**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**ĐỒ ÁN TỐT NGHIỆP**

**Xây dựng mô hình chuyển đổi giọng nói trong tổng hợp tiếng nói Tiếng Việt dựa trên công nghệ học sâu**

**PHAN TRUNG KIÊN**

kien.pt166322@sis.hust.edu.vn

**Ngành Cử nhân Công nghê thông tin**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | PGS. TS. Đỗ Phan Thuận  Chữ ký của GVHD |
| **Bộ môn:** | Khoa học Máy tính |
| **Viện:** | Công nghệ thông tin và truyền thông |

**HÀ NỘI, 6/2020**

**ĐỀ TÀI TỐT NGHIỆP**

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

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

Giáo viên hướng dẫn

Ký và ghi rõ họ tên

**Lời cảm ơn**

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

**Tóm tắt nội dung đồ án**

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

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

Sinh viên thực hiện

Ký và ghi rõ họ tên

**MỤC LỤC**

[CHƯƠNG 1. TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI VÀ CHUYỂN ĐỔI GIỌNG NÓI 1](#_Toc43038839)

[1.1 Giới thiệu về tổng hợp tiếng nói 1](#_Toc43038840)

[1.1.1 Định nghĩa và quá trình phát triển tổng hợp tiếng nói 1](#_Toc43038841)

[1.1.2 Ứng dụng của tổng hợp tiếng nói 1](#_Toc43038842)

[1.1.3 Thành phần của tổng hợp tiếng nói 2](#_Toc43038843)

[1.2 Các phương pháp tổng hợp tiếng nói 3](#_Toc43038844)

[1.2.1 Tổng hợp mô phỏng hệ thống phát âm 3](#_Toc43038845)

[1.2.2 Tổng hợp tần số formant 3](#_Toc43038846)

[1.2.3 Tổng hợp ghép nối 4](#_Toc43038847)

[1.2.4 Tổng hợp dùng tham số thống kê 4](#_Toc43038848)

[1.2.5 Tổng hợp bằng phương pháp lai ghép 4](#_Toc43038849)

[1.2.6 Tổng hợp tiếng nói dựa trên phương pháp học sâu (DNN) 4](#_Toc43038850)

[1.3 Giới thiệu về chuyển đổi giọng nói 4](#_Toc43038851)

[1.4 Tình hình phát triển và các vấn đề với tổng hợp tiếng nói tiếng Việt 4](#_Toc43038852)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 5](#_Toc43038853)

[2.1 Tổng quan về học sâu 5](#_Toc43038854)

[2.1.1 Mạng nơ ron nhân tạo 5](#_Toc43038855)

[2.1.2 Logistic regression 5](#_Toc43038856)

[2.1.3 Mạng nơ ron học sâu 5](#_Toc43038857)

[2.2 Tổng hợp tiếng nói dựa trên công nghệ học sâu 6](#_Toc43038858)

[2.2.1 Trích chọn đặc trưng ngôn ngữ 6](#_Toc43038859)

[2.2.2 Mô hình âm học dựa trên mạng nơ ron học sâu 7](#_Toc43038860)

[2.2.3 Vocoder 7](#_Toc43038861)

[2.3 Chuyển đổi giọng nói dựa trên công nghệ học sâu 10](#_Toc43038862)

[CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT CHUYỂN ĐỔI GIỌNG NÓI TIẾNG VIỆT 11](#_Toc43038863)

[3.1 Mô hình cho quá trình tổng hợp tiếng nói 11](#_Toc43038864)

[3.1.1 Tổng quan mô hình 11](#_Toc43038865)

[3.1.2 Mô đun trích chọn đặc trưng ngôn ngữ 11](#_Toc43038866)

[3.1.3 Mô đun tạo tham số đặc trưng âm học 11](#_Toc43038867)

[3.1.4 Mô đun tổng hợp tiếng nói từ đặc trưng âm học 11](#_Toc43038868)

[3.2 Sử dụng phương pháp Transfer Learning 11](#_Toc43038869)

[3.2.1 Sử dụng mô hình gốc một người nói 11](#_Toc43038870)

[3.2.2 Sử dụng mô hình gốc nhiều người nói 11](#_Toc43038871)

[3.3 Sử dụng vec-tơ định danh người nói 11](#_Toc43038872)

[3.3.1 One-hot encoding 11](#_Toc43038873)

[3.3.2 X-vector 11](#_Toc43038874)

[CHƯƠNG 4. THỬ NGHIỆM VÀ VÀ ĐÁNH GIÁ 12](#_Toc43038875)

[4.1 Dữ liệu 12](#_Toc43038876)

[4.1.1 Chuẩn hóa văn bản 12](#_Toc43038877)

[4.1.2 Phân phối bộ dữ liệu 13](#_Toc43038878)

[CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN 14](#_Toc43038879)

[TÀI LIỆU THAM KHẢO 15](#_Toc43038880)

[PHỤ LỤC 16](#_Toc43038881)

**DANH MỤC HÌNH VẼ**

[Hình 1.1 Sơ đồ tổng quát một hệ thống tổng hợp tiếng nói [\*] 2](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038606)

[Hình 2.2 Thông tin đặc trưng ngôn ngữ ở mức âm vị [\*] 7](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038607)

[Hình 2.1 Biểu diễn đặc trưng ngôn ngữ học của văn bản [\*] 7](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038608)

[Hình 2.3 Tổng quan về hệ thống WORLD vocoder [\*] 8](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038609)

[Hình 2.4 Tổng hợp tiếng nói với WORLD vocoder 10](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038610)

[Hình 4.1 Các bước chuẩn hóa văn bản đầu vào 12](https://husteduvn-my.sharepoint.com/personal/kien_pt166322_sis_hust_edu_vn/Documents/DATN_PhanTrungKien_20166322%20(1).docx" \l "_Toc43038611)

**DANH MỤC HÌNH VẼ**

[Bảng 1.1 Thống kê các thiết bị và giá thành](#_Toc20580109) **[Error! Bookmark not defined.](#_Toc20580109)**

# TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI VÀ VẤN ĐỀ ĐẶT RA CHO ĐỒ ÁN

## Giới thiệu về tổng hợp tiếng nói

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

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

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

Vào đầu thế kỷ XX, sự ra đời của các hệ thống điện đã mang lại một sự thay đổi lớn trong các thiết bị tổng hợp tiếng nói, ví dụ như máy VOCODER của phòng thí nghiệm Bell (1930) được điều khiển bằng bàn phím và có thể phát âm rõ ràng, máy ghi âm và tổng hợp giọng nói của nhà vật lý John Larry Kelly, Jr có thể tạo ra bài hát Daisy Bell với âm nhạc phụ họa bởi Max Mathews.

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

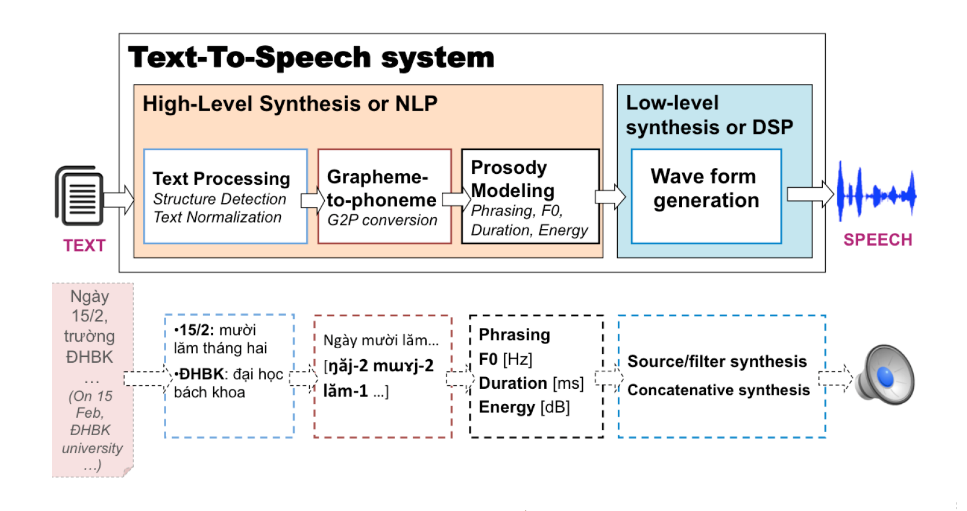
### Ứng dụng của tổng hợp tiếng nói

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

Tiếp theo có thể kể đến các ứng dụng trợ lý ảo như Siri của Apple, Google Assistant của Google, Cotarna của Mircrosoft, ... đều áp dụng công nghệ tổng hợp tiếng nói để nâng cao trải nghiệm tương tác giữa người sử dụng và máy. Hay như các tổng đài trả lời tự động áp dụng các công nghệ nhận dạng tiếng nói, xử lý ngôn ngữ tự nhiên và tổng hợp tiếng nói để giúp khách hàng tránh khỏi tình trạng chờ đợi khi số lượng nhân viên tư vấn là hạn chế.

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

### Thành phần của tổng hợp tiếng nói

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

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

#### Xử lý ngôn ngữ tự nhiên trong tổng hợp tiếng nói

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

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

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

Phân tích cách phát âm là quá trình xác định cách phát âm chính xác cho từng từ trong văn bản, quá trình này còn được gọi là chuyển đổi văn bản sang chuỗi âm vị. Có hai cách cơ bản để xác định cách cho văn bản, cách thứ nhất và cũng là cách đơn giản hơn đó là dựa vào từ điển, sử dụng một từ điển có chứa tất cả các từ của một ngôn ngữ và chưa cách phát âm đúng tương ứng cho mỗi từ. Việc xác định cách phát âm cho văn bản chỉ đơn giản là tra từ điển và thay thế đoạn văn bản bằng chuỗi âm vị đã lưu trong từ điển. Ưu điểm của cách này đó là tốc độ nhanh và tính chính xác, nhưng nhược điểm đó là yêu cầu lượng từ vựng lưu trữ lớn và không hoạt động trong trường hợp từ không có trong từ điểm. Cách thứ hai là dựa trên các quy tắc và sử dụng quy tắc để để tìm ra cách phát âm tương ứng. Cách này phù hợp với mọi văn bản nhưng độ phức tạp có thể tăng cao nếu ngôn ngữ có nhiều trường hợp bất quy tắc.

#### Xử lý tổng hợp tín hiệu tiếng nói

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

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

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

## Các phương pháp tổng hợp tiếng nói

### Tổng hợp mô phỏng hệ thống phát âm

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

### Tổng hợp tần số formant

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

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

### Tổng hợp ghép nối

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

Có ba kiểu tổng hợp ghép nối:

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

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

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

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

### Tổng hợp dùng tham số thống kê HMM

### Tổng hợp bằng phương pháp lai ghép

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

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

## Giới thiệu về chuyển đổi giọng nói

Cùng với sự phát triển của các kỹ thuật tổng hợp tiếng nói trong những năm gần đây, yêu cầu về chất lượng của hệ thống tổng hợp tiếng nói cũng ngày càng nâng cao. Bên cạnh độ tự nhiên, hệ thống tổng hợp tiếng nói cũng được kỳ vọng sẽ có khả năng tạo ra giọng nói của người nói tùy ý với dữ liệu đào tạo tối thiểu. Để đáp ứng vấn đề đó, thích ứng người nói và chuyển đổi giọng nói đã trở thành các hướng nghiên cứu chính trong lĩnh vực tổng hợp tiếng nói [\*]

## Vấn đề đặt ra với đồ án

# CƠ SỞ LÝ THUYẾT

## Tổng quan về học sâu

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

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

### Mạng nơ ron nhân tạo

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

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

### Logistic regression

##### Logistic regression là mô hình neural network đơn giản nhất chỉ với input layer và output layer.

Mô hình của logistic regression là:

Trong đó w là hệ số cần tối ưu và x là dữ liệu đầu vào, b là bias.

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

### Mạng nơ ron học sâu

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

## Tổng hợp tiếng nói dựa trên công nghệ học sâu

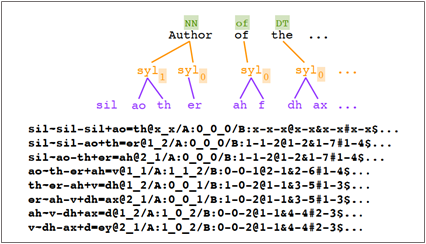
Mô hình âm học dựa trên mô hình Markov ẩn (HMM) và mô hình GMM là hai loại phổ biến nhất được sử dụng trong quá trình tạo tín hiệu tiếng nói từ chuỗi kỹ tự đầu vào (thường là chuỗi âm vị) thông qua việc tạo trực tiếp các đặc trưng âm học của tiếng nói [\*]. Tuy nhiên những mô hình kiểu này có những giới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

Có thể thấy rằng

### Trích chọn đặc trưng ngôn ngữ

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

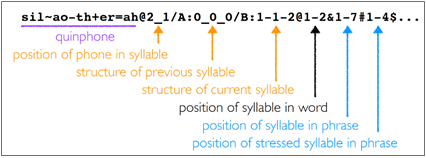
* Văn bản đầu vào sẽ được chuyển thành một chuỗi âm vị nhờ từ điển phiên âm tương ứng với ngôn ngữ đang tổng hợp
* Văn bản đầu vào sẽ được cho qua một hệ thống xử lý ngôn ngữ tự nhiên để trích chọn các thông tin về ngôn ngữ, hệ thống xử lý ngôn ngữ tự nhiên này được xây dựng trên cơ sở ba mô hình: mô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



Hình 2.1 Biểu diễn đặc trưng ngôn ngữ học của văn bản [\*]

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

* Thông tin mức âm vị: bao gồm các âm vị hiện tại, phía trước, phía sau, thông tin về vị trí âm vị trong âm tiết, từ cụm từ, …
* Thông tin mức âm tiết: bao gồm thông tin về số lượng âm vị của âm tiết hiện tại, phía trước, phía sau, thông tin về thanh điệu và vị trí của âm tiết trong từ, cụm từ, …
* Thông tin mức từ: bao gồm các thông tin về nhãn từ loại, số lượng âm tiết của từ hiện tại, và các từ kề nó.
* Thông tin mức cụm từ: bao gồm số lượng từ, âm tiết trong cụm từ hiện tại, phía trước, phía sau.
* Thông tin mức câu: bao gồm thông tin vế số lượng âm tiết, số lượng từ, số lượng cụm từ trong câu.



Hình 2.2 Thông tin đặc trưng ngôn ngữ ở mức âm vị [\*]

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

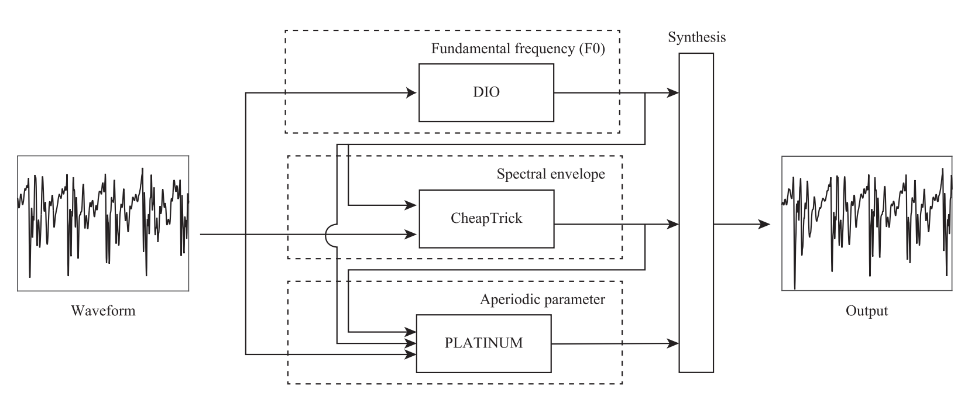
### Mô hình âm học dựa trên mạng nơ ron học sâu

### Vocoder

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

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

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



Hình 2.3 Tổng quan về hệ thống WORLD vocoder [\*]

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

* Sử dụng các bộ lọc thông thấp với các tần số cắt khác nhau để lọc tín hiệu, nếu tín hiệu được lọc nào có chứa thành phần tần số cơ bản thì nó sẽ có dạng hình sin với chu kỳ T0. Bởi vì chưa biết F0, nên ta sử dụng nhiều bộ lọc với các tần số cắt khác nhau.
* Tìm các ứng viên cho tần số cơ bản F0 và độ tin cậy của nó trong mỗi tín hiệu được lọc.
* Chọn ra ứng viên nào có độ tin cậy cao nhất làm F0.

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

|  |  |  |
| --- | --- | --- |
|  |  |  |

Trong đó là tín hiệu và là hàm cửa sổ. Sau khi tính được năng lượng phổ nêu trên, chúng được làm mịn với một cửa sổ chữ nhật có độ dài , như sau:

Với là

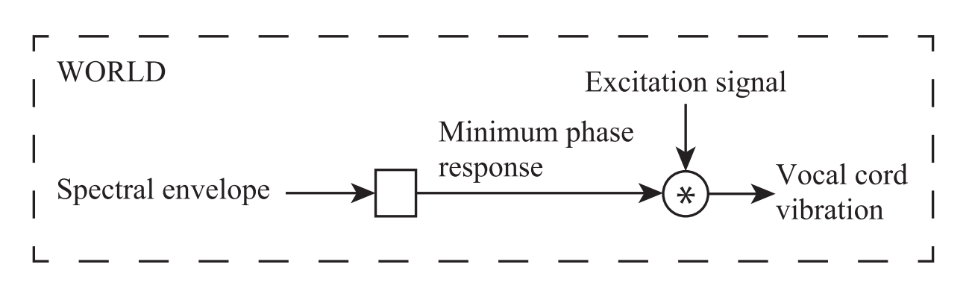
Đường bao phổ được tính như sau:

Trong đó, là hàm nâng cho việc làm mịn logarit năng lượng phổ, là hàm nâng cho việc hồi phục phổ và là các tham số cho việc phục hồi phổ. Các ký hiệu [] và đại diện cho biến đổi Fourier và biến đổi Fourier ngược.

Cuối cùng, phương pháp PLATINUM ước lượng tín hiệu kích thích. Đầu tiên, tín hiệu đi qua cửa sổ có độ dài , phổ của tín hiệu sau khi đưa qua cửa sổ được chia ra bởi phổ tối thiểu . được tính theo biểu thức sau:

Tín hiệu kích thích được biểu diễn như sau:

Sau khi đã có được thông tin đặc trưng cần thiết, âm thanh tổng hợp được tính bằng cách nhân chập tín hiệu kích thích và đáp ứng pha tối thiểu, điều này được minh họa trong Hình 2.4.



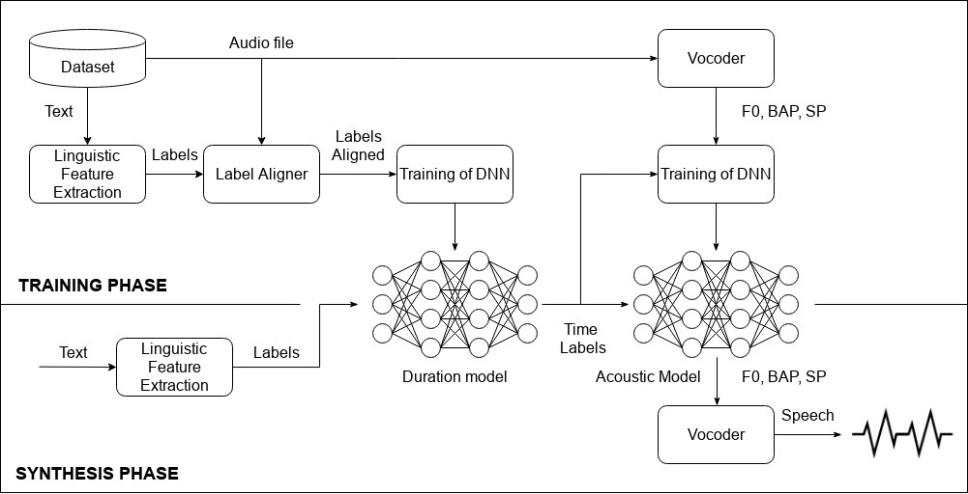
Hình 2.4 Tổng hợp tiếng nói với WORLD vocoder

## Chuyển đổi giọng nói dựa trên công nghệ học sâu

# PHƯƠNG PHÁP ĐỀ XUẤT CHUYỂN ĐỔI GIỌNG NÓI TIẾNG VIỆT

## Mô hình cho quá trình tổng hợp tiếng nói

### Tổng quan mô hình



### Chuẩn bị đặc trưng đầu vào (feature)

Để tạo dữ liệu đầu vào cho mô hình thời gian và mô hình âm học, đồ án sử dụng

### Chuẩn bị nhãn (label)

### Mô hình dự đoán

Hai mô

#### Mô hình thời gian (Duration model)

#### Mô hình âm học (Acoustic model)

### Tổng hợp tiếng nói từ đặc trưng âm học

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

## Sử dụng phương pháp Transfer Learning

### Sử dụng mô hình gốc một người nói

### Sử dụng mô hình gốc nhiều người nói

## Sử dụng vec-tơ định danh người nói

### Tổng quan

### One-hot encoding

Với phương pháp này, đặc trưng đầu vào cho duration model và acoustic model được gắn thêm vec-tơ định danh. Vec-tơ này có dạng độ dài là số lượng người nói trong tập dữ liệu. Trong đó:

* Với mỗi câu trong tập dữ liệu, sẽ bằng 1 nếu người nói là người thứ và bằng 0 với mọi .

### X-vector

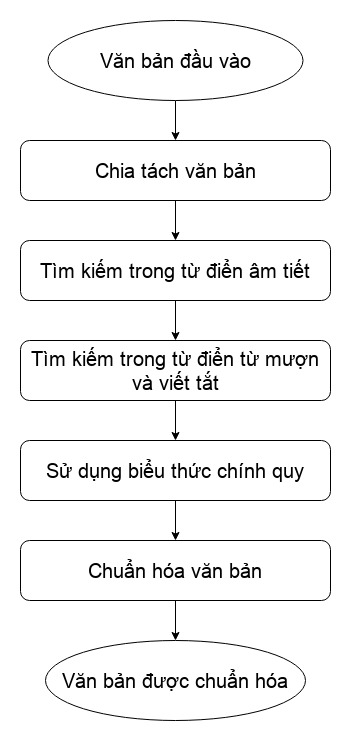
# THỬ NGHIỆM VÀ VÀ ĐÁNH GIÁ

## Dữ liệu

### Chuẩn hóa văn bản

Quá trình chuẩn hóa văn bản đầu vào có nhiệm vụ chính là làm cho văn bản đầu vào có thể đọc được một cách rõ ràng, nhất quán, chuẩn hóa các thành phần không chuẩn như từ mượn, từ viết tắt, số, ngày tháng, ... Các bước chuẩn hóa văn bản được thể hiện trong hình \*\*.

Hình 4.1 Các bước chuẩn hóa văn bản đầu vào

Trong đó:

* Văn bản đầu vào sẽ được phân tách thành các các thành phần dựa theo khoảng trắng, từng thành phần này được tìm kiếm trong từ điểm âm tiết, nếu có trong từ điển thì nó là thành phần có thể đọc được, nếu không có sẽ tiếp tục được tìm kiếm trong từ điển từ mượn, từ viết tắt
* Những thành phần không có trong từ điển âm tiết được tìm kiếm ở trong từ điển viết tắt, nếu được tìm thấy thì thành phần này sẽ được chuyển thành một chuỗi các từ chuẩn theo từ điển âm tiết, nếu không tìm thấy sẽ được chuyển sang bước tiếp theo.
* Áp dụng biểu thức chính quy:

### Phân phối bộ dữ liệu

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Trần Bách, Lưới điện và hệ thống điện, Nhà xuất bản Khoa học Kỹ thuật, 2004. |
| [2] | Abe Masayuki, “A Practical Approach to Accurate Fault Location on Extra High Voltage Teed Feeders,” *IEEE Transaction on Power Delivery,* pp. 159-168, 1995. |
| [3] | Microsoft, "Add citations in a Word document," 2017. |

# PHỤ LỤC

1. **Chi tiết số liệu thí nghiệm**

Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có).

1. **Chi tiết các bước tính toán**

Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có).

1. **Chi tiết sơ đồ mô phỏng**

Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có). Trình phụ lục tại đây (nếu có).