**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**ĐỒ ÁN TỐT NGHIỆP**

**Xây dựng mô hình tổng hợp tiếng nói và chuyển đổi giọng 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. CÁC QUI ĐỊNH CHUNG 1](#_Toc20584278)

[1.1 Giới thiệu chung 1](#_Toc20584279)

[1.2 Sử dụng các định dạng văn bản theo qui định 1](#_Toc20584280)

[1.2.1 Qui định về căn lề văn bản 1](#_Toc20584281)

[1.2.2 Tạo chương mới 3](#_Toc20584282)

[1.2.3 Tạo tiêu đề các cấp 3](#_Toc20584283)

[1.2.4 Định dạng phần nội dung các chương, mục 3](#_Toc20584284)

[1.2.5 Hình vẽ - Đồ thị 4](#_Toc20584285)

[1.2.6 Bảng biểu 6](#_Toc20584286)

[1.2.7 Phương trình 8](#_Toc20584287)

[1.3 Tạo tham chiếu chéo giữa các đoạn văn bản 11](#_Toc20584288)

[1.4 Tạo danh mục tài liệu tham khảo 11](#_Toc20584289)

[1.5 Cập nhật lại các chú thích và tham chiếu 15](#_Toc20584290)

[1.6 Tạo danh mục hình vẽ 15](#_Toc20584291)

[1.7 Tạo danh mục bảng biểu 16](#_Toc20584292)

[1.8 Tạo trang mục lục 16](#_Toc20584293)

[1.9 Qui cách đóng quyển 17](#_Toc20584294)

[CHƯƠNG 2. SỬ DỤNG CÁC BIỂU ĐỒ 19](#_Toc20584295)

[2.1 Giới thiệu về biểu diễn bằng đồ thị 19](#_Toc20584296)

[2.2 Đồ thị kiểu bánh 19](#_Toc20584297)

[2.3 Đồ thị kiểu thanh ngang 20](#_Toc20584298)

[2.4 Đồ thị kiểu cột đứng 20](#_Toc20584299)

[2.5 Đồ thị kiểu đường 21](#_Toc20584300)

[2.6 Đồ thị kiểu diện tích 21](#_Toc20584301)

[CHƯƠNG 3. KẾT LUẬN 23](#_Toc20584302)

[3.1 Kết luận 23](#_Toc20584303)

[3.2 Hướng phát triển của đồ án trong tương lai 23](#_Toc20584304)

[TÀI LIỆU THAM KHẢO 24](#_Toc20584305)

[PHỤ LỤC 25](#_Toc20584306)

**DANH MỤC HÌNH VẼ**

[Hình 2.1 Đồ thị kiểu bánh 19](#_Toc20580104)

[Hình 2.2 Đồ thị kiểu thanh ngang 19](#_Toc20580105)

[Hình 2.3 Đồ thị kiểu cột đứng 20](#_Toc20580106)

[Hình 2.4 Đồ thị kiểu đường 20](#_Toc20580107)

[Hình 2.5 Đồ thị kiểu diện tích 21](#_Toc20580108)

**DANH MỤC HÌNH VẼ**

[Bảng 1.1 Thống kê các thiết bị và giá thành 8](#_Toc20580109)

# TỔNG QUAN VỀ TỔNG HỢP TIẾNG NÓI

## 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 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 [1]. 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. 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) [2], phiên bản cải tiến VODER của Homer Dudley (1939) [3]. Từ đó đến nay, công nghệ tổng hợp tiếng nói đã có những bước tiến bộ vượt bậc, chất lượng giọng nói tổng hợp ngày càng có độ tự nhiên và dễ nghe, bên cạnh đó các ứng dụng của nó cũng được áp dụng rộng rãi.

### Ứng dụng của tổng hợp tiếng nói

### 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 [4]. Phần xử lý ngôn ngữ tự nhiên: 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: Tạo ra tín hiệu tiếng nói từ các thành phầm phát âm được nêu trên [5]. Một hệ thống tổng hợp tiếng nói gồm hai thành phần trên được mô tả ở hình 1.

##### 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 [6]. 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

##### 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. [6]

## 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 [6]. 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 phaircos một không gian rất lớn và tốc độ tính toán, truy vấn của hệ thống mạnh, do đó điều này là không kinh tế.

### Tổng hợp dùng tham số thống kê

### 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

# CHUYỂN ĐỔI GIỌNG NÓI TRONG TỔNG HỢP TIẾNG NÓI VÀ VẤN ĐỀ ĐẶT RA VỚI ĐỒ ÁN

## Tổng quan 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 [\*]

## Các phương pháp chuyển đổi giọng nói

## Vấn đề với chuyển đổi giọng nói tiếng Việt

# CÔNG NGHỆ HỌC SÂU TRONG CHUYỂN ĐỔI GIỌNG NÓI

## Công nghệ học sâu trong tổng hợp tiếng nói

### 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.

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

### Tổng quan về mạng nơ ron

##### Logistic regression

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

##### DNN

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

## Công nghệ học sâu trong chuyển đổi giọng nói

# PHƯƠNG PHÁP ĐỀ XUẤT CHUYỂN ĐỔI GIỌNG NÓI TIẾNG VIỆT

## 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

### One-hot encoding

### X-vector

# THỬ NGHIỆM VÀ VÀ ĐÁNH GIÁ

## Dữ liệu

### Chuẩn bị

### Phân phối bộ dữ liệu

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỒ ÁN

# TÀI LIỆU THAM KHẢO

# 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).