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

ĐỀ TÀI TỐT NGHIỆP

Biểu mẫu của Đề tài/khóa luận tốt nghiệp theo qui định của viện, tuy nhiên cần đảm bảo giáo viên giao đề tài ký và ghi rõ họ và tên.

Trường hợp có 2 giáo viên hướng dẫn thì sẽ cùng ký tên.

Giáo viên hướng dẫn
Ký và ghi rõ họ tên

Lời cảm ơn

Đây là mục tùy chọn, nên viết phần cảm ơn ngắn gọn, tránh dùng các từ sáo rỗng, giới hạn trong khoảng 100-150 từ.

Tóm tắt nội dung đồ án

Tóm tắt nội dung của đồ án tốt nghiệp trong khoảng tối đa 300 chữ. Phần tóm tắt cần nêu được các ý: vấn đề cần thực hiện; phương pháp thực hiện; công cụ sử dụng (phần mềm, phần cứng…); kết quả của đồ án có phù hợp với các vấn đề đã đặt ra hay không; tính thực tế của đồ án, định hướng phát triển mở rộng của đồ án (nếu có); các kiến thức và kỹ năng mà sinh viên đã đạt được.

(Đối với luận văn thạc sĩ, phần tóm tắt được in trong một bản nộp riêng)

Sinh viên thực hiện
Ký và ghi rõ họ tên

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

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]). 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.

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, ví dụ như máy VOCODER của phòng thí nghiệm Bell (1930) được điều khiển bằng bàn phím và có thể phát âm rõ ràng, máy ghi âm và tổng hợp giọng nói của nhà vật lý John Larry Kelly, Jr có thể tạo ra bài hát Daisy Bell với âm nhạc phụ họa bởi Max Mathews.

Từ đó đến nay, công nghệ tổng hợp tiếng nói đã có những bước tiến bộ vượt bậc nhờ vào các kỹ thuật học máy, học sâu. Chất lượng giọng nói tổng hợp ngày càng có độ tự nhiên, dễ nghe, thậm chí nhiều hệ thống đạt được độ tự nhiên tiệm cận với giọng nói con người.

1.1.2 Ứng dụng của tổng hợp tiếng nói

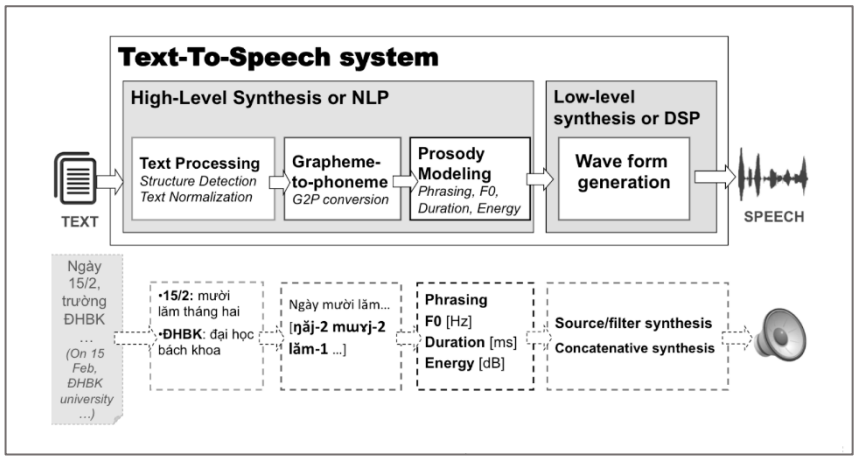
Bên cạnh sự phát triển về mặt chất lượng tiếng nói tổng hợp, các hệ thống tổng hợp tiếng nói cũng được áp dụng ngày càng rộng rãi trong nhiều lĩnh vực của cuộc sống. Phổ biến nhất có thể kể tới các ứng dụng sách nói, báo nói, với số lượng rất lớn sách mới xuất bản mỗi năm và tin tức cập nhập mỗi ngày , 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, Google Assistant của Google, Cotarna của Mircrosoft, ... đều áp dụng công nghệ tổng 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 giúp khách hàng tránh khỏi tình trạng chờ đợi khi số lượng nhân viên tư vấn là hạn chế.

Có thể thấy rằng việc phát triển công nghệ tổng hợp tiếng nói là rất cần thiết để cải thiện chất lượng cuộc sống cũng như đóng góp vào phát triển kinh tế.

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 [1]. 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 [2]. 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 [*]

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

Trong một hệ thống tổng hợp tiếng nói, khối xử lý ngôn ngữ tự nhiên 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 [1]. Quá trình xử lý ngôn ngữ tự nhiên thường bao gồm 3 bước:

- Xử lý và chuẩn hóa văn bản (Text Processing)
- Phân tích cách phát âm (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 [1].

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 [3]. 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). 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 chết 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 mạnh, 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)

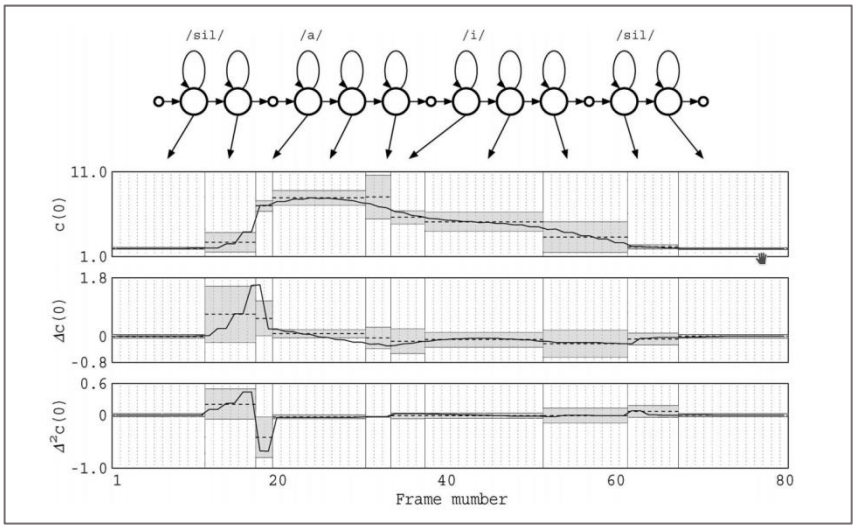
Tổng hợp chọn đơn vị dùng một cơ sở dữ liệu lớn các giọng nói ghi âm. Trong đó, mỗi câu được tách thành các đơn vị khác nhau như: các tiếng đơn lẻ, các âm tiết, hình vị, từ, nhóm từ hoặc câu văn. Một bảng tra các đơn vị được lập ra dựa trên các phần đã tách và các thông số âm học như tần số cơ bản, thời 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

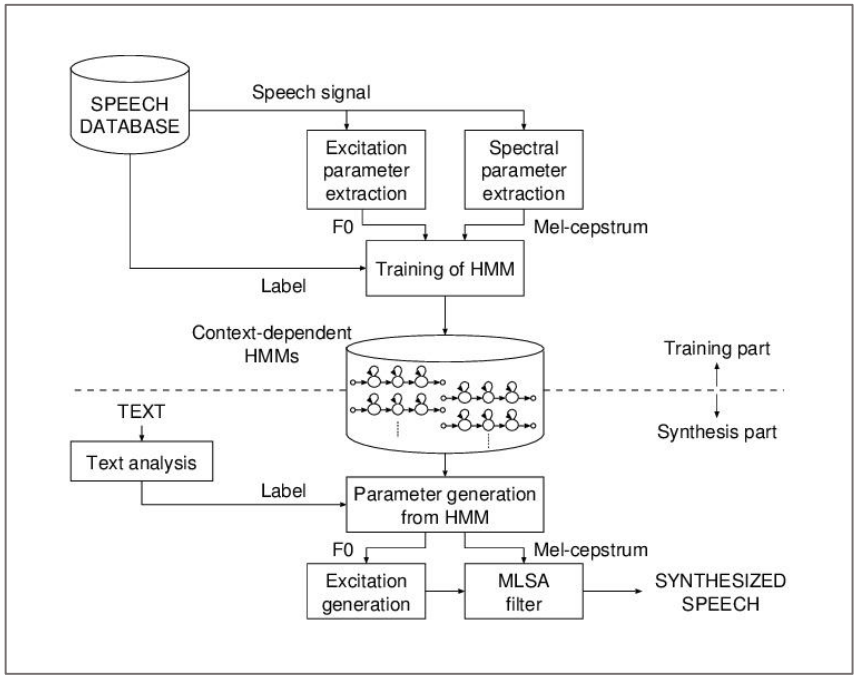
Phương pháp tổng hợp tiếng nói dùng tham số thống kê là phương pháp dựa trên mô hình Markov ẩn (HMM) [*]. Ở đây, HMM là một mô hình thống kê, được sử dụng để mô hình hóa các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể.



Hình 1.2 Ví dụ về thống kê và các tham số được tạo từ HMM cấp câu bao gồm các HMM cấp âm vị cho /a/ và /i/.

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 một hệ thống tổng hợp tiếng nói HMM

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 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 TTS dựa trên HMM 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 dựa trên HMM vẫn còn những tồn tại. 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 (DNN)

- Tổng quan

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 [*], 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 [*], 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. Kết quả ban đầu cho thấy phiên bản thứ hai của hệ thống Hoa Súng có sự cải thiện về mặt chất lượng, tuy nhiên dữ liệu kiểm thử không được thiết kế để bao trùm toàn bộ đơn vị âm, thêm

nữa không có sự kết nối giữa quá trình chọn đơn vị âm và quá trình chọn đơn vị như một bán âm tiết trong việc tính toán chi phí mục tiêu và chi phí ghép nối. Kết quả là tổng chi phí không được tối ưu hóa cho những câu cần bán âm tiết.

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, sản phẩm của tập đoàn FPT hay hệ thống tổng hợp tiếng nói tiếng Việt Mica TTS (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 của tập đoàn Viettel, hệ thống tổng hợp tiếng nói của Zalo 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, Báo Mới, …), trợ lý ảo (Google Assistant) hay tổng đài trả lời tự động (Viettel Cyber Callbot).

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 [4]. 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. Thích ứng giọng nói không phải là một chủ đề mới mà đã được nghiên cứu kỹ đặc biệt trên hệ thống tổng hợp tiếng nói dựa theo mô hình Markov ẩn [5] và các hệ thống nhận dạng tiếng nói.

Đối với các hệ thống tổng hợp dựa theo mô hình Markov ẩn, Maximum likelihood linear regression (MLLR) và Constrained maximum likelihood linear regression là hai kỹ thuật phổ biến áp dụng một số dạng chuyển đổi tuyến tính cho phân phối Gausian của mô hình gốc. 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.

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 [*] 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 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 [6]. 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.

Neural là tính từ của neuron (nơ-ron), network chỉ cấu trúc đồ thị nên neural network (NN) là một hệ thống tính toán lấy cảm hứng từ sự hoạt động của các nơ-ron trong hệ thần kinh.

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.

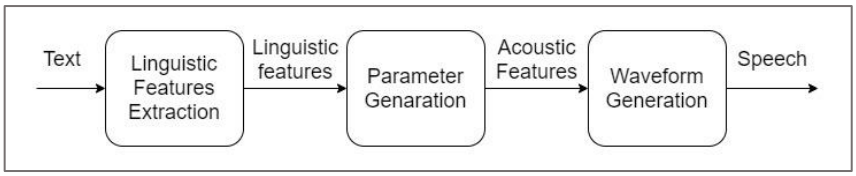
θ là hàm kích hoạt (activation functoion). Có nhiều hàm kích hoạt thường được sử dụng như là hàm sigmoid, hàm tanh, hàm ReLU.

2.1.3 Mạng nơ ron học sâu

Mạng nơ ron học sâu (DNN-Deep neural Network) là một mạng nơ ron nhân tạo (ANN) với nhiều đơn vị lớp ẩn giữa lớp đầu vào và đầu ra.

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 (DNN) ngày càng được sử dụng rộng rãi và cho thấy ưu điểm so với các phương pháp thông thường (như HMM hoặc GMM). Hình 2.1 mô tả một kiến trúc cơ bản của hệ thống tổng hợp tiếng nói dựa trên phương pháp học sâu.



Hình 2.1 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:

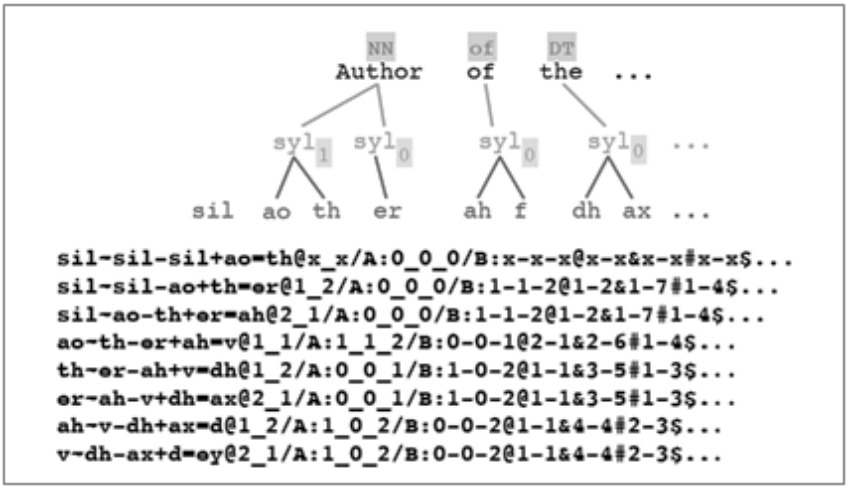
- Mô đun trích chọn đặc trưng ngôn ngữ: văn bản đầu vào được xử lý, phân tích và trích chọn bởi bộ Linguistic Features Extraction ra thành các vec tơ đặc trưng ngôn ngữ, các vec tơ này thường bao gồm các thông tin về chuỗi âm vị, vị trí tương đối của âm vị trong câu, cụm từ hay từ, số lượng âm vị trong 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 [*]. Để lấy được các đặc trưng ngôn ngữ trên, văn bản đầu vào sẽ được xử lý theo các bước như sau:

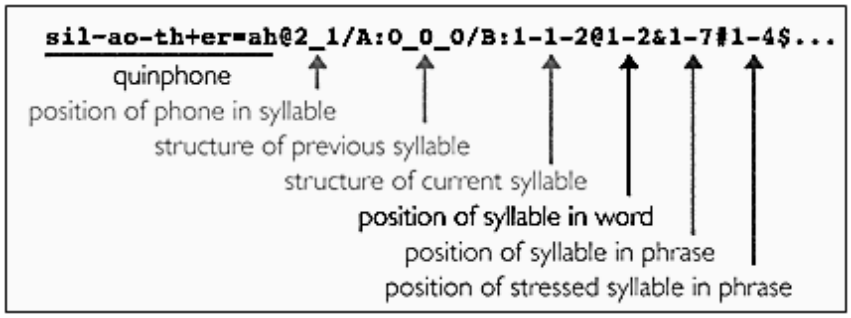
- Văn bản đầu vào sẽ được chuyển thành một chuỗi âm vị nhờ từ điển phiên âm tương ứng với ngôn ngữ đang tổng hợp
- Văn bản đầu vào sẽ được cho qua một hệ thống xử lý ngôn ngữ tự nhiên để trích chọn các thông tin về ngôn ngữ, hệ thống xử lý ngôn ngữ tự nhiên này được xây dựng trên cơ sở ba mô hình: mô 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 2.2 Biểu diễn đặc trưng ngôn ngữ học của văn bản [*]

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 các âm vị hiện tại, phía trước, 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, phía trước, 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, và các từ kề nó.
- Thông tin mức cụm từ: bao gồm số lượng từ, âm tiết trong cụm từ hiện tại, phía trước, phía sau.
- Thông tin mức câu: bao gồm thông tin vế số lượng âm tiết, số lượng từ, số lượng cụm từ trong câu.

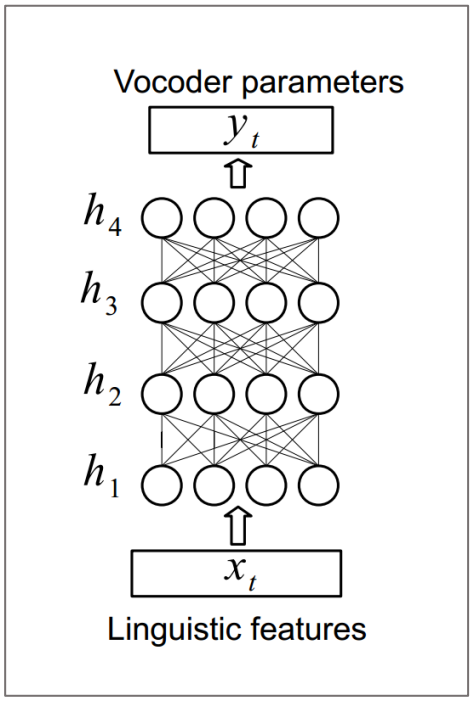


Hình 2.3 Thông tin đặc trưng ngôn ngữ ở mức âm vị [*]

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 ** minh họa thời gian cho từng trạng thái trong mỗi âm vị (thông thường ta sử dụng 5 trạng thái cho mỗi 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 2.4, 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).



Hình 2.4 Một mình họa về mạng nơ ron học sâu với bốn lớp ẩn

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 phần

***. 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. Chi tiết cách áp dụng câu hỏi để chuyển hóa các thông tin đặc trưng ngôn ngữ thành véc tơ nhị phân được thể hiện trong hình ** 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 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 phương trình **.

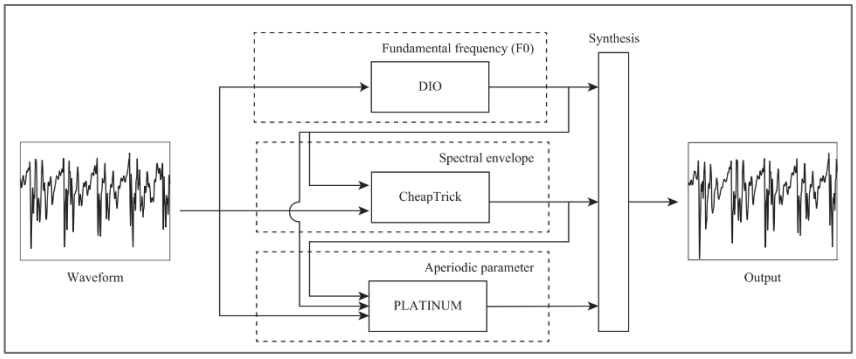
Trong đó ** là hàm kích hoạt phi tuyến (thường là hàm tanh), ... và ... là ma trận trọng số.

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 [*], WORLD vocoder [*], Magphase vocoder [*]. 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



Hình 2.5 Tổng quan về hệ thống WORLD vocoder [*]

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 [*], đường bao phổ được ước lượng bởi phương pháp CheapTrick [*] và tín hiệu kích thích được ước lượng bởi phương pháp PLATINUM [*] và được sử dụng như tham số không tuần. Hình ** mô tả quá trình xử lý của WORLD vocoder trong hai giai đoạn phân tích và tổng hợp tín hiệu tiếng nói.

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:

$$\int_0^{3T_0} \big(y(t)w(t)\big)^2 dt = 1.125 \int_0^{T_0} y^2(t)dt \qquad PT\ 2.2$$

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:

$$P_s(\omega) = \frac{3}{2\omega_0} \int_{-\frac{\omega_0}{3}}^{\frac{\omega_0}{3}} P(\omega + \lambda)d\lambda \qquad PT\ 2.3$$

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

$$P_l(\omega) = \exp\big(\mathcal{F}\big[l_s(\tau)l_q(\tau)p_s(\tau)\big]\big) \qquad PT\ 2.4$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_o}\right) \qquad PT\ 2.5$$

$$l_s(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_o \tau} \qquad \textit{PT 2.6}$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right) \qquad \textit{PT 2.7}$$

$$p_s(\tau) = \mathcal{F}^{-1}\left[\log\left(P_s(\omega)\right)\right] \qquad \textit{PT 2.8}$$

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:

$$S_m(\omega) = \exp(\mathcal{F}[c_m(\tau)]) \qquad \textit{PT 2.9}$$

$$c_m(\tau) = \begin{cases} 2c(\tau) & (\tau > 0) \\ c(\tau) & (\tau = 0) \\ 0 & (\tau < 0) \end{cases} \qquad \textit{PT 2.10}$$

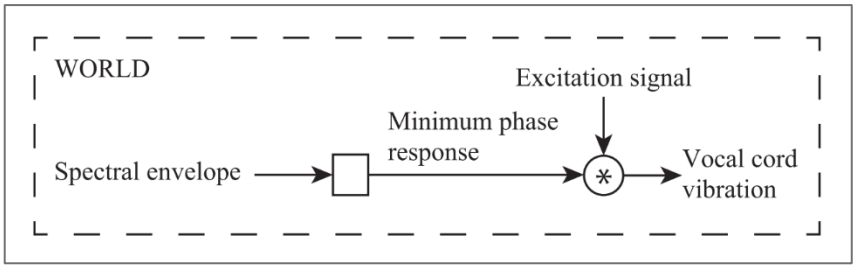
$$c(\tau) = \mathcal{F}^{-1}\left[\log\left(P_1(\omega)\right)\right] \qquad \textit{PT 2.11}$$

Tín hiệu kích thích được biểu diễn như sau:

$$x_p(t) = \mathcal{F}^{-1}\left[X_p(\omega)\right] \qquad \textit{PT 2.12}$$

$$X_p(\omega) = \frac{X(\omega)}{S_m(\omega)} \qquad \textit{PT 2.13}$$

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 2.6.



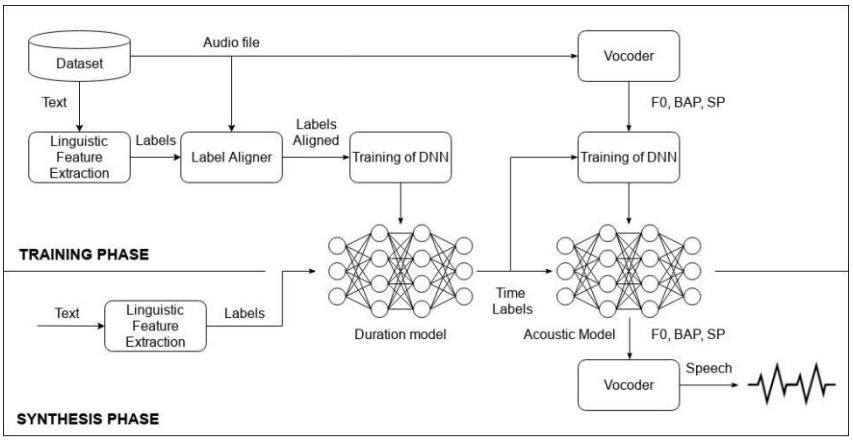
Hình 2.6 Tổng hợp tiếng nói với WORLD vocoder

2.3 Chuyển đổi giọng nói dựa trên công nghệ học sâu

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

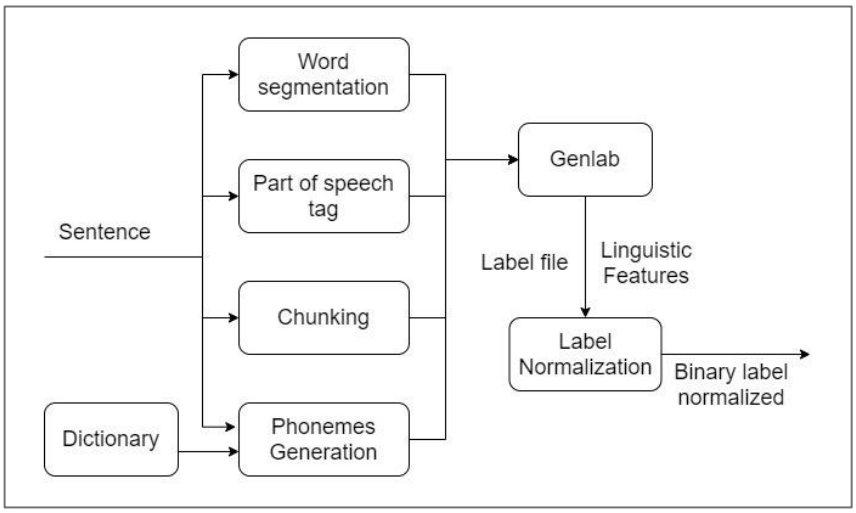


Hình 3.1 Tổng quan mô hình tổng hợp tiếng nói

3.1.2 Trích chọn đặc trưng ngôn ngữ

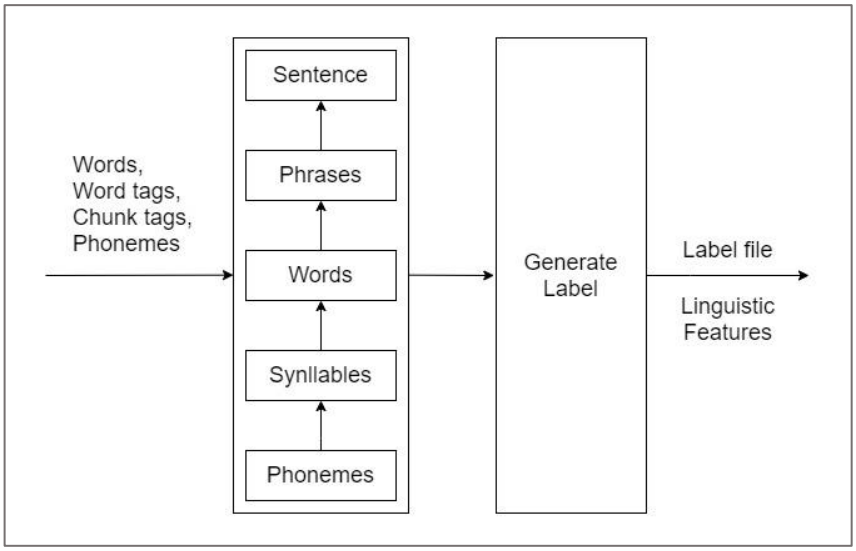
Để tạo dữ liệu đầu vào cho mô hình thời gian và mô hình âm học, đồ án sử dụng công cụ Vita [*] để trích chọn đặc trưng ngôn ngữ. Một từ điển phiên âm âm tiết tiếng Việt với khoảng 6700 từ được sử dụng cho quá trình tạo chuỗi âm vị.

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 **, 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 ** 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 phần *** 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 phần 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

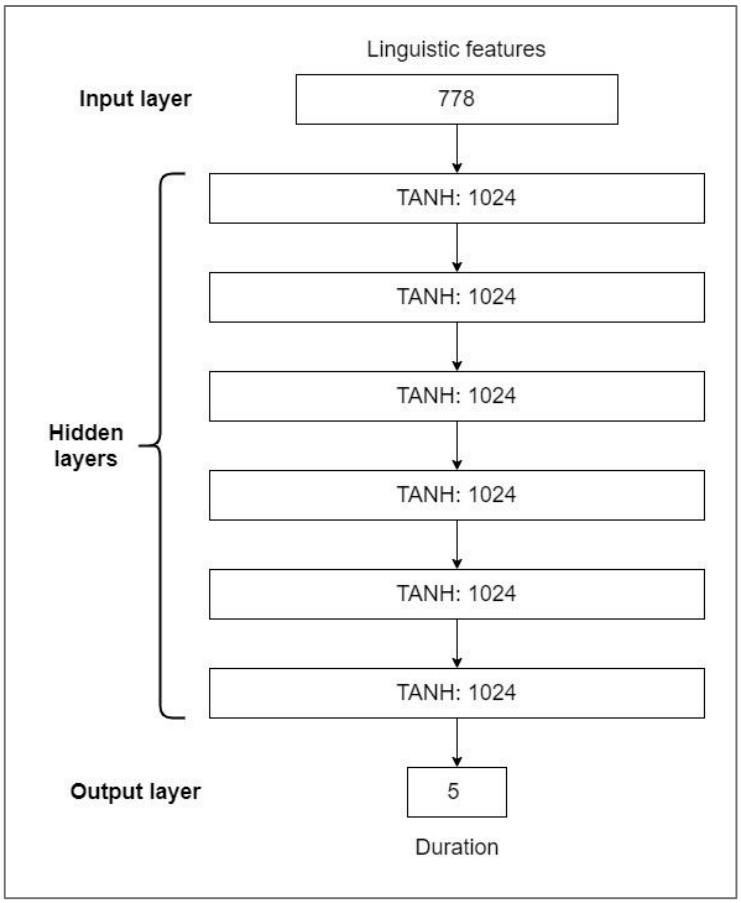
Để trích chọn đặc trưng âm học, đồ án sử dụng WORLD vocoder. Các đặc trưng được trích xuất bao gồm 60 chiều hệ số mel (MCCs), 5 chiều hệ số BAP và tần số cơ bản F0 trên thang đo log (log F0).

3.1.4 Mô hình dự đoán

Mô hình dự đoán có nhiệm vụ lấy đầu vào là các đặc trưng ngôn ngữ học được trích chọn ở phần *** 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 [*].

a) Mô hình thời gian (Duration model)

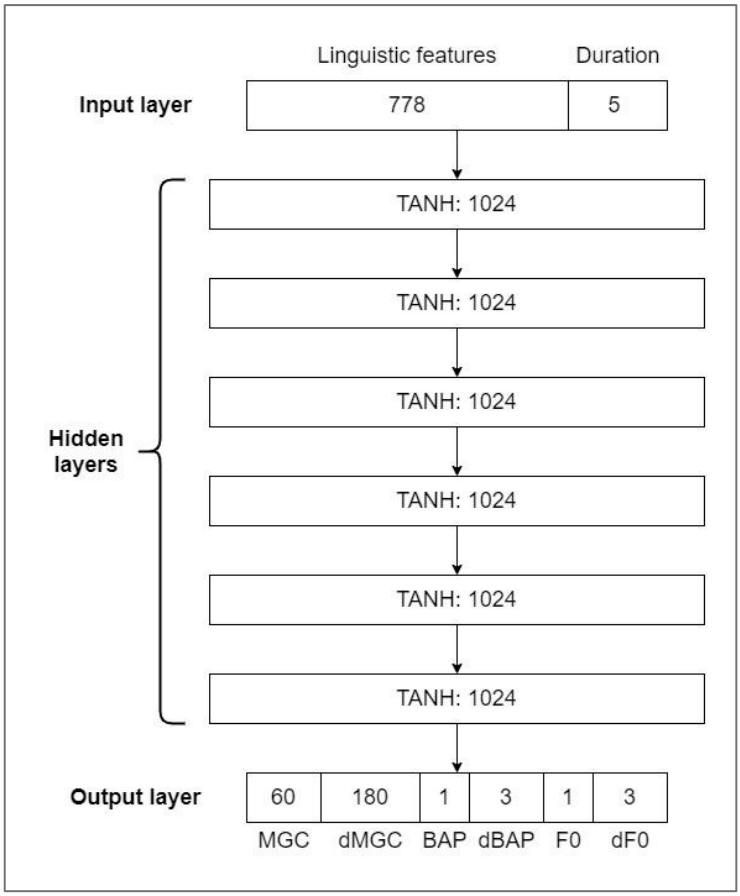
Mô hình khoảng thời gian có nhiệm vụ nhận các đặc trưng ngôn ngữ từ mô đun trích chọn đặc trưng ngôn ngữ và dự đoán thông tin về thời gian xuất hiện của mỗi âm vị. Mô hình này là một mạng nơ ron học sâu với 6 lớp ẩn, mỗi lớp có 1024 nốt và hàm kích hoạt là hàm tanh (**). Đầu vào của mạng là véc tơ 778 chiều chứa đặc trưng ngôn ngữ. Đầu ra của mạng là vec tơ 5 chiều chưa thông tin ước lượng khoảng thời gian xuất hiện của âm vị. Hình 3.4 mô tả chi tiết cấu trúc của mô hình thời gian.



Hình 3.4 Cấu trúc mô hình thời gian (Duration Model)

b) Mô hình âm học (Acoustic model)

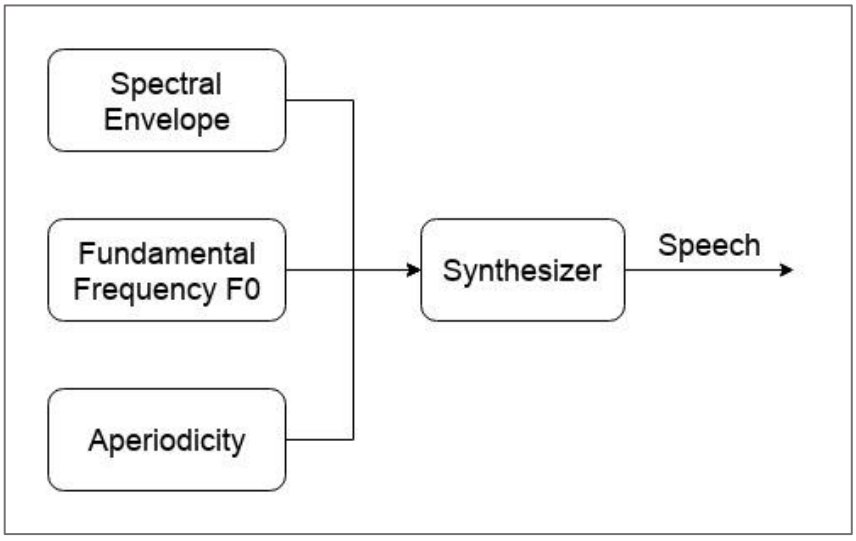
Mô hình âm học có nhiệm vụ lấy đầu vào là các đặc trưng ngôn ngữ cùng thông tin về thời gian xuất hiện của từng âm vị và dự đoán đầu ra là các đặc trưng âm học của tín hiệu tiếng nói. Mô hình này là một mạng nơ ron học sâu với 6 lớp ẩn, mỗi lớp ẩn có 1024 nốt và có hàm kích hoạt là hàm tanh (**). Đầu vào là véc tơ 783 chiều chứa đặc trưng ngôn ngữ và thông tin thời gian. Đầu ra là véc tơ 248 chiều chứa đặc trưng âm học được ước lượng. Hình 3.5 mô tả chi tiết cấu trúc của mô hình âm học.



Hình 3.5 Cấu trúc mô hình âm học (Acoustic model)

3.1.5 Tổng hợp tiếng nói từ đặc trưng âm học

Trong phần này, đồ án sử dụng WORLD vocoder cho nhiệm vụ tổng hợp tiếng nói từ các tham số đặc trưng, chi tiết về bộ vocoder này được trình bày trong phần **.



Hình 3.6 Tổng hợp tiếng nói từ các đặc trưng âm học bằng WORLD vocoder

Trên Hình 3.6 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 (Aperiodicity) là đầu ra của mô hình âm học. Đầu ra của quá trình này chính là tín hiệu tiếng nói.

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

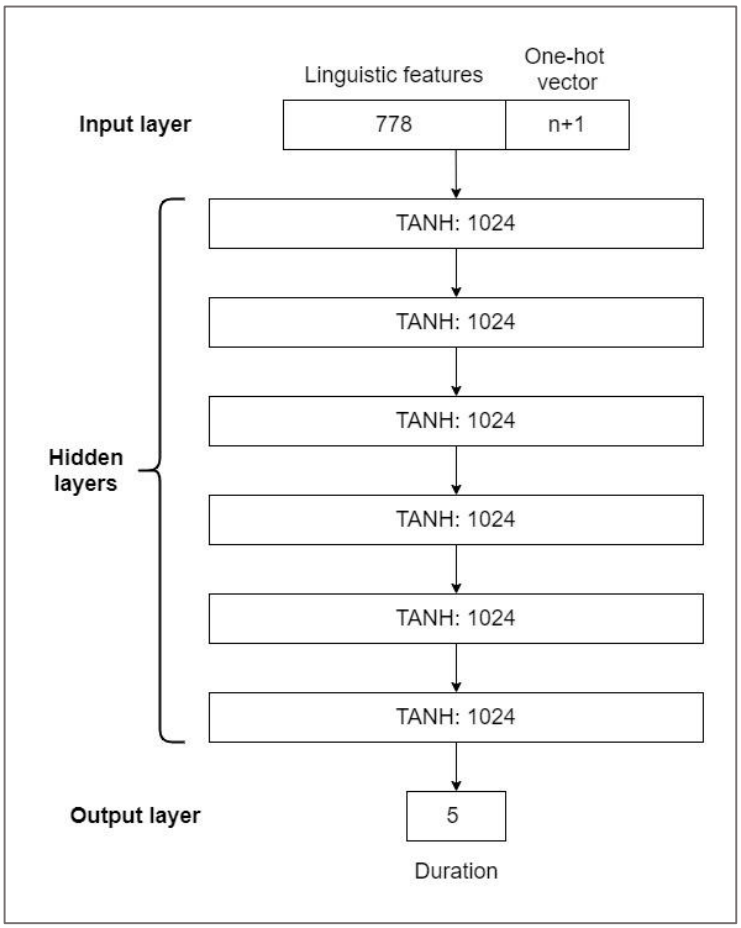
3.2.1 Sử dụng mô hình gốc một người nói

3.2.2 Sử dụng mô hình gốc nhiều người nói

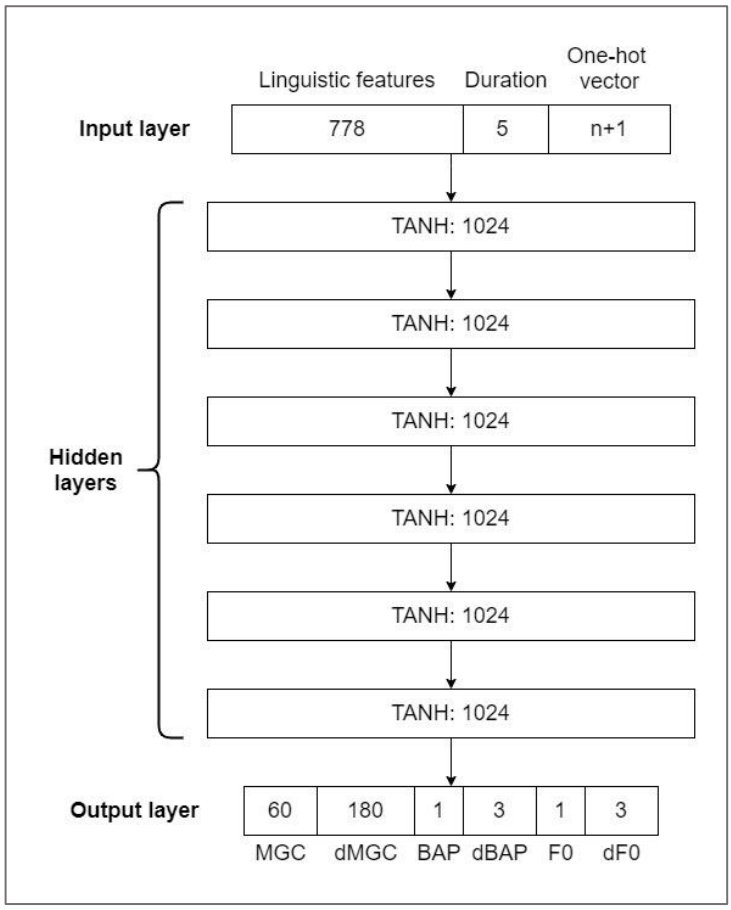
3.3 Sử dụng vec-tơ mã hóa người nói cho thích ứng giọng nói

3.3.1 One-hot encoding

Với phương pháp này, đặc trưng đầu vào cho duration model và acoustic model được gắn thêm vec-tơ định danh. Vec-tơ này có dạng $X = [x_1, x_2, \dots, x_{n+1}]$ độ dài 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 3.7 và Hình 3.8.



Hình 3.7 Cấu trúc mô hình thời gian cho phương pháp one-hot encoding



Hình 3.8 Cấu trúc mô hình âm học cho phương pháp one-hot encoding

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.

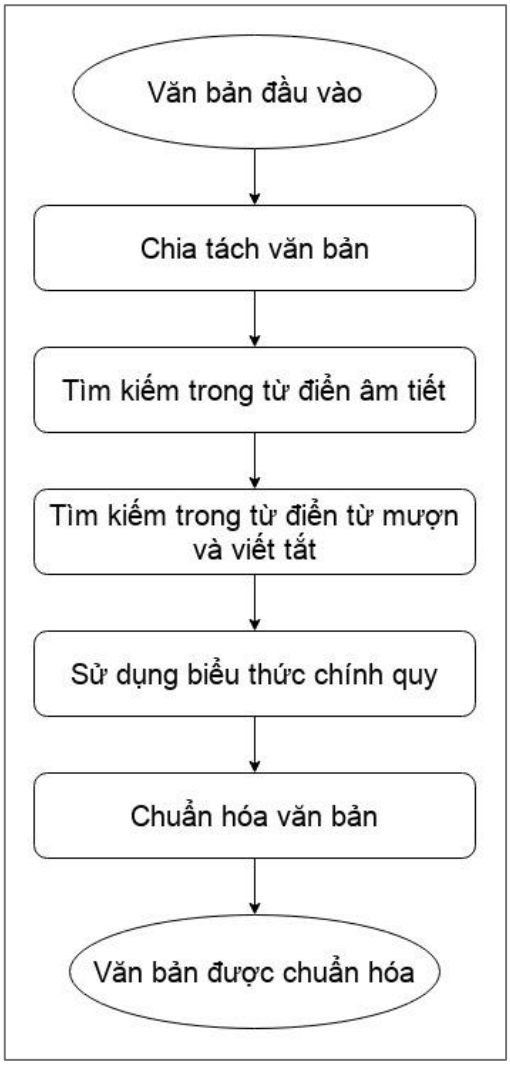
3.3.2 X-vector

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, ... Các bước chuẩn hóa văn bản được thể hiện trong hình **.



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 Phân phối bộ dữ liệu

Để huấn luyện và đánh giá mô hình, đồ án sử dụng 4 bộ dữ liệu được mô tả trong Bảng 4.1~~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
W-01(LTY)	12624	8 giờ 35 phút	Nữ	Miền Nam
W-02 (TTVT)	3716	3 giờ 32 phút	Nữ	Miền Bắc
M-01 (PTQ)	3767	4 giờ 41 phút	Nam	Miền Bắc
VTR-60	15600	20 giờ 39 phút		

Trong đó, bộ dữ liệu VTR-60 bao gồm các mẫu ghi âm của 60 người, được mô tả trong Bảng 4.2~~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

Các mô hình được huấn luyện và thử nghiệm trên hệ thống máy tính có CPU E5-2640 với 32 nhân, tần số 2.6 GHz; RAM 128 Gb; GPU Quadro K22000 với 4 Gb GPU Memory.

Thời gian huấn luyện cho từng mô hình được mô tả ở Bảng 4.3~~Bảng 4.3~~.

Bảng 4.3 Thời gian huấn luyện các mô hình

Mô hình	Bộ dữ liệu	Thời gian huấn luyện
Merlin	F-01(LTY)	
	F-02 (TTVT)	
	M-01 (PTQ)	
	VTR-60	
Transfer	F-02 (TTVT)	
	M-01 (PTQ)	
One hot vector	VTR-60	
	F-02 (TTVT)	
	M-01 (PTQ)	
X-vector	VTR-60	
	F-02 (TTVT)	
	M-01 (PTQ)	

4.3 Đánh giá kết quả

4.3.1 Đánh giá điểm MOS (Mean Opinion Score) của các mô hình

Mô hình	Bộ dữ liệu	MOS
Gốc	F-02	
Gốc	F-M02	
Transfer	F-01 -> F-02	
Transfer	F-01 -> M-01	
Transfer	VTR-60 -> F-02	
Transfer	VTR-60 -> M-01	
One-hot vector	VTR-60 -> F-02	
One-hot vector	VTR-60 -> M-01	
X-vector	VTR-60 -> F-02	
X-vector	VTR-60 -> M-01	

4.3.2 Đánh giá thời gian tổng hợp của các mô hình

Mô hình	Bộ dữ liệu	MOS
Gốc	F-02	
Gốc	F-M02	
Transfer	F-01 -> F-02	
Transfer	F-01 -> M-01	
Transfer	VTR-60 -> F-02	
Transfer	VTR-60 -> M-01	
One-hot vector	VTR-60 -> F-02	
One-hot vector	VTR-60 -> M-01	
X-vector	VTR-60 -> F-02	
X-vector	VTR-60 -> M-01	

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN

5.1 Kết luận

5.2 Phương hướng phát triển và cải thiện đồ án

TÀI LIỆU THAM KHẢO

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

[2] 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.

[3] J. Dang , K. Honda, "Construction and control of a physiological articulatory," *J. Acoust. Soc. Am,* vol. 115, p. 853–870, 2014.

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

[5] 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 Trans. Acoust. Speech, Signal Process.,* vol. 17, pp. 66-83, 2009.

[6] D. K. Ninh, "A Speaker-Adaptive HMM-based Vietnamese Textto-Speech System," *2019 11th International Conference on Knowledge and Systems Engineering (KSE),* pp. 1-5, Otc 2019.

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