Mục lục

Danh mục hình vẽ

1. Tính cấp thiết

Hiện nay, tổng hợp tiếng nói có nhiều ứng dụng thực tiễn trong cuộc sống. Ví dụ như hệ thống này có thể giúp người có thị lực kém (hoặc khiếm thị) nghe được máy đọc ra văn bản; đặc biệt là các văn bản có thể xử lý trên máy tính. Hệ thống như vậy có thể lắp đặt trong phần mềm xử lý văn bản hay trình duyệt mạng.

Trong phạm vi của báo cáo này, tôi trình bày tìm hiểu về tổng hợp tiếng nói theo phương pháp học sâu.

2. Mục tiêu

Nắm được tổng hợp tiếng nói theo phương pháp học sâu.

3. Cách tiếp cận

Đề tài tiếp cận vấn đề nhằm đạt đến mục tiêu từ các góc độ sau:

- Tiếp cận từ thực tiễn: khảo sát, đánh giá, phát hiện những điểm yếu hiện tại trong các phương pháp tổng hợp tiếng nói
- Tiếp cận từ cơ sở lý thuyết: khảo sát, đánh giá mức độ tốt của các phương pháp tổng hợp, để từ đó cải tiến, đề xuất định dạng mới, phù hợp với mục tiêu đề ra
- Tiếp cận từ những xu hướng phát triển công nghệ hiện đại

4. Phương pháp nghiên cứu:

Phương pháp phân tích lý thuyết: Nghiên cứu Lý thuyết tổng quan về tổng hợp tiếng nói.

Phương pháp thực nghiệm: thực hiện trực tiếp các tính năng mới, phân tích đánh giá kết quả.

5. Phạm vi nghiên cứu:

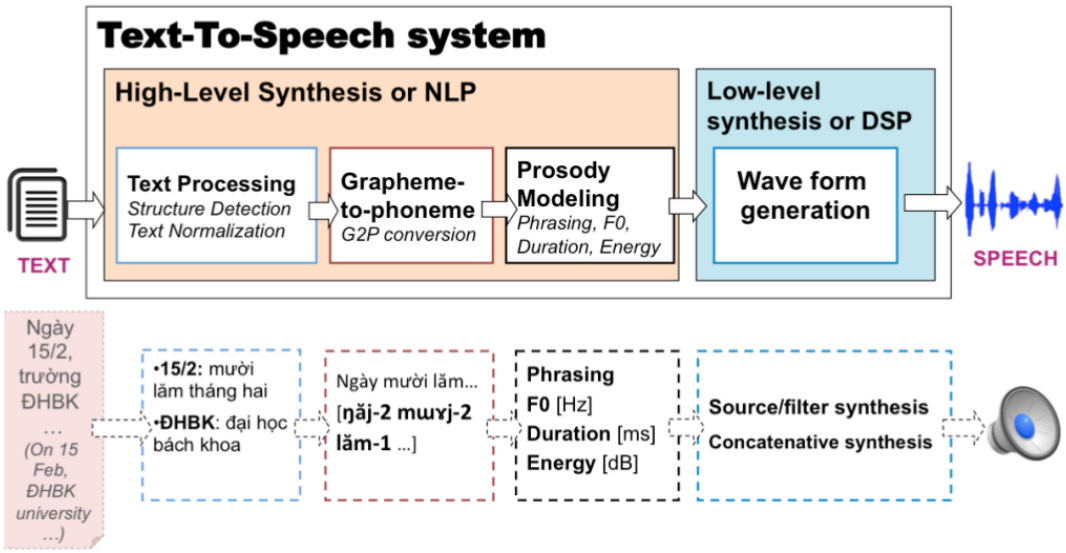
Phương pháp tổng hợp tiếng nói theo phương pháp học sâu.

6. Nội dung nghiên cứu và kết quả đạt được:

6.1. 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. Hiện nay, để xây dựng một hệ tổng hợp tiếng nói, ta có thể sử dụng các phương pháp tổng hợp như sau: Phương pháp tổng hợp mô phỏng hệ thống phát âm, Phương pháp tổng hợp tần số formant, Phương pháp tổng hợp dựa trên ghép nối, Phương pháp tổng hợp dựa trên tham số v.v.

Qua quá trình phát triển, hiện nay về cơ bản một hệ thống tổng hợp tiếng nói bao gồm hai thành phần chính: (1) khối phân tích xử lý ngôn ngữ tự nhiên hay còn gọi là khối tổng hợp mức cao; và (2) là khối xử lý tổng hợp tiếng nói có nhiệm vụ tổng hợp tiếng nói hay còn gọi là khối tổng hợp mức thấp.



Hình 1. Sơ đồ tổng quát một hệ thống tổng hợp tiếng nói (nguồn (Trang et al. 2014))

Tổng hợp mức cao có nhiệm vụ chuyển đổi các ký tự văn bản đầu vào thành một dạng chuỗi các ngữ âm đã được thiết kế trước của hệ thống TTS. Nghĩa là, chuyển đổi chuỗi văn bản đầu vào thành dạng biểu diễn ngữ âm, xác định cách đọc nội dung văn bản. Quá trình này đòi hỏi khả năng dự đoán ngôn điệu từ văn bản đầu vào với thông tin ngữ âm và ngữ điệu tương ứng. Từ các thông tin ngôn điệu và ngữ âm là chuỗi các nhãn phụ thuộc ngữ cảnh mức âm vị của văn bản đầu vào. Khối tổng hợp mức thấp sẽ chọn các tham số thích hợp từ tập các giá trị tần số cơ bản, phổ tín hiệu, trường độ âm thanh (bao gồm âm vị và âm tiết). Sau đó tiếng nói ở dạng sóng tín hiệu sẽ được tạo ra bằng một kĩ thuật tổng hợp.

<u>Khối xử lý ngôn ngữ tự nhiên</u>

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 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 thường bao gồm ba 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).

Mô hình hóa các thông tin ngôn điệu, ngữ âm cho văn bản (Prosody modeling).

Chuẩn hoá văn bản

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, v.v. thành dạng viết đầy đủ và chính xác. Chuẩn hoá văn bản quyết định xem làm thế nào có thể đọc được những từ không chuẩn, những từ này vốn là những từ mà không thể áp dụng quy tắc "ký tự - thành – âm thanh" chẳng hạn như từ "Nato" (Na tô), "WTO" (vê kép tê ô). Quá trình này có vai trò quan trong đến việc quyết định chất lượng của một hệ tổng hợp tiếng. Hầu hết các văn bản không phải lúc nào cũng bao gồm toàn những từ ở dạng chuẩn có thế phát âm chính xác, chúng thường chứa nhiều cấu trúc đặc biệt, những từ khồng chuẩn mà không thể đọc bằng việc áp dụng quy tắc "ký tự - thành – âm thanh" thông thường. Ví dụ, chúng có thể là những số liệu, những chữ viết tắt (như GD viết tắt cho "Giáo Dục"), những cấu trúc về biểu diễn thời gian và ngày tháng (như 10h30 hoặc 1/1/2016), tên nước ngoài hoặc địa danh (như New York), những chữ số La Mã, v.v.

Chuẩn hóa văn bản là một vấn đề khó với nhiều nhập nhằng trong cách đọc. Trong một số ngôn ngữ khác nhau, các từ có thể được phát âm khác nhau tuỳ theo ngữ cảnh. Ví dụ, với việc chọn phát âm chữ số cũng là một vấn đề. Lý do ở đây là do có nhiều cách phát âm chữ số khác nhau và phụ thuộc vào ngữ cảnh khác nhau. Ví dụ số 12345 có thể được đọc là "Mười hai nghìn ba trăm bốn mươi lăm" nếu nó là số tự nhiên chỉ về số lượng, nhưng cũng có thế được đọc là "một hai ba bốn năm" nếu nó là mã số tài khoản. Khi đó thì hệ thống tổng hợp phải có nhiệm vụ đoán văn cảnh bằng việc quan sát các từ kế cận, các số hay dấu câu bên cạnh, hoặc dùng trường hợp mặc định khi không thể phân định.

Chuyển đổi hình vị sang âm vị (Phân tích cách phát âm)

Chuyền đổi từ hình vị sang âm vị nhằm mục đích xác định đúng cách phát âm của hình vị đó. 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. 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 phát 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.

Mô hình hóa các thông tin ngôn điệu

Xác định đúng được ngữ điệu, trọng âm và khoảng thời gian giữa các tiếng từ văn bản viết là những vấn đề khó khăn nhất đối với một hệ tổng hợp tiếng. Các đặc tính này được gọi là ngôn điệu và có thể được xem xét như là giai điệu, nhịp điệu và sự nhấn mạnh của tiếng nói ở mức cảm giác. Ngữ điệu có nghĩa là sự thay đổi của tần số cơ bản trong thời gian nói. Ngôn điệu của tiếng nói liên tục phụ thuộc vào nhiều yếu tố như nghĩa của các câu, đặc trưng và cảm xúc của người nói. Mô hình hóa 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. 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.

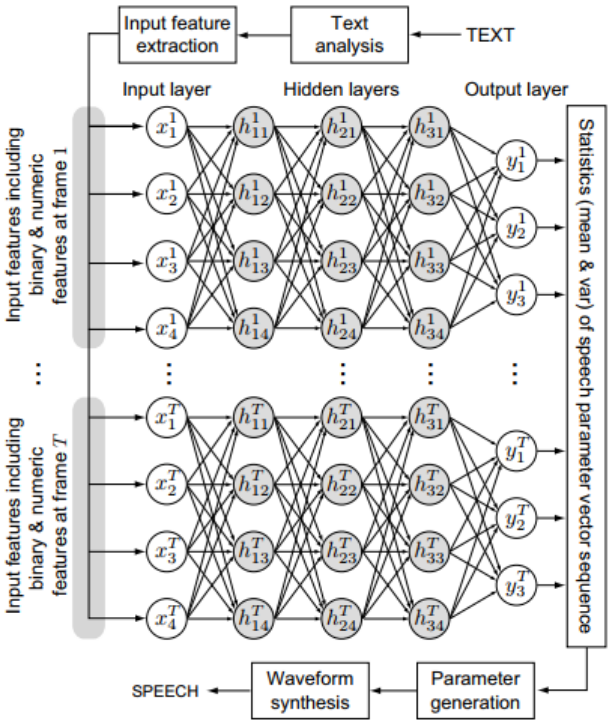
Khối xử lý tổng hợp tiếng nói

Chuỗi các nhãn của văn bản và thông tin ngôn điệu của nó được đưa sang khối xử lý tổng hợp sau khi qua khối xử lý ngôn ngữ tự nhiên của hệ thống TTS. Tại đây, các thành phần chức năng của khối này có nhiệm vụ tạo ra dạng sóng tín hiệu tiếng nói. Tiếng nói có thể được sinh ra theo nhiều cách khác nhau và các phương pháp tổng hợp có thể được áp dụng tuỳ theo vào tiêu chí cụ thể. Việc phân loại phương pháp tổng hợp tiếng nói cơ bản phụ thuộc vào tiếng nói được tổng hợp từ từ các tham số nhân tạo (các tần số formant), hay từ các mẫu tiếng nói được thu âm trước. Các phương pháp tổng hợp tiếng nói có thể kể tới gồm: Phương pháp tổng hợp mô phỏng hệ thống phát âm, Phương pháp tổng hợp tần số formant, Phương pháp tổng hợp dựa trên ghép nối, Phương pháp tổng hợp dựa trên tham số (Chi tiết các phương pháp được trình bày trong công việc 5.4.).

6.2 Tổng hợp tiếng nói theo 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 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, v.v. Sau đó mô hình âm học dựa trên DNN (thay vì HMM) 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 **Error! Reference source not found.**. 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.

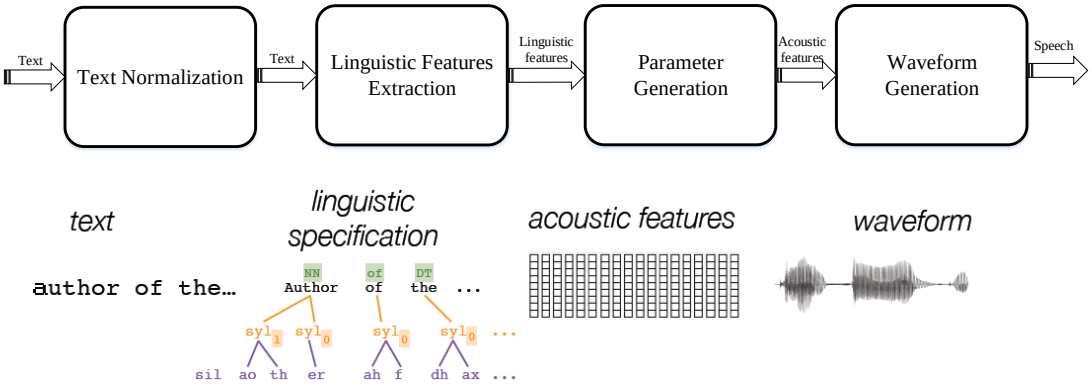


Hình 2. Tổng hợp tiếng nói dựa trên DNN (Ze et al. 2013)

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.

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 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 hay các sản phẩm của Google, Baidu, v.v. Vì vậy để lựa chọn một phương pháp có khả năng áp dụng trên nhiều ngôn ngữ theo yêu cầu của đề tài, nhóm thực hiện đề tài lựa chọn triển khai phương pháp tổng hợp tiếng nói dựa trên phương pháp học sâu.

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) (Ze et al. 2013). **Error! Reference source not found.** được thể hiện lại trong phương pháp này như sau:



*Hình 3. **Mô hình chung tổng hợp tiếng nói dựa trên phương pháp học sâu** (Simon King et al. 2017)*

Một hệ thống tổng hợp tiếng nói gồm các mô đun chính và đây cũng là các mô đun trong tổng hợp tiếng nói dựa trên công nghệ học sâu:

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 nước ngoài sang dạng phiên âm,…

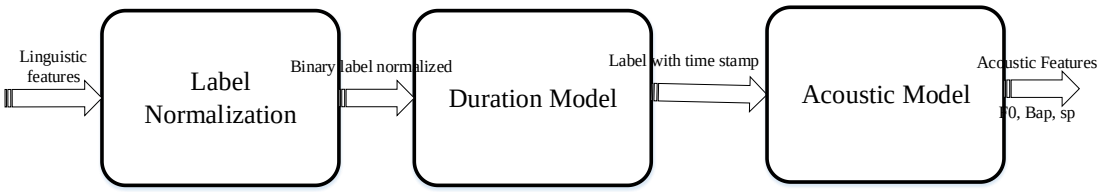
Mô đun trích chọn đặc trưng ngôn ngữ: văn bản đầu vào được xử lý, phân tích và trích chọn bởi bộ Linguistic Features Extraction ra thành các vec tơ đặc trưng ngôn ngữ học, các vec tơ này thường bao gồm các thông tin về chuỗi âm vị, vị trí tương đối của âm vị trong câu, cụm từ hay từ, số lượng âm vị trong câu, trong cụm từ hay trong từ,…

Bộ Parameter Generation 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. 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 (hay gọi là Vocoder).

Mô đun tạo tham tham số đặc trưng âm học (Parameter Generation) 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 **Error! Reference source not found.**, trong đó mô đun này được cấu tạo bởi ba phần chính đó là bộ chuẩn hóa đặc trưng đầu vào (Label Normalization), mô hình khoảng thời gian (Duration Model), mô hình âm học (Acoustic model).



Hình 4. Cấu trúc mô đun tạo tham số đặc trưng

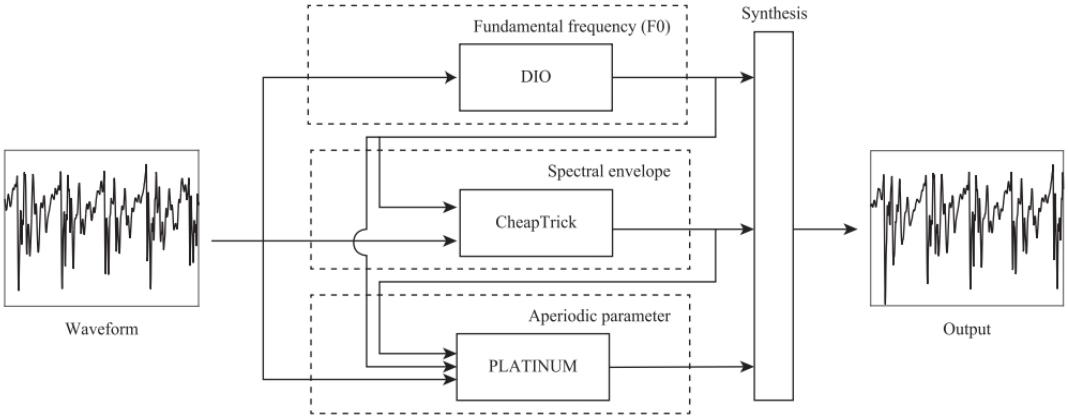
Mô hình khoảng thời gian (Duration Model), nhận đầu vào là các véc tơ đặc trưng ngôn ngữ học, và đầu ra của bộ này là các véc tơ đặc trưng ngôn ngữ học cộng thêm với các thông tin về thời gian xuất hiện (thời điểm bắt đầu và kết thúc) của mỗi âm vị. Mô đun này được huấn luyện bằng một mô hình mạng nơ ron với mỗi đầu vào là các véc tơ đặc trưng ngôn ngữ học và đầu ra là thông tin về thời gian xuất hiện của âm vị tương ứng.

Mô hình âm học (Acoustic Model), 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. Mô hình âm học này cũng được mô hình hóa sử dụng một mạng nơ ron học sâu.

Mô đun tạo tín hiệu tiếng nói (hay gọi là 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.

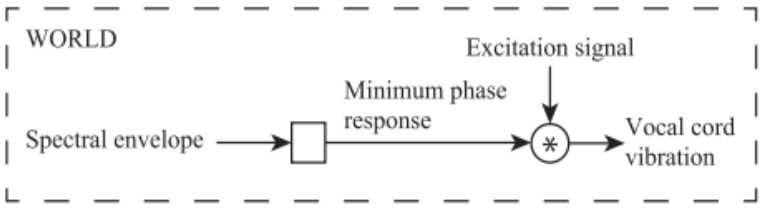
Có rất nhiều loại vocoder khác nhau được phát triển để cải thiện chất lượng phân tích và tổng hợp tiếng nói như Straight vocoder (Kawahara 2006), World vocoder (Morise et al. 2016), Magphase vocoder (Espic et al. 2017),… Trong phần này sẽ chỉ trình bày về một vocoder vô cùng mạnh mẽ, được phát triển 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 báo cáo này, đó là WORLD vocoder.

WORLD vocoder được sử dụng để trích chọn các đặc trưng âm học và tổng hợp tiếng nói từ những đặc trưng này, bao gồm: Đường bao phổ của tín hiệu, Các thành phần không tuần hoàn (Aperiodicities), và tần số cơ bản F0.



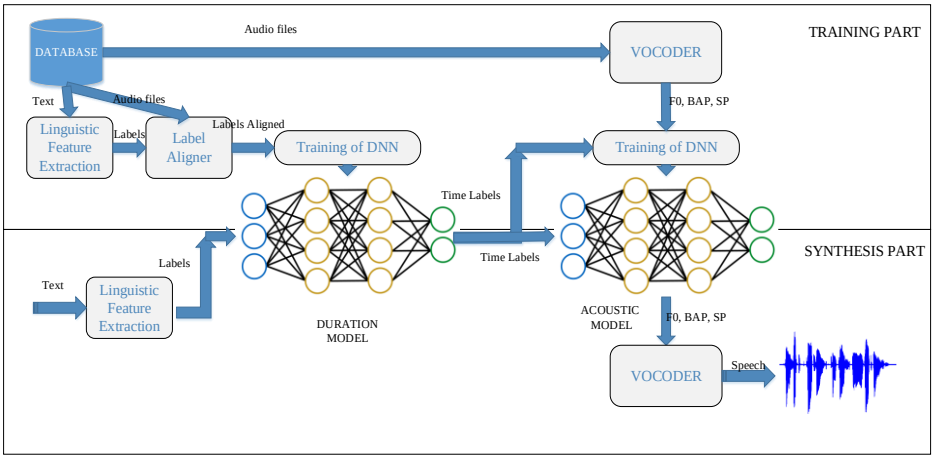
Hình 5. Tổng quan về hệ thống WORLD vocoder (Morise et al. 2016)

Error! Reference source not found. mô tả quá trình xử lý của WORLD vocoder trong hai giai đoạn phân tích và tổng hợp tín hiệu tiếng nói. Trong giai đoạn phân tích, tần số cơ bản F0 được ước lượng bởi phương pháp DIO (Morise et al. 2009), đường bao phổ được ước lượng bởi phương pháp CheapTrick(Morise 2015), và tín hiệu kích được ước lượng bởi phương pháp PLATINUM (Morise 2012), nó được sử dụng như một tham số không tuần hoàn. Trong giai đoạn tổng hợp, âm thanh tổng hợp được tính bằng cách nhân chập tín hiệu kích thích và đáp ứng pha tối thiểu (**Error! Reference source not found.**).



Hình 6. Tổng hợp tiếng nói với WORLD vocoder (Morise et al. 2016)

Như vậy có thể thấy được 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 (**Error! Reference source not found.**).



Hình 7. Quá trình huấn luyện và tổng hợp một hệ thống tổng hợp tiếng nói dựa trên mô hình mạng nơ ron học sâu

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 huấn luyện 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, đầu ra sẽ là các đặc trưng ngôn ngữ học được biểu diễn dưới dạng các nhãn. 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 force alignment. 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. Hai thông tin này được đưa vào huấn luyện mô hình khoảng thời gian Duration Model: với đầu vào là các nhãn đặc trưng âm học và đầu ra mạng là thời gian xuất hiện của từng âm vị tương ứng với nhãn. 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 mới của âm vị 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. 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) được đưa vào huấn luyện cho mô hình âm học (Acoustic model): đầu vào là nhãn mang thông tin về đặc trưng ngôn ngữ và thời gian xuất hiện của âm vị, đầu ra là đặc trưng âm học từ các tệp âm thanh.

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.

6.3. Kết luận và kiến nghị

- Phương pháp tổng hợp tiếng nói theo phương pháp học sâu là một phương pháp hiện đại, có chất lượng tốt cho hệ thống TTS.

- Tôi đề nghị sinh viên nghiên cứu về tổng hợp tiếng nói theo phương pháp học sâu để giúp hữu ích cho các môn Trí tuệ nhân tạo, Hệ chuyên gia.

Danh sách từ viết tắt

Chữ cái viết tắt/ Ký hiệu	Tiếng Anh	Tiếng Việt
TTS	Text to speech	Tổng hợp tiếng nói
VB	Visual Basic	Ngôn ngữ lập trình VB

Tài liệu tham khảo

Espic F, Botinhao CV, King S (2017) Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis. ISCA, pp 1383–1387

Kawahara H (2006) Straight, exploitation of the other aspect of Vocoder: Perceptually isomorphic decomposition of speech sounds. Acoust Sci Technol 27:349–353. https://doi.org/10.1250/ast.27.349

Morise M (2015) CheapTrick, a spectral envelope estimator for high-quality speech synthesis. Speech Commun 67:1–7. https://doi.org/10.1016/j.specom.2014.09.003

Morise M (2012) PLATINUM: A method to extract excitation signals for voice synthesis system. Acoust Sci Technol 33:123–125. https://doi.org/10.1250/ast.33.123

Morise M, Kawahara H, Katayose H (2009) 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. Audio Engineering Society

Morise M, Yokomori F, Ozawa K (2016) WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Trans Inf Syst E99.D:1877–1884. https://doi.org/10.1587/transinf.2015EDP7457

Simon King, Oliver Watts, Srikanth Ronanki, Felipe Espic, Zhizheng Wu (2017) Deep Learning for Text-to-Speech Synthesis, using the Merlin toolkit

Trang NTT, Rilliard A, D'Alessandro C (2014) Prosodic phrasing modeling for Vietnamese TTS using syntactic information. INTERSPEECH

Ze H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. IEEE, pp 7962–7966