TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP
Xây dựng mô hình thích ứng giọng nói trong tổng hợp tiếng nói tiếng Việt dựa trên công nghệ học sâu

PHAN TRUNG KIÊN
kien.pt166322@sis.hust.edu.vn

Ngành Cử nhân Công nghệ thông tin

Giảng viên hướng dẫn: PGS. TS. Đỗ Phan Thuận
Chữ ký của GVHD

Bộ môn: Khoa học Máy tính
Viện: Công nghệ thông tin và truyền thông

HÀ NỘI, 6/2020

PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

1. Thông tin về sinh viên

Họ tên sinh viên: **PHAN TRUNG KIÊN**

Điện thoại liên lạc: **0869629420** Email: **kien.pt166322@sis.hust.edu.vn**

Lớp: **CN CNTT-1-K61** Hệ đào tạo: **Cử nhân công nghệ thông tin**

Đồ án tốt nghiệp được thực hiện tại: **Trường đại học Bách Khoa Hà Nội**

Thời gian làm ĐATN: từ ngày 15/02/2020 đến 24/06/2020

2. Mục đích nội dung của ĐATN

Xây dựng mô hình thích ứng giọng nói trong tổng hợp tiếng nói tiếng Việt dựa trên công nghệ học sâu.

3. Các nhiệm vụ cụ thể của ĐATN

- Tìm hiểu cơ sở lý thuyết lĩnh vực tổng hợp tiếng nói và công nghệ học sâu.
- Xây dựng mô hình học sâu cho bài toán thích ứng giọng nói.
- Chuẩn bị dữ liệu, huấn luyện mô hình và đánh giá kết quả.

4. Lời cam đoan của sinh viên

Tôi – Phan Trung Kiên cam kết Đồ án tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của PGS.TS. Đỗ Phan Thuận. Các kết quả đạt được và nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu kham khảo. Tôi xin hoàn toàn chịu trách nhiệm nếu vi phạm quy chế của nhà trường.

Hà Nội, ngày tháng năm 2020

Tác giả ĐATN

Phan Trung Kiên

Xác nhận của giảng viên về mức độ hoàn thành ĐATN và cho phép bảo vệ:

……
……

Hà Nội, ngày tháng năm 2020

Giảng viên hướng dẫn

PGS. TS. Đỗ Phan Thuận

Lời cảm ơn

Lời cảm ơn đầu tiên tôi xin được dành cho gia đình mình và đặc biệt là mẹ tôi, người đã một mình nuôi dạy tôi từ thủa còn thơ cho đến ngày hôm nay. Nếu không có những công lao đó, chắc tôi chẳng thể khôn lớn nên người cũng như chẳng thể bước chân vào giảng đường đại học Bách Khoa để chinh phục ước mơ của mình.

Tôi xin chân thành cảm ơn PGS. TS. Đỗ Phan Thuận, người thầy đã giúp đỡ tôi trong 2 năm vừa qua cũng như hướng dẫn tôi trong quá trình hoàn thành đồ án. Nhờ thầy mà tôi đã học được nhiều kiến thức quý giá cũng như định hướng được con đường phía trước cho bản thân mình.

Tôi cũng xin cảm ơn anh Đỗ Văn Hải, anh Nguyễn Tiến Thành, anh Nguyễn Văn Thịnh và những đồng nghiệp khác tại Trung tâm không gian mạng Viettel đã giúp đỡ tôi trong quá trình thực hiện đồ án này.

Bên cạnh đó tôi cũng muốn gửi lời cảm ơn tới những người bạn đã gắn bó với tôi trong suốt quãng thời gian sinh viên. Cuối cùng tôi xin cảm ơn Bách Khoa, cảm ơn những người thầy, người cô nơi đây, cảm ơn mái nhà thứ hai đã cho tôi không chỉ là kiến thức mà còn là những kỷ niệm quý giá mà tôi sẽ không bao giờ quên. Xin cảm ơn!

Hà Nội, ngày 25 tháng 6 năm 2020

Phan Trung Kiên

Tóm tắt nội dung đồ án

Hiện nay, lĩnh vực tổng hợp tiếng nói đã được nghiên cứu và phát triển rất mạnh mẽ, đặc biệt là các phương pháp tổng hợp tiếng nói dựa trên công nghệ học sâu. Nhiều hệ thống đã được thử nghiệm và mang lại chất lượng giọng nói có chất lượng vượt trội, thậm chí khó có thể phân biệt được với giọng nói con người. Ở Việt Nam, nhiều hệ thống tổng hợp tiếng nói dựa trên công nghệ học sâu cũng đã mang lại chất lượng giọng nói rất cao như hệ tống tổng hợp tiếng nói của Viettel, hay hệ thống tổng hợp tiếng nói của Zalo, …

Với sự phát triển đó bên cạnh độ tự nhiên, các hệ thống tổng hợp tiếng nói cũng được kỳ vọng sẽ có khả năng tạo ra giọng nói của người nói tùy ý với dữ liệu đào tạo tối thiểu. Để đáp ứng vấn đề đó, thích ứng giọng nói và chuyển đổi giọng nói đã trở thành các hướng nghiên cứu chính trong lĩnh vực tổng hợp tiếng nói. Thích ứng giọng nói là nhiệm vụ tạo ra giọng nói mới cho hệ thống tổng hợp tiếng nói bằng cách điều chỉnh các tham số của một mô hình ban đầu với một lượng ít dữ liệu ghi âm của người nói mới. Đây không phải là một chủ đề mới mà là một chủ đề đã được nghiên cứ kỹ lưỡng từ lâu, đặc biệt là trên hệ thống tổng hợp tiếng nói dựa trên mô hình Markov ẩn.

Đối với tổng hợp tiếng nói tiếng Việt, đã có những nghiên cứu về chủ đề thích ứng giọng nói trên mô hình HMM, tuy nhiên với các hệ thống dựa trên công nghệ học sâu thì chưa có nghiên cứu cụ thể nào. Chính vì lý do này đề tài được đề xuất nhằm thử nghiệm các phương pháp thích ứng giọng với mục đích mở rộng sự đa dạng giọng nói cho các hệ thống tổng hợp tiếng nói với lượng dữ liệu cần bổ sung là tối thiểu. Để thực hiện điều này tác giả đề xuất sử dụng hai phương pháp transfer learning và sử dụng véc tơ mã hóa người nói.

Sau đây là bố cục chính của đồ án

- CHƯƠNG 1 - TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI VÀ VẤN ĐỀ ĐẶT RA CHO ĐỒ ÁN: Chương này giới thiệu chung về tổng hợp tiếng nói, tình hình nghiên cứu và phát triển hệ thống tổng hợp tiếng nói tiếng Việt cũng như đặt ra vấn đề cho đồ án.
- CHƯƠNG 2 - CƠ SỞ LÝ THUYẾT: Chương này chủ yếu nói về phương pháp học sâu là ứng dụng của nó trong tổng hợp tiếng nói cũng như thích ứng giọng nói.
- 34CHƯƠNG 3 - PHƯƠNG PHÁP ĐỀ XUẤT CHUYỂN ĐỔI GIỌNG NÓI TIẾNG VIỆT: Chương này chủ yếu nói về kiến trúc của hệ thống tổng hợp tiếng nói cũng như các phương pháp đề xuất cho thích ứng giọng nói trên các hệ thống đó.
- CHƯƠNG 4 - THỬ NGHIỆM VÀ VÀ ĐÁNH GIÁ: Chương này nói về cách thức cài đặt, thử nghiệm và đánh giá kết quả hệ thống thích ứng giọng nói.
- CHƯƠNG 5 - KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN: Chương này có nội dung kết luận về kết quả đạt được của đồ án cũng như những phương hướng nghiên cứu, cải thiện.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT VÀ THUẬT NGỮ

Từ viết tắt	Từ đầy đủ	Ý nghĩa
TTS	Text to speech	Tổng hợp văn bản thành giọng nói
HMM	Hidden Markov model	Mô hình Markov ẩn
GMM	Gaussian mixture model	Mô hình Gaussian hỗn hợp
ANN	Artificial neural network	Mạng nơ ron nhân tạo
DNN	Deep neural network	Mạng nơ ron học sâu
Seq2seq	Sequence to sequence	
PSOLA	Pitch synchronous overlap and add	Kỹ thuật chồng đồng bộ cao độ tần số cơ bản
MOS	Mean opinion score	Điểm ý kiến trung bình
F0	Fundamental frequency	Tần số cơ bản
MAP	Maximum a posterior	
MLLR	Maximum likelihood linear regression	
MSE	Mean squared error	Sai số toàn phương trung bình
LHUC	Learning hidden unit contribution	
FST	Feature space transformation	
PCA	Principal Component Analysis	Phương pháp Phân tích thành phần chính

CHƯƠNG 1. TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI VÀ VẤN ĐỀ ĐẶT RA CHO ĐỒ ÁN

1.1 Giới thiệu về tổng hợp tiếng nói

1.1.1 Định nghĩa và quá trình phát triển tổng hợp tiếng nói

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói của con người từ văn bản hoặc các mã hóa việc phát âm. Ở thời điểm hiện tại, khi nhắc đến hệ thống tổng hợp tiếng nói, đa số ám chỉ hệ thống chuyển đổi văn bản thành giọng nói (text-to-speech).

Từ lâu trước khi các kỹ thuật xử lý tín hiệu điện tử được phát minh, các nhà nghiên cứu giọng nói đã cố gắng xây dựng các máy móc bắt chước giọng nói của người. Các hệ thống đầu tiên ra đời vào cuối thế kỷ XVIII đầu thế kỷ XIX là các máy cơ học mô phỏng thanh quản con người. Năm 1779, nhà khoa học người Đan Mạch Christian Kratzenstein, lúc đó làm việc tại Viện Hàn lâm Khoa học Nga, xây dựng một mô hình có thể bắt chước giọng nói người với năm nguyên âm ([a], [e], [I], [o] và [u]) [1]. Máy này sau đó được cải tiến thành 'Máy Phát âm Cơ khí-Âm học' của Wolfgang von Kempelen ở Viên, Áo, theo mô tả máy tạo ra mô hình lưỡi và môi cho phép tạo ra phụ âm thêm vào cho nguyên âm [2].

Vào đầu thế kỷ XX, sự ra đời của các hệ thống điện đã mang lại một sự thay đổi lớn trong các thiết bị tổng hợp tiếng nói. Năm 1930, Phòng thí nghiệm Bell tạo ra máy VOCODER, một máy phân tích và tổng hợp giọng nói điều khiển bằng bàn phím, được mô tả là phát âm rõ ràng, Homer Dudley cải tiến cỗ máy này thành VODER, và trưng bày nó tại New York World's Fair 1939 [3]. Từ cuối những năm 1950, các hệ thống tổng hợp tiếng nói dựa trên máy tính đầu tiên xuất hiện. Noriko Umeda và cộng sự đã phát triển hệ thống chuyển văn bản thành tiếng Anh nói chung đầu tiên vào năm 1968, tại Phòng thí nghiệm kỹ thuật điện tại Nhật Bản [4].

Từ đó đến nay, công nghệ tổng hợp tiếng nói đã có những bước tiến bộ vượt bậc, chất lượng giọng nói tổng hợp ngày càng có độ tự nhiên cao và khó phân biệt với giọng người thật. Mục tiêu phát triển của các hệ thống tổng hợp tiếng nói cũng mở rộng như việc thể hiện cảm xúc cho lời nói hay sự đa dạng giọng nói trong một hệ thống.

1.1.2 Ứng dụng của tổng hợp tiếng nói

Bên cạnh sự phát triển về mặt chất lượng giọng nói tổng hợp, các hệ thống tổng hợp tiếng nói cũng được áp dụng ngày càng rộng rãi trong nhiều lĩnh vực của cuộc sống. Phổ biến nhất có thể kể tới các ứng dụng sách nói, báo nói. Với số lượng rất lớn sách mới xuất bản mỗi năm và tin tức cập nhập mỗi ngày, việc thu âm bằng giọng phát thanh viên trở thành tốn kém và bất khả thi. Hệ thống tổng hợp tiếng nói với độ tự nhiên và tốc độ xử lý nhanh chính là giải pháp cho vấn đề này.

Tiếp theo có thể kể đến các ứng dụng trợ lý ảo như Siri của Apple[1], Google Assistant của Google[2], Cortana của Mircrosoft[3], ... đều áp dụng công nghệ tổng

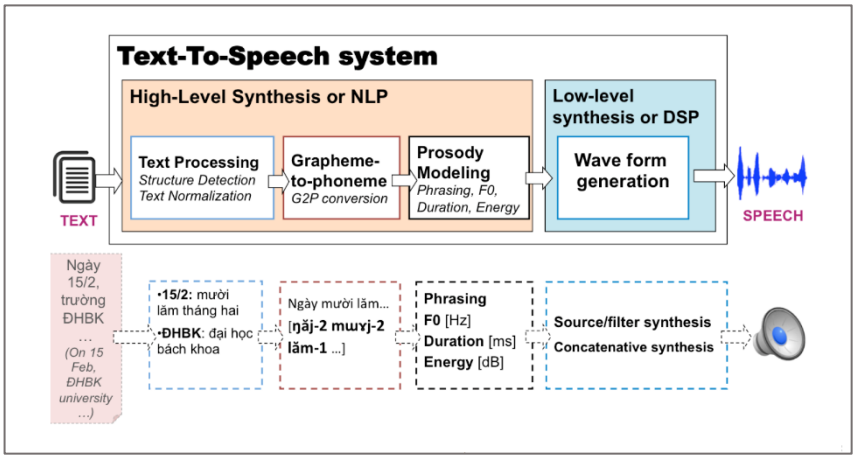
[1] https://www.apple.com/ios/siri
[2] https://assistant.google.com
[3] https://www.microsoft.com/en-us/cortana

hợp tiếng nói để nâng cao trải nghiệm tương tác giữa người sử dụng và máy. Hay như các tổng đài trả lời tự động áp dụng các công nghệ nhận dạng tiếng nói, xử lý ngôn ngữ tự nhiên và tổng hợp tiếng nói để phục vụ nhiều cuộc gọi cùng một lúc mà không bị giới hạn bởi số lượng tư vấn viên.

Có thể thấy rằng việc phát triển công nghệ tổng hợp tiếng nói là cần thiết vì tính ứng dụng và giá trị mà công nghệ này mang lại.

1.1.3 Thành phần của tổng hợp tiếng nói

Hiện nay, đa số các hệ thống tổng hợp tiếng nói đều bao gồm hai thành phần chính: phần xử lý ngôn ngữ tự nhiên và phần xử lý tổng hợp tiếng nói [5]. Phần xử lý ngôn ngữ tự nhiên có nhiệm vụ chuẩn hóa, xứ lý các văn bản đầu vào thành các thành phần có thể phát âm được. Phần xử lý tổng hợp tiếng nói có nhiệm vụ tạo ra tín hiệu tiếng nói từ các thành phần phát âm được nêu trên. Hình 1.1 mô tả một hệ thống tổng hợp tiếng nói gồm hai thành phần trên.



Hình 1.1 Sơ đồ tổng quát một hệ thống tổng hợp tiếng nói [23]

a) Xử lý ngôn ngữ tự nhiên trong tổng hợp tiếng nói

Trong hệ thống tổng hợp tiếng nói, khối xử lý ngôn ngữ tự nhiên có nhiệm vụ phát sinh các thông tin về ngữ âm và ngữ điệu cho việc đọc văn bản đầu vào. Thông tin ngữ âm cho biết những âm nào sẽ được phát ra, trong ngữ cảnh cụ thể nào, thông tin ngữ điệu mô tả điệu tính của các âm được phát [5]. Quá trình xử lý ngôn ngữ tự nhiên thường bao gồm 3 bước:

- Xử lý và chuẩn hóa văn bản (Text Processing)
- Chuyển đổi hình vị sang âm vị (Grapheme to phoneme).
- Phát sinh các thông tin ngôn điệu, ngữ âm cho văn bản (Prosody modeling).

Chuẩn hóa văn bản là quá trình chuyển đổi văn bản thô ban đầu thành một văn bản dạng chuẩn, có thể đọc được một cách dễ dàng, ví dụ như chuyển đổi các số, từ viết tắt, ký tự đặc biệt,... thành dạng viết đầy đủ và chính xác.

Phân tích cách phát âm là quá trình xác định cách phát âm chính xác cho từng từ trong văn bản, quá trình này còn được gọi là chuyển đổi văn bản sang chuỗi âm vị. Có hai cách cơ bản để xác định cách cho văn bản, cách thứ nhất và cũng là cách đơn giản hơn đó là dựa vào từ điển, sử dụng một từ điển có chứa tất cả các từ của một ngôn ngữ và chưa cách phát âm đúng tương ứng cho mỗi từ. Việc xác định

cách phát âm cho văn bản chỉ đơn giản là tra từ điển và thay thế đoạn văn bản bằng chuỗi âm vị đã lưu trong từ điển. Ưu điểm của cách này đó là tốc độ nhanh và tính chính xác, nhưng nhược điểm đó là yêu cầu lượng từ vựng lưu trữ lớn và không hoạt động trong trường hợp từ không có trong từ điểm. Cách thứ hai là dựa trên các quy tắc và sử dụng quy tắc để để tìm ra cách phát âm tương ứng. Cách này phù hợp với mọi văn bản nhưng độ phức tạp có thể tăng cao nếu ngôn ngữ có nhiều trường hợp bất quy tắc.

Phát sinh các thông tin ngôn điệu cho văn bản là việc xác định vị trí trọng âm của từ được phát âm, sự lên xuống giọng ở các vị trí khác nhau trong câu và xác định các biến thể khác nhau của âm phụ thuộc vào ngữ cảnh khi được phát âm trong một ngôn ngữ lưu liên tục, ngoài ra quá trình này còn phải xác định điểm dừng nghỉ, lấy hơi khi phát âm hoặc đọc một đoạn văn bản. Thông tin về thời gian thường được đo bằng mi li giây và được ước lượng dựa trên các quy tắc hoặc các thuật toán học máy. Cao độ (pitch) là một tương quan về mặt cảm nhận của tần số cơ bản F0, được biểu thị theo đơn vị Hz hoặc phân số của tông (tones) (nửa tông, một phần hai tông). Tần số cơ bản F0 là một đặc trưng quan trọng trong việc tạo ngôn điệu của tín hiệu tiếng nói, do đó việc tạo các đặc trưng cao độ là một vấn đề phức tạp và quan trọng trong tổng hợp tiếng nói.

b) Xử lý tổng hợp tín hiệu tiếng nói

Khối xử lý tổng hợp tính hiệu tiếng nói đảm nhiệm việc tạo ra tín hiệu tiếng nói từ các thông tin ngữ âm và ngữ điệu do khối phân tích xử lý ngôn ngữ tự nhiên cung cấp.

Chất lượng tiếng nói tổng hợp được đánh giá thông qua hai khía cạnh: mức độ dễ hiểu nội dung và mức độ tự nhiên. Mức độ dễ hiểu đề cập đến nội dung của tiếng nói tổng hợp có thể hiểu được dễ dàng hay không. Mức độ tự nhiên của tiếng nói tổng hợp là sự so sánh độ giống nhau giữa giọng nói tổng hợp và giọng nói tự nhiên của con người.

Một hệ thống tổng hợp tiếng nói lý tưởng cần phải vừa dễ hiểu vừa tự nhiên và mục tiêu xây dựng hệ thống tổng hợp tiếng nói là cải thiện đến mức tối đá hai tính chất này [6].

1.2 Các phương pháp tổng hợp tiếng nói

1.2.1 Tổng hợp mô phỏng hệ thống phát âm

Tổng hợp mô phỏng hệ thống phát âm là các kỹ thuật tổng hợp giọng nói dựa trên mô hình máy tính mô phỏng cơ quan phát âm của con người và các quá trình phát âm tại đó. Về mặt lý thuyết, đây được xem là phương pháp cơ bản nhất để tổng hợp tiếng nói, nhưng cũng vì thế mà phương pháp này khó thực hiện và tính toán nhất, do đó khó có thể tổng hợp được tiếng nói chất lượng cao [7]. Tổng hợp mô phỏng phát âm đã từng chỉ là hệ thống dành cho nghiên cứu khoa học cho mãi đến những năm gần đây. Lý do là rất ít mô hình tạo ra âm thanh chất lượng đủ cao hoặc có thể chạy hiệu quả trên các ứng dụng thương mại. Một ngoại lệ là hệ thống dựa trên NeXT; vốn được phát triển và thương mại hóa bởi Trillium Sound Research Inc, ở Calgary, Alberta, Canada.

1.2.2 Tổng hợp tần số formant

Tổng hợp tần số formant, hay còn được gọi là tổng hợp formant, là kỹ thuật tổng hợp tiếng nói âm học cơ bản nhất, sử dụng lý thuyết mô hình nguồn lọc để tạo tiếng nói. Mô hình này mô phỏng hiện tượng cộng hưởng của các cơ quan phát âm bằng một tập các bộ lọc. Các bộ lọc này còn được gọi là các bộ cộng hưởng formant, chúng có thể được kết hợp song song hoặc nối tiếp với nhau hoặc kết hợp cả hai. Phương pháp tổng hợp formant không phải sử dụng trực tiếp mẫu giọng thật nào khi thực hiện tổng hợp tiếng nói. Thay vào đó, tín hiệu âm thanh được tổng hợp dựa trên một mô hình tuyến âm (vocal tract). Các thông số như tần số cơ bản, giọng nói và mức độ tiếng ồn được thay đổi theo thời gian để tạo ra dạng sóng của lời nói nhân tạo [8]. Tuy nhiên, phương pháp phân tích tổng hợp vẫn cần mẫu giọng thật ở bước phân tích để có thể trích rút được các đặc trưng formant, trường độ hay năng lượng tiếng nói.

Hệ thống tổng hợp tiếng nói dựa trên phương pháp tổng hợp tần số formant có những ưu điểm, nhược điểm có thể kể đến như: Nhược điểm của hệ thống này là tạo ra giọng nói không tự nhiên, nghe cảm giác rất phân biệt với giọng người thật và phụ thuộc nhiều vào chất lượng của quá trình phân tích tiếng nói của từng ngôn ngữ. Tuy nhiên độ tự nhiên cao không phải lúc nào cũng là mục đích của hệ thống và hệ thống này cũng có các ưu điểm riêng của nó, hệ thống này khá dễ nghe, không có tiếng cọ sát do ghép âm tạo ra, các hệ thống này cũng nhỏ gọn vì không chứa cơ sở dữ liệu mẫu âm thanh lớn.

1.2.3 Tổng hợp ghép nối

Tổng hợp ghép nối là phương pháp tổng hợp tiếng nói bằng cách ghép vào nhau các đoạn tín hiệu tiếng nói của một giọng nói đã được ghi âm. Các giọng nói sau khi được ghi âm sẽ được chia thành các câu, các câu sẽ chia thành các đơn vị âm. Các đơn vị âm phổ biến là âm vị, âm tiết, bán âm tiết, âm đôi, âm ba, từ, cụm từ. Trong quá trình chạy, hệ thống tổng hợp ghép nối sẽ sắp xếp và nối các đơn vị âm đã có để thu được đoạn tiếng nói yêu cầu. Do đặc tính tự nhiên của tiếng nói được lưu trữ trong các đơn vị âm, nên tổng hợp ghép nối là phương pháp có khả năng tổng hợp được giọng nói với độ dễ hiểu và độ tự nhiên cao. Tuy nhiên, sự gián đoạn tại các điểm ghép nối có thể khiến cho âm thanh biến dạng, mặc dù đã sử dụng biện pháp và thuật toán làm trơn tín hiệu tại chỗ ghép nối. Bên cạnh đó, tập hợp các đơn vị luôn bị hạn chế về số lượng cũng như nội dung, điều này dẫn đến tiếng nói tổng hợp nghe thô rát. Ngoài ra, để có thể lưu trữ được tất cả các đơn vị âm cần thiết cho một lượng đủ lớn các giọng người nói khác nhau, với nhiều ngữ cảnh và đặc trưng trạng thái, thì cần phải có một không gian rất lớn và tốc độ tính toán, truy vấn của hệ thống cao, do đó điều này là không kinh tế.

Có ba kiểu tổng hợp ghép nối:

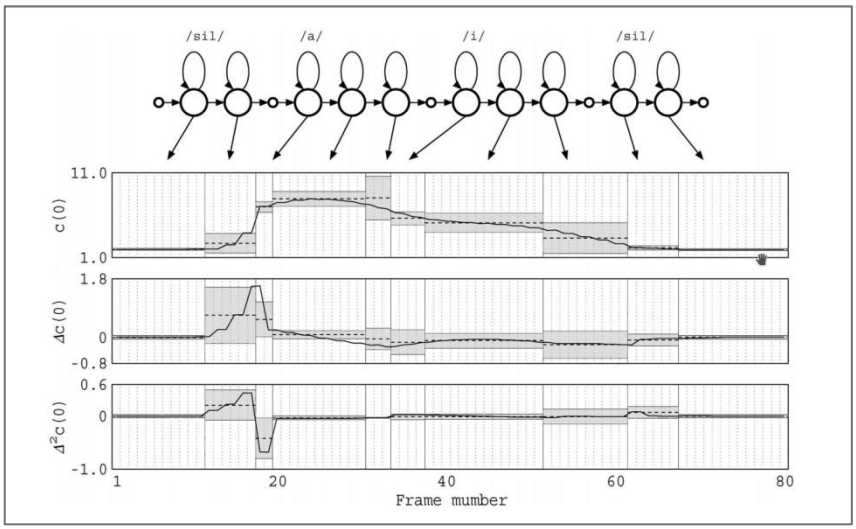
- Tổng hợp chọn đơn vị (unit selection)
- Tổng hợp âm kép (diphone)
- Tổng hợp chuyên biệt (domain-specific)

Tổng hợp chọn đơn vị dùng một cơ sở dữ liệu lớn các giọng nói ghi âm. Trong đó, mỗi câu được tách thành các đơn vị khác nhau như: các tiếng đơn lẻ, các âm tiết, hình vị, từ, nhóm từ hoặc câu văn. Một bảng tra các đơn vị được lập ra dựa trên các phần đã tách và các thông số âm học như tần số cơ bản, thời lượng, vị trí của âm tiết và các tiếng gần nó. Khi thực hiện tổng hợp, các câu phát biểu tạo ra bàng cách xác định chuỗi đơn vị phù hợp nhất từ cơ sở dữ liệu. Quá trình này được gọi là chọn đơn vị và thường sử dụng thuật toán cây quyết định để thực hiện. Ưu điểm của phương pháp này là có thể tạo được giọng nói có độ tự nhiên cao tuy nhiên nhược điểm đó là cần một cơ sở dữ liệu lớn chứa các đơn vị để lựa chọn.

Tổng hợp âm kép dùng một cơ sở dữ liệu giọng nói nhỏ chưa tất cả các âm kép xuất hiện trong ngôn ngữ đang xét. Số lượng âm kép phụ thuộc vào đặc tính ghép âm học của ngôn ngữ. Trong tổng hợp âm kép, chỉ có một mẫu của âm kép được lưu trữ trong cơ sở dữ liệu. Khi chạy, lời văn được chồng lên các đơn vị này bằng kỹ thuật xử lý tín hiệu số như mã tiên đoán tuyến tính, PSOLA hay MBROLA. Chất lượng của âm thanh tổng hợp theo cách này không cao bằng phương pháp chọn đơn vị nhưng tự nhiên hơn so với phương pháp tổng hợp cộng hưởng tần số. Ưu điểm của nó là có kích thước dữ liệu nhỏ.

Tổng hợp chuyên biệt ghép nối các từ và đoạn văn đã được ghi âm để tạo ra lời phát biểu. Nó được dùng trong các ứng dụng có các văn bản chuyên biệt cho một chuyên ngành, sử dụng lượng từ vựng hạn chế, như các thông báo chuyến bay hay dự báo thời tiết. Công nghệ này rất đơn giản, và đã được thương mại hóa từ lâu, đã đi vào các đồ vật như đồng hồ biết nói hay máy tính bỏ túi biết nói. Mức độ tự nhiên của các hệ thống này có thể rất cao vì số lượng các câu nói không nhiều và khớp với lời văn và âm điệu của giọng nói ghi âm. Tuy nhiên các hệ thống này bị hạn chế bởi cơ sở dữ liệu chuyên ngành, không phục vụ mọi mục đích mà chỉ hoạt động với các câu nói mà chúng đã được lập trình sẵn.

1.2.4 Tổng hợp dùng tham số thống kê HMM

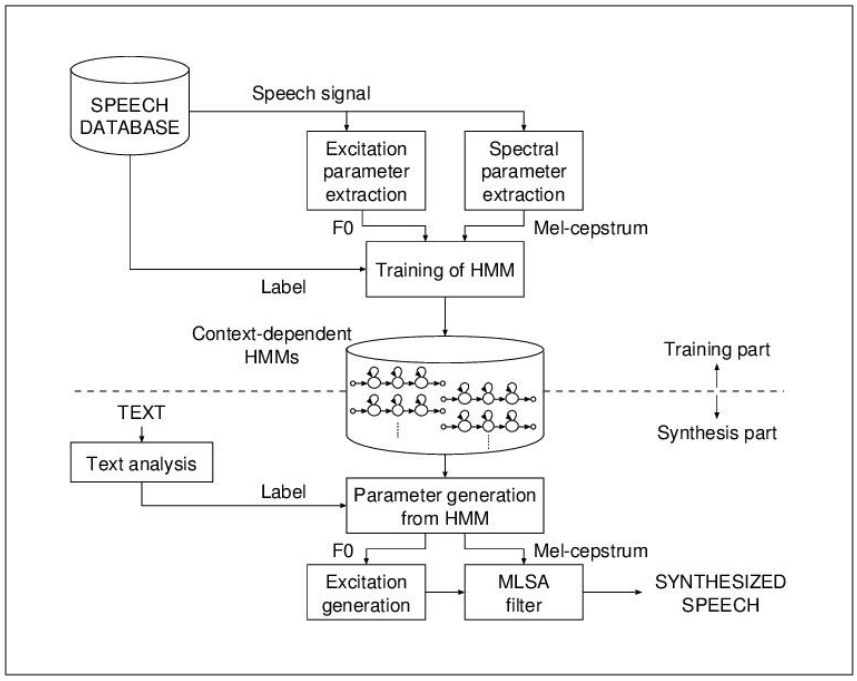


Hình 1.2 Ví dụ về thống kê và các tham số HMM cấp câu bao gồm các HMM cấp âm vị là /a/ và /i/ [36].

Phương pháp tổng hợp tiếng nói dùng tham số thống kê HMM là phương pháp dựa trên mô hình Markov ẩn (HMM). Ở đây, HMM là một mô hình thống kê, được sử dụng để mô hình hóa các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể.

Hình 1.2 là ví dụ về thống kê và các tham số HMM cấp câu được tạo bằng cách ghép hai HMM cấp âm vị (cụ thể là âm vị /a/ và /i/). Đường nét đứt và tô bóng hiển thị kỳ vọng, độ lệch chuẩn cuả phân phối Gausian (phân phối chuẩn) ở mỗi trạng thái.

Mô hình Markov ẩn là một mô hình học máy dựa trên thông kế, do đó hệ thống tổng hợp tiếng nói dựa trên mô hình Markov ẩn hoạt động bao gồm hai qúa trình là quá trình huấn luyện và quá trình tổng hợp. Hình 1.3 mô tả hai quá trình nêu trên.



Hình 1.3 Quá trình huấn luyện và tổng hợp của một hệ thống tổng hợp tiếng nói HMM [39]

Quá trình huấn luyện mô hình bao gồm các bước: trích chọn đặc trưng tiếng nói, trích chọn đặc trưng ngôn ngữ và huấn luyện mô hình. Các đặc trưng tiếng nói được trích trong quá trình huấn luyện là Mel-cepstrum và tần số cơ bản F0. Các đặc trưng ngôn ngữ của văn bản được mô tả bằng cách sử dụng một bộ phân cụm (thường là cây quyết định) để gom các cụm trạng thái của mô hình Markov ẩn có đặc tính ngôn ngữ gần nhau nhất và bầu chọn ra một trạng thái tiêu biểu để thay thế cho các trạng thái còn lại trong cụm.

Hệ thống tổng hợp tiếng nói dùng tham số thống kê HMM là một hệ thống có khả năng tạo ra tiếng nói mang các phong cách nói khác nhau, với đặc trưng của nhiều người nói khác nhau, thậm chí mang cả cảm xúc của người nói. Ưu điểm của phương pháp này là cần ít bộ nhớ lưu trữ và tài nguyên hệ thống hơn so với tổng hợp dựa trên ghép nối và có thể điều chỉnh tham số để thay đổi ngữ điệu, thay đổi các đặc trưng người nói. Tuy nhiên, mức độ tự nhiên trong tiếng nói tổng hợp của

các hệ thống này thường bị suy giảm so với tổng hợp tiếng nói dựa trên ghép nối. Mặc dù có nhiều ưu điểm, nhưng hệ thống tổng hợp tiếng nói dùng tham số thống kê HMM vẫn còn tồn tại nhiều nhược điểm. Trong hệ thống này, phổ tín hiệu và tần số cơ bản được ước lượng từ các giá trị xấp xỉ trung bình của phổ và tần số cơ bản, phát xạ từ các HMM được huấn luyện từ nhiều dữ liệu khác nhau. Các đặc trưng ngôn điệu của tiếng nói thu âm gốc có thể bị thay thế bởi các đặc trưng "trung bình" này, khiến cho tiếng nói tổng hợp nghe có vẻ "đều đều", quá "mịn" hay quá "ổn định". Đặc điểm quá "mịn" của tiếng nói tổng hợp dựa trên HMM vẫn có thể chấp nhận được khi chỉ chú ý đến tính chất nghe hiểu. Nhưng chính những hạn chế này khiến cho tiếng nói tổng hợp dựa trên HMM nghe như bị "nghẹt mũi" và làm giảm ngôn điệu, sắc thái cảm xúc hay phong cách nói trong câu nói.

1.2.5 Tổng hợp bằng phương pháp lai ghép

Tổng hợp lai ghép là hướng tiếp cận tổng hợp phương pháp lai ghép giữa tổng hợp ghép nối chọn đơn vị và tổng hợp tham số thống kê HMM nhằm tận dụng ưu điểm của mỗi phương pháp trong hệ thống mới.

Một cách tiếp cận là sử dụng các mô hình HMM để làm mịn các điểm ghép nối của phương pháp tổng hợp lựa chọn đơn vị. Mặc dù cách tiếp cận này có thể cải thiện sự gián đoạn tại vị trí ghép nối, nhưng nó lại tạo ra thành phần không mong muốn khi có sự nhầm lẫn giữa các hệ số làm mịn và tín hiệu nguồn kích thích.

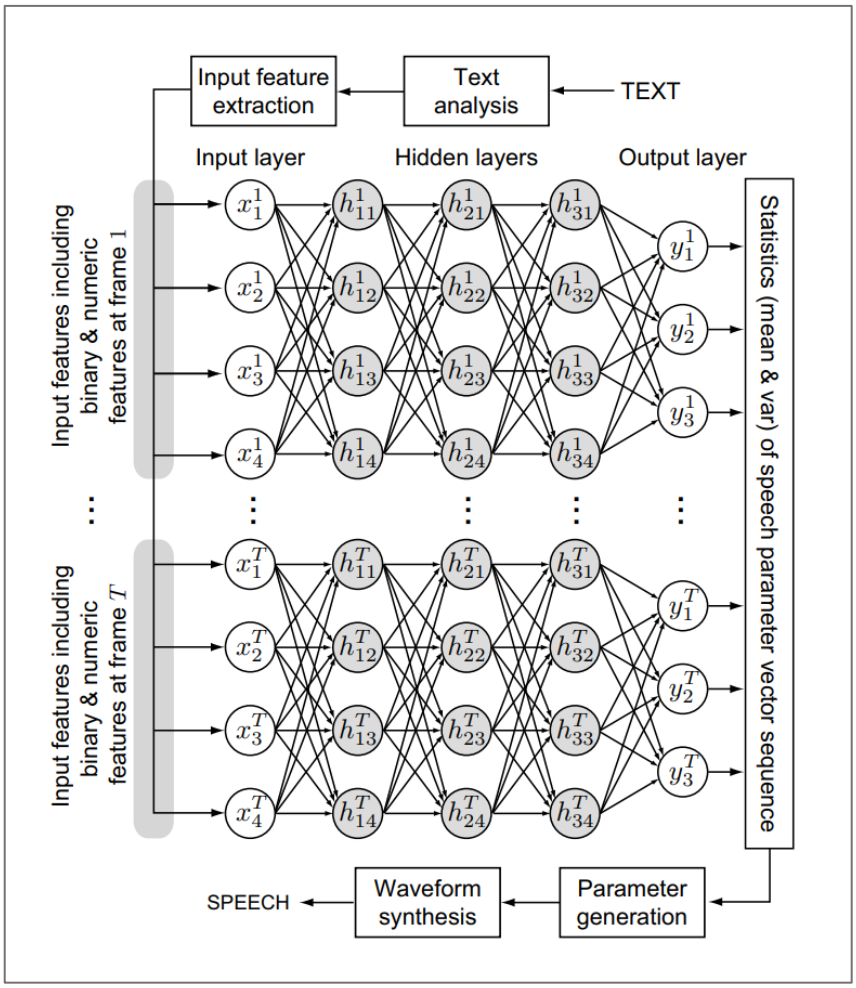
Một hình thức lai ghép khác là sử dụng các tham số phổ, tần số cơ bản và thời gian trạng thái sinh ra từ các HMM để tính toán chi phí mục tiêu và chi phí ghép nối cho quá trình ghép nối lựa chọn đơn vị. Phương pháp lai ghép này có thể cải thiện chất lượng và tính ổn định của tiếng nói tổng hợp và vẫn bảo toàn tính ưu việt của hệ thống TTS dựa trên HMM là thích nghi, thay đổi đặc trưng người nói trong điều kiện dữ liệu huấn luyện hạn chế.

1.2.6 Tổng hợp tiếng nói dựa trên phương pháp học sâu

Tổng hợp tiếng nói dựa trên phương pháp học sâu đã phát triển mạnh mẽ trong những năm gần đây, phương pháp được xây dựng dựa trên việc mô hình hóa mô hình âm học bằng một mạng nơ ron nhân tạo (ANN). Trong đó văn bản đầu vào sẽ được chuyển hóa thành một véc tơ đặc trưng ngôn ngữ. Mô hình âm học dựa trên mạng nơ ron nhân tạo sẽ lấy đầu vào là véc tơ đặc trưng ngôn ngữ này và tạo ra các đặc trưng âm học tương ứng ở đầu ra. Một bộ tổng hợp tín hiệu tiếng nói (vocoder) sẽ tạo ra tín hiệu tiếng nói từ những đặc trưng âm học thu được ở trên.

Mạng nơ ron nhân tạo được sử dụng cho mô hình âm học có thể là mạng nơ ron học sâu (DNN), mạng nơ ron tích chập (CNN), mạng nơ ron hồi tiếp (RNN) hay mạng bộ nhớ dài-ngắn hạn (LSTM). Trong đó được sử dụng phổ biến nhất là mạng nơ ron học sâu DNN. Hình 1.4 mô tả một mô hình tổng hợp tiếng nói dựa trên DNN. Trong đó, văn bản đầu vào sẽ đi qua một bộ phân tích văn bản (Text analysis) để trích chọn các đặc trưng ngôn ngữ và được chuyển hóa thành các véc tơ đặc trưng qua nhờ vào bộ Input feature extraction. Các véc tơ này được đưa vào mạng theo từng khung (frame), đầu ra của mạng sẽ là các đặc trưng chứa thông tin về phổ và tín hiệu kích thích. Các đặc trưng này thông qua bộ tạo tham số

(Parameter Generation) sẽ được chuyển thành các đặc trưng âm học và được đưa v ào bộ tạo tín hiệu tiếng nói (Waveform generation) để tạo ra tín hiệu tiếng nói chính là đầu ra của toàn bộ mô hình.



Hình 1.4 Một mô hình tổng hợp tiếng nói dựa vào DNN [35]

Ngoài ra một mô hình deep learning mới hơn đang dần thay thế cho các mô hình DNN đó là mô hình Sequence to sequence (seq2seq). Seq2seq được giới thiệu bởi nhóm nghiên cứu của Google vào năm 2014 trong bài báo "Sequence to Sequence with Neural Networks" [9]. Mặc dù mục đích ban đầu của mô này là để áp dụng trong bài toán dịch máy, tuy nhiên hiện nay seq2seq cũng được áp dụng nhiều trong các hệ thống khác như nhận diện giọng nói, tóm tắt văn bản, đặt tiêu đề cho ảnh, tổng hợp tiếng nói, … Seq2seq gồm 2 phần chính là Encoder và Decoder. Cả hai thành phần này đều được hình thành từ các mạng Neural Networks, trong đó Encoder có nhiệm vụ chuyển đổi dữ liệu đầu vào (input sequence) thành một biểu diễn với số chiều thấp còn Decoder có nhiệm vụ tạo ra đầu ra (output sequence) từ biểu diễn được tạo ra ở phần Encoder. Trong tổng hợp tiếng nói, các hệ thống sử dụng mô hình seq2seq đã cho thấy chất lượng giọng nói tổng hợp vượt trội so với các phương pháp cũ với độ tự nhiên tương đương với giọng nói con người. Ở Bảng 1.1 ta có thể thấy rằng điểm đánh giá độ tự nhiên của mô hình Tacotron 2 [10] là vượt trội so với các mô hình khác và tương đương với giọng nói tự nhiên.

Bảng 1.1 Đánh giá một số mô hình tổng hợp tiếng nói trên bộ dữ liệu North American English

Mô hình	MOS
Giọng nói tự nhiên	4.582 ± 0.053
Tacotron 2 [10]	4.526 ± 0.066
Wavenet (Linguistic) [10]	4.210 ± 0.081
Wavenet (L+F) [11]	4.341 ± 0.051
HMM-driven concatenative [11]	3.860 ± 0.137
LSTM-RNN parametric [11]	3.670 ± 0.098

Ưu điểm của các hệ thống tổng hợp tiếng nói dựa trên phương pháp học sâu đó chính là độ tự nhiên của giọng nói tổng hợp. Tuy nhiên nhược điểm của phương pháp này đó là lượng dữ liệu cần để huấn luyện mô hình rất lớn, cùng với đó là thời gian huấn luyện mất hàng chục tiếng thậm chí hàng tuần và yêu cầu về hiệu năng máy tính rất lớn. Do đó chi phí để xây dựng những hệ thống này là rất lớn.

1.3 Tình hình phát triển và các vấn đề với tổng hợp tiếng nói tiếng Việt

Ở Việt Nam trong những năm vừa qua, nhờ vào sự phát triển của công nghệ thông tin cũng như sự phát triển kinh tế đã tạo điều kiện về mặt công nghệ cũng như cơ sở vật vật chất để có thể nghiên cứu và triển khai các ứng dụng về khoa học công nghệ. Lĩnh vực tổng hợp tiếng nói tiếng Việt cũng không nằm ngoài xu thế phát triển đó, nhiều hệ thống tổng hợp tiếng nói tiếng Việt đã có những thành tựu đáng kể. Các hệ thống đầu tiên ra đời như VietVoice, VnSpeech, Vais, hệ thống tổng hợp tiếng nói của tập đoàn FPT hay hệ thống tổng hợp tiếng nói Hoa súng. Trong đó các hệ thống này được xây dựng dựa theo hai hướng phổ biến là tổng hợp ghép nối và tổng hợp sử dụng tham số thống kê.

Đối với phương pháp tổng hợp tiếng nói ghép nối dành cho tiếng Việt thì đã có rất nhiều hệ thống được phát triển, có thể kể đến như hệ thống Hoa Súng [12], được phát triển lần đầu vào năm 2007, dữ liệu để xây dựng hệ thống này được gọi là VNSpeech Corpus, nó được thu thập và lọc từ nhiều nguồn khác nhau như truyện, sách, … Dữ liệu này bao gồm nhiều loại khác nhau như: các từ với đầy đủ sáu thanh điệu, các số, câu thoại, đoạn văn ngắn, … Đến năm 2011 hệ thống được mở rộng [13], sử dụng kỹ thuật lựa chọn âm vị không đồng nhất. Phiên bản này cũng sử dụng cùng bộ dữ liệu ở phiên bản trước, nhưng được đánh chú thích ở mức độ âm tiết với những thông tin cần thiết như các thành phần âm vị, thanh điệu, thời gian, năng lượng, và những đặc trưng ngữ cảnh khác.

Đối với phương pháp tổng hợp tiếng nói sử dụng tham số thống kê, hay là tổng hợp tiếng nói dựa trên mô hình Markov ẩn (HMM). Ở Việt Nam cũng đã có nhiều hệ thống tổng hợp tiếng nói phát triển dựa trên phương pháp này, có thể kể đến như sản phẩm Vais[4], sản phẩm của tập đoàn FPT[5] hay hệ thống tổng hợp tiếng nói

[4] https://vais.vn/
[5] https://speech.openfpt.vn/

tiếng Việt Mica TTS[6] (Viện Mica Đại học Bách Khoa Hà Nội). Dữ liệu sử dụng cho hệ thống này bao gồm 3000 câu giàu ngữ âm và được gán nhãn bán tự động mức âm vị. Báo cáo kết quả của hệ thống này cho thấy độ hiểu đạt gần mức 100% và chất lượng tổng hợp đạt điểm 3.23 trên 5 thông qua một đánh giá sơ bộ.

Trong những năm gần đây, cùng với sự phát triển của công nghệ học sâu, nhiều hệ thống tổng hợp mới được ra đời. Ở Việt Nam, nhưng mô hình đó cũng được nhanh chóng áp dụng cho tổng hợp tiếng nói tiếng Việt. Các hệ thống tiêu biểu có thể kể đến như hệ thống tổng hợp tiếng nói Viettel AI[7] của tập đoàn Viettel, hệ thống tổng hợp tiếng nói của Zalo[8] hay giọng nói tổng hợp của Google. Nhưng hệ thống này đã đạt được độ tư nhiên cao cũng như được ứng dụng rộng rãi trong các ứng dụng như báo nói (Dân Trí[9], Báo Mới[10], …), trợ lý ảo (Google Assistant[11]) hay tổng đài trả lời tự động (Viettel Cyber Callbot[12]).

1.4 Giới thiệu về thích ứng giọng nói

Cùng với sự phát triển của các kỹ thuật tổng hợp tiếng nói, yêu cầu về chất lượng của hệ thống tổng hợp tiếng nói cũng ngày càng nâng cao. Bên cạnh độ tự nhiên, hệ thống tổng hợp tiếng nói cũng được kỳ vọng sẽ có khả năng tạo ra giọng nói của người nói tùy ý với dữ liệu đào tạo tối thiểu. Để đáp ứng vấn đề đó, thích ứng giọng nói và chuyển đổi giọng nói đã trở thành các hướng nghiên cứu chính trong lĩnh vực tổng hợp tiếng nói [14]. Thích ứng giọng nói là nhiệm vụ tạo ra giọng nói mới cho hệ thống tổng hợp tiếng nói bằng cách điều chỉnh các tham số của một mô hình ban đầu với một lượng ít dữ liệu ghi âm của người nói mới. Đây không phải là một chủ đề mới mà là một chủ đề đã được nghiên cứ kỹ lưỡng từ lâu, đặc biệt là trên hệ thống tổng hợp tiếng nói dựa trên mô hình Markov ẩn và các hệ thống nhận dạng tiếng nói.

Đối với các hệ thống tổng hợp dựa trên mô hình Markov ẩn, nhiều kỹ thuật thích ứng giọng nói đã được phát triển để cải thiện tính tự nhiên và mức độ tương tự của giọng nói tổng hợp. Các kỹ thuật có thể chia thành hai loại đó là Maximum likelihood linear regression (MLLR) [15] và Maximum a posteriori (MAP) [16]. Nhóm các kỹ thuật MLLR cố gắng học một phép biến đổi tuyến tính có thể biến đổi giọng nói trung bình thành âm thanh như giọng đích, trong khi các kỹ thuật MAP sử dụng các mô hình độc lập với người nói (speaker-independent) làm mô hình mẫu trước khi ước tính mô hình giọng đích. Những kỹ thuật thích ứng giọng nói này đã được chứng minh là có hiệu quả trong việc bắt chước giọng nói của một người bằng cách sử dụng một lượng nhỏ dữ liệu thích ứng Có nhiều yếu tố ảnh hưởng đến chất lượng của hệ thống thích ứng giọng nói theo phương pháp này như là trạng thái của mô hình ban đầu hay các tiêu chí ước tính.

[6] http://sontinh.mica.edu.vn/tts2/
[7] https://viettelgroup.ai/service/tts
[8] https://zalo.ai/demo/text-to-speech
[9] https://dantri.com.vn/
[10] https://baomoi.com/
[11] https://assistant.google.com/
[12] https://viettelgroup.ai/product/cyberbot

Đối với các hệ thống dựa trên mạng nơ ron học sâu, huấn luyện một mô hình có thể thích ứng giọng nói bằng cách sử dụng một véc tơ mã hóa người nói là một phương pháp được sử dụng phổ biến. Mô hình Deep Voice [17] thêm véc tơ mã hóa người nói vào nhiều phần của mạng để tạo ra mô hình nhiều người nói và có thể thích ứng giọng nói đối với người mới. Mô hình Voiceloop [18] đào tạo một mã hóa người nói cùng với mô hình âm học để có thể thích ứng giọng nói với người nói mới bằng việc sử dụng cả mẫu ghi âm cũng như nhãn của nó.

1.5 Vấn đề đặt ra với đồ án

Như đã đề cập trong mục 1.4, thích ứng giọng nói là một chủ đề nhận được nhiều sự quan tâm trong lĩnh vực tổng hợp tiếng nói. Đối với tổng hợp tiếng nói tiếng Việt, nhiều hệ thống cũng đã được nghiên cứu ví dụ như các hệ thống thích ứng giọng nói dựa trên mô hình HMM [19]. Tuy nhiên, đối với các mô hình sử dụng mạng nơ ron học sâu cho tổng hợp tiếng nói tiếng Việt, do mới được phát triển gần đây nên chưa có nhiều nghiên cứu về chủ đề thích ứng giọng nói.

Tại các công ty có sản phẩm về tổng hợp tiếng nói, nhu cầu của khách hàng về sự đa dạng của tiếng nói tổng hợp cũng ngày càng cao. Để phát triển một giọng nói tổng hợp mới, đáp ứng yêu cầu về chất lượng, cần tốn hàng chục giờ dữ liệu sạch, tương đương với hàng trăm giờ thu âm trong phòng thu với các thiết bị chuyên dụng. Vì thế chi phí cho việc phát triển hệ thống là rất tốn kém về chi phí cũng như thời gian.

Bởi những lý do trên, đồ án này tập trung vào tìm kiếm, thử nghiệm những kỹ thuật thích ứng giọng nói cho hệ thống tổng hợp tiếng nói tiếng Việt dựa trên mạng nơ ron học sâu.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Tổng quan về học sâu

Học sâu là một nhánh của lĩnh vực học máy, dựa trên một tập hợp các thuật toán để cố gắng mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng nhiều lớp xử lý với cấu trúc phức tạp, hoặc bằng cách khác bao gồm nhiều biến đổi phi tuyến. Chương này sẽ chủ yếu trình bày kiến thức cơ bản về kỹ thuật học sâu và ứng dụng của nói trong bài toán tổng hợp tiếng nói cũng như chuyển đổi giọng nói.

2.1.1 Mạng nơ ron nhân tạo

Mạng nơ ron nhân tạo (ANN) là một mô hình toán học hay mô hình tính toán được xây dựng mô phỏng theo các mạng nơ ron sinh học. ANN bao gồm các đơn vị (hay nút) được kết nối gọi là nơ ron nhân tạo. Mỗi kết nối, giống như các khớp thần kinh trong bộ não, có thể truyền tín hiệu đến các tế bào thần kinh khác. Mỗi nơ ron nhân tạo nhận tín hiệu sau đó xử lý và nó có thể báo hiệu các nơ ron được kết nối với nó. Trong ANN, tín hiệu tại một kết nối là một số thực và đầu ra của mỗi nơ ron được tính bằng một số hàm phi tuyến tính của các tổng đầu vào (input) của nó. Những kết nối được gọi là cạnh (edge). Các nơ ron và cạnh thường có trọng số (weight) được điều chỉnh trong quá trình học. Trọng số làm tăng hoặc giảm cường độ tín hiệu tại mỗi kết nối. Các nơ ron có thể có một ngưỡng sao cho tín hiệu chỉ được gửi nếu tín hiệu tổng hợp vượt qua ngưỡng đó. Thông thường các nơ ron được tổng hợp thành các lớp (layer). Các lớp khác nhau có thể thực hiện các biến đổi khác nhau trên đầu vào của chúng. Tín hiệu truyền từ lớp đầu tiên (input layer) đến lớp cuối cùng (output layer) sau khi đã đi qua các lớp nhiều lần.

Mục tiêu ban đầu của ANN là giải quyết các vấn đề tương tự như bộ não của con người. Nhưng theo thời gian, sự chú ý chuyển sang thực hiện các nhiệm vụ cụ thể, dẫn đến sự sai lệch so với bộ não sinh học. ANN đã được sử dụng cho nhiều nhiệm vụ khác nhau bao gồm thị giác máy tính, xử lý ngôn ngữ tự nhiên, nhận dạng tiếng nói, tổng hợp tiếng nói, chuẩn đoán y tế, ...

2.1.2 Logistic regression

Logistic regression là mô hình mạng nơ ron nhân tạo đơn giản nhất chỉ với input layer và output layer.

Mô hình của logistic regression là:

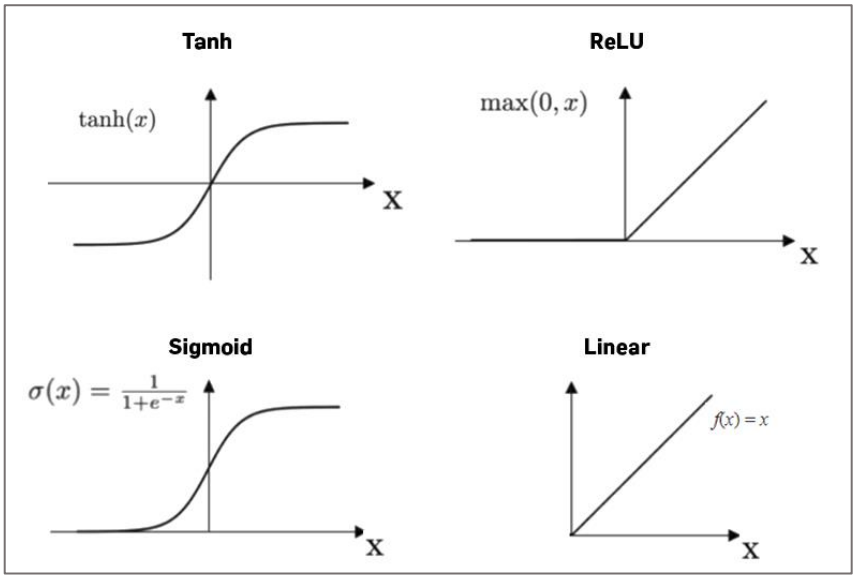
$$\hat{y} = \theta(w^T x + b) \qquad PT\ 2.1$$

Trong đó w là hệ số cần tối ưu và x là dữ liệu đầu vào, b là bias. θ là hàm kích hoạt (activation functoion). Có nhiều hàm kích hoạt thường được sử dụng như là hàm sigmoid (PT 2.1), hàm tanh (phương trình PT 2.3), hàm ReLU (phương trình PT 2.4). Hình 2.1 mô tả hình dạng của các hàm đó.

$$\sigma(s) = \frac{1}{1 + e^{-s}} \qquad \textit{PT 2.2}$$

$$tanh(s) = \frac{e^{s} - e^{-s}}{e^{s} + e^{-s}} \qquad \textit{PT 2.3}$$

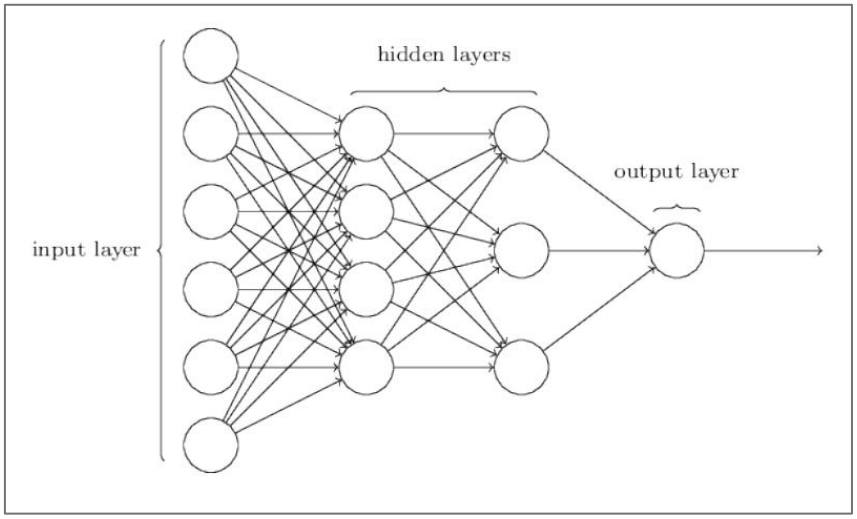
$$ReLU(s) = \max(0, s) \qquad \textit{PT 2.4}$$



Hình 2.1 Một số hàm kích hoạt thường gặp

2.1.3 Mạng nơ ron học sâu

Mạng nơ ron học sâu (DNN) là một mạng nơ ron nhân tạo (ANN) với nhiều đơn vị lớp ẩn giữa lớp đầu vào và đầu ra. Mạng nơ ron học học sâu mô hình hóa một hàm toán học để biến đổi đầu vào thành đầu ra cho dù mối quan hệ giữa chúng là tuyến tính hay phi tuyến tính.



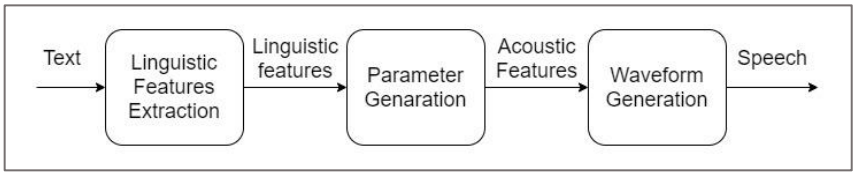
Hình 2.2 Mạng nơ ron với hai lớp ẩn [37]

Hình 2.2 mô tả một mạng nơ ron đơn giản với bốn lớp, trong đó có hai lớp ẩn. Lớp ngoài cùng bên trái gọi là lớp đầu vào (input layer), lớp ngoài cùng bên phải là lớp đầu ra (output layer). Hai lớp ở giữa hai lớp này được gọi là lớp ẩn (hidden layer). Việc thiết lập đầu vào và đầu ra của một mạng nơ ron thường đơn giản và phụ thuộc vào bài toán. Ví dụ như trong bài toán phân loại ảnh, kích thước của lớp đầu vào thường bằng $w \times h \times d$ với w và h và chiều rộng và chiều dài của ảnh, d là chiều sâu của ảnh (ảnh grayscale có $d = 1$, ảnh RGB có $d = 3$), kích thước của lớp đầu ra thường bằng số lớp cần phân loại. Trong khi đó việc thiết kế các lớp ẩn rất phức tạp và cực kỳ quan trọng để tạo được những đầu ra theo hướng mong muốn.

Ưu điểm của mạng nơ ron học sâu đó là khả năng mô hình hóa những quan hệ phi tuyến phức tạp giữa đầu vào và đầu ra. Tuy nhiên nhược điểm của mạng nơ ron học sâu đó là cần nhiều dữ liệu và thời gian huấn luyện để có thể đạt được kết quả mong muốn.

2.2 Tổng hợp tiếng nói dựa trên công nghệ học sâu

Mô hình âm học dựa trên mô hình Markov ẩn (HMM) và mô hình GMM là hai loại phổ biến nhất được sử dụng trong quá trình tạo tín hiệu tiếng nói từ chuỗi kỹ tự đầu vào (thường là chuỗi âm vị) thông qua việc tạo trực tiếp các đặc trưng âm học của tiếng nói. Tuy nhiên những mô hình kiểu này có những giới hạn trong việc biểu diễn mối quan hệ phức tạp và phi tuyến giữa chuỗi ký tự đầy vào và các đặc trưng âm học. Với sự phát triển của công nghệ học sâu, các mạng nơ ron học sâu ngày càng được sử dụng rộng rãi và cho thấy ưu điểm so với các phương pháp thông thường (như HMM hoặc GMM). Hình 2.3 mô tả một kiến trúc cơ bản của hệ thống tổng hợp tiếng nói dựa trên phương pháp học sâu.



Hình 2.3 Kiến trúc cơ bản của hệ thống thống hợp tiếng nói dựa trên phương pháp học sâu

Có thể thấy rằng hệ thống gồm ba mô đun chính, trong đó:

- Mô đun trích chọn đặc trưng ngôn ngữ: văn bản đầu vào được xử lý, phân tích và trích chọn bởi bộ Linguistic Features Extraction ra thành các vec tơ đặc trưng ngôn ngữ, các vec tơ này thường bao gồm các thông tin về chuỗi âm vị, vị trí tương đối của âm vị trong câu, cụm từ hay từ, số lượng âm vị trong câu, trong cụm từ hay trong từ,…
- Mô đun Parameter Generation có nhiệm vụ chuyển hóa các đặc trưng ngôn ngữ ở đầu vào thành thành các đặc trưng âm học tương ứng, với hệ thống tổng hợp tiếng nói được xây dựng dựa trên phương pháp học sâu, thì mô đun này sử dụng mạng nơ ron học sâu DNN để dự đoán các đặc trưng âm học từ đặc trưng ngôn ngữ đầu vào.

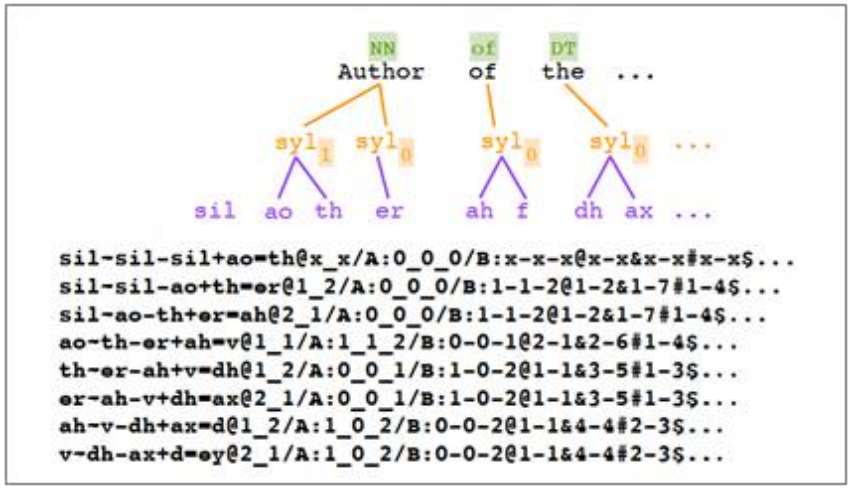
- Mô đun tạo tín hiệu tiếng nói: Các đặc trưng âm học được mô đun này chuyển hóa thành tín hiệu tiếng nói

Chi tiết từng mô đun sẽ được trình trong các mục 2.2.1, 2.2.2 và 2.2.3

2.2.1 Trích chọn đặc trưng ngôn ngữ

Các đặc trưng ngôn ngữ được sử dụng để làm đầu vào cho mô hình âm học bao gồm các thông tin như: âm vị hiện tại, vị trí của âm vị trong câu, cụm từ, vị trí từ trong câu, số lượng âm vị trong từ hay thanh điệu hiện tại là gì, … Các thông tin này cũng được phân theo các mức như: mức âm vị, mức âm tiết, mức từ, mức cụm từ, mức câu [20]. Để lấy được các đặc trưng ngôn ngữ trên, văn bản đầu vào sẽ được xử lý theo các bước như sau:

- Văn bản đầu vào sẽ được chuyển thành một chuỗi âm vị nhờ từ điển phiên âm tương ứng với ngôn ngữ đang tổng hợp
- Văn bản đầu vào sẽ được cho qua một hệ thống xử lý ngôn ngữ tự nhiên để trích chọn các thông tin về ngôn ngữ, hệ thống xử lý ngôn ngữ tự nhiên này được xây dựng trên cơ sở ba mô hình: mô hình tách từ (word segmentation) để tách văn bản thành chuỗi các từ, mô hình gán nhãn từ loại (part of speech tag) để gán nhãn các từ thành từ loại tương ứng (danh từ, động từ, đại từ, giới từ, trạng từ, …) và mô hình phân tách cụm từ (text chunking) để tách văn bản thành các cụm từ và kèm theo thông tin về vị trí của các từ trong cụm.
- Từ chuỗi âm vị được chuyển hóa và các kết quả của việc tách từ, gán nhãn từ loại, tách cụm từ ta tiền hành tính toán các thông tin đặc trưng ngôn ngữ của văn bản.

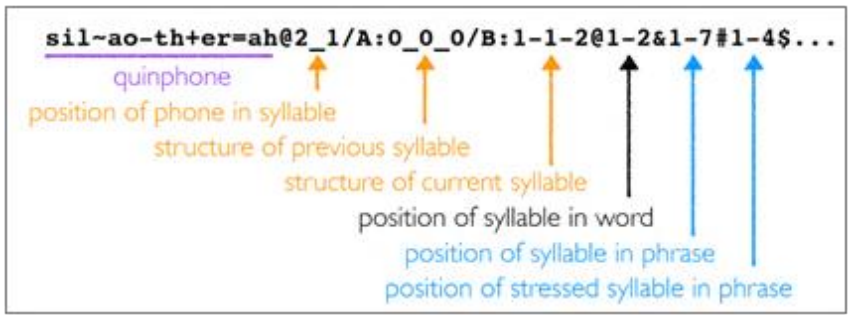


Hình 2.4 Biểu diễn đặc trưng ngôn ngữ học của văn bản [38]

Các đặc trưng ngôn ngữ trích chọn được từ quá trình trên bao gồm các thông tin như:

- Thông tin mức âm vị: bao gồm thông tin âm vị hiện tại, âm vị phía trước, âm vị phía sau, thông tin về vị trí âm vị trong âm tiết, từ, cụm từ, …
- Thông tin mức âm tiết: bao gồm thông tin về số lượng âm vị của âm tiết hiện tại, âm tiết phía trước, âm tiết phía sau, thông tin về thanh điệu và vị trí của âm tiết trong từ, cụm từ, …

- Thông tin mức từ: bao gồm các thông tin về nhãn từ loại, số lượng âm tiết của từ hiện tại, từ phía trước, từ phía sau.
- Thông tin mức cụm từ: bao gồm số lượng từ, âm tiết trong cụm từ hiện tại, cụm từ phía trước, cụm từ phía sau.
- Thông tin mức câu: bao gồm thông tin vế số lượng âm tiết, số lượng từ, số lượng cụm từ trong câu.



Hình 2.5 Thông tin đặc trưng ngôn ngữ ở mức âm vị [38]

Ngoài các đặc trưng ngôn ngữ, các mô hình tiếp theo (mô hình âm học và mô hình thời gian) vẫn cần thêm thêm thông để có thể huấn luyện. Một thông tin cần thiết phải thêm vào đó là thời gian xuất hiện của mỗi âm vị trong câu nói. Để lấy được thông tin về thời gian này, ta sử dụng mô hình Markov ẩn (HMM), quá trình này được gọi là force alignment. Kết quả của quá trình force alignment sẽ là khoảng thời gian xuất hiện của mỗi trạng thái trong mỗi âm vị. Hình 2.6 minh họa thời gian cho từng trạng thái của một âm vị với hai chỉ số đầu là thời gian bắt đầu và kết thúc trạng thái đó. Thông thường ta sử dụng 5 trạng thái cho mỗi âm vị.

```
    3050000 3100000 xx~#-p+l=i:1_4/A/0_0_0/B/1-1-4:1-1&1-4#1-3$1-4>0-1<0-
1|i/C/1+1+3/D/0_0/E/content+1:1+3&1+2#0+1/F/content_1/G/0_0/H/4=3:1=1&L-
L%/I/0_0/J/4+3-1[2]

    3100000 3150000 xx~#-p+l=i:1_4/A/0_0_0/B/1-1-4:1-1&1-4#1-3$1-4>0-1<0-
1|i/C/1+1+3/D/0_0/E/content+1:1+3&1+2#0+1/F/content_1/G/0_0/H/4=3:1=1&L-
L%/I/0_0/J/4+3-1[3]

    3150000 3250000 xx~#-p+l=i:1_4/A/0_0_0/B/1-1-4:1-1&1-4#1-3$1-4>0-1<0-
1|i/C/1+1+3/D/0_0/E/content+1:1+3&1+2#0+1/F/content_1/G/0_0/H/4=3:1=1&L-
L%/I/0_0/J/4+3-1[4]

    3250000 3350000 xx~#-p+l=i:1_4/A/0_0_0/B/1-1-4:1-1&1-4#1-3$1-4>0-1<0-
1|i/C/1+1+3/D/0_0/E/content+1:1+3&1+2#0+1/F/content_1/G/0_0/H/4=3:1=1&L-
L%/I/0_0/J/4+3-1[5]

    3350000 3900000 xx~#-p+l=i:1_4/A/0_0_0/B/1-1-4:1-1&1-4#1-3$1-4>0-1<0-
1|i/C/1+1+3/D/0_0/E/content+1:1+3&1+2#0+1/F/content_1/G/0_0/H/4=3:1=1&L-
L%/I/0_0/J/4+3-1[6]

    305000 310000 are the starting and ending time.
    [2], [3], [4], [5], [6] mean the HMM state index.
```

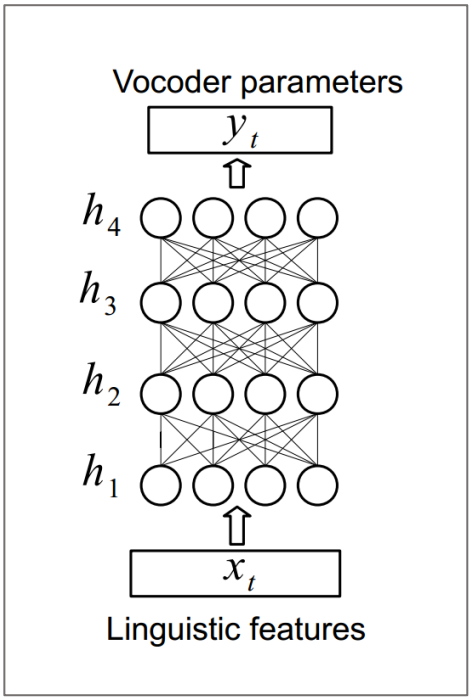
Hình 2.6 Thông tin đặc trưng ngôn ngữ ở mức trạng thái âm vị

2.2.2 Mô hình âm học dựa trên mạng nơ ron học sâu

Trong tổng hợp tiếng nói dựa trên phương pháp học sâu, mô hình âm học được mô hình hóa bằng một mạng nơ ron học sâu như Hình 2.7, trong đó đầu vào của mạng

là một vec tơ đặc trưng ngôn ngữ và đầu ra là các đặc trưng âm học hay chính là tham số của vocoder (trình bày tại mục 2.2.3).

Như đã nói ở trên, đầu vào của mạng nơ ron là một véc tơ đặc trưng ngôn ngữ, véc tơ này được mã hóa từ các đặc trưng ngôn ngữ mà ta trích chọn được trong mục 2.2.1. Có nhiều phương pháp khác nhau để chuyển hóa thông tin đặc trưng ngôn ngữ thành một véc tơ đầu vào cho mạng nơ ron học sâu, một trong số đó là sử dụng một tệp các câu hỏi. Các câu hỏi này được dùng để lấy các thông tin mà các đặc trưng ngôn ngữ đem lại. Bằng cách trả lời các câu hỏi này, ta thu được véc tơ nhị phân biể diễn các đặc trưng ngôn ngữ học. Đầu ra của mạng nơ ron là các véc tơ đặc trưng âm học, véc tơ này là đầu vào cho vocoder để tổng hợp tiếng nói. Các véc tơ đặc trưng âm học bao gồm các thông tin như: tần số cơ bản F0, đường bao phổ của tín hiệu tiếng nói, thông tin về các thành phần không tuần hoàn. Ở bước



Hình 2.7 Một mình họa về mạng nơ ron học sâu với bốn lớp ẩn trong tổng hợp tiếng nói [21]

huấn luyện mô hình âm học, các véc tơ đặc trưng âm học này được trích chọn từ các mẫu thu âm nhờ vào vocoder.

Véc tơ đầu vào sẽ được sử dụng để dự đoán kết quả đầu ra thông qua các lớp của các đơn vị ẩn, mỗi đơn vị thực hiện một hàm không tuyến tính như các phương trình sau:

$$h_t = \mathcal{H}(W^{xh}x_t + b^h) \qquad \textit{PT 2.5}$$

$$y_t = W^{hy}h_t + b^y \qquad \textit{PT 2.6}$$

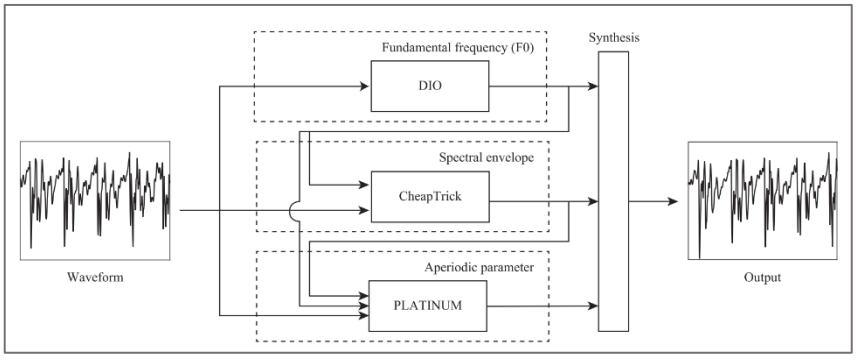
Trong đó $\mathcal{H}(\cdot)$ là hàm kích hoạt phi tuyến ở lớp ẩn (thường là hàm tanh), W^{xh} và W^{hy} là ma trận trọng số, b^h và b^y là hai véc tơ bổ sung (bias vector). $W^{hy}h_t$ là thành phần hồi quy tuyến tính để dự đoán đặc trưng đích từ hàm kích hoạt trong lớp ẩn trước [21].

2.2.3 Vocoder

Vocoder (viết tắt của voice encoder) là một hệ thống hệ thống phân tích và tổng hợp tín hiệu tiếng nói của con người. Trong tổng hợp tiếng nói dựa trên mạng nơ ron học sâu, vocoder được sử dụng trong hai quá trình: huấn luyện mô hình và tổng hợp tiếng nói. Trong quá trình huấn luyện mô hình, vocoder được sử dụng để phân tích dữ liệu âm thanh thành các đặc trưng âm học (chẳng hạn như phổ, tần số cơ bản, cepstra, ...), các đặc trưng này được sử dụng để huấn luyện mạng nơ ron học sâu. Trong quá trình tổng hợp, các đặc trưng âm học của tiếng nói được tạo ra bởi mạng nơ ron học sâu sẽ là đầu vào cho vocoder để tạo thành tín hiêu tiếng nói.

Qua quá trình phát triển, nhiều loại vocoder đã được phát minh nhằm cải thiện chất lượng phân tích và tổng hợp tiếng nói, tiêu biểu như STRAIGHT vocoder [22], WORLD vocoder [23], Magphase vocoder [24]. Trong phần này sẽ chỉ trình bày về WORLD vocoder, vocoder được sử dụng trong mô hình tổng hợp tiếng nói của đồ án này.

Như đã nói ở trên, WORLD vocoder được sử dụng để trích chọn các đặc trưng âm học và tổng hợp tiếng nói từ những đặc trưng này. Các đặc trưng âm học mà WORLD vocoder trích chọn bao gồm: đường bao phổ của tín hiệu, các thành phần không tuần hoàn (Aperiodicities) và tần số cơ bản F0. Trong đó tần số cơ bản F0 được ước lượng bằng phương pháp DIO [25], đường bao phổ được ước lượng bởi phương pháp CheapTrick [26] và tín hiệu kích thích được ước lượng bởi phương pháp PLATINUM [27] và được sử dụng như tham số không tuần hoàn. Hình 2.8 mô tả quá trình xử lý của WORLD vocoder trong hai giai đoạn phân tích và tổng hợp tín hiệu tiếng nói.



Hình 2.8 Tổng quan về hệ thống WORLD vocoder [23]

Tần số cơ bản, hay là tần số âm cơ bản, là tần số thấp nhất của dạng sóng tuần hoàn. Phương pháp DIO ước lượng tần số cơ bản F0 bằng ba bước:

- Sử dụng các bộ lọc thông thấp với các tần số cắt khác nhau để lọc tín hiệu, nếu tín hiệu được lọc nào có chứa thành phần tần số cơ bản thì nó

sẽ có dạng hình sin với chu kỳ T0. Bởi vì chưa biết F0, nên ta sử dụng nhiều bộ lọc với các tần số cắt khác nhau.

- Tìm các ứng viên cho tần số cơ bản F0 và độ tin cậy của nó trong mỗi tín hiệu được lọc.
- Chọn ra ứng viên nào có độ tin cậy cao nhất làm F0.

WORLD ước lượng đường bao phổ bằng phương pháp CheapTrick, dựa trên ý tưởng việc phân tích đồng bộ cao độ và sử dụng một cửa sổ hanning (hanning window) với độ dài 3T0. Các bước để ước lượng đường bao theo phổ theo phương pháp CheapTrick như sau: Năng lượng phổ được tính trên cơ sở mỗi khung tín hiệu được lấy bởi cửa số hanning nêu trên. Tổng năng lượng trong một khung tín hiệu được coi là tạm thời ổn định và được tính dựa theo công thực sau:

$$\int_0^{3T_0} \big(y(t)w(t)\big)^2 dt = 1.125 \int_0^{T_0} y^2(t)dt \qquad PT\ 2.7$$

Trong đó $y(t)$ là tín hiệu và $w(t)$ là hàm cửa sổ. Sau khi tính được năng lượng phổ nêu trên, chúng được làm mịn với một cửa sổ chữ nhật có độ dài $2\omega_0/3$, như sau:

$$P_s(\omega) = \frac{3}{2\omega_0} \int_{-\frac{\omega_0}{3}}^{\frac{\omega_0}{3}} P(\omega + \lambda)d\lambda \qquad PT\ 2.8$$

Với ω_0 là $2\pi/T_0$. Đường bao phổ $P_1(\omega)$ được tính như sau:

$$P_l(\omega) = \exp\big(\mathcal{F}\big[l_s(\tau)l_q(\tau)p_s(\tau)\big]\big) \qquad PT\ 2.9$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_o}\right) \qquad PT\ 2.10$$

$$l_s(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_o \tau} \qquad PT\ 2.11$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right) \qquad PT\ 2.12$$

$$p_s(\tau) = \mathcal{F}^{-1}\big[\log\big(P_s(\omega)\big)\big] \qquad PT\ 2.13$$

Trong đó, $l_s(\tau)$ là hàm nâng cho việc làm mịn logarit năng lượng phổ, $l_q(\tau)$ là hàm nâng cho việc hồi phục phổ và $\tilde{q}_0,\ \tilde{q}_1$ là các tham số cho việc phục hồi phổ. Các ký hiệu $\mathcal{F}[]$ và $\mathcal{F}^{-1}[]$ đại diện cho biến đổi Fourier và biến đổi Fourier ngược. Cuối cùng, phương pháp PLATINUM ước lượng tín hiệu kích thích. Đầu tiên, tín hiệu đi qua cửa sổ có độ dài $2T_0$, phổ của tín hiệu sau khi đưa qua cửa sổ được chia ra bởi phổ tối thiểu $S_m(\omega)$. $S_m(\omega)$ được tính theo biểu thức sau:

$$S_m(\omega) = \exp(\mathcal{F}[c_m(\tau)]) \qquad PT\ 2.14$$

$$c_m(\tau) = \begin{cases} 2c(\tau) & (\tau > 0) \\ c(\tau) & (\tau = 0) \\ 0 & (\tau < 0) \end{cases} \qquad PT\ 2.15$$

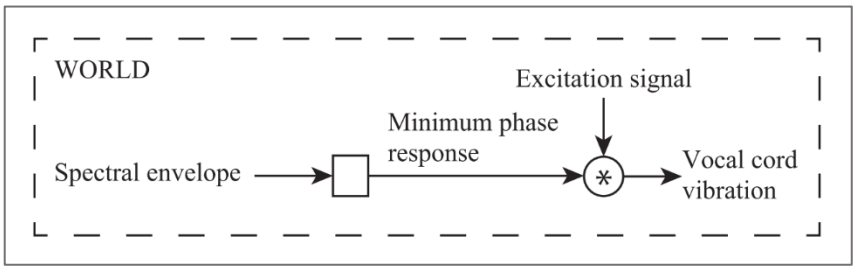
$$c(\tau) = \mathcal{F}^{-1}\left[\log\left(P_1(\omega)\right)\right] \qquad PT\ 2.16$$

Tín hiệu kích thích được biểu diễn như sau:

$$x_p(t) = \mathcal{F}^{-1}\left[X_p(\omega)\right] \qquad PT\ 2.17$$

$$X_p(\omega) = \frac{X(\omega)}{S_m(\omega)} \qquad PT\ 2.18$$

Sau khi đã có được thông tin đặc trưng cần thiết, âm thanh tổng hợp được tính bằng cách nhân chập tín hiệu kích thích và đáp ứng pha tối thiểu, điều này được minh họa trong Hình 2.9.



Hình 2.9 Tổng hợp tiếng nói với WORLD vocoder [23]

2.3 Thích ứng giọng nói dựa trên công nghệ học sâu

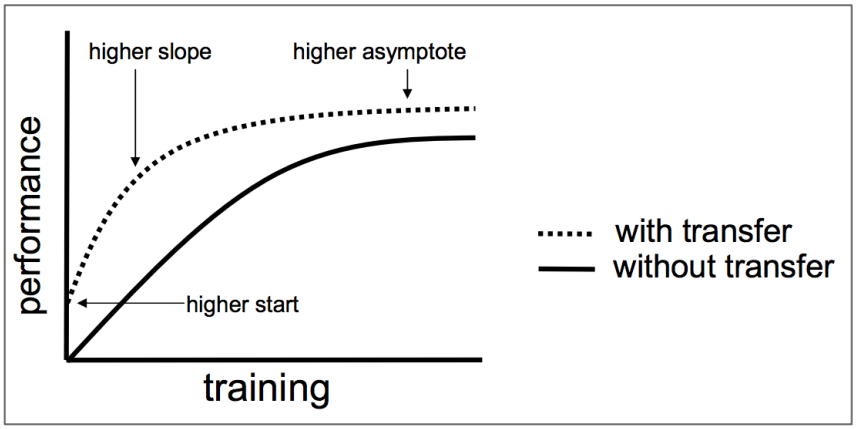
Với các mô hình sử dụng tham số thông kê (ví dụ như HMM hay GMM), MLLR và MAP hai phương pháp thành công nhất để thích ứng từ mô hình tổng quát sang một giọng nói cụ thể. Đối với các mô hình sử dụng mạng nơ ron học sâu, đặc biệt

với các mạng có rất nhiều lớp ẩn, việc thích ứng giọng nói là một nhiệm vụ không đơn giản. Lý do là bởi số lượng cũng như sự phân bố phức tạp của các trọng số trong mô hình DNN, nên không thể sử dụng biến đổi đơn giản như MLLR để cập nhật các trọng số của mô hình. Tuy nghiên, các nghiên cứu gần đây đã chỉ ra việc thích ứng giọng nói cho các mô hình DNN là khả thi. Nhiều phương pháp đã được đề xuất và mang lại hiệu quả nhất định cho thích ứng giọng nói, ví dụ như: Transfer learning, sử dụng véc tơ mã hóa người nói [28] [29], Learning hidden unit contribution (LHUC) [30], Feature space transformation (FST) [31]. Trong phần này sẽ trình bày hai phương pháp đó là transfer learning và sử dụng véc tơ mã hóa người nói cũng là hai phương pháp sẽ được áp dụng trong chương tiếp theo.

2.3.1 Transfer learning

Transfer learning là một kỹ thuật trong học máy khi mà một mô hình đã được huấn luyện cho một nhiệm vụ nhất định được sử dụng lại để làm điểm khởi đầu cho mô hình với một nhiệm vụ mới. Transfer learning cho phép rút ngắn quá trình huấn luyện cũng như nâng cao hiệu năng cho quá trình huấn luyện mô hình mới. Ví dụ, trong nhận dạng giọng nói, một mô hình âm học được đào tạo cho một ngôn ngữ có thể được sử dụng để nhận dạng giọng nói bằng ngôn ngữ khác, với rất ít hoặc không có dữ liệu đào tạo lại. Kỹ thuật transfer learning có thể mang lại các lợi ích như:

- Higher start: Hiệu năng trước khi tinh chỉnh mô hình có thể cao hơn so với việc khởi tạo ngẫu nhiên các tham số cho mô hình ban đầu.
- Higher slope: Tốc độ hội tụ của mô hình mới có thể nhanh hơn.
- Higher asymptote: Hiệu năng của mô hình khi hội tụ có thể tốt hơn.



Hình 2.10 Các lợi ích của Transfer learning đối với việc huấn luyện mô hình

Đối với bài toán thích ứng giọng nói, transfer learning cũng đã đạt được một số thành công nhất định. Có hai hướng tiếp cận thường được sử dụng cho quá trình thích ứng giọng nói bằng transfer learning:

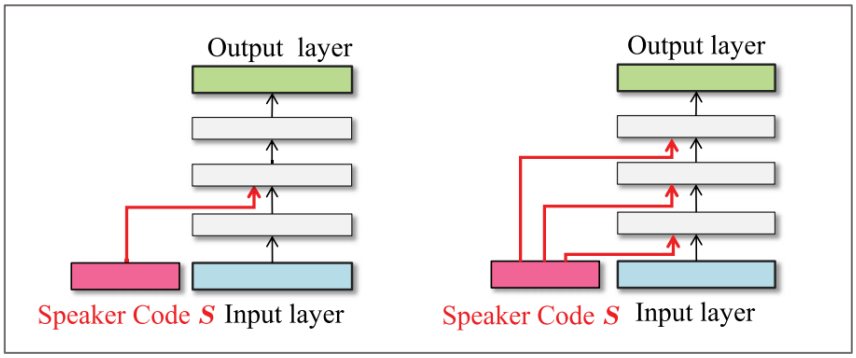
- Hướng tiếp cận đầu tiên và cũng là đơn giản nhất đó là trong quá trình huấn luyện mô hình thích ứng trên tập dữ liệu mới, toàn bộ trọng số của mô hình cũ sẽ được cập nhật. Hướng tiếp cận này kỳ vọng với việc sử

dụng mô hình cũ làm điểm khởi tạo cho mô hình thích ứng thì có thể tăng tốc độ hội tụ cũng như chất lượng của mô hình thích ứng.

- Hướng tiếp cận thứ hai đó là trong quá trình huấn luyện mô hình mới, chỉ một phần các trọng số của mô hình cũ được cập nhật. Sự thích ứng có thể được chỉ thực hiện trên lớp đầu vào, ở các lớp ẩn hoặc chỉ ở lớp đầu ra.

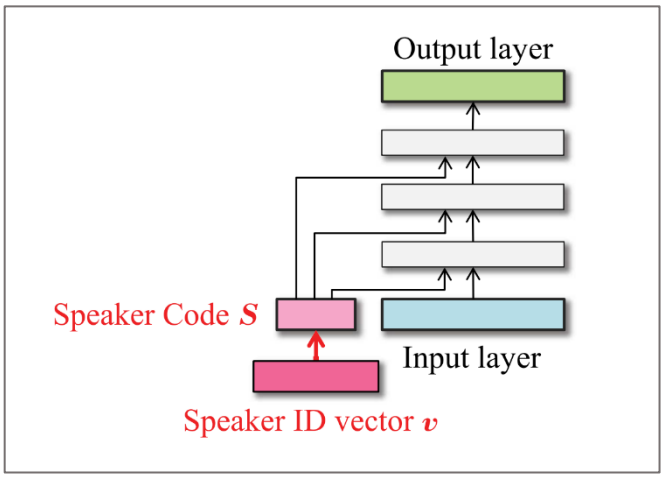
2.3.2 Sử dụng véc tơ mã hóa người nói

Phương pháp sử dụng véc tơ mã hóa người nói từ lâu đã được sử dụng trong bài toán nhận dạng tiếng nói [32] và đã cho hiệu quả đáng kể trong việc xác minh người nói độc lập với văn bản. Véc tơ mã hóa người nói là một véc tơ có số chiều thấp, được gắn với các đặc trưng đầu vào để định danh người nói. Từ ý tưởng đó mà phương pháp này cũng được áp dụng cho bài toán tổng hợp tiếng nói.



Hình 2.11 Sử dụng vec tơ mã hóa người nói kết nối với một hoặc nhiều lớp ẩn [28]

Có nhiều phương pháp đã được dùng để mã hóa thông tin của một người nói. Phương pháp đơn giản nhất đó là sử dụng véc tơ one-hot để mã hóa người nói. Véc tơ one-hot sẽ có dạng $X = [x_1, x_2, \dots, x_n]$ với n là số lượng người nói trong tập dữ liệu. Với mỗi câu trong tập dữ liệu, x_i sẽ bằng 1 nếu người nói là người thứ i và $x_j = 0 \; \forall \; j \neq i$. Véc tơ này sẽ được gắn vào các đặc trưng đầu vào của mô hình hoặc được kết nối trực tiếp vào một hoặc nhiều lớp ẩn như mô tả ở Hình 2.11.

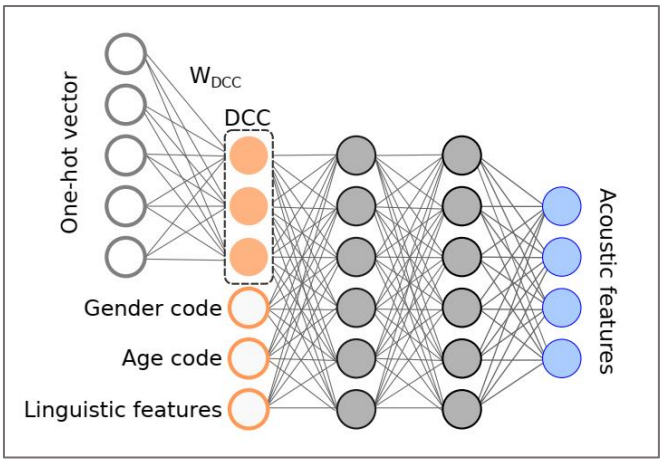


Hình 2.12 Sử dụng véc tớ nhúng [28]

Một phương pháp khác đó là sử dụng véc tơ nhúng, từ mô hình DNN ban đầu sẽ được thêm một tập các nốt để mã hóa thông tin người nói như mô tả ở Hình 2.12. Đầu vào cho các nốt này là véc tơ one-hot mã hóa người nói, đầu ra sẽ được kết nối tới các lớp ẩn của mô hình DNN. Trong quá trình huấn luyện mô hình, các trọng số của các nốt thêm vào được cập nhật cùng lúc với các trọng số của mô hình DNN ban đầu. Trong quá trình thích ứng giọng nói, các trọng số toàn bộ mô hình sẽ được giữ nguyên, véc tơ nhúng S sẽ được tính toán nhờ thuật toán back propagation.

Ngoài ra, một số phương pháp khác sử dụng mạng DNN để trích xuất véc tơ mã hóa người nói. Trong quá trình huấn luyện mạng DNN này sẽ dùng để phân loại người nói (speaker classification). Sau khi thu được mô hình đã qua huấn luyện, phương pháp này sẽ cắt bỏ lớp phân loại ở cuối mô hình đi và sử dụng véc tơ đặc trưng thu được ở cuối mô hình mới này để làm véc tơ mã hóa người nói.

Ngoài việc mã hóa người nói, một số thông tin như tuổi, giới tính cũng được thêm vào véc tơ đầu vào để tăng hiệu năng cho mô hình thích ứng giọng nói [29].

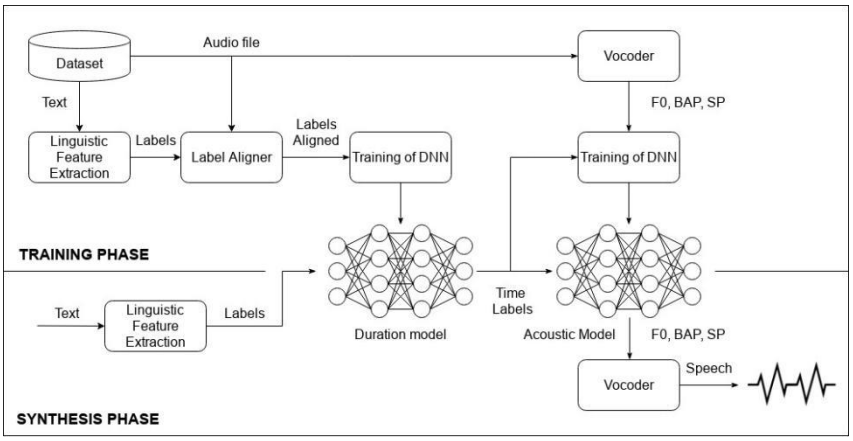


Hình 2.13 Các thông tin về tuổi và giới tính được thêm vào cùng với véc tơ mã hóa người nói [29]

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT CHUYỂN ĐỔI GIỌNG NÓI TIẾNG VIỆT

3.1 Mô hình cho quá trình tổng hợp tiếng nói

3.1.1 Tổng quan mô hình tổng hợp tiếng nói



Hình 3.1 Tổng quan mô hình tổng hợp tiếng nói

Mô hình tổng hợp tiếng nói cho đồ án này bao gồm 3 thành phần chính:

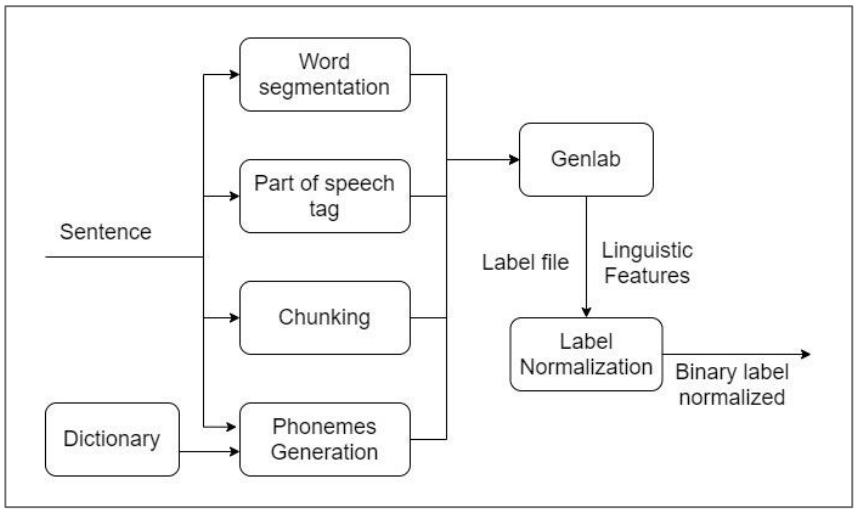
- Mô đun trích chọn đặc trưng ngôn ngữ
- Mô hình thời gian và mô hình âm học
- Mô đun trích chọn đặc trưng và tổng hợp tiếng nói (vocoder)

Chi tiết về từng thành phần được mô tả trong các mục 3.1.1, 3.1.3 và 3.1.4

3.1.2 Trích chọn đặc trưng ngôn ngữ

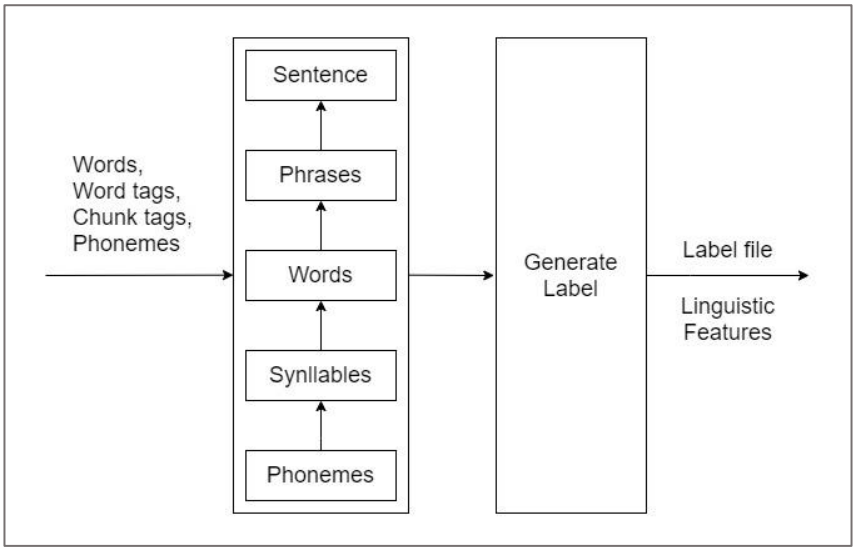
Để tạo dữ liệu đầu vào cho mô hình thời gian và mô hình âm học, đồ án sử dụng công cụ Vita [33] để trích chọn đặc trưng ngôn ngữ. Một từ điển phiên âm âm tiết tiếng Việt với khoảng 6700 từ được sử dụng cho quá trình tạo chuỗi âm vị.

Việc trích chọn đặc trưng ngôn ngữ này được xây dựng dựa trên ba mô hình: mô hình tách từ, mô hình gán nhãn từ loại và mô hình tách cụm từ. Quá trình hoạt động của hệ thống được biểu diễn như Hình 3.2, trong đó văn bản đầu vào được đưa qua bộ tách từ (Word segmentation), bộ gán nhãn từ loại (Part of speech tag) để gán nhãn, tách cụm từ bởi bộ tách cụm từ (Chunking) và qua bộ tạo chuỗi âm vị (Phonemes Genaration). Kết quả đầu ra các bộ này sẽ được đưa vào bộ Genlab để tạo label file, label file là tệp chứa các đặc trưng ngôn ngữ học của câu văn.



Hình 3.2 Hoạt động của bộ trích chọn đặc trưng ngôn ngữ

Bộ Genlab là bộ tạo đặc trưng ngôn ngữ học, cấu trúc bộ Genlab được thể hiện trên Hình 3.3 trong đó các chuỗi từ, chuỗi từ đã gán nhãn, chuỗi cụm từ được gán nhãn, chuỗi âm vị sẽ được đưa vào một cấu trúc dữ liệu đặc biệt bao gồm một đối tượng đại diện cho câu (Sentence) lưu trữ các cụm từ (Phrases), các cụm từ lưu trữ các từ (Words), các từ lưu trữ các âm tiết (Syllables), các âm tiết lưu trữ các âm vị (Phonemes). Sau đó từ cấu trúc dữ liệu này, hay nói các khác là từ đối tượng câu trở thành đầu vào cho bộ Generate Labels, nơi mà dùng để trích chọn các thông tin về đặc trưng ngôn ngữ học như đã nêu trong mục 2.2.1 sẽ được tính toán, ước lượng và lưu trong tệp chứa các nhãn (Label file). Cấu trúc từng dòng trong label file được nếu trong phụ lục A.



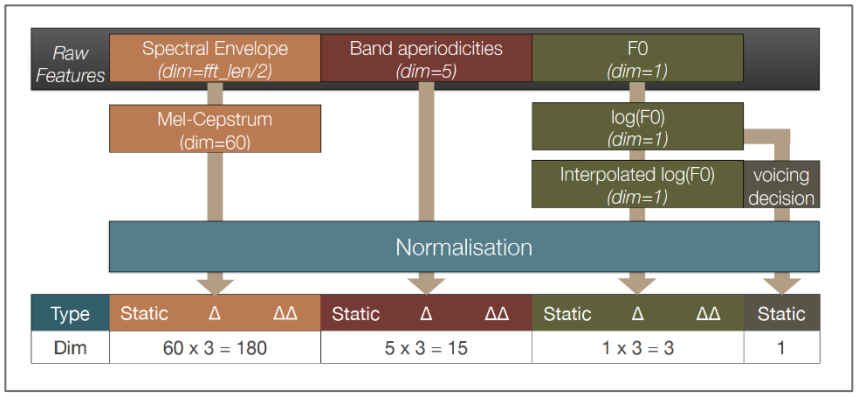
Hình 3.3 Cấu trúc và hoạt động của bộ Genlab

Bộ chuẩn hóa đặc trưng đầu vào (Label Normalization) có nhiệm vụ nhận các véc tơ đặc trưng ngôn ngữ và trả về véc tơ đặc trưng ngôn ngữ dưới dạng nhị phân đã chuẩn hóa. Phương pháp để chuyển từ véc tơ đặc trung ngôn ngữ sang dạng nhị phân là sử dụng tập câu hỏi đã được trình bày trong mục 2.2.1. Sau đó các véc tơ

này được chuẩn hóa cực tiểu cực đại (min-max normalization) về khoảng [0.01, 0.99].

3.1.3 Trích chọn đặc trưng âm học

Để trích chọn đặc trưng âm học, đồ án sử dụng WORLD vocoder. Các đặc trưng được WORLD vocoder trích xuất theo từng khung có độ dài 5ms, bao gồm Spectral Envelope, Band aperiodicties, F0 (như mô tả ở mục 2.2.3). Sau đó các đặc trưng này được biến đổi về các véc tơ đặc trưng bao gồm 60 chiều hệ số mel (MCCs), 5 chiều hệ số BAP và tần số cơ bản F0 trên thang đo log (log F0). Ngoài ra deltas và delta-deltas của các véc tơ trên cũng được tính. Tất cả véc tơ trên sau đó đi qua bộ chuẩn hóa để đưa các giá trị về phân phối có kỳ vọng là 0 và phương sai là 1. Quá trình này được mô tả tại Hình 3.4



Hình 3.4 Chuyển đổi từ đặc trưng gốc của vocoder sang đặc trưng âm học [38]

3.1.4 Mô hình dự đoán

Mô hình dự đoán có nhiệm vụ lấy đầu vào là các đặc trưng ngôn ngữ được trích chọn ở phần 3.1.2 và dự đoán đầu ra là các đặc trưng âm học. Cấu trúc của mô hình dự đoán bao gồm 2 thành phần đó là mô hình thời gian và mô hình âm học. Công cụ để xây dựng 2 mô hình này là Merlin [21].

a) Mô hình thời gian (Duration model)

Mô hình khoảng thời gian có nhiệm vụ nhận các đặc trưng ngôn ngữ từ mô đun trích chọn đặc trưng ngôn ngữ và dự đoán thông tin về thời gian xuất hiện của các trạng thái của âm vị. Số trạng thái của một âm vị là 5.

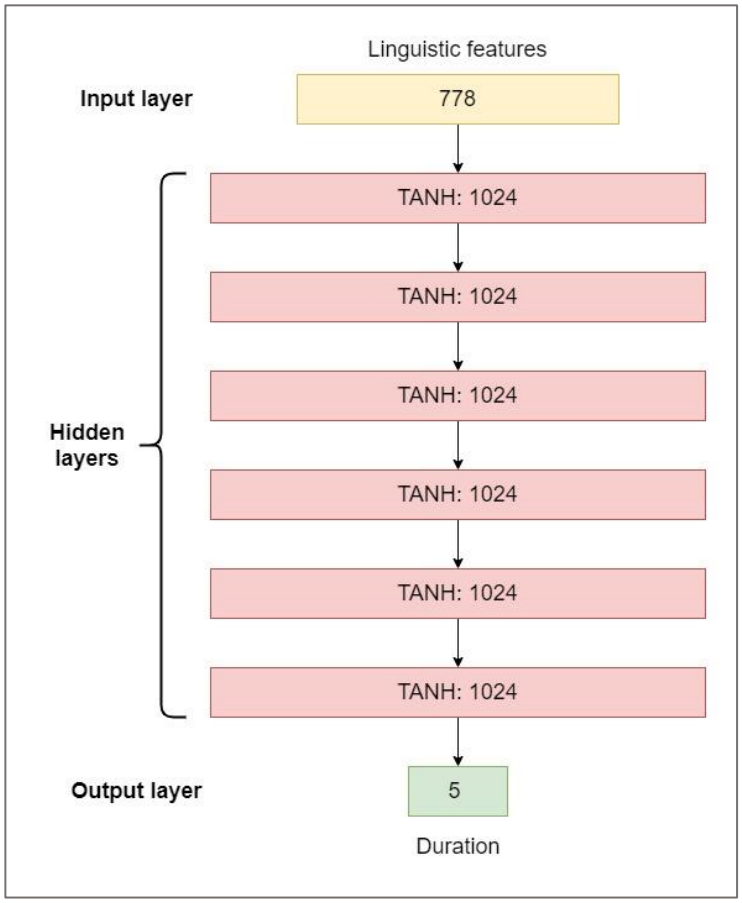
Mô hình này là một mạng nơ ron học sâu với các đặc điểm:

- Đầu vào của mạng là các véc tơ 778 chiều chứa đặc trưng ngôn ngữ của từng âm vị. Với mỗi câu trong tập dữ liệu, số lượng véc tơ đầu vào sẽ là số âm vị có trong câu đó.
- Có 6 lớp ẩn, mỗi lớp có 1024 nốt và hàm kích hoạt là hàm tanh (phương trình PT 2.3).
- Đầu ra của mạng là vec tơ 5 chiều chưa thông tin ước lượng khoảng thời gian xuất hiện của từng trạng thái trong âm vị. Tương tự như số lượng

véc tơ đầu vào, số véc tơ đầu ra của mô hình cho từng câu bằng số âm vị có trong câu đó.

- Hàm mất mát của mô hình là hàm MSE.

Hình 3.5 mô tả chi tiết cấu trúc của mô hình thời gian.



Hình 3.5 Cấu trúc mô hình thời gian (Duration Model)

b) Mô hình âm học (Acoustic model)

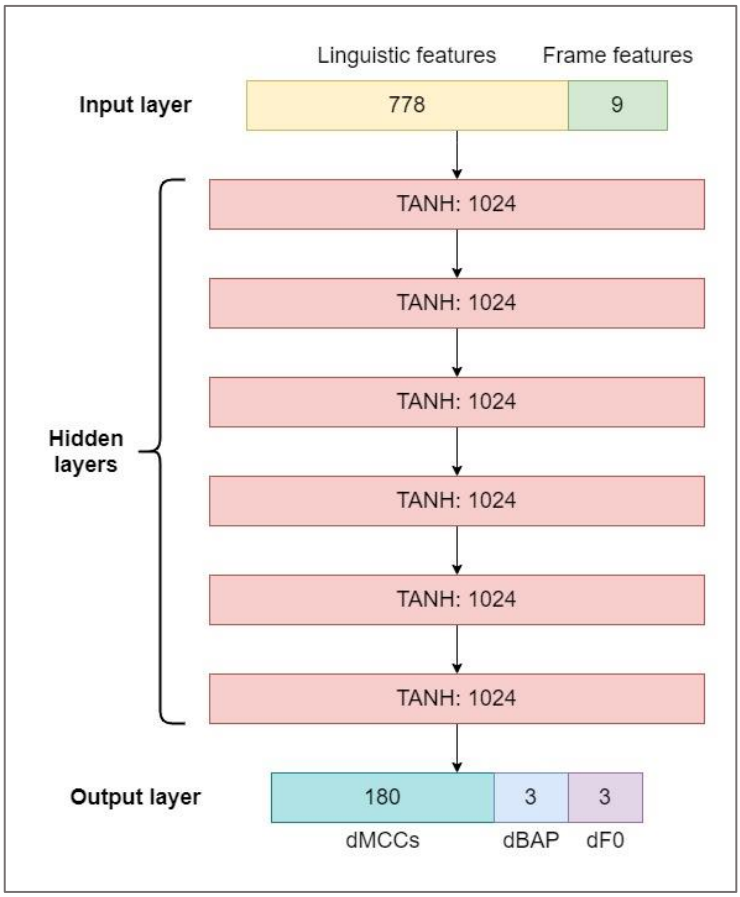
Mô hình âm học có nhiệm vụ lấy đầu vào là các đặc trưng ngôn ngữ cùng thông tin về thời gian xuất hiện từng trạng thái của âm vị và dự đoán đầu ra là các đặc trưng âm học của tín hiệu tiếng nói. Do đầu ra của mô hình âm học là các đặc trưng âm học cho từng khung tín hiệu có độ dài 5ms, nên đầu vào của mô hình này cũng phải là các đặc trưng ngôn ngữ học theo từng khung 5ms. Vì thế mà từ thông tin về thời gian xuất hiện các trạng thái của âm vị, đặc trưng ngôn ngữ được chia ra thành từng khung và được gắn thêm các thông tin về khung bao gồm: vị trí của khung trong trạng thái (tính từ đầu trạng thái), ví trí của khung trong trạng thái (tính từ cuối trạng thái), số khung của trạng thái hiện tại, số vị trí của trạng thái hiện tại trong âm vị, số khung của âm vị hiện tại, vị trí của khung trong âm vị, vị trí của trạng thái trong âm vị (tính từ đầu âm vị), ví trí của trạng thái trong âm vị (tính từ cuối âm vị).

Mô hình âm học là một mạng nơ ron học sâu với các đặc điểm:

- Đầu vào là véc tơ 787 chiều chứa, trong đó 778 chiều chứa đặc trưng ngôn ngữ của âm vị và 9 chứa các đặc trưng của khung.

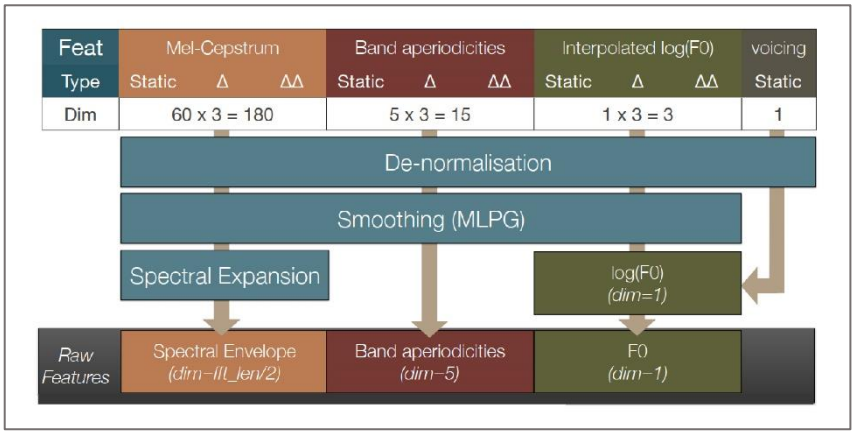
- Có 6 lớp ẩn, mỗi lớp có 1024 nốt và hàm kích hoạt là hàm tanh (phương trình PT 2.3).
- Đầu ra là véc tơ 186 chiều chứa đặc trưng âm học được ước lượng bao gồm MCCs, BAP, log F0, deltas và deltas-deltas của của 3 đại lượng đó.

Hình 3.6 mô tả chi tiết cấu trúc của mô hình âm học.



Hình 3.6 Cấu trúc mô hình âm học (Acoustic model)

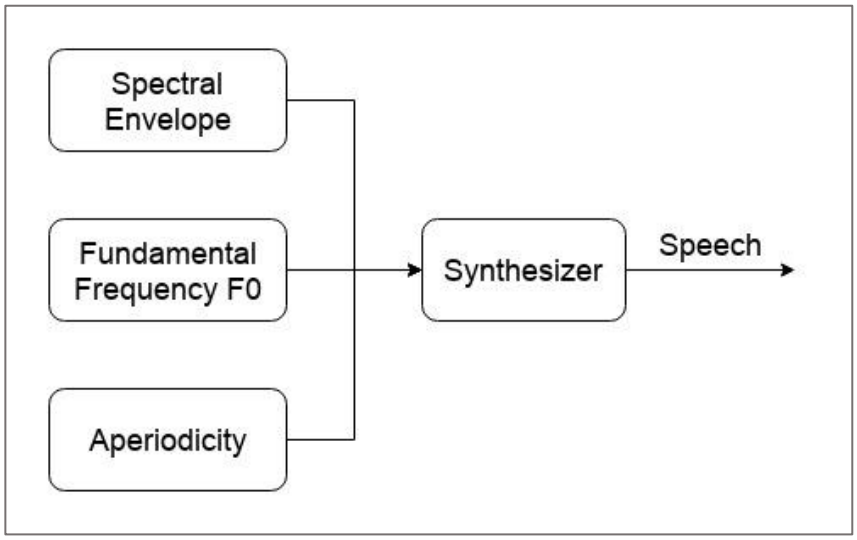
3.1.5 Tổng hợp tiếng nói từ đặc trưng âm học



Hình 3.7 Chuyển đổi từ đặc trưng âm học sang các đặc trưng gốc của vocoder [38]

Trong phần này, đồ án sử dụng tiếp tục sử dụng WORLD vocoder cho nhiệm vụ tổng hợp tiếng nói từ các tham số đặc trưng. Để tổng hợp tiếng nói bằng WORLD, trước tiên các đặc trưng âm học từ đầu ra của mô hình được chuyển đổi về các đặc trưng gốc của WORLD vocoder như mô tả trong Hình 3.7.

Trên Hình 3.8 ta thấy bộ Synthesizer của WORLD vocoder nhận các đầu vào đặc trưng âm học là đường bao phổ Spectral Envelope được tính từ 60 chiều của các hệ số mel, tần số cơ bản F0 và các tham số không tuần hoàn (Band aperiodicity) là đầu ra của mô hình âm học. Đầu ra của quá trình này chính là tín hiệu tiếng nói.



Hình 3.8 Tổng hợp tiếng nói từ các đặc trưng âm học bằng WORLD vocoder

3.2 Sử dụng phương pháp Transfer Learning cho thích ứng giọng nói

Như đã đề cập trong mục 2.3.1, transfer learning là một phương pháp đơn giản nhưng mang lại hiệu quả cho bài toán thích ứng giọng nói. Chính vì thế đồ án lựa chọ phương pháp này để thử nghiệm trước tiên.

Trong phương pháp này, đồ án thử nghiệm theo 2 hướng:

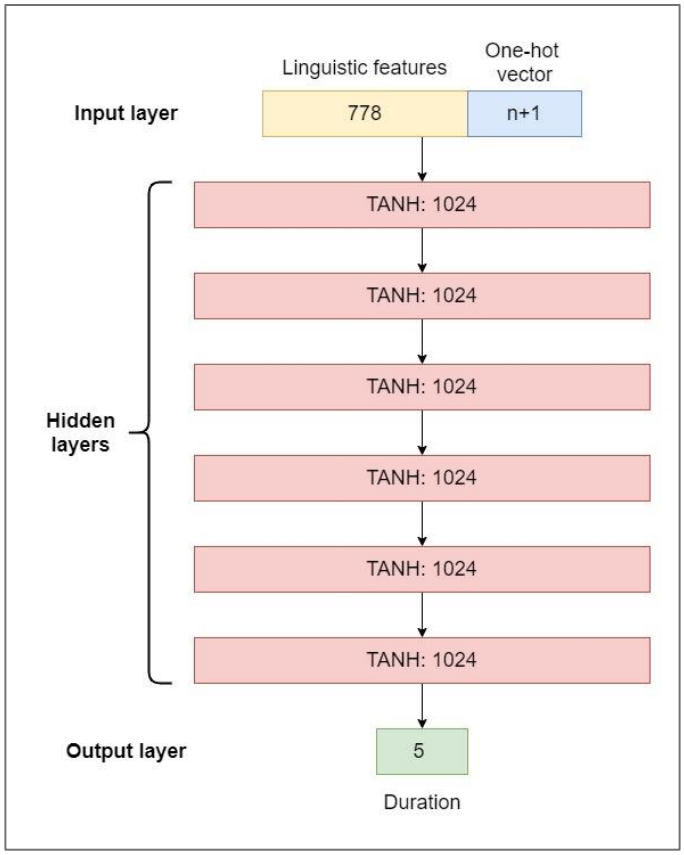
- Thích ứng giọng nói mới từ mô hình một người nói
- Thích ứng giọng nói mới từ mô hình nhiều người nói (Average voice model)

Với cả hai hướng này, mô hình thời gian và mô hình âm học đều được giữ nguyên kiến trúc. Trước tiên hai mô hình sẽ được huấn luyện trên tập dữ liệu gốc (một người nói hoặc nhiều người nói). Sau khi hoàn thành quá trình huấn luyện trên tập dữ liệu gốc, hai mô hình sẽ được huấn luyện tiếp tục trên tập dữ liệu thích ứng để thu được mô hình cho giọng nói thích ứng.

3.3 Sử dụng vec-tơ mã hóa người nói cho thích ứng giọng nói

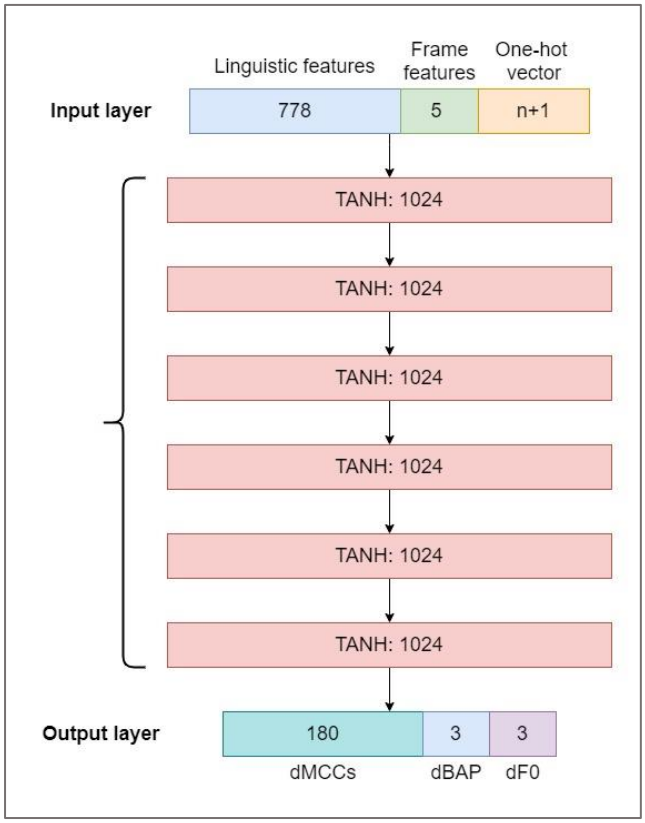
3.3.1 Phương pháp sử dụng one-hot vector

Với phương pháp này, đặc trưng đầu vào cho duration model và acoustic model được gắn thêm vec-tơ mã hóa người nói. Vec-tơ này có dạng $X = [x_1, x_2, \dots, x_{n+1}]$ với n là số lượng người nói trong tập dữ liệu. Với mỗi câu trong tập dữ liệu, x_i sẽ bằng 1 nếu người nói là người thứ i và x_j bằng 0 với mọi $j \neq i$. Véc tơ này sẽ được gắn với véc tơ đặc trưng ngôn ngữ để đưa vào huấn luyện, chi tiết của mô hình thời gian và mô hình âm học được mô tả ở Hình 3.9 và Hình 3.11.



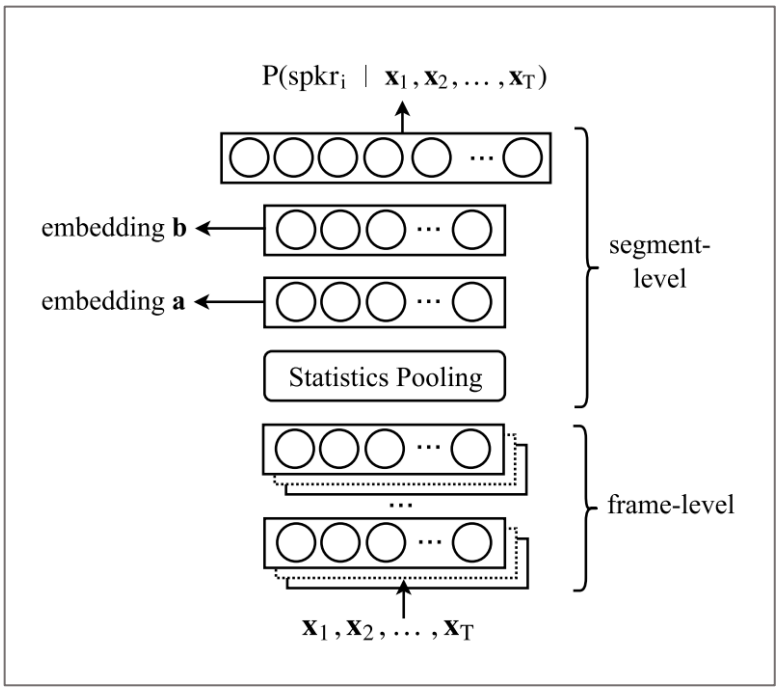
Hình 3.9 Cấu trúc mô hình thời gian cho phương pháp sử dụng one-hot vector

Trong quá trình thích ứng giọng nói sang người nói mới, véc tơ X được gán giá trị sao cho $x_i = 0 \; \forall \; i \leq n$ và $x_{n+1} = 1$. Mô hình sẽ được huấn luyện tiếp với dữ liệu của người nói mới để thu được giọng nói thích ứng.



Hình 3.11 Cấu trúc mô hình âm học cho phương pháp sử dụng one-hot vector

3.3.2 Phương pháp sử dụng x-vector



Hình 3.10 Mô hình mạng DNN được sử dụng để trích xuất x-vector [34]

Trong phương pháp này, véc tơ mã hóa người nói được sử dụng là x-vector. X-vector là véc tơ được trích xuất từ mô hình DNN được mô tả trong bài báo "*Deep Neural Network Embeddings for Text-Independent Speaker Verification"* [34]. Đồ án sử dụng mô hình đã được huấn luyện từ trước với tập dữ liệu VoxCeleb[13]. Đây là tập dữ liệu bao gồm các hơn một triệu đoạn tiếng nói, được trích xuất từ các video phỏng vấn được đăng tải lên YouTube. Trong bộ dữ liệu có hơn 7000 người nói với nhiều ngôn ngữ khác nhau, tổng độ dài bộ dữ liêu là hơn 2000 giờ.

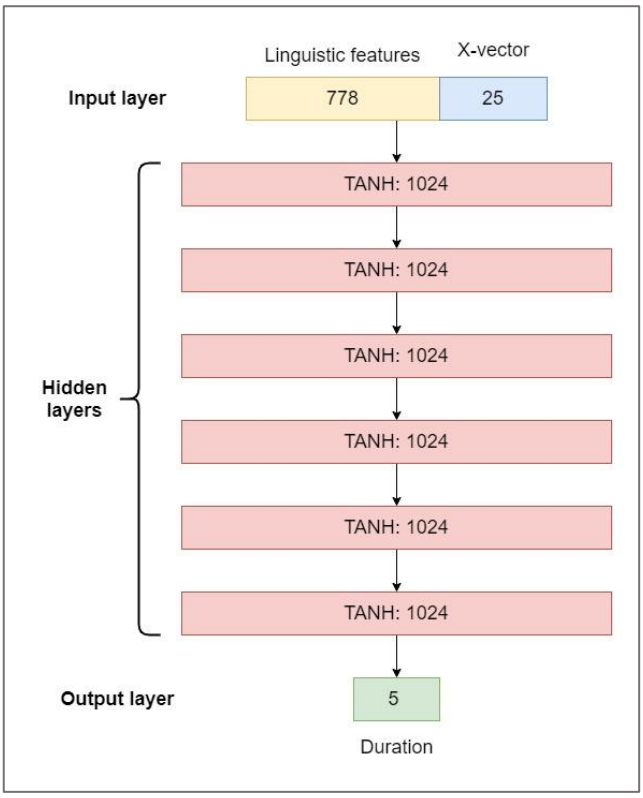
X-vector thu được từ quá trình trên là một véc tơ 200 chiều chưa thông tin mã hóa của người nói. Do số chiều này là khá lớn nếu so sánh tỷ lệ với các đặc trưng ngôn ngữ được sử dụng trong mô hình thời gian và mô hình âm học. Do đó, phương pháp phân tích thành phần chính (PCA) được sử dụng để giảm số chiều véc tơ xuống còn 25 chiều. Gọi $X \in \mathbb{R}^{N\times 200}$ là ma trận với mỗi hàng là một x-vector của của một người nói trong bộ dữ liệu, N là số lượng người nói của bộ dữ liệu. Phương pháp PCA được thực hiện như sau:

- Tính vector kỳ vọng của toàn bộ dữ liệu.
- Trừ mỗi điểm dữ liệu đi véc tơ kỳ vọng của toàn bộ dữ liệu thu được ma trận chuẩn hóa $\hat{X}$.
- Tính ma trận hiệp phương sai.
- Tính các trị riêng và véc tơ riêng có norm bằng 1 của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
- Chọn K véc tơ riêng ứng với K trị riêng lớn nhất để tạo thành một hệ trực giao (trong đồ án sử dụng K = 25). K véc tơ này, còn được gọi là các thành phần chính, tạo thành một không gian con mới.
- Chiếu bộ dữ liệu đã chuẩn hóa $\hat{X}$ xuống không gian con này thu được dữ liệu mới.
- Các thông tin về vector kỳ vọng và K véc tơ riêng đươc lưu lại cho tính toán sau này.

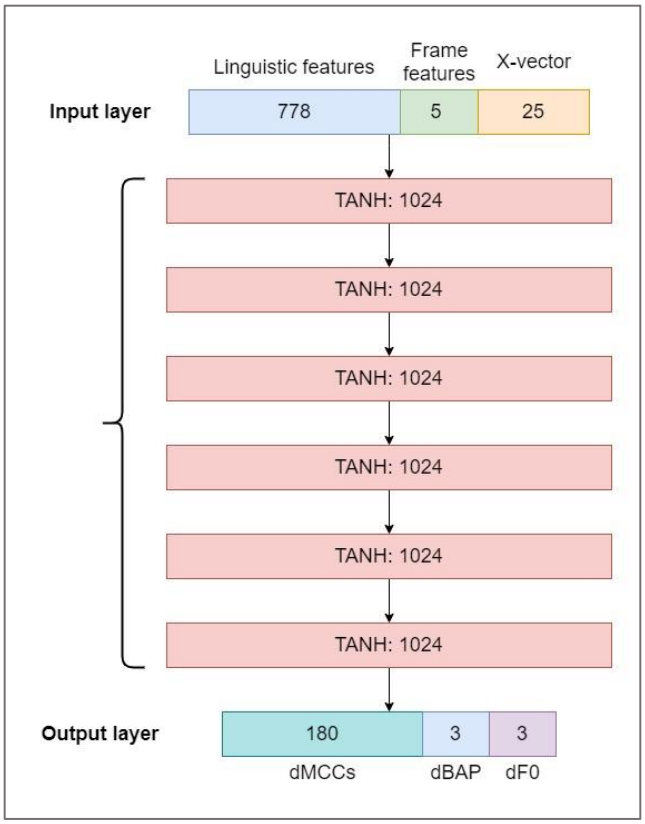
Dữ liệu x-vector mới sau quá trình này sẽ được gắn vào cùng các đặc trưng âm học để huấn luyện mô hình thời gian và mô hình âm học được mô tả tại Hình 3.12 và Hình 3.13

Trong quá trình thích ứng giọng nói, x-vector của người nói mới cũng được trích xuất nhờ vào mô hình DNN, sau đó được chuẩn hóa và đưa về 25 chiều nhờ vào các thông tin được của phương pháp PCA ở bước trước.

[13] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/



Hình 3.12 Cấu trúc mô hình thời gian cho phương pháp sử dụng x-vector



Hình 3.13 Cấu trúc mô hình âm học cho phương pháp sử dụng x-vector

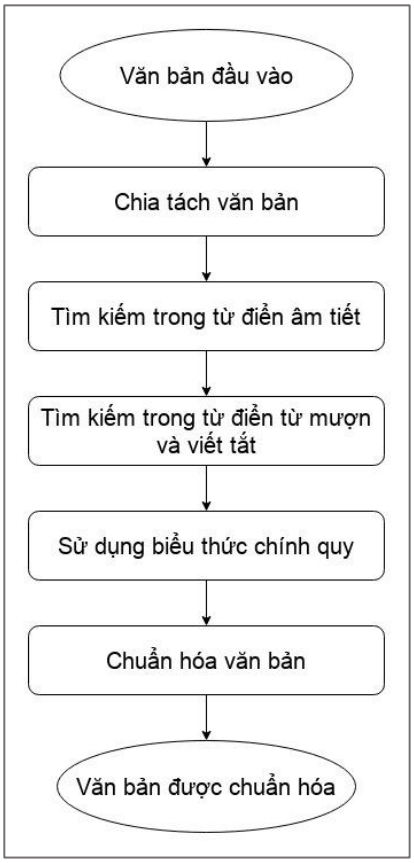
CHƯƠNG 4. THỬ NGHIỆM VÀ VÀ ĐÁNH GIÁ

4.1 Xử lý dữ liệu

4.1.1 Chuẩn hóa văn bản

Quá trình chuẩn hóa văn bản đầu vào có nhiệm vụ chính là làm cho văn bản đầu vào có thể đọc được một cách rõ ràng, nhất quán, chuẩn hóa các thành phần không chuẩn như từ mượn, từ viết tắt, số, ngày tháng. Ngoài ra các dấu câu như dấu chấm, dấu phẩy, dấu chấm phẩy, … được tách ra thành một thành phần độc lập thay vì gắn liền với từ đứng trước nó.

Các bước chuẩn hóa văn bản được thể hiện trong Hình 4.1.



Hình 4.1 Các bước chuẩn hóa văn bản đầu vào

Trong đó:

- Văn bản đầu vào sẽ được phân tách thành các các thành phần dựa theo khoảng trắng, từng thành phần này được tìm kiếm trong từ điểm âm tiết, nếu có trong từ điển thì nó là thành phần có thể đọc được, nếu không có sẽ tiếp tục được tìm kiếm trong từ điển từ mượn, từ viết tắt
- Những thành phần không có trong từ điển âm tiết được tìm kiếm ở trong từ điển viết tắt, nếu được tìm thấy thì thành phần này sẽ được chuyển thành một chuỗi các từ chuẩn theo từ điển âm tiết, nếu không tìm thấy sẽ được chuyển sang bước tiếp theo.
- Áp dụng biểu thức chính quy: Bước này áp dụng cho những thành phần mà không có trong cả hai từ điển nêu ở trên ví dụ như: ngày tháng viết

tắt, số, … Sử dụng biểu thức chính quy để tìm kiếm các mẫu có sẵn phù hợp với các thành phần này, sau đó thay thế chúng theo đúng mẫu phù hợp, ví dụ thành phần ngày tháng có dạng "…/..." sẽ được thay thế bằng "ngày … tháng …".

- Cuối cùng là bước chuẩn hóa văn bản: Ở bước này lưu các từ đã được chuẩn hóa ở các bước trước và phân văn bản thành từng câu.

4.1.2 Bộ dữ liệu

Để huấn luyện và đánh giá mô hình, đồ án sử dụng 6 bộ dữ liệu được mô tả trong Bảng 4.1.

Bảng 4.1 Các bộ dữ liệu sử dụng

Tên bộ dữ liệu	Số lượng câu	Tổng thời gian	Giới tính	Phương ngữ
FEMALE-1	12624	8 giờ 35 phút	Nữ	Miền Nam
FEMALE-2	3716	3 giờ 32 phút	Nữ	Miền Bắc
MALE	3767	4 giờ 41 phút	Nam	Miền Bắc
FEMALE-2 (30p)	530	30 phút	Nữ	Miền Bắc
MALE (30p)	418	30 phút	Nam	Miền Bắc
VTR-60	15600	20 giờ 39 phút		

Trong đó, bộ dữ liệu FEMALE-1, FEMALE-2 và MALE là các bộ dữ liệu có chất lượng cao, được ghi âm từ phòng thu chuyên nghiệp và phát thanh viên chuyên nghiệp. Để đánh giá mô hình thích ứng giọng nói, bộ dữ liệu FEMALE-2 (30p) là 30 phút dữ liệu được chọn ngẫu nhiên từ bộ FEMALE-2, bộ dữ liệu MALE (30p) cũng là 30 phút dữ liệu được chọn ngẫu nhiên từ bộ dữ liệu MALE.

Bộ dữ liệu VTR-60 là bộ dữ liệu được chọn ra từ bộ dữ liệu hơn 500 giờ cho bài toán nhận diện tiếng nói. Do chất lượng thu âm và chất lượng giọng nói không đồng đều nên đồ án chỉ sử dụng khoảng 20 giờ dữ liệu bao gồm các mẫu ghi âm của 60 người, chi tiết được mô tả trong Bảng 4.2.

Bảng 4.2 Thông tin chi tiết bộ dữ liệu VTR-60

Tên bộ dữ liệu	Giới tính		Phương ngữ		Số lượng câu / người
	Nam	Nữ	Miền Bắc	Miền Nam	
VTR-60	30	30	30	30	160

4.2 Huấn luyện mô hình

Trước khi đưa vào huấn luyện, mỗi bộ dữ liệu được chia làm 3 tập đó là tập huấn luyện (training set), tập kiểm định (validation set) và tập kiểm tra (test set) với tỷ lệ lần lượt là 90%, 5% và 5%.

Các mô hình được tối ưu với thuật toán Stochastic gradient descent với learning rate là 0.002, batch size là 256 và số epoch là 25. Thời gian huấn luyện cho từng mô hình được mô tả ở Bảng 4.3.

Hệ thống máy tính được sử dụng để huấn luyện và thử nghiệm có cấu hình như sau: CPU E5-2640 với 32 nhân, tần số 2.6 GHz; RAM 128 Gb; GPU Quadro K22000 với 4 Gb GPU Memory.

Bảng 4.3 Thời gian huấn luyện các mô hình

Mô hình		Bộ dữ liệu	Thời gian huấn luyện
Mô hình gốc		FEMALE-1	15 giờ 30 phút
		FEMALE-2	8 giờ 45 phút
		MALE	9 giờ 50 phút
		VTR-60	18 giờ 15 phút
		FEMALE-2 (30p)	25 phút
		MALE (30p)	30 phút
Transfer		FEMALE-2	27 phút
		MALE	26 phút
One hot vector	Mô hình gốc	VTR-60	20 giờ 37 phút
	Mô hình thích ứng	FEMALE-2	31 phút
		MALE	32 phút
X-vector	Mô hình gốc	VTR-60	19 giờ 22 phút
	Mô hình thích ứng	FEMALE-2	35 phút
		MALE	34 phút

4.3 Đánh giá kết quả

4.3.1 Đánh giá điểm MOS của các mô hình

Mục tiêu của đánh giá này là kiểm tra chất lượng hệ thống thích ứng giọng nói. Với giả định rằng dữ liệu để thích ứng là hạn chế, điểm MOS sẽ được dùng để so sánh hiệu quả giữa các phương pháp thích ứng giọng nói và so sánh với trường hợp có nhiều dữ liệu.

Phương pháp đánh giá điểm MOS như sau:

- Mời 18 người tham gia đánh giá và cho điểm chất lượng hệ thống.
- Tiêu chí cho điểm chất lượng hệ thống dựa trên độ tự nhiên và độ nghe hiểu của giọng nói tổng hợp.
- Tập dữ liệu đánh giá là tập gồm 25 tệp âm thanh được tổng hợp từ 25 văn bản khác nhau được lấy từ báo chí.
- Mỗi người đánh giá sẽ được nghe 10 tệp được chọn ra ngẫu nhiên cho mỗi mô hình và chấm điểm cho từng tệp.
- Điểm số được chấm ở thang điểm 5 với các mức: 1 – Rất tệ (Không nghe hiểu được), 2 – Tệ (Chỉ nghe hiểu được một số từ), 3 – Bình thường

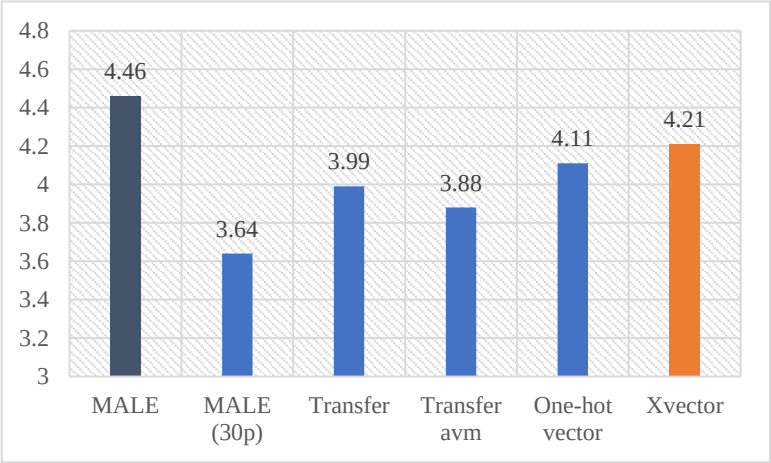
(Không nghe rõ nhưng vẫn hiểu nội dung), 4 – Tốt (Nghe rõ ràng tuy nhiên chưa được tự nhiên), 5 – Rất tốt (Giống như người thật nói).

- Kết quả cuối sẽ là trung bình các điểm số được đánh giá cho từng mô hình.

Bảng 4.4 Đánh giá điểm MOS cho cá mô hình

	Bộ dữ liệu		MOS
	FEMALE-2		4.32
	FEMALE-2 (30p)		3.44
	MALE		4.46
	MALE (30p)		3.64
Phương pháp	Dữ liệu gốc	Dữ liệu thích ứng	
Transfer	FEMALE-1	MALE (30p)	3.99
Transfer avm	VTR-60	MALE (30p)	3.88
One-hot vector	VTR-60	MALE (30p)	4.12
X-vector	VTR-60	MALE (30p)	4.21
Transfer	FEMALE-1	FEMALE-2 (30p)	3.41
Transfer avm	VTR-60	FEMALE-2 (30p)	3.46
One-hot vector	VTR-60	FEMALE-2 (30p)	3.80
X-vector	VTR-60	FEMALE-2 (30p)	3.82

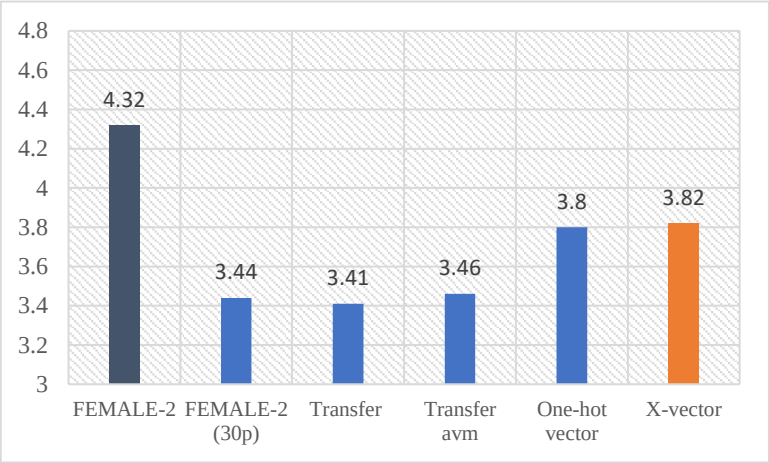
Kết quả đánh giá được nêu trong Bảng 4.4, để đảm bảo tính công bằng kết quả được đánh giá theo 2 phần riêng biệt, dựa trên tập dữ liệu được sử dụng để thích ứng bao gồm bộ FEMALE-2 (30p) và bộ MALE (30p).



Hình 4.2 Đánh giá điểm MOS các mô hình giọng nam

Đối với mô hình giọng nam, ta có thể thấy trong trường hợp dữ liệu hạn chế (chỉ 30 phút), mô hình tổng hợp tiếng nói gốc có kết quả tệ hơn rất nhiều so với trường hợp nhiều dữ liệu (4 giờ 41 phút). Sau khi áp dụng phương pháp transfer learning, kết quả được cải thiện rõ rệt. Mô hình thích ứng từ giọng một người nói cho kết

quả tốt hơn so với mô hình thích ứng từ giọng trung bình. Sau khi áp dụng thêm phương pháp sử dụng véc tơ mã hóa người nói, kết quả cũng được cải thiện rõ rệt, đặc biệt là phương pháp sử dụng x-vector.



Hình 4.3 Đánh giá điểm MOS các mô hình giọng nữ

Đối với mô hình giọng nữ, ta cũng có thể thấy trong trường hợp dữ liệu hạn chế (chỉ 30 phút), mô hình tổng hợp tiếng nói gốc có kết quả tệ hơn rất nhiều so với trường hợp nhiều dữ liệu (3 giờ 32 phút). Sau khi áp dụng phương pháp transfer learning, kết quả được cải thiện rõ rệt. Mô hình thích ứng từ giọng trung bình cho kết quả tốt hơn so với mô hình thích ứng từ giọng một người nói. Sau khi áp dụng thêm phương pháp sử dụng véc tơ mã hóa người nói, kết quả cũng được cải thiện rõ rệt, hai phương pháp sử dụng one-hot vector và x-vector có kết quả tương đương nhau.

Tổng quát có thể thấy rằng phương pháp sử dụng x-vector là phương pháp hiệu quả nhất. Mặc dù vậy, kết quả khi sử dụng 30 phút để thích ứng vẫn thấp hơn so với trường hợp sử dụng đầy đủ dữ liệu.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN

5.1 Kết luận

Qua những kết quả được đánh giá ở mục 4.3, có thể thấy rằng đồ án đã đạt được mục đích khi đã xây đựng được mô hình thích ứng giọng nói cho tổng hợp tiếng nói tiếng Việt dựa trên công nghệ học sâu, các mô hình cũng mang lại những cải thiện cho chất lượng giọng nói thích ứng. Mặc dù vậy đồ án vẫn còn nhiều hạn chế như là mức độ tự nhiên của giọng nói thích ứng chưa thể so sánh với giọng nói con người, việc đánh giá các mô hình cũng chưa đầy đủ và chính xác bởi những người thực hiện đánh giá đều không phải chuyên gia và có thể có nhiều sai sót. Bên cạnh đó là việc chưa áp dụng các phương pháp vào các hệ thống thực tế, do đó thiếu các đánh giá về hiệu năng cũng như là hiệu quả kinh tế.

Sau quá trình hoàn thành đồ án này, bản thân tôi đã đạt thu được rất nhiều kiến thức và kinh nghiệm như:

- Kiến thức về công nghệ tổng hợp tiếng nói, cũng như các phương pháp thích ứng giọng nói
- Kiến thức và kinh nghiệm về xây dựng mô hình thích ứng giọng nói dựa trên công nghệ học sâu.
- Các kỹ năng cần thiết như là kỹ năng tự tìm kiếm tài liệu, tổng hợp thông tin, kỹ năng chế bản, kỹ năng trình bày và viết báo cáo.

5.2 Phương hướng phát triển đồ án

Mặc dù các phương pháp thích ứng giọng nói được sử dụng là hiệu quả tuy nhiên chất lượng của giọng nói thích ứng vẫn chưa thực sự đạt độ tự nhiên khi so sánh với giọng nói con người. Vì vậy hướng phát triển tiếp theo của đồ án là tiếp tục cải thiện chất lượng của giọng nói thích ứng.

Vì vậy tác giả đề xuất một số phương án để có thể cải thiện nhược điểm và phát triển đồ án này trong thời gian tới như sau:

- Thay thế mô hình DNN bằng mô hình Seq2seq đang dần được sử dụng rộng rãi hiện nay để nâng cao độ tự nhiên của giọng nói thích ứng.
- Áp dụng các phương pháp chưa được thử nghiệm như là LHUC, FST, ... (được đề cập trong mục 2.3)
- Hướng tới kết hợp bài toán thích ứng giọng nói và bài toán tổng hợp tiếng nói cho nhiều người nói (multi-speaker) để tạo ra hệ thống có khả năng tổng hợp nhiều giọng nói cũng như linh hoạt trong việc thêm các giọng mới vào hệ thống.
- Áp dụng mô hình thích ứng giọng nói cho các sản phẩm tại đơn vị đang làm việc là Trung tâm Không gian mạng Viettel để cải thiện sản phẩm cũng như nhận được đánh giá phản hồi từ phía người sử dụng. Qua đó có thể đánh giá khách quan hơn về hiệu quả của các phương pháp.

TÀI LIỆU THAM KHẢO

[1] J. J. Ohala, "Christian Gottlieb Kratzenstein: pioneer in speech synthesis," *Proc 17th*, 2011.

[2] W. v. Kempelen, "Mechanism of the human speech with description of its speaking machine".

[3] Lawrence J. Raphael, Gloria J. Borden, Katherine S. Harris, Speech Science Primer: Physiology, Acoustics, and Perception of Speech.

[4] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of the Acoustical Society of America,* 1987.

[5] P. T. Sơn, P. T. Nghĩa, "Một số vấn đề về tổng hợp tiếng nói tiếng Việt," p. 5, 2014.

[6] P. Taylor, "Text-to-speech synthesis," *Cambridge University Press,* 2009.

[7] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, vol. 115, pp. 853-870, 2004.

[8] Dartmouth College, "Music and Computers," 1993.

[9] Ilya Sutskever, Oriol Vinyals, Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014.

[10] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A Saurous, Yannis Agiomvrgiannakis, Yonghui Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 2016.

[12] D. D. Tran, "Synthèse de la parole à partir du texte en langue vietnamienne," *PhD Thesis, Grenoble INP*, 2007.

[13] Thao Van Do, Do-Dat Tran, Thu-Trang Thi Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology*, 2011, pp. 165-171.

[14] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King, "A study of speaker adaptation for DNN-based speech synthesis," *Proc. Interspeech*, 2015.

[15] J-L Gauvain, Chin-Hui Lee, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden," *Computer Speech & Language*, vol. 9, p. 291–298, 1994.

[16] 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*, vol. 17, pp. 66-83, 2009.

[17] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[18] Yaniv Taigman, Lior Wolf, Adam Polyak, Eliya Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *International Conference on Learning Representations (ICLR)*, 2018.

[19] D. K. Ninh, "A Speaker-Adaptive HMM-based Vietnamese Text-to-Speech System," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019.

[20] Thi Thu Trang Nguyen, Do Dat Tran, Albert Rilliard, Christophe d'Alessandro, Thi Ngoc Yen Pham, "Intonation issues in HMM-based speech synthesis for Vietnamese," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[21] Zhizheng Wu, Oliver Watts, Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System," *SSW*, pp. 202-207, 2016.

[22] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Society of Japan*, vol. 27, no. 6, pp. 349-353, 2006.

[23] Masanori Morise, Fumiya Yokomori, Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, pp. 1877-1884, 2016.

[24] Felipe Espic, Cassia Valentini-Botinhao, Simon King, "Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis.," *INTERSPEECH*, pp. 1383-1387, 2017.

[25] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2009.

[26] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1-7, 2015.

[27] M. Morise, "PLATINUM: A method to extract excitation signals for voice synthesis system," *Acoust. Sci. Technol.*, vol. 33, pp. 123-125, 2012.

[28] Nobukatsu Hojo, Yusuke Ijima, Hideyuki Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, pp. 462-472, 2018.

[29] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, Junichi Yamagishi, "Adapting and Controlling DNN-Based Speech Synthesis Using Input Codes," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4905-4909, 2017.

[30] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. IEEE Spoken Language Technology Workshop*, 2014.

[31] Tomoki Toda, Alan W Black, Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222-2235, 2007.

[32] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*, p. 7942–7946, 2013.

[33] Q. T. Do, "Vita: A Toolkit for Vietnamese segmentation, chunking, part of speech tagging and morphological analyzer," 2015.

[34] David Snyder, Daniel Garcia-Romero, Daniel Povey, Sanjeev Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," *INTERSPEECH 2017*, 2017.

[35] Heiga Zen, Andrew Senior, Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013.

[36] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proc. IEEE*, vol. 101, p. 1234–1252, May 2013.

[37] M. Nielsen, Neural networks and deep learning, 2015.

[38] Simon King, Oliver Watts, Srikanth Ronanki, Zhizheng Wu, Felipe Espic, "Deep Learning for Text-to-Speech Synthesis, using the Merlin toolkit," [Online].

[39] Heiga Zen, Tomoki Toda, "An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Ninth European Conference on Speech Communication and Technology (Eurospeech)*, 2005.

PHỤ LỤC

Phụ lục A: Cấu trúc của một nhãn biểu diễn ngữ cảnh của âm vị

Cấu trúc mỗi nhãn (tương ứng là mỗi dòng trong tệp chứa các nhãn):

p1^p2-p3+p4=p5@p6_p7/A:a1_a2/B:b1-b2@b3-b4&b5-b6/C:c1+c2/D:d1-d2/E:e1+e2/F:f1-f2/G:g1-g2/H:h1=h2@h3=h4/I:i1_i2/J:j1+j2-j3

Giải thích các trường cho nhãn trên như sau:

Trường	Mô tả
P1	Âm vị phía trước của âm vị phía trước âm vị hiện tại
P2	Âm vị phía trước âm vị hiện tại
P3	Âm vị hiện tại
P4	Âm vị tiếp theo
P5	Âm vị phía sau âm vị tiếp theo
P6	Vị trí của âm vị hiện tại trong từ hiện tại (tính từ phía trước)
P7	Vị trí của âm vị hiện tại trong từ hiện tại (tính từ phía sau)
A1	Thanh điệu ở âm tiết phía trước
A2	Số lượng âm vị trong âm tiết phía trước
B1	Thanh điệu của âm tiết hiện tại
B2	Số lượng âm vị trong âm tiết hiện tại
B3	Vị trí của âm tiết trong từ hiện tại (tính từ phía trước)
B4	Vị trí của âm tiết trong từ hiện tại (tính từ phía sau)
B5	Vị trí của âm tiết hiện tại trong cụm từ hiện tại (tính từ phía trước)
B6	Vị trí của âm tiết hiện tại trong cụm từ hiện tại (tính từ phía sau)
C1	Thanh điệu của từ tiếp theo
C2	Số lượng âm vị trong âm tiết tiếp theo
D1	Nhãn từ loại của từ phía trước
D2	Số lượng âm vị trong từ phía trước
E1	Nhãn từ loại của từ hiện tại
E2	Số lượng âm vị trong từ hiện tại
F1	Nhãn của từ loại tiếp theo

F2	Số lượng âm vị trong từ tiếp theo
G1	Số lượng âm vị trong cụm từ phía trước
G2	Số lượng từ trong cụm từ phía trước
H1	Số lượng âm vị trong cụm từ hiện tại
H2	Số lượng từ trong cụm từ hiện tại
H3	Vị trí của cụm hiện tại trong câu (tính từ phía trước)
H4	Vị trí của cụm hiện tại trong câu (tính từ phía sau)
I1	Số lượng âm vị trong cụm từ tiếp theo
I2	Số lượng từ trong cụm từ tiếp theo
J1	Số lượng âm vị trong câu
J2	Số lượng từ trong câu
J3	Số lượng cụm từ trong câu