NGUYỄN VĂN THỊNH

HỆ THỐNG THÔNG TIN

CLC2017B

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**
**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**
**----------**

**Nguyễn Văn Thịnh**

# NGHIÊN CỨU PHÁT TRIỂN HỆ THỐNG TỔNG HỢP TIẾNG NÓI TIẾNG VIỆT SỬ DỤNG CÔNG NGHỆ HỌC SÂU

LUẬN VĂN THẠC SĨ KHOA HỌC
HỆ THỐNG THÔNG TIN

Hà Nội 2018

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**
**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**
**----------**

**Nguyễn Văn Thịnh**

# NGHIÊN CỨU PHÁT TRIỂN HỆ THỐNG TỔNG HỢP TIẾNG NÓI TIẾNG VIỆT SỬ DỤNG CÔNG NGHỆ HỌC SÂU

Chuyên ngành : Hệ Thống Thông Tin

LUẬN VĂN THẠC SĨ KHOA HỌC
HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC :
TS. Mạc Đăng Khoa

Hà Nội 2018

# LỜI CẢM ƠN

Đầu tiên, tôi xin được gửi lời cảm ơn chân thành tới Viện nghiên cứu quốc tế MICA nơi đã tạo điều kiện cho tôi thực hiện luận văn này. Tiếp đến, tôi xin cảm ơn trung tâm không gian mạng VIETTEL, nơi tôi làm việc, đã tạo điều kiện và giúp đỡ tôi trong việc hoàn thành hệ thống mà tôi trình bày trong luận văn thạc sỹ này. Tôi xin chân thành cảm ơn TS. Mạc Đăng Khoa người thầy, người hướng dẫn tôi trong suốt thời gian qua để tôi có thể hoàn thành luận văn cho mình.

Thêm nữa, tôi xin chân thành cảm ơn anh Nguyễn Tiến Thành, chị Nguyễn Hằng Phương cùng toàn thể các bộ viện nghiên cứu quốc tế MICA đã giúp đỡ tôi trong quá trình làm luận văn tại viện nghiên cứu quốc tế MICA.

Tôi xin gửi lời cảm ơn trận trọng đến anh Nguyễn Quốc Bảo cùng toàn thể đồng nghiệp của tôi tại nhóm voice trung tâm không gian mạng VIETTEL, ban giám đốc trung tâm cùng toàn thể anh chị em trong trung tâm đã giúp đỡ hỗ trợ tôi trong quá trình tôi hoàn thành luận văn thạc sỹ này.

Cuối cùng tôi xin gửi lời cảm ơn tới cô Đỗ Thị Ngọc Diệp, người đã hướng dẫn tôi từ khi còn là sinh viên đại học và hỗ trợ, giúp đỡ tôi đến khi tôi hoàn thành luận văn này.

Hà Nội, ngày 27 tháng 03 năm 2018

Nguyễn Văn Thịnh

# MỤC LỤC

# DANH MỤC HÌNH ẢNH

# DANH MỤC BẢNG

# DANH MỤC TỪ VIẾT TẮT VÀ THUẬT NGỮ

| **Từ viết tắt** | **Từ đầy đủ** | Ý nghĩa. |
|---|---|---|
| HMM | Hidden markov model | Mô hình markov ẩn |
| DNN | Deep Neural Network | Mạng nơ ron học sâu |
| PSOLA | Pitch Synchronous Overlap and Add | Kỹ thuật chồng đồng bộ cao độ tần số cơ bản |
| TTS | Text To Speech | Tổng hợp văn bản thành tiếng nói. |
| MSLA | Mel Log Spectral Approximation | xấp xỉ phổ mel. |
| GMM | Gaussian mixture model | Mô hình gauss hỗn hợp |
| VLSP | Vietnamese language and speech processing | Xử lý ngôn ngữ và tiếng nói tiếng Việt |
| MOS | Mean opinion score | Điểm ý kiến trung bình |
| F0 | Fundamental frequency | Tần số cơ bản |

# MỞ ĐẦU

Hiện nay, lĩnh vực tổng hợp tiếng nói đã được nghiên cứu và phát triển ở rất nhiều nơi trên thế giới, nhiều công nghệ và phương pháp khác nhau được thử nghiệm, triển khai thành công, thậm chí có những công trình đã đạt đến mức khó có thể phân biệt được với giọng đọc của con người. Còn ở Việt Nam, cũng đã có nhiều công trình nghiên cứu và sản phẩm về lĩnh vực tổng hợp tiếng nói, có thể kể đến như các nghiên cứu của Viện công nghệ thông tin thuộc Viện hàn lâm khoa học công nghệ Việt Nam ([1], [2]), các nghiên cứu này đều dựa trên kiến trúc của hệ thống HTS[3] để xây dựng hệ thống tổng hợp tiếng nói, và mô hình được áp dụng là mô hình Markov ẩn. Các công trình nghiên cứu và hệ thống thực tế về tổng hợp tiếng nói ở Việt nam hiện nay chủ yếu được phát triển dựa trên hai phương pháp: tổng hợp tiếng nói ghép nối và tổng hợp tiếng nói thống kê dựa trên mô hình Markov ẩn (HMM). Hai phương pháp nêu trên là hai phương pháp đã được nghiên cứu và phát triển nhiều năm trên thế giới cũng như ở Việt Nam, đã có nhiều sản phẩm, hệ thống thành công với nó. Tuy nhiên hai phương pháp này vẫn còn nhiều mặt hạn chế như chất lượng tiếng nói tổng hợp không thật đối với HMM và cơ sở dữ liệu cần lưu trữ lớn cũng như chỉ cho chất lượng tốt trong miền hẹp đối với tổng hợp ghép nối. Mặt khác trên thế giới hiện nay đã bắt đầu phát triển một công nghệ tổng hợp tiếng nói mới, đó là tổng hợp tiếng nói dựa trên phương pháp học sâu, nó cũng đã cho thấy những kết quả tích cực, chất lượng tổng hợp của hệ thống ở mức cao, gần với tự nhiên[4]. Vì hai lý do trên, để tài được đề xuất thực hiện nhằm thử nghiệm áp dụng công nghệ học sâu vào trong tổng hợp tiếng nói tiếng Việt với mong muốn tạo được một hệ thống tổng hợp tiếng nói có chất lượng cao.

Đề tài này tập trung nghiên cứu áp dụng công nghệ tổng hợp tiếng nói dựa trên mạng nơ ron học sâu cho tổng hợp tiếng nói tiếng Việt, sao cho đạt được một hệ thống có chất lượng giọng tổng hợp tốt hơn so với các hệ thống tổng hợp tiếng Việt sử dụng các công nghệ khác cũ hơn. Để làm được điều này, tác giả đã đề ra các nhiệm vụ chính cần hoàn thành như sau:

- Nghiên cứu về phương pháp tổng hợp tiếng nói dựa trên công nghệ học sâu và cách áp dụng.
- Triển khai xây dựng hệ thống tổng hợp tiếng nói dựa trên công nghệ này.
- Áp dụng một số giải pháp tiền xử lý dữ liệu để nâng cao chất lượng giọng tổng hợp.

Luận văn này được xây dựng trong quá trình làm việc tại trung tâm không gian mạng VIETTEL và thời gian làm việc tại phòng Giao tiếp tiếng nói thuộc Viện nghiên cứu quốc tế MICA. Với môi trường làm việc nghiêm túc, được sự hướng dẫn của TS. Mạc Đăng Khoa cùng với sự trợ giúp của đồng nghiệp và các anh, chị, thầy, cô ở Viện Nghiên cứu quốc tế MICA tôi đã đúc rút được kinh nghiệm và hoàn thành luận văn này.

Sau đây là bố cục chính của luận văn

- CHƯƠNG 1 TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI: Chương này giới thiệu chung về tổng hợp tiếng nói, tình hình nghiên cứu và phát triển các hệ thống tổng hợp tiếng nói, và các phương pháp tổng hợp tiếng nói phổ biến hiện nay.

- CHƯƠNG 2: PHƯƠNG PHÁP HỌC SÂU ÁP DỤNG TRONG TỔNG HỢP TIẾNG NÓI: Chương này chủ yếu nói về phương pháp học sâu và cách áp dụng nó trong trong tổng hợp tiếng nói.
- CHƯƠNG 3: XÂY DỰNG HỆ THỐNG TỔNG HỢP TIẾNG NÓI TIẾNG VIỆT VỚI CÔNG NGHỆ HỌC SÂU: Chương này chủ yếu nói về kiến trúc hệ thống tổng hợp tiếng nói tiếng Việt dựa trên phương pháp học sâu, cách triển khai xây dựng từng mô đun dựa trên kiến trúc này và cách thu thập, phương pháp xử lý, lọc dữ liệu cho hệ thống tổng hợp tiếng nói.
- CHƯƠNG 4: CÀI ĐẶT THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ: Chương này chủ yếu nói về cách thức cài đặt, thử nghiệm và đánh giá kết quả hệ thống tổng hợp tiếng nói đã được xây dựng.
- Phần KẾT LUẬN: Phần này là phần kết luận về luận văn cũng như những phương hướng nghiên cứu, cải thiện.

# LỜI CAM ĐOAN

Tôi là Nguyễn Văn Thịnh, là tác giả của luận văn này. Trong đề tài Nghiên cứu phát triển hệ thống tổng hợp tiếng nói tiếng Việt sử dụng công nghệ học sâu, hệ thống được xây dựng bao gồm bốn mô đun chính: Mô đun chuẩn hóa văn bản (Text normalization), mô đun trích chọn đặc trưng ngôn ngữ (Linguistic Feature Extraction), mô đun tạo tham số đặc trưng (Parameter Generation) và mô đun tạo tín hiệu tiếng nói (Waveform Generation). Trong bốn mô đun trên, tác giả tham gia và có đóng góp chính trong việc xây dựng ba mô đun là mô đun trích chọn đặc trưng ngôn ngữ, mô đun tạo tham số đặc trưng, mô đun tạo tín hiệu tiếng nói.

Tác giả xin cam đoan toàn bộ những gì nêu trên cũng như toàn bộ các phần triển khai trong luận văn là thật.

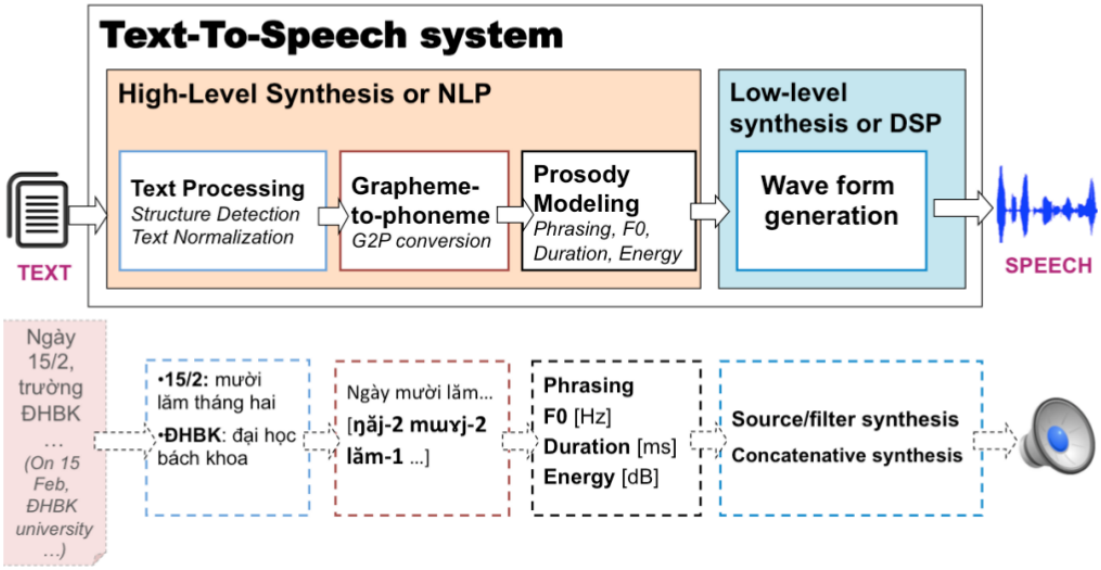
# CHƯƠNG 1: TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI

## 1.1 Giới thiệu về tổng hợp tiếng nói

### 1.1.1 Tổng quan về tổng hợp tiếng nói

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói của con người từ văn bản, hệ thống tổng hợp tiếng nói là hệ thống nhận đầu vào là một văn bản và tạo ra tín hiệu tiếng nói tương ứng ở đầu ra. Nghiên cứu về tổng hợp tiếng nói đã bắt đầu từ rất lâu, năm 1779 nhà khoa học người đan mạch Christian Kratzenstein đã xây dựng mô phỏng đơn giản hệ thống cấu âm của con người, mô hình này đã có thể phát ra được âm thanh của một số nguyên âm dài[5]. Đến tận thế kỷ 19 các nghiên cứu tổng hợp tiếng nói vẫn còn ở mức đơn giản, phải sang thế kỷ 20 khi mà có sự lớn mạnh của hệ thống điện, điện tử thì mới thực sự xuất hiện những hệ thống tổng hợp tiếng nói chất lượng, có thể kể đến như hệ thống VODER lần đầu được giới thiệu năm 1939[6]. Cho đến hiện nay, có rất nhiều các sản phẩm như sách nói, đồ chơi,.. sử dụng công nghệ tổng hợp tiếng nói. Đặc biệt các mô đun tổng hợp tiếng nói còn được tích hợp trong các trợ lý ảo trên điện thoại và máy tính như Siri[1] hay Cortana[2].

Qua quá trình phát triển, hiện nay về cơ bản một hệ thống tổng hợp tiếng nói bao gồm hai thành phần chính: phần xử lý ngôn ngữ tự nhiên và phần xử lý tổng hợp tiếng nói[7]. Phần xử lý ngôn ngữ tự nhiên: chuẩn hóa, xử lý các văn bản đầu vào thành các thành phần có thể phát âm được. Phần xử lý tổng hợp tiếng nói: Tạo ra tín hiệu tiếng nói từ các thành phần phát âm được nêu trên[8]. Trên hình 1 mô tả một hệ thống tổng hợp tiếng nói gồm hai thành phần nêu trên.




Hình 1: Sơ đồ tổng quát một hệ thống tổng hợp tiếng nói [9]

### 1.1.2 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 có nhiệm vụ trích chọn các thông tin về ngữ âm, ngữ điệu của văn bản đầu vào. Thông tin ngữ

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

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[7]. Quá trình xử lý ngôn ngữ tự nhiên thường bao gồm ba bước (xem trên hình 1):

- 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 hóa 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. Chuẩn hóa văn bản là một vấn đề khó với nhiều nhập nhằng trong cách đọc, ví như chữ số có nhiều cách đọc khác nhau tùy theo văn cảnh khác nhau, như 3579 có thể được đọc là "ba nghìn năm trăm bẩy chín" nếu coi nó là một số nhưng cũng có thể đọc là "ba năm bẩy chín" nếu như nó là một mã xác thực, các từ viết tắt cũng vậy, cũng có nhiều cách đọc phụ thuộc vào quy ước của người viết.

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 và đơn giản nhất là dựa vào từ điển, sử dụng một từ điển lớn có chứa tất cả các từ của một ngôn ngữ và chứa cách phát âm đúng tương ứng cho từng từ. Việc xác định cách phất âm đúng cho từng từ chỉ đơn giản là tra từ điển và thay đoạn văn bản bằng chuỗi âm vị đã ghi trong từ điển. 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 ưu nhược điểm khác nhau, cách dựa trên từ điển nhanh và chính xác, nhưng sẽ không hoạt động nếu từ phát âm không có trong từ điển. Và lượng từ vựng cần lưu là lớn. Cách dùng quy tắc 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 các điểm dừng nghỉ lấy hơi khi phát âm hoặc đọc một đoạn văn bản[10]. 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 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.

### 1.1.3 Tổng hợp tín hiệu tiếng nói

Khối xử lý tổng hợp tiếng nói đảm nhận việc 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 công nghệ 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 một 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 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 tính chất này. Hiện nay có ba phương pháp chính, phổ biến nhất là: tổng hợp mô hình hóa hệ thống phát âm, tổng hợp cộng hưởng tần số và tổng hợp ghép nối, ngoài ra cũng có các phương pháp khác phát triển từ ba phương pháp trên [11].

## 1.2 Các phương pháp tổng hợp tiếng nói

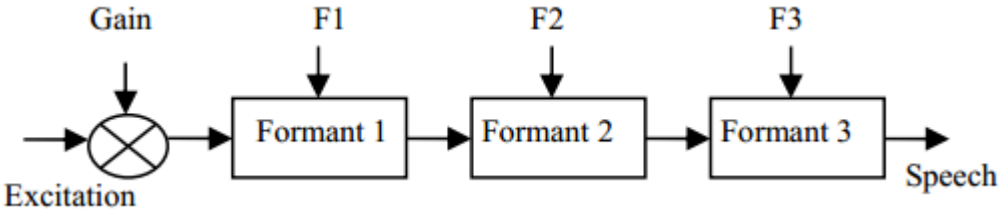
### 1.2.1 Tổng hợp mô phỏng hệ thống phát âm

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 quá trình tạo ra tiếng nói trên đó. Vì mục tiêu của phương pháp này là mô phỏng quá trình tạo tiếng nói sao cho càng giống cơ chế của con người càng tốt, nên về mặt lý thuyết đây được xem là phương pháp cơ bản nhất để tổng hợp tiếng nói, nhưng cũng vì vậy mà phương pháp này khó thực hiện nhất và khó có thể tổng hợp được tiếng nói chất lượng cao[12]. 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 NeXT, vốn được phát triển thương mại hóa bởi Trillium Sound Research Inc, Canada. Để thực hiện được phương pháp tổng hợp dựa trên việc mô phỏng hệ thống phát âm đòi hỏi thời gian, chi phí và công nghệ. Phương pháp này khó có thể ứng dụng tại Việt Nam thời điểm hiện nay.

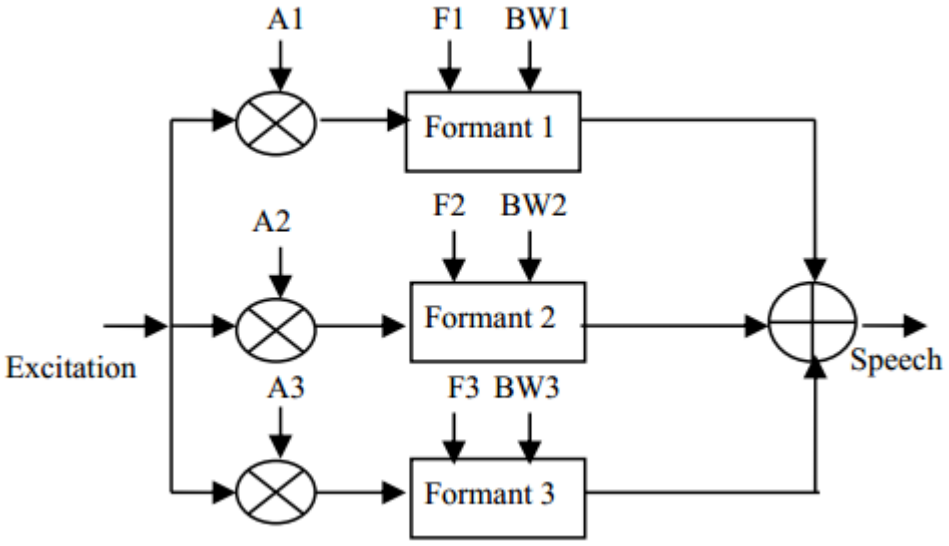
### 1.2.2 Tổng hợp tần số formant

Tổng hợp tiếng nói formant là phương pháp tổng hợp tiếng nói không sử dụng mẫu giọng thật nào khi chạy, 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 của các cơ quan phát âm bằng một tập hợp các bộ lọc. Các bộ lọc này được gọi là các bộ lọc cộng hưởng formant, chúng có thể được kết hợp song song hoặc nối tiếp với nhau hoặc kết hợp cả hai.

Tổng hợp nối tiếp là bộ tổng hợp formant có các tầng nối tiếp, đầu ra của bộ cộng hưởng này là đầu vào của bộ cộng hưởng kia, cấu trúc cơ bản bộ tổng hợp nối tiếp được biểu diễn trên hình 2.




Hình 2: Cấu trúc cơ bản bộ tổng hợp formant nối tiếp[13].




Hình 3: Cấu trúc cơ bản bộ tổng hợp formant song song[13].

Tổng hợp song song (trên hình 3) bao gồm các bộ cộng hưởng mắc song song. Đầu ra là kết hợp của tín hiệu nguồn và tất cả các formant. Cấu trúc song song cần nhiều thông tin để điều khiển hơn cấu trúc nối tiếp.

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 âm tiết sau khi được tạo thành sẽ được tiếp tục ghép lại với nhau tạo thành đoạn 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 giữ 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 thỉnh thoảng tạo ra những tiếng cọ xát không tự nhiên ở phần ghép nối. 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, âm tiết, từ, nhóm từ hoặc câu văn. Một bảng tra các đơn vị được lập ra dựa trên các phần đã

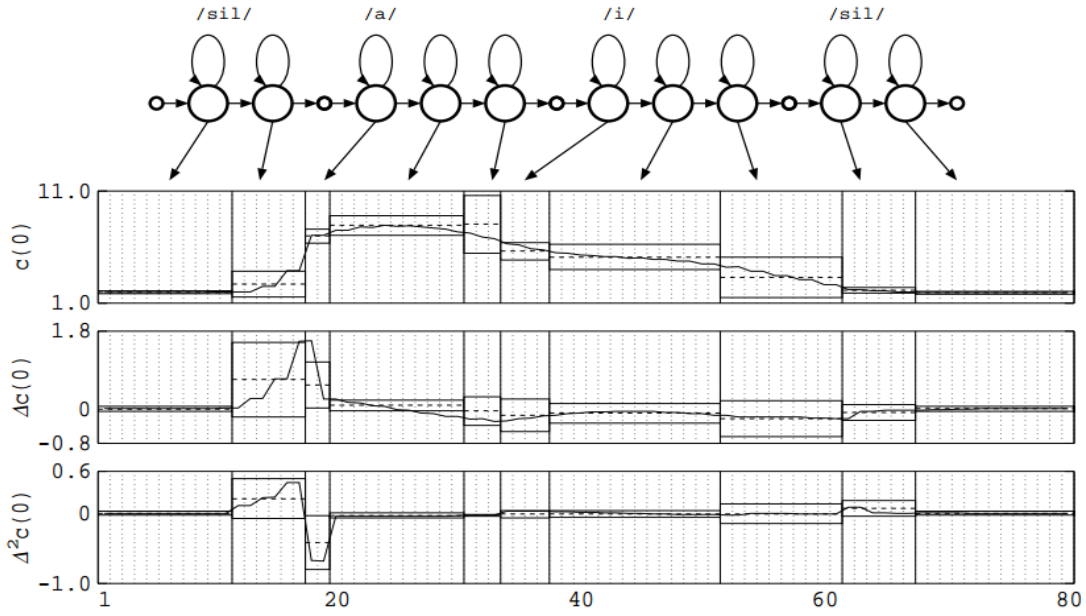
tách và các thông số âm học như tần số cơ bản, thời lượng, vị trí của âm tiết và các tiếng gần nó. Khi chạy các câu nói được tạo ra bằng cách xác định chuỗi đơn vị phù hợp nhất từ cơ sở dữ liệu. Quá trình này được gọi là chọn đơn vị và thường cần dùng đến cây quyết định được thực hiện. Thực tế, các hệ thống chọn đơn vị có thể tạo ra được giọng nói rất giống với người thật, tuy nhiên để đạt 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 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 chứa trong cơ sở dữ liệu, khi chạy thì lời văn được chồng lên các đơn vị này bằng kỹ thuật xử lý tín hiệu số nhờ mã tuyên đoán tuyến tính hay PSOLA [14]. Chất lượng âm thanh tổng hợp theo cách này thường không cao bằng phương pháp chọn đơn vị nhưng tự nhiên hơn cộng hưởng tần số và ưu điểm của nó là có kích thước dữ liệu nhỏ.

Tổng hợp chuyên biệt (Domain-specific) là phương pháp ghép nối từ các đoạn văn bản 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 nghành, sử dụng 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. Mức độ tự nhiên của hệ thống này có thể rất cao vì số lượng các câu nói không nhiều và khớp với lời văn, âm điệu của giọng nói ghi âm. Tuy nhiên hệ thống kiểu này 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

Tiếp theo đây chúng ta sẽ xem xét đến một phương pháp tổng hợp tiếng nói được nghiên cứu phổ biến và rộng rãi hiện nay đó là phương pháp tổng hợp dựa trên mô hình Markov ẩn (HMM) [15]. 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 4: Mô hình markov ẩn áp dụng trong tổng hợp tiếng nói

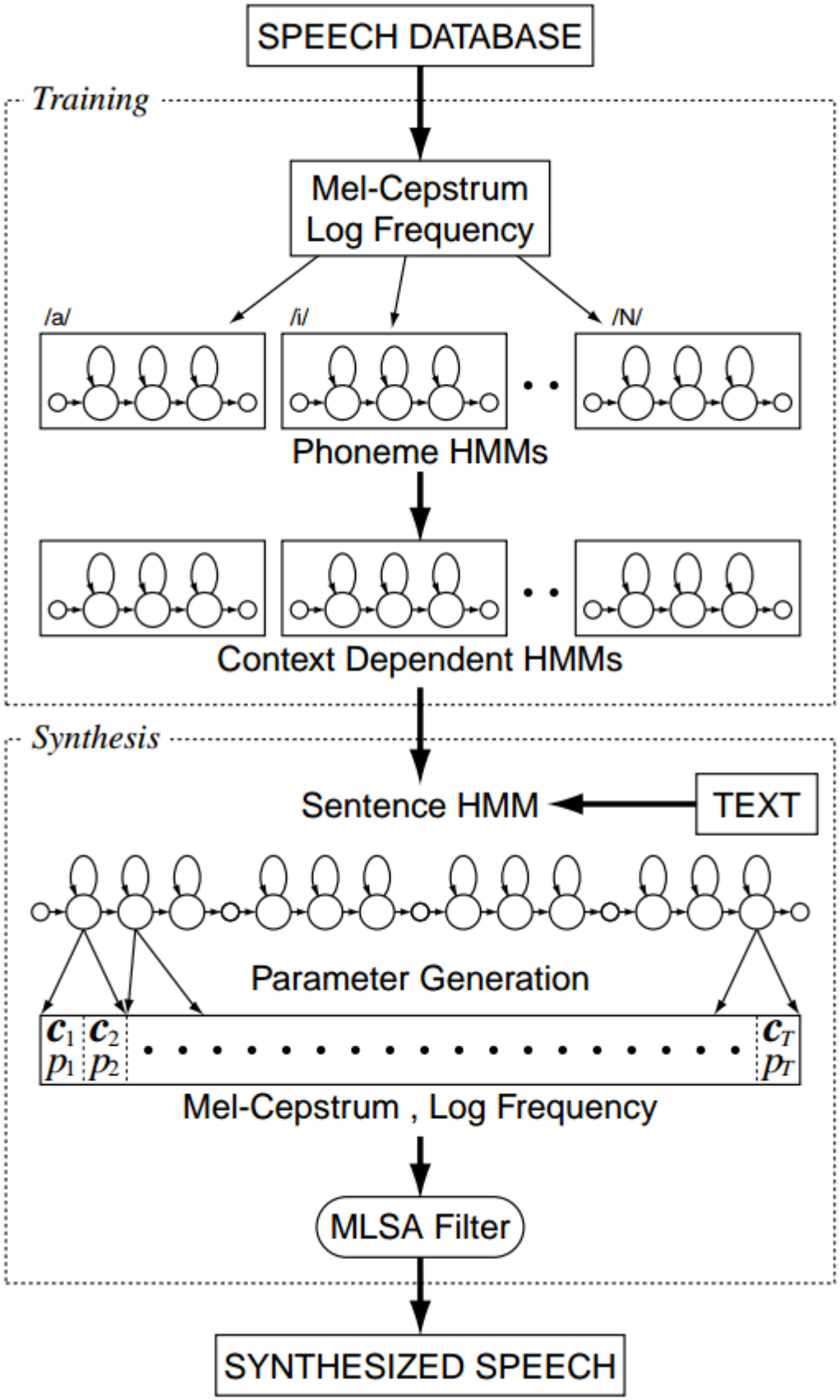
Hình 4 mô tả cách áp dụng mô hình markov ẩn trong tổng hợp tiếng nói, trong đó mỗi mô hình markov ẩn được sử dụng để mô hình hóa một âm vị, và các mô hình markov ẩn được móc nối với nhau để mô hình hóa chuỗi âm vị. Mô hình markov ẩn là một mô hình học máy dựa trên thống kê, do đó hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn hoạt động bao gồm hai quá trình là quá trình huấn luyện và quá trình tổng hợp. Hình 5 mô tả quá trình tổng hợp và huấn luyện một hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn.

Quá trình tổng hợp dựa trên mô hình markov ẩn sẽ là quá trình mà nhận đầu vào là một đoạn văn bản, chuyển hóa đoạn văn bản này thành chuỗi âm vị, sau đó dựa vào các mô hình markov ẩn mô hình hóa chuỗi các âm vị tương ứng ta sẽ tìm ra được các tham số mel và tần số cơ bản f0. Từ các tham số mel xây dựng nên chuỗi các bộ lọc MLSA (Mel Log Spectral Approximation) và kết hợp với tín hiệu kích thích được tạo từ f0 sẽ tạo ra được tín hiệu tiếng nói[16], [17].

Quá trình huấn luyện dựa trên mô hình markov ẩn bao gồm các bước: Trích chọn đặc trưng tiếng nói và huấn luyện mô hình dựa trên các véc tơ đặc trưng trích được. Các đặc trưng tiếng nói được trích trong quá trình huấn luyện là các véc tơ như véc tơ hệ số mel và véc tơ mô tả f0. Nhưng đến đây việc mô hình hóa như vậy sẽ lại nảy sinh một vấn đề đó là tần số cơ bản f0 chỉ tồn tại ở âm hữu thanh còn các âm vô thanh lại là nhiễu. Do đó, để giải quyết vấn đề này người ta đã sử dụng một mô hình mở rộng hơn, đó là Multi-Space Probability Distribution Hidden Markov Model[16]. Mô hình này thường bao gồm: một không gian véc tơ được sử dụng để mô hình hóa véc tơ mel và hai không gian véc tơ để mô hình hóa tần số cơ bản f0. Mỗi không gian véc tơ trong mô hình thì được đặc trưng bởi một phân bố xác xuất, mỗi quan sát của một trạng thái lại được mô tả như sau: o=(X,x) trong đó X là tập các không gian véc tơ, còn x là véc tơ đặc trưng. Mục tiêu của quá trình huấn luyện là từ dữ liệu đầu vào cải thiện các tham số của mô hình markov ẩn mà mô hình hóa cho mỗi âm vị.

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 dựa trên mô hình markov ẩn là một hệ thống có khả năng tạo tiếng nói mang phong cách nói khác nhau, với đặc trưng của nhiều người nói khác nhau, thậm chí là mang 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 ghép nối, và có thể điều chỉnh tham số để thay đổi ngữ điệu. Tuy nhiên, một số nhược điểm của hệ thống này đó là độ tự nhiên trong tiếng nói tổng hợp của hệ thống bị suy giảm hơn so với tổng hợp ghép nối, phổ tín hiệu và tần số cơ bản được ước lượng từ các giá trị trung bình của các 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".




Hình 5: Quá trình huấn luyện và tổng hợp một hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn.

### 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 nó trong hệ thống. Như đã nói, 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 các hệ thống tổng hợp lai ghép thành hai loại 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 các mô hình markov ẩn.
- Âm thanh sau cùng được làm mịn bằng phương pháp làm mịn phổ.

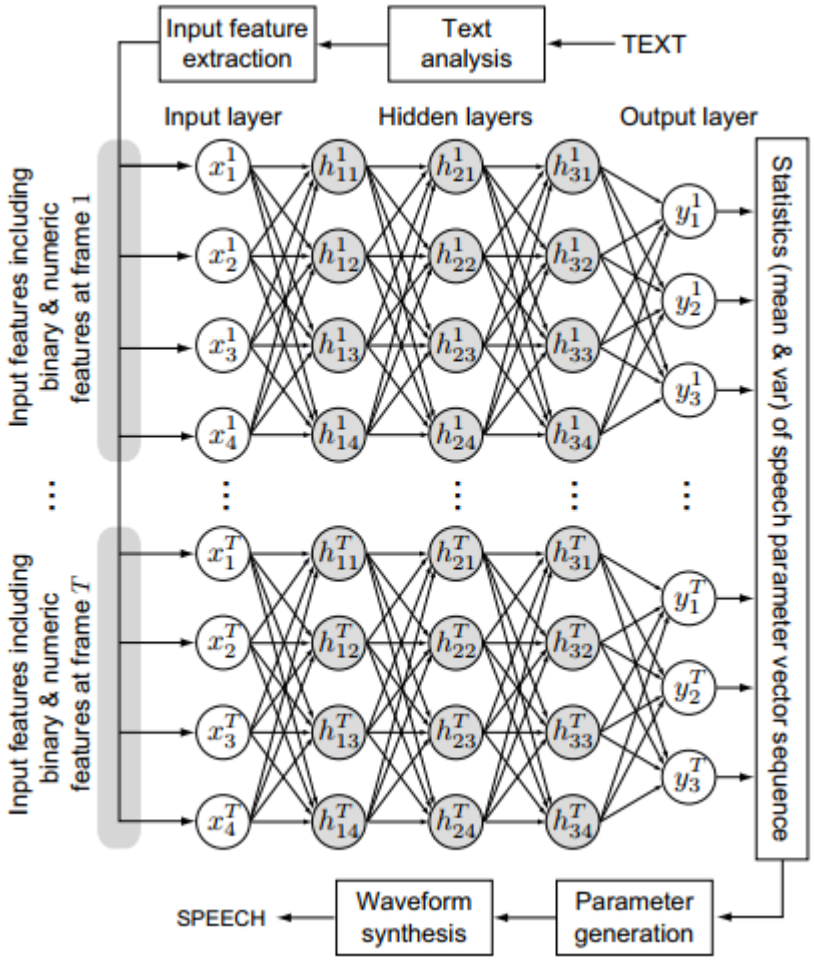
Khác với hệ thống tổng hợp hướng ghép nối, hệ thống tổng hợp hướng HMM sử dụng các thuật toán sinh tham số từ các HMM và phần tổng hợp ghép nối được sử dụng để tăng cường chất lượng chuỗi tham số này.

Hai hướng tổng hợp lai ghép nêu trên đều có ưu nhược điểm khác nhau, và được sử dụng tùy 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 còn tồn tại.

### 1.2.6 Tổng hợp tiếng nói dựa trên phương pháp học sâu (DNN)

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 này được xây dựng dựa trên việc mô hình hóa mô hình âm học bằng một mạng nơ ron học sâu DNN. Trong đó Văn bản đầu vào sẽ được chuyển hóa thành một véc tơ đặc trưng ngôn ngữ, các véc tơ đặc trưng này mang các thông tin về âm vị, ngữ cảnh xung quanh âm vị, thanh điệu,… Sau đó mô hình âm học dựa trên DNN 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 này sẽ tạo thành tín hiệu tiếng nói nhờ một bộ tổng hợp tín hiệu tiếng nói (thường là vocoder).

Kiến trúc tổng quan của một hệ thống tổng hợp tiếng nói dựa trên mạng nơ ron học sâu DNN được mô tả trong hình 6. Trong đó, văn bản cần được tổng hợp sẽ đi qua bộ phân tích văn bản (Text analysis) 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 các đặc trưng đầu ra $\{y_m^t\}$ thông qua một 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 này 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 (Waveform generation) để tạo ra tín hiệu tiếng nói thực.




Hình 6: Tổng hợp tiếng nói dựa trên DNN[18]

Mạng nơ ron học sâu DNN 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 DNN 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 DNN cũng có những hạn chế đó là vì sự mạnh mẽ của nó nên nó rất nhạy cảm với các thông tin sai lệch và không tốt như nhiễu, và nó cũng cần rất nhiều dữ liệu để huấn luyện mô hình. Nhờ sự mạnh mẽ trong mô hình hóa mô hình âm học, DNN đã được áp dụng trong nhiều ứng dụng tổng hợp tiếng nói trên thế giới như các sản phẩm của Google, Baidu, Microsoft hay trong hệ thống Merlin của CSTR đã đạt được độ tự nhiên rất cao.

| | | |
|---|---|---|
| HMM | 1 mix<br>2 mix | 3.537 ± 0.113<br>3.397 ± 0.115 |
| DNN | 4x1024<br>5x1024<br>6x1024<br>7x1024 | 3.635 ± 0.127<br>3.681 ± 0.109<br>3.652 ± 0.108<br>3.637 ± 0.129 |

Bảng 1: Đánh giá so sánh HMM và DNN

Kết quả đánh giá so sánh hệ thống tổng hợp tiếng nói dựa trên HMM so với DNN của Google[19] được thể hiện trong bảng 1. Đánh giá này sử dụng phương pháp

trung bình điểm ý kiến MOS trên thang điểm 5, với 173 câu kiểm tra chia theo 5 chủ đề, mỗi chủ đề khoảng 30 câu. Từ kết quả này cho thấy tổng hợp tiếng nói dựa trên DNN có chất lượng tốt hơn HMM.

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

Việt nam đang trong thời kỳ phát triển nhanh chóng của công nghệ thông tin. Điều đó cho phép chúng ta có những nền tảng khoa học kỹ thuật và nền tảng cơ sở vật chất để có thể nghiên cứu cũng như triển khai các ứng dụng về khoa học công nghệ trong cuộc sống. Trong nhiều năm trở lại đây, tổng hợp tiếng Việt đã có những thành tựu đáng kể, các hệ thống tổng hợp tiếng nói tiếng việt được ra đời như VietVoice[3], VnSpeech[4], Vais[5], Hệ thống tổng hợp tiếng nói của tập đoàn FPT hay hệ thống tổng hợp tiếng nói Hoa súng. Trong đó các hệ thống tổng hợp tiếng nói tiếng Việt 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[20], được phát triển lần đầu vào năm 2007, dữ liệu để xây dựng hệ thống này được gọi là VNSpeechCorpus, 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[21], 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[6] hay hệ thống tổng hợp tiếng nói tiếng Việt Mica TTS[7] (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.

Như đã nêu ở trên, hiện tại ở Việt Nam mới chỉ phát triển các hệ thống tổng hợp tiếng nói dựa trên những phương pháp đã cũ như tổng hợp ghép nối hay tổng hợp sử

[3] http://www.vietvoice.net/
[4] http://www.vnspeech.com
[5] https://vais.vn/
[6] https://speech.openfpt.vn/
[7] http://sontinh.mica.edu.vn/tts2

dụng tham số thống kê. Trong khi đó trên thế giới đã có những phương pháp mới cho tổng hợp tiếng nói được phát triển và đạt được kết quả cao, điển hình là tổng hợp dựa trên mạng nơ ron học sâu DNN, ví dụ như hệ thống tổng hợp tiếng nói của CSTR[22] hay các sản phẩm của Google, Baidu,… Do đó lý do để lựa chọn mô hình mạng nơ ron học sâu (DNN) trong việc xây dựng hệ thống tổng hợp tiếng nói tiếng Việt là để:

- Thử nghiệm kỹ thuật mới, hiện đại và phổ biến trên thế giới hiện nay nhằm so sánh với các công nghệ tổng hợp tiếng nói tiếng Việt hiện có.
- Tìm hiểu các vấn đề có thể xảy ra khi sử dụng DNN cho tổng hợp tiếng Việt và đưa ra những cách khắc phục.

# CHƯƠNG 2: PHƯƠNG PHÁP HỌC SÂU ÁP DỤNG TRONG TỔNG HỢP TIẾNG NÓI

## 2.1 Kỹ thuật học sâu sử dụng mạng nơ ron nhân tạo

Học sâu là một nhánh của lĩnh vực học máy, dựa trên một tập hợp các thuật toán nhằm cố gắng mô hình hóa dữ liệu trừu tượng ở mức cao nhất bằng cách sử dụng nhiều lớp xử lý với cấu trúc phức tạp, hoặc bao gồm nhiều biến đổi phi tuyến[23]. Chương này sẽ chủ yếu trình bầy về hướng tiếp cận "kỹ thuật học sâu sử dụng mạng nơ ron nhân tạo" hay chính là tìm hiểu về "mạng nơ ron học sâu", vì nó là phương pháp được áp dụng cho việc xây dựng hệ thống tổng hợp tiếng nói tiếng việt của đề tài.

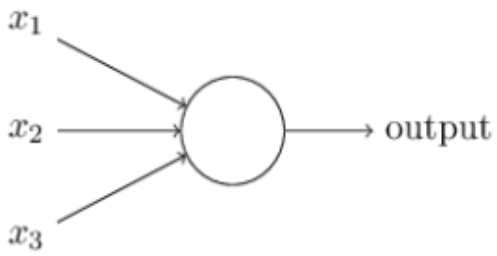
Trước khi đi vào mạng nơ ron học sâu, ta sẽ xem xét sơ lược về mạng nơ ron nhân tạo. Mạng nơ ron nhân tạo là một mô hình toán học được xây dựng dựa trên ý tưởng của các mạng nơ ron sinh học trong bộ não của con người. Nó gồm một nhóm các nơ ron nhân tạo (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối, sau đó tính giá trị mới tại các nút. Để hiểu rõ hơn chúng ta sẽ xem xét tìm hiểu về hai loại nơ ron nhân tạo cơ bản là perceptron, sigmoid và kiến trúc mạng nơ ron cơ bản.

### 2.1.1 Những mạng nơ ron cơ bản

#### 2.1.1.1 Perceptron

Perceptron bắt đầu được phát triển vào những năm 1950 và 1960 bởi Frank Rosenblatt, ngày nay nó phổ biến trong nhiều mô hình mạng nơ ron khác nhau và nhiều công trình hiện đại về mạng nơ ron[24].

Perceptron nhận một số đầu vào nhị phân: x1, x2,… tạo ra một đầu ra nhị phân duy nhất:

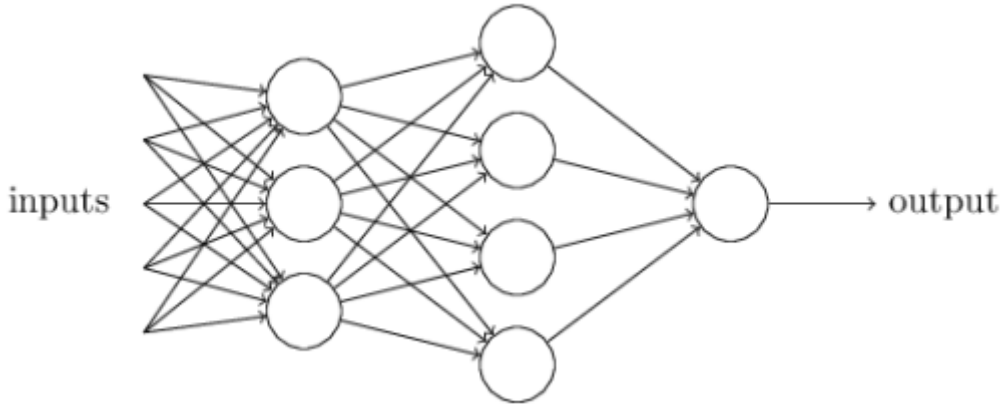



Hình 7: Một perceptron với ba đầu vào[24]

Trong hình 7 thể hiện một perceptron với ba đầu vào x1,x2,x3 và một đầu ra output (trong thực tế có thể có số lượng đầu vào khác). Rosenblatt đề xuất một quy tắc đơn giản để tính toán đầu ra, ông ấy giới thiệu các trọng số w1,w2,… thể hiện tầm quan trọng của các yếu tố đầu vào với đầu ra tương ứng. Đầu ra của nơ ron, 0 hoặc 1, được xác định bằng cách xem xét tổng $\sum_i w_i x_i$ nhỏ hơn hoặc lớn hơn một ngưỡng nhất định. Cũng như các trọng số, ngưỡng là số thực và là tham số của nơ ron. Khi đó đầu ra được tính như sau:

$$output = \begin{cases} 0 \text{ if } \sum_j w_j x_j <= threshold \\ 1 \text{ if } \sum_j w_j x_j > threshold \end{cases} \quad (2.1.1.1)$$

Bằng cách thay đổi trọng số và ngưỡng, chúng ta có thể có được các mô hình khác nhau. Nhưng một perceptron không phải là một mô hình tối ưu, do đó một mạng lưới perceptron có thể đưa ra quyết định chính xác hơn:




Hình 8: Mạng nơ ron gồm nhiều perceptron[24]

Trong mạng nơ ron hình 8, lớp đầu tiên gồm ba perceptron đưa ra quyết định từ ba đầu vào, lớp thứ hai gồm bốn perceptron đưa ra quyết định từ đầu vào là đầu ra của lớp đầu tiên, mỗi perceptron của lớp này cũng có ba đầu vào. Lớp perceptron thứ hai có thể đưa ra quyết định phức tạp và trừu tượng hơn lớp đầu tiên. Và thậm chí quyết định phức tạp hơn có thể được thực hiện bởi các perceptron trong lớp thứ ba, thứ tư.... Bằng cách này, một mạng lưới nhiều lớp của perceptron có thể tham gia vào việc ra quyết định phức tạp.

Perceptron và mạng perceptron cho thấy rằng sự điều chỉnh hay sự học có thể xẩy ra khi phản ứng với các kích thích mà không cần sự can thiệp trực tiếp của một lập trình viên. Các thuật toán học cho phép chúng ta sử dụng mạng nơ ron nhân tạo theo các hoàn toàn khác với các cổng logic thông thường. Mạng nơ ron có thể học và giải quyết vấn đề một cách đơn giản trong khi vấn đề đó lại vô cùng khó khăn đối với mạng thông thường.

### 2.1.1.2 Nơ ron Sigmoid

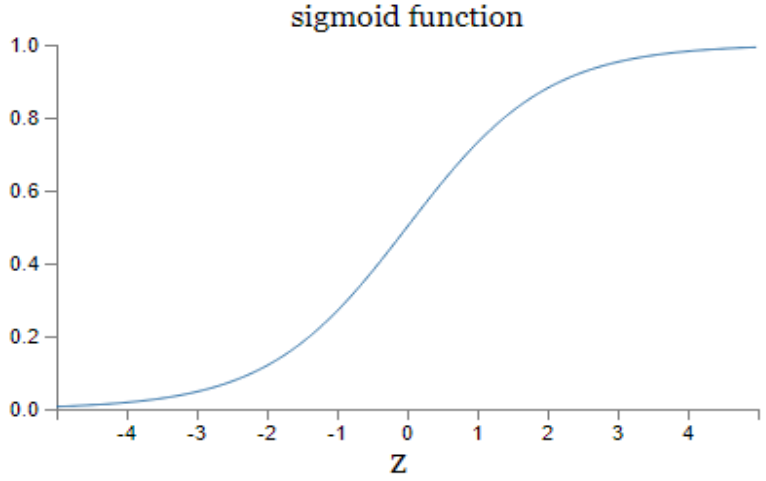
Với Perceptron, một chút thay đổi trọng số của bất kỳ perceptron trong một mạng cũng có thể dẫn đến kết quả hoàn toàn thay đổi. Tuy nhiên, trong thực tế đôi khi chỉ cần một thay đổi nhỏ ở trọng số để cho ra kết quả tốt hơn, do đó để khắc phục vấn đề của perceptron ta sử dụng nơ ron nhân tạo được gọi là sigmoid. Cũng giống như perceptron, các nơ-ron sigmoid có đầu vào, x1, x2,... Nhưng thay vì đầu vào chỉ có 0 hoặc 1 thì nó có thể là bất cứ giá trị nào trong khoảng 0 1 . Ví dụ, 0,638 là một đầu vào có giá trị trong một nơ-ron sigmoid. Các nơ-ron sigmoid cũng có trọng số cho mỗi đầu vào là w1, w2 …và định hướng (bias) b. Thêm nữa, đầu ra cũng không phải là 0 hoặc 1. Thay vào đó, đầu ra là σ(w·x + b), trong đó σ được gọi là hàm sigmoid và được xác định bằng:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.1.2.1)$$

Một nơ ron sigmoid với đầu vào x1, x2,… trọng số w1, w2,… khi đó bias b là:

$$b = \frac{1}{1 + \exp(-\sum_i w_i x_i - b)} \quad (2.1.2.2)$$

Đồ thị hàm sigmoid được biểu diễn trên hình 9:

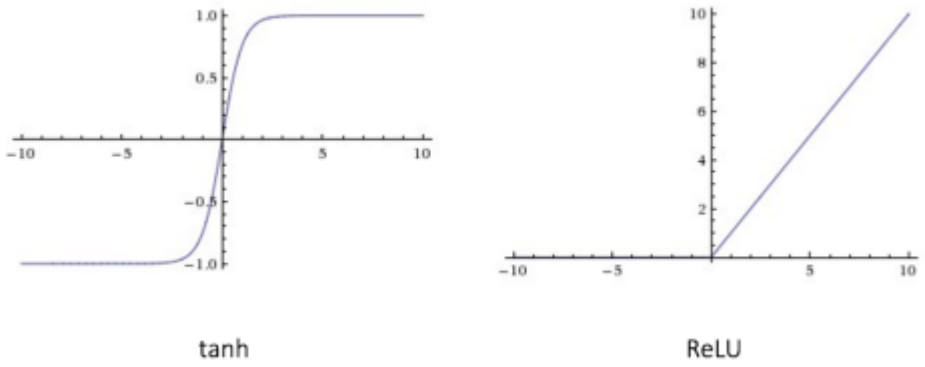



Hình 9: Hàm sigmoid[24]

Ngoài hàm sigmoid trong nơ ron sigmoid còn nhiều hàm kích hoạt khác trong các nơ ron nhân tạo như hàm tanh (công thức 2.1.2.3) và hàm Relu (công thức 2.1.2.4). Đồ thị hàm relu và tanh được biểu diễn trên hình 10.

$$\tanh(x) = 2\sigma(2x) - 1 \quad (2.1.2.3)$$

$$f(x) = max(0, x) \quad (2.1.2.4)$$

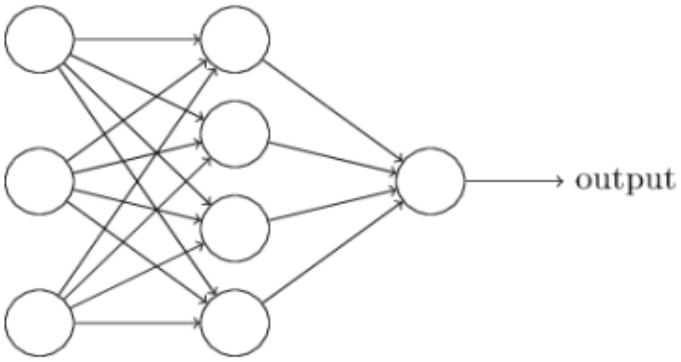



Hình 10: Hàm kích hoạt tanh và relu[8]

## 2.1.2 Mạng nơ ron học sâu

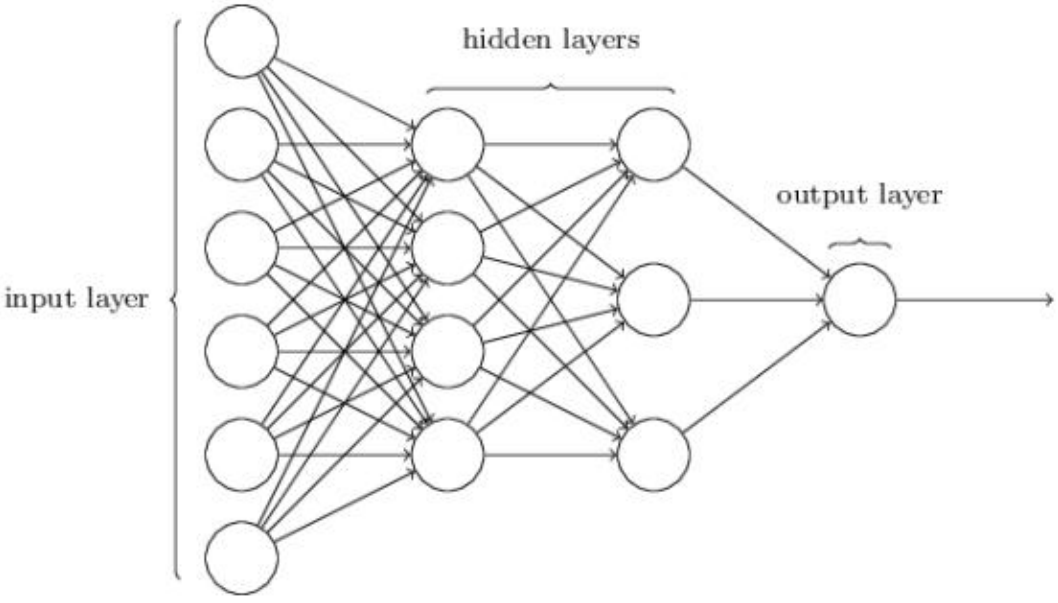
Trước khi xem xét thế nào là mạng nơ ron học sâu, ta xem xét qua một mạng nơ ron cơ bản như trên hình 11.

[8] https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/




Hình 11: Mạng nơ ron một lớp ẩn [24]

Đây là mạng nơ ron với duy nhất một lớp ẩn, lớp ngoài cùng bên trái gọi là lớp đầu vào và các nơ ron trong lớp này được gọi là nơ ron đầu vào, đây cũng chính là nơi nhận đầu vào của mạng nơ ron. Lớp ngoài cùng bên phải là lớp đầu ra (output), lớp này trả về giá trị đầu ra tương ứng với những đầu vào được nhận từ lớp đầu vào. Lớp ở giữa được gọi là lớp ẩn, lớp này không nhận đầu vào cũng như đầu ra, mạng trên có duy nhất một lớp ẩn nhưng các mạng khác có thể có nhiều lớp ẩn. Hình 12 là một mạng nơ ron với hai lớp ẩn:



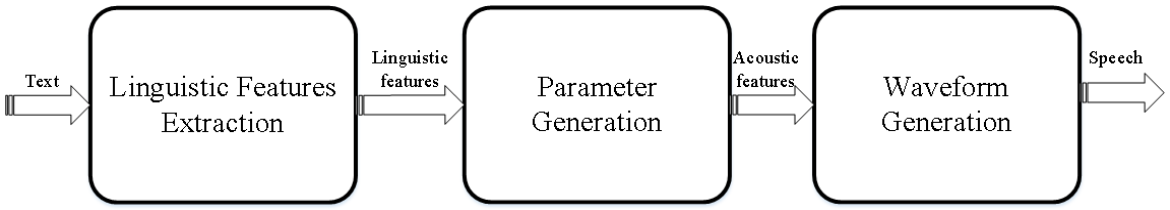

Hình 12: Mạng nơ ron hai lớp ẩn[24]

Trong khi việc thiết lập đầu vào và đầu ra của một mạng nơ ron thường đơn giản thì việc tạo ra các lớp ẩn tốn nhiều công sức, với mỗi mô hình mạng khác nhau và các kiến trúc với những lớp ẩn khác nhau được tạo ra để đáp ứng những yêu cầu phù hợp. Do đó việc thiết kế các lớp ẩn là cực kỳ quan trọng để tạo được những đầu ra theo hướng mong muốn. Các nơ ron trong mạng cũng rất đa dạng có thể là perceptron, có thể là sigmoid hoặc cũng có thể là nhiều loại nơ ron khác như tanh, relu,… tùy theo yêu cầu bài toán mà hình thành các lớp ẩn với kiến trúc khác nhau và nơ ron khác nhau.

Một mạng nơ ron nhiều lớp ẩn, hay có số lớp ẩn lớn hơn hai dược gọi là mạng nơ ron học sâu DNN (deep neural network). Với những mạng nơ ron học sâu, chúng có ưu điểm là có thể được sử dụng để xây dựng một hệ thống các khái niệm phức tạp[24].

## 2.2 Tổng hợp tiếng nói dựa trên phương pháp 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 ký tự đầu vào (thường là chuỗi âm vị) thông qua việc tạo trực tiếp các đặc trưng âm học của tiếng nói[25]. Tuy nhiên những mô hình kiểu này có những giới hạn trọng việc biểu diễn mối quan hệ phức tạp và phi tuyến giữa chuỗi ký tự đầu vào và các đặc trưng âm học[25]. Trong hướng tiếp cận này, mạng nơ ron học sâu (DNN) sẽ được sử dụng để mô hình hóa mối quan hệ giữa chuỗi ký tự đầu vào và các đặc trưng âm học ở đầu ra, việc sử dụng DNN có thể giải quyết một số giới hạn của những phương pháp thông thường (như HMM hoặc GMM)[18]. Hình 13 mô tả một kiến trúc cơ bản của một hệ thống tổng hợp tiếng nói dựa trên phương pháp học sâu.




Hình 13: Kiến trúc cơ bản của hệ thống tổng hợp tiếng nói.

Dựa trên kiến trúc của hệ thống tổng hợp tiếng nói trên hình 13, có thể thấy rằng một hệ thống tổng hợp tiếng nói gồm ba mô đun chính và đây cũng là ba mô đun trong tổng hợp tiếng nói dựa trên công nghệ học sâu:

- 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 học, 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,…
- Bộ Parameter Generation có nhiệm vụ chuyển hóa các đặc trưng ngôn ngữ ở đầu vào thành thành các đặc trưng âm học tương ứng, trong trường hợp hệ thống tổng hợp tiếng nói được xây dựng dựa trên phương pháp học sâu, thì bộ này sử dụng mạng nơ ron học sâu DNN để mô hình hóa các mô hình.
- Mô đun tạo tín hiệu tiếng nói: Các đặc trưng âm học sẽ được chuyển hóa thành tín hiệu tiếng nói nhờ bộ Waveform Generation.

Chi tiết từng mô đun trong hình 10 sẽ được trình bầy lần lượt ở các chương sau, trong đó vocoder sẽ làm nhiệm vụ tạo tín hiệu tiếng nói, hay đó chính là bộ Waveform Generation. Còn mô hình âm học chính là phần lõi chính cho mô đun Parameter Generation.

## 2.3 Trích chọn các đặc trưng ngôn ngữ

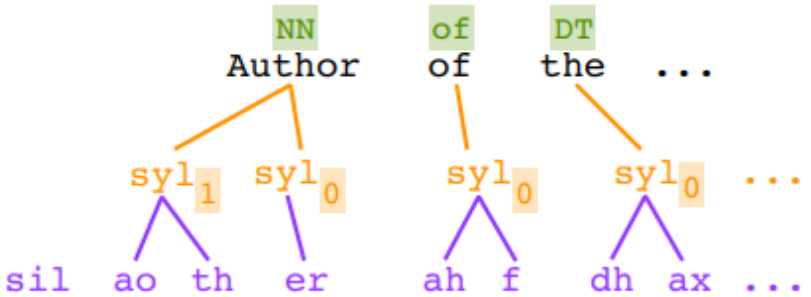
Đặc trưng ngôn ngữ học của văn bản được sử dụng làm đầu vào cho mô hình âm học bao gồm các thông tin như: âm vị hiện tại, vị trí của âm vị trong câu, cụm từ, vị trí từ trong câu, số lượng âm vị trong từ hay thanh điệu hiện tại là gì,… Các thông

tin này cũng được phân theo các mức: Mức âm vị, mức âm tiết, mức từ, mức cụm từ, mức câu[26]. Để trích chọn được các đặc trưng ngôn ngữ học nên trên, văn bản đầu vào sẽ được xử lý thông qua một quy trình như sau:

- Văn bản đầu vào sẽ được chuyển thành một chuỗi âm vị nhờ vào một 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 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[27].
- 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.

Đầu ra đặc trưng ngôn ngữ của quá trình này bao gồm những thông tin như sau:

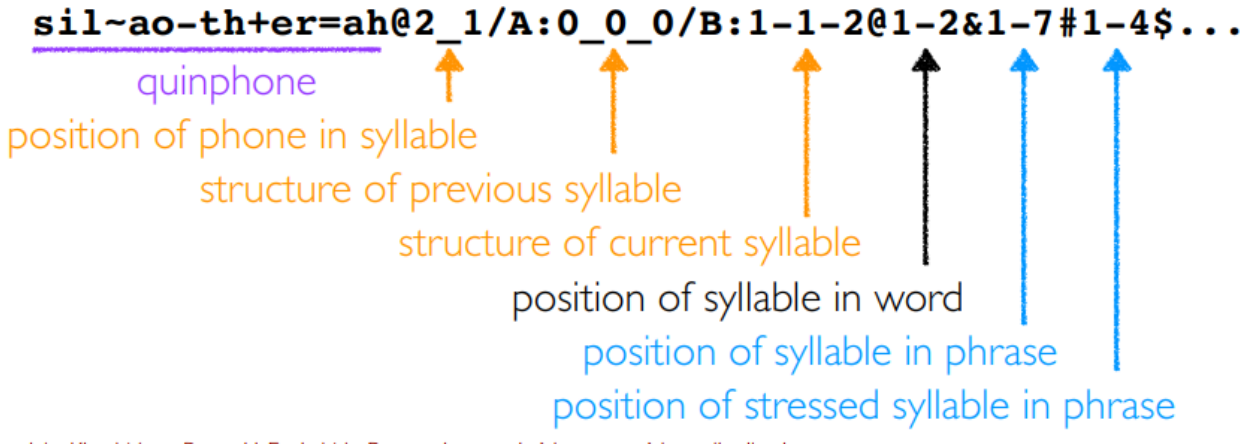
- Thông tin mức âm vị: thông tin mức âm vị bao gồm có các âm vị hiện tại, phía trước, phía sau, thông tin về vị trí các âm vị trên trong âm tiết, từ, cụm từ,…
- Thông tin mức âm tiết: gồm có thông tin về thanh điệu và số lượng âm vị của các âm tiết hiện tại, phía trước, phía sau. Vị trí của âm tiết trong từ,…
- Thông tin mức từ: bao gồm các thông tin về nhãn từ loại, số lượng âm tiết của từ hiện tại, phía trước, phía sau,…
- Thông tin mức cụm từ: Số lượng các từ, âm tiết trong cụm hiện tại, phía trước, phía sau.
- Thông tin mức câu: bao gồm các thông tin về số lượng từ, số lượng âm tiết, số lượng cụm từ trong câu.




```
sil~sil-sil+ao=th@x_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$...
sil~sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
sil~ao-th+er=ah@2_1/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
ao~th-er+ah=v@1_1/A:1_1_2/B:0-0-1@2-1&2-6#1-4$...
th~er-ah+v=dh@1_2/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...
er~ah-v+dh=ax@2_1/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...
ah~v-dh+ax=d@1_2/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
v~dh-ax+d=ey@2_1/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
```

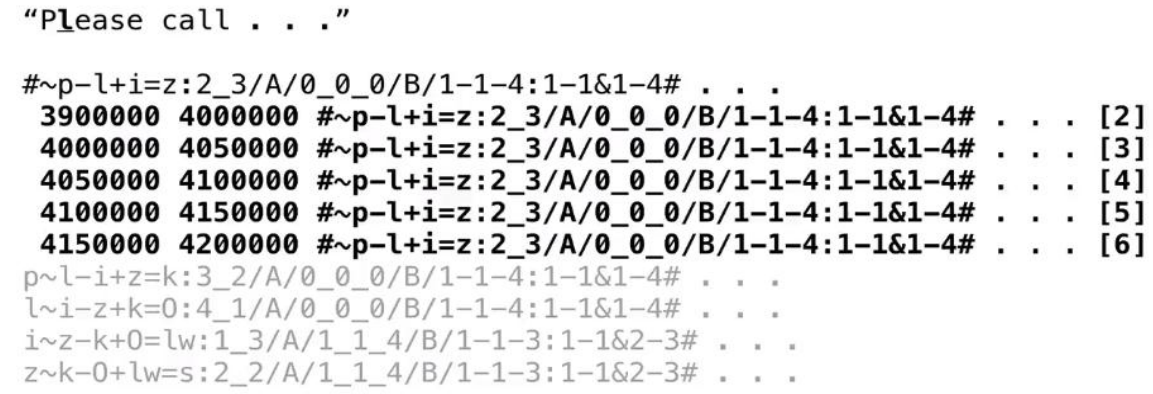
Hình 14: Biểu diễn đặc trưng ngôn ngữ học của văn bản[28]

Kết quả đầu ra của quá trình trích chọn các đặc trưng âm học được thể hiện trong hình 14, trong đó văn bản đầu vào được phân tích thành một chuỗi âm vị, mỗi âm vị tương ứng bởi một dòng có chứa các thông tin đặc trưng ngôn ngữ ở phía dưới. Chi tiết nội dung của từng dòng được mô tả trong phụ lục A, và được biểu diễn trên hình 15. Ở đây cần lưu ý một chút, có sự khác biệt về cấu trúc cho mỗi dòng trong phụ lục A và ở hình 15, điều này xảy ra là vì cấu trúc mỗi dòng ở phụ lục A đặc trưng cho tiếng việt còn ở hình 15 là đặc trưng cho tiếng Anh, do đó với mỗi ngôn ngữ khác nhau thì cấu trúc mỗi dòng tương ứng mỗi âm vị cũng khác nhau. Nhưng điểm chung của chúng là đều thể hiện các thông tin như: Vị trí của âm vị trong âm tiết, cấu trúc của âm tiết phía trước, cấu trúc âm tiết phía sau, vị trí của âm tiết trong từ, vị trí của âm tiết trong cụm từ, vân vân… Và đó cũng chính là các thông tin đặc trưng ngôn ngữ mà ta cần.




Hình 15: Thông tin đặc trưng ngôn ngữ liên quan đến từng âm vị[28]

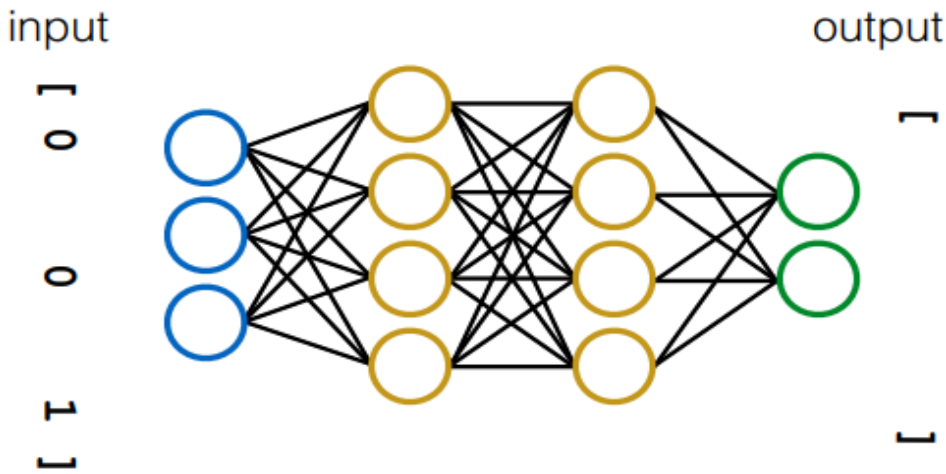
Mặc dù đã hoàn thành trích chọn đặc trưng ngôn ngữ, nhưng những thông tin trích chọn được vẫn là chưa đủ cho huấn luyện các mô hình tiếp theo (mô hình âm học và mô hình thời gian) của hệ thống tổng hợp tiếng nói. Một thông tin cực kỳ quan trọng và cần thiết nữa cần được thêm vào, đó là thời gian xuất hiện của mỗi âm vị trong câu nói. Để lấy được thông tin về thời gian tương ứng mỗi âm vị, ta sử dụng mô hình markov ẩn, quá trình này được gọi là force alignment[4], [27]. Kết quả của quá trình forced alignment sẽ cho ra khoảng thời gian xuất hiện của mỗi trạng thái trong mỗi âm vị. Hình 16 minh họa thời gian cho từng trạng thái trong mỗi âm vị (thông thường sử dụng 5 trạng thái theo mô hình markov ẩn).




Hình 16: Thời gian xuất hiện mỗi trạng thái của từng âm vị

## 2.4 Mô hình âm học dựa trên mạng nơ ron học sâu

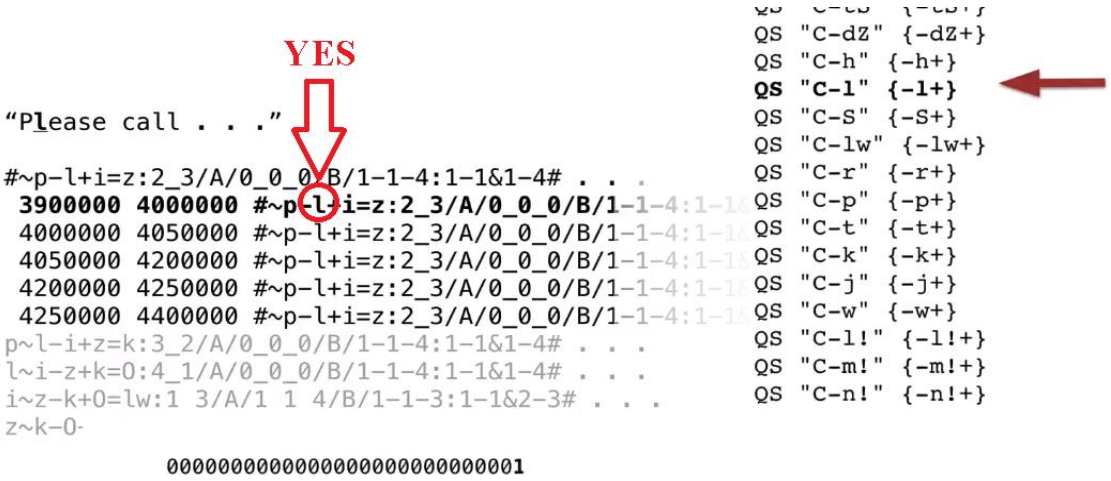
Trong tổng hợp tiếng nói dựa trên phương pháp học sâu, mô hình âm học được mô hình hóa bằng một mạng nơ ron học sâu như hình 17, trong đó đầu vào của mạng này là một véc tơ đặc trưng ngôn ngữ học và đầu ra là các đặc trưng âm học hay chính là các tham số của vocoder (sẽ trình bày chi tiết về vocoder ở phần sau) và được sử dụng làm đầu vào cho vocoder trong quá trình tổng hợp tiếng nói.




Hình 17: Mạng nơ ron feat forward.

Như đã nói ở trên, đầu vào của mạng nơ ron là một véc tơ đặc trưng ngôn ngữ học, véc tơ này được chuyển hóa từ các đặc trưng ngôn ngữ học mà ta trích chọn được trong phần 2.3. Có nhiều cách khác nhau để chuyển hóa các thông tin đặc trưng ngôn ngữ học thành một véc tơ đầu vào cho một mạng nơ ron học sâu, một trong số đó là sử dụng một tệp các câu hỏi. Các câu hỏi này được dùng để khai phá các thông tin mà các đặc trưng ngôn ngữ đem lại, nội dung của các câu hỏi 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 này, ta tìm được véc tơ nhị phân biểu 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 18 và theo một quy trình như sau:

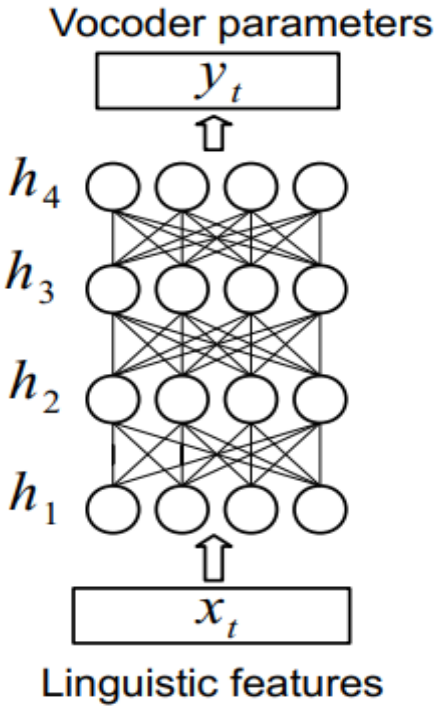
- Đưa từng dòng chứa các thông tin đặc trưng ngôn ngữ tương ứng với từng âm vị, vào trả lời chuỗi các câu hỏi.
- Với mỗi câu trả lời đúng thì được giá trị là 1 và trả lời sai giá trị là 0 (như trên hình ứng với câu hỏi âm vị hiện tại là "l" thì đúng âm vị hiện tại trong dòng cũng là "l" nên kết quả nhận được là 1).
- Trả lời hết chuỗi các câu hỏi ta được một véc tơ nhị phân làm đầu vào cho mạng nơ ron




Hình 18: Chuyển hóa véc tơ đặc trưng thành các véc tơ nhị phân.

Đầu ra của mạng nơ ron là các véc tơ đặc trưng âm học chứa các tham số vocoder, các véc tơ này chính 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 pha huấn luyện mô hình âm học, thì các véc tơ đặc trưng âm học này của dữ liệu huấn luyện được trích chọn trong quá trình phân tích đặc trưng âm học của vocoder.

Trong việc mô hình hóa mô hình âm học thì mạng nơ ron feat forward là mạng được sử dụng phổ biến, hình 19 minh họa cho một mạng nơ ron feat forward (mạng DNN) áp dụng trong tổng hợp tiếng nói. Trong đó mạng nơ ron lấy các véc tơ nhị phân mang thông tin đặc trưng ngôn ngữ làm đầu vào và đầu ra chính là tham số vocoder nói trên.




Hình 19: Mạng nơ ron học sâu áp dụng trong tổng hợp tiếng nói[4].

Mạng nơ ron feat forward là một mạng đơn giản, với đủ các lớp thì nó còn được gọi là mạng nơ ron học sâu. 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 sau:

$$h_t = H(W^{xh} x_t + b^h) \quad (2.4.1)$$

$$y_t = W^{hy} h_t + b^y \quad (2.4.2)$$

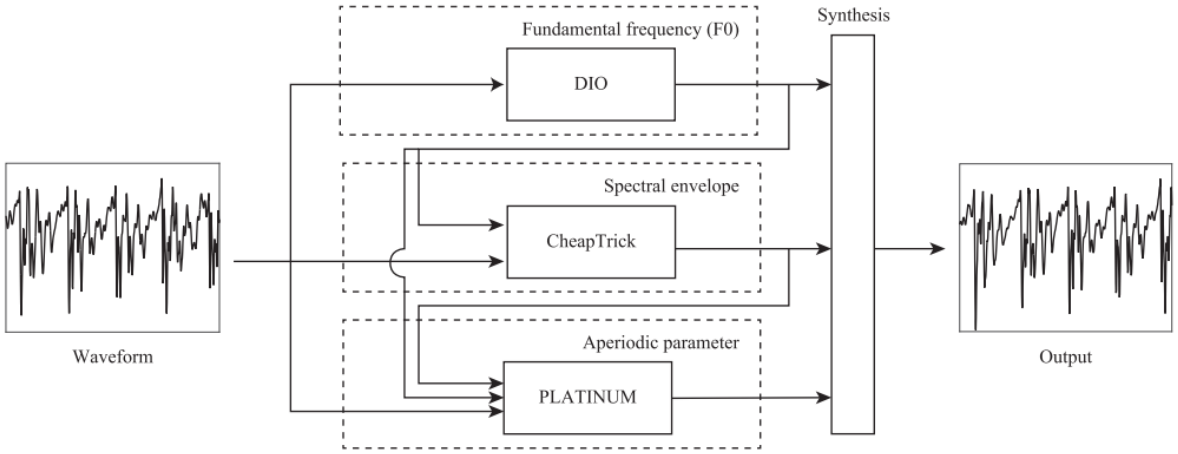
Trong đó H() là hàm kích hoạt phi tuyến (thường là hàm tanh), $W^{xh}$ và $W^{hy}$ là ma trận trọng số, $b^h$ và $b^y$ là các véc tơ bổ sung (bias vector), $W^{hy}h_t$ là thành phần hồi quy tuyến tính để dự đoán đặc trưng đích từ hàm kích hoạt trong lớp ẩn trước.

## 2.5 Vocoder

Vocoder là một 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 và tổng hợp tiếng nói. Trong quá trình huấn luyện, vocoder được sử dụng để phân tích dữ liệu âm thanh thành các đặc trưng âm học, 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 hiệu tiếng nói.

Có rất nhiều loại vocoder khác nhau được phát triển để cải thiện chất lượng phân tích và tổng hợp tiếng nói như Straight vocoder[29], World vocoder[30], Magphase vocoder[31],… Trong phần này sẽ chỉ trình bày về một vocoder vô cùng mạnh mẽ, được phát triển để cải thiện chất lượng âm thanh trong những ứng dụng thời gian thực và cũng được sử dụng để xây dựng hệ thống tổng hợp tiếng nói trong luận văn này, đó là WORLD vocoder.

Như đã nói ở trên, WORLD vocoder được sử dụng để trích chọn các đặc trưng âm học và tổng hợp tiếng nói từ những đặc trưng này. Các đặc trưng âm học mà WORLD vocoder trích chọn được 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 bởi phương pháp DIO[32], đường bao phổ được ước lượng bởi phương pháp CheapTrick[33], và tín hiệu kích được ước lượng bởi phương pháp PLATINUM[34], nó được sử dụng như một tham số không tuần hoàn. Hình 20 mô tả quá trình xử lý của WORLD vocoder trong hai giai đoạn phân tích và tổng hợp tín hiệu tiếng nói.




Hình 20: Tổng quan về hệ thống WORLD vocoder[30].

Để ước lượng tần số cơ bản F0 bằng phương pháp DIO ta trải qua các bước sau:

- 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 ra.
- Chọn ra ứng viên nào có độ tin cậy cao nhất làm F0.

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

$$\int_0^{3T_0} (y(t)\,\mathrm{w}(t))^2\,dt = 1.125\int_0^{T_0} y^2(t)\,dt \quad (2.5.1)$$

Trong đó y(t) là tín hiệu và w(t) là hàm cửa sổ. Sau khi tính được năng nượng phổ nêu trên, chúng được làm mịn với một cửa sổ chữ nhật có độ dài $2w_0/3$, như sau:

$$P_s(\mathrm{w}) = \frac{3}{2w_0}\int_{-\frac{w_0}{3}}^{\frac{w_0}{3}} P(\mathrm{w}+\lambda)\,\mathrm{d}\lambda \quad (2.5.2)$$

Với $w_0$ là $2\pi / T_0$ .

Và cuối cùng, đường bao phổ $P_l(w)$ được tính như sau:

$$P_l(\mathrm{w}) = \exp(F[l_s(\tau)l_q(\tau)p_s(\tau)]) \quad (2.5.3)$$

$$l_s(\tau) = \frac{\sin(\pi f_0\tau)}{\pi f_0\tau} \quad (2.5.4)$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1\cos(\frac{2\pi\tau}{T_0}) \quad (2.5.5)$$

$$p_s(\tau) = F^{-1}[\log(P_s(w))] \quad (2.5.6)$$

Trong đó $l_s(\tau)$ biểu diễn làm nâng cho việc làm mịn, $l_q(\tau)$ biểu diễn hàm nâng cho việc phục hồi phổ, và $\tilde{q}_0$ và $\tilde{q}_1$ là các tham số cho việc phục hồi phổ. Các ký hiệu F[] và $F^{-1}$[] đại diện cho biến đổi fourier và biến đổi fourier ngược.

Cuối cùng ta xem xét đến phương pháp PLATINUM để ước lượng tín hiệu kích thích. Đầu tiên ta cho tín hiệu 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 $S_m(w)$. $S_m(w)$ được tính theo biểu thức sau:

$$S_m(\mathrm{w}) = \exp(F[c_m(\tau)]) \quad (2.5.7)$$

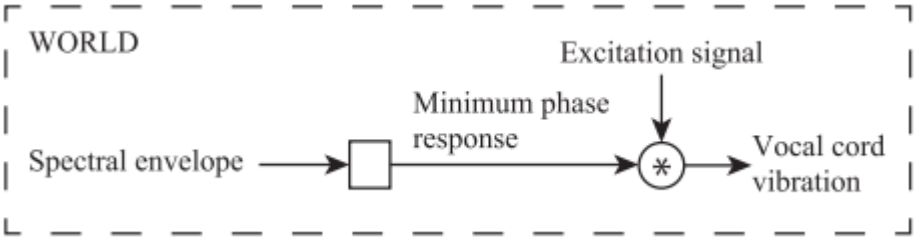
$$c_m(\tau) = \begin{cases} 2c(\tau) \leftrightarrow (\tau > 0) \\ c(\tau) \leftrightarrow (\tau = 0) \\ 0 \leftrightarrow (\tau < 0) \end{cases} \quad (2.5.8)$$

Tín hiệu kích thích được biểu diễn như sau:

$$x_p(t) = F^{-1}[X_p(\mathrm{w})] \quad (2.5.9)$$

$$X_p(\mathrm{w}) = \frac{X_p(\mathrm{w})}{S_m(\mathrm{w})} \quad (2.5.10)$$

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 21.



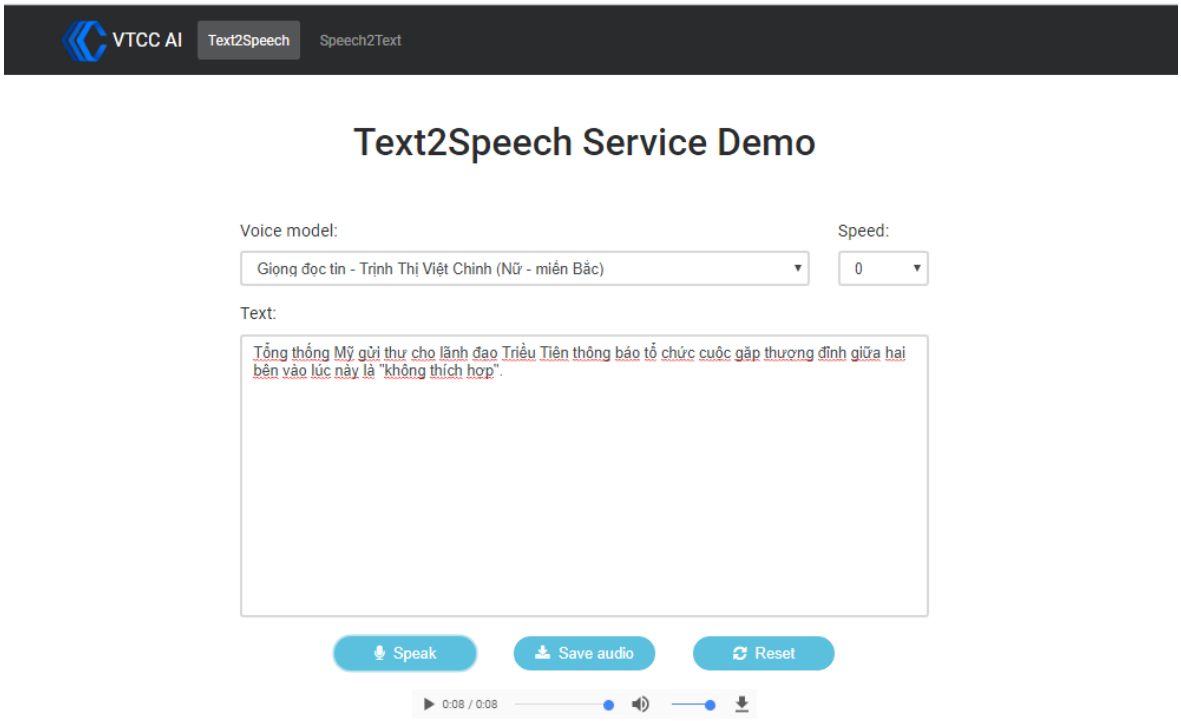

Hình 21: Tổng hợp tiếng nói với WORLD vocoder

Trong tổng hợp tiếng nói dựa trên phương pháp học sâu, vocoder được sử dụng trong hai quá trình. Vocoder được dùng để trích chọn (hay phân tích) các đặc trưng âm học của dữ liệu dùng trong quá trình huấn luyện của mô hình âm học, và nó còn được sử dụng để tổng hợp tín hiệu tiếng nói từ những đặc trưng âm học được sinh ra bởi mô hình âm học trong quá trình tổng hợp.

# CHƯƠNG 3: XÂY DỰNG HỆ THỐNG TỔNG HỢP TIẾNG NÓI TIẾNG VIỆT VỚI CÔNG NGHỆ HỌC SÂU

## 3.1 Giới thiệu hệ thống Viettel TTS

Hệ thống tổng hợp tiếng nói Viettel TTS[9] là một phần trong kế hoạch xây dựng một hệ thống chăm sóc khách hàng tự động trên tổng đài Viettel, hệ thống này đã bắt đầu được xây dựng được khoảng chưa đầy hai năm. Những ngày đầu tiên, hệ thống được xây dựng dựa trên công nghệ tổng hợp ghép nối, sau đó phát triển qua nhiều phiên bản theo phương pháp tổng hợp thống kê HMM và sử dụng nhiều công cụ khác nhau như MarryTTS[10] hay HTS[11]. Cho đến hiện nay hệ thống được phát triển theo phương pháp tổng hợp tiếng nói dựa trên công nghệ học sâu. Kiến trúc chung của hệ thống dựa trên công nghệ học sâu này được biểu diễn trên hình 22. Trong đó toàn bộ bốn mô đun được phát triển trong thời gian thực hiện luận văn này. Một số hình ảnh thử nghiệm của sản phẩm tổng hợp tiếng nói Viettel TTS như trên hình 22.

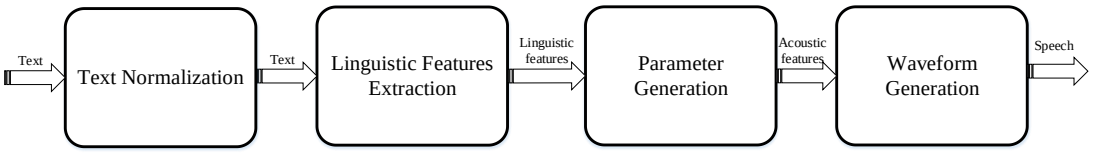



Hình 22: Hệ thống tổng hợp tiếng nói Viettel TTS

## 3.2 Kiến trúc tổng quan của hệ thống Viettel TTS

Dựa trên kiến trúc cơ bản của hệ thống tổng hợp tiếng nói áp dụng mạng nơ ron học sâu, chúng tôi đề xuất một hệ thống tổng hợp tiếng nói gồm có bốn mô đun được mô tả trên hình 23.

[9] http://vtcc.vn/production/vtcc-ai/
[10] http://mary.dfki.de/
[11] http://hts.sp.nitech.ac.jp/




Hình 23: Kiến trúc hệ thống tổng hợp tiếng nói.

Hệ thống tổng hợp tiếng nói trong hình 23 được xây dựng với bốn mô đun tương ứng như sau:

- Text normalization: Mô đun chuẩn hóa văn bản đầu vào, mô đun này nhận đầu vào là văn bản thô sau đó chuyển hóa nó thành văn bản có thể đọc được như là: chuyển các từ viết tắt thành chuỗi các từ, chuyển số thành chữ, chuyển các từ tiếng anh sang dạng phiên âm tiếng việt,…
- Linguistic Feature Extraction: Mô đun trích chọn đặc trưng ngôn ngữ, mô đun này trích chọn các đặc trưng âm học của văn bản sau khi chuẩn hóa, và đầu ra của mô đun là các tệp mà chứa các dòng có định dạng như trong phụ lục A, những dòng này mang thông tin về bối cảnh của mỗi âm vị.
- Parameter Generation: Mô đun tạo tham số, mô đun này có thành phần chính là mô hình âm học, nhận đầu vào là các đặc trưng âm học được lưu trong các tệp nhãn được tạo ra bởi "Linguistic Feature Extraction" và tạo ra các tham số đặc trưng âm học ở đầu ra.
- Waveform Generation: đây là mô đun tổng hợp tiếng nói từ các đặc trưng âm học đầu vào.

Các công việc thực hiện trong luận văn này nằm trong khuôn khổ dự án xây dựng hệ thống tổng hợp tiếng nói tiếng Việt Viettel TTS. Công việc cụ thể trong đồ án như sau:

- Đề xuất và triển khai hệ thống tổng hợp tiếng nói theo công nghệ học sâu, sử dụng công cụ Merlin[12].
- Triển khai xây dựng các mô đun Linguistic Feature Extraction, Parameter Generation, Waveform Generation.
- Xử lý dữ liệu sẵn có của Viettel (bộ dữ liệu này được thu thập từ mạng internet).
- Huấn luyện các mô hình học máy.
- Đánh giá kết quả

Các phần tiếp theo sẽ trình bày chi tiết các công việc trên.

## 3.3 Xây dựng các mô đun của hệ thống tổng hợp tiếng nói

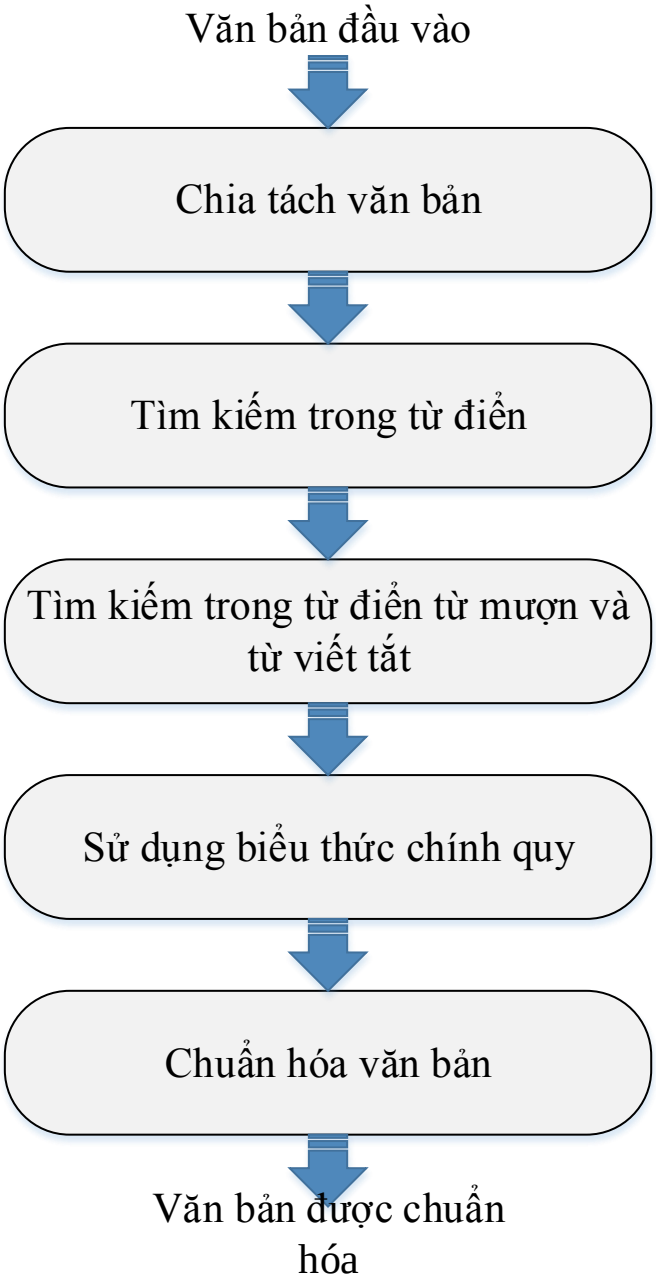
### 3.3.1 Mô đun chuẩn hóa văn bản đầu vào

Mô đun chuẩn hóa văn bản đầu vào có nhiệm vụ chính là làm cho văn bản đầu vào có thể đọc được một cách trơn tru, chuẩn hóa lại các thành phần không chuẩn như từ mượn, từ viết tắt, ngày tháng, số,… Quy trình chuẩn hóa văn bản đầu vào được thể hiện trong hình 24. Trong đó:

- Văn bản đầu vào sẽ được phân tách thành một danh sách các thành phần theo khoảng trắng, từng thành phần này được đưa vào tìm kiếm trong từ điển âm tiết, nếu có trong từ điển thì nó là thành phần có thể đọc được, nếu không có sẽ tiếp tục được đưa vào tìm kiếm trong từ điển từ viết tắt.

[12] http://www.cstr.ed.ac.uk/projects/merlin/

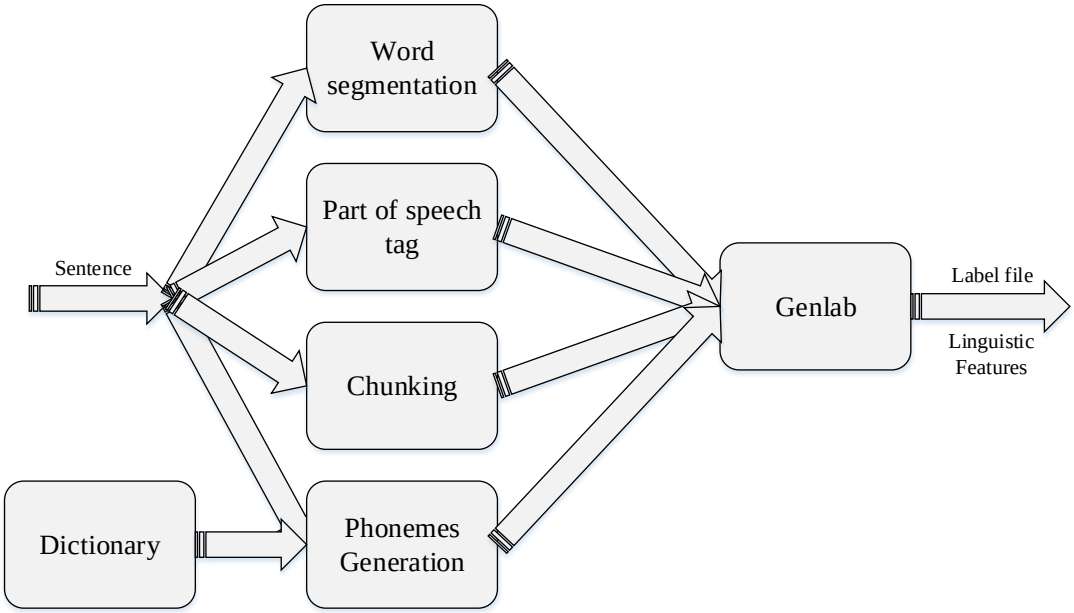
- Những thành phần không có trong từ điển thường được đưa vào từ điển viết tắt, tiếp tục nếu những thành phần này được tìm thấy trong từ điển viết tắt thì chúng sẽ được chuyển thành chuỗi các từ chuẩn theo đúng như trong từ điển, còn nếu không tìm thấy sẽ chuyển sang bước tiếp theo.
- Áp dụng các biểu thức chính quy: Bước này áp dụng cho những thành phần mà không có trong cả hai loại từ điển nêu trên như: ngày tháng, 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 dạng "../.."sẽ được thay thế bằng "ngày … tháng …".
- Cuối cùng là bước chuẩn hóa văn bản: Ở đây lưu lại những từ đã được chuẩn hóa ở các bước trước, tách những từ dính nhau.




Hình 24: Quá trình chuẩn hóa văn bản đầu vào

### 3.3.2 Mô đun trích chọn đặc trưng ngôn ngữ

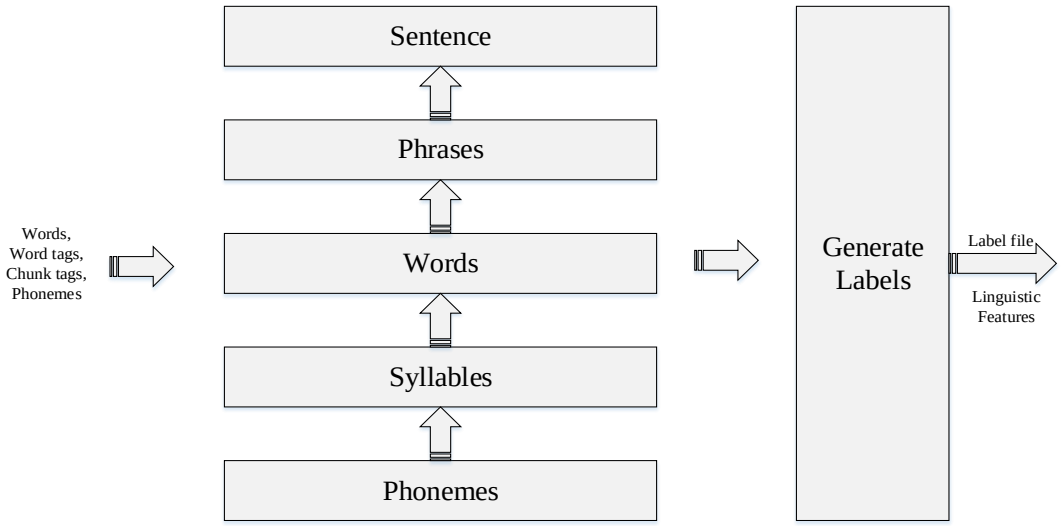
Mô đun trích chọn đặc trưng ngôn ngữ được xây dựng dựa trên ba mô hình: Mô hình tách từ, mô hình gán nhãn từ loại và mô hình tách cụm từ. Quá trình hoạt động của hệ thống được biểu diễn như hình 25, trong đó văn bản đầu vào sẽ được đưa qua các bộ gán nhãn từ loại (part of speech tag) để gán nhãn, tách từ bởi bộ tách từ (Word segmentation), Tách cụm từ với bộ tách cụm từ (Chunking) và tạo chuỗi âm vị với bộ Phonemes Generation. Sau đó, kết quả đầu ra của 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 (được mô tả trong phần 2.3).




Hình 25: Hoạt động của bộ trích chọn đặc trưng ngôn ngữ học

Trong hệ thống này, các mô hình gán nhãn từ loại, tách từ, tách cụm từ được xây dựng dựa trên mô hình Trường điều kiện ngẫu nhiên (Conditional Random Fields)[35] và sử dụng công cụ vita[36]. Từ điển được sử dụng trong quá trình tạo chuỗi âm vị là một từ điển phiên âm âm tiết tiếng Việt với khoảng 6700 từ.

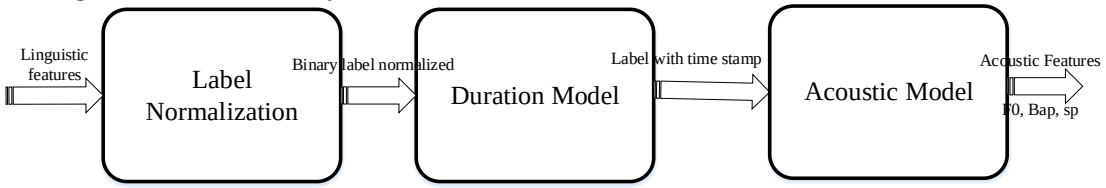
Bộ Genlab là bộ tạo đặc trưng ngôn ngữ học (lưu trong Label file), cấu trúc bộ Genlab được thể hiện trên hình 26, trong đó các chuỗi từ, chuỗi gán nhãn từ, chuỗi gán nhãn cụm từ (chuỗi cụm từ và nhãn), chuỗi âm vị sẽ được đưa vào một cấu trúc dữ liệu đặc biệt bao gồm một đối tượng đại diện cho câu (Sentence) 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 cách khác là từ đối tượng câu (Sentence) 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 2.3 sẽ được tính toán, ước lượng ra và lưu trong tệp chứa các nhãn (Label file). Cấu trúc từng dòng trong label file được nêu trong phụ lục A.




Hình 26: Cấu trúc và hoạt động của bộ Genlab

### 3.3.3 Mô đun tạo tham số đặc trưng âm học

Mô đun tạo tham tham số đặc trưng âm học có nhiệm vụ lấy đầu vào là các véc tơ đặc trưng ngôn ngôn ngữ học được trích ở phần trước, hay chính là các dòng được lưu trong label file. Đầu ra của mô đun này là các đặc trưng âm học bao gồm các thông tin như: F0 là tần số cơ bản, SP là đường bao phổ, BAP chứa thông tin về các thành phần không tuần hoàn. Cấu trúc của mô đun tạo tham số đặc trưng âm học được mô tả trong hình 27, trong đó mô đun này được cấu tạo bởi ba phần chính đó là mô hình âm học (Acoustic model), mô hình khoảng thời gian (Duration Model), bộ chuẩn hóa đặc trưng đầu vào (Label Normalization). Công cụ được sử dụng để xây dựng cả ba mô đun này là Merlin[4].




Hình 27: Cấu trúc mô đun tạo tham số đặc trưng

Bộ chuẩn hóa đặc trưng đầu vào có nhiệm vụ nhận các véc tơ đặc trưng ngôn ngữ từ mô đun trích chọn đặc trưng ngôn ngữ, và trả về đầu ra là các véc tơ mô tả đặc trưng ngôn ngữ nhưng dưới dạng nhị phân. Phương pháp được sử dụng để chuyển hóa véc tơ đặc trưng ngôn ngữ học sang dạng véc tơ nhị phân là sử dụng tập câu hỏi đã được trình bày trong phần 2.4. Tập câu hỏi được sử dụng cho tiếng việt bao gồm 743 câu hỏi về các thông tin như "âm vị hiện tại", "âm vị trước, sau", "nhãn từ loại", "thanh điệu", "số lượng âm vị trong âm tiết, từ, cụm từ",…

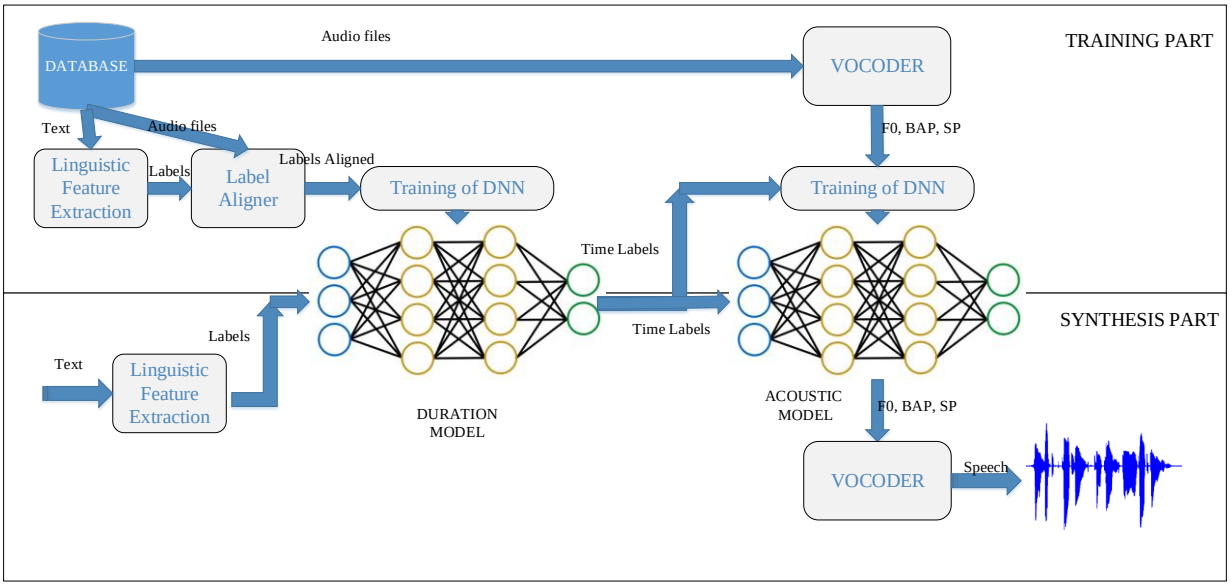
Bộ mô hình khoản thời gian Duration Model, nhận đầu vào là các véc tơ đặc trưng ngôn ngữ học, và đầu ra của bộ này là các véc tơ đặc trưng ngôn ngữ học cộng thêm với các thông tin về thời gian xuất hiện (thời điểm bắt đầu và kết thúc) của mỗi âm vị. Mô đun này cũng sử dụng mô hình mạng nơ ron giống như mô hình mạng nơ ron giống như mô hình âm học được mô tả trong phần 2.4 chỉ khác một chút là với

mỗi đầu vào của mạng nơ ron này là các véc tơ đặc trưng ngôn ngữ học thì đầu ra của mạng là thông tin về thời gian xuất hiện của âm vị tương ứng.

Mô hình âm học nhận đầu vào là các véc tơ chứa đặc trưng ngôn ngữ học và thông tin về thời gian xuất hiện của từng âm vị tương ứng trong véc tơ đặc trưng ngôn ngữ học, và trả về đầu ra là các véc tơ đặc trưng âm học của tín hiệu tiếng nói. Véc tơ đặc trưng âm học chứa các thông tin cụ thể như sau: Véc tơ 60 chiều  của các hệ số Mel mang các thông tin về đường bao phổ, véc tơ 5 chiều của các tham số không tuần hoàn (Bap), và lô ga rit của tần số cơ bản F0. Các véc tơ đặc trưng ngôn ngữ học sẽ là đầu vào cho mô đun vocoder để tạo tín hiệu tiếng nói.

Hình 28 biểu diễn hai quá trình chính của hệ thống tổng hợp tiếng nói dựa trên mạng nơ ron học sâu là quá trình huấn luyện các mô hình và tổng hợp tiếng nói từ các mô hình đã huấn luyện. Quá trình huấn luyện hệ thống tổng hợp tiếng nói bao gồm các giai đoạn sau: Giai đoạn một là huấn luyện mô hình khoảng thời gian Duration model và giai đoạn hai là huấn luyện mô hình âm học. Trong giai đoạn một, dữ liệu đầu vào gồm có các tập âm thanh và văn bản tương ứng, các tập văn bản này sẽ được trích chọn các đặc trưng ngôn ngữ thông qua bộ Linguistic Feature Extraction đã nêu ở phần 3.3.2,  đầu ra sẽ là các đặc trưng ngôn ngữ học được biểu diễn dưới dạng các nhãn (cấu trúc nhãn xem phụ lục A). các nhãn này sẽ được đưa vào bộ Label Aligner cùng với các tệp âm thanh. Bộ label aligner là bộ tính toán thời gian xuất hiện của âm vị sử dụng mô hình markov ẩn (được nêu trong mục 2.3). Kết quả đầu ra của Label Aligner là các nhãn đặc trưng âm học và kèm thêm thông tin về thời gian xuất hiện của từng âm vị tương ứng với nhãn đó. Nhãn mới có chứa thông tin về thời gian xuất hiện sẽ được đưa vào huấn luyện mô hình khoảng thời gian Duration Model. Sau khi huấn luyện xong, mô hình khoảng thời gian sẽ được sử dụng để ước lượng lại thời gian xuất hiện của âm vị, thay thế cho kết quả của bộ Label Aligner dùng HMM. Thông tin về thời gian xuất hiện của âm vị mới được ước lượng bởi mô hình khoảng thời gian sẽ thay thế thông tin về thời gian cũ trong nhãn. Giai đoạn hai, Bộ Vocoder (cụ thể ở đây là WORLD vocooder) sẽ được sử dụng để trích chọn các đặc trưng âm học từ các tệp âm thanh đầu vào, các đặc trưng âm học này bao gồm các thông tin về tần số cơ bản F0, đường bao phổ SP và các tham số không tuần hoàn BAP. Từ các đặc trưng âm học này, kết hợp với các nhãn mang thông tin về đặc trưng ngôn ngữ và thời gian xuất hiện của âm vị (đầu ra của mô hình khoảng thời gian duration model) đưa vào huấn luyện cho mô hình âm học (Acoustic model.

Quá trình tổng hợp tiếng nói từ văn bản, văn bản đầu vào sẽ được đưa qua bộ Linguistic Feature Extraction để tạo các nhãn (Labels) mang các thông tin đặc trưng âm học. Các nhãn đặc trưng âm học được đưa qua mô hình khoảng thời gian (Duration Model), kết quả nhận được là các nhãn mới có thêm các thông tin về thời gian xuất hiện của âm vị tương ứng. Các nhãn mới này sẽ được đưa qua mô hình âm học, từ mô hình âm học ta có được các đặc trưng âm học như tần số cơ bản F0, đường bao phổ SP, tham số không tuần hoàn BAP. Các đặc trưng âm học này sẽ được đưa vào vocoder để tạo ra tín hiệu tiếng nói.

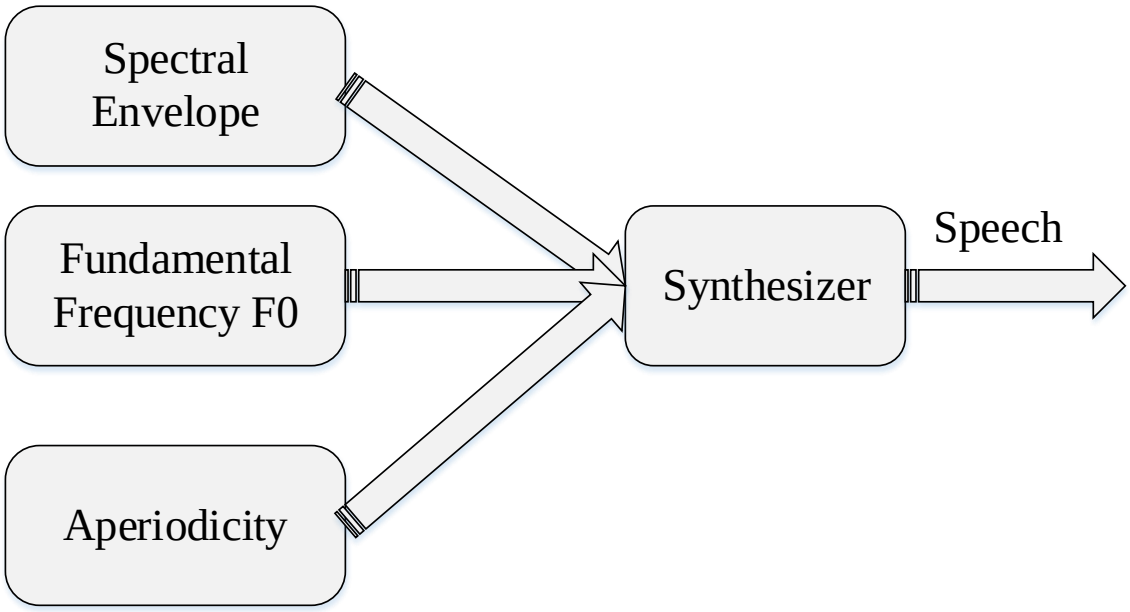



Hình 28: Quá trình huấn luyện và tổng hợp một hệ thống tổng hợp tiếng nói dựa trên mô hình mạng nơ ron học sâu.

Trong toàn bộ hệ thống nêu trên, phần tạo tham số đặc trưng bao gồm hai mô hình mạng nơ ron mô tả mô hình âm học và mô hình thời gian, trong thực tế triển khai hai mô hình này đều là các mạng nơ ron với 6 lớp ẩn, mỗi lớp có 1024 nút, hàm kích hoạt trong mỗi nút là hàm tanh, sử dụng phương pháp tối ưu SDG (Stochastic Gradient Descent) để tối ưu mạng.

### 3.3.4 Mô đun tổng hợp tiếng nói từ các đặc trưng âm học

Trong mô đun này, chúng tôi sử dụng WORLD vocoder[30] cho nhiệm vụ tổng hợp tiếng nói từ các tham số đặc trưng, chi tiết về bộ vocoder này được trình bầy trong phần 2.5.




Hình 29: Tổng hợp tiếng nói từ các đặc trưng âm học bằng WORLD vocoder.

Trên hình 29 ta thấy, bộ Synthesizer của WORLD vocoder nhận các đầu vào đặc trưng âm học là đường bao phổ Spectral Envelope được tính từ véc tơ 60 chiều của

các hệ số mel đầu ra của mô đun Parameter Generation, tần số cơ bản F0 được tính từ log của F0 là đầu ra của mô đun Parameter Generation và các tham số không tuần hoàn. Đầu ra của quá trình này chính là tín hiệu tiếng nói tự nhiên.

## 3.4 Xây dựng cơ sở dữ liệu và huấn luyện hệ thống

### 3.4.1 Thu thập dữ liệu cho hệ thống tổng hợp tiếng nói

Dữ liệu là một trong những phần quan trọng nhất ảnh hưởng đến chất lượng tổng hợp tiếng nói, do đó để có được một hệ thống tổng hợp tiếng nói chất lượng ta cần phải chuẩn bị những bộ dữ liệu chất lượng. Để xây dựng được bộ cơ sở dữ liệu có chất lượng tốt, cần trải qua hai quá trình: Thu thập dữ liệu cho hệ thống tổng hợp tiếng nói, xử lý dữ liệu thu thập được (tiền xử lý dữ liệu).

Thu thập dữ liệu: Do có nguồn tài nguyên và nhân lực hạn chế nên việc thu âm dữ liệu với nhiều giọng đọc khác nhau là không thể, do đó lựa chọn thu thập dữ liệu âm thanh từ mạng Internet là hợp lý. Nguồn dữ liệu âm thanh chủ yếu được thu thập từ các trang phát thanh trực tuyến như netnews.vn hay các kênh trên youtube.com chuyên về đọc truyện đêm khuya. Dữ liện sau khi thu thập gồm có các tệp âm thanh được nén dạng mp3 và có thời gian phát khá dài, các tệp này sẽ được chuyển thành định dạng wav để phục vụ cho quá trình xử lý dữ liệu trong phần 3.5. Dữ liệu thu được vào khoảng 7 giờ ghi âm.

### 3.4.2 Huấn luyện hệ thống

Trước khi đưa hệ thống này vào hoạt động thì cần huấn luyện mô hình học máy của nó, cụ thể là hai mô hình khoảng thời gian và mô hình âm học, vì hai mô hình này có quá trình huấn luyện giống hệt nhau nên sẽ chỉ trình bày tập trung vào quá trình huấn luyện mô hình âm học.

Dữ liệu được sử dụng để huấn luyện cho cả hai mô hình là dữ liệu được tải về từ Internet trong phần trước. Sau khi xử lý, dữ liệu còn lại được đưa vào huấn luyện bao gồm 3504 tệp dữ liệu âm thanh và tương ứng 3504 câu văn bản, tương đương với khoảng 6,5 giờ ghi âm. Trong đó, 3154 tệp dùng để huấn luyện, 174 tệp dùng làm tập kiểm tra (test) và 174 tệp dùng làm tệp chuẩn (valid), bảng 2 chứa chi tiết mô tả dữ liệu âm thanh cho huấn luyện hệ thống. Quá trình huấn luyện hệ thống gồm hai giai đoạn đã được nêu trong mục 3.2.3

| Số tệp âm thanh | Số giờ ghi âm (giờ) | Số tệp huấn luyện | Số Tệp chuẩn | Số tệp kiểm tra |
|---|---|---|---|---|
| 3504 | 6.5 | 3154 | 174 | 174 |

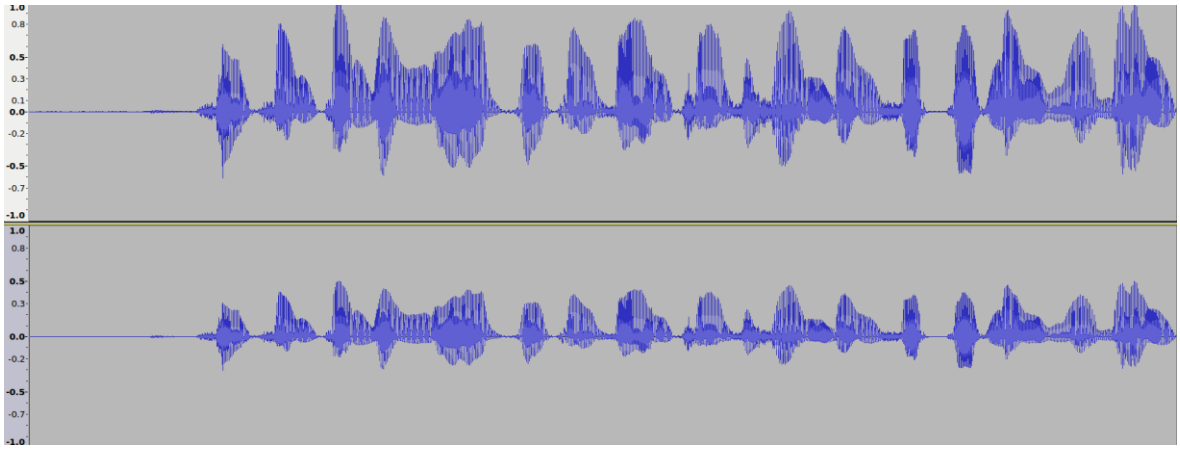
Bảng 2: Dữ liệu huấn luyện hệ thống tổng hợp tiếng nói

## 3.5 Xử lý dữ liệu huấn luyện để nâng cao chất lượng đầu ra

Dữ liệu tải về từ internet có rất nhiều vấn đề như nhiễu nền, giọng đọc lúc to lúc nhỏ, nhiều từ vay mượn hay tệp âm thanh quá dài. Điều này dẫn đến các vấn đề về chất lượng đầu ra của hệ thống như sau:

- Giọng đọc lúc to lúc nhỏ dẫn đến kết quả tổng hợp tiếng nói cũng bị như vậy, thậm trí còn trầm trọng hơn khi lúc thì quá to và lúc thì quá nhỏ. Để giải quyết vấn đề này cần cân bằng cường độ âm thanh của dữ liệu huấn luyện.
- Có nhiều tiếng nhiễu ở giữa các khoảng nghỉ và các dấu phẩy, dấu chấm. Nguyên nhân là do dữ liệu còn nhiễu ở các khoảng này, giải pháp để giải quyết là lọc nhiễu dữ liệu ở các khoảng này, tuy nhiên không lọc vào tín hiệu vì có thể gây méo.
- Các tệp âm thanh không đủ cả câu, điều này ảnh hưởng đến ngữ điệu của tiếng nói đầu ra vì các đặc trưng ngôn ngữ không mang đủ thông tin về ngữ điệu tương ứng cả câu. Giải pháp cho vấn đề này là căc tệp âm thanh theo đúng một câu tương ứng một câu văn bản.
- Có các từ viết tắt và từ tiếng nước ngoài trong văn bản, nếu sử dụng các âm vị để phiên âm từ này thì sẽ gây méo  tín hiệu tiếng nói của âm vị đó, do đó các câu có chứa những từ này cần được loại bỏ.

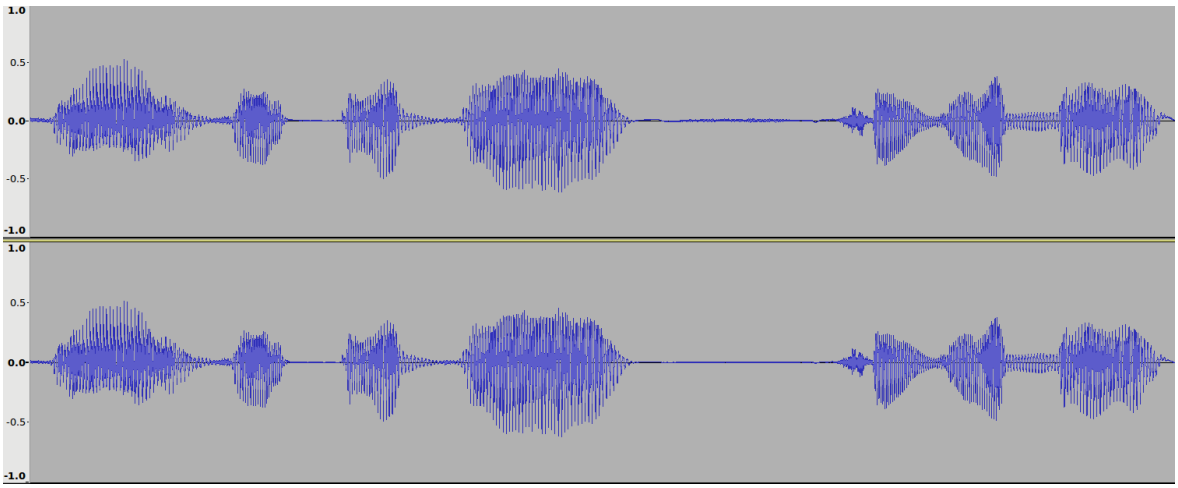
Cân bằng dữ liệu âm thanh là quá trình chuyển đổi cường độ âm của các tệp âm thanh trong toàn bộ dữ liệu về gần một ngưỡng cân bằng . Để làm được điều này, đầu tiên ta tìm cường độ âm thanh lớn nhất của mỗi tệp âm thanh trên toàn bộ tập dự liệu, sau đó tính cường độ âm thanh trung bình của các cường độ âm thanh lớn nhất trên toàn bộ dữ liệu, và cuối cùng là chuyển hóa toàn bộ cường độ âm thanh của các tệp trong dữ liệu sao cho cường độ lớn nhất của chúng gần ngưỡng trung bình mà ta tìm được. Công cụ được sử dụng để tìm cường độ trung bình và chuyển hóa cường độ âm thanh nêu trên là SOX[13]. Hình 30 cho thấy cường độ âm thanh trước và sau khi cân bằng, ở đây ngưỡng cân bằng nhỏ hơn cường độ lớn nhất của tệp âm thanh do đó quá trình cân bằng sẽ làm giảm cường độ của tệp âm thanh.



Hình 30: Tín hiệu âm thanh trước (trên) và sau khi cân bằng (dưới)

Lọc dữ liệu âm thanh: đây là quá trình khá tốn công sức vì nếu lọc tên toàn bộ tập anh thanh thì tín hiệu âm thanh của tệp đó sẽ bị thay đổi, thậm trí là bị méo gây nên việc không đảm bảo chất lượng dữ liệu đầu vào do đó chỉ được phép lọc ở những khoảng nghỉ giữa các từ và không được lọc vào phần tín hiệu tạo âm. Hình 31 cho thấy tệp âm thanh trước và sau khi lọc dữ liệu, phần nhiễu trong khoảng lặng đã biến mất.

[13] http://sox.sourceforge.net/sox.html

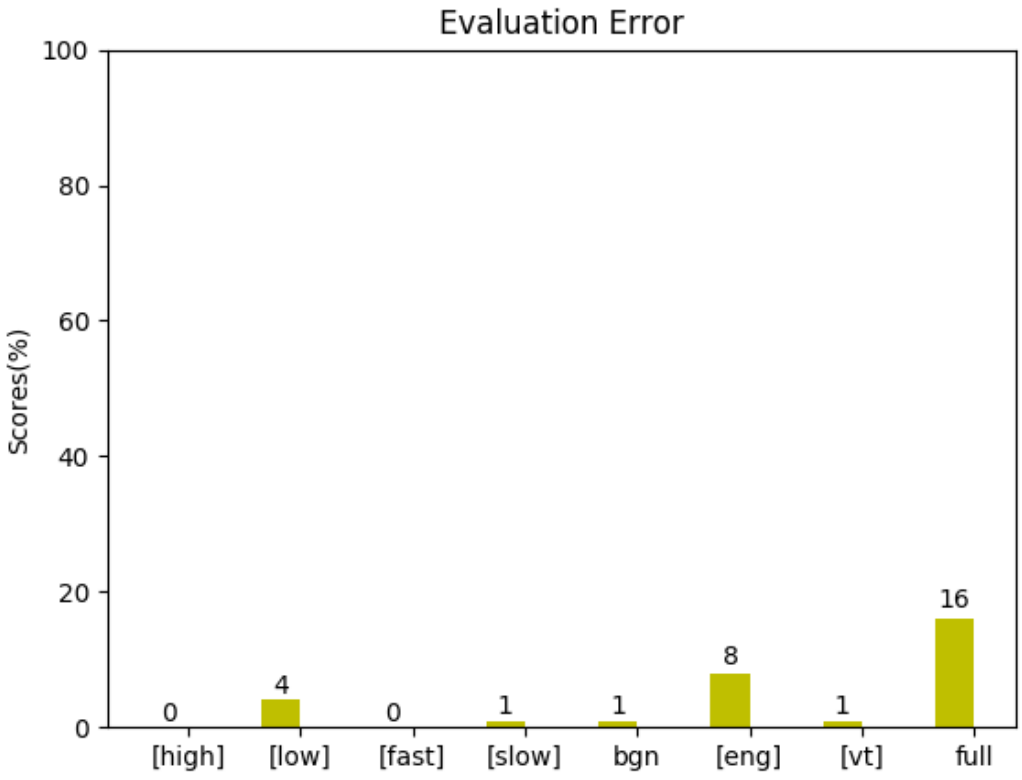



Hình 31: Tín hiệu âm thanh trước (ở trên) và sau (ở dưới) sau khi lọc nhiễu

Để giải quyết các vấn đề trên, một quy trình xử lý dữ liệu được đề xuất như sau:

- Cắt tệp âm thanh dài thành các tập âm thanh nhỏ sao cho mỗi tập tương ứng như một câu trong văn bản.
- Gán nhãn dữ liệu theo các nhãn: tệp có nhiễu nền, câu quá to hoặc quá nhỏ, câu quá nhanh hoặc quá chậm, từ tiếng anh hoặc tiếng nước ngoài khác, từ viết tắt.
- Lọc bỏ những câu có nhãn không mong muốn (như nhiễu nền).
- Lọc nhiễu ở những đoạn câm lặng hay các khoảng nghỉ (không lọc vào âm thanh vì gây méo phổ).
- Cân bằng các tệp âm thanh về mức âm lượng vừa phải.
- Thêm khoảng lặng vào đầu và cuối mỗi tệp âm thanh để đảm bảo các tệp có khoảng lặng đầu cuối là như nhau.

Hình 32 cho thấy phân bố dữ liệu sau khi gán nhãn của một giọng mẫu được tải về từ internet, trong đó có rất nhiều từ mượn trong tiếng anh (eng) và nó chiếm tới 8% tổng số câu trong dữ liệu, ngoài ra còn có các câu có âm lượng nhỏ (low: chiếm tới 4% tổng số câu). Tổng cộng các câu được gán nhãn chiếm 16% dữ liệu. Đối với những câu có từ tiếng anh thì nên được loại bỏ vì các từ này khi đưa vào huấn luyện sẽ làm sai lệch cách đọc của âm vị. Đối với những câu quá nhỏ hoặc quá lớn thì tiếng hành cân bằng âm lượng, với những từ viết tắt cần viết lại đúng theo người đọc trong tệp âm thanh tương ứng. Những câu quá chậm hoặc quá nhanh thì cần được loại bỏ. Sau quá trình xử lý dữ liệu theo những bước nên trên ta thu được một bộ dữ liệu khá chất lượng và đảm bảo được cho quá trình huấn luyện.

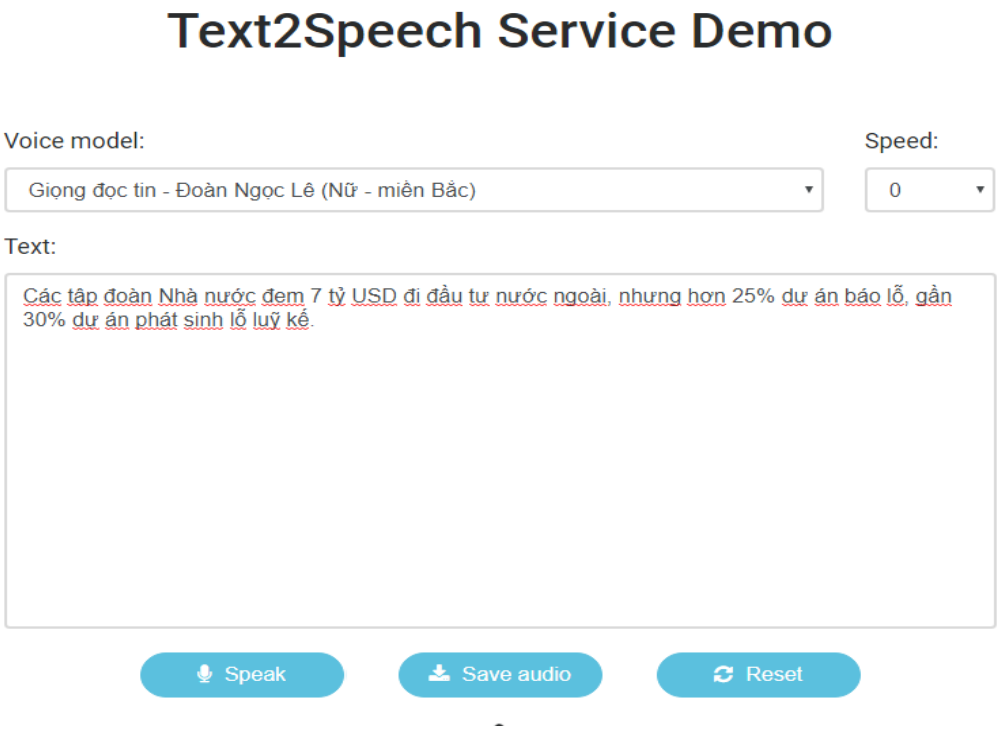



Hình 32: Phân bố dữ liệu sau khi gán nhãn

# CHƯƠNG 4: CÀI ĐẶT THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

## 4.1 Cài đặt thử nghiệm hệ thống

Sau khi đã có được một hệ thống hoàn chỉnh và huấn luyện được mô hình, chúng tôi tiến hành cài đặt hệ thống trên máy tính có CPU E5-2640 với 32 nhân, tần số 2,6Ghz, Ram 128Gb. Một số kết quả chạy thử nghiệm hệ thống được biểu diễn trên hình 33 và 34.




Hình 33: Hình ảnh chạy thử nghiệm hệ thống tổng hợp tiếng nói 1.


Hình 34: Hình ảnh chạy thử nghiệm hệ thống tổng hợp tiếng nói 2

## 4.2 Đánh giá kết quả thử nghiệm hệ thống

### 4.2.1 Đánh giá chất lượng bộ tổng hợp dùng DNN so với HMM

Mục tiêu của đánh giá này là để kiểm tra xem việc áp dụng mạng nơ ron học sâu vào tổng hợp tiếng nói có thực sự giúp chất lượng tổng hợp được cải thiện hay không.

Phương pháp đánh giá như sau: phương pháp MOS (mean opinion score)

- Mời 6 người tham gia đánh giá và cho điểm chất lượng hệ thống
- Tiêu chí cho điểm chất lượng hệ thống là dựa trên độ tự nhiên.
- Tập dữ liệu đánh giá là tập gồm 30 tệp âm thanh được tổng hợp từ ba mươi văn bản khác nhau được lấy từ báo chí.
- Kết quả đánh giá được cho trên thang điểm 5, kết quả cuối cùng là điểm trung bình của tất của người nghe.

Các hệ thống được đánh giá: Hệ thống được đánh giá bao gồm hai hệ thống được xây dựng theo hai mô hình HMM và DNN. Hai hệ thống này được huấn luyện từ cùng một bộ dữ liệu có 3504 câu, và cùng một tập đánh giá nêu trên.

Kết quả đánh giá được nêu trong bảng 3, trong đó điểm số trung bình của hệ thống DNN là 4.23 lớn hơn hẳn HMM là 3.96 điều này cho thấy rõ ràng là việc áp dụng mạng nơ ron học sâu đã góp phần cải thiện chất lượng hệ thống tổng hợp tiếng nói cả về độ hiểu và độ tự nhiên.

| | DNN | HMM |
|---|---|---|
| Điểm MOS | 4.23 | 3.96 |

Bảng 3: Kết quả so sánh bộ tổng hợp DNN và HMM

### 4.2.2 Đánh giá kết quả của việc cải thiện cơ sở dữ liệu huấn luyện

Mục tiêu của đánh giá này là để kiểm tra xem việc xử lý dữ liệu huấn luyện của hệ thống có giúp nâng cao chất lượng tổng hợp tiếng nói của hệ thống hay không.

Phương pháp đánh giá như sau: phương pháp MOS (mean opinion score)

- Mời 6 người tham gia đánh giá và cho điểm chất lượng hệ thống
- Tiêu chí cho điểm chất lượng hệ thống là dựa trên độ tự nhiên.
- Tập dữ liệu đánh giá là tập gồm 30 tệp âm thanh được tổng hợp từ ba mươi văn bản khác nhau được lấy từ báo chí.
- Kết quả đánh giá được cho trên thang điểm 5, kết quả cuối cùng là điểm trung bình của tất của người nghe.

Các hệ thống được đánh giá: Hệ thống được đánh giá bao gồm hai hệ thống được xây dựng theo hai mô hình DNN. Hai hệ thống này được huấn luyện từ cùng một bộ dữ liệu có 3504 câu, nhưng hệ thống DNN1 là hệ thống được huấn luyện trên bộ dữ liệu chưa được xử lý và hệ thống DNN2 là hệ thống được huấn luyện trên bộ dữ liệu đã qua xử lý theo các bước trong phần (3.5).

Kết quả đánh giá được nêu trong bảng 4, kết quả này cho thấy một sự cải thiện chất lượng cực kì đáng kể khi mà điểm số của DNN2 là 4.61 cao hơn nhiều so với của DNN1 là 4.11, độ lệch này thậm chí còn lớn hơn độ lệch khi mà sử dụng DNN thay cho HMM (0.5 so với 0.26). Điều này cho thấy chất lượng dữ liệu huấn luyện là cực

kỳ quan trọng và việc tiền xử lý dữ liệu đầu vào góp phần cải thiện đáng kể chất lượng tổng hợp của hệ thống.

| | DNN1 | DNN2 |
|---|---|---|
| Điểm MOS | 4.11 | 4.61 |

Bảng 4: Kết quả so sánh chất lượng tổng hợp tiếng nói của hệ thống có dữ liệu huấn luyện đã được xử lý (DNN2) và chưa được xử lý (DNN1).

### 4.2.3 Đánh giá so sánh chất lượng hệ thống tổng hợp tiếng nói so với các hệ thống tổng hợp tiếng Việt hiện có

Mục tiêu của đánh giá này là kiểm tra chất lượng hệ thống tổng hợp tiếng nói được xây dựng trong luận văn này và so sánh với các hệ thống tổng hợp tiếng nói sẵn có cho tiếng Việt hiện nay.

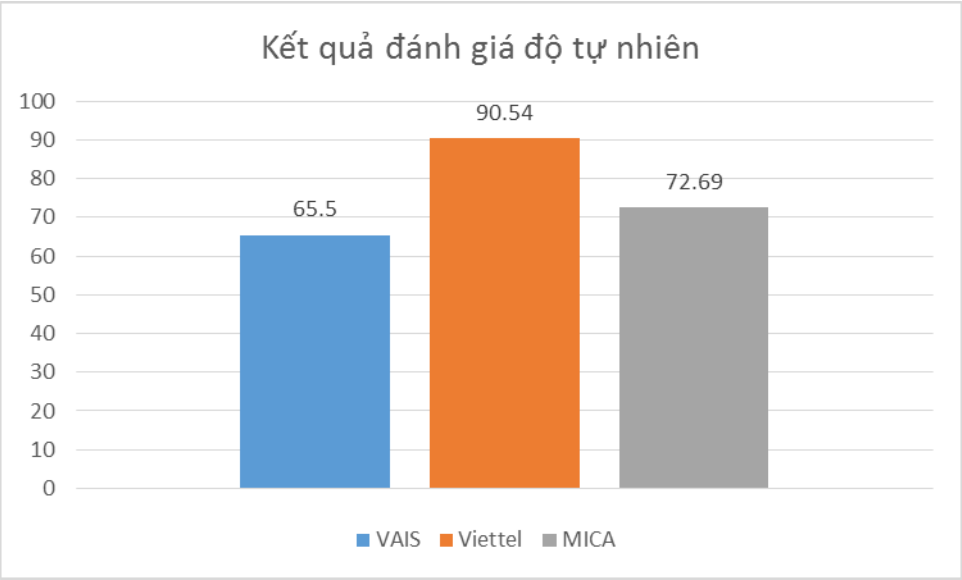
Kết quả này được lấy từ kết quả đánh giá hệ thống khi tham gia cuộc thi tổng hợp tiến nói trong VLSP Workshop (Hệ thống của chúng tôi đã đoạt giải nhất cuộc thi này). Phương pháp đánh giá:

- Đánh giá theo các tiêu chí độ tự nhiên và độ hiểu trên thang 100 và đánh giá dựa trên ý kiến trung bình của người nghe (MOS Mean Opinion Score) xét trên cả độ tự nhiên và độ hiểu, trên thang điểm 5.
- Mời 20 người tham gia đánh giá cho hệ thống dựa trên trên một tập gồm 30 câu văn được lấy từ báo chí, thông tin về 20 người này được thể hiện trong bảng 5.

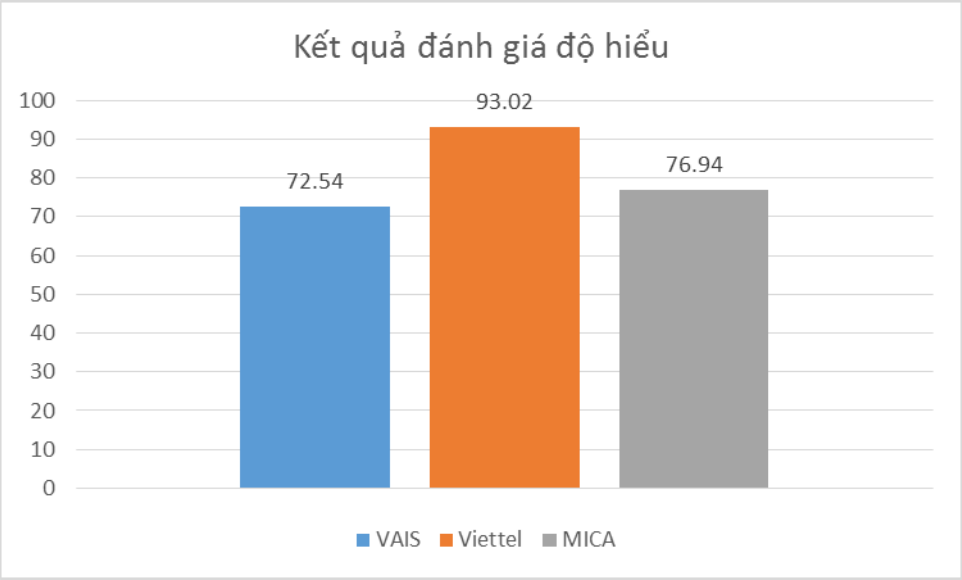
| Thông tin người nghe | | | | | | |
|---|---|---|---|---|---|---|
| Giới tính | | Phương ngữ | | | Chuyên gia ngữ âm | |
| Nữ | Nam | Bắc | Nam | Trung | Yes | No |
| 12 | 08 | 10 | 07 | 03 | 13 | 07 |

Bảng 5: Thông tin người nghe đánh giá hệ thống tổng hợp tiếng nói

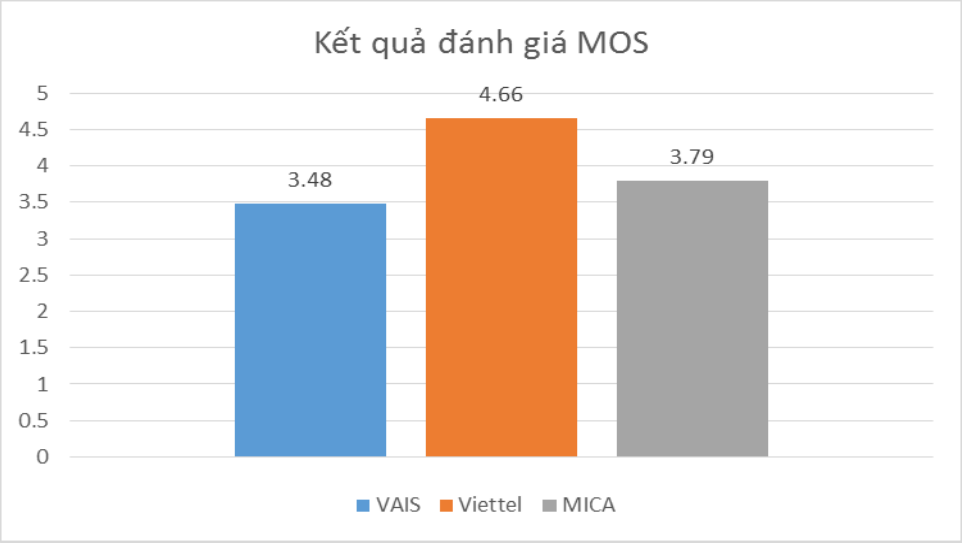
Kết quả đánh giá chất lượng giọng tổng hợp được thể hiện trong các hình 35, 36, 37. Trong đó đội Viettel cũng chính là đội chúng tôi với hệ thống hiện tại mà tôi mô tả trong luận văn này, luôn đạt điểm cao nhất trong các đánh giá so với các đội khác cùng tham gia VLSP. Với điểm về độ hiểu và độ tự nhiên luôn đạt mức trên 90% tức là đã rất gần với giọng tự nhiên. Đặc biệt với đánh giá MOS (Mean Opinion Score) đánh giá chất lượng giọng dựa trên ý kiến trung bình của người nghe, hệ thống của chúng tôi đạt điểm 4.66, một số điểm rất là cao ngay cả khi so sánh với các hệ thống tiếng Anh.




Hình 35: Đánh giá độ tự nhiên



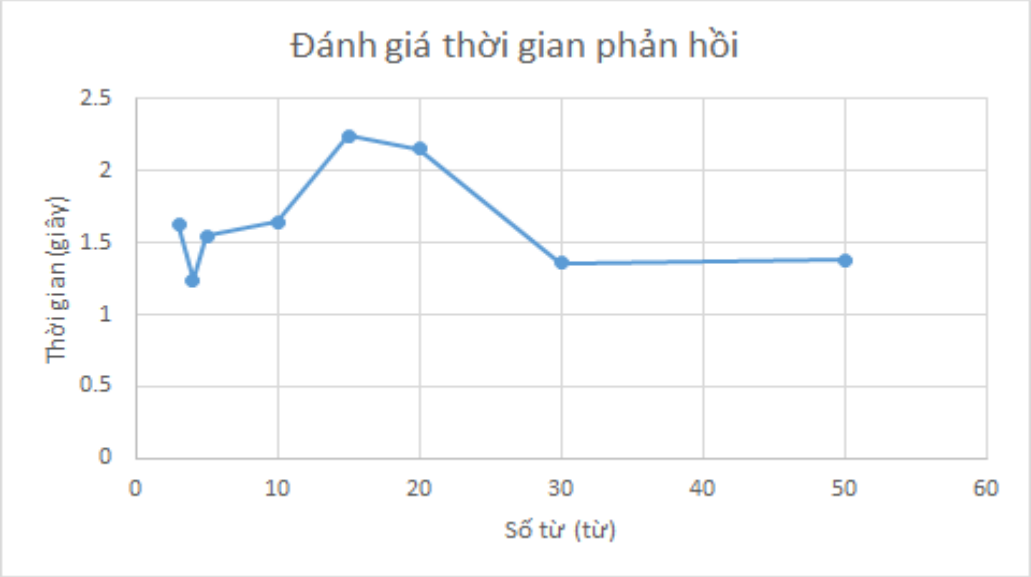

Hình 36: Đánh giá độ hiểu




Hình 37: Đánh giá MOS

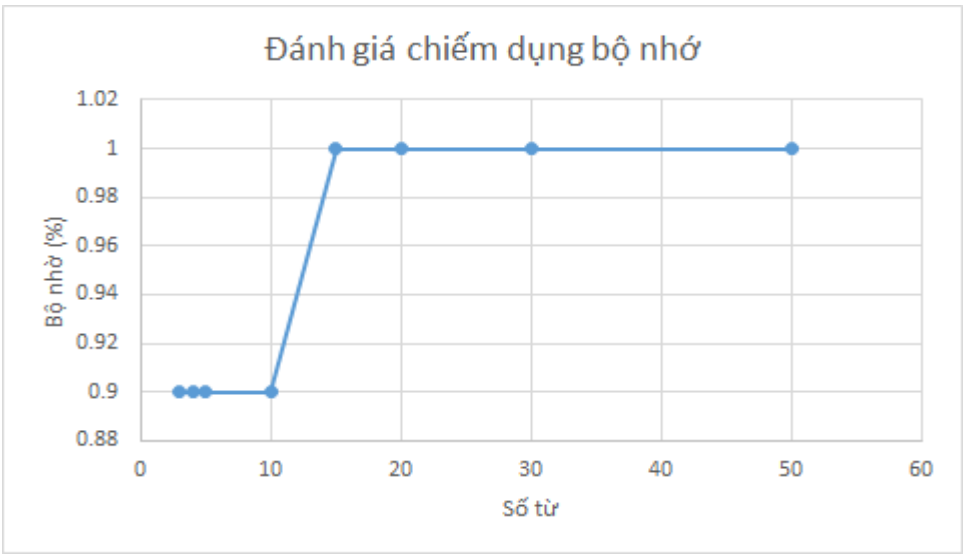
### 4.2.4 Đánh giá hiệu năng hệ thống

Đánh giá về hiệu năng của hệ thống này bao gồm đánh giá về thời gian phản hồi, dung lượng chiếm dụng bộ nhớ và tài nguyên của hệ thống. Môi trường đánh giá là môi trường cài đặt hệ thống trong phần 4.1. Phương pháp đánh giá là sử dụng 30 câu có độ dài ngắn khác nhau để đưa vào hệ thống và tính toán các tham số cần thiết.




Hình 38: Đánh giá thời gian đáp ứng của hệ thống

Đánh giá thời gian phản hồi của hệ thống là đánh giá thời gian hệ thống cần để chuyển hóa văn bản thành tiếng nói, khoảng thời gian này được tính từ thời điểm người dùng chuyển đoạn văn bản lên hệ thống cho tới lúc nhận kết quả trả về. Kết quả đánh giá được thể hiện trong hình 38. Từ kết quả đánh giá ta thấy thời gian đáp ứng của hệ thống không quá phụ thuộc vào độ dài câu, vì trong hệ thống này sử dụng giải pháp chia cắt câu thành các đoạn đặc trưng ngôn ngữ học và tổng hợp chúng theo thời gian thực. Tuy nhiên thời gian phản hồi trung bình vẫn còn khá lớn (lớn hơn 1,5 giây) do đó chưa đủ để đáp ứng bài toán thương mại và cần phải cải thiện thêm.




Hình 39: Đánh giá chiếm dụng bộ nhớ

Đánh giá dung lượng chiếm dụng bộ nhớ, dung lượng bộ nhớ mà hệ thống chiếm dụng được tính tại thời điểm tạo các tham số đặc trưng âm học bằng mô hình DNN, đây cũng là thời điểm chiếm dụng bộ nhớ nhiều nhất. Kết quả đánh giá được thể hiện trong hình 39. Kết quả đánh giá cho thấy, dung lượng chiếm dụng bộ nhớ không quá nhiều chỉ khoảng 1% trên toàn bộ dung lượng vật lý của môi trường.

# KẾT LUẬN

## A. Tổng kết

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

- Tìm hiểu và làm chủ được công nghệ tổng hợp tiếng nói, xây dựng thành công hệ thống tổng hợp tiếng nói tiếng Việt đầu tiên sử dụng công nghệ học sâu.
- Phân tích được một số vấn đề trong việc xây dựng cơ sở dữ liệu huấn luyện trong tổng hợp tiếng nói dựa trên phương pháp học sâu, kiểm định kết quả cải thiện thông qua các đánh giá.

Hệ thống tổng hợp tiếng nói được phát triển trong khuôn khổ luận văn này đã được ứng dụng và triển khai tại tập đoàn công nghiệp viễn thông quân đội Viettel, là một mô đun cấu thành nên nền tảng trí tuệ nhân tạo (AI) của Viettel, và đã được tích hợp vào các hệ thống như hệ thống trợ lý ảo Viettel và hệ thống chăm sóc khách hàng tự động. Ngoài ra, hệ thống tổng hợp tiếng nói này cũng đã được gửi đi tham dự cuộc thi về tổng hợp tiếng nói trong hội nghị VLSP[14] 2018 và đã giành giải nhất, vượt qua các đội Mica và vais (Đánh giá về cả ba hệ thống này được nêu trong chương 4). Báo cáo về hệ thống tổng hợp tiếng nói này dành cho hội thảo VLSP được nêu trong phụ lục B.

Ngoài ra, trong quá trình làm luận văn, tác giả có có một bài báo được công bố và trình bày tại Hội nghị quốc tế về Nhận dạng ký tự và Xử lý ngôn ngữ tự nhiên cho các ngôn ngữ Asean (Regional Conference on **O**ptical character recognition and **N**atural language processing technologies for **A**SEAN languages - **ONA 2017**))[15]. Chi tiết về các báo cáo khoa học trong cuộc thi tổng hợp tiếng nói tại VLSP 2018 và bài báo tại hội nghị ONA 2017 xin xem trong Phụ lục B.

## B. Phương hướng phát triển và cải thiện hệ thống

Hệ thống tổng hợp tiếng nói trong khuôn khổ của luận văn tuy đạt được chất lượng đầu ra tương đối tốt so với các hệ thống hiện tại, tuy nhiên vẫn còn một số vấn đề cần cải thiện như:

- Thời gian đáp ứng còn chậm
- Chưa đạt được chất giọng tốt trong tổng hợp tiếng nói theo phương ngữ miền Nam của tiếng Việt

Vì vậy, công việc tiếp theo của luận văn là tiếp tục cải thiện các nhược điểm của hệ thống cũng như nâng cấp các khả năng khác của hệ thống cụ thể như:

- Cải thiện thời gian đáp ứng bằng cách song song hóa và lọc bỏ các khâu không cần thiết.
- Thêm các giải pháp mới cho bài toàn chuẩn hóa văn bản đầu vào.
- Thêm từ điển dành riêng cho các phương ngữ khác như phương ngữ Nam và Trung để cải thiện chất lượng tổng hợp các phương ngữ này.

[14] http://vlsp.org.vn/
[15] http://ona2017.org/

# TÀI LIỆU THAM KHẢO


[1] A.-T. Dinh, T.-S. Phan, T.-T. Vu, and C.-M. Luong, "Vietnamese HMM-based Speech Synthesis with prosody information," *Th ISCA Speech Synth. Workshop*, p. 4, 2013.

[2] T.-S. Phan, T.-C. Duong, A.-T. Dinh, T.-T. Vu, and C.-M. Luong, "Improvement of naturalness for an HMM-based Vietnamese speech synthesis using the prosodic information," 2013, pp. 276–281.

[3] H. Zen *et al.*, "The HMM-based Speech Synthesis System (HTS) Version 2.0," p. 6, 2007.

[4] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," 2016, pp. 202–207.

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

[6] D. Suendermann, H. Höge, and A. Black, "Challenges in Speech Synthesis," in *Speech Technology*, Huggins and F. Chen, Eds. Boston, MA: Springer US, 2010, pp. 19–32.

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

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[9] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech: Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," PhD Thesis, Paris 11, 2015.

[10] Q. Nguyễn Hồng, "Phân tích văn bản cho tổng hợp tiếng nói tiếng Việt," Đại Học Bách Khoa Hà Nội, 2006.

[11] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[12] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 853–870, 2004.

[13] S. Lukose and S. S. Upadhya, "Text to speech synthesizer-formant synthesis," 2017, pp. 1–4.

[14] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," 1986, vol. 11, pp. 2015–2018.

[15] S.-J. Kim, "HMM-based Korean speech synthesizer with two-band mixed excitation model for embedded applications," PhD Thesis, Ph. D. dissertation, School of Engineering, Information and Communication University, Korea, 2007.

[16] T. Masuko, "HMM-Based Speech Synthesis and Its Applications," p. 185, 2002.

[17] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," 1992, pp. 137–140 vol.1.

[18] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," 2013, pp. 7962–7966.

[19] H. Zen, "Statistical Parametric Speech Synthesis," *Autom. Speech Recognit.*, p. 93.

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

[21] T. Van Do, D.-D. Tran, and T.-T. T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology*, 2011, pp. 165–171.

[22] S. Ronanki, M. S. Ribeiro, F. Espic, and O. Watts, "The CSTR entry to the Blizzard Challenge 2017."

[23] T. Q. Cường, "Nghiên Cứu Áp Dụng Kỹ Thuật Học Sâu (Deep Learning) Cho Bài Toán Nhận Dạng Ký Tự Latinh," TRƯỜNG ĐẠI HỌC HÀNG HẢI VIỆT NAM, HẢI PHÒNG, 2016.
[24] M. A. Nielsen, *Neural networks and deep learning*. Determination Press, 2015.
[25] Z.-H. Ling *et al.*, "Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.
[26] N. T. T. Trang, T. D. Dat, A. Rilliard, C. D'alessandro, and P. T. N. Yen, "Intonation Issues In HMM-Based Speech Synthesis For Vietnamese," *St Petersburg*, p. 7, 2014.
[27] D. Jurafsky and J. H. Martin, *Speech and language processing*, vol. 3. Pearson, 2014.
[28] C. King, "• Prof. of Speech Processing • Director of CSTR • Co-author of Festival • CSTR website: www.cstr.ed.ac.uk • Teaching website: speech.zone," p. 424.
[29] H. Kawahara, "Straight, exploitation of the other aspect of Vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.
[30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.
[31] F. Espic, C. V. Botinhao, and S. King, "Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis," 2017, pp. 1383–1387.
[32] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2009.
[33] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, Mar. 2015.
[34] M. Morise, "PLATINUM: A method to extract excitation signals for voice synthesis system," *Acoust. Sci. Technol.*, vol. 33, no. 2, pp. 123–125, 2012.
[35] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.
[36] Q. T. Do, *Vita: A Toolkit for Vietnamese segmentation, chunking, part of speech tagging and morphological analyzer*. 2015.

# 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 sau của 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 hiên tại (tính từ phía trước) |
| B4 | Vị trí của âm tiết trong từ hiên tại (tính từ phía sau) |
| B5 | Vị trí của âm tiết trong cụm từ hiện tại (tính từ phía trước) |
| B6 | Vị trí của âm tiết 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 của từ loại trong từ hiện tại |
| E2 | Số lượng âm vị trong từ hiện tại |
| | |
| F1 | Nhãn của từ loại trong từ tiếp theo |
| F2 | Số lượng âm vị trong từ tiếp theo |
| | |
| G1 | Số lượng âm vị trong cụm phía trước |
| G2 | Số lượng từ trong cụm phía trước |
| | |
| H1 | Số lượng âm vị trong cụm hiện tại |
| H2 | Số lượng từ trong cụm 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 tiếp theo |
| I2 | Số lượng từ trong cụm tiếp theo |
| | |
| J1 | Số lượng âm vị trong câu |
| J2 | Số lượng từ trong câu |
| J3 | Số lượng cụm từ trong câu |

## Phụ lục B: Các công bố khoa học của luận văn

1. **Van-Thinh NGUYEN**, Thi-Ngoc-Diep DO, Dang-Khoa MAC, Eric CASTELLI (2017). *Optimizing data transmission on mobile platform for speech translation system* First Regional Conference on OCR and NLP for ASEAN Languages, Phnom Penh – Cambodia
2. **Van Thinh NGUYEN**, Khac Tan PHAM, Huy Kinh PHAN and Quoc Bao NGUYEN (2018), *Development of a Vietnamese Speech Synthesis System for VLSP 2018,* The Fifth International Workshop on Vietnamese Language and Speech Processing (VLSP 2018), Hanoi, March 2018