Tổng quan về tổng hợp tiếng nói

(Biên soạn – Nguyễn Văn Thịnh, Trung tâm Không Gian Mạng Viettel VTCC)

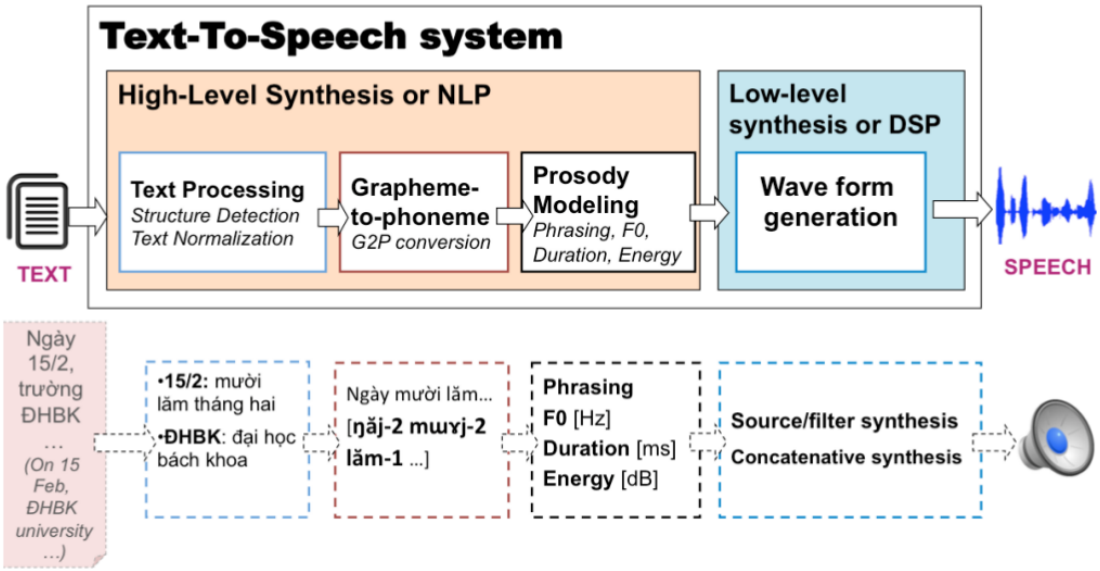
Mục lục

1.1 Giới thiệu về tổng hợp tiếng nói

1.1.1 Tổng quan về tổng hợp tiếng nói

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

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



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

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

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

Trong một hệ thống tổng hợp tiếng nói, khối xử lý ngôn ngữ tự nhiên có nhiệm vụ trích chọn các thông tin về ngữ âm, ngữ điệu của văn bản đầu vào. Thông tin ngữ âm cho biết những âm nào được phát ra trong hoàn cảnh cụ thể nào, thông tin ngữ điệu mô tả điệu tính của các âm được phát[7]. Quá trình xử lý ngôn ngữ tự nhiên thường bao gồm ba bước (xem trên hình 1):

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

Chuẩn hóa văn bản là quá trình chuyển hóa văn bản thô ban đầu thành một văn bản dạng chuẩn, có thể đọc được một cách dễ dàng, ví dụ như chuyển đổi các số, từ viết tắt, ký tự đặc biệt,… thành dạng viết đầy đủ và chính xác. Chuẩn hóa văn bản là một vấn đề khó với nhiều nhập nhằng trong cách đọc, ví như chữ số có nhiều cách đọc khác nhau tùy theo văn cảnh khác nhau, như 3579 có thể được đọc là "ba nghìn năm trăm bảy chín" nếu coi nó là một số nhưng cũng có thể đọc là "ba năm bảy chín" nếu như nó là một mã xác thực, các từ viết tắt cũng vậy, cũng có nhiều cách đọc phụ thuộc vào quy ước của người viết.

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

Phát sinh các thông tin ngôn điệu cho văn bản là việc xác định vị trí trọng âm của từ được phát âm, sự lên xuống giọng ở các vị trí khác nhau trong câu và xác định các biến thể khác nhau của âm phụ thuộc vào ngữ cảnh khi được phát âm trong một ngôn ngữ lưu liên tục, ngoài ra quá trình này còn phải xác định các điểm dừng nghỉ lấy hơi khi phát âm hoặc đọc một đoạn văn bản[10]. Thông tin về thời gian (duration) được đo bằng đơn vị xen ti giây (centi second) hoặc mi li giây (mili second), và được ước lượng dựa trên các quy tắc hoặc các thuật toán học máy. Cao độ (pitch) là một tương quan về mặt cảm nhận của tần số cơ bản F0, được biểu thị theo đơn vị Hz hoặc phân số của tông (tones) (nửa tông, một phần hai tông). Tần số cơ bản F0 là một đặc trưng quan trọng trong việc tạo ngôn điệu của tín hiệu tiếng nói, do đó việc tạo các đặc trưng cao độ là một vấn đề phức tạp và quan trọng trong tổng hợp tiếng nói.

1.1.3 Tổng hợp tín hiệu tiếng nói

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

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

Một hệ thống tổng hợp tiếng nói lý tưởng cần vừa tự nhiên, vừa dễ hiểu và mục tiêu xây dựng một hệ thống tổng hợp là làm gia tăng tối đa hai tính chất này. Hiện nay có ba phương pháp chính, phổ biến nhất là: tổng hợp mô hình hóa hệ thống phát âm, tổng hợp cộng hưởng tần số và tổng hợp ghép nối, ngoài ra cũng có các phương pháp khác phát triển từ ba phương pháp trên [11].

1.2 Các phương pháp tổng hợp tiếng nói

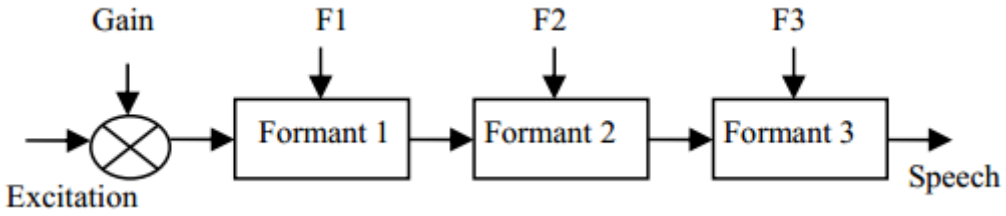
1.2.1 Tổng hợp mô phỏng hệ thống phát âm

Tổng hợp mô phỏng hệ thống phát âm là các kỹ thuật tổng hợp giọng nói dựa trên mô hình máy tính mô phỏng cơ quan phát âm của con người và quá trình tạo ra tiếng nói trên đó. Vì mục tiêu của phương pháp này là mô phỏng quá trình tạo tiếng nói sao cho càng giống cơ chế của con người càng tốt, nên về mặt lý thuyết đây được xem là phương pháp cơ bản nhất để tổng hợp tiếng nói, nhưng cũng vì vậy mà phương pháp này khó thực hiện nhất và khó có thể tổng hợp được tiếng nói chất lượng cao[12]. Tổng hợp mô phỏng phát âm đã từng chỉ là hệ thống dành cho nghiên cứu khoa học cho mãi đến những năm gần đây. Lý do là rất ít mô hình tạo ra âm thanh chất lượng đủ cao hoặc có thể chạy hiệu quả trên các ứng dụng thương mại. Một ngoại lệ là hệ thống NeXT, vốn được phát triển thương mại hóa bởi Trillium Sound Research Inc, Canada. Để thực hiện được phương pháp tổng hợp dựa trên việc mô phỏng hệ thống phát âm đòi hỏi thời gian, chi phí và công nghệ. Phương pháp này khó có thể ứng dụng tại Việt Nam thời điểm hiện nay.

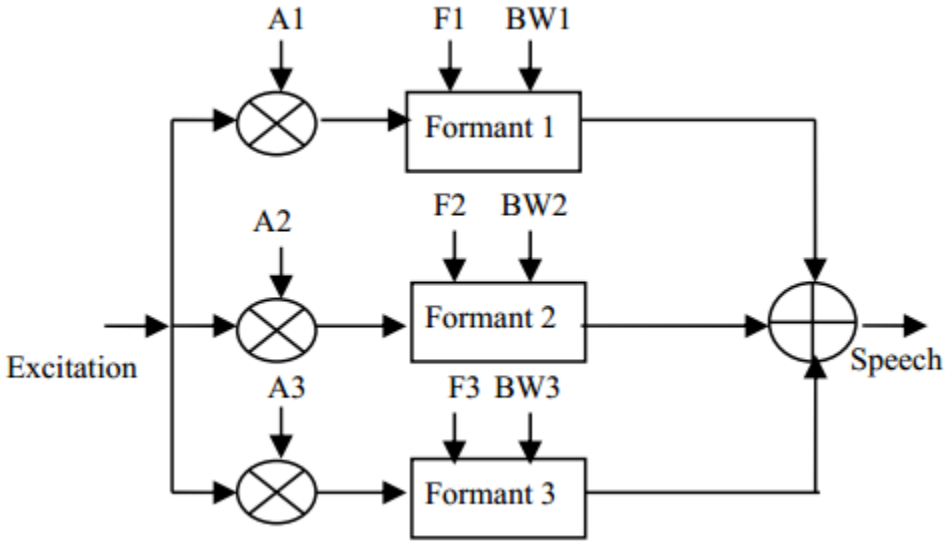
1.2.2 Tổng hợp tần số formant

Tổng hợp tiếng nói formant là phương pháp tổng hợp tiếng nói không sử dụng mẫu giọng thật nào khi chạy, thay vào đó tín hiệu tiếng nói được tạo ra bởi một mô hình tuyến âm. Mô hình này mô phỏng hiện tượng cộng hưởng của các cơ quan phát âm bằng một tập hợp các bộ lọc. Các bộ lọc này được gọi là các bộ lọc cộng hưởng formant, chúng có thể được kết hợp song song hoặc nối tiếp với nhau hoặc kết hợp cả hai.

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



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



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

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

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

1.2.3 Tổng hợp ghép nối

Tổng hợp ghép nối là phương pháp tổng hợp tiếng nói bằng cách ghép vào nhau các đoạn tín hiệu tiếng nói của một giọng nói đã được ghi âm. Các âm tiết sau khi được tạo thành sẽ được tiếp tục ghép lại với nhau tạo thành đoạn tiếng nói. Đơn vị âm phổ biến là âm vị, âm tiết, bán âm tiết,

âm đôi, âm ba, từ, cụm từ. Do đặc tính tự nhiên của tiếng nói được lưu giữ trong các đơn vị âm, nên tổng hợp ghép nối là phương pháp có khả năng tổng hợp tiếng nói với mức độ dễ hiểu và tự nhiên, chất lượng cao. Tuy nhiên, giọng nói tự nhiên được ghi âm có sự thay đổi từ lần phát âm này sang lần phát âm khác, và công nghệ tự động hóa việc ghép nối các đoạn của sóng âm thỉnh thoảng tạo ra những tiếng cọ xát không tự nhiên ở phần ghép nối. Có ba kiểu tổng hợp ghép nối:

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

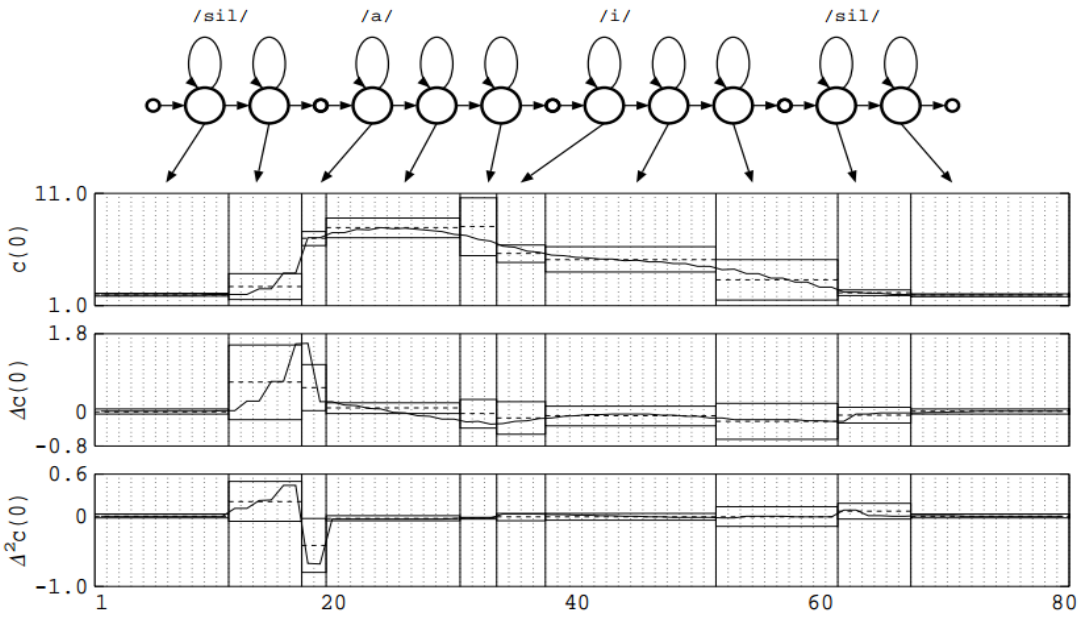
Tổng hợp chọn đơn vị dùng một cơ sở dữ liệu lớn các giọng nói ghi âm. Trong đó, mỗi câu được tách thành các đơn vị khác nhau như: các tiếng đơn lẻ, âm tiết, từ, nhóm từ hoặc câu văn. Một bảng tra các đơn vị được lập ra dựa trên các phần đã tách và các thông số âm học như tần số cơ bản, thời lượng, vị trí của âm tiết và các tiếng gần nó. Khi chạy các câu nói được tạo ra bằng cách xác định chuỗi đơn vị phù hợp nhất từ cơ sở dữ liệu. Quá trình này được gọi là chọn đơn vị và thường cần dùng đến cây quyết định được thực hiện. Thực tế, các hệ thống chọn đơn vị có thể tạo ra được giọng nói rất giống với người thật, tuy nhiên để đạt độ tự nhiên cao thường cần một cơ sở dữ liệu lớn chứa các đơn vị để lựa chọn.

Tổng hợp âm kép là dùng một cơ sở dữ liệu chứa tất cả các âm kép trong ngôn ngữ đang xét. Số lượng âm kép phụ thuộc vào đặc tính ghép âm học của ngôn ngữ. Trong tổng hợp âm kép chỉ có một mẫu của âm kép được chứa trong cơ sở dữ liệu, khi chạy thì lời văn được chồng lên các đơn vị này bằng kỹ thuật xử lý tín hiệu số nhờ mã tuyên đoán tuyến tính hay PSOLA [14]. Chất lượng âm thanh tổng hợp theo cách này thường không cao bằng phương pháp chọn đơn vị nhưng tự nhiên hơn cộng hưởng tần số và ưu điểm của nó là có kích thước dữ liệu nhỏ.

Tổng hợp chuyên biệt (Domain-specific) là phương pháp ghép nối từ các đoạn văn bản đã được ghi âm để tạo ra lời nói. Phương pháp này thường được dùng cho các ứng dụng có văn bản chuyên biệt, cho một chuyên nghành, sử dụng từ vựng hạn chế như các thông báo chuyến bay hay dự báo thời tiết. Công nghệ này rất đơn giản và đã được thương mại hóa từ lâu. Mức độ tự nhiên của hệ thống này có thể rất cao vì số lượng các câu nói không nhiều và khớp với lời văn, âm điệu của giọng nói ghi âm. Tuy nhiên hệ thống kiểu này bị hạn chế bởi cơ sở dữ liệu chuyên biệt không áp dụng được cho miền dữ liệu mở.

1.2.4 Tổng hợp dùng tham số thống kê

Tiếp theo đây chúng ta sẽ xem xét đến một phương pháp tổng hợp tiếng nói được nghiên cứu phổ biến và rộng rãi hiện nay đó là phương pháp tổng hợp dựa trên mô hình Markov ẩn (HMM) [15]. Ở đây HMM là một mô hình thống kê, được sử dụng để mô hình hóa các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể.



Hình 4: Mô hình markov ẩn áp dụng trong tổng hợp tiếng nói

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

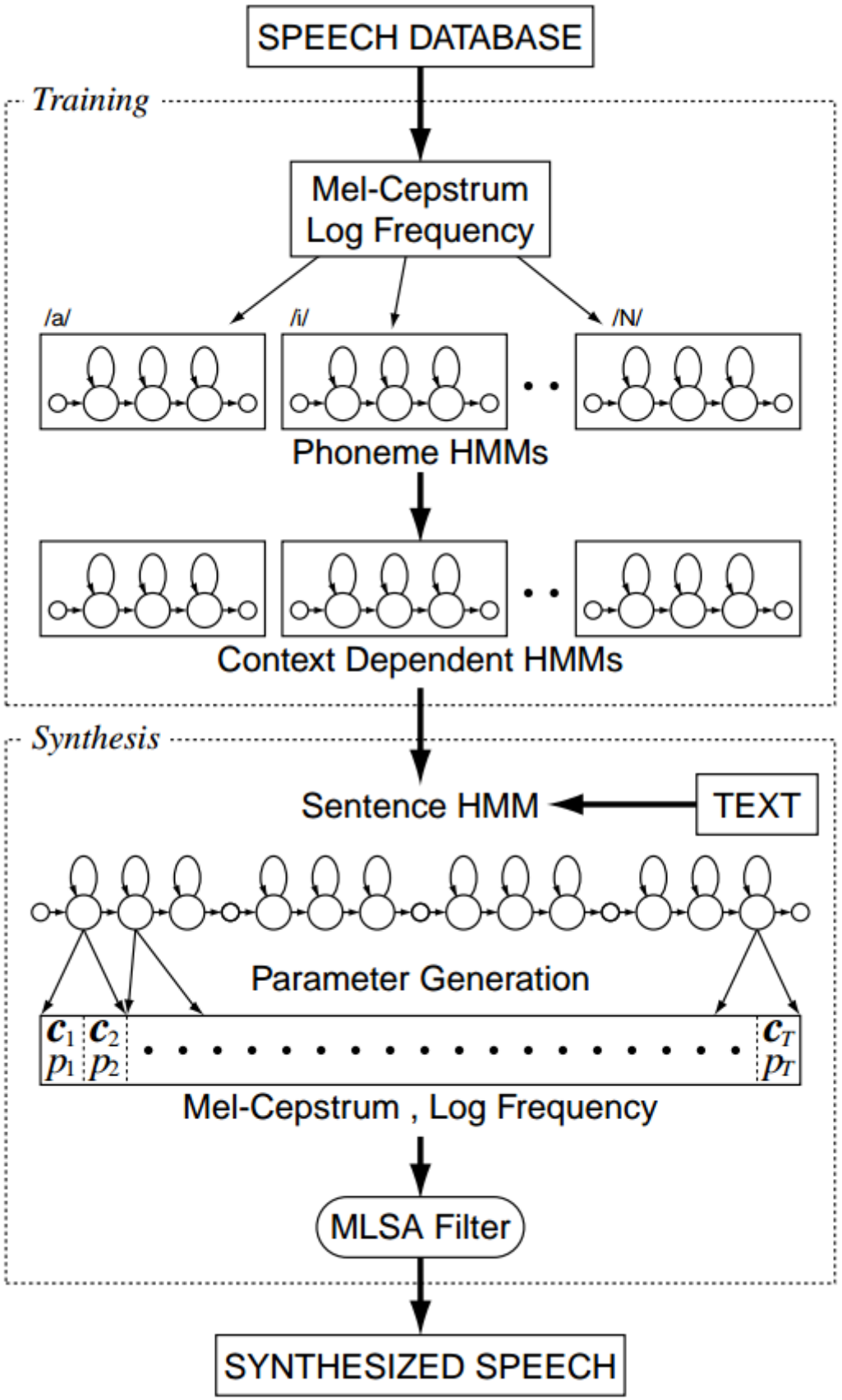
Quá trình tổng hợp dựa trên mô hình markov ẩn sẽ là quá trình mà nhận đầu vào là một đoạn văn bản, chuyển hóa đoạn văn bản này thành chuỗi âm vị, sau đó dựa vào các mô hình markov ẩn mô hình hóa chuỗi các âm vị tương ứng ta sẽ tìm ra được các tham số mel và tần số cơ bản f0. Từ các tham số mel xây dựng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

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

trưng. Mục tiêu của quá trình huấn luyện là từ dữ liệu đầu vào cải thiện các tham số của mô hình markov ẩn mà mô hình hóa cho mỗi âm vị.

Các đặc trưng ngôn ngữ của văn bản được mô tả bằng cách sử dụng một bộ phân cụm (thường là cây quyết định) để gom các cụm trạng thái của mô hình markov ẩn có đặc tính ngôn ngữ gần nhau nhất và bầu chọn ra một trạng thái tiêu biểu để thay thế cho các trạng thái còn lại trong cụm.

Hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn là một hệ thống có khả năng tạo tiếng nói mang phong cách nói khác nhau, với đặc trưng của nhiều người nói khác nhau, thậm chí là mang cảm xúc của người nói. Ưu điểm của phương pháp này là cần ít bộ nhớ lưu trữ và tài nguyên hệ thống hơn so với tổng hợp ghép nối, và có thể điều chỉnh tham số để thay đổi ngữ điệu. Tuy nhiên, một số nhược điểm của hệ thống này đó là độ tự nhiên trong tiếng nói tổng hợp của hệ thống bị suy giảm hơn so với tổng hợp ghép nối, phổ tín hiệu và tần số cơ bản được ước lượng từ các giá trị trung bình của các mô hình markov ẩn được huấn luyện từ dữ liệu khác nhau, điều này khiến cho tiếng nói tổng hợp nghe có vẻ đều đều mịn và đôi khi trở thành bị "nghẹt mũi".


SPEECH DATABASE
Training
Mel-Cepstrum
Log Frequency
/a/
/i/
/N/
Phoneme HMMs
Context Dependent HMMs
Synthesis
Sentence HMM
TEXT
Parameter Generation
c_1 c_2 c_T
p_1 p_2 p_T
Mel-Cepstrum , Log Frequency
MLSA Filter
SYNTHESIZED SPEECH

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

1.2.5 Tổng hợp tiếng nói bằng phương pháp lai ghép

Tổng hợp lai ghép là phương pháp tổng hợp bằng cách lai ghép giữa tổng hợp ghép nối chọn đơn vị và tổng hợp dựa trên mô hình markov ẩn, nhằm tận dụng ưu điểm của mỗi phương pháp và áp dụng nó trong hệ thống. Như đã nói, hệ thống tổng họp lai ghép kết hợp ưu nhược điểm của từng hệ thống thành phần, tùy theo thành phần nào đóng vai trò chủ đạo mà có thể phân loại các hệ thống tổng hợp lai ghép thành hai loại sau: Tổng hợp hướng ghép nối và tổng hợp hướng HMM.

Hệ thống tổng hợp hướng ghép nối sử dụng các HMM để hỗ trợ quá trình ghép nối, ý tưởng chính của phương pháp này như sau:

- Đơn vị dùng để lựa chọn trong "tổng hợp ghép nối chọn đơn vị" cũng sẽ là đơn vị được tổng hợp ra.
- Đường biên giữa các đơn vị sẽ được làm mịn bằng các mô hình markov ẩn.
- Âm thanh sau cùng được làm mịn bằng phương pháp làm mịn phổ.

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

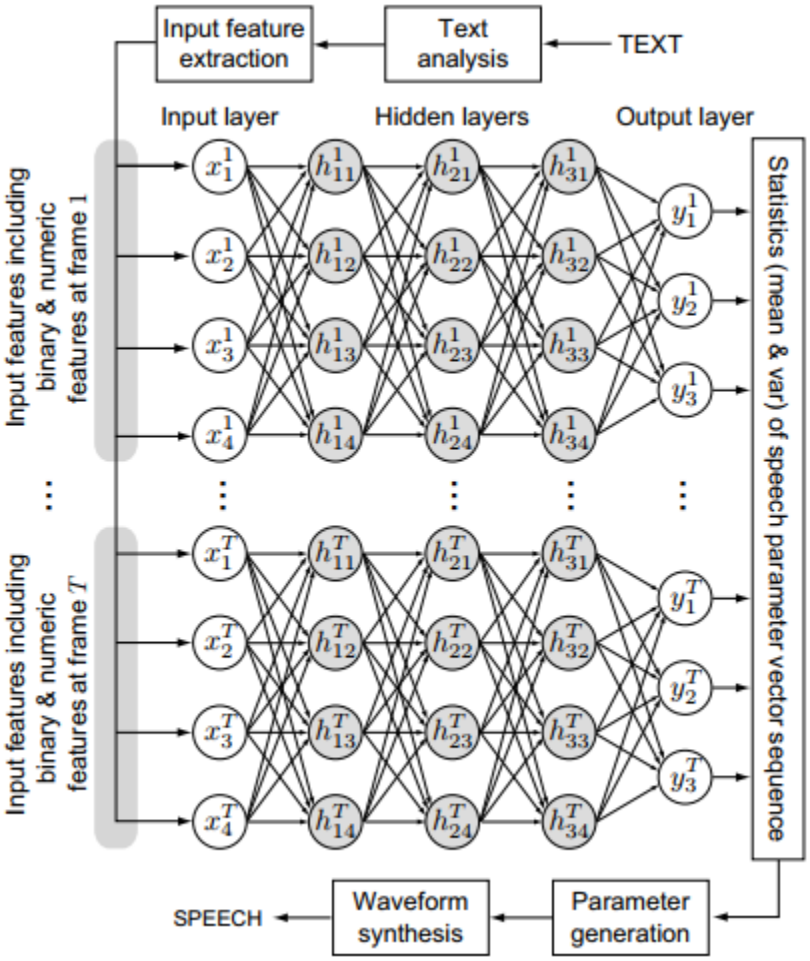
Hai hướng tổng hợp lai ghép nêu trên đều có ưu nhược điểm khác nhau, và được sử dụng tùy vào yêu cầu chất lượng tiếng nói hay yêu cầu cụ thể về hệ thống. Ưu điểm cơ bản của hệ thống lai ghép hướng ghép nối đó là giảm tác động không mong muốn do dữ liệu không đủ và giảm sự phụ thuộc vào dữ liệu, hay cũng chính là cải thiện các nhược điểm của tổng hợp ghép nối. Mặc dù đã giải quyết cơ bản những vấn đề về ghép nối nhưng vấn đề trở ngại tại những điểm ghép nối vẫn còn tồn tại.

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

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

Kiến trúc tổng quan của một hệ thống tổng hợp tiếng nói dựa trên mạng nơ ron học sâu DNN được mô tả trong hình 6. Trong đó, văn bản cần được tổng hợp sẽ đi qua bộ phân tích văn bản (Text analysis) để trích chọn các đặc trưng ngôn ngữ học và được chuyển hóa thành các véc tơ

nhị phân bởi bộ Input feature extraction, các véc tơ nhị phân đầu vào $\{x_n^t\}$ với x_n^t là đặc trưng thứ n tại khung t (frame t), các véc tơ này tương ứng tạo ra các đặc trưng đầu ra $\{y_m^t\}$ thông qua một mạng nơ ron DNN đã được huấn luyện, với mỗi y_m^t là đặc trưng đầu ra thứ m tại khung t. Các đặc trưng đầu ra này chứa các thông tin về phổ và tín hiệu kích thích, thông qua bộ tạo tham số (Parameter Generation) sẽ được chuyển thành các tham số đặc trưng âm học và được đưa vào bộ tạo tín hiệu tiếng nói (Waveform generation) để tạo ra tín hiệu tiếng nói thực.



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

Mạng nơ ron học sâu DNN dựa trên các lớp nơ ron nhân tạo, có khả năng mô hình hóa những mối quan hệ phi tuyến phức tạp giữa đầu vào và đầu ra. Đặc biệt trong trường hợp sử dụng DNN có thể mô hình hóa một cách mạnh mẽ mối quan hệ phi tuyến, phức tạp giữa các đặc trưng ngôn ngữ học của văn bản và đặc trưng âm học của tín hiệu tiếng nói, tuy nhiên việc sử dụng DNN cũng có những hạn chế đó là vì sự mạnh mẽ của nó nên nó rất nhạy cảm với các thông tin sai lệch và không tốt như nhiễu, và nó cũng cần rất nhiều dữ liệu để huấn luyện mô hình. Nhờ sự mạnh mẽ trong mô hình hóa mô hình âm học, DNN đã được áp dụng trong nhiều ứng dụng tổng hợp tiếng nói trên thế giới như các sản phẩm của Google, Baidu, Microsoft hay trong hệ thống Merlin của CSTR đã đạt được độ tự nhiên rất cao.

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

Bảng 1: Đánh giá so sánh HMM và DNN

Kết quả đánh giá so sánh hệ thống tổng hợp tiếng nói dựa trên HMM so với DNN của Google[19] được thể hiện trong bảng 1. Đánh giá này sử dụng phương pháp trung bình điểm ý kiến MOS trên thang điểm 5, với 173 câu kiểm tra chia theo 5 chủ đề, mỗi chủ đề khoảng 30 câu. Từ kết quả này cho thấy tổng hợp tiếng nói dựa trên DNN có chất lượng tốt hơn HMM.

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

Việt nam đang trong thời kỳ phát triển nhanh chóng của công nghệ thông tin. Điều đó cho phép chúng ta có những nền tảng khoa học kỹ thuật và nền tảng cơ sở vật chất để có thể nghiên cứu cũng như triển khai các ứng dụng về khoa học công nghệ trong cuộc sống. Trong nhiều năm trở lại đây, tổng hợp tiếng Việt đã có những thành tựu đáng kể, các hệ thống tổng hợp tiếng nói tiếng việt được ra đời như VietVoice[3], VnSpeech[4], Vais[5], Hệ thống tổng hợp tiếng nói của tập đoàn FPT hay hệ thống tổng hợp tiếng nói Hoa súng. Trong đó các hệ thống tổng hợp tiếng nói tiếng Việt được xây dựng dựa theo hai hướng phổ biến là tổng hợp ghép nối và tổng hợp sử dụng tham số thống kê.

Đối với phương pháp tổng hợp tiếng nói ghép nối: Dành cho tiếng Việt thì đã có rất nhiều hệ thống được phát triển, có thể kể đến như hệ thống Hoa Súng[20], được phát triển lần đầu vào năm 2007, dữ liệu để xây dựng hệ thống này được gọi là VNSpeechCorpus, nó được thu thập và lọc từ nhiều nguồn khác nhau như truyện, sách,… Dữ liệu này bao gồm nhiều loại khác nhau như: các từ với đầy đủ sáu thanh điệu, các số, câu thoại, đoạn văn ngắn,… Đến năm 2011 hệ thống được mở rộng[21], sử dụng kỹ thuật lựa chọn âm vị không đồng nhất. Phiên bản này cũng sử dụng cùng bộ dữ liệu ở phiên bản trước, nhưng được đánh chú thích ở mức độ âm tiết với những thông tin cần thiết như các thành phần âm vị, thanh điệu, thời gian, năng lượng, và những đặc trưng ngữ cảnh khác. Kết quả ban đầu cho thấy phiên bản thứ hai của hệ thống hoa súng có sự cải thiện về mặt chất lượng, tuy nhiên dữ liệu kiểm thử không được thiết kế để bao trùm toàn bộ đơn vị âm, thêm nữa không có sự kết nối giữa quá trình chọn đơn vị âm và quá trình chọn

[3] http://www.vietvoice.net/
[4] http://www.vnspeech.com
[5] https://vais.vn/

đơn vị như một bán âm tiết trong việc tính toán chi phí mục tiêu và chi phí ghép nối. Kết quả là tổng chi phí không được tối ưu hóa cho những câu cần bán âm tiết.

Đối với phương pháp tổng hợp tiếng nói sử dụng tham số thống kê, hay là tổng hợp tiếng nói dựa trên mô hình Markov ẩn (HMM). Ở Việt Nam cũng đã có nhiều hệ thống tổng hợp tiếng nói phát triển dựa trên phương pháp này, có thể kể đến như sản phẩm Vais, sản phẩm của tập đoàn FPT[6] hay hệ thống tổng hợp tiếng nói tiếng Việt Mica TTS[7] (Viện Mica Đại học Bách Khoa Hà Nội). Dữ liệu sử dụng cho hệ thống này bao gồm 3000 câu giàu ngữ âm và được gán nhãn bán tự động mức âm vị. Báo cáo kết quả của hệ thống này cho thấy độ hiểu đạt gần mức 100% và chất lượng tổng hợp đạt điểm 3.23 trên 5 thông qua một đánh giá sơ bộ.

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

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

[6] https://speech.openfpt.vn/
[7] http://sontinh.mica.edu.vn/tts2

TÀI LIỆU THAM KHẢO

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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