**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**BỘ MÔN KỸ THUẬT MÁY TÍNH**

**──────── \* ───────**



**Báo cáo bài tập lớn**

**Môn học: Công nghệ Nhận dạng và Tổng hợp tiếng nói**

**Đề tài: Deep Learning áp dụng vào Tổng hợp tiếng nói**

**GVHD: PGS.TS. Trịnh Văn Loan**

**Sinh viên thực hiện:**

**Lê Minh Trung – 20173422**

**Hà Nội, 6/2021**

**Mục lục**

[**CHƯƠNG 1: GIỚI THIỆU VỀ TỔNG HỢP TIẾNG NÓI** 4](#_Toc74946661)

[**1. Tổng quan về tổng hợp tiếng nói** 4](#_Toc74946662)

[**1.1. Khối xử lý ngôn ngữ tự nhiên** 5](#_Toc74946663)

[**1.2. Khối tổng hợp tín hiệu tiếng nói** 6](#_Toc74946664)

[**2. Phương pháp tổng hợp tiếng nói dựa trên phương pháp học sâu** 6](#_Toc74946665)

[**3. Tình hình phát triển hệ thống tổng hợp tiếng nói ở Việt Nam** 8](#_Toc74946666)

[**CHƯƠNG 2: MẠNG NƠ RON HỌC SÂU** 9](#_Toc74946667)

[**1. Mạng nơ ron thần kinh** 9](#_Toc74946668)

[**2. Mạng nơ ron học sâu** 10](#_Toc74946669)

[**CHƯƠNG 3: HỆ THỐNG TỔNG HỢP TIẾNG NÓI THEO PHƯƠNG PHÁP HỌC SÂU** 11](#_Toc74946670)

[**1. Pha huấn luyện** 12](#_Toc74946671)

[**1.1. Khối trích chọn đặc trưng ngôn ngữ** 12](#_Toc74946672)

[**1.2. Mô hình thời gian** 13](#_Toc74946673)

[**1.3. Mô hình âm học** 13](#_Toc74946674)

[**1.4. Khối trích trọn đặc trưng tiếng nói** 14](#_Toc74946675)

[**2. Pha kiểm thử** 17](#_Toc74946676)

[**2.1. Khối trích trọn đặc trưng ngôn ngữ** 17](#_Toc74946677)

[**2.2. Mô hình thời gian** 17](#_Toc74946678)

[**2.3. Mô hình âm học** 18](#_Toc74946679)

[**2.4. Khối tạo tiếng nói** 19](#_Toc74946680)

[**CHƯƠNG 4: KẾT LUẬN** 21](#_Toc74946681)

[**1. Kết quả đạt được** 21](#_Toc74946682)

[**2. Kết luận** 21](#_Toc74946683)

[**3. Hướng phát triển** 21](#_Toc74946684)

**Danh mục hình ảnh**

[**Hình 1: Tổng quan về hệ thống tổng hợp tiếng nói 4**](#_Toc74946265)

[**Hình 2: Mô hình hệ thống tổng hợp tiếng nói theo phương pháp học sâu 7**](#_Toc74946266)

[**Hình 3: Mạng nơ ron thần kinh 9**](#_Toc74946267)

[**Hình 4: Mạng nơ ron nhân tạo 10**](#_Toc74946268)

[**Hình 5: Kiến trúc hệ thống tổng hợp tiếng nói theo phương pháp học sâu 11**](#_Toc74946269)

[**Hình 6: Mô hình trích xuất đặc trưng ngôn ngữ 13**](#_Toc74946270)

[**Hình 7: Mô hình WORLD vocoder 15**](#_Toc74946271)

[**Hình 8: Đặc trưng Spectral Envelop của tín hiệu tiếng nói 15**](#_Toc74946272)

[**Hình 9: Tần số F0 của tín hiệu tiếng nói 16**](#_Toc74946273)

[**Hình 10: Đặc trưng Aperiodic Energy của tín hiệu tiếng nói 17**](#_Toc74946274)

[**Hình 11: Trích xuất đặc trưng âm thanh 17**](#_Toc74946275)

[**Hình 12: Cấu trúc mạng nơ ron mô hình thời gian 19**](#_Toc74946276)

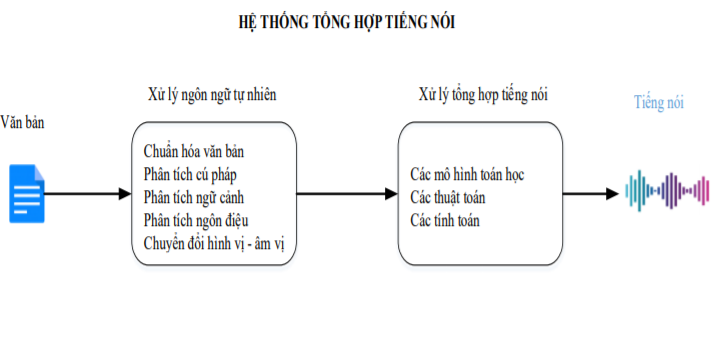
[**Hình 13: Cấu trúc mạng nơ ron mô hình âm học Acoutic 20**](#_Toc74946277)

[**Hình 14: Tổng hợp tiếng nói từ đặc trưng âm học 20**](#_Toc74946278)

# **CHƯƠNG 1: GIỚI THIỆU VỀ TỔNG HỢP TIẾNG NÓI**

## **1. Tổng quan về tổng hợp tiếng nói**

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



Hình 1: Tổng quan về hệ thống tổng hợp tiếng nói

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

### **1.1. Khối xử lý ngôn ngữ tự nhiên**

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

* Xử lý và chuẩn hóa văn bản (Text processing)
* Phân tích cách phát âm (Grapheme to phoneme)
* Phát sinh các thông tin ngôn điệu, ngữ âm cho văn bản (Prosody modeling)

Chuẩn hóa văn bản là quá trình chuyển đổi văn bản thô ban đầu thành một văn bản dạng chuẩn, có thể đọc được một cách dễ dàng, ví dụ như chuyển đổi các số, từ viết tắt và các ký tự đặc biệt… thành dạng viết đầy đủ và chính xác. Đây là một vấn đề rất khó do có nhiều cách đọc khác nhau phụ thuộc vào từng ngữ cảnh, ví dụ như 2020 có thể đọc là “hai nghìn không trăm hai mươi” hoặc “hai nghìn hai mươi” hoặc “hai không hai không”.

Phân tích cách phát âm là quá trình xác định cách phát âm chính xác cho văn bản, các hệ thống tổng hợp tiếng nói dùng hai cách cơ bản để xác định cách phát âm cho văn bản, quá trình này còn được gọi là chuyển đổi văn bản sang chuỗi âm vị. Cách thứ nhất là dựa vào từ điển, sử dụng một từ điển lớn có chứa tất cả các từ và cách phát âm của chúng. Cách thứ hai là dựa trên các quy tắc và sử dụng các quy tắc để tìm ra cách phát âm tương ứng. Mỗi cách đều có các ưu nhược điểm khác nhau, cách dùng từ điển sẽ nhanh và chính xác tuy nhiên không hoạt động được với các từ chưa có trong từ điển và lượng từ vựng cần lưu trữ là lớn. Cách dùng quy tắc phù hợp hơn với mọi văn bản nhưng độ phức tạp có thể tăng cao nếu ngôn ngữ có nhiều bất quy tắc.

Phát sinh các thông tin ngôn điệu cho văn bản là việc xác định vị trí trọng âm của từ được phát âm, sự lên xuống giọng ở các vị trí khác nhau trong câu và xác định các biến thể khác nhau âm phụ thuộc vào ngữ cảnh khi được phát âm trong một ngôn ngữ liên tục, ngoài ra quá trình này còn phải xác định các điểm dừng lấy hơi khi phát âm hoặc đọc một đoạn văn bản. 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). Tần số cơ bản F0 là một đặc trưng quan trọng trong việc tạo ngôn điệu của tín hiệu tiếng nói, tạo ra các đặc trưng cao độ là một vấn đề phức tạp và quan trọng trong tổng hợp tiếng nói.

### **1.2. Khối tổng hợp tín hiệu tiếng nói**

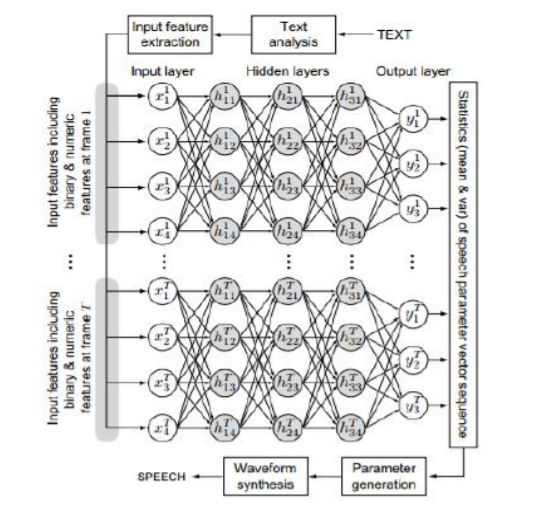
Khối tổng hợp tiếng nói có chức năng tạo ra tiếng nói từ các thông tin về ngữ âm, ngữ điệu do khối xử lý ngôn ngữ tự nhiên cung cấp. Trong thực tế, có hai cách tiếp cận cơ bản liên quan đến tổng hợp tiếng nói: Tổng hợp tiếng nói sử dụng mô hình nguồn âm và tổng hợp dựa trên việc ghép nối các đơn vị âm.

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

Một hệ thống tổng hợp tiếng nói lý tưởng cần vừa tự nhiên, vừa dễ hiểu và mục tiêu xây dựng một hệ thống tổng hợp là làm gia tăng tối đa hai yêu cầu này.

## **2. Phương pháp tổng hợp tiếng nói dựa trên phương pháp học sâu**

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



Hình 2: Mô hình hệ thống tổng hợp tiếng nói theo phương pháp học sâu

Văn bản cần được tổng hợp sẽ đi qua bộ phân tích văn bản để trích chọn các đặc trưng ngôn ngữ học và được chuyển hóa thành các véc tơ nhị phân bởi bộ Input feature extraction, các véc tơ nhị phân đầu vào {xn t} với xn t là đặc trưng thứ n tại khung t (frame t), các véc tơ này tương ứng tạo ra {ym t} các đặc trưng đầu vào thông qua mạng nơ ron DNN đã được huấn luyện, với mỗi ym t là đặc trưng đầu ra thứ m tại khung t. Các đặc trưng đầu ra chứa các thông tin về phổ và tín hiệu kích thích, thông qua bộ tạo tham số (Parameter Generation) sẽ được chuyển thành các tham số đặc trưng âm học và được đưa vào bộ tạo tín hiệu tiếng nói để tạo ra tín hiệu tiếng nói.

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

## **3. Tình hình phát triển hệ thống tổng hợp tiếng nói ở Việt Nam**

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

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

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

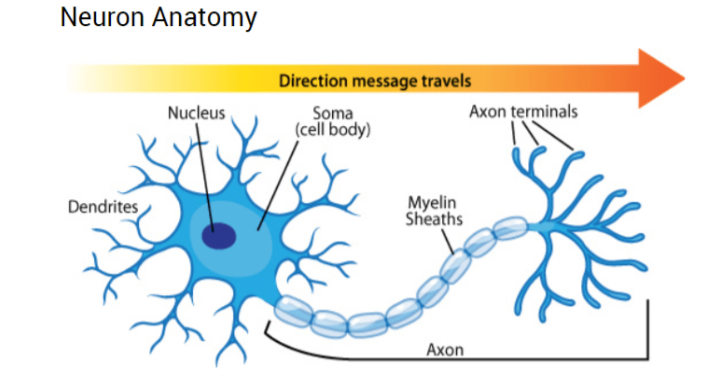
Phương pháp tổng hợp tiếng nói sử dụng tham số thống kê theo mô hình Markov ẩn, ở Việt Nam có nhiều hệ thống phát triển dựa trên phương pháp này như sản phẩn VAIS, sản phẩm của tập đoàn FPT.

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

# **CHƯƠNG 2: MẠNG NƠ RON HỌC SÂU**

## **1. Mạng nơ ron thần kinh**

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



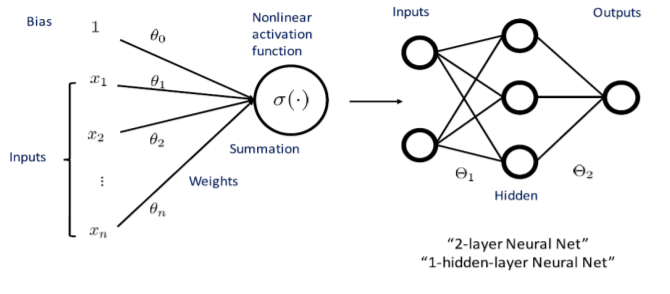
Hình 3: Mạng nơ ron thần kinh

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

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

## **2. Mạng nơ ron học sâu**

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



Hình 4: Mạng nơ ron nhân tạo

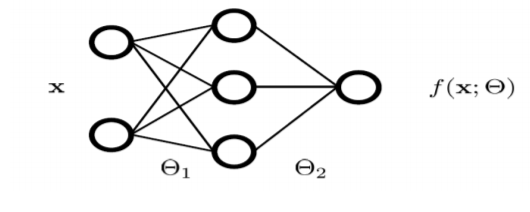
Lớp đầu tiên là input layer, các lớp ở giữa gọi là hidden layer, lớp cuối cùng được gọi là output layer. Tổng số lớp trong mô hình được quy ước là số lớp bớt 1 do không tính input layer.

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

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

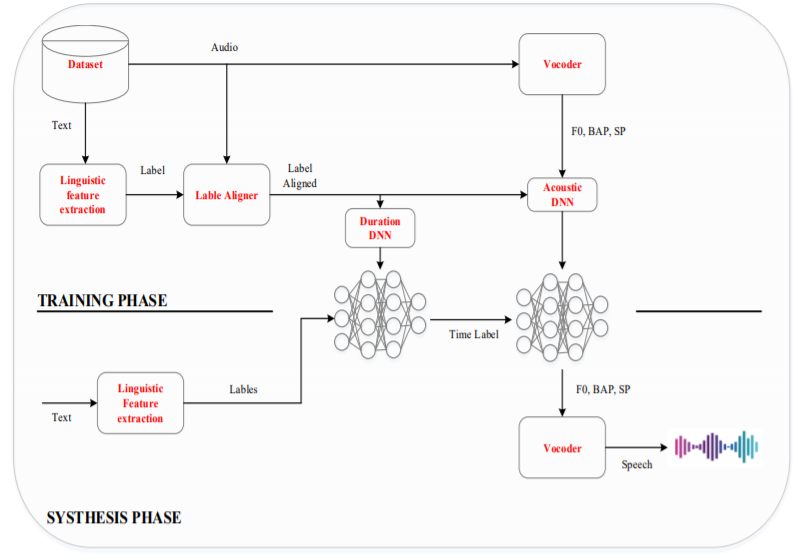
Các dữ liệu và hàm được sử dụng trong mạng nơ ron DNN:

* Dữ liệu huấn luyện {(x1, y1), .., (xn , yn)}; trong đó xi: dữ liệu đầu vào thứ i; yi dữ liệu đích mà DNN muốn sinh ra.
* Hàm số đặc trưng của mạng nơ ron DNN là f (x; Ɵ) ϵ R với biến số là Ɵ



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

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



Hình 5: Kiến trúc hệ thống tổng hợp tiếng nói theo phương pháp học sâu

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

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

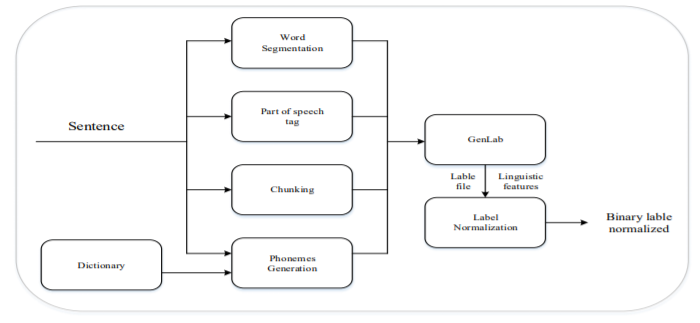
## **1. Pha huấn luyện**

### **1.1. Khối trích chọn đặc trưng ngôn ngữ**

Trước khi đưa vào hệ thống tổng hợp tiếng nói, văn bản được chuẩn hóa để có thể đọc một cách rõ ràng, chuẩn hóa các từ mượn, từ viết tắt, số và ngày tháng bằng bộ từ điển tiếng Việt. Các dấu câu như chấm, phẩy, chấm phẩy được tách ra thành câu độc lập. Khối trích chọn đặc trưng ngôn ngữ thực hiện trích xuất các đặc trưng ngôn ngữ của văn bản đã được chuẩn hóa. Chuỗi các từ đơn của văn bản được chuyển đổi thành chuỗi âm vị, các âm vị sau đó được nhóm lại với nhau theo các từ đơn hoặc các cụm từ có nghĩa. Các thông tin đặc trưng ngôn ngữ bao gồm: part-of-speech tag (loại từ), word segmentation (tách từ), text chunking (cụm từ), âm vị và âm tiết, nhãn thời gian của âm vị. Đặc trưng ngôn ngữ bao gồm các thông tin như sau:

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

Mô hình trích xuất đặc trưng ngôn ngữ như sau:



Hình 6: Mô hình trích xuất đặc trưng ngôn ngữ

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

### **1.2. Mô hình thời gian**

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

* Đưa từng dòng chứa các đặc trưng ngôn ngữ tương ứng với từng âm vị vào trả lời chuỗi các câu hỏi.
* Câu trả lời đúng có giá trị là 1 và câu trả lời sai giá trị là 0, như trong hình vẽ thì âm vị hiện tại là chữ ‘l’ thì đúng với âm vị hiện tại trong dòng ‘l’ nên câu trả lời có giá trị là 1.
* Trả lời hết chuỗi câu hỏi ta được véc tơ đầu vào cho mạng nơ ron

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

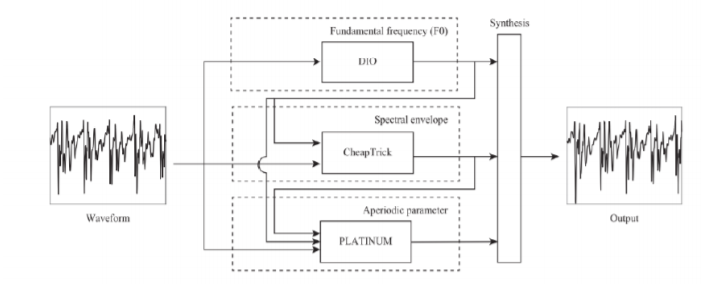
### **1.3. Mô hình âm học**

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

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

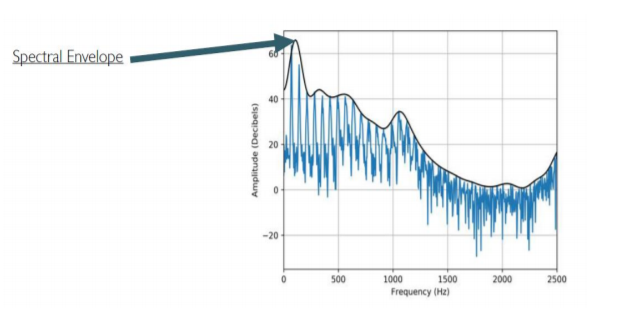
### **1.4. Khối trích trọn đặc trưng tiếng nói**

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



Hình 7: Mô hình WORLD vocoder

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

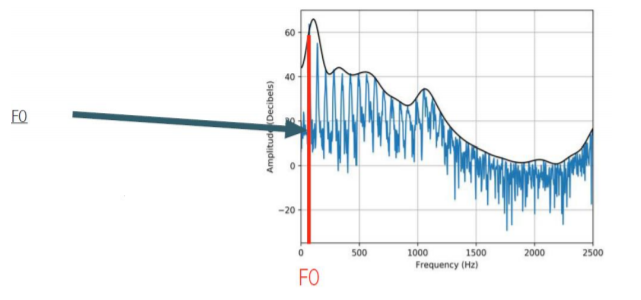


Hình 8: Đặc trưng Spectral Envelop của tín hiệu tiếng nói

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

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

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

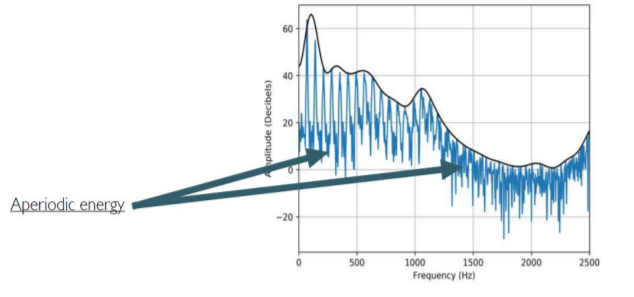


Hình 9: Tần số F0 của tín hiệu tiếng nói

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

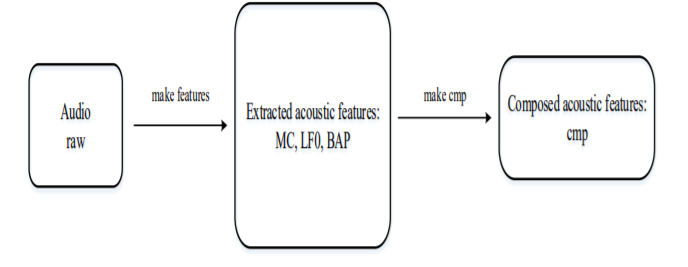
* Sử dụng bộ lọc thông thấp với các tần số cắt khác nhau để lọc tín hiệu, nếu tín hiệu có chứa thành phần tần số cơ bản thì sẽ có dạng hình Sin với chu kỳ T0. Do chưa biết F0 nên sử dụng bộ lọc với tần số cắt khác nhau.
* Tìm các ứng viên cho tần số cơ bản F0 và độ tin cậy của tần số được trích ra.
* Chọn ra ứng viên có độ tin cậy cao nhất là tần số cơ bản F0.

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



Hình 10: Đặc trưng Aperiodic Energy của tín hiệu tiếng nói

Tín hiệu tiếng nói được trích xuất thành các đặc trưng: Hệ số Mel, tần số cơ bản F0, tín hiệu kích thích không tuần hoàn. Gộp 3 đặc trưng thành một và biểu diễn thành véc tơ đặc trưng âm học:



Hình 11: Trích xuất đặc trưng âm thanh

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

## **2. Pha kiểm thử**

### **2.1. Khối trích trọn đặc trưng ngôn ngữ**

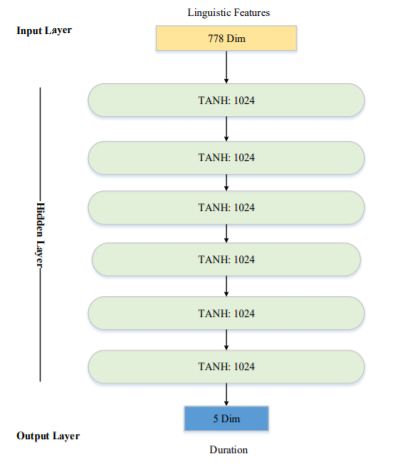
Trong pha tổng hợp, khối trích chọn đặc trưng ngôn ngữ có chức năng trích xuất đặc trưng ngôn ngữ của văn bản cần tạo ra tiếng nói. Đặc trưng ngôn ngữ là các nhãn chứa thông tin về âm vị, được biểu diễn thành các véc tơ đặc trưng ngôn ngữ. Tuy nhiên, đặc trưng ngôn ngữ chưa được gán thời gian bắt đầu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

### **2.2. Mô hình thời gian**

Véc tơ từ khối trích chọn đặc trưng ngôn ngữ được đưa vào mô hình thời gian. Dựa vào mô hình được sinh ra trong pha huấn luyện, âm vị được ước lượng thời gian bắt đầu và thời gian kết thúc của trạng thái khi qua mô hình thời gian. Mô hình thời gian là mạng nơ ron học sâu có đặc điểm như sau:

* Đầu vào là véc tơ 778 chiều chứa các đặc trưng ngôn ngữ của từng âm vị. Với mỗi câu trong tập dữ liệu, số lượng véc tơ đầu vào sẽ là số âm vị có trong câu.
* Có 6 lớp ẩn, mỗi lớp có 1024 neutron và sử dụng hàm TANH là hàm kích hoạt.
* Đầu ra là véc tơ 5 có chiều chứa thông tin ước lượng khoảng thời gian xuất hiện của từng trạng thái trong âm vị. Số lượng véc tơ đầu ra bằng số âm vị có trong câu.

Cấu trúc mô hình thời gian như sau:



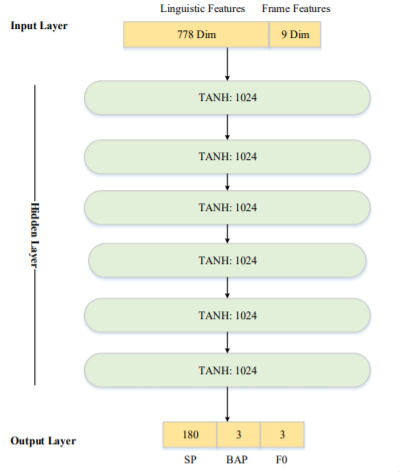
Hình 12: Cấu trúc mạng nơ ron mô hình thời gian

### **2.3. Mô hình âm học**

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

* Véc tơ đầu vào có 787 chiều, trong đó 778 chiều chứa đặc trưng ngôn ngữ của âm vị và 9 chiều để đánh số thứ tự khung (mỗi âm vị được chia nhỏ thành nhiều khung thời gian có độ dài 5ms tương ứng với đặc trưng âm học của WORLD).
* Có 6 lớp ẩn, mỗi lớp có 1024 nơ ron và sử dụng hàm TANH là hàm kích hoạt.
* Đầu ra là véc tơ 186 chiều chứa các đặc trưng âm học được ước lượng bao gồm: Đường bao phổ tín hiệu (SP), (tín hiệu kích thích) BAP, Logarit của tần số cơ bản F0 (log F0), deltas và deltas deltas của 3 đại lượng.

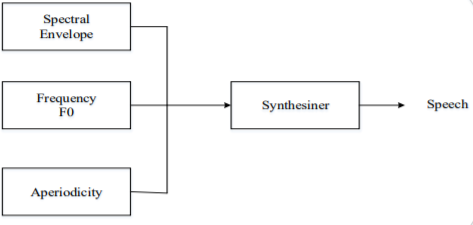
Mô hình âm học có cấu trúc như sau:



Hình 13: Cấu trúc mạng nơ ron mô hình âm học Acoutic

### **2.4. Khối tạo tiếng nói**

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



Hình 14: Tổng hợp tiếng nói từ đặc trưng âm học

# **CHƯƠNG 4: KẾT LUẬN**

## **1. Kết quả đạt được**

Sau khi hoàn thành bài tập lớn, em đã đạt được một số kết quả như sau:

* Nắm vững cơ sở lý thuyết tổng hợp tiếng nói nói chung và tổng hợp tiếng nói tiếng Việt nói riêng.
* Nắm vững cơ sở lý thuyết về mạng nơ ron nhân tạo

## **2. Kết luận**

Học sâu có khả năng tận dụng lượng lớn dữ liệu đào tạo đã trở thành một kỹ thuật quan trọng để tổng hợp giọng nói. Gần đây, ngày càng có nhiều nghiên cứu được thực hiện về các kỹ thuật học sâu hoặc thậm chí là các khuôn khổ đầu cuối và đạt được hiệu suất tối tân. Báo cáo cung cấp một cái nhìn tổng quan về những tiến bộ hiện tại về tổng hợp tiếng nói nói chung và phương pháp tổng hợp tiếng nói theo phương pháp mạng nơ ron học sâu

nói riêng.

## **3. Hướng phát triển**

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

**TÀI LIỆU THAM KHẢO**

1. <http://lib.uet.vnu.edu.vn/bitstream/123456789/1136/1/Bao%20cao%20luan%20van%20thac%20sy_final%20-%20chinh%20sua%20theo%20nhan%20xet%20cua%20HD.pdf?fbclid=IwAR1mV2ontsDefjaH0um_8LXZ6qZHdOcyje2ESQI6kMoesTVSXT42pI7iEQk>
2. <https://www.mdpi.com/2076-3417/9/19/4050/htm?fbclid=IwAR2AX_rvVgn_umqj3jLUDuroG5fhIzzKHsHgJ8XXuz_PL1_-Fb-m-HbSKw8>