

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

LÊ THANH TÙNG

**NGHIÊN CỨU HỆ THỐNG TỔNG HỢP TIẾNG NÓI
THEO PHƯƠNG PHÁP HỌC SÂU**

LUẬN VĂN THẠC SĨ NGÀNH HỆ THỐNG THÔNG TIN

HÀ NỘI - 2020

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

LÊ THANH TÙNG

**NGHIÊN CỨU HỆ THỐNG TỔNG HỢP TIẾNG NÓI
THEO PHƯƠNG PHÁP HỌC SÂU**

Ngành: Hệ Thống Thông Tin

Chuyên ngành: Hệ Thống Thông Tin

Mã số: 60480104

LUẬN VĂN THẠC SĨ NGÀNH HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. NGUYỄN PHƯƠNG THÁI

HÀ NỘI - 2020

LỜI CẢM ƠN

Lời đầu tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới PGS.TS Nguyễn Phương Thái, đã tận tình hướng dẫn và chỉ bảo tôi trong suốt quá trình thực hiện luận văn tốt nghiệp.

Tôi xin chân thành cảm ơn các thầy, cô trong trường đại học Công Nghệ - Đại học Quốc gia Hà Nội đã cho tôi nền tảng kiến thức tốt và tạo mọi điều kiện thuận lợi cho tôi học tập và nghiên cứu.

Tôi cũng xin gửi lời cảm ơn đến TS. Đỗ Văn Hải và các bạn trong nhóm Xử lý Tiếng nói – Trung tâm Không gian Mạng – Tập đoàn Công nghiệp Viễn thông Quân đội đã hỗ trợ tôi rất nhiều về kiến thức chuyên môn trong quá trình thực hiện luận văn.

Cuối cùng, tôi xin được gửi lời cảm ơn vô hạn tới gia đình và bạn bè, những người đã luôn bên cạnh, giúp đỡ và động viên tôi trong quá trình học tập cũng như trong suốt quá trình thực hiện luận văn.

Tôi xin chân thành cảm ơn!

Hà Nội, ngày tháng năm 2020

Học viên

Lê Thanh Tùng

LỜI CAM ĐOAN

Tôi xin cam đoan bài luận văn tìm hiểu về mô hình tổng hợp tiếng nói theo phương pháp học sâu và thực nghiệm được trình bày trong luận văn là do tôi đề ra và thực hiện dưới sự hướng dẫn của PGS.TS Nguyễn Phương Thái.

Tất cả các tài liệu tham khảo từ các nghiên cứu liên quan đều có nguồn gốc rõ ràng từ danh mục tài liệu tham khảo trong luận văn. Trong luận văn, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

Hà Nội, ngày tháng năm 2020

Học viên

Lê Thanh Tùng

MỤC LỤC

LỜI CẢM ƠN.....	1
LỜI CAM ĐOAN	2
MỤC LỤC	3
DANH MỤC HÌNH VẼ	5
DANH MỤC BẢNG BIỂU	6
MỞ ĐẦU	7
CHƯƠNG 1: GIỚI THIỆU VỀ TỔNG HỢP TIẾNG NÓI.....	8
1.1. Tổng quan về tổng hợp tiếng nói	8
1.1.1. Khôi xử lý ngôn ngữ tự nhiên	9
1.1.2. Khôi tổng hợp tín hiệu tiếng nói	10
1.2. Các phương pháp tổng hợp tiếng nói.....	10
1.2.1. Tổng hợp mô phỏng hệ thống phát âm	10
1.2.2. Tổng hợp tần số formant	10
1.2.3. Tổng hợp ghép nối	11
1.2.4. Tổng hợp dùng tham số thống kê.....	12
1.2.5. Tổng hợp tiếng nói bằng phương pháp lai ghép	15
1.2.6. Tổng hợp tiếng nói dựa trên phương pháp học sâu.....	16
1.2.7. Tổng hợp tiếng nói theo phương pháp End-to-End	17
1.2.8. Các phương pháp và độ đo đánh giá hiệu năng hệ thống tổng hợp tiếng nói	18
1.3. Tình hình phát triển hệ thống tổng hợp tiếng nói ở Việt Nam	18
CHƯƠNG 2: MẠNG NƠ RON HỌC SÂU VÀ ĐẶC TRƯNG NGÔN NGỮ	19
TRONG TỔNG HỢP TIẾNG NÓI	19
2.1. Mạng nơ ron học sâu.....	19
2.1.1. Mạng nơ ron thần kinh	19
2.1.2. Mạng nơ ron học sâu	20
2.2. Bài toán học máy	23
2.3.1. Pha huấn luyện	24
2.3.2. Pha kiểm thử.....	24
2.3. Đặc trưng của ngôn ngữ tiếng Việt.....	24
2.3.1. Tổng quan về âm học	24

2.3.2. Các đặc trưng của âm học	25
CHƯƠNG 3: HỆ THỐNG TỔNG HỢP TIẾNG NÓI THEO	29
PHƯƠNG PHÁP HỌC SÂU	29
3.1. Pha huấn luyện	30
3.1.1. Khởi trích chọn đặc trưng ngôn ngữ	30
3.1.2. Mô hình thời gian	32
3.1.3. Mô hình âm học	33
3.1.4. Khởi trích trọn đặc trưng tiếng nói	33
3.2. Pha kiểm thử	36
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ	39
4.1. Thực nghiệm	39
4.1.1. Môi trường thực nghiệm	39
4.1.2. Bộ dữ liệu sử dụng trong thực nghiệm	40
4.1.3. Mô hình huấn luyện	41
4.1.4. Tạo ra tiếng nói tiếng Việt từ mô hình mạng nơ ron học sâu	41
4.2. Đánh giá kết quả	43
4.2.1. Phương pháp đánh giá	43
4.2.2. Bảng so sánh tiếng nói tổng hợp từ 2 mô hình DNN và HMM	43
4.2.3. Kết quả đánh giá	44
CHƯƠNG 5: KẾT LUẬN	45
5.1. Kết quả đạt được của luận văn	45
5.2. Đánh giá hệ thống	45
5.3. Hướng phát triển	45
TÀI LIỆU THAM KHẢO	46

DANH MỤC HÌNH VẼ

Hình 1.1: Tổng quan về hệ thống tổng hợp tiếng nói.....	8
Hình 1.2: Tổng hợp tiếng nói theo phương pháp formant [1]	11
Hình 1.3: Tổng hợp tiếng nói theo phương pháp ghép nối [1].....	11
Hình 1.4: Huấn luyện tiếng nói theo phương pháp tổng hợp tham số [1].....	14
Hình 1.5: Tổng hợp tiếng nói theo phương pháp tham số thống kê [1]	14
Hình 1.6: Mô hình hệ thống tổng hợp tiếng nói theo phương pháp học sâu [3]	16
Hình 1.7: Sơ đồ Encoder và Decoder trong mô hình Seq2Seq	17
Hình 2.1 Mạng nơ ron thần kinh [10].....	19
Hình 2.2 Mạng nơ ron nhân tạo.....	20
Hình 2.3 Mô hình bài toán học máy [10]	23
Hình 2.4 Cụm từ Âm tiết Tiếng Việt [18]	25
Hình 3. 1 Kiến trúc hệ thống tổng hợp tiếng nói theo phương pháp học sâu.....	29
Hình 3. 2 Mô hình trích xuất đặc trưng ngôn ngữ.....	30
Hình 3. 3 Nhãn đặc trưng của ngôn ngữ.....	31
Hình 3. 4 Chuyển đổi nhãn thành véc tơ	32
Hình 3. 5 Mô hình WORLD vocoder [16]	33
Hình 3. 6 Đặc trưng Spectral Envelop của tín hiệu tiếng nói [19]	34
Hình 3. 7 Tần số F0 của tín hiệu tiếng nói [19].....	34
Hình 3. 8 Đặc trưng Aperiodic Energy của tín hiệu tiếng nói [19]	35
Hình 3. 9 Trích xuất đặc trưng âm thanh.....	35
Hình 3. 10 Cấu trúc mạng nơ ron mô hình thời gian.....	36
Hình 3. 11 Cấu trúc mạng nơ ron mô hình âm học Acoustic	38
Hình 3. 12 Tổng hợp tiếng nói từ đặc trưng âm học	38

DANH MỤC BẢNG BIỂU

Bảng 2.1: Các đặc trưng âm học [18].....	25
Bảng 2.2: Nhãn âm vị theo cấu trúc HTS.....	26
Bảng 2.3: Mô tả nhãn âm vị	28
Bảng 4.1 Cấu hình phần cứng máy chủ thử nghiệm	39
Bảng 4.2 Các phần mềm sử dụng trong hệ thống.....	40
Bảng 4.3 Bộ dữ liệu thử nghiệm.....	40
Bảng 4.4 Bảng so sánh tiếng nói tổng hợp	43

MỞ ĐẦU

Tổng hợp tiếng nói từ văn bản là quá trình chuyển đổi tự động một văn bản thành lời nói. Hệ thống được sử dụng cho mục đích này gọi là hệ thống tổng hợp tiếng nói, hệ thống tổng hợp tiếng nói gồm hai thành phần cơ bản: Phần xử lý ngôn ngữ tự nhiên và phần xử lý tổng hợp tiếng nói.

Tổng hợp tiếng nói đã được ứng dụng nhiều trong các lĩnh vực của đời sống như ứng dụng cho người mù, cho người bị điếc hoặc gặp khó khăn trong phát âm, ứng dụng giáo dục, các trung tâm hỗ trợ khách hàng, hệ thống tương tác người máy.

Tổng hợp tiếng nói dựa trên phương pháp học sâu đã bắt đầu phát triển mạnh mẽ trong vài năm trở lại đây, phương pháp được xây dựng dựa trên việc mô hình hóa mô hình âm học bằng một mạng nơ ron học sâu. Văn bản đầu vào được chuyển hóa thành một véc tơ đặc trưng ngôn ngữ, véc tơ mang thông tin về âm vị, ngữ cảnh xung quanh âm vị, thanh điệu. Sau đó mô hình âm học dựa trên mạng nơ ron lấy đầu vào véc tơ đặc trưng ngôn ngữ và tạo ra các đặc trưng âm học tương ứng ở đầu ra. Từ các đặc trưng âm học sẽ tạo thành tín hiệu tiếng nói nhờ một bộ tổng hợp tiếng nói vocoder. Mạng nơ ron học sâu được sử dụng trong các sản phẩm Google, Baidu, Microsoft hay hệ thống Merlin của CSTR đã đạt được độ tự nhiên tiếng nói rất cao.

Cụ thể trong luận văn này, tác giả nghiên cứu hệ thống tổng hợp tiếng nói tiếng Việt theo phương pháp học sâu.

Nội dung luận văn chia làm các chương như sau:

Chương 1: Luận văn giới thiệu tổng quan về tổng hợp tiếng nói, các phương pháp được áp dụng để tổng hợp tiếng nói từ văn bản.

Chương 2: Luận văn giới thiệu mạng nơ ron nhân tạo, đặc trưng ngôn ngữ trong tổng hợp tiếng nói.

Chương 3: Luận văn giới thiệu về hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học.

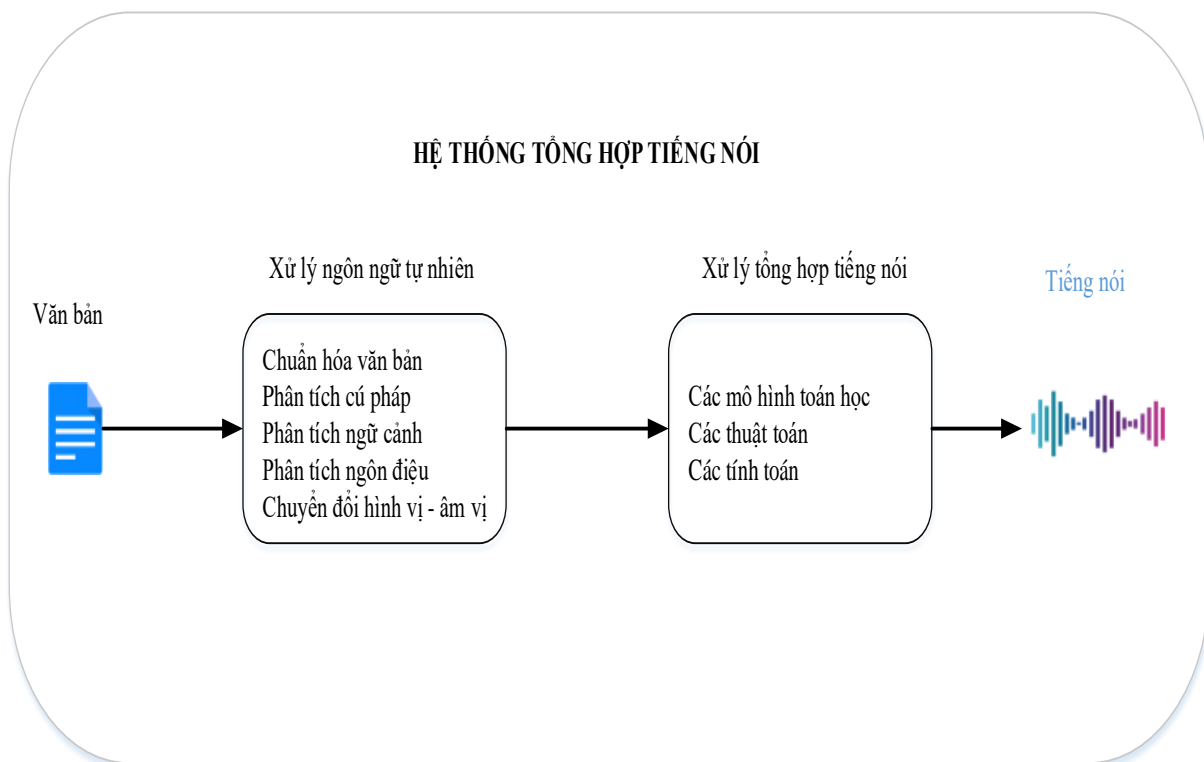
Chương 4: Thực nghiệm và đánh giá hệ thống tổng hợp tiếng nói trên tập dữ liệu tiếng Việt.

Chương 5: Kết luận.

CHƯƠNG 1: GIỚI THIỆU VỀ TỔNG HỢP TIẾNG NÓI

1.1. Tổng quan về tổng hợp tiếng nói

Tổng hợp tiếng nói (Speech Synthesis) là quá trình tạo ra tiếng nói con người một cách nhân tạo. Tổng hợp tiếng nói từ văn bản (Text-To-Speech) là quá trình chuyển đổi tự động một văn bản có nội dung bất kỳ thành lời nói. Hệ thống được sử dụng cho mục đích này gọi là hệ thống tổng hợp tiếng nói. Một hệ thống tổng hợp tiếng nói gồm hai thành phần cơ bản: Phần xử lý ngôn ngữ tự nhiên (Natural Language Processing) và phần xử lý tổng hợp tiếng nói (Speech Synthesis Processing) [1].



Hình 1.1: Tổng quan về hệ thống tổng hợp tiếng nói

Khối xử lý ngôn ngữ tự nhiên có nhiệm vụ chuyển chuỗi các ký tự văn bản đầu vào thành một dạng chuỗi các nhãn ngữ âm đã được thiết kế trước của hệ thống tổng hợp tiếng nói. Tức là thực hiện chuyển đổi văn bản đầu vào thành chuỗi dạng biểu diễn ngữ âm. Từ thông tin ngôn điệu và ngữ âm là chuỗi các nhãn phụ thuộc ngữ cảnh mức âm vị của văn bản đầu vào, khối xử lý tổng hợp tiếng nói chọn ra các tham số thích hợp từ tập các giá trị tần số cơ bản, phổ tín hiệu, trường độ âm thanh (âm vị, âm tiết). Cuối cùng, tiếng nói ở dạng sóng tín hiệu được tạo ra bằng một kỹ thuật tổng hợp.

1.1.1. Khối xử lý ngôn ngữ tự nhiên

Trong hệ thống tổng hợp tiếng nói, khối xử lý ngôn ngữ tự nhiên có nhiệm vụ trích chọn các thông tin về ngữ âm, ngữ điệu của văn bản đầu vào. Thông tin ngữ âm cho biết những âm nào được phát ra trong hoàn cảnh cụ thể nào, thông tin ngữ điệu mô tả điệu tính của các âm được phát. Quá trình xử lý ngôn ngữ tự nhiên gồm có 3 bước:

- Xử lý và chuẩn hóa văn bản (Text processing)
- Phân tích cách phát âm (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 và các ký tự đặc biệt... thành dạng viết đầy đủ và chính xác. Đây là một vấn đề rất khó do có nhiều cách đọc khác nhau phụ thuộc vào từng ngữ cảnh, ví dụ như 2020 có thể đọc là “hai nghìn không trăm hai mươi” hoặc “hai nghìn hai mươi” hoặc “hai không hai không”.

Phân tích cách phát âm là quá trình xác định cách phát âm chính xác cho văn bản, các hệ thống tổng hợp tiếng nói dùng hai cách cơ bản để xác định cách phát âm cho văn bản, quá trình này còn được gọi là chuyển đổi văn bản sang chuỗi âm vị. Cách thứ nhất là dựa vào từ điển, sử dụng một từ điển lớn có chứa tất cả các từ và cách phát âm của chúng. Cách thứ hai là dựa trên các quy tắc và sử dụng các quy tắc để tìm ra cách phát âm tương ứng. Mỗi cách đều có các ưu nhược điểm khác nhau, cách dùng từ điển sẽ nhanh và chính xác tuy nhiên không hoạt động được với các từ chưa có trong từ điển và lượng từ vựng cần lưu trữ là lớn. Cách dùng quy tắc phù hợp hơn với mọi văn bản nhưng độ phức tạp có thể tăng cao nếu ngôn ngữ có nhiều 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 âm phụ thuộc vào ngữ cảnh khi được phát âm trong một ngôn ngữ liên tục, ngoài ra quá trình này còn phải xác định các điểm dừng lấy hơi khi phát âm hoặc đọc một đoạn văn bản [2]. Thông tin về thời gian (duration) được đo bằng đơn vị xen ti giây (centi second) hoặc mi li giây (mili second), 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 F_0 , được biểu thị theo đơn vị Hz hoặc phân số của tông (tones). Tần số cơ bản F_0 là một đặc trưng quan trọng trong việc tạo ngôn điệu của tín hiệu tiếng nói, tạo ra các đặc trưng cao độ là một vấn đề phức tạp và quan trọng trong tổng hợp tiếng nói.

1.1.2. Khối tổng hợp tín hiệu tiếng nói

Khối tổng hợp tiếng nói có chức năng tạo ra tiếng nói từ các thông tin về ngữ âm, ngữ điệu do khối xử lý ngôn ngữ tự nhiên cung cấp. Trong thực tế, có hai cách tiếp cận cơ bản liên quan đến tổng hợp tiếng nói: Tổng hợp tiếng nói sử dụng mô hình nguồn âm và tổng hợp dựa trên việc ghép nối các đơn vị âm.

Chất lượng tiếng nói của hệ thống tổng hợp được đánh giá thông qua hai khía cạnh: Độ dễ hiểu và độ tự nhiên. Độ dễ hiểu đề cập đến nội dung của tiếng nói được tổng hợp có thể hiểu một cách có 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 vừa tự nhiên, vừa dễ hiểu và mục tiêu xây dựng một hệ thống tổng hợp là làm gia tăng tối đa hai yêu cầu này.

1.2. Các phương pháp tổng hợp tiếng nói

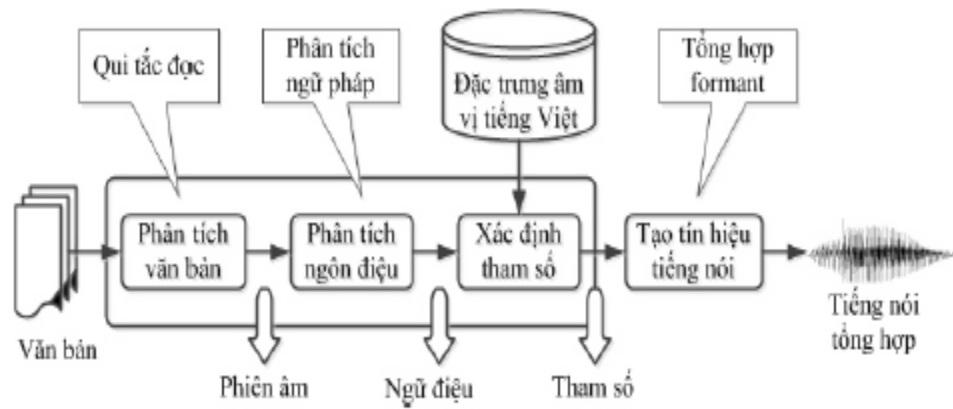
1.2.1. Tổng hợp mô phỏng hệ thống phát âm

Tổng hợp mô phỏng hệ thống phát âm là kỹ thuật tổng hợp giọng nói dựa trên mô hình máy tính mô phỏng cơ quan phát âm của con người. Vì mục tiêu của phương pháp là mô phỏng quá trình tạo ra tiếng nói càng giống cơ chế của con người càng tốt, nên về mặt lý thuyết đây là phương pháp cơ bản nhất để tổng hợp tiếng nói, nhưng phương pháp này khó thực hiện nhất và 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 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. Để thực hiện được phương pháp tổng hợp tiếng nói dựa trên mô phỏng hệ thống phát âm đòi hỏi thời gian, chi phí và công nghệ.

1.2.2. Tổng hợp tần số formant

Tổng hợp tiếng nói formant là phương pháp tổng hợp tiếng nói không sử dụng mẫu giọng thật, thay vào đó tín hiệu tiếng nói được tạo ra bởi một mô hình tuyến âm. Mô hình này mô phỏng hiện tượng cộng hưởng các cơ quan phát âm bằng tập hợp các bộ lọc. Các bộ lọc được gọi là các bộ lọc cộng hưởng formant, có thể kết hợp song song hay nối tiếp nhau hoặc cả hai.

Phương pháp tổng hợp tần số formant không phải sử dụng tiếp mẫu giọng thật khi tổng hợp tiếng nói, tín hiệu âm thanh được tổng hợp dựa trên 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.

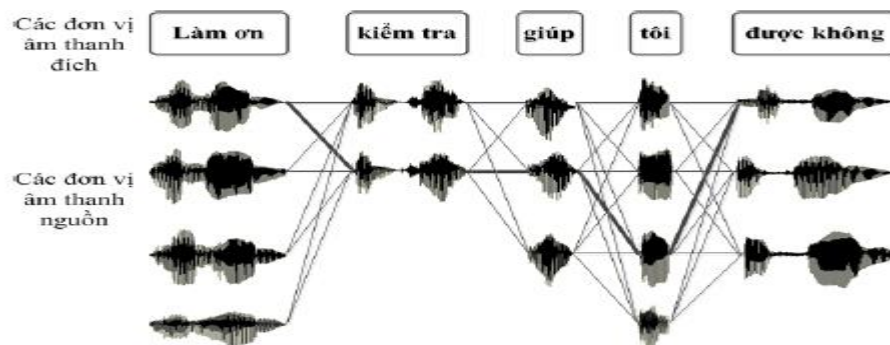


Hình 1.2: Tổng hợp tiếng nói theo phương pháp formant [1]

Hệ thống tổng hợp tiếng nói dựa trên các phương pháp tổng hợp tần số formant có những ưu điểm, nhược điểm như sau: 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, hệ thống có các ưu điểm của riêng như khá dễ nghe và cũng nhỏ gọn vì không chứa cơ sở dữ liệu âm thanh lớn.

1.2.3. Tổng hợp ghép nối

Tổng hợp ghép nối là phương pháp tổng hợp tiếng nói bằng cách ghép các đoạn tín hiệu tiếng nói của một giọng nói đã được ghi âm. Các âm tiết sau khi được tạo thành sẽ được tiếp tục ghép lại với nhau tạo thành tiếng nói. Đơn vị âm phổ biến là âm vị, âm tiết, bán âm tiết, âm đôi, âm ba, từ, cụm từ. 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 tiếng nói với mức độ dễ hiểu và tự nhiên, chất lượng cao. Tuy nhiên, giọng nói tự nhiên được ghi âm có sự thay đổi từ lần phát âm này sang lần phát âm khác và công nghệ tự động hóa việc ghép nối các đoạn của sóng âm thì thoảng tạo ra những tiếng cọt xọt không tự nhiên ở phần ghép nối.



Hình 1.3: Tổng hợp tiếng nói theo phương pháp ghép nối [1]

Có 3 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, âm tiết, 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 âm tiết và các tiếng gần nó. Khi chạy các câu nói được tạo ra bằng 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 cần dùng đến cây quyết định để thực hiện. Thực tế, các hệ thống chọn đơn vị có thể tạo ra được giọng nói rất giống với người thật, tuy nhiên để đạt được độ tự nhiên cao thường cần một cơ sở dữ liệu lớn chứa các đơn vị để lựa chọn.

Tổng hợp âm kép là dùng một cơ sở dữ liệu chứa tất cả các âm kép trong ngôn ngữ. 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 chứa trong cơ sở dữ liệu, khi chạy thì lời văn bản được chèn lên các đơn vị này bằng kỹ thuật xử lý tín hiệu số nhờ mã tuyên đoán tuyến tính hay PSOLA [4]. Chất lượng âm thanh tổng hợp theo cách này thường không cao bằng phương pháp lựa chọn theo đơn vị nhưng tự nhiên hơn cộng hưởng tần số và ưu điểm của nó là kích thước dữ liệu nhỏ.

Tổng hợp chuyên biệt (Domain specific) là phương pháp ghép nối từ các đoạn văn đã được ghi âm để tạo ra lời nói. Phương pháp này thường được dùng cho các ứng dụng có văn bản chuyên biệt, cho một chuyên ngành, sử dụng từ vựng hạn chế như các thông báo chuyến bay hay dự báo thời tiết. Các công nghệ này rất đơn giản và đã được thương mại hóa từ lâu. Mức độ tự nhiên của hệ thống này rất cao vì số lượng câu nói không nhiều, khớp với lời văn, âm điệu của giọng nói ghi âm. Tuy nhiên hệ thống bị hạn chế bởi cơ sở dữ liệu chuyên biệt không áp dụng được cho miền dữ liệu mở.

1.2.4. Tổng hợp dùng tham số thống kê

Một phương pháp tổng hợp tiếng nói được nghiên cứu phổ biến và rộng rãi là phương pháp tổng hợp tiếng nói dựa trên mô hình Markov ẩn HMM [1]. Ở đây HMM là một mô hình thống kê, được sử dụng để mô hình hóa các tham số tiếng nói của đơn vị ngữ âm, trong một ngữ cảnh cụ thể.

Mô hình Markov ẩn là một mô hình học máy dựa trên thống kê, do đó hệ thống tổng hợp tiếng nói dựa trên mô hình Markov ẩn bao gồm 2 quá trình là huấn luyện và tổng hợp.

Trong quá trình huấn luyện, đầu vào là các câu nói được thu âm sẵn và mô tả mức âm vị, tiếp đó các HMM phụ thuộc vào ngữ cảnh của từng âm vị được huấn luyện từ các đặc trưng tham số phổ và tham số nguồn kích thích. Các tham số phổ được mô hình thông qua việc sử dụng các HMM phân bố liên tục, trong khi các tham số kích thích lại được mô hình bằng cách sử dụng các HMM phân bố xác suất đa không gian (Multi-Space probability Distribution HMMs, MSD-HMM) để khắc phục sự đan xen của các âm hữu thanh và vô thanh. Đồng thời các mật độ thời gian trạng thái cũng được mô hình bởi các phân bố Gaussian đơn.

1.2.4.1. Pha huấn luyện

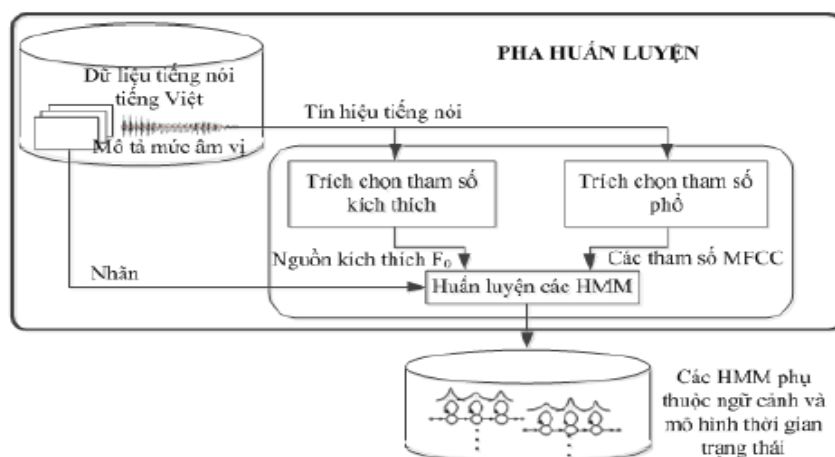
Trong pha huấn luyện, dữ liệu đầu vào gồm tiếng nói và văn bản mô tả tiếng nói. Tín hiệu tiếng nói được sử dụng để lấy ra 2 trích chọn đặc trưng là: Tham số nguồn kích thích (F0) và tham số phổ (MFCC). Văn bản mô tả tiếng nói được sử dụng để trích chọn ra các đặc trưng ngôn ngữ.

Tham số nguồn kích thích được tính toán bằng cách lấy logarit tần số cơ bản F0 và các giá trị delta và delta-delta của nó. Chuỗi các tham số log F0 của các vùng âm vô thanh được mô hình bởi HMM dựa trên xác suất phân bố đa không gian [6].

Tham số phổ tín hiệu MFCC là đặc trưng thanh điệu của tiếng nói, và thời gian trạng thái và các hệ số delta và delta-delta tương ứng của chúng. Các hệ số delta và delta-delta tương ứng với các tham số thanh điệu, thời gian trạng thái được tính toán nhằm phản ánh sự biến thiên của tiếng nói theo thời gian. Phổ tín hiệu MFCC được mô hình hóa thành chuỗi các véc tơ MFCC, và được mô hình bởi các HMM mật độ liên tục. Kỹ thuật phân tích cho phép tổng hợp tiếng nói từ các MFCC nhờ sử dụng bộ lọc Mel Log Spectral Approximation [8]. Các MFCC được trích chọn thông qua phân tích Mel-cepstral bậc 24, sử dụng cửa sổ Hamming 40 ms, độ dịch khung là 8 ms. Các xác suất đầu ra của các MFCC tương ứng với các phân bố Gauss đa biến [7].

Mật độ thời gian trạng thái được mô hình thông qua phân bố Gauss đơn. Chiều của các mật độ này chính là số trạng thái của HMM, chiều thứ n của mật độ trạng thái tương ứng với trạng thái thứ n của HMM. Cấu trúc các HMM bao gồm các trạng thái từ trái qua phải, không bỏ qua trạng thái.

Văn bản mô tả tiếng nói được trích chọn thành các đặc trưng ngôn ngữ theo cấu trúc của bộ nhãn HTS [14], mỗi HMM tương ứng với một âm vị trong bộ nhãn HTS. Một âm vị có các yếu tố phụ thuộc ngữ cảnh như trọng âm, phương ngữ và thanh điệu. Các yếu tố này có ảnh hưởng đến phổ, cao độ và thời gian trạng thái.

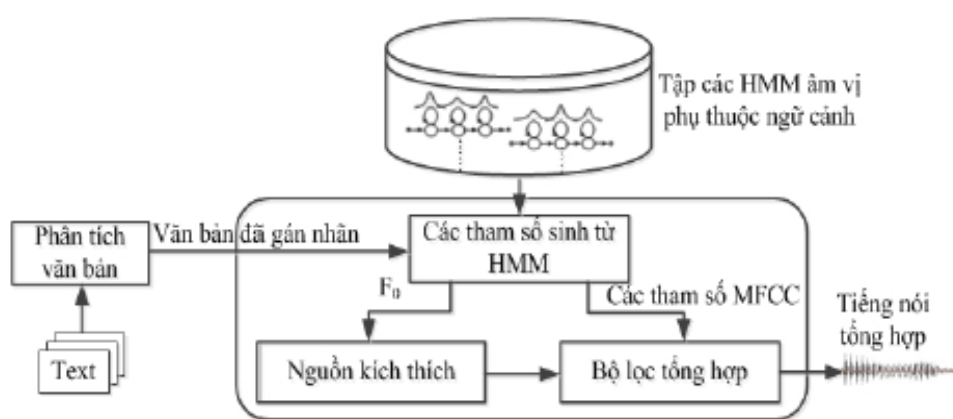


Hình 1.4: Huấn luyện tiếng nói theo phương pháp tổng hợp tham số [1]

Bộ nhận đặc trưng ngôn ngữ và các đặc trưng âm học như tham số kích thích và tham số phổ được huấn luyện để tạo ra các HMM phụ thuộc vào ngữ cảnh và mô hình thời gian trạng thái của các HMM.

1.2.4.2. Pha tổng hợp

Trong pha tổng hợp, các tham số tiếng nói sinh ra từ tập các HMM phụ thuộc ngữ cảnh theo thứ tự chuỗi nhãn ngữ cảnh tương ứng với phát âm của văn bản cần tổng hợp. Các tham số kích thích và MFCC sinh ra được sử dụng để tạo tín hiệu tiếng nói dạng sóng thông qua bộ lọc tổng hợp. Ưu điểm của phương pháp này là trích rút được các đặc trưng âm thanh của các phát âm phụ thuộc ngữ cảnh trong kho dữ liệu tiếng nói. Các đặc tính tiếng nói có thể dễ dàng thay đổi bằng cách điều chỉnh tham số HMM.



Hình 1.5: Tổng hợp tiếng nói theo phương pháp tham số thống kê [1]

Văn bản được chuyển thành chuỗi các nhãn âm vị HTS phụ thuộc vào ngữ cảnh. Dựa vào chuỗi âm vị, tập hợp các HMM mức âm vị được lấy ra và ghép nối thành chuỗi âm vị

tương ứng. Sau đó, độ dài của mỗi trạng thái trong tập các HMM mức câu được tính toán để tối đa hóa xác suất độ dài trạng thái của chuỗi các trạng thái. Tùy thuộc vào thời gian trạng thái mà chuỗi các MFCC và giá trị tham số kích thích được tạo ra từ HMM mức câu bằng cách sử dụng thuật toán sinh tham số tiếng nói. Cuối cùng, tiếng nói được tổng hợp trực tiếp từ các MFCC và các giá trị tham số kích thích thông qua bộ lọc MSLA [8].

Hệ thống tổng hợp tiếng nói dựa trên mô hình Markov ẩn là một hệ thống có khả năng tạo tiếng nói theo phong cách khác nhau, với đặc trưng của nhiều người nói khác nhau. Ưu điểm của phương pháp này là cần ít bộ nhớ lưu trữ và tài nguyên hệ thống thấp hơn nhiều so với tổng hợp ghép nối, có thể điều chỉnh tham số để thay đổi ngữ điệu. Tuy nhiên một số nhược điểm của hệ thống đó là độ tự nhiên trong tiếng nói bị suy giảm so với tổng hợp ghép nối, phổ tín hiệu và tần số cơ bản được ước lượng từ các giá trị trung bình của mô hình Markov ẩn được huấn luyện từ dữ liệu khác nhau, điều này khiến cho tiếng nói tổng hợp nghe có vẻ đều đều mịn và đôi khi trở thành bị nghẹt mũi.

1.2.5. Tổng hợp tiếng nói bằng phương pháp lai ghép

Tổng hợp lai ghép là phương pháp tổng hợp bằng cách lai ghép giữa tổng hợp ghép nối chọn đơn vị và tổng hợp dựa trên mô hình Markov ẩn, nhằm tận dụng ưu điểm của mỗi phương pháp và áp dụng trong hệ thống. Hệ thống tổng hợp lai ghép kết hợp ưu nhược điểm của từng hệ thống thành phần, tùy theo thành phần nào đóng vai trò chủ đạo mà có thể phân loại thành 2 loại như sau: Tổng hợp hướng ghép nối và tổng hợp hướng HMM.

Hệ thống tổng hợp hướng ghép nối sử dụng các HMM để hỗ trợ quá trình ghép nối, ý tưởng chính của phương pháp này như sau:

- Đơn vị dùng để lựa chọn trong “tổng hợp ghép nối chọn đơn vị” cũng sẽ là đơn vị được tổng hợp ra.
- Đường biên giữa các đơn vị sẽ được làm mịn bằng mô hình Markov ẩn.
- Âm thanh sau cùng được làm mịn bằng phương pháp làm mịn phổ.

Khác với hệ thống tổng hợp hướng ghép nối, hệ thống tổng hợp hướng HMM sử dụng các thuật toán sinh tham số từ các HMM và phần tổng hợp ghép nối được sử dụng để tăng cường chất lượng chuỗi tham số này bằng cách bổ sung vào nguồn dữ liệu tiếng nói thêm các tiếng nói mới hình thành do được ghép nối.

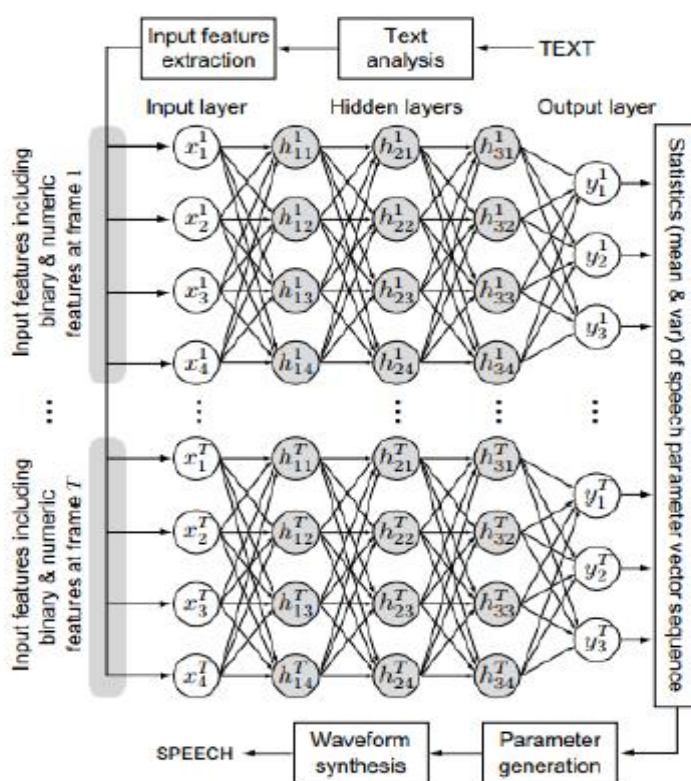
Hai hướng tổng hợp lai ghép nêu trên đều có ưu nhược điểm khác nhau, được sử dụng vào yêu cầu chất lượng tiếng nói hay yêu cầu cụ thể về hệ thống. Ưu điểm cơ bản của hệ thống lai ghép hướng ghép nối đó là giảm tác động không mong muốn do dữ liệu không đủ và giảm sự phụ thuộc vào dữ liệu, hay cũng chính là cải thiện các nhược điểm của tổng hợp

ghép nối. Mặc dù đã giải quyết cơ bản những vấn đề về ghép nối nhưng vấn đề trở ngại tại những điểm ghép nối vẫn tồn tại.

1.2.6. Tổng hợp tiếng nói dựa trên phương pháp học sâu

Tổng hợp tiếng nói dựa trên phương pháp học sâu đã bắt đầu phát triển mạnh mẽ trong vài năm trở lại đây, phương pháp được xây dựng dựa trên việc mô hình hóa mô hình âm học bằng một mạng nơ ron học sâu. Trong đó, văn bản đầu vào được chuyển hóa thành một véc tơ đặc trưng ngôn ngữ, các véc tơ đặc trưng này mang thông tin về âm vị, ngữ cảnh xung quanh âm vị, thanh điệu... Sau đó, mô hình âm học dựa trên mạng nơ ron học sâu lấy đầu vào là véc tơ đặc trưng ngôn ngữ và tạo ra các đặc trưng âm học tương ứng ở đầu ra. Từ các đặc trưng âm học của mô hình âm học sẽ tạo thành tín hiệu tiếng nói nhờ một bộ tổng hợp tín hiệu tiếng nói.

Kiến trúc tổng quan của một hệ thống tổng hợp tiếng nói dựa trên mạng nơ ron học sâu được mô tả như sau:



Hình 1.6: Mô hình hệ thống tổng hợp tiếng nói theo phương pháp học sâu [3]

Văn bản cần được tổng hợp sẽ đi qua bộ phân tích văn bản để trích chọn các đặc trưng ngôn ngữ học và được chuyển hóa thành các véc tơ nhị phân bởi bộ Input feature extraction, các véc tơ nhị phân đầu vào $\{x_n^t\}$ với x_n^t là đặc trưng thứ n tại khung t (frame t), các véc

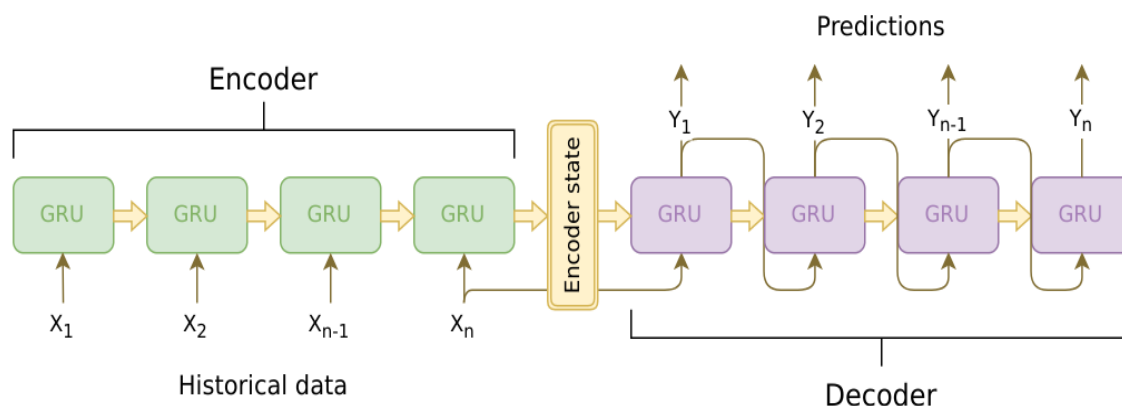
tơ này tương ứng tạo ra $\{y_m^t\}$ các đặc trưng đầu vào thông qua mạng nơ ron DNN đã được huấn luyện, với mỗi y_m^t là đặc trưng đầu ra thứ m tại khung t. Các đặc trưng đầu ra chứa các thông tin về phổ và tín hiệu kích thích, thông qua bộ tạo tham số (Parameter Generation) sẽ được chuyển thành các tham số đặc trưng âm học và được đưa vào bộ tạo tín hiệu tiếng nói để tạo ra tín hiệu tiếng nói.

Mạng nơ ron học sâu dựa trên các lớp nơ ron nhân tạo, có khả năng mô hình hóa những mối quan hệ phi tuyến phức tạp giữa đầu vào và đầu ra. Đặc biệt trong trường hợp sử dụng mạng nơ ron có thể mô hình hóa một cách mạnh mẽ mối quan hệ phi tuyến, phức tạp giữa các đặc trưng ngôn ngữ học của văn bản và đặc trưng âm học của tín hiệu tiếng nói, tuy nhiên việc sử dụng mạng nơ ron cũng có những hạn chế đó là vì sự mạnh mẽ nên rất nhạy cảm với thông tin sai lệch và không tốt như nhiều, và cần nhiều dữ liệu để huấn luyện mô hình.

1.2.7. Tổng hợp tiếng nói theo phương pháp End-to-End

Phương pháp End-to-End được Google đề xuất năm 2017 dựa trên mô hình Seq2Seq được ứng dụng rộng rãi trong dịch máy.

Seq2Seq gồm 2 thành phần là Encoder và Decoder, cả 2 thành phần đều là mạng nơ ron. Encoder có nhiệm vụ chuyển đổi dữ liệu đầu vào (input sequence) thành một biểu diễn đặc trưng ngôn ngữ còn Decoder có nhiệm vụ tạo ra âm thanh đầu ra (output sequence) từ đặc trưng ngôn ngữ được tạo ra ở phần Encoder.



Hình 1.7: Sơ đồ Encoder và Decoder trong mô hình Seq2Seq

Đây là phương pháp tổng hợp tiếng nói tốt nhất hiện nay, tiêu biểu là hệ thống Tacotron [5], tạo ra tiếng nói gần với tiếng nói tự nhiên của con người nhất.

Phương pháp End-to-End ưu điểm là có ít module xử lý do vậy sai lệch giữa kết quả dự đoán và đầu vào là nhỏ, cho ra giọng nói có chất lượng gần với tự nhiên nhất. 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.

1.2.8. Các phương pháp và độ đo đánh giá hiệu năng hệ thống tổng hợp tiếng nói

Hiệu năng của hệ thống tổng hợp tiếng nói được đo bằng phương pháp so sánh tiếng nói tổng hợp với tiếng nói thu âm gốc theo 2 tiêu chí là: Nghe rõ nội dung và tính tự nhiên của giọng nói.

Có 2 phương pháp đánh giá tiêu chí nghe rõ nội dung và tính tự nhiên của giọng nói tổng hợp. Phương pháp thứ nhất là đánh giá khách quan, thực hiện so sánh trực quan trên ảnh phổ và trên đường bao cao độ, sự biến dạng của thang tần số Mel và sai lệch căn bậc hai trung bình phương của $\log F_0$ của tiếng nói tổng hợp và tiếng nói thu âm gốc. Phương pháp thứ hai là đánh giá chủ quan dựa trên tiêu chí điểm đánh giá trung bình MOS (Mean Opinion Score) của người nghe, đánh giá MOS thực hiện hiện bằng cách cho nghe tiếng nói tổng hợp, cho điểm đánh giá theo cảm nhận của người nghe theo 2 tiêu chí đánh giá.

1.3. Tình hình phát triển hệ thống tổng hợp tiếng nói ở Việt Nam

Việt Nam đang đẩy mạnh phát triển công nghệ thông tin trong cuộc cách mạng công nghiệp 4.0. Điều đó cho phép những nền tảng khoa học kỹ thuật và nền tảng cơ sở vật chất được nghiên cứu cũng như triển khai các ứng dụng về khoa học công nghệ trong cuộc sống. Hệ thống tổng hợp tiếng nói tiếng Việt đã có những thành tựu đáng kể và có những sản phẩm tiêu biểu cho các phương pháp tổng hợp tiếng nói.

Phương pháp tổng hợp tiếng nói tần số formant có ứng dụng tiêu biểu là phần mềm đọc văn bản tiếng Việt VnSpeech giới thiệu năm 2009.

Phương pháp tổng hợp tiếng nói ghép nối có ứng dụng tiêu biểu là hệ thống Hoa Súng của Viện nghiên cứu MICA của Đại học Bách khoa Hà nội được giới thiệu năm 2007.

Phương pháp tổng hợp tiếng nói sử dụng tham số thống kê theo mô hình Markov ẩn, ở Việt Nam có nhiều hệ thống phát triển dựa trên phương pháp này như sản phẩm VAIS, sản phẩm của tập đoàn FPT.

Trong thời gian gần đây, trí tuệ nhân tạo được ứng dụng mạnh mẽ vào tổng hợp tiếng nói. Các hệ thống tổng hợp tiếng nói ứng dụng trí tuệ nhân tạo lần lượt ra đời, có thể kể đến: Hệ thống tổng hợp tiếng nói của Viettel, hệ thống tổng hợp tiếng nói của Zalo.

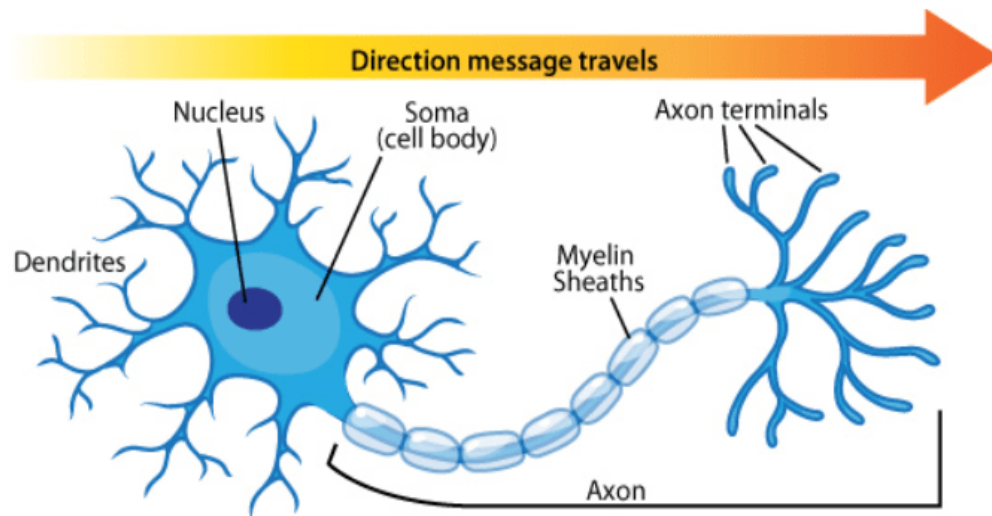
CHƯƠNG 2: MẠNG NƠI HỌC SÂU VÀ ĐẶC TRƯNG NGÔN NGỮ TRONG TỔNG HỢP TIẾNG NÓI

2.1. Mạng nơron học sâu

2.1.1. Mạng nơron thần kinh

Mạng nơron là một hệ thống tính toán lấy cảm hứng từ hoạt động của các nơron trong hệ thần kinh.

Neuron Anatomy



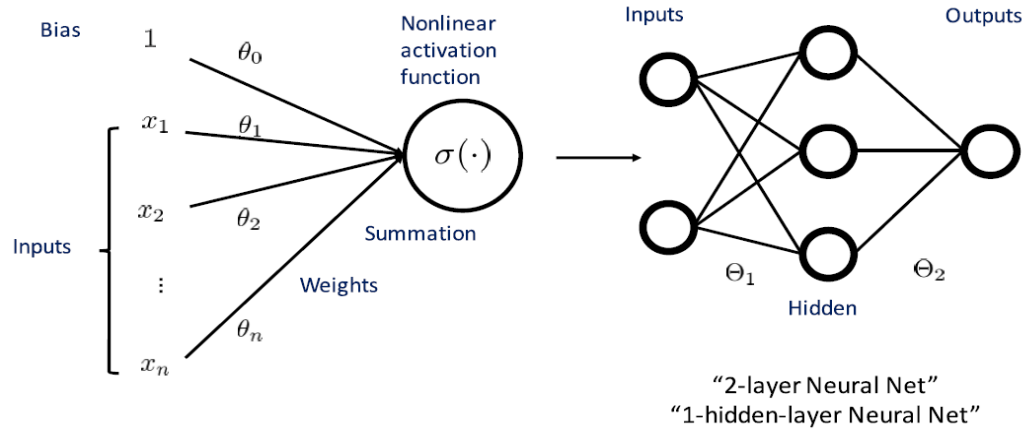
Hình 2.1 Mạng nơron thần kinh [10]

Nơron là đơn vị cơ bản cấu tạo hệ thống thần kinh và là một phần quan trọng của bộ não. Não chúng ta gồm khoảng 10 triệu nơron và mỗi nơron liên kết với 10.000 nơron khác. Ở mỗi nơron có phần thân (soma) chứa nhân, các tín hiệu đầu vào sợi nhánh (dendrites) và các tín hiệu đầu ra qua sợi trục (axon) kết nối với nơron khác. Mỗi nơron nhận xung điện từ các nơron khác qua sợi nhánh. Nếu các xung điện đủ lớn để kích hoạt nơron, thì tín hiệu sẽ đi qua sợi trục để kích hoạt các nơron khác.

Dựa vào cấu tạo và hoạt động của hệ thống nơron con người, có 3 loại mạng nơron nhân tạo thông dụng được sử dụng: Mạng nơron truyền thẳng (feed-forward), mạng nơron hồi quy (Recurrent) và mạng tự tổ chức (self-organizing) [24]. Trong phạm vi luận văn nghiên cứu về phương pháp tổng hợp tiếng nói theo phương pháp học sâu sử dụng mạng nơron truyền thẳng nhiều lớp. Phần tiếp theo xin giới thiệu về cách thức hoạt động của mạng nơron học sâu.

2.1.2. Mạng nơ ron học sâu

Mạng nơ ron học sâu là mạng nơ ron truyền thẳng (Feedforward Neural Network) gồm 1 lớp đầu vào, 1 lớp đầu ra và có nhiều lớp ẩn. Mạng nơ ron học sâu còn được gọi là mạng nơ ron DNN (Deep Neural Network). Mô hình mạng nơ ron truyền thẳng như sau:



Hình 2.2 Mạng nơ ron nhân tạo

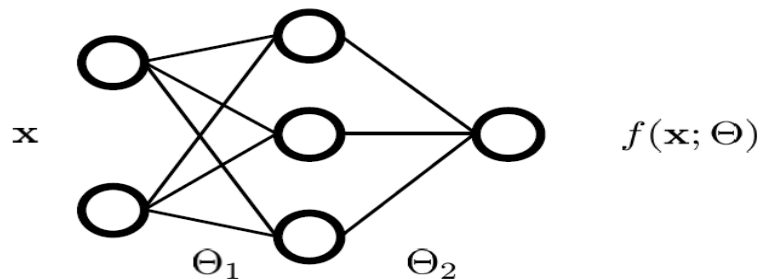
Lớp đầu tiên là input layer, các lớp ở giữa gọi là hidden layer, lớp cuối cùng được gọi là output layer. Tổng số lớp trong mô hình được quy ước là số lớp bớt 1 do không tính input layer.

Mỗi node trong hidden layer và output layer có đặc trưng:

- Liên kết với tất cả các node ở lớp trước đó với các hệ số Θ riêng.
- Mỗi node có hệ số Bias riêng.
- Tính tổng và áp dụng activation function tại node để quyết định xem có gửi đến node kế tiếp không.

Các dữ liệu và hàm được sử dụng trong mạng nơ ron DNN:

- Dữ liệu huấn luyện $\{(x_1, y_1), \dots, (x_n, y_n)\}$; trong đó x_i : dữ liệu đầu vào thứ i ; y_i dữ liệu đích mà DNN muốn sinh ra.
- Hàm số đặc trưng của mạng nơ ron DNN là $f(x; \Theta) \in \mathbb{R}$ với biến số là Θ



Forward Propagation: Thực hiện tính toán giá trị đầu ra \bar{Y} của mạng nơ ron DNN

$$\bar{Y} = \sigma (\Theta_k^T \sigma (\Theta_{k-1}^T \sigma (\dots \sigma (\Theta_1^T))))$$

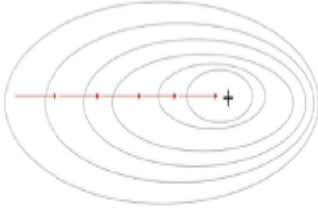
$\sigma (.)$ là hàm activation hay hàm Sigmoid hoặc TANH; k là số lớp của mạng nơ ron DNN

Back Propagation: Tìm tham số Θ để có giá trị \bar{Y} gần với giá trị mong muốn Y nhất.

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^n L(f(x_i; \Theta), y_i) := L(\Theta)$$

$L(.,.)$ là hàm loss function.

Thực hiện phép đạo hàm để cập nhập tham số Θ để tìm ra giá trị $(Y - \bar{Y})$ là nhỏ nhất.

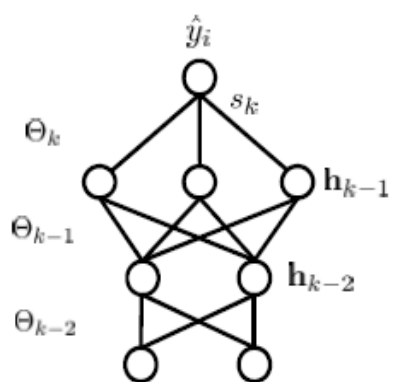
$$\begin{array}{c} \text{parameters} \quad \text{loss function} \\ \Theta^{(t+1)} = \Theta^{(t)} - \underbrace{\gamma}_{\text{learning rate}} \underbrace{\nabla L(\Theta^{(t)})}_{:= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \Theta^{(t)})} \end{array}$$


A contour plot illustrating the optimization process. It shows several concentric elliptical contours representing levels of constant loss. A red line with arrows indicates the path of an optimization algorithm, starting from the outer contours and moving towards the center, which represents the minimum loss. A small '+' mark is at the center of the innermost contour.

Thực hiện quá trình cập nhập tham số Θ :

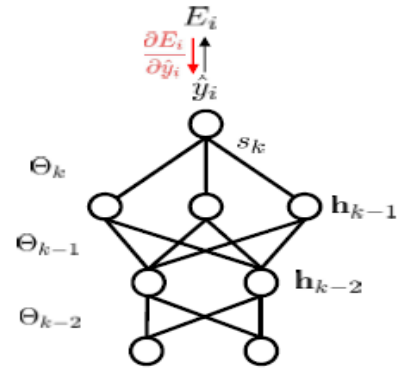
- Cập nhập tham số tại lớp cuối cùng Θ_k
- Tính toán các sai số tại các lớp trước
- Cập nhập tham số tại các lớp trước

Quá trình tính toán [9]:

Dữ liệu đầu vào (x_i, y_i)	 <p>The diagram shows a neural network structure with three layers of nodes. The top layer has one node labeled \hat{y}_i. The middle layer has three nodes, with the rightmost one labeled h_{k-1}. The bottom layer has three nodes, with the rightmost one labeled h_{k-2}. Weights are indicated by labels Θ_k, Θ_{k-1}, and Θ_{k-2} next to the layers. A specific node in the top layer is labeled s_k.</p>
Đưa ra giá trị x_i vào mạng DNN, ta tính được giá trị $\bar{Y} = f(x_i; \Theta)$	
Tại lớp i, giá trị tại các node là: $s_i = \Theta_i^T h_{i-1}$	
Tính toán giá trị sai khác $L(\bar{Y}_i, y_i) = \frac{1}{2} (y_i - \bar{Y}_i)^2 := E_i$	

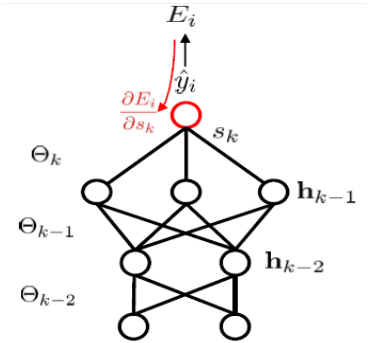
Tính đạo hàm của E_i theo \tilde{Y}_i :

$$\begin{aligned}\frac{dE_i}{d\tilde{Y}_i} &= \frac{d}{d\tilde{Y}_i} \frac{1}{2} (y_i - \tilde{Y}_i)^2 \\ &= - (y_i - \tilde{Y}_i)\end{aligned}$$



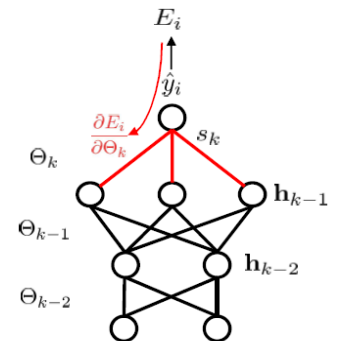
Tính toán đạo hàm E_i theo s_k :

$$\begin{aligned}\frac{dE_i}{ds_k} &= \frac{dE_i}{d\tilde{Y}_i} \frac{d\tilde{Y}_i}{ds_k} \\ &= \frac{dE_i}{d\tilde{Y}_i} \frac{d}{ds_k} \sigma(s_k) \\ &= (\tilde{Y}_i - y_i) \sigma'(s_k)\end{aligned}$$



Tính toán đạo hàm E_i theo Θ_k

$$\begin{aligned}\frac{dE_i}{d\Theta_k} &= \frac{dE_i}{d\tilde{Y}_i} \frac{d\tilde{Y}_i}{sk} \frac{dsk}{d\Theta_k} \\ &= \frac{dE_i}{d\tilde{Y}_i} \frac{d\tilde{Y}_i}{sk} \frac{d}{d\Theta_k} (\Theta_k^T h_{k-1}) \\ &= (\tilde{Y}_i - y_i) \sigma'(s_k) h_{k-1}\end{aligned}$$



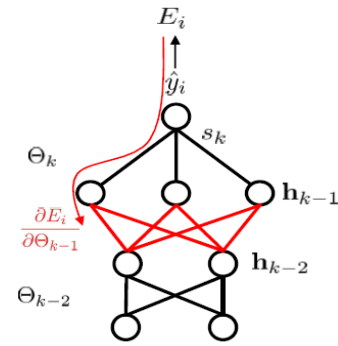
Thực hiện cập nhật tham số:

$$\Theta_k \leftarrow \Theta_k - \gamma \frac{dE_i}{d\Theta_k}$$

(trong đó γ là learning rate)

Tính toán đạo hàm E_i theo Θ_{k-1}

$$\begin{aligned}\frac{dE_i}{d\Theta_{k-1}} &= \frac{dE_i}{d\tilde{y}_i} \frac{d\tilde{y}_i}{s_k} \frac{ds_k}{dh_{k-1}} \frac{dh_{k-1}}{dsk-1} \frac{dsk-1}{d\Theta_{k-1}} \\ &= \frac{dE_i}{d\tilde{y}_i} \frac{d\tilde{y}_i}{s_k} \frac{ds_k}{dh_{k-1}} \frac{dh_{k-1}}{dsk-1} \frac{d}{d\Theta_{k-1}} (\Theta_{k-1}^T h_{k-2})\end{aligned}$$



Thực hiện cập nhật tham số:

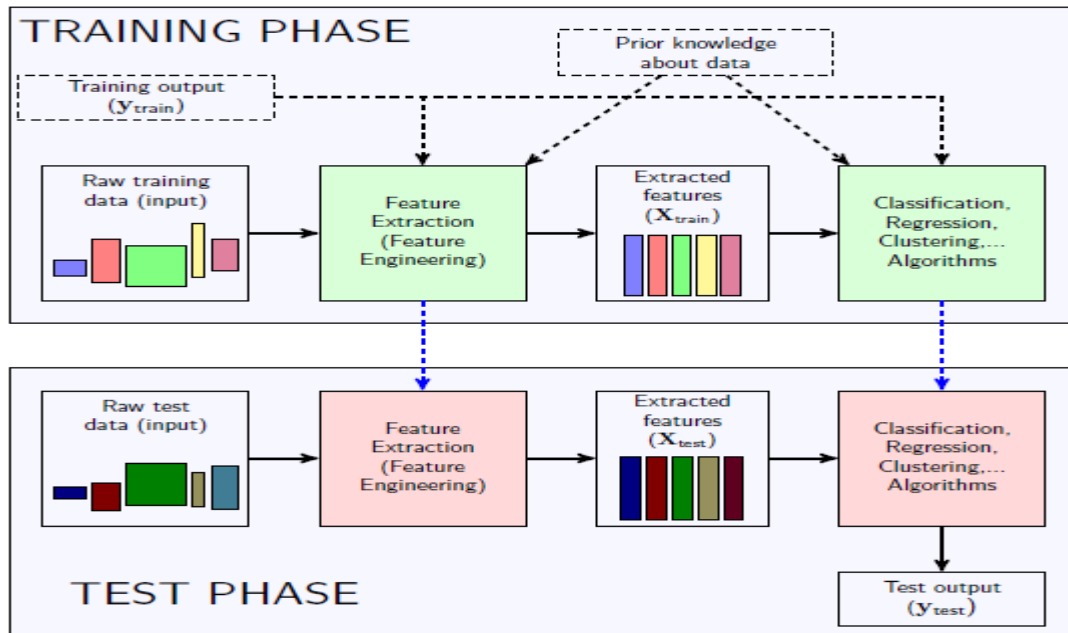
$$\Theta_{k-1} \leftarrow \Theta_{k-1} - \gamma \frac{dE_i}{d\Theta_{k-1}} \text{ (trong đó } \gamma \text{ là learning rate)}$$

Tương tự, thực hiện cập nhật lại tham số của các lớp còn lại.

2.2. Bài toán học máy

Mạng nơ ron học sâu DNN sử dụng bài toán học máy để biểu diễn mối quan hệ giữa cặp dữ liệu đầu vào và đầu ra.

Bài toán học máy chia thành 2 pha cần thực hiện là: Pha huấn luyện (Training Phase) và pha kiểm thử (Test Phase) [10]. Pha huấn luyện dùng dữ liệu trong tập huấn luyện, pha kiểm thử dùng dữ liệu trong tập kiểm thử.



Hình 2.3 Mô hình bài toán học máy [10]

2.3.1. Pha huấn luyện

Dữ liệu thô (raw training data) bao gồm các thông tin về dữ liệu. Ví dụ: Một văn bản là từng từ, từng câu; của một file âm thanh là đoạn tín hiệu và mô tả nội dung file âm thanh...Dữ liệu thô không được biểu diễn ở dạng véc tơ và có số chiều khác nhau.

Dữ liệu đầu ra (training output) trong nhiều trường hợp không được sử dụng, ví dụ như bài toán học không giám sát vì dữ liệu không được gán nhãn. Tuy nhiên trong bài toán tổng hợp tiếng nói thì dữ liệu đầu ra đóng vai trò rất quan trọng, dữ liệu đầu ra được sử dụng để lấy ra các trích trợn đặc trưng âm thanh của ngôn ngữ. Kết quả được xem là ánh xạ từ đặc trưng ngôn ngữ sang đặc trưng âm thanh.

Đặc trưng về dữ liệu (Prior knowledge about data) là các thông tin đã biết về dữ liệu, ví dụ đặc trưng của ngôn ngữ tiếng Việt có bao nhiêu âm vị, âm tiết ...

Trong bước huấn luyện, khối trích chọn đặc trưng (Feature Extraction) thực hiện lấy ra các đặc trưng của dữ liệu, biểu diễn các đặc trưng dữ liệu thành dạng véc tơ, véc tơ đặc trưng có kích thước như nhau bất kể dữ liệu đầu vào có kích thước khác nhau.

Các véc tơ đặc trưng được gọi là đặc trưng của dữ liệu (extracted feature) được đưa vào huấn luyện bằng các thuật toán tạo ra mô hình mạng nơ ron sử dụng trong pha kiểm thử.

2.3.2. Pha kiểm thử

Khi có dữ liệu thô mới, thực hiện trích chọn ra các đặc trưng của dữ liệu thô mới và biểu diễn dưới dạng véc tơ đặc trưng. Véc tơ đặc trưng được đưa vào mô hình mạng nơ ron sinh ra trong pha huấn luyện để tạo ra đầu ra tương ứng của dữ liệu thô mới, so sánh sự sai khác giữa dữ liệu mới sinh ra và dữ liệu thô ban đầu để đánh giá mô hình.

Trong bước kiểm thử ngoài dữ liệu mới sinh ra (y_{test}), còn có kết quả đánh giá hiệu năng của mô hình có hiệu quả hay không có hiệu quả. Phương pháp đánh giá phổ biến nhất là độ chính xác (accuracy), ngoài ra còn sử dụng các phương pháp như ma trận nhầm lẫn (confusion matrix) và True/False Positive/Negative để đo hiệu năng của mô hình.

2.3. Đặc trưng của ngôn ngữ tiếng Việt

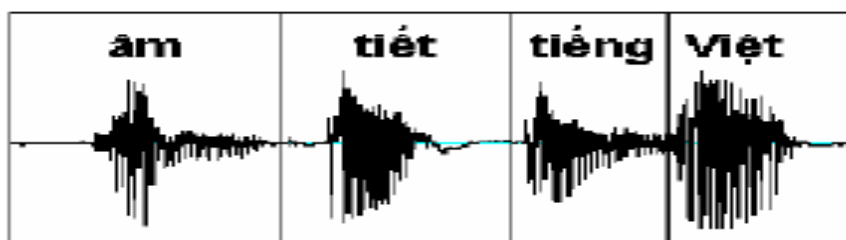
2.3.1. Tổng quan về âm học

Khi giao tiếp con người phát ra những chuỗi âm thanh. Trong chuỗi âm thanh đem chia cắt sẽ được đơn vị cấu thành nhỏ hơn là âm tiết, âm tố và âm vị.

Âm vị là đơn vị tối thiểu của hệ thống ngữ âm dùng để phân biệt âm thanh (là nghĩa của từ) các đơn vị có nghĩa của ngôn ngữ. Âm vị còn có thể được xem là một tổng thể đặc trưng được thể hiện đồng thời.

Âm tiết là đơn vị phân chia tự nhiên trong lời nói, là đơn vị phát âm nhỏ nhất. Mỗi âm tiết là một tiếng, âm tiết được định nghĩa là một đơn vị mà khi phát âm được đặc trưng bởi sự căng lên rồi chùng xuống của cơ trong bộ máy phát âm.

Trong mỗi âm tiết, chỉ có một âm tố có khả năng tạo thành âm tiết gọi âm tố âm tiết tính, còn lại là các yếu tố đi kèm, không tự tạo thành âm tiết được. Âm tố âm tiết tính thường được phân bố ở trung tâm, làm hạt nhân âm tiết. Đó thường là các nguyên âm. Điều này dẫn đến hệ quả là một âm tiết khi được phát ra thì phần năng lượng tập trung nhiều nhất ở phần giữa âm (có biên độ cao), đi về đầu và cuối âm tiết thì năng lượng giảm dần, ví dụ như phổ tần số của cụm từ “âm tiết tiếng Việt” như sau:



Hình 2.4 Cụm từ Âm tiết Tiếng Việt [18]

2.3.2. Các đặc trưng của âm học

Các đặc trưng của âm học và độ đo đặc trưng cho ngôn ngữ như trong bảng sau:

Âm học	Cảm thụ	Ngôn ngữ
Tần số cơ bản (F0)	Cao độ	Thanh điệu, ngữ điệu, độ nhấn
Biên độ, Năng lượng, Cường độ	Độ to nhỏ	Độ nhấn
Trường độ	Độ dài	Độ nhấn
Biên độ động	Độ mạnh	Độ nhấn

Bảng 2.1: Các đặc trưng âm học [18]

Đường nét F0 và cường độ âm thanh có thể tính toán từ tín hiệu lời nói. Độ dài được phỏng đoán bằng cách chia tín hiệu thành các đoạn nhỏ theo định nghĩa về mặt ngữ âm hoặc âm vị.

2.3.3. Phân mức ngôn ngữ

Có nhiều yếu tố ngữ cảnh (âm tố, trọng âm, phương ngữ, thanh điệu) có ảnh hưởng đến phổ, cao độ và thời gian trạng thái.

Các yếu tố ngữ cảnh phụ thuộc ngôn ngữ sử dụng trong tổng hợp tiếng nói chính là các nhân ngữ cảnh và các yếu tố phân cụm ngữ cảnh. Do tiếng Việt là ngôn ngữ có thanh điệu,

nên cần có một tập phát âm phụ thuộc thanh điệu và tập ngữ âm và yếu tố điệu tính tương ứng để xây dựng cây quyết định. Vấn đề phân cụm ngữ cảnh dựa vào cây được thiết kế để có được thanh điệu chính xác là vấn đề rất quan trọng trong bài toán tổng hợp các ngôn ngữ thanh điệu, trong đó có tiếng Việt [12], [13].

2.3.3.1. Mức âm vị

- Âm vị trước, âm vị hiện tại, hai âm vị phía sau.
- Vị trí hiện tại của âm vị trong âm tiết.

2.3.3.2. Mức âm tiết

- Thanh điệu của âm tiết trước, âm tiết hiện tại và âm tiết phía sau.
- Số lượng âm vị của âm tiết trước, âm tiết hiện tại và âm tiết phía sau.
- Vị trí của âm tiết trong từ hiện tại.
- Mức độ trọng âm (thể hiện điệu tính).

2.3.3.3. Mức từ

- Loại từ của từ trước, từ hiện tại và từ phía sau.
- Số lượng âm tiết trong từ trước, từ hiện tại và từ phía sau.
- Vị trí của từ trong cụm từ.
- Số lượng từ trong nhóm từ {trước, sau} tính từ vị trí hiện tại.
- Khoảng cách đến từ trước và từ sau tính từ vị trí hiện tại.

2.3.3.4. Mức cụm từ

- Số lượng âm tiết, từ trong cụm từ trước, cụm từ hiện tại và cụm từ phía sau.
- Vị trí của cụm từ hiện tại trong câu nói.

2.3.3.5. Mức câu nói

- Số lượng âm tiết, từ, cụm từ trong câu nói.

2.3.3.6. Nhãn âm vị

Nhãn âm vị phụ thuộc vào ngữ cảnh được định nghĩa như sau [14]:

$p1^{p2-p3+p4=p5}@p6_p7$

/A: a1_a2_a3

/B: b1-b2-b3@b4-b5& b6-b7#b8-b9\$b10-b11!b12-b13;b14-b15/b16

/C: c1+c2+c3

/D: d1_d2/E:e1+e2@e3+e4&e5+e6#e7+e8/F:f1_f2

/G: g1_g2/H:h1=h2@h3=h4/h5/I:i1_i2

/J: j1+j2-j3

Bảng 2.2: Nhãn âm vị theo cấu trúc HTS

Trong đó:

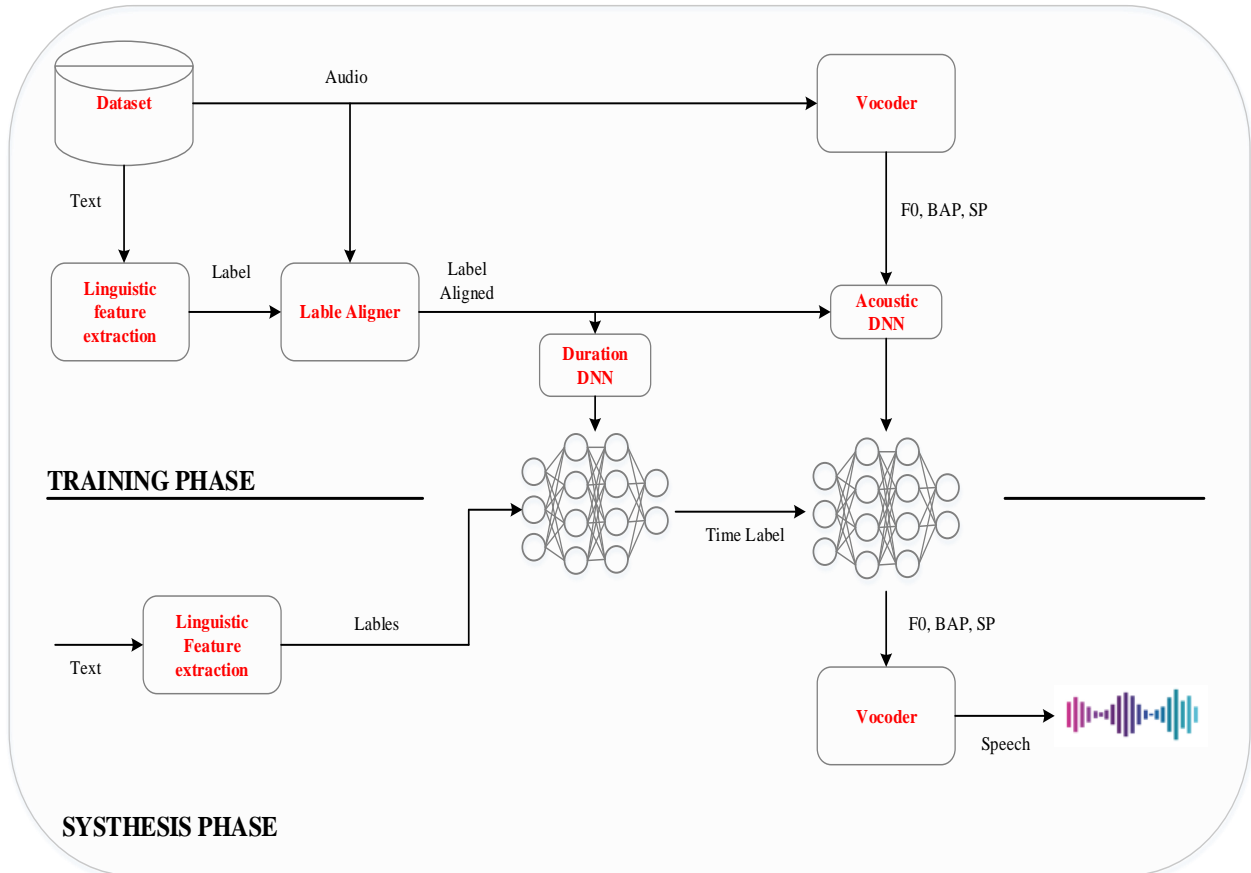
<ul style="list-style-type: none"> - p1: Âm vị đứng trước của âm vị trước đó so với âm vị hiện tại - p2: Âm vị đứng trước âm vị hiện tại - p3: Âm vị hiện tại - p4: Âm vị đứng sau âm vị hiện tại - p5: Âm vị đứng sau của âm vị đứng sau so với âm vị hiện tại - p6: Vị trí của âm vị hiện tại trong âm tiết (từ trái qua phải) - p7: Vị trí của âm vị hiện tại trong âm tiết (từ phải qua trái)
<ul style="list-style-type: none"> - a1: Thanh điệu của âm tiết đứng trước âm tiết hiện tại - a2: Thanh điệu của âm tiết đứng trước âm tiết trước đó so với âm tiết hiện tại - a3: Số âm vị trong âm tiết đứng trước âm tiết hiện tại
<ul style="list-style-type: none"> - b1: Âm tiết hiện tại có được nhấn mạnh hay không (0: Không nhấn mạnh, 1: nhấn mạnh) - b2: Âm tiết hiện tại có là trọng âm hay không (0: Không là trọng âm, 1: trọng âm) - b3: Số âm vị trong âm tiết hiện tại - b4: Vị trí của âm tiết hiện tại trong từ hiện tại (forward) - b5: Vị trí của âm tiết hiện tại trong từ hiện tại (backward) - b6: Vị trí của âm tiết hiện tại trong cụm từ hiện tại (forward) - b7: Vị trí của âm tiết hiện tại trong cụm từ hiện tại (backward) - b8: Số lượng âm tiết được nhấn trọng âm trước âm tiết hiện tại trong cụm từ hiện tại - b9: Số lượng âm tiết được nhấn mạnh sau âm tiết hiện tại trong cụm từ hiện tại - b10: Số lượng âm tiết có trọng âm trước âm tiết hiện tại trong cụm từ hiện tại - b11: Số lượng âm tiết có trọng âm sau âm tiết hiện tại trong cụm từ hiện tại - b12: Khoảng cách mỗi âm tiết từ âm tiết được nhấn trọng âm trước đó đến âm tiết hiện tại - b13: Khoảng cách mỗi âm tiết từ âm tiết hiện tại đến âm tiết có trọng âm tiếp theo - b14: Khoảng cách mỗi âm tiết từ âm tiết có trọng âm trước đến âm tiết hiện tại - b15: Khoảng cách mỗi âm tiết từ âm tiết hiện tại đến âm tiết có trọng âm tiếp theo

- b16: Tên của nguyên âm của âm tiết hiện tại
<ul style="list-style-type: none"> - c1: Âm tiết tiếp theo có được nhấn trọng âm hay không (0: không nhấn trọng âm, 1: nhấn trọng âm) - c2: Âm tiết tiếp theo có trọng âm hay không (0: không có dấu, 1: có dấu) - c3: Số lượng âm vị trong âm tiết tiếp theo
<ul style="list-style-type: none"> - d1: Đoán một phần của từ trước đó - d2: Số lượng âm tiết trong từ trước đó
<ul style="list-style-type: none"> - e1: Đoán từ hiện tại - e2: Số lượng âm tiết trong từ hiện tại - e3: Vị trí của từ hiện tại trong cụm từ hiện tại (forward) - e4: Vị trí của từ hiện tại trong cụm từ hiện tại (backward) - e5: Số lượng từ trước từ hiện tại trong cụm từ hiện tại - e6: Số lượng từ sau từ hiện tại trong cụm từ hiện tại - e7: Khoảng cách mỗi từ, từ từ trước đó đến từ hiện tại - e8: Khoảng cách mỗi từ, từ từ hiện tại đến từ tiếp theo
<ul style="list-style-type: none"> - f1: Đoán từ tiếp theo - f2: Số lượng âm tiết trong từ tiếp theo
<ul style="list-style-type: none"> - g1: Số lượng âm tiết trong cụm từ trước - g2: Số lượng từ trong cụm từ trước
<ul style="list-style-type: none"> - h1: Số lượng âm tiết trong cụm từ hiện tại - h2: Số lượng từ trong cụm từ hiện tại - h3: Vị trí của cụm từ hiện tại trong câu văn (forward) - h4: Vị trí của cụm từ hiện tại trong câu văn (backward) - h5: Âm tiết kết thúc của từ hiện tại
<ul style="list-style-type: none"> - i1: Số lượng âm tiết trong cụm từ tiếp theo - i2: số lượng từ trong cụm từ tiếp theo - j1: số lượng âm tiết trong câu văn - j2: số lượng từ trong câu văn - j3: số lượng các cụm từ trong câu văn

Bảng 2.3: Mô tả nhãn âm vị

CHƯƠNG 3: HỆ THỐNG TỔNG HỢP TIẾNG NÓI THEO PHƯƠNG PHÁP HỌC SÂU

Hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu có kiến trúc gồm pha huấn luyện và pha tổng hợp như sau [15]:



Hình 3.1 Kiến trúc hệ thống tổng hợp tiếng nói theo phương pháp học sâu

Hệ thống bao gồm các khối: Khối trích trồn đặc trưng ngôn ngữ (Linguistic Features Extraction), Khối gán thời gian cho đặc trưng ngôn ngữ (Label Aligner), Mô hình thời gian (Duration Model), Mô hình âm học (Acoustic Model), Khối trích trồn đặc trưng âm học/ Tạo tiếng nói (Vocoder).

Tổng hợp tiếng nói được tách thành 2 pha riêng biệt: Pha huấn luyện và pha tổng hợp. Pha huấn luyện lấy các trích trồn các đặc trưng ngôn ngữ từ văn bản và âm học từ dữ liệu mẫu, thực hiện ghép đôi cho các trích chọn đặc trưng ngôn ngữ và đặc trưng âm học thành cặp dữ liệu (trong đó văn bản là đầu vào, tiếng nói là kết quả đầu ra). Trong pha tổng hợp, khi có văn bản mới cần tạo ra tiếng nói thì hệ thống sẽ lấy ra các đặc trưng âm học

tương ứng với đặc trưng ngôn ngữ của văn bản mới, sau đó đặc trưng âm học được đưa vào bộ tổng hợp để tạo ra tiếng nói.

3.1. Pha huấn luyện

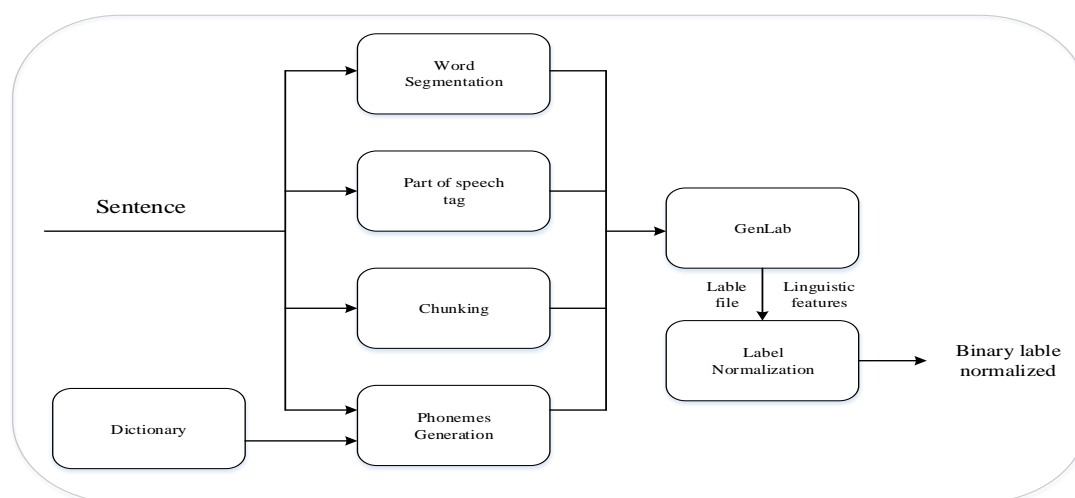
3.1.1. Khối trích chọn đặc trưng ngôn ngữ

Trước khi đưa vào hệ thống tổng hợp tiếng nói, văn bản được chuẩn hóa để có thể đọc một cách rõ ràng, chuẩn hóa các từ mượn, từ viết tắt, số và ngày tháng bằng bộ từ điển tiếng Việt. Các dấu câu như chấm, phẩy, chấm phẩy được tách ra thành câu độc lập.

Khối trích chọn đặc trưng ngôn ngữ thực hiện trích xuất các đặc trưng ngôn ngữ của văn bản đã được chuẩn hóa. Chuỗi các từ đơn của văn bản được chuyển đổi thành chuỗi âm vị, các âm vị sau đó được nhóm lại với nhau theo các từ đơn hoặc các cụm từ có nghĩa. Các thông tin đặc trưng ngôn ngữ bao gồm: part-of-speech tag (loại từ), word segmentation (tách từ), text chunking (cụm từ), âm vị và âm tiết, nhãn thời gian của âm vị. Đặc trưng ngôn ngữ bao gồm các thông tin như sau:

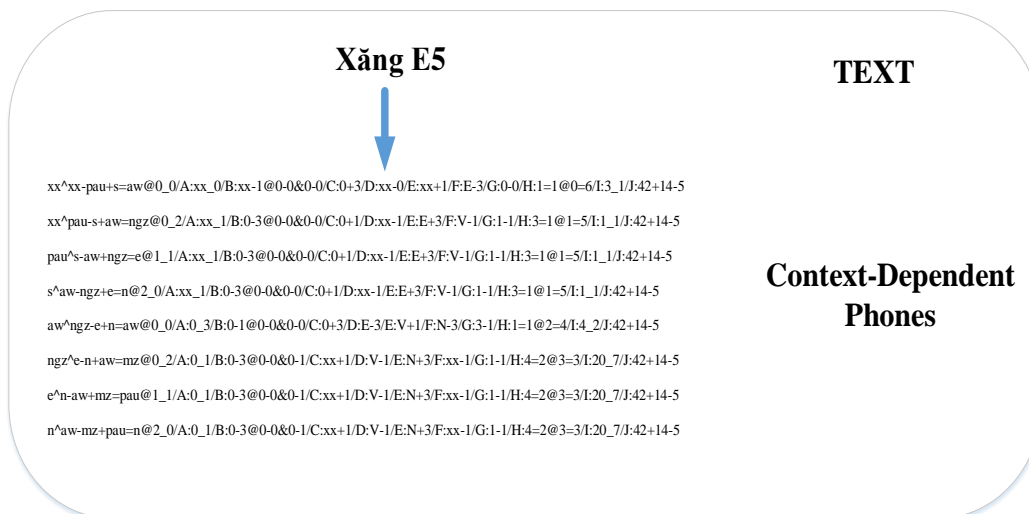
- Mức âm vị: Âm vị hiện tại, âm vị trước, âm vị sau và vị trí của âm vị trong âm tiết.
- Mức âm tiết: Tổng số âm vị, giọng điệu của âm tiết, âm tiết trước và âm tiết sau, vị trí của âm tiết trong từ và trong câu.
- Mức từ: Loại của từ, từ trước và sau, tổng số âm vị, âm tiết hiện tại và âm tiết trước và sau của từ tiếp theo.
- Nhóm từ: Tổng số từ và âm vị trong nhóm từ hiện tại, nhóm từ trước và từ sau đó.
- Mức câu: Tổng số từ, tổng số âm vị và tổng số nhóm từ.

Mô hình trích xuất đặc trưng ngôn ngữ như sau:



Hình 3.2 Mô hình trích xuất đặc trưng ngôn ngữ

Nhãn của đặc trưng ngôn ngữ được sinh ra có định dạng như sau:



Hình 3.3 Nhãn đặc trưng của ngôn ngữ

Nhãn của âm vị được đưa vào khối Label Aligner (là ứng dụng HTK - Hidden Markov Toolkit) thực hiện gán thời gian bắt đầu và kết thúc. Theo thiết kế của HTK, âm vị được đưa vào HTK được chia thành 5 trạng thái theo thời gian. Mỗi trạng thái của âm vị được gán thời gian bắt đầu, kết thúc dựa vào thông tin thời gian được trích xuất từ dữ liệu âm thanh nhờ khối trích chọn đặc trưng âm học.

Hệ thống tổng hợp tiếng nói theo phương pháp học sâu cho ngôn ngữ tiếng Việt sử dụng phần mềm mã nguồn mở Vita_ana để trích xuất các đặc trưng ngôn ngữ của văn bản.

Kết quả tạo ra nhãn gồm 5 trạng thái của âm vị được gán nhãn thời gian bắt đầu và kết thúc như sau:

```
0 50000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[2]
50000 350000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[3]
350000 6100000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[4]
6100000 6150000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[5]
6150000 6250000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[6]
6250000 6600000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[2]
6600000 7200000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[3]
7200000 7400000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[4]
7400000 7600000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[5]
7600000 7750000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[6]
```

Trong pha huấn luyện, đặc trưng ngôn ngữ từ khối Label Aligner được gắn thông tin thời gian xuất hiện và thời gian kết thúc. Các đặc trưng ngôn ngữ được biểu diễn thành véc tơ nhị phân dựa vào bộ các câu hỏi được thiết kế riêng cho tiếng Việt. Các câu hỏi được dùng để khai phá thông tin mà đặc trưng ngôn ngữ mang lại, có thể là: “âm vị hiện tại là gì”, “âm vị phía trước là gì”, “âm vị phía sau là gì”, “có bao nhiêu âm vị trong từ”, “có bao nhiêu âm vị trong câu”... Bằng cách trả lời các câu hỏi, ta có được véc tơ nhị phân biểu diễn đặc trưng của ngôn ngữ. Cách áp dụng câu hỏi để chuyển hóa các thông tin đặc trưng ngôn ngữ thành véc tơ nhị phân theo quy tắc như sau:

- YES**

```

"Please call . . ."
#~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1&1-4# . . .
3900000 4000000 #~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1
4000000 4050000 #~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1
4050000 4200000 #~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1
4200000 4250000 #~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1
4250000 4400000 #~p-l+i=z:2_3/A/0_0_0/B/1-1-4:1-1
p~l-i+z=k:3_2/A/0_0_0/B/1-1-4:1-1&1-4# . . .
l~i-z+k=0:4_1/A/0_0_0/B/1-1-4:1-1&1-4# . . .
i~z-k+0=lw:1 3/A/1 1 4/B/1-1-3:1-1&2-3# . . .
z~k-0.

00000000000000000000000000000001
    
```

Các véc tơ đặc trưng ngôn ngữ được đưa vào huấn luyện để tạo ra mô hình thời gian, là mạng nơ ron học sâu DNN gồm có 6 lớp ẩn. Mô hình thời gian được sử dụng để gán thời gian bắt đầu và thời gian kết thúc cho các trạng thái của âm vị trong pha tổng hợp tiếng nói.

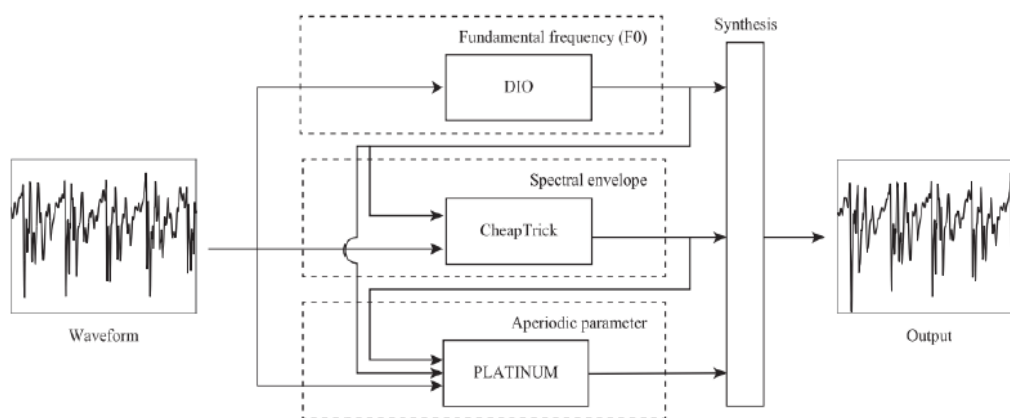
3.1.3. Mô hình âm học

Trong pha huấn luyện với mô hình âm học, dữ liệu đầu vào là các véc tơ đặc trưng ngôn ngữ và véc tơ đặc trưng âm học. Các véc tơ đặc trưng ngôn ngữ và đặc trưng âm học theo từng cặp được huấn luyện, tạo ra mạng nơ ron học sâu cho mô hình âm học. Do đầu ra của mô hình âm học là các đặc trưng âm học cho khung tín hiệu có độ dài 5ms, nên đầu vào cũng phải là các đặc trưng ngôn ngữ có theo từng khung 5ms (frame). Từ thông tin về thời gian xuất hiện của âm vị, đặc trưng ngôn ngữ được chia thành các khung và được gắn thêm các thông tin về khung: Vị trí của khung trong trạng thái (tính từ trạng thái đầu), vị trí của khung trong trạng thái (tính từ trạng thái cuối), số khung của vị trí hiện tại, 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), vị trí của trạng thái trong âm vị (tính từ cuối âm vị).

Chức năng của mạng nơ ron học sâu cho mô hình âm học là dự đoán các trung âm thanh từ đặc trưng ngôn ngữ.

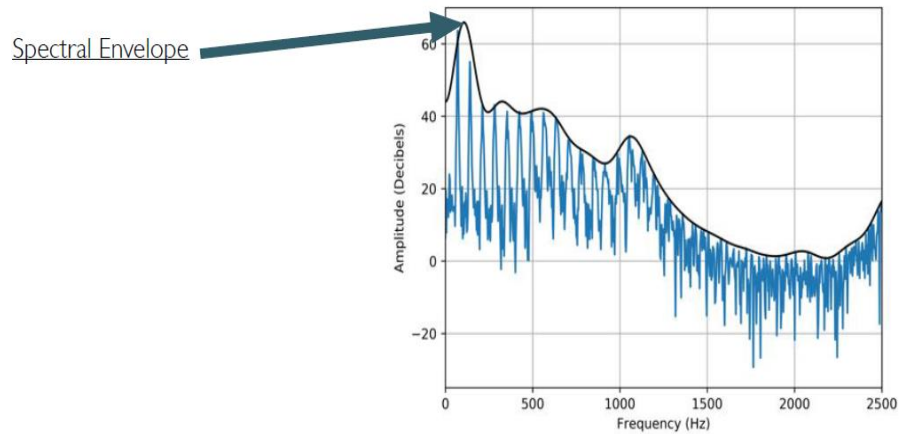
3.1.4. Khối trích chọn đặc trưng tiếng nói

Khối trích chọn đặc trưng tiếng nói (vocoder) có chức năng phân tích âm thanh thành các đặc trưng âm học, được sử dụng để huấn luyện mô hình mạng nơ ron âm học. Hệ thống tổng hợp tiếng nói tiếng Việt sử dụng phần mềm WORLD vocoder [16]. Các đặc trưng tiếng nói mà WORLD vocoder trích chọn được bao gồm: Đường bao phổ tín hiệu (Spectral Envelope), tín hiệu kích thích không tuần hoàn (Aperiodic Energy) và tuần số cơ bản F0.



Hình 3.5 Mô hình WORLD vocoder [16]

Đường bao phổ tín hiệu là đặc trưng cho độ to của giọng nói, được ước lượng bằng công cụ CheapTrick [20]. Đường bao phổ tín hiệu được mô tả như hình sau:

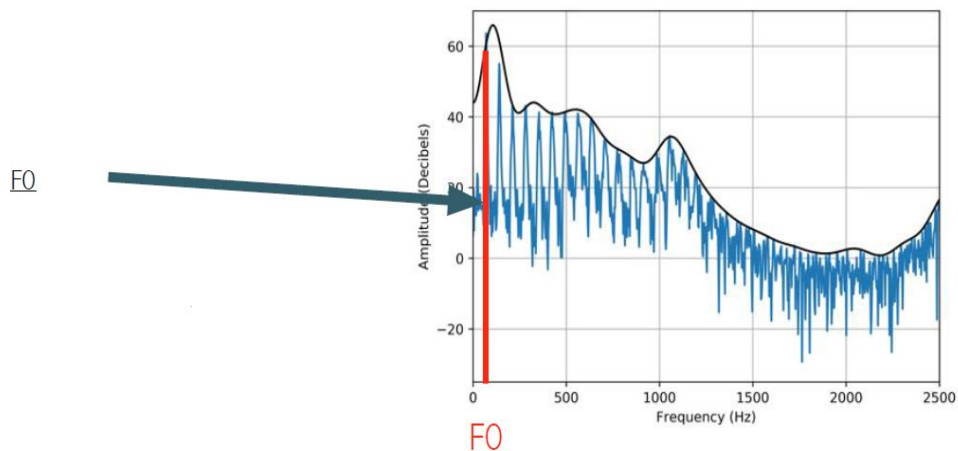


Hình 3.6 Đặc trưng Spectral Envelop của tín hiệu tiếng nói [19]

Đặc trưng bao phổ tín hiệu được chuyển đổi thành hệ số Mel (Mel coefficients - MC) theo các bước như sau:

- Bước 1: Biến đổi tín hiệu tiếng nói thành âm phổ (spectrum) bằng phép biến đổi Fast Fourier Transform.
- Bước 2: Sử dụng bộ lọc lấy được đường bao phổ (spectral envelop) của tín hiệu.
- Bước 3: Sử dụng phép biến đổi Inverse Fast Fourier Transform, trích xuất được các hệ số Mel từ đường bao phổ.

Tần số cơ bản F0 đặc trưng cho độ to của giọng nói, được lấy mẫu và logarit để chuyển đổi thành log F0 bằng công cụ DIO. Tần số cơ bản của âm thanh như hình sau:

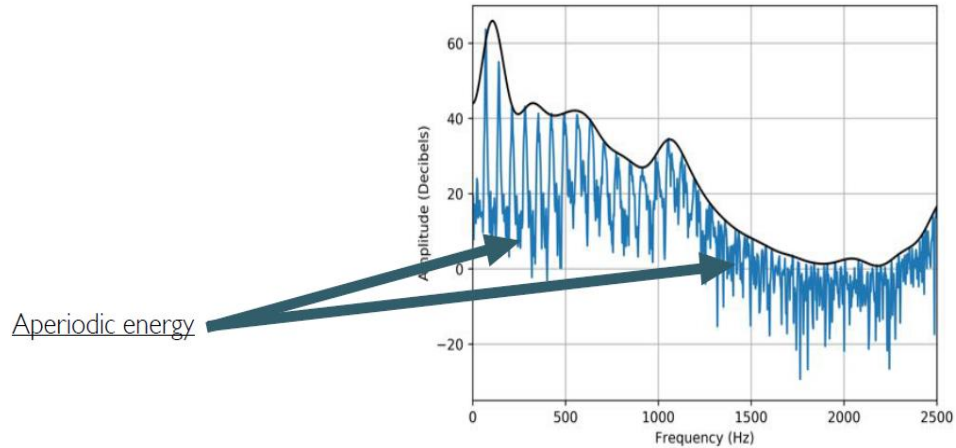


Hình 3.7 Tần số F0 của tín hiệu tiếng nói [19]

Tần số cơ bản F0 được trích xuất bằng công cụ DIO qua các bước sau:

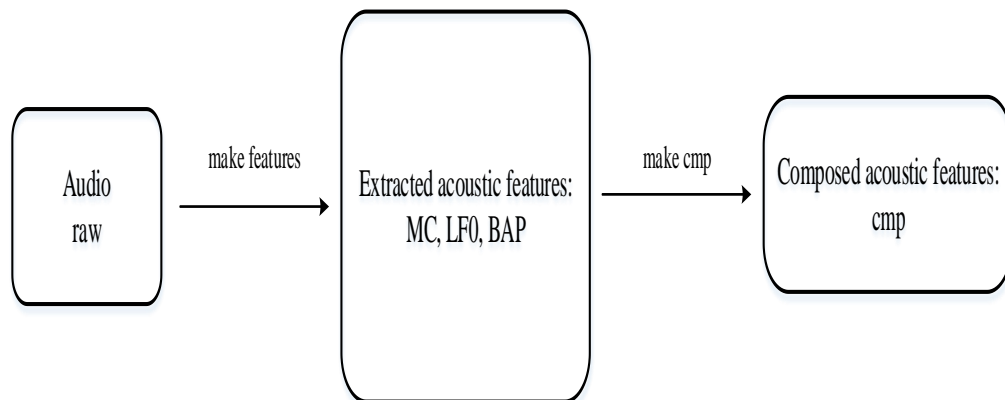
- Sử dụng bộ lọc thông thấp với các tần số cắt khác nhau để lọc tín hiệu, nếu tín hiệu có chứa thành phần tần số cơ bản thì sẽ có dạng hình Sin với chu kỳ T_0 . Do chưa biết F_0 nên sử dụng bộ lọc với tần số cắt khác nhau.
- Tìm các ứng viên cho tần số cơ bản F_0 và độ tin cậy của tần số được trích ra.
- Chọn ra ứng viên có độ tin cậy cao nhất là tần số cơ bản F_0 .

Tín hiệu kích thích không tuần hoàn đặc trưng cho độ dài và độ mạnh của giọng nói, được trích xuất bằng công cụ PLATINUM [22].



Hình 3.8 Đặc trưng Aperiodic Energy của tín hiệu tiếng nói [19]

Tín hiệu tiếng nói được trích xuất thành các đặc trưng: Hệ số Mel, tần số cơ bản F_0 , tín hiệu kích thích không tuần hoàn. Gộp 3 đặc trưng thành một và biểu diễn thành véc tơ đặc trưng âm học [17]:



Hình 3.9 Trích xuất đặc trưng âm thanh

Các véc tơ đặc trưng âm học được vào mô hình âm học, cùng với đặc trưng ngôn ngữ từ khối Label Aligner để huấn luyện thành mô hình âm học là mạng nơ ron học sâu có 6 lớp ẩn.

3.2. Pha kiểm thử

3.2.1. Khối trích chọn đặc trưng ngôn ngữ

Trong pha tổng hợp, khối trích chọn đặc trưng ngôn ngữ có chức năng trích xuất đặc trưng ngôn ngữ của văn bản cần tạo ra tiếng nói. Đặc trưng ngôn ngữ là các nhãn chứa thông tin về âm vị, được biểu diễn thành các véc tơ đặc trưng ngôn ngữ. Tuy nhiên, đặc trưng ngôn ngữ chưa được gán thời gian bắt đầu và thời gian kết thúc cho âm vị.

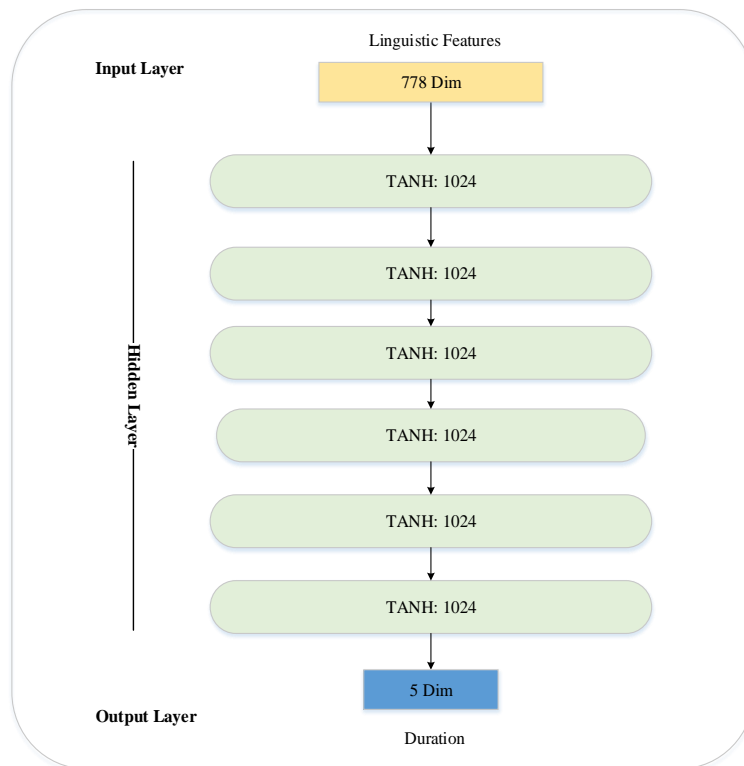
3.2.2. Mô hình thời gian

Véc tơ từ khối trích chọn đặc trưng ngôn ngữ được đưa vào mô hình thời gian. Dựa vào mô hình được sinh ra trong pha huấn luyện, âm vị được ước lượng thời gian bắt đầu và thời gian kết thúc của trạng thái khi qua mô hình thời gian.

Mô hình thời gian là mạng nơ ron học sâu có đặc điểm như sau:

- Đầu vào là véc tơ 778 chiều chứa các 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.
- Có 6 lớp ẩn, mỗi lớp có 1024 neutron và sử dụng hàm TANH [21] là hàm kích hoạt.
- Đầu ra là véc tơ 5 có chiều chứa thông tin ước lượng khoảng thời gian xuất hiện của từng trạng thái trong âm vị. Số lượng véc tơ đầu ra bằng số âm vị có trong câu.

Cấu trúc mô hình thời gian như sau:



Hình 3.10 Cấu trúc mạng nơ ron mô hình thời gian

Sau mô hình thời gian, đặc trưng ngôn ngữ của văn bản được gán thời gian bắt đầu và thời gian kết thúc cho các âm vị:

```
0 1550000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[2]
1550000 1900000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[3]
1900000 6000000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[4]
6000000 9300000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[5]
9300000 12900000 xx^xx-pau+s=aw@0_0/A:xx_0/B:xx-1@0-0&0-0/C:0+3/D:xx-0/E:xx+1/F:E-3/G:0-0/H:1=1@0=6/I:3_1/J:42+14-5[6]
12900000 13150000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[2]
13150000 13350000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[3]
13350000 13650000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[4]
13650000 13850000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[5]
13850000 14100000 xx^pau-s+aw=ngz@0_2/A:xx_1/B:0-3@0-0&0-0/C:0+1/D:xx-1/E:E+3/F:V-1/G:1-1/H:3=1@1=5/I:1_1/J:42+14-5[6]
```

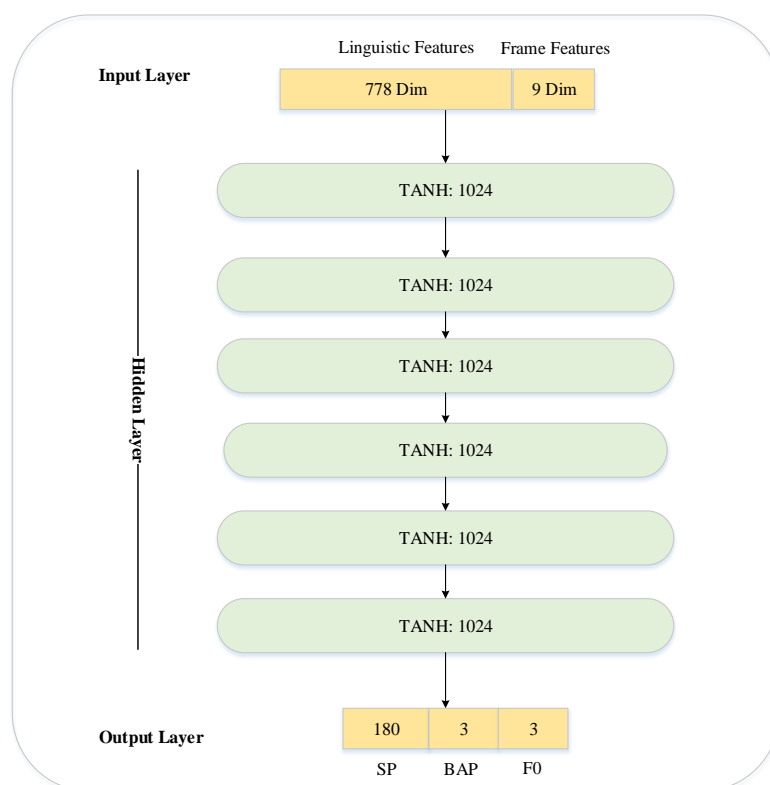
3.2.3. Mô hình âm học

Mô hình âm học được tạo ra trong pha huấn luyện được sử dụng để dự đoán đặc trưng âm học. Đầu vào là trạng thái âm vị được gán nhãn thời gian sau mô hình thời gian, đầu ra là các đặc trưng âm học tương ứng theo khung thời gian 5ms.

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

- Véc tơ đầu vào có 787 chiều, trong đó 778 chiều chứa đặc trưng ngôn ngữ của âm vị và 9 chiều để đánh số thứ tự khung (mỗi âm vị được chia nhỏ thành nhiều khung thời gian có độ dài 5ms tương ứng với đặc trưng âm học của WORLD).
- Có 6 lớp ẩn, mỗi lớp có 1024 nơ ron và sử dụng hàm TANH là hàm kích hoạt.
- Đầu ra là véc tơ 186 chiều chứa các đặc trưng âm học được ước lượng bao gồm: Đường bao phổ tín hiệu (SP), (tín hiệu kích thích) BAP, Logarit của tần số cơ bản F_0 ($\log F_0$), deltas và deltas deltas của 3 đại lượng.

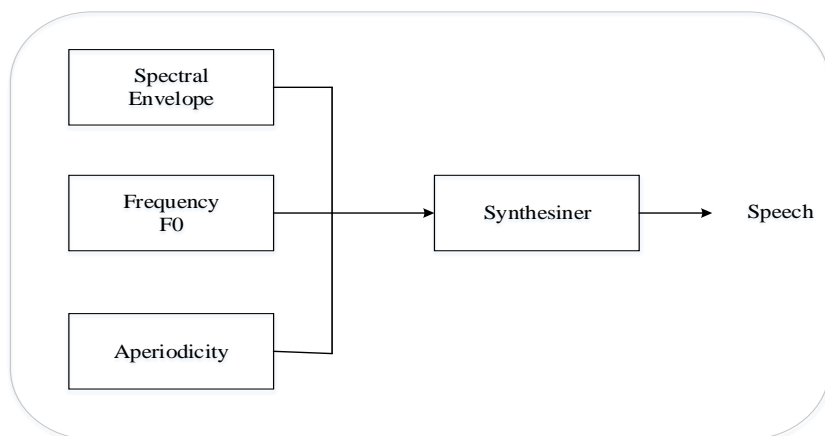
Mô hình âm học có cấu trúc như sau:



Hình 3.11 Cấu trúc mạng nơ ron mô hình âm học Acoustic

3.2.4. Khối tạo tiếng nói

Khối tạo tiếng nói là công cụ WORLD được sử dụng trong pha huấn luyện để tạo ra tiếng nói. Các đặc trưng âm học sinh ra từ mô hình âm học gồm: SP, BAP, F0. Khối tổng hợp tiếng nói sinh ra tín hiệu tiếng nói tương ứng với văn bản đầu vào.



Hình 3.12 Tổng hợp tiếng nói từ đặc trưng âm học

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Thực nghiệm

4.1.1. Môi trường thực nghiệm

Hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu và hệ thống tổng hợp tiếng nói theo phương pháp tham số thống kê HMM được cài đặt trên máy tính ảo có cấu hình hạn chế như sau:

Thành phần	Chỉ số
CPU	1.90 GHz Core i5 Intel (8 cores)
RAM	8GB
OS	Ubuntu 16.04 LTS
Disk	120GB

Bảng 4.1 Cấu hình phần cứng máy chủ thử nghiệm

Phần mềm sử dụng cho hệ thống tổng hợp theo phương pháp mạng nơ ron học sâu:

STT	Tên phần mềm	Tác giả	Nguồn
1	Python2.7		https://www.python.org/download/releases/2.7/
2	GCC 5.4		https://ftp.gnu.org/gnu/gcc/gcc-5.4.0/
3	vita_ana	Truong Do	https://bitbucket.org/vaisvn/hts_for_vietnamese/src/master/tools/textana/
4	Merlin Toolkit	Centre for Speech Technology Research (CSTR), University of Edinburgh	https://github.com/CSTR-Edinburgh/merlin
5	WORLD Vocoder	University of Yamanashi	https://github.com/mmori-se/World

STT	Tên phần mềm	Tác giả	Nguồn
6	Hidden Markov Model Toolkit	University of Cambridge	http://htk.eng.cam.ac.uk/

Bảng 4.2 Các phần mềm sử dụng trong hệ thống

Trong đó:

- Python2.7 và GCC5.4 là môi trường chạy phần mềm Merlin và HMM.
- Vita_ana tương ứng với khối trích chọn đặc trưng ngôn ngữ, có chức năng trích chọn đặc trưng ngôn ngữ từ văn bản.
- Hidden Markov Model Toolkit tương ứng với khối Label Aligned, có chức năng gán thời gian bắt đầu và kết thúc cho nhãn của âm vị.
- Merlin Toolkit là mạng nơ ron học sâu của mô hình thời gian và mô hình âm học.
- WORLD vocoder có chức năng trích chọn đặc trưng âm thanh và tổng hợp lại tín hiệu tiếng nói từ các đặc trưng âm học.

Hệ thống tổng hợp theo phương pháp tham số thống kê HMM sử dụng mã nguồn của VAIS [23].

4.1.2. Bộ dữ liệu sử dụng trong thực nghiệm

Để đánh giá hệ thống tổng hợp tiếng nói dựa trên mạng nơ ron học sâu, luận văn sử dụng bộ dữ liệu tiếng nói của VAIS và Trung tâm Không gian mạng Viettel như sau:

TÊN BỘ DỮ LIỆU	SỐ LƯỢNG CÂU	SỐ LƯỢNG TỪ	TỔNG THỜI GIAN	GIỚI TÍNH	PHƯƠNG NGỮ
Data500	500	7960 (1805 từ không lặp)	45 phút	Nữ	Miền Bắc
Data1000	1000	14383 (2515 từ không lặp)	83 phút	Nữ	Miền Bắc
Data3156	3156	47340 (5600 từ không lặp)	263 phút	Nữ	Miền Bắc

Bảng 4.3 Bộ dữ liệu thử nghiệm

Bộ dữ liệu thử nghiệm đã được tiền xử lý như: Loại bỏ các âm thanh bị nhiễu, chuẩn hóa âm thanh và nội dung văn bản, chuyển đổi số thành chữ, chuyển đổi từ viết tắt thành viết đầy đủ, chia nhỏ thành các câu có độ dài từ 15 đến 20 từ.

4.1.3. Mô hình huấn luyện

Trước khi đưa vào huấn luyện, bộ dữ liệu tiếng nói được chia thành 3 tập: Tập huấn luyện (training set), tập kiểm định (validation set) và tập kiểm tra (test set) với tỷ lệ là 90%:5%:5%. Trong đó:

- Tập dữ liệu tập huấn luyện được sử dụng để tạo ra mạng nơ ron học sâu cho mô hình thời gian và mô hình âm học trong pha huấn luyện.
- Tập dữ liệu kiểm định được sử dụng để tinh chỉnh hệ số θ , là hệ số liên kết giữa các nút nơ ron trong mạng nơ ron học sâu để được kết quả gần với giá trị đầu vào của tập kiểm định nhất. Tập dữ liệu kiểm định được sử dụng để tối ưu mô hình mạng nơ ron để cho ra kết quả tốt nhất.
- Tập kiểm tra được sử dụng để đánh giá độ chính xác của mô hình mạng nơ ron học sâu sinh ra. Đây là bước đánh giá độ đo của mô hình.

Mô hình thời gian và mô hình âm học được tối ưu bằng thuật toán Stochastic Gradient Descent [19]. Trong đó có thể điều chỉnh các tham số sau:

- Learning Rate: 0.002, là tốc độ điều chỉnh hệ số θ của mạng nơ ron để có được mô hình tối ưu nhất. Giá trị learning rate kiểm soát tốc độ thay đổi hệ số θ để phù hợp với bài toán. Giá trị learning rate cao giúp mạng nơ ron được huấn luyện nhanh hơn do cần ít lần tích tiến để về điểm tối ưu, nhưng có thể làm giảm độ chính xác do không thể tiến về điểm tối ưu.
- Batch size: 256, là số mẫu đồng thời được đưa vào huấn luyện mô hình. Đối với máy chủ có bộ nhớ ít, phải giảm số mẫu đưa vào đồng thời để tránh bị tràn bộ nhớ.
- Epoch: 25, là số lần đưa toàn bộ dữ liệu vào huấn luyện mô hình hay chính là số vòng lặp huấn luyện mô hình. Trong quá trình thực nghiệm, có thể giảm số vòng lặp Epoch nếu kết quả tinh chỉnh mô hình mạng nơ ron không có sự thay đổi lớn giữa các vòng lặp.

4.1.4. Tạo ra tiếng nói tiếng Việt từ mô hình mạng nơ ron học sâu

Quá trình tạo ra tiếng nói tiếng Việt trên hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu gồm các bước như sau:

4.1.4.1. Pha huấn luyện

- **Thiết lập tập dữ liệu:** Tập dữ liệu gồm 1000 mẫu (mỗi mẫu là cặp audio và text) được chia thành 3 tập con ngẫu nhiên: 900 mẫu cho tập training, 50 mẫu cho tập validation, 50 mẫu cho tập test.
- **Trích chọn đặc trưng ngôn ngữ:** Sử dụng phần mềm vita_ana để trích chọn các đặc trưng ngôn ngữ của tập dữ liệu, các đặc trưng ngôn ngữ được biểu diễn dưới dạng nhãn theo quy chuẩn của HTS [14].
- **Trích chọn đặc trưng âm học:** Sử dụng công cụ WORLD vocoder để trích chọn các đặc trưng âm học. Đặc trưng âm học được dùng để gán thời gian xuất hiện cho đặc trưng ngôn ngữ, và đồng thời được sử dụng để huấn luyện mô hình âm học.
- **Gán nhãn thời gian cho đặc trưng ngôn ngữ:** Sử dụng Hidden Markov Model Toolkit để gán thời gian cho âm vị. Đầu vào là các đặc trưng ngôn ngữ và đặc trưng âm học, đầu ra là các đặc trưng ngôn ngữ của âm vị đã được gán thời gian.
- **Huấn luyện mạng nơ ron cho mô hình thời gian:** Đặc trưng ngôn ngữ đã được gán thời gian được biểu diễn thành các véc tơ đặc trưng ngôn ngữ theo bộ câu hỏi HTS được thiết kế riêng cho tiếng Việt. Các véc tơ đặc trưng được đưa vào huấn luyện để tạo ra mạng nơ ron học sâu cho mô hình thời gian. Mô hình thời gian được sử dụng để dự đoán thời gian cho các âm vị trong pha tổng hợp.
- **Huấn luyện mạng nơ ron cho mô hình âm học:** Véc tơ đặc trưng ngôn ngữ và véc tơ đặc trưng âm học được đưa vào huấn luyện để tạo ra mạng nơ ron học sâu cho mô hình âm học. Mô hình âm học được sử dụng để dự đoán đặc trưng âm học tương ứng với đặc trưng âm vị trong pha tổng hợp.

4.1.4.2. Pha tổng hợp

- **Chuẩn hóa văn bản:** Văn bản chuyển thành tiếng nói cần được chuẩn hóa, ví dụ: Chuẩn hóa từ viết tắt thành từ viết đầy đủ, chuẩn hóa số thành chữ, ngày tháng dạng số thành dạng chữ...
- **Trích chọn đặc trưng ngôn ngữ:** Văn bản chuẩn hóa được trích chọn ra đặc trưng ngôn ngữ bằng phần mềm vita_ana, biểu diễn thành dạng véc tơ ngôn ngữ đặc trưng ngôn ngữ nhờ bộ câu hỏi HTS.
- **Gán nhãn thời gian cho đặc trưng ngôn ngữ:** Véc tơ đặc trưng ngôn ngữ được đưa vào mạng nơ ron học sâu của mô hình thời gian. Sau mô hình thời gian, véc tơ đặc trưng ngôn ngữ được gán thêm thông tin thời gian xuất hiện của âm vị. Mô hình thời gian là mạng nơ ron truyền thẳng 6 lớp ẩn, véc tơ đầu vào là đặc trưng ngôn ngữ có 778 chiều, véc tơ đầu ra thời gian xuất hiện của đặc trưng ngôn ngữ có 5 chiều.

- **Dự đoán đặc trưng âm học:** Sau khi gán nhãn thời gian, véc tơ đặc trưng ngôn ngữ được đưa vào mạng nơ ron học sâu của mô hình âm học. Sau mô hình âm học, các đặc trưng âm học tương ứng đặc trưng ngôn ngữ được sinh ra. Mô hình âm học là mạng nơ ron truyền thẳng 6 lớp ẩn, véc tơ đầu vào là đặc trưng ngôn ngữ đã được chia nhỏ thành khung 5ms có 787 chiều (778 chiều đặc trưng ngôn ngữ và 9 chiều mới để xác định các khung 5ms), véc tơ đầu ra là đặc trưng âm học: BAP, SP, F0.
- **Tổng hợp tiếng nói:** Các đặc trưng âm học được vào phần mềm WORLD vocoder, tại đây tín hiệu tiếng nói tiếng Việt được tổng hợp ra tương ứng với văn bản đầu vào.

4.2. Đánh giá kết quả

4.2.1. Phương pháp đánh giá

Tiếng nói tổng hợp được đánh giá bằng phương pháp so sánh điểm MOS (Mean Opinion Score), là điểm trung bình theo cảm nhận của người nghe. Phương pháp đánh giá như sau:

- Mời 10 người tham gia đánh giá và cho điểm chất lượng.
- Tiêu chí cho điểm chất lượng dựa vào độ tự nhiên và độ nghe dễ hiểu của giọng nói tổng hợp.
- Điểm số được chấm ở thang điểm 5 với các mức: 1 - Rất tệ (không nghe hiểu được), 2 - Tệ (chỉ nghe hiểu 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).
- So sánh điểm trung bình của 10 người tham gia đánh giá với tiếng nói tổng hợp của phương pháp thống kê tham số HMM và phương pháp mạng nơ ron học sâu.

4.2.2. Bảng so sánh tiếng nói tổng hợp từ 2 mô hình DNN và HMM

Điểm đánh giá độ dễ hiểu và độ tự nhiên của giọng nói tổng hợp của 2 phương pháp như sau:

BẢNG SO SÁNH GIỌNG NÓI TỔNG HỢP GIỮA HMM VÀ DNN		
Data	MOS	
	HMM	DNN
Data500	3.8	4.2
Data1000	4.1	4.3
Data3156	4.3	4.5

Bảng 4.4 Bảng so sánh tiếng nói tổng hợp

Nhận xét: Giọng nói được tổng hợp bằng phương pháp mạng nơ ron học sâu có mức độ dễ hiểu (Intelligibility) tương đương với phương pháp tham số thống kê HMM. Tuy nhiên, phương pháp mạng nơ ron học sâu cho ra tiếng nói có độ tự nhiên (Naturalness) gần giống với giọng nói nguyên bản; giọng nói tổng hợp từ phương pháp tham số thống kê HMM có tốc độ đều đều, làm giảm sắc thái cảm xúc trong câu nói.

4.2.3. Kết quả đánh giá

So sánh phương pháp tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu và phương pháp tổng hợp tiếng nói dựa vào thống kê tham số của HMM:

- Cả 2 phương pháp đều cho tiếng nói có thể nghe hiểu được nội dung.
- Phương pháp tổng hợp dựa trên mạng nơ ron học sâu cho ra tiếng nói tự nhiên hơn, gần giống với tiếng nói nguyên bản nhất.
- Phương pháp tổng hợp theo phương pháp tham số thống kê HMM cho ra tiếng nói nghe có vẻ đều đều, quá mịn hay quá ổn định làm giảm ngôn điệu, sắc thái cảm xúc hay phong cách nói trong câu.

CHƯƠNG 5: KẾT LUẬN

5.1. Kết quả đạt được của luận văn

Sau toàn bộ quá trình hoàn thành luận văn, tôi đã đạt được một số kết quả như sau:

- Nắm vững cơ sở lý thuyết tổng hợp tiếng nói nói chung và tổng hợp tiếng nói tiếng Việt nói riêng.
- Nắm vững cơ sở lý thuyết về mạng nơ ron nhân tạo.
- Xây dựng hệ thống tổng hợp tiếng nói tiếng Việt theo phương pháp mạng nơ ron học sâu.

Hệ thống tổng hợp tiếng nói tiếng Việt theo phương pháp học sâu đã được ứng dụng và triển khai trong Tập đoàn Công nghiệp Viễn thông Quân đội Viettel như: Hệ thống trợ lý ảo Viettel, hệ thống callbox chăm sóc khách hàng.

5.2. Đánh giá hệ thống

Sau quá trình thử nghiệm về đánh giá hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu. Ngoài những ưu điểm đã được đưa ra ở chương 4, tôi nhận thấy hệ thống có các nhược điểm như sau:

- Hệ thống nhạy cảm với dữ liệu nhiễu, cần phải có bước xử lý làm sạch dữ liệu trước khi đưa vào mô hình huấn luyện.
- Hệ thống gồm nhiều khối được ghép nối với nhau, sai số của hệ thống là tổng sai số của các khối. Do vậy, hệ thống có sai số lớn hơn hệ thống có kiến trúc E2E.
- Hệ thống phù hợp với các ứng dụng chạy trên máy chủ có cấu hình thấp, không cần GPU để huấn luyện dữ liệu.

5.3. Hướng phát triển

Hệ thống tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu cho ra tiếng nói tổng hợp có chất lượng khá tốt. Vì vậy, hướng phát triển tiếp theo của luận văn là tiếp tục thử nghiệm để tối ưu để đưa phương pháp tổng hợp tiếng nói theo phương pháp học sâu vào các hệ thống công nghiệp như: Thiết bị định vị dẫn đường, tổng đài chăm sóc khách hàng và cung cấp dịch vụ báo nói cho các trang báo điện tử.

TÀI LIỆU THAM KHẢO

- [1] P. T. Son, P. T. Nghĩa, "Một số vấn đề về tổng hợp tiếng nói tiếng Việt," in Hội thảo quốc gia 2014 về Điện tử, Truyền thông và Công nghệ thông tin, 2014.
- [2] P. T. Son and D. T. Cường, "Trích trộn các tham số đặc trưng tiếng nói cho hệ thống tổng hợp tiếng nói tiếng Việt dựa vào mô hình Markov ẩn," Tạp chí Tin học và Điều khiển, 2013.
- [3] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," IEEE, 2013.
- [4] T. Masuko, "HMM-Based Speech Synthesis and Its Applications," 2002.
- [5] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous., "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," 2017.
- [6] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," IEICE, 2002.
- [7] Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura, "An HMM-based Vietnamese speech synthesis system," in 2009 Oriental COCODA International Conference on Speech Database and Assessments, Urumqi, China, 2009.
- [8] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in International Conference on Acoustics, Speech, and Signal Processing, Boston Massachusetts, 1983.
- [9] Hyungwon Choi and Yunhun Jang, "http://alinlab.kaist.ac.kr," [Online].
- [10] Vũ Hữu Tiệp, "machinelearningcoban," 8 June 2018. [Online]. Available: <https://machinelearningcoban.com/ebook/>.
- [11] N. H. Huy, "Nghiên cứu các đặc trưng tín hiệu và ràng buộc ngôn điệu để nâng cao chất lượng tổng hợp và nhận dạng tiếng Việt," Học viện Khoa học và Công nghệ, 2016.
- [12] Đ. T. Thuật, Ngữ âm tiếng Việt, NXB Đại học Quốc gia Hà Nội.
- [13] H. C. Tín, Giáo trình Cơ sở ngữ âm học, Đại học Cần thơ.
- [14] H. W. Group, "An example of context-dependent label format," 2015.
- [15] O. W. S. K. Zhizheng Wu, "Merlin: An open source neural network speech synthesis system," The Centre for Speech Technology Research, University of Edinburgh, 2017.

- [16] F. Y. K. O. Masannori Morise, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Application," IEICE, 2016.
- [17] "<http://www.cs.columbia.edu/~ecooper/tts/data.html>," Columbia University. [Online].
- [18] N. H. HUY, "Nghiên cứu các đặc trưng tín hiệu và ràng buộc ngôn điệu để nâng cao chất lượng tổng hợp và nhận dạng tiếng Việt," Học viện Khoa học và Công nghệ, Hà Nội, 2016.
- [19] Simon King, Oliver Watts, Srikanth Ronanki, Felipe Espic, Zhizheng Wu, "<http://media.speech.zone/>," 2017.
[Online].
Available:http://media.speech.zone/images/Interspeech2017_tutorial_Merlin_for_publication_watermarked_compressed_v2.pdf.
- [20] M. Morise, CheapTrick, a spectral envelope estimator for high-quality, Yamanashi, 2015.
- [21] Chigozie Enyinna Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall, "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning," 2018.
- [22] M. Morise, "PLATINUM: A method to extract excitation signals for voice synthesis system," The Acoustical Society of Japan, Kusatsu, 2011.
- [23] "bitbucket.org," VAIS, 1 2017. [Online]. Available: https://bitbucket.org/vaisvn/hts_for_vietnamese/src/master/.
- [24] Đặng Ngọc Đức, Lương Chi Mai, "Tăng cường độ chính xác của hệ thống," Tạp chí Bưu chính Viễn thông, Hà Nội, 2004.