghiên cứu [111] đã đề xuất một mô hình i-vector, được tính toán bằng cách sử dụng hỗn hợp mô hình GMM (Gaussian Mixture Model) và mô hình UBM (Universal Background Model[3]). Vào năm 2018, [112] đã đề xuất x-vector, cho phép trích xuất các đặc trưng bằng cách sử dụng mô hình DNN được huấn luyện trực tiếp để phân biệt người nói. Nghiên cứu cho thấy rằng đặc trưng x-vector vượt trội hơn đáng kể so với i-vector trong nhiệm vụ nhận dạng người nói khoảng 3% EER. Trong các hệ thống nhận dạng người nói x-vector thường được tính toán ở cấp độ câu. Để làm được điều đó, đầu tiên các vectơ được trích xuất ở cấp độ khung (frame level), sau đó chúng được tổng hợp trên tất cả các khung tiếng nói (frame) để tạo ra một vectơ có chiều cố định. Quá trình này dẫn đến độ nhạy đối với các vectơ ngoại lai. Nếu có tiếng ồn trong giọng nói, các vectơ đầu ra có thể khác nhiều so với vectơ được tính toán trong môi trường sạch. Để giảm thiểu vấn đề này, [113] đã đề xuất mô hình Thin-ResNet sử dụng CNN và NetVLAD (hoặc GhostVLAD) để tạo ra các vectơ có độ dài cố định, vượt trội hơn hiệu suất của i-vectơ và x-vectơ. Đây cũng chính là ý tưởng tạo ra vector biểu diễn đặc trưng giọng nói ở cấp độ khung và cũng học cách tổng hợp các vectơ theo thời gian.

*** Kiến trúc vector nhúng**

Hệ thống x-vector: Kiến trúc dựa trên hệ thống vector nhúng DNN được mô tả trong [114]. Mô-đun của được xây dựng bởi bộ công cụ Kaldi với dữ liệu huấn luyện từ VoxCeleb. 5 lớp đầu tiên hoạt động và trượt trên khung tiếng nói với ngữ cảnh thời gian ngắn. Khung tiếng nói hiện tại có nhãn là *t* ở giữa các số xung quanh của ngữ cảnh khác. Lớp càng cao, nó càng học được nhiều ngữ cảnh trên khung tiếng nói gốc.

Thin ResNet: ResNet đã sửa đổi mã hóa các phổ 2D, tiếp theo là một lớp NetVLAD để tổng hợp các đặc trưng để tạo ra một bộ mô tả đầu ra có chiều cố định. Thin-ResNet bao gồm một số khối là một chồng phép toán tích chập với

[3] UBM là một mô hình thường sử dụng cho xác thực sinh trắc học và nhận dạng người nói để đại diện cho đặc trưng chung, độc lập với con người nói để so sánh với các mô hình về đặc trưng của từng người khi đưa ra quyết định chấp nhận hoặc từ chối.

kích thước nhân là 1 và 3. Kiến trúc có 3 triệu tham số so với ResNet-34 tiêu chuẩn (22 triệu) [115] bởi vì trong mỗi khối dư, mô hình loại bỏ số lượng kênh để giảm kích thước và tăng tốc độ xử lý.

*** Vector biểu diễn đặc trưng giọng nói cho xác thực người nói (speaker verification)**

Xác thực hoặc định danh người nói (Speaker verification) gồm hai thành phần: Thứ nhất là một bộ vector biểu diễn đặc trưng giọng nói được huấn luyện để tạo ra vectơ người nói có chiều cố định ở các cấp độ câu được sử dụng để đại diện cho đặc trưng của người nói. Thứ hai, một bộ phân loại để phân biệt (định danh) các vectơ người nói với nhau. Trong bài báo [116] đã đề xuất một mô hình tăng cường khả năng định danh người nói trong môi trường nhiễu sử dụng bộ phân lớp DNN. Nghiên cứu đề xuất: 1) Đánh giá hiệu quả của các mô hình vector biểu diễn đặc trưng giọng nói khác nhau và bộ phân loại (classifier) trong các điều kiện khác nhau; và 2) đề xuất một bộ phân loại mạng nơ-ron bên trên vector nhúng và huấn luyện nó với dữ liệu tăng cường. Kết quả thử nghiệm chỉ ra rằng pipeline được đề xuất hoạt động tốt hơn pipeline truyền thống **5% F1** trên bộ thử nghiệm sạch và **9% F1** trên bộ thử nghiệm nhiễu (Bảng 10).

Bảng 10: Kết quả kết hợp hệ thống trích xuất và phân lớp trong hệ thống xác minh người nói [116]

	F1(%)					EER(%)				
	Clean	SNR 20	SNR 15	SRN 10	SNR 5	Clean	SNR 20	SNR 15	SRN 10	SNR 5
x-vector+LDA+PLDA ⋆	89.0	88.4	88.2	86.6	84.6	11.4	12.3	12.6	14.2	15.8
ThinResNet+LDA+PLDA	89.3	88.4	87.4	86.1	84.8	11.5	11.9	12.5	14.1	15.6
x-vector+LDA+DNN	92.0	92.0	91.0	90.0	89.0	9.1	10	10.1	11.1	12.1
x-vector+DNN †	93.0	92.0	91.0	91.0	90.0	9.5	9.3	10.5	10.8	11.0
ThinResNet+LDA+DNN	**95.0**	**95.0**	**94.0**	**94.0**	92.0	**5.6**	**6.2**	6.5	**6.8**	**7.6**
ThinResNet+DNN †	**95.0**	**95.0**	**94.0**	**94.0**	**93.0**	**5.6**	**6.2**	**6.1**	6.9	7.7

⋆ notation indicates that (x-vector+LDP+PLDA) is the baseline model.
† notation means that the size of the vector is 512-d, and the others have size 150-d.

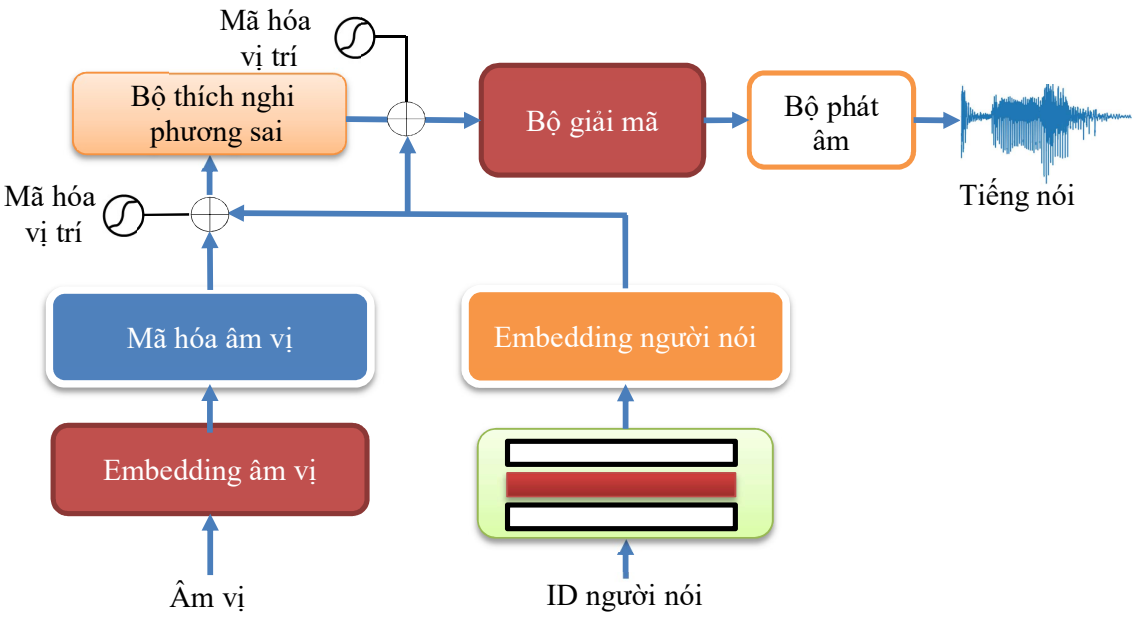
Nghiên cứu này đã nghiên cứu hiệu quả của các bộ trích xuất và phân loại khác nhau đối với nhiệm vụ xác minh người nói trong các môi trường nhiễu khác nhau. Các thử nghiệm cho thấy khi giảm tỷ lệ SNR, độ chính xác của hệ thống cũng giảm xuống 4,6% F1, cho thấy rằng các vector nhúng nhạy cảm với môi trường nhiễu, khoảng cách giữa tín hiệu sạch và nhiễu. Để giải quyết vấn đề, luận án cũng đề xuất bộ phân loại DNN, giúp rút ngắn khoảng cách hiệu suất giữa tín hiệu sạch và nhiễu. Khi kết hợp bộ phân loại DNN với bộ trích xuất vector nhúng dựa trên

ThinResNet, hệ thống đã đạt được hiệu suất tốt nhất với **94% F1**, trái ngược với **86,1% F1** sử dụng bộ phân loại PLDA trong cùng một môi trường.

Như vậy có thể thấy rằng vector biểu diễn đặc trưng giọng nói là yếu tố quan trọng nhất trong định danh người nói, nó bao hàm các đặc trưng riêng để phân biệt của những người nói khác nhau. Tuy nhiên, cả Thin-ResNet và x-vector đều được huấn luyện trước cho mục đích nhận dạng giọng nói, nên các mô hình chỉ tập trung vào việc phân biệt giữa các người nói khác nhau và có thể bỏ qua những thông tin không có ích cho mục đích này, dẫn đến kết quả không tốt cho mô hình TTS [117]. Do đó, để tạo ra một đại diện người nói lý tưởng cho mô hình TTS, cần mô hình hóa tất cả các đặc trưng của người nói mà không nhất thiết phải phân biệt rõ ràng giữa từng người nói. Do vậy, phần sau đây sẽ đề xuất một vector biểu diễn đặc trưng giọng nói một cách đầy đủ phục vụ cho tổng hợp thích nghi giọng nói.

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

Bộ mã hóa là một thành phần cho phép mã hóa các chuỗi độ dài khác nhau thành các vector biểu diễn có số chiều cố định. Trong mô hình TTS đa người nói cơ bản [118] [5] [119], trong bộ mã hóa của người nói, một thành phần quan trọng là vector nhúng của người nói, biểu diễn tín hiệu giọng của người nói dưới dạng vector đặc trưng.



Hình 29: Sơ đồ kiến trúc của hệ thống tổng hợp giọng nói đa người nói cơ bản sử dụng vector biểu diễn đặc trưng giọng nói cơ bản

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. Để làm điều đó, hệ thống xử lý tiếng nói phải trước tiên chuyển đổi mỗi đoạn âm thanh có độ dài khác nhau thành một vector có độ dài cố định biểu thị danh tính của người nói, được gọi là vector nhúng của người nói, và thực hiện phân cụm dựa trên các vector này. Vector nhúng của người nói cũng được sử dụng rộng rãi trong các tác vụ xử lý tiếng nói như xác định danh tính người nói, phân tích đa người nói, điều chỉnh tiếng nói và tổng hợp ngôn ngữ. Phương pháp truyền thống thường sử dụng một module nhúng để trích xuất vector đại diện. Có thể mô tả toán học phương pháp truyền thống như công thức sau đây:

$$e_{spk} = Encoder(Speaker\_ID) \quad (3.11)$$

Hệ thống tổng hợp tiếng nói đa người dùng vector nhúng người nói cơ bản biểu diễn trong Hình 29. Tuy nhiên, phương pháp cơ bản này không thể nắm bắt được các đặc trưng 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. Mô hình của FastSpeech2 [30] được xây dựng dựa trên cấu trúc Transformer với mạng tự chú ý và mạng truyền xuôi trong mỗi khối Transformer. Cả việc nhân ma trận trong quá trình tự chú ý và mạng truyền xuôi hai lớp đều tốn nhiều tham số, không hiệu quả cho việc thích nghi.

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. Ví dụ: Trong nghiên cứu [45] [12] sử dụng phương pháp thích nghi giọng nói bằng bộ mã hóa người nói để trích xuất một vector hoặc một chuỗi vector để biểu diễn đặc trưng của chuỗi tiếng nói hoặc chỉ tinh chỉnh vector biểu diễn đặc trưng giọng nói nên không thể đạt đủ chất lượng. Trong nghiên cứu [120], đã giới thiệu vector đại diện phong cách tổng quát GST (Global style token). Vector này được huấn luyện không có nhãn để mô hình hóa đặc trưng âm học và kiểm soát quá trình tổng hợp theo nhiều phong cách nói khác nhau, bao gồm cả tốc độ, cách nói và tính độc lập của văn bản. Phương pháp này đôi khi thể hiện sự chuyển đổi phong cách nói thành công. Tuy nhiên, vì huấn luyện xen kẽ chỉ đảm bảo tiếp xúc

với một số kết hợp có thể có của các lớp phong cách nói trong quá trình huấn luyện, nên nó có thể dẫn đến mất khả năng biểu diễn phong cách của người nói. Một nghiên cứu khác [121] đã sử dụng bộ chuẩn hóa lớp thích nghi theo phong cách nói SALN (Style-Adaptive Layer Normalization) để điều chỉnh mức tăng và độ lệch của kiểu nhập văn bản với phong cách nói có được từ âm thanh ngắn dùng làm mẫu tham khảo, cho phép mô tả chung về vectơ phong cách nói đại diện cho phong cách nói của người nói từ đầu vào âm thanh tham chiếu được mã hóa bởi bộ mã hóa phong cách.

Trong luận án này 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 *stv* cố định từ phổ Mel, như được mô tả trong Hình 30. 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:

$$\hat{y} = G(x, stv; \theta) \tag{3.12}$$

trong đó, vector biểu diễn đặc trưng giọng 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:

$$stv = EMV(Mel\_X) \tag{3.13}$$

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. 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 giọng nói. Sau đó, Bộ thích nghi phương sai (Variance adaptor) 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, bộ phát âm sẽ chuyển đổi các phổ Mel này thành tín hiệu tiếng nói. Kiến trúc tổng thể được mô tả trong Hình 24.

Chức năng và kiến trúc chi tiết của các mô-đun EMV đề xuất sẽ được trình bày dưới đây. Trong khối mô-đun EMV này, cấu trúc 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:

1) Tại khối **Encoder Feature**, đầu tiên tiếng nói đầu vào chuyển đổi thành phổ Mel được đưa đến lớp kết nối đầy đủ (fully-connected FC) và các hàm kích hoạt Mish (hàm kích hoạt tự kiểm soát không đơn điệu) để chuyển đổi mỗi khung của phổ Mel thành các chuỗi ẩn. Hàm kích hoạt Mish được lựa chọn do đạt được hiệu năng vượt trội hơn ReLU và Swish trong nhiều thử nghiệm [122] và do Mish thể hiện được sự ổn định khi thay đổi các siêu tham số như kích thước lô (batch size), độ sâu của mô hình, tỉ lệ học, bộ kiểm soát hóa (regularizer),… Hàm kích hoạt Mish được định nghĩa bằng công thức toán học như sau:

$$Mish(x) = x * \tanh(softplus(x)) \tag{3.14}$$

trong đó x là đầu vào của hàm kích hoạt Mish và $\tanh$ là hàm tanh thông thường. Hàm $softplus(.)$ được định nghĩa bởi công thức:

$$softplus(x) = \log(1 + \exp(x)) \tag{3.15}$$

Mish có một số đặc điểm quan trọng: 1) Đồng biến: Hàm Mish có tính chất đồng biến, tức là đầu ra tăng khi đầu vào tăng và giảm khi đầu vào giảm. Khả năng chịu được đầu vào lớn: Đặc tính tanh trong hàm Mish giúp hàm này chịu được đầu vào lớn hơn so với các hàm kích hoạt khác như sigmoid hoặc ReLU. Điều này giúp tránh hiện tượng "triệt tiêu gradient" và cho phép mạng nơ-ron học được hiệu quả khi xử lý đầu vào lớn; 2) Hình dạng giống hàm ReLU: Hàm Mish có hình dạng gần giống với hàm ReLU ở các vùng dương, tạo điều kiện cho học nhanh chóng và giúp tránh vấn đề "dead neurons" (nơron không hoạt động) mà thường xảy ra với ReLU; 3) Khả năng biến đổi dữ liệu: Hàm Mish có tính chất phi tuyến, cho phép nó biến đổi dữ liệu một cách linh hoạt và phức tạp hơn so với các hàm kích hoạt tuyến tính như sigmoid và tanh.

Sau đó kết quả sẽ chuyển qua hai lớp kết nối đầy đủ (FC). Mục đích của khối **Encoder Feature** là chuyển đổi đặc trưng đầu vào thành đặc trưng bộ mã hóa.

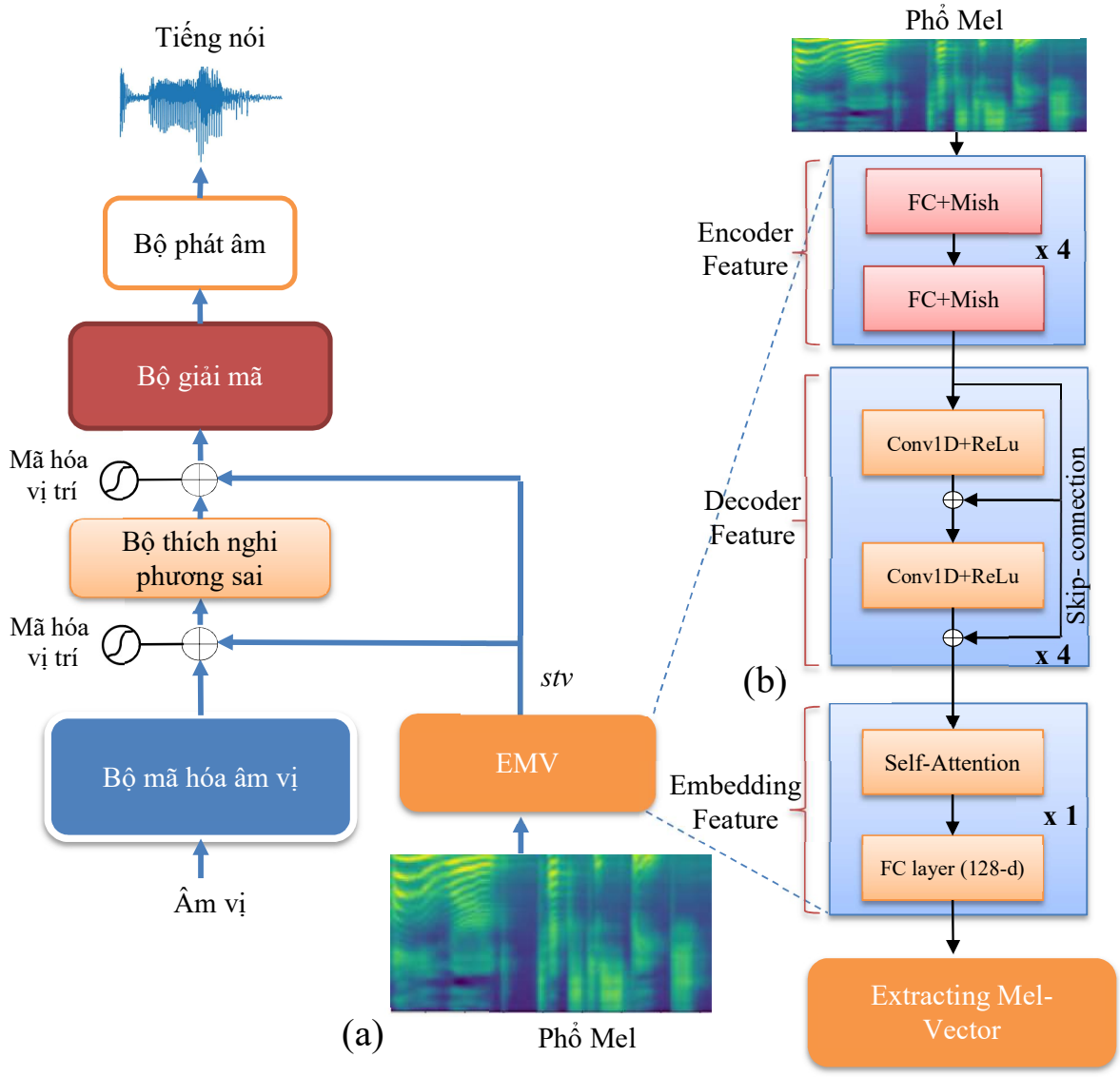
2) Tiếp theo, vector này sẽ được đưa qua khối giải mã đặc trưng **Decoder Feature**. Bằng cách sử dụng Conv1D +ReLu với kết nối dư để nắm bắt (capture)

được chuỗi thông tin từ tiếng nói mẫu đã cho, mục tiêu của mô-đun này sẽ chuyển đổi đầu ra của **Encoder Feature** thành đặc trưng bộ giải mã. Ngoài ra, kết nối tắt (skip-connection) cũng tích hợp để sử dụng các đặc trưng có giá trị của các khối trước đó và giải quyết được vấn đề triệt tiêu gradient.

3) Cuối cùng, đầu ra của Decoder Feature sẽ được chuyển sang mô-đun **Embedding Feature**, mô-đun này có thành phần tự chú ý tập trung vào các đặc trưng quan trọng với kết nối dư cộng lớp affine để mã hóa các thông tin một cách tổng quát toàn bộ đặc trưng giọng nói và phong cách nói. Áp dụng nó ở cấp khung 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 phong cách nói một chiều *stv*. Như vậy mô-đun này sẽ tạo ra một vectơ đại diện cho phổ Mel và vectơ này sẽ thêm thông tin đặc trưng giọng nói vào mô hình chuyển văn bản thành giọng nói. Vectơ đại diện EMV sẽ điều khiển đầu ra của mô hình TTS và tạo ra giọng nói tổng hợp tương tự như vectơ đầu vào. Chi tiết kiến trúc của EMV được thể hiện trong Bảng 11.

Bảng 11: Kiến trúc Trích xuất Mel-Vector (EMV)

Lớp	**Đầu vào x Đầu ra**
Mel	T x80
FC + Mish	T x128
FC + Mish	T x128
ConV1D + ReLu	128x T (chuyển vị)
ConV1D + ReLu	128x T
Self − Attention	T x128 (chuyển vị)
FC layer	1x128



Hình 30: a) Sơ đồ kiến trúc của mô hình dựa trên thích nghi Multi-TTS tiếng Việt với mô-đun Trích xuất Mel-vector (EMV) và b) Cấu trúc chi tiết của mô-đun EMV

3.3.3. Hàm mất mát huấn luyện

Mô hình đề xuất được huấn luyện với phương pháp học đa nhiệm (được biết đến như là joint-training hoặc multi-stask learning) với việc huấn luyện nhiều hàm mất mát đồng thời. Hàm mất mát của mô hình là tổng của hàm mất mát L1 dự đoán phổ Mel từ vector có điều kiện EMV và các hàm mất mát L2 dự đoán trường độ và các hàm mất mát dự đoán bổ sung của các đặc trưng âm thanh như hàm mất mát dự đoán cao độ và cường độ. Hàm mất mát tổng quát huấn luyện của mô hình như sau:

$$L_{final} = L_{mel} + L_{duration} + L_{pich} + L_{energy} \quad (3.16)$$

1. Hàm mất mát L_{mel}: Khoảng cách giữa phổ Mel dự đoán và phổ Mel mục tiêu được mô tả như sau:

$$L_{mel} = \mathbb{E}[\| \hat{y} - y \|_1] \quad (3.17)$$

2. 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:

$$L_{duration} = ||d - \hat{d}||_2^2, L_{pich} = ||p - \hat{p}||_2^2, L_{energy} = ||\varepsilon - \hat{\varepsilon}||_2^2 \quad (3.18)$$

3.3.4. Thử nghiệm đánh giá và kết quả

3.3.4.1. Thử nghiệm đánh giá

Để đánh giá các hệ thống TTS đa người nói, phần này sử dụng mô hình tổng hợp giọng nói đa người nói hiện đại như FastSpeech2 [30] và bộ phát âm HifiGAN làm mô hình cơ sở (baseline model), như được mô tả trong Hình 28. Kiến trúc FastSpeech2 bao gồm các phần chính: 1) Bộ mã hóa âm vị chuyển đổi chuỗi nhúng âm vị thành chuỗi ẩn âm vị; 2) Bộ thích nghi phương sai nhằm mục đích thêm thông tin phương sai như thời lượng, cao độ và cường độ vào chuỗi ẩn âm vị; 3) Bộ giải mã phổ Mel chuyển đổi song song thứ tự ẩn được điều chỉnh thành thứ tự phổ Mel; 4) Bộ phát âm biến đổi từ phổ Mel sang dạng sóng.

Công cụ MFA (Montreal Forced Align) được sử dụng để trích xuất trường độ âm vị tiếng Việt [36]. Quá trình đa người mã hóa cơ bản lấy ID của người nói làm đầu vào để mã hóa người nói tương ứng thành các vectơ đặc trưng của người nói. Variance adapter đoán trường độ, cao độ và cường độ được tối ưu hóa với hàm sai số bình phương trung bình (MSE) để cực tiểu hóa sai số đầu ra của mô hình với âm thanh gốc. Kỹ thuật phân phối dữ liệu cũng được sử dụng mặc định trong các mô hình TTS đề xuất và cơ sở để giữ các tham số đặc tính tiếng nói thích nghi [CT2].

Trong mô hình Multi-TTS dựa trên thích nghi đề xuất như được mô tả trong Hình 30, sẽ thay thế vector biểu diễn đặc trưng giọng nói cơ sở bằng mô-đun EMV để mã hóa các đặc trưng của người nói trực tiếp từ phổ Mel.

Tiến hành hai thử nghiệm để đánh giá hiệu suất của các mô hình cho tiếng Việt:

Bộ dữ liệu: Một bộ dữ liệu tiếng Việt đa người nói có nhãn đã được sử dụng để đánh giá mô hình: tổng cộng 54 người nói, bao gồm 26 giọng nam, 28 giọng nữ và phương ngữ Bắc-Nam, với mỗi người nói đọc khoảng 500 câu. 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 TTS.

Thử nghiệm 1: Đánh giá chất lượng âm thanh tổng hợp được tạo ra bởi mô hình Multi-TTS cơ sở (đã sử dụng kỹ thuật thích nghi bằng phân phối dữ liệu [CT1]) và mô hình đề xuất Multi-TTS dựa trên thích nghi. Để ước tính lượng âm thanh tối thiểu để huấn luyện mô hình TTS dựa trên thích nghi, bộ dữ liệu 16 phút không cần dùng để đánh giá. Âm thanh gốc cũng được thêm vào để đánh giá nhằm đảm bảo tính khách quan. Kết quả đánh giá được trình bày trong Bảng 11.

Thang đo MOS được sử dụng để đánh giá chất lượng giọng nói do hệ thống tạo ra. Âm thanh được tổng hợp từ mô hình TTS đa người nói cơ sở và mô hình TTS nhiều giọng nói thích nghi sẽ được 26 người nghe đánh giá với độ tin tưởng 95%. Mỗi người nghe sẽ nghe một bộ gồm 120 câu âm thanh được trộn lẫn giữa âm thanh gốc, âm thanh được tạo bởi mô hình cơ sở và thích nghi.

Ngoài ra, chỉ số WER cũng sử dụng để xác thực tính dễ hiểu của giọng nói được tạo. Sử dụng mô hình Wave2vec ASR đã được huấn luyện trước để tính tỷ lệ lỗi từ (WER) [66]. Cả MCD và WER đều không phải là thước đo tuyệt đối để đánh giá chất lượng giọng nói, vì vậy chỉ sử dụng chúng để so sánh tương đối.

Thử nghiệm 2: Đánh giá mức độ giống nhau của âm thanh tổng hợp được tạo bởi 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. Phân tích phổ Mel của âm thanh tổng hợp được so sánh với âm thanh gốc để thấy sự khác biệt giữa các mẫu. Kết quả đánh giá được thể hiện trong Bảng 12.

Độ méo Mel-Cepstral (MCD) được sử dụng để đo mức độ khác nhau của hai chuỗi Mel-cepstral nhằm đánh giá hiệu suất thích nghi của giọng tổng hợp. Chỉ số SIM cũng được sử dụng để so sánh sự giống nhau của âm thanh tổng hợp và âm thanh gốc. 26 người nghe có 4 lựa chọn để đánh giá từng cặp âm thanh với độ tin tưởng 95%. 90 cặp âm thanh được sử dụng để đánh giá, mỗi mô hình tạo ra 30 câu âm thanh (được trộn ngẫu nhiên giữa các âm thanh được tạo bởi các mô hình thích nghi đề xuất, mô hình cơ bản và âm thanh gốc).

3.3.4.2. Kết quả

Bảng 12: Bảng đánh giá chất lượng giữa mô hình Multi-TTS cơ sở (sử dụng vector biểu diễn đặc trưng giọng nói cơ bản) và Mô hình Multi-TTS dựa trên thích nghi (sử dụng mô-đun EMV) với độ tin tưởng 95%

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 (Groundtruth)	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

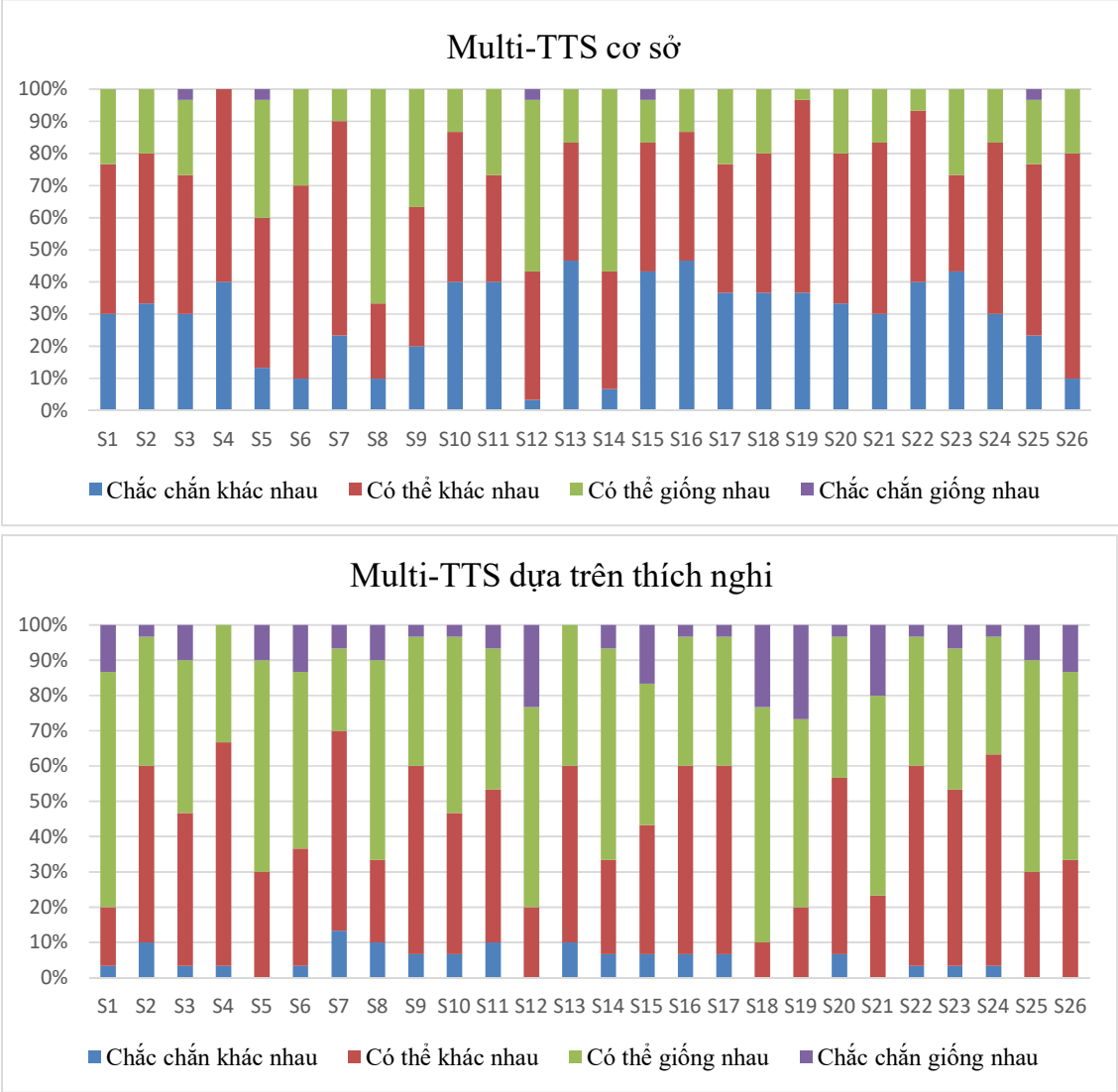
Trên Bảng 12 chỉ ra rằng, chỉ cần 1 phút dữ liệu tiếng nói đích thì mô hình Multi-TTS dựa trên thích nghi đã có thể tổng hợp được âm thanh có điểm MOS đạt 3.81 so với điểm 4.6 của người nói. Điểm số này cao hơn hẳn so với MOS sinh từ mô hình Multi-TTS cơ sở (sử dụng 16 phút giọng đích). Điểm số WER cũng thể hiện rằng mô hình Multi-TTS dựa trên thích nghi tổng hợp giọng tốt hơn mô hình Multi-TTS cơ sở.

Bảng 13: 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 với độ tin tưởng 95%

Mô hình thử nghiệm (1 phút dữ liệu thích nghi)	**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**

Bảng 13, 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 25% so với điểm SIM 1.96 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 mô hình Multi-TTS cơ sở.



Hình 31: So sánh sự tương đồng của của mô hình Multi-TTS cơ sở (trên) và mô hình đề xuất (dưới) trên tất cả các cặp câu đánh giá

Tiến hành phân tích đánh giá SIM theo phương pháp [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 31, 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 đề xuất và mô hình cơ sở là rất rõ ràng.

Bảng 14: Bảng phân tích ANOVA về điểm đánh giá tương đồng giữa mô hình Multi-TTS cơ sở và mô hình đề xuất

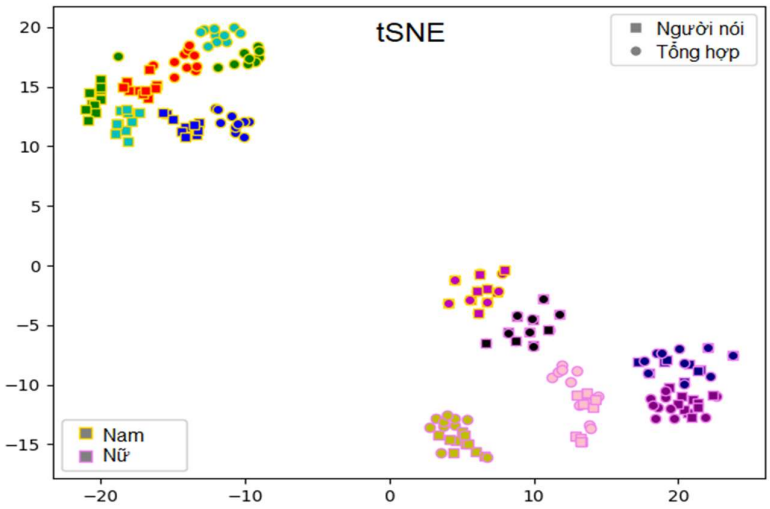
Mô hình	**Nguồn phương sai**	*Tổng bình phương (SS)*	*Bậc tự do (df)*	*Bình phương trung bình (MS)*	*Giá trị thống kê (F)*	*Giá trị p*	*F tới hạn*
Multi-TTS Cơ sở	Giữa các nhóm	57.237	25	2.289	4.636	1.55E-12	1.521
	Trong các nhóm	372.367	754	0.494			
Multi-TTS dựa trên thích nghi	Giữa các nhóm	53.888	25	2.156	4.608	1.97E-12	1.521
	Trong các nhóm	352.7	754	0.468			

Phân tích ANOVA một chiều (phụ thuộc một biến duy nhất là mô hình TTS) trong đánh giá độ tương đồng SIM giữa các mô hình cho kết quả ở Bảng 13. Các giả thuyết cho phân tích ANOVA ban đầu như sau :

- **Giả thuyết không(H_0):** Tất cả các phương pháp có chất lượng tương đồng bằng nhau (không có sự khác biệt).
- **Giả thuyết thay thế(H_1):** Có sự khác biệt ít nhất một cặp phương pháp.

Qua kết quả phân tích 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.

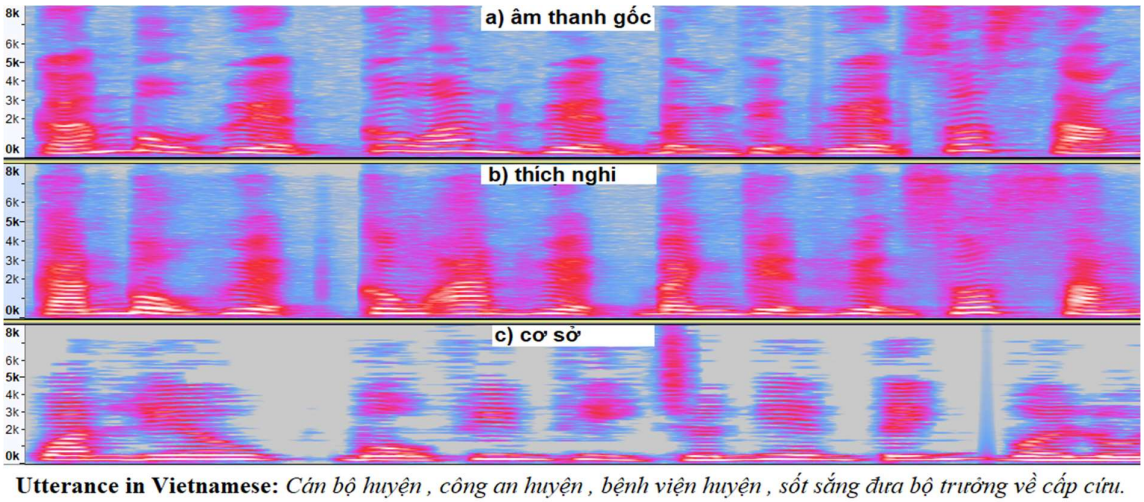
Trực quan hóa các vectơ kiểu để hiểu rõ hơn về hiệu quả của vectơ EMV trong việc mã hóa các kiểu của từng người nói. Trong Hình 32, minh họa phép chiếu t-SNE của các vectơ phong cách nói giữa giọng người nói và giọng tổng hợp tương ứng. Sử dụng giọng nói của 10 người nói (5 nam và 5 nữ), có thể thấy 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 cụm rõ ràng giữa từng người nói với nhau và các điểm tổng hợp lân cận với giọng người nói.



Hình 32: Hình ảnh t-SNE phân bố của giọng nói của người nói và giọng nói tổng hợp (sử dụng EMV)

Các giọng nói gần như được phân biệt giới tính rõ ràng trong hình ảnh t-SNE, tất cả các giọng nói nữ xuất hiện ở bên phải và tất cả các giọng nói nam xuất hiện ở bên trái. Các điểm biểu diễn giọng tổng hợp và giọng người nói tương đối gần nhau, cho thấy rằng bộ mã hóa EMV đã học cách biểu diễn tốt không gian đặc trưng của từng người nói.

Ngoài ra, Hình 33 mô tả phổ Mel của âm thanh được tổng hợp từ mô hình Multi-TTS cơ sở, mô hình Multi-TTS dựa trên thích nghi và giọng người nói. Có thể thấy, chỉ với 1 phút dữ liệu thích nghi, phổ âm thanh tổng hợp khá giống với phổ âm thanh thật và hoàn toàn khác với phổ âm thanh được tạo ra từ mô hình Multi-TTS cơ sở.



Hình 33: So sánh phổ Mel của a) âm thanh gốc, b) âm thanh được tạo ra từ mô hình thích nghi và c) âm thanh được tạo ra từ mô hình cơ sở với mẫu giọng nói thích nghi dài 1 phút

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].

Trong Chương 3 cũng đề xuất một mô hình thích nghi nhằm nâng cao chất lượng cho hệ thống Multi-TTS tiếng Việt với kiến trúc mô-đun vector đặc trưng EMV để khắc phục các nhược điểm của các vector nhúng biểu diễn đặc trưng giọng nói cơ bản. Mô hình được đề xuất 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 (x-vector, d-vector, Thin ResNet), 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)

Như đã trình bày ở Chương 3, các kỹ thuật tổng hợp tiếng nói dựa trên thích nghi hiện nay dựa trên hai 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 giọng 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 mất mát 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.

Để giải quyết các tồn tại này và làm rõ câu hỏi nghiên cứu: *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? Lượng mẫu thích nghi tối thiểu cần bao nhiêu và kích thước mẫu sẽ ảnh hưởng đến mô hình như thế nào và ưu nhược điểm của phương pháp này so với các phương pháp đã trình bày trong Chương 3?* 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 là mô hình khử nhiễu khuếch tán phổ Mel (Mel-spectrogram denoiser) cho phép tổng hợp giọng mới mà không cần phải huấn luyện lại với chất lượng tốt hơn [CT1]; Tiến hành các thử nghiệm và phân tích kết quả đánh giá cho mô hình đề xuất; Ứng dụng thích nghi trong tổng hợp tiếng nói [CT7].

4.1. Các nghiên cứu liên quan

Các mô hình tổng hợp TTS trên có thể tổng hợp tốt với các giọng tồn tại trong dữ liệu huấn luyện hoặc nhìn thấy trong quá trình huấn luyện. Tuy nhiên, nếu không huấn luyện lại thì chất lượng tổng hợp vẫn còn là một thách thức lớn [12] [73].

Hiện nay, có hai kỹ thuật thích nghi chính thường được sử dụng đã được trình bày ở Chương 3: Một là tinh chỉnh toàn bộ hoặc một phần lớp mạng với dữ liệu thích nghi dựa trên một mô hình huấn luyện trước (đã được huấn luyện với lượng lớn dữ liệu); và hai là sử dụng một vector để mô hình hóa được đặc trưng của người nói với một lượng mẫu thích nghi nhỏ (Few-shot TTS) [5] [45]. Hai phương pháp này có ưu điểm là cho chất lượng tổng hợp tốt và giọng tổng hợp có độ tương đồng cao với giọng đích. Tuy nhiên, để học giọng mới vẫn cần tinh chỉnh một lượng mẫu của giọng đích để cập nhật các tham số mô hình và quá trình huấn luyện người dùng có trong quá trình huấn luyện trong thời gian dài (nhiều giờ thậm chí nhiều ngày). Điều này dẫn đến tốn tài nguyên tính toán và tốn thời gian để sinh giọng mới, hạn chế nhiều khả năng ứng dụng vào thực tế. Hướng tiếp cận để giải quyết vấn đề này chính là thích nghi trong hợp tiếng nói với mẫu nhỏ không huấn luyện lại mô hình (Zero-shot TTS).

4.1.1. Zero-shot TTS

Trong bài toán Zero-shot TTS, một vấn đề chính là có khoảng cách về độ tương đồng giữa những người nói đã xuất hiện và chưa xuất hiện trong quá trình huấn luyện mô hình. Zero-shot TTS đã được đề xuất đầu tiên bởi Sercan O. Arik [45]. Ý tưởng sử dụng bộ mã hóa người nói làm tín hiệu điều hòa đã được khám phá thêm bởi Ye Jia [75], trong đó đề xuất một mạng phân biệt người nói được huấn luyện trước bởi hàng ngàn người nói độc lập và sử dụng mạng này để trích xuất ra vector nhúng có chiều dài cố định của vài giây giọng nói thích nghi tuy nhiên phương pháp này được chứng minh là kém hiệu quả do vector nhúng trích xuất từ nhiệm vụ phân biệt người nói chỉ tốt cho nhiệm vụ phân biệt và không phù hợp cho việc tái tạo đầy đủ đặc trưng giọng nói. Nghiên cứu [12] đã cố gắng thu hẹp khoảng cách về chất lượng giữa người nói có trong tập huấn luyện và người nói không có trong tập huấn luyện trong mô hình Zero-shot TTS bằng cách sử dụng các vector nhúng nhiều thông tin hơn. Nghiên cứu này đã đề xuất một vector biểu diễn đặc trưng giọng nói là LDEs (Learnable Dictionary Encoding) dựa trên mạng nơ-ron để tăng cường độ tương đồng và độ tự nhiên của giọng và sử dụng x-vector nhằm tăng khả năng mở rộng cho nhiệm vụ xác minh người nói. Với việc sử dụng các phần

vector nhúng của người nói được chú ý để mã hóa phong cách nói tổng quát hơn thay vì âm thanh của người nói [102] cũng như các phương pháp giải mã khác nhau trong không gian âm học chẳng hạn như luồng tổng quát hóa đặc trưng [76], những nỗ lực tiếp theo đã được thực hiện nhằm thu hẹp khoảng cách về chất lượng giữa những giọng nói đã xuất hiện và chưa xuất hiện trong quá trình huấn luyện. Ngoài ra, việc thích nghi các mô hình Multi-TTS để sao chép giọng với ít giọng đích yêu cầu sự đa dạng (gồm số nhiều giọng chất lượng cao và nhiều thuộc tính giọng nói) của đa người nói trong dữ liệu huấn luyện và rất quan trọng để đạt được sự khái quát hóa cao trên tập người nói chưa từng xuất hiện [45]. Do đó, đây vẫn là các thách thức lớn và chưa phải là một nhiệm vụ được giải quyết hoàn toàn.

Ngoài ra, trong mô hình TTS thích nghi, các mô hình âm học trước đây chủ yếu sử dụng tổn thất đơn giản (ví dụ: L1 hoặc L2) để tái tạo lại các đặc trưng âm học. Tuy nhiên, sự tối ưu hóa này dựa trên các giả định phân phối không chính xác, dẫn đến đầu ra bị mờ và làm mịn quá mức. Một số phương pháp hiện tại cố gắng giải quyết vấn đề này bằng mạng đối thủ chung (GAN) [121], việc huấn luyện một mạng GAN tương đối phức tạp và đôi khi có thể thất bại do bộ phân biệt (Discriminator) không ổn định. Những vấn đề này cản trở tính tự nhiên của giọng nói tổng hợp. Các nghiên cứu gần đây cũng chứng minh tính hiệu quả của mô hình khuếch tán so với mô hình GAN trong nhiều tác vụ [123].

Đặt $S = \{(x_i^s, e_i^s, y_i^s)_{i=1}^{N_s} | x_i^s \in X^s, e_i^s \in E^s, y_i^s \in Y^s\}$ và

$$U = \{(x_j^u, e_j^u, y_j^u)_{j=1}^{N_u} | x_j^u \in X^u, e_j^u \in E^u, y_j^u \in Y^u$$

biểu diễn tương ứng lần lượt các tập dữ liệu xuất hiện và không xuất hiện trong quá trình huấn luyện, trong đó $x_i^s, x_j^u \in \mathbb{R}^D$ biểu diễn các đặc trưng ẩn của văn bản với *D* chiều trong không gian đặc trưng $\mathcal{X}$ có thể quan sát được bằng cách sử dụng một mô hình huấn luyện trước; $e_i^s, e_j^u \in \mathbb{R}^K$ biểu diễn vector nhúng đặc trưng giọng nói K chiều trong không gian $\mathcal{E}$;

$$Y^s = \{y_1^s, y_2^s, .. y_{C_s}^s\} \text{ và} \quad (4.1)$$

$$Y^u = \{y_1^u, y_2^u, .. y_{C_u}^u\} \quad (4.2)$$

biểu diễn âm thanh tương ứng của người nói đã xuất hiện và chưa xuất hiện trong không gian đặc trưng $\mathcal{Y}$ trong đó C_s, C_u là số người nói, $\mathcal{Y} = Y^s \cup Y^u$ biểu diễn hợp của hai tập người nói đã xuất hiện và chưa xuất hiện và giao của $Y^s \cap Y^u = \emptyset$. Mục tiêu huấn luyện mô hình: $f: \mathcal{X} \to \mathcal{Y}$ trong đó hàm mục tiêu:

$$\hat{y} = argmax(\, f_c(x) - y), c \in Y^s. \tag{4.3}$$

4.1.2. Mô hình khuếch tán (Diffusion model)

Phương pháp tổng hợp tiếng nói dựa trên nhiễu thêm nhiễu vào âm thanh gốc, và mạng được học để dự đoán tín hiệu gốc chưa bị nhiễu. Số lượng nhiễu khác nhau được thêm vào trong quá trình huấn luyện, vì vậy mô hình được kỳ vọng sẽ học quá trình loại bỏ nhiễu với bất kỳ lượng nhiễu nào. Trong quá trình suy luận, bắt đầu từ một tín hiệu ban đầu chứa một lượng nhiễu nhất định, mạng sẽ tinh chỉnh tín hiệu theo các vòng lặp để đạt được chất lượng cao.

Mô hình thống kê khuếch tán khử nhiễu (DDPM - Denoising diffusion probabilistic model, hay gọi tắt là diffusion model - mô hình khuếch tán) đã cho thấy hiệu quả cao trong nhiệm vụ sinh ảnh và sinh âm thanh [62]. Mô hình khuếch tán là một chuỗi Markov đã được tham số hóa và được huấn luyện bằng cách sử dụng suy diễn biến phân để tạo ra các mẫu khớp giống với dữ liệu gốc sau một khoảng thời gian hữu hạn [62].

Mô hình khuếch tán gồm hai quá trình ngược nhau như mô tả tại Hình 34: 1) Quá trình khuếch tán xuôi (diffusion process) là một chuỗi Markov với các tham số cố định nhằm chuyển đổi dữ liệu phức tạp thành phân phối Gauss đẳng hướng (isotropic) bằng cách dần dần thêm các nhiễu Gauss vào ảnh sạch x_0 từng bước cho đến khi ảnh bị phá vỡ cấu trúc thành x_T; 2) Quá trình phục hồi (reverse process) hay còn gọi là quá trình khuếch tán ngược là một chuỗi Markov được triển khai bằng mạng nơ-ron để học cách phục hồi x_T đã bị phá vỡ cấu trúc về dữ liệu gốc x_0 từ nhiễu trắng Gauss thông qua huấn luyện lặp đi lặp lại. Mục tiêu huấn luyện gồm có hai mục tiêu, một là là tối thiểu hóa khoảng cách giữa nhiễu khuếch tán x_T và giải mã quá trình phục hồi x_T; và hai là tối đa hóa làm sao để log xác suất phục hồi giữa x_0lớn nhất dựa trên bộ giải mã nhiễu x_0.

Trong suốt quá trình khuếch tán xuôi, xác định phân bố dữ liệu $x_0 \sim q(x_0)$ và biến ẩn x_t ở bước t, nghiệm cuối gần đúng được xác định trong chuỗi Markov và được định nghĩa như sau [62]:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad (4.4)$$

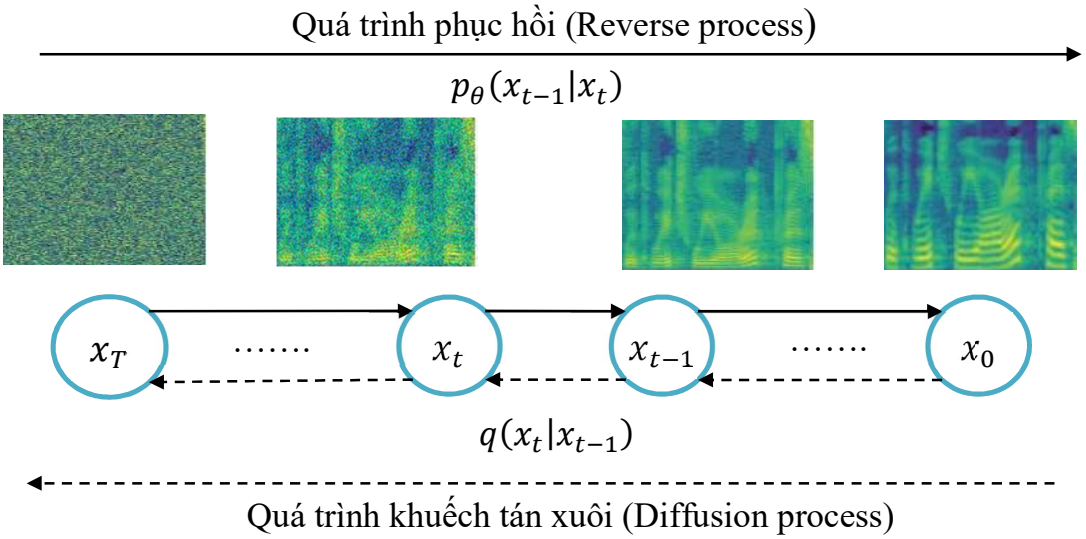
$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \text{ với } \beta_t \in (0,1) \quad (4.5)$$

trong đó β_t là tốc độ khuếch tán ở bước *t.* Quy trình ngược lại, mô hình khuếch tán ngược nhằm phục hồi x_0 từ x_T bị phá hủy. Điều này được huấn luyện bằng cách ước tính nhiễu trong quá trình huấn luyện, được định nghĩa như một chuỗi Markov để học chuyển đổi Gauss bắt đầu từ $p(x_T) = N(x_T; 0; I)$:

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad (4.6)$$

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t; t), \Sigma_\theta(x_t; t)) \quad (4.7)$$

Mô hình khuếch tán có tính linh hoạt cao và cho phép sử dụng bất kỳ kiến trúc nào có kích thước đầu vào và đầu ra giống nhau. Đây là đặc trưng quan trọng nhằm ứng dụng mô hình khuếch tán trong tổng hợp tiếng nói để đạt được giọng tổng hợp có chất lượng cao và giống nguyên mẫu nhất có thể.



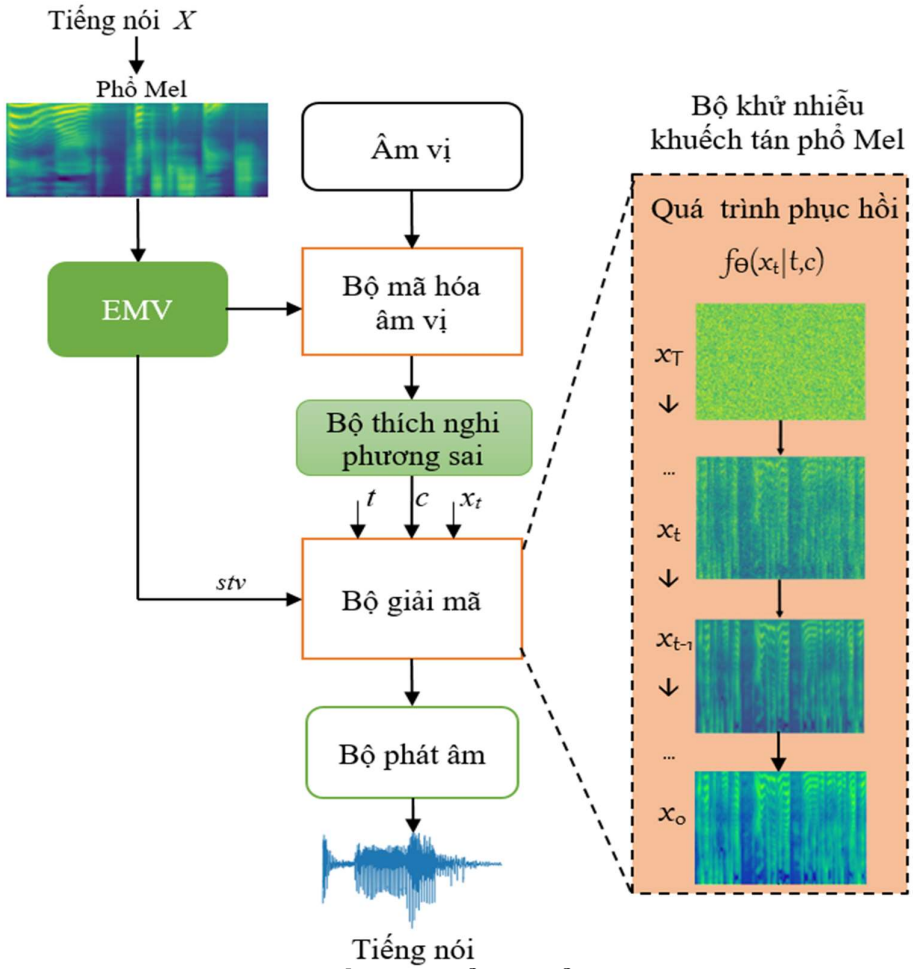
Hình 34: Mô tả trực quan tiến trình phục hồi và tiến trình khuếch tán của mô hình khuếch tán (Diffusion model)

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

Trong phần này, luận án đề xuất một mô hình Adapt-TTS để tổng hợp một giọng nói mới chưa từng xuất hiện trong tập huấn luyện với chỉ một câu nói (Zero-shot TTS). Mô hình Adapt-TTS đề xuất với hai thành phần cải tiến và những đóng góp chính của như:

1) Đề xuất kiến trúc Extracting Mel-vector (EMV) cho phép biểu diễn đặc trưng giọng để khái quát hóa tốt hơn. Kiến trúc này đã tỏ ra rất hiệu quả trong việc học cái đặc trưng giọng nói từ mẫu rất nhỏ giọng đích [CT2].
2) Đề xuất mô hình khử nhiễu khuếch tán 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.
3) Đề xuất ứng dụng nhân bản giọng đề xuất thể hiện tính khả thi của mô hình ứng dụng trong thực tế.

4.2.1. Mô hình tổng quát



Hình 35: Kiến trúc tổng thể Adapt-TTS

Kiến trúc của Adapt-TTS bao gồm các thành phần chính: Mô-đun EMV để trích xuất các đặc trưng giọng nói và phong cách nói thành một vector đặc trưng, Bộ mã hóa âm vị (Phoneme Encoder) dùng để biến đổi chuỗi âm vị thành các chuỗi âm vị ẩn, sau đó Bộ thích nghi phương sai (Variance adaptor) 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, bộ phát âm sẽ chuyển đổi các phổ Mel này thành tín hiệu tiếng nói. Kiến trúc tổng thể được mô tả trong Hình 35. Chức năng và kiến trúc chi tiết của các mô-đun cải tiến đề xuất sẽ trình bày dưới đây.

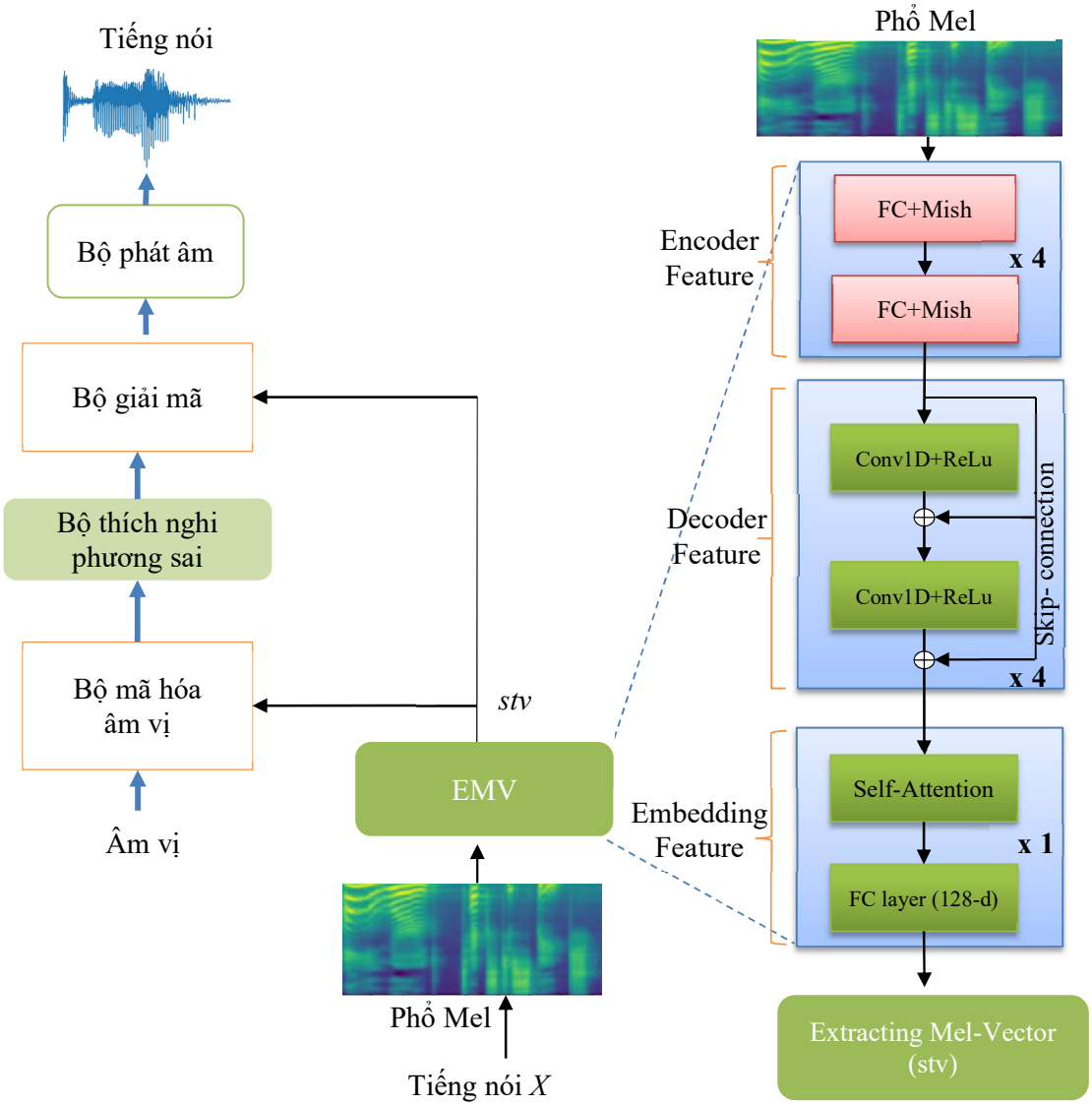
4.2.2. Mã hóa đặc trưng với EMV

Chương 4 đề xuất một mô-đun tương đồng với Chương 3 gọi là EMV có thể trích xuất một vectơ cố định từ biểu đồ phổ Mel của người nói có kích thước 128 chiều, để biểu các đặc trưng của người nói (như đặc trưng giọng 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 nhúng *stv* có chứa phong cách và đặc trưng của người nói *X*.

$$stv = EMV(Mel\_X), stv \in \mathbb{R}^N \quad (4.8)$$

Cụ thể hơn, trong mô-đun này, sử dụng ba thành phần (mã hóa đặc trưng, giải mã đặc trưng và nhúng đặc trưng). Trong khối **mã hóa đặc trưng (Encoder Feature)**, đầu vào của khối này là phổ Mel (*Mel*), sau đó nó sẽ chuyển qua hai lớp FC và các hàm kích hoạt Mish. Mục đích của khối Bộ mã hóa đặc trưng là chuyển đổi đặc trưng đầu vào thành đặc trưng của bộ mã hóa. Tiếp theo, vector này sẽ được đưa qua khối **Giải mã đặc trưng (Decoder feature)**. Bằng cách sử dụng Conv1D và ReLu, mô-đun này sẽ chuyển đổi đặc trưng của bộ mã hóa thành đặc trưng bộ giải mã. Ngoài ra, tích hợp kết nối tắt sẽ sử dụng các đặc trưng có giá trị của các khối trước đó. Cuối cùng, bộ giải mã đặc trưng sẽ chuyển sang mô-đun **nhúng đặc trưng (Embedding feature)**, mô-đun này có mô-đun tự chú ý và

lớp affine. Mô-đun này sẽ tạo ra một vectơ đại diện cho phổ Mel *stv* và vectơ này sẽ thêm vào mô hình chuyển văn bản thành giọng nói. Vectơ đại diện sẽ điều khiển đầu ra của mô hình TTS và tạo ra giọng nói tổng hợp tương tự như vectơ đầu vào. Chi tiết kiến trúc của EMV được trình bày trong Hình 36.

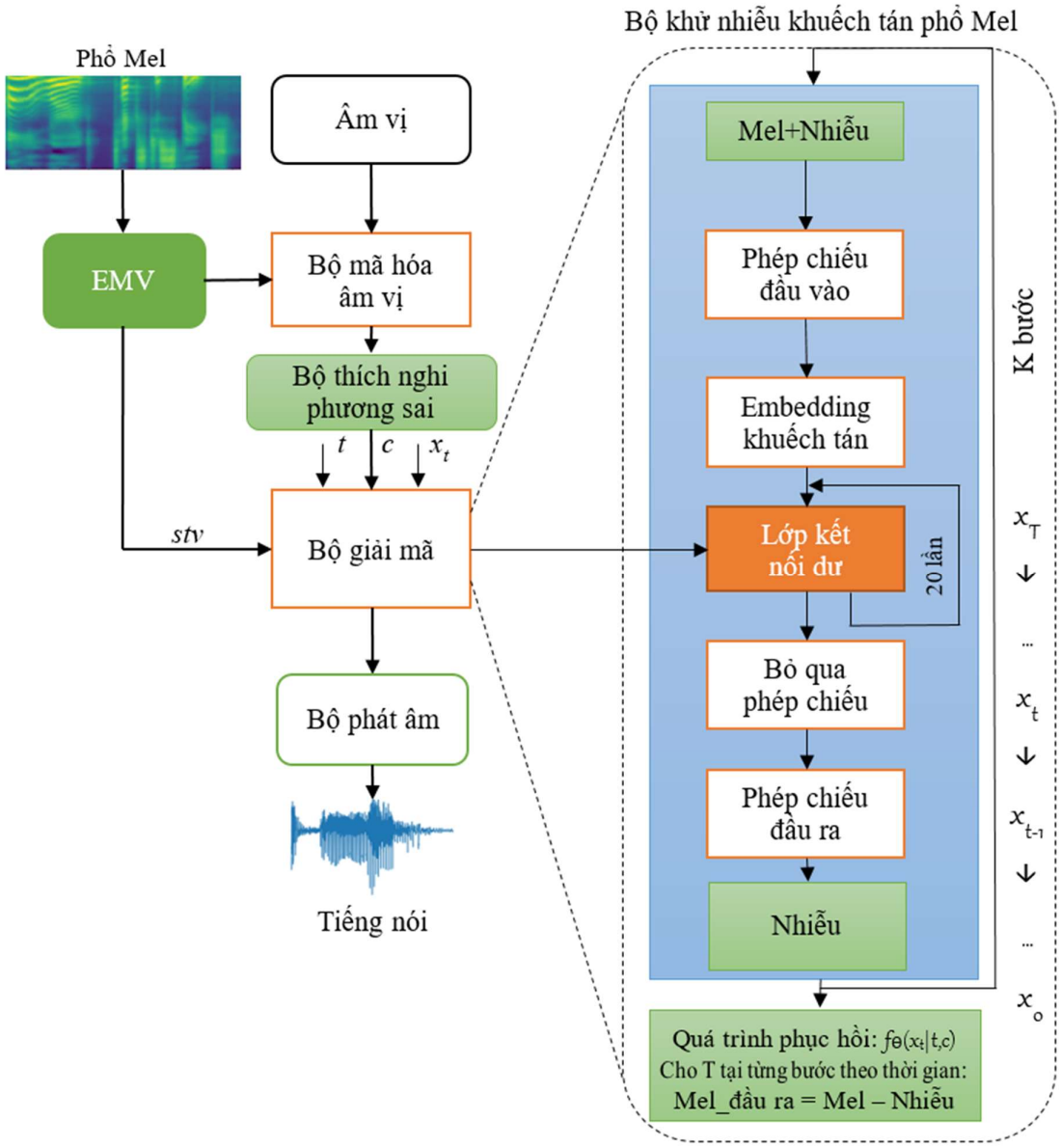


Hình 36: Cấu trúc chi tiết của mô-đun EMV

4.2.3. Bộ khử nhiễu khuếch tán phổ Mel (Mel-spectrogram denoiser)

Bộ giải mã nhận đầu vào từ chuỗi âm vị ẩn thông qua Bộ thích nghi phương sai để thêm các thông tin của giọng nói về như trường độ, cao độ, cường độ, sau đó kết hợp với vector biểu diễn đặc trưng người EMV nhằm cung cấp đầy đủ các tham số ẩn nhằm tái tạo các đặc trưng của giọng nói. Trong bộ giải mã, khối khử nhiễu khuếch tán phổ Mel sẽ nhận đầu vào là chuỗi x_t , biến c là đầu ra của bộ

thích nghi phương sai và bước thời gian t để thực hiện khử nhiễu và tổng hợp âm thanh chất lượng cao dựa trên trên mô hình khuếch tán. Quá trình suy diễn của mô hình khuếch tán cho Multi-TTS là tối ưu hóa hàm mục tiêu $f_\theta(x_t|t,c)$ để chuyển đổi các phân phối nhiễu thành một phân phối phổ Mel tương ứng với thông tin phương sai cho trước, mô hình này gồm hai tiến trình chính (Hình 37):



Hình 37: Kiến trúc chi tiết của khối khử nhiễu khuếch tán

Quá trình khuếch tán xuôi (Diffusion process) [62]**:**

Ý nghĩa của quá trình này dần dần phá vỡ cấu trúc của phân bố dữ liệu một cách hệ thống. Đầu tiên, ảnh phổ Mel dần dần bị làm nhiễu (corrupted) với nhiễu Gauss và biến đổi thành các biến 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,...,T* là chỉ số cho bước thời gian khuếch tán. Khi đó, quá trình khuếch tán xuôi biến đổi phổ Mel x_0 thành một nhiễu Guass x_T thông qua một chuỗi Markov. Mỗi một bước chuyển trạng thái được định nghĩa trước với biến lập lịch $\beta_1, \beta_2, \beta_3, \ldots, \beta_T$. Cụ thể hơn, mỗi phép biến đổi thực hiện phù hợp với xác suất chuyển trạng thái Markov $q(x_t|x_{t-1}, c)$ được giả định độc lập với biến c và nó được định nghĩa như sau:

$$q(x_t|x_{t-1}, c) = N(x_t, \sqrt{1-\beta_t x_{t-1}}, \beta_t I) \quad (4.9)$$

Toàn bộ khuếch tán xuôi $q(x_{1:T}|x_0, c)$ là quá trình Markov và có thể được phân tích như sau:

$$q(x_1 \ldots, x_T|x_0, c) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (4.10)$$

Quá trình phục hồi (Reverse process) [62]**:** Học cách khôi phục phân bố dữ liệu, là quy trình tạo phổ Mel từ nhiễu Gauss. Quá trình khôi phục cần xác định phân phối có điều kiện $p_\theta(x_{0:T}|x_T, c)$, và xác suất của trạng thái $x_0 \ldots, x_{T-1}$ có thể được thừa số hóa dựa trên tính chất chuỗi Markov:

$$p_\theta(x_0 \ldots, x_{T-1}|x_T, c) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c) \quad (4.11)$$

Thông qua các biến đổi ngược $p_\theta(x_{t-1}|x_t, c)$, các biến ẩn dần dần được khôi phục thành phổ Mel tương ứng với bước thời gian khuếch tán và điều kiện văn bản c. Nói cách khác, phân phối $p_\theta(x_0|c)$ thu được từ quy trình ngược lại.

Đặt $q(x_0|c)$ là phân bố phổ Mel. Để mô hình xấp xỉ tốt với $q(x_0|c)$, quá trình phục hồi lại nhằm mục đích tối đa hóa log-likelihood của phổ Mel: $E_{logq(x_0|c)}[logp_\theta(x_0|c)]$. Vì $p_\theta(x_0|c)$ là xác suất của mô hình ban đầu, nên sử dụng thủ thuật tham số hóa được trình bày trong [62] để tính cận dưới biến thiên của log-likelihood ở dạng đóng.

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:

$$minL_\theta = E_{x_0,\epsilon,t}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c)||_1\right] \quad (4.12)$$

trong đó *t* được lấy thống nhất từ toàn bộ bước thời gian khuếch tán, ϵ là tham số nhiễu ở bước thứ $(t-1)$ và ϵ_t mẫu khuếch tán ở bước thứ t.

Bộ khử nhiễu khuếch tán phổ Mel không cần bất kỳ hàm mất mát phụ nào ngoại trừ hàm hàm mất mát L1 giữa đầu ra của mô hình $\epsilon_\theta(\cdot)$ và nhiễu Gauss $\epsilon \sim N(0, I)$.

Trong quá trình khuếch tán ngược phục hồi một phổ Mel từ một biến ẩn bằng cách lặp lại từ dự đoán mỗi bước biến đổi thuận với $\epsilon_\theta(x_t, t, c)$ và loại bỏ phần bị nhiễu như sau:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, c)\right) + \sigma_t z \tag{4.13}$$

trong đó $z \sim N(0, I)$ và $\sigma_t = \eta\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}$. Thuật ngữ kỳ hạn thời gian η là hệ số tỷ lệ của phương sai. Lưu ý rằng bước thời gian khuếch tán *t* cũng được sử dụng làm đầu vào cho Bộ khử nhiễu khuếch tán phổ Mel, cho phép chia sẻ tham số cho tất cả các bước thời gian khuếch tán. Kết quả là, phân bố phổ Mel cuối cùng $p(x_0|c)$ thu được thông qua lấy mẫu lặp trên tất cả các bước thời gian đặt trước.

4.2.4. 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:

$$p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t, y) \tag{4.14}$$

$$\epsilon_\theta(x_t, t) \rightarrow \epsilon_\theta(x_t, t, y) \tag{4.15}$$

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

$$p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t, c, stv) \tag{4.16}$$

$$\epsilon_\theta(x_t, t) \rightarrow \epsilon_\theta(x_t, t, c, stv) \tag{4.17}$$

4.2.5. Hàm mất mát huấn luyện

- Hàm mất mát L_θ*:* Thay vì sử dụng dữ liệu sạch x_0 ban đầu trong quá trình huấn luyện Adapt-TTS thì ta nhận được giá trị mục tiêu $\widehat{x_0}$ với phương sai giảm dần bằng cách chạy hai bước lấy mẫu mô hình khử nhiễu phổ Mel của bộ huấn luyện:

$$L_\theta = ||\hat{x}_0 - x_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon)||_1, \epsilon \sim N(0,I) \quad (4.18)$$

- Hàm mất mát chỉ số tương tự về cấu trúc (SSIM)*:* Giá trị của SSIM nằm trong khoảng từ 0 và 1, trong đó 1 biểu diễn chất lượng cảm nhận hoàn hảo so với giọng nói gốc, luận án áp dụng hàm mất mát chỉ số tương tự cấu trúc trong huấn luyện mô hình TTS như sau:

$$L_{SSIM} = 1 - SSIM\left(x_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right), \hat{x}_0\right) \quad (4.19)$$

- Hàm mất mát phương sai: Để tăng cường tính tự nhiên và tính biểu cảm của giọng nói tổng hợp cần cung cấp thêm thông tin phương sai về âm thanh bao gồm trường độ, cao độ và cường độ. Hàm mất mát phương sai cũng được thêm vào để huấn luyện bộ tạo âm học như sau:

$$L_{duration} = ||d - \hat{d}||_2^2, L_{pich} = ||p - \hat{p}||_2^2, L_{energy} = ||\varepsilon - \hat{\varepsilon}||_2^2, \quad (4.20)$$

trong đó sử dụng d, p và e để biểu thị trường độ, cao độ và cường độ của mẫu âm thanh mục tiêu và sử dụng $\hat{d}, \hat{p}$ và $\hat{\varepsilon}$ để biểu thị tương ứng các giá trị mẫu âm thanh dự đoán. Tất cả các trọng số mất mát được đặt bằng 0,1.

***Tóm tắt quá trình huấn luyện và suy diễn:**

Quá trình huấn luyện: Bên cạnh hàm loss tái tạo mẫu được mô tả ở trên, để đánh giá chất lượng của đầu ra dự đoán về cao độ, cường độ và trường độ, các giá trị loss của thông tin biến thể được tính toán bằng cách sử dụng chỉ số lỗi bình phương trung bình (MSE) giữa với âm thanh gốc và âm thanh nhân tạo. Ngoài ra, để đánh giá mức độ giống nhau của phổ Mel dự đoán với âm thanh thực tế, mức suy hao được tính toán bằng cách sử dụng sai số tuyệt đối trung bình (MAE) và thước đo chỉ số tương tự cấu trúc (SSIM), cung cấp thước đo độ trung thực của

âm thanh. Giá trị loss cuối cùng trong quá trình huấn luyện bộ khử nhiễu phổ Mel bao gồm các phần sau đây:

$$L_{final} = L_{\theta} + L_{SSIM} + L_{duration} + L_{pich} + L_{energy} \quad (4.21)$$

3. Giá trị hàm mất mát L_{θ}: Sai số bình phương trung bình MSE giữa mẫu phổ mel dự đoán và mục tiêu;
4. Giá trị hàm mất mát chỉ số tương đồng cấu trúc L_{SSIM}(SSIM): Một trừ đi chỉ số SSIM giữa mẫu phổ Mel dự đoán và mục tiêu;
5. Giá trị hàm mất mát phương sai $L_{duration}$, L_{pich}, L_{energy}: Sai số bình phương trung bình giữa trường độ các âm vị, cao độ và cường độ của mẫu dự đoán và mục tiêu.

Quá trình suy diễn: Trong quá trình suy diễn, lấy phổ Mel đầu vào x_0 không bị nhiễu và sau đó thêm nhiễu bằng cách sử dụng phân phối hậu nghiệm, từ đó tạo ra các mặt phẳng phổ Mel với các bước 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, ảnh phổ Mel được tạo ra từ x_0 được chuyển đổi thành dạng sóng bằng cách sử dụng một bộ phát âm được huấn luyện trước.

4.3. Thử nghiệm đánh giá và kết quả

4.3.1. Thử nghiệm đánh giá

Bộ dữ liệu: Để đánh giá mô hình, một bộ dữ liệu đa người nói tiếng Việt có nhãn đã được sử dụng. Bộ dữ liệu bao gồm 54 người nói, với 26 giọng nam và 28 giọng nữ. Bộ dữ liệu cũng bao gồm cả phương ngữ Bắc và Nam, với mỗi người nói ghi lại khoảng 500 câu nói. Để đánh giá chất lượng âm thanh tổng hợp tạo từ các mô hình đề xuất, 05 bộ dữ liệu đã được chuẩn bị, trong đó, 04 bộ được tổng hợp từ các âm thanh tham chiếu với khoảng thời gian tương ứng là 1 giây, 3 giây, 5 giây và một bộ là các âm thanh gốc để đối sánh.

Cách đánh giá kết quả: Luận án sẽ sử dụng hai mô hình để tổng hợp: một là mô hình cơ sở (Baseline) đề xuất bởi Yi Ren [30] (đã sử dụng kỹ thuật thích nghi bằng phân phối dữ liệu [CT1]) và hai là mô hình Adapt-TTS. Luận án sử dụng 30 người nghe là các cán bộ, giáo viên và sinh viên đang học tập, làm việc

tại các trường Đại học tại Việt Nam để nghe và chấm điểm các âm thanh được cung cấp thông qua một ứng dụng đánh giá nền web. Luận án đánh giá hệ thống tổng hợp bằng cách phối hợp đánh giá cả theo phương pháp đánh giá khách quan (Subjective sử dụng các chỉ số định lượng như WER) và đánh giá chủ quan (Objective sử dụng các chỉ số định tính như MOS/SIM).

Thử nghiệm 1: Đánh giá chất lượng tổng hợp tiếng nói: Sử dụng thang đo MOS (Mean Opinion Score) để đánh giá chất lượng tổng hợp. Để đảm bảo khách quan, các âm thanh gốc được đưa vào trộn lẫn cùng để xác định thang điểm tối đa đối với giọng người nói. Sử dụng chỉ số WER nhằm đo lường tỷ lệ phần trăm các từ bị nhận diện sai trong văn bản nhận dạng từ âm thanh tổng hợp so với văn bản nhận dạng từ âm thanh gốc.

Thử nghiệm 2: Đánh giá độ tương đồng giữa giọng tổng hợp và giọng người nói: Sử dụng chỉ số SIM (similarity) để đo sự tương đồng giữa âm thanh tổng hợp và âm thanh gốc của người nói mục tiêu. Sử dụng 30 người nghe, mỗi người nghe 140 cặp câu âm thanh để đánh giá, các câu này được tổ hợp từ các mô hình cơ sở và đề xuất, và kết hợp đánh giá cả bộ âm thanh gốc, mỗi mô hình sẽ tạo ra 20 câu. Từ chỉ số đó đánh giá được hiệu quả của mô hình thích nghi.

4.3.2. Kết quả

a) Chất lượng tổng hợp

Bảng 15: Kết quả đánh giá chất lượng tổng hợp MOS/WER của các mô hình cơ sở và mô hình đề xuất với các giọng chưa có trong tập huấn luyện với độ tin tưởng 95%

Trường độ/ Mô hình	Mô hình cơ sở		Mô hình Adapt-TTS	
	MOS (↑)	WER(↓)	MOS(↑)	WER(↓)
Âm thanh gốc (Groundtruth)	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

Bảng 15 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 giọng 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 mẫu âm của một giọng nói (reference-speaker) mô hình đề xuất đã có thể tổng hợp được âm thanh đạt WER 3.38.

b) Độ tương đồng

Bảng 16: 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	Mô hình 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

Bảng 16 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 **2.22/3.90** của giọng người nói. Trong khi mô hình cơ sở chỉ đạt chỉ số SIM 1.24/3.9.

Hình 38: 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

Tiến hành phân tích đánh giá SIM theo phương pháp [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 38, 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.

Bảng 17: Bảng phân tích ANOVA về điểm đánh giá tương đồng giữa mô hình cơ sở và mô hình đề xuất Adapt-TTS với 3 giây âm thanh mẫu

Mô hình	**Nguồn phương sai**	*Tổng bình phương (SS)*	*Bậc tự do (df)*	*Bình phương trung bình (MS)*	*Giá trị thống kê (F)*	*Giá trị p*	*F tới hạn*
Cơ sở	Giữa các nhóm	9.415	29	0.325	1.675	0.015712	1.487
	Trong các nhóm	110.450	570	0.194			
Adapt-TTS	Giữa các nhóm	36.415	29	1.256	2.099	0.000774	1.487
	Trong các nhóm	340.850	570	0.598			

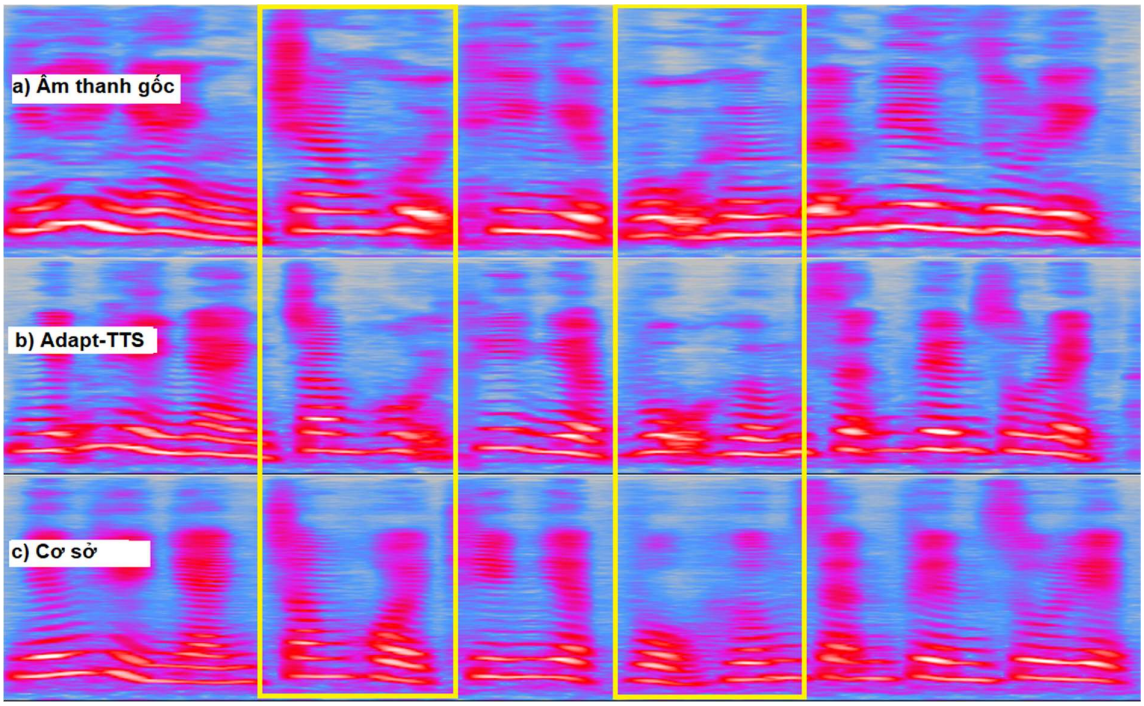
Phân tích ANOVA một chiều (phụ thuộc một biến duy nhất là mô hình TTS) trong đánh giá độ tương đồng SIM giữa các mô hình cho kết quả ở Bảng 17. Các giả thuyết cho phân tích ANOVA ban đầu như sau :

- **Giả thuyết không(H_0):** Tất cả các phương pháp có chất lượng tương đồng bằng nhau (không có sự khác biệt).
- **Giả thuyết thay thế(H_1):** Có sự khác biệt ít nhất một cặp phương pháp.

Qua kết quả phân tích 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.

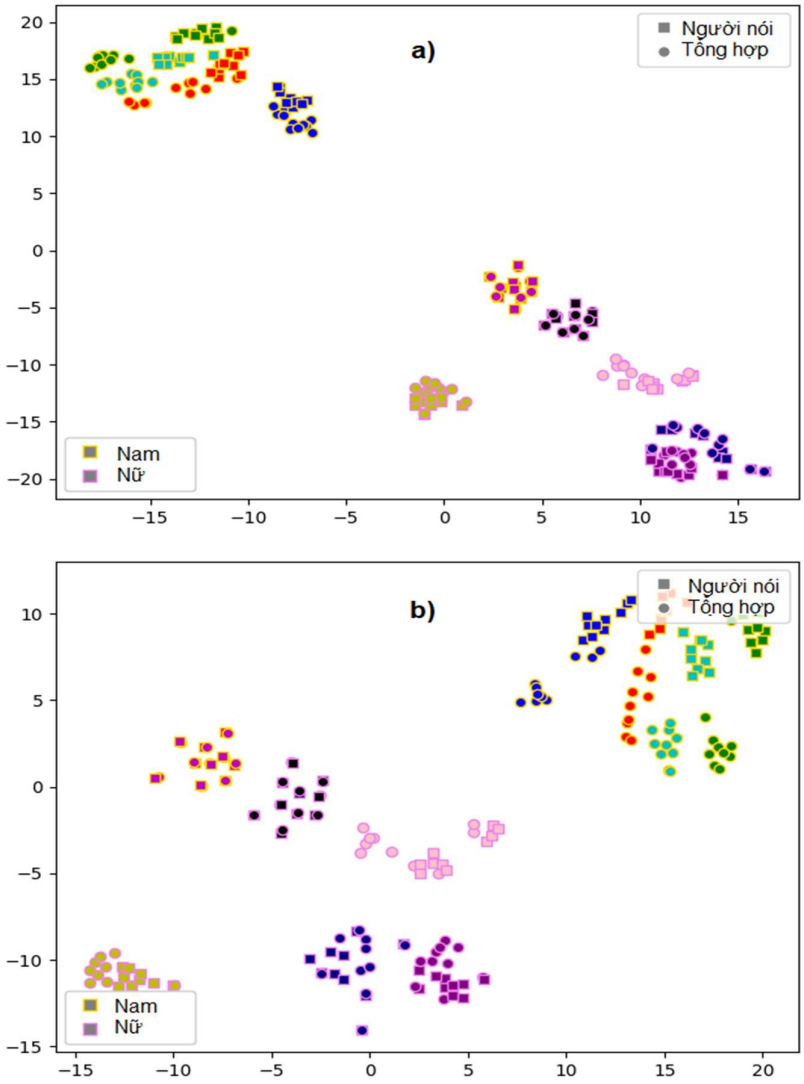
Ngoài ra dựa vào so sánh ảnh phổ âm thanh với 3 giây mẫu thích nghi, có thể thấy: Ảnh phổ Mel (khoanh vùng chữ nhật) giữa âm thanh được tạo bởi Adapt-TTS và âm thanh gốc có độ tương đồng cao hơn hẳn so với âm thanh được tạo

bởi mô hình cơ sở (Hình 39). Ngoài ra âm thanh tạo bởi mô hình cơ sở có vân mờ hơn và mang nhiều nhiễu.



Hình 39: Ảnh phổ Mel của 3 âm thanh: a) âm thanh gốc b) âm thanh tạo bởi Adapt-TTS và c) âm thanh tạo bởi mô hình cơ sở với 3 giây mẫu thích nghi

Để hiểu sâu hơn về hiệu quả của mô hình Adapt-TTS, minh họa các vectơ EMV thông qua phương pháp trực quan hóa bằng cách tính toán ma trận khoảng cách giữa các điểm dữ liệu của âm thanh tổng hợp và âm thanh gốc. Hình 40 trình bày phép chiếu t-SNE mẫu âm thanh những người nói không nhìn thấy trong bộ dữ liệu đa người nói của tiếng Việt và âm thanh tổng hợp tương ứng, 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 (Hình 40b). Biểu đồ t-SNE của mô hình Adapt-TTS (Hình 40a) cho thấy âm thanh tổng hợp và âm thanh gốc của cùng một người nói co cụm lại gần nhau. Đặc trưng giới tính cũng được phân cụm rõ ràng ở hai vùng khác nhau.



Hình 40: 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 đun mã hóa, phân phối do bộ mã hóa 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 mã hóa. Bộ mã hóa đặc trưng giọng nói với EMV và cho phép thích nghi các đặc trưng giọng nói (trường độ, cao độ, cường độ) bằng Variance Adapter. Các phổ Mel được dự đoán chính xác hơn nhờ bộ giải mã với thành phần Bộ khử nhiễu khuếch tán phổ Mel theo mô hình khuếch tán tạo âm thanh chất lượng cao và ít nhiễu. Từ đó cho phép mô

hình đề xuất Adapt-TTS tạo tiếng nói giống với giọng nói của những người nói không được huấn luyện bằng cách thực hiện chuyển đổi giọng nói Zero-shot TTS.

Khả năng học nhanh giọng này cho phép ứng dụng rộng rãi trong thực tế, ví như học giọng mới tức thì của cha mẹ để trao tiếp với trẻ nhỏ trên các loa thông minh, thiết bị di động; hoặc xây dựng các ứng dụng đọc báo cáo tự động bằng giọng nói giống người trình bày thông qua học nhanh một số mẫu giọng nhỏ; và tiềm năng có thể áp dụng được cho nhiều nhu cầu khác trong cuộc sống…

4.4. 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]. Ở phần Phụ lục, luận án sẽ trình bày xây dựng một ứng dụng thích nghi dựa trên các kỹ thuật đã đề xuất trong cả ba chương để chứng minh tính khả thi của các mô hình đề xuất và cung cấp minh chứng có liên quan.

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.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

I. Tạp chí khoa học

- **[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.
- **[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.

II. Kỷ yếu hội thảo chuyên ngành

- **[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.
- **[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.
- **[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.
- **[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.
- **[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.*

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Y. A. Chung, Y. Wang, W. N. Hsu, Y. Zhang and R. J. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 6940-6944, 2019.

[2] Y. Yan, X. Tan, B. Li, Q. Tao, S. Zhao, Y. Shen and T.-Y. Liu, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] Y. Yan, B. L. Xu Tan, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang and T.-Y. Liu, "Adaspeech 3: Adaptive text to speech for spontaneous style," in *INTERSPEECH*, 2021.

[4] J. K. T. Yamagishi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst,* Vols. Vols. E90-D, 2007.

[5] Q. Xie, X. Tian, G. Liu, K. Song, L. Xie, Z. Wu and X. Xu, "The multi-speaker multi-style voice cloning challenge 2021," in *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, June.

[6] N. T. T. Trang, N. H. Ky, P. Q. Minh and V. Manh, "Remaining problems with state-of-the-art techniques in proceedings of the seventh international workshop on Vietnamese language and speech processing," VLSP 2020, 2020.

[7] P. T. Son, V. T. Thang and C. T. Dương, "Nghiên cứu nâng cao chất lượng tổng hợp tiếng nói tiếng Việt dựa trên mô hình Markov ẩn và đặc trưng ngôn ngữ," *Kỷ yếu Hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông, Hà Nội,* pp. 238-242, 2013.

[8] D. K. Ninh, "A speaker-adaptive hmm-based vietnamese text-to-speech system," *2019 11th International Conference on Knowledge and Systems Engineering (KSE),* pp. 1-5, 2019.

[9] H. Zen, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *2013 ieee international conference on acoustics, speech and signal processing. IEEE,* pp. 7962-7966, 2013.

[10] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Koray, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[11] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *In NeurIPS*, 2019.

[12] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[13] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao and T. Y. Liu, "AdaSpeech: Adaptive Text to Speech for Custom Voice.," *arXiv preprint arXiv:2103.00993,* 2021.

[14] Z. Wu, P. Swietojanski, C. Veaux, S. Renals and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] I. Tokuda, "The Source–Filter Theory of Speech," in *Oxford Research Encyclopedia of Linguistics*, 2021.

[16] Damper, C. H. Shadle and R. I, "Prospects for articulatory synthesis: A position paper," in *In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[17] P. Seeviour, J. Holmes and M. Judd, "Automatic generation of control signals for a parallel formant speech synthesizer," in *In ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1976.

[18] A. J. Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings olume 1, pages 373–376. IEEE,*, 1996.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *IEE2000 E International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2000.

[20] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *Third International Conference on Spoken Language Processing*, 1994.

[21] H. Kawahara, I. Masuda-Katsuse and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* vol. Volume 27, no. Issues 3–4, pp. 187-207, 1999.

[22] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proc. ICASSP-83,* p. 93–96, 1983.

[23] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," in *Acoustical science and technology*, 2006.

[24] M. Morise, F. Yokomori and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," in *IEICE TRANSACTIONS on Information and Systems*, 2016.

[25] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping and Jonathan, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in neural information processing systems,* p. 2962–2970, 2017.

[26] W. Wang, S. Xu and B. Xu, "First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention," *In Interspeech,* pp. 2243-2247, 2016.

[27] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, Q. L. S. Bengio, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech,* p. 4006–4010, 2017.

[28] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang and R. Skerrv-Ryan, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[29] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," arXiv preprint arXiv:1710.07654, 2017.

[30] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," In International Conference on Learning Representations, 2021, 2021.

[31] W. Ping, K. Peng and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," International Conference on Learning Representations, 2018., 2018.

[32] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen and K. Simonyan, "End-to-end adversarial text-to-speech," ICLR, 2021., 2021.

[33] T. Tho, T. C. Chu, V. Hoang, T. Bui and S. Truong, "An Efficient and High Fidelity Vietnamese Streaming End-to-End Speech Synthesis," *In INTERSPEECH 2022,* pp. 466-470, 2022.

[34] Bahdanau, K. C. Dzmitry' and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems. Curran Associates, Inc,* pp. volume 30, pages 5998–6008, 2017.

[36] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *In Interspeech*, 2017.

[37] J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems,* pp. 502-509, 2005.

[38] M. Tachibana, J. Yamagishi, T. Masuko and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *EICE transactions on information and systems,* pp. 2484-2491, 2005.

[39] T. Nose, J. Yamagishi, T. Masuko and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems,* pp. 1406-1413, 2007.

[40] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing,* pp. 66-83, 2009.

[41] Saito, Yuki, S. Takamichi and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* pp. 84-96, 2017.

[42] J. Kong, J. Kim and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," arXiv preprint arXiv:2010.05646, 2020.

[43] "ZeroSpeech, Zero resource speech challenge," ZeroSpeech Zero resource speech challenge, 2020. [Online]. Available: https://www.zerospeech.com/.

[44] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed and H. Zen, "Sample efficient adaptive text-to-speech," in *International Conference on Learning Representations*, 2018.

[45] S. Ö. Arık, J. Chen, K. Peng, W. Ping and Y. Zhou, "Neural voice cloning with a few samples," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[46] D. Tan, H. Huang, G. Zhang and T. Lee, "Cuhk-ee voice cloning system for icassp 2021 m2voc challenge," [Online]. Available: arXiv preprint arXiv:2103.04699, 2021..

[47] D. Paul, M. P. Shifas, Y. Pantazis and Y. Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion," in *Interspeech 2020*, 2020.

[48] Q. Hu, T. Bleisch, P. Petkov, T. Raitio, E. Marchi and V. Lakshminarasimhan, "Whispered and lombard neural speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[49] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. J. A. Rosenberg and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Proc. Interspeech 2019*, 2019.

[50] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *Interspeech 2019*, 2019.

[51] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz and R. Hoory, "High quality, lightweight and adaptable tts using lpcnet," in *Proc. Interspeech 2019*, 2019.

[52] M. N. Chừ, Cơ sở ngôn ngữ học và tiếng Việt, Nxb Giáo dục, 1997.

[53] X. Võ, Giáo trình Ngữ âm tiếng Việt hiện Đại, Đại học Quy Nhơn, 2009.

[54] N. Trang, P. Thanh and T. Đạt, "A method for Vietnamese Text Normalization to improve the quality of speech synthesis," *Symposium on Information and Communication Technology, SoICT 2010,* 2010.

[55] Đ. T. Thuật, "Ngữ âm tiếng Việt," in *NXB Đại học Quốc gia Hà Nội*, 2003.

[56] R. Prenger, R. Valle and B. Catanzaro., "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3617–3621. IEEE,*, 2019.

[57] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning pages 2410–2419.PMLR*, 2018.

[58] S. Kim, S.-G. Lee, J. Song, J. Kim and S. Yoon, "Flowavenet: A generative flow for raw audio," in *International Conference on Machine Learning, 3370–3378. PMLR*, 2019.

[59] D. P. Kingma and P. Dhariwal, "Glow: generative flow with invertible 1x1 convolutions.," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems,pages 10236–10245*, 2018.

[60] I. J. Goodfellow, J. Pouget-Abadi, M. M. B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[61] M. W. Diederik P Kingma, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114,, 2013.

[62] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," arXiv preprint arXiv:2006.11239,, 2020.

[63] P.85 and I.-T. Recommendation, A method for subjective performance assessment of the quality of speech output devices, International Telecommunications Union publication, 1994.

[64] J. Kominek, T. Schultz and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[65] W. M, Z. Wu and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 Evaluation Results," in *In Interspeech (pp. 1637-1641).*, 2016.

[66] S. Schneider, A. Baevski, R. Collobert and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862., 2019.

[67] G. S. H. S. D. N. M. Picheny, "Speaker adaptation of neural network acoustic models using I-vectors," *Proc IEEE ASRU,* pp. 55-59, 2013.

[68] B. Potard, P. Motlicek and D. Imsen, "Preliminary work on speaker adaptation for dnn-based speech synthesis," *Idiap, Tech.Rep,* 2015.

[69] B. Bollepalli, L. Juvela and P. Alku, "Lombard speech synthesis using transfer learning in a tacotron text-to-speech system," in *Interspeech, 2019*, 2019.

[70] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu and H.-y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," [Online]. Available: arXiv preprint arXiv:2103.04088, 2021..

[71] T. M. EA Platanios. M Sachan. G Neubig, "Contextual parameter generation for universal neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[72] H. B. Moss, V. Aggarwal, N. Prateek, J. González and R. Barra-Chicote., "Boffin tts: Few-shot speaker adaptation by bayesian optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[73] Wu, Yihan, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin and T.-Y. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," arXiv preprint arXiv:2204.00436., 2022.

[74] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim and Q. J. Wu, "A review of generalized zero-shot learning methods," in *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[75] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang and I. L. Moreno, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[76] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. de Oliveira, A. C. Junior and M. A. Ponti, "Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model," arXiv preprint arXiv:2104.05557., 2021.

[77] T. D. Dat, "Synthèse de la parole à partir du texte en langue vietnamienne," PhD thesis,Grenoble, INPG, 2007.

[78] D. V. Thao, T. D. Dat and N. T. T. Trang, "Nonuniform unit selection in Vietnamese Speech Synthesis," *Proceedings of the 2nd SoICT 2011,* pp. 165-171, 2011.

[79] V. H. Quân and C. X. Nam, "Tổng hợp tiếng nói tiếng Việt theo phương pháp ghép nối cụm từ," *Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Tạp chí CNTT và TT,* pp. Tập V-1(1), tr. 70-76, 2009.

[80] V. T. Thang, L. C. Mai and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," *Proceedings of the Oriental COCOSDA,* 2009.

[81] H. Linyu, Y. Jian, Z. Libo and K. Liping., "A trainable Vietnamese speech synthesis system based on HMM," *Proceedings of the International Conference on ElectricInformation and Control Engineering (ICEICE),* p. 3910–3913, 2011.

[82] D. Anh-Tuan, P. Thanh-Son, V. Tat-Thang and L. C. Mai, "Vietnamese hmm-based speech synthesis with prosody information," *8th ISCA Workshop on Speech Synthesis, Barcelona, Spain,* p. 51–54, 2013.

[83] K. Yun, J. Osborne, T. L. M. Lee and E. Chow, "Automatic speech recognition for launch control center communication using recurrent neural networks with data augmentation and custom language model," *Disruptive Technologies in Information Sciences, International Society for Optics and Photonics,* p. vol. 10652, 2018.

[84] N. V. Thinh, D. Q. Bao, P. H. Khanh and D. Hai, "Development of vietnamese speech synthesis system using deep neural networks," *Journal of Computer Science and Cybernetics,* pp. 349-363, 2018.

[85] N. T. T. Trang and N. H. Ky, "Vlsp 2021-tts challenge: Vietnamese spontaneous speech synthesis.," in *VNU Journal of Science: Computer Science and Communication Engineering, 38(1).*, 2022.

[86] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process,* pp. vol. 2, no. 2, pp. 291–298, 1994.

[87] M. Tonomura, T. Kosaka and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," *Comput. Speech Lang,* pp. vol. 10, no. 2, pp., 1995.

[88] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "Adaptation of pitch and spectrum for HMMbased speech synthesis using MLLR," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 2, pp. 805-808). IEEE.*, 2001.

[89] P. T. Son, "Nghiên cứu nâng cao chất lượng tổng hợp tiếng nói tiếng Việt dựa trên mô hình Markov ẩn và đặc trưng ngôn ngữ," *Luận án tiến sĩ,* pp. 77-78, 2014.

[90] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito and S.Imai, "Spectral estimation of speech based on mel-cepstral representation," *Trans. IEICE, vol. J74-A,* p. 1240–1248, 1991.

[91] S. Imai, T. Fukada, T. K and K. T., "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP-92,* p. 137–140, 1992.

[92] W. P and C. Leggetter, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang,* pp. vol. 9, no. 2, pp. 171–185, 1995.

[93] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst,* pp. vol. E85-D, no.3, pp. 455–464.

[94] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc.ICASSP-95,* p. 660–663, 1995.

[95] L. Maguer, Sébastien, I. Steiner and A. Hewer, "An HMM/DNN Comparison for Synchronized Text-to-Speech and Tongue Motion Synthesis," in *In Interspeech, pp. 239-243.*, 2017.

[96] Yang, Hongwu, W. Zhang and P. Zhi, "A DNN-based emotional speech synthesis by speaker adaptation," in *In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 633-637. IEEE,*, 2018.

[97] N. T. T. Trang, N. Ky, P. Q. Minh and V. D. Manh, "Vietnamese text-to-speech shared task vlsp 2020: Remaining problems with state-of-the-art techniques," in *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, 2021.

[98] L. C. Mai, "Special issue in vlsp 2018," *Computer Science and Cybernetics, ,* vol. Vol. 34, pp. No. 4, 2018., 2018.

[99] N. T. T. Trang and N. X. Tung, "Text-to-speech shared task in vlsp campaign 2019: evaluating vietnamese speech synthesis on common datasets,," VLSP, 2019.

[100] L. V. Bac, T. D. Dat, E. Castelli and L. Besacier, "Spoken and written language resources for Vietnamese," *Proceedings of LREC,* 2004.

[101] J.-s. Zhang and S. Nakamura, "An efficient algorithm to search for a minimum sentence set for collecting speech database," in *Proc. ICPhS*, 2003.

[102] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. R. Y. Jia and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *ICML, 2018,* p. 5167–5176, 2018.

[103] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[104] N. Tits, K. E. Haddad and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *Proceedings of SAI Intelligent Systems Conference*, 2019.

[105] D. B. Hamed Hemati, "Using ipa-based tacotron for data efficient cross-lingual speaker adaptation and pronunciation enhancement," arXiv:2011.06392, 2020., 2020.

[106] Pan, S. Jialin and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering,* pp. vol. 22, no. 10,pp. 1345–1359, 2010.

[107] A. S. Razavian, H. Azizpour, J. Sullivan and a. S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 806-813*, 2014.

[108] K. P. A. G. S. O. A. A. K. S. N. J. R. a. J. M. W. Ping, "Deep voice 3: 2000-speaker neural text-to-speech," *International Conference on Learning Representations,* 2018.

[109] D. S. Y. Z. R. J. S.-R. E. B. J. S. Y. X. Y. J. F. R. a. R. A. S. Y. Wang, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *ICML, 2018,* p. 5167–5176, 2018.

[110] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng and J. Raiman, "Deep voice: Real-time neural text-to-speech.," arXiv preprint arXiv:1702.07825, 2017.

[111] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* pp. vol. 19, no. 4, pp. 788–798, 2010.

[112] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 5329–5333, 2018.

[113] W. Xie, A. Nagrani, J. S. Chung and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 5791–5795, 2019.

[114] David, Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Interspeech, 2017,* p. 999–1003, 2017.

[115] J. S. Chung, A. Nagrani and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," arXiv preprint arXiv:1806.05622, 2018.

[116] T. Q. Chung, N. Q. Minh, P. N. Phuong, D. Q. Truong and L. C. Mai, "Improving Speaker Verification in Noisy Environment Using DNN Classifier," *RIVF 2021,* 2021.

[117] X. Tan, T. Qin, F. Soong and T. Y. Liu, "A Survey on Neural Speech Synthesis," arXiv preprint arXiv:2106.15561, 2021.

[118] H. Sung-Feng, C.-J. Lin, D.-R. Liu, Y.-C. Chen and H.-y. Lee, "Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1558-1571*, 2022.

[119] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, "Fastspeech2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[120] T. Wang, J. Tao, R. Fu, J. Yi, Z. Wen and R. Zhong, "Spoken Content and Voice Factorization for Few-shot Speaker Adaptation," *Proc. Interspeech 2020,* pp. 796-800, 2020.

[121] D. Min, D. B. Lee, E. Yang and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning (pp. 7748-7759). PMLR.*, 2021.

[122] D. Misra, "Mish: A self regularized non-monotonic activation function," arXiv:1908.08681, 2019.

[123] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in neural information processing systems*, 2021.

[124] T. Schultz, Speaker characteristics, Speaker classification I , (Springer, 2007) pp.53-54, 2007.

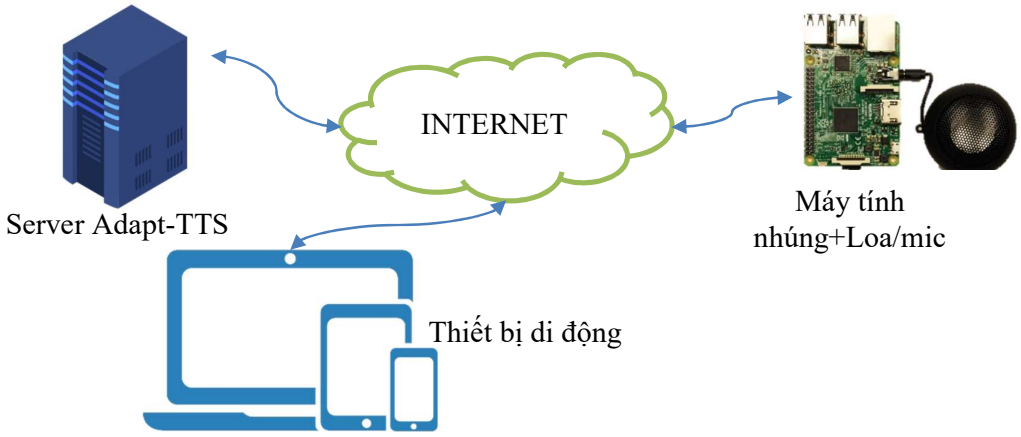
PHỤ LỤC

i) Xây dựng ứng dụng nhân bản giọng tiếng Việt

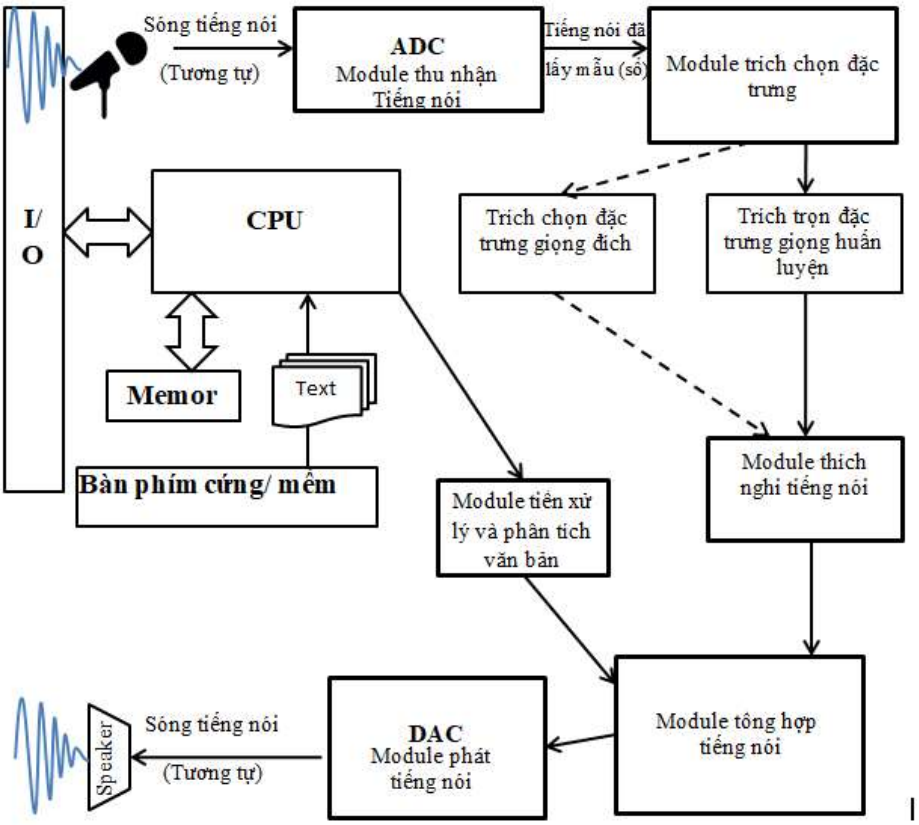
Trong phần phụ lục, sẽ trình bày các đặc tả xây dựng một ứng dụng thích nghi dựa trên các kỹ thuật đã đề xuất trong cả ba chương: Tính năng tổng hợp tiếng tiếng Việt nói đơn người nói có cảm xúc nhờ các kỹ thuật phân đoạn ngữ điệu như đã trình bày ở Chương 2. Tính năng tổng hợp nhân bản các giọng bằng cách huấn luyện với mẫu dữ liệu nhỏ từ 1- 4 phút như đã trình bày ở Chương 3; Tính năng tổng hợp nhân bản giọng với mẫu dữ liệu siêu nhỏ 1-3 giây mà mà không cần huấn luyện lại như đã trình bày ở Chương 4. Luận án sẽ thử nghiệm xây dựng một mô hình phần cứng chuyên dụng với các tài nguyên phần cứng được thiết kế đáp ứng được việc cài đặt ứng dụng thích nghi tiếng nói chuyên biệt.

Đầu tiên người dùng muốn giả giọng cần cung cấp một lượng nhỏ mẫu tiếng nói để mô hình huấn luyện thông qua đọc một vài câu văn bản trên màn hình ứng dụng (theo hướng dẫn của phần mềm). Sau đó, mẫu giọng ở dạng sóng tiếng nói được chuyển đổi từ dạng tương tự sang số sẽ được lấy mẫu, lọc nhiễu và chuyển sang module trích chọn đặc trưng theo mẫu giọng đích hoặc giọng huấn luyện sau đó mới chuyển sang module thích nghi. Sau khi đã huấn luyện được người nói mới, người dùng có thể tương tác tạo ra các giọng thích nghi bằng cách nhập các dữ liệu văn bản để đưa vào module xử lý ngôn ngữ tự nhiên, sau đó kết hợp với module thích nghi để tạo ra giọng nói tổng hợp phát ra loa ngoài (ở dạng tín hiệu tương tự) có các đặc trưng giống giọng nói đích.

*** Thiết kế phần cứng**

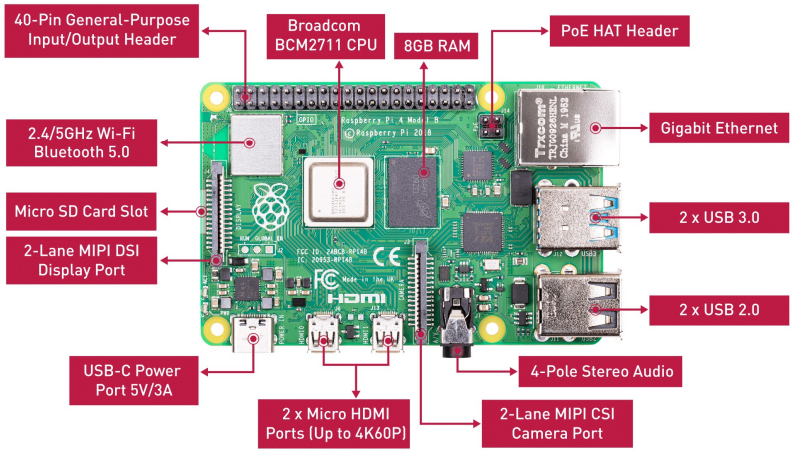


Hình 41: Sơ đồ khối hệ thống kết nối tổng thể



Hình 42: Sơ đồ khối hệ thống thích nghi giọng nói xây dựng trên hệ thống nhúng

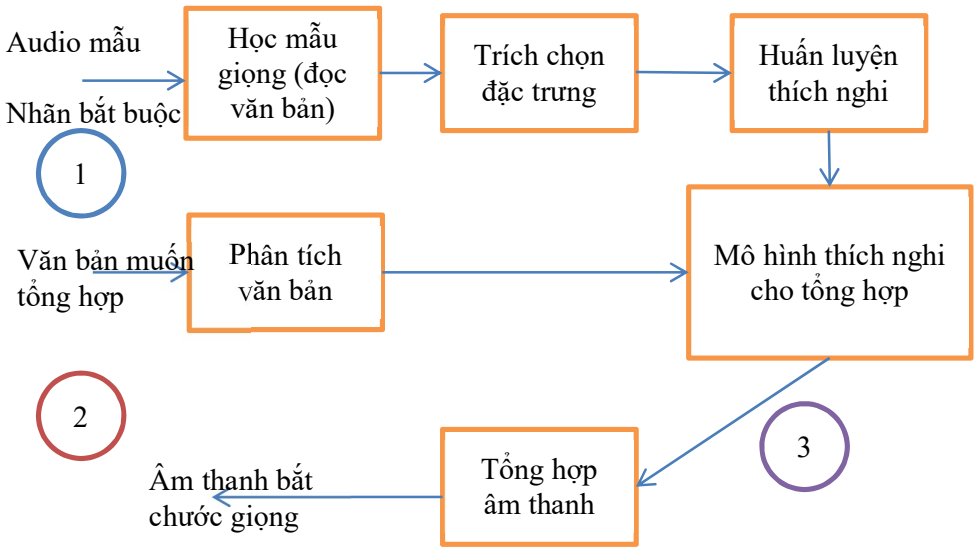
Phần cứng được sử dụng là máy tính nhúng Raspberry pi 4 Model B với cấu hình : Broadcom DCM2711 CPU Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz, Ram 8GB, Bộ nhớ lên đến 128GB, tích hợp đầy đủ các cổng kết nối Ethernet/Wifi, giao tiếp màn hình LCD/mini HDMI, các chân vào/ra logic. Các màn hình cảm ứng, pin rời và các nút bấm cứng được lắp thêm thành một thiết bị cầm tay nhỏ gọn.



Hình 43: Các cổng giao tiếp trên Raspberry Pi 4 Model B

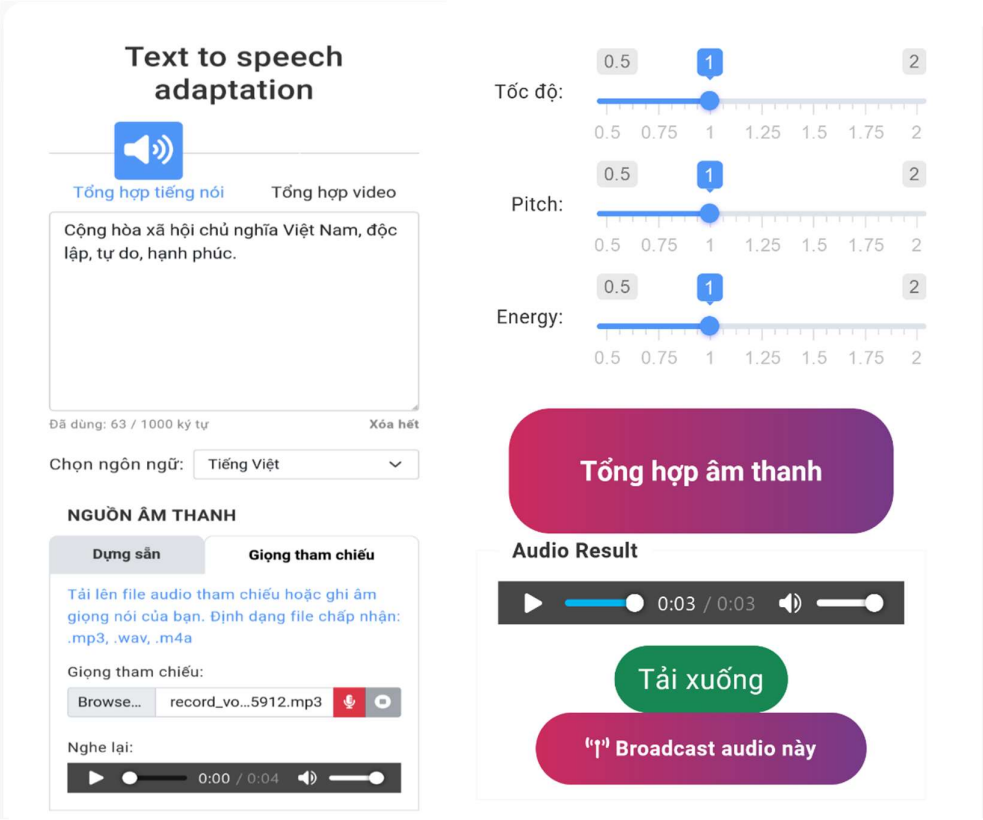
*** Thiết kế phần mềm**

Phần mềm được thiết kế theo luồng sau :



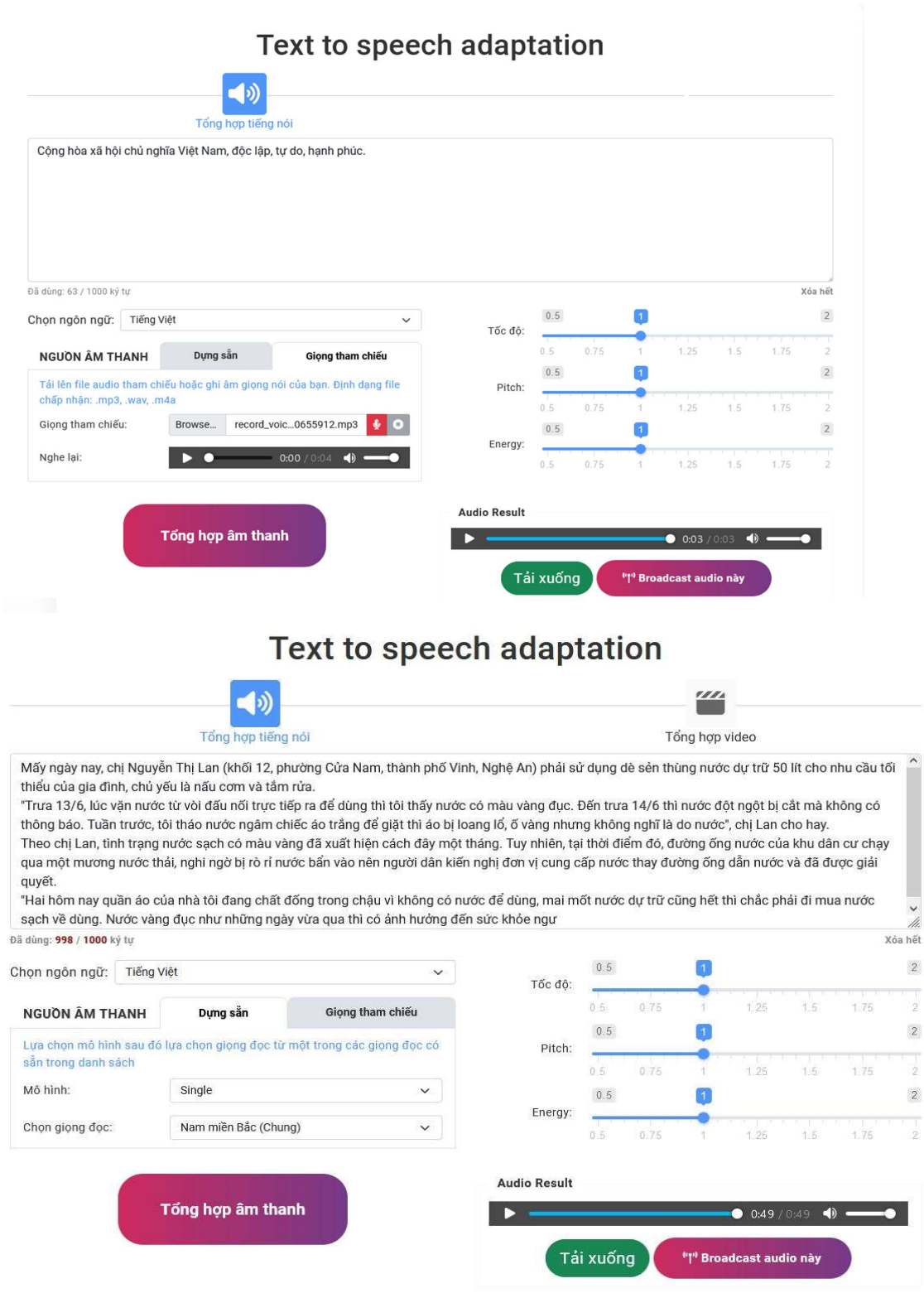
Hình 44: Sơ đồ luồng nghiệp vụ phần mềm ứng dụng bắt chước giọng

*** Giao diện trên di động**



Hình 45: Giao diện trên di động

*** Giao diện trên thiết bị nhúng hoặc để bàn**



Hình 46: Giao diện trên máy tính nhúng

ii) Địa chỉ demo

Địa chỉ demo ứng dụng tại liên kết sau : http://demo.aimed.edu.vn

iii) Chứng nhận quyền tác giả

BỘ VĂN HÓA, THỂ THAO VÀ DU LỊCH
CỤC BẢN QUYỀN TÁC GIẢ

GIẤY CHỨNG NHẬN
ĐĂNG KÝ QUYỀN TÁC GIẢ

CỤC BẢN QUYỀN TÁC GIẢ CHỨNG NHẬN

Tác phẩm: *Phần mềm Chuyển đổi văn bản thành giọng nói ADAPT - TTS*

Loại hình: *Chương trình máy tính (Không bao gồm dữ liệu)*

Tác giả: *Phạm Ngọc Phương*
Tổ 10, P. Gia Sàng, TP. Thái Nguyên, T. Thái Nguyên

Quốc tịch: *Việt Nam*
Số CCCD: *019084001343 10/04/2021*

Đã đăng ký quyền tác giả tại Cục Bản quyền Tác giả

Hà Nội, ngày 26 tháng 09 năm 2022
KT. CỤC TRƯỞNG
PHÓ CỤC TRƯỞNG

CỘNG HÒA X.H.C.N VIỆT NAM
CỤC BẢN QUYỀN TÁC GIẢ
BỘ VĂN HÓA THỂ THAO VÀ DU LỊCH

Số: 7590/2022/QTG
Cấp cho Chủ sở hữu

Phạm Thị Kim Oanh