Một số vấn đề về tổng hợp tiếng nói tiếng Việt

Phan Thanh Sơn
Khoa CNTT, Đại học Thông tin liên lạc
Nha Trang, Việt Nam
Email: ptson@tcu.edu.vn

Phùng Trung Nghĩa
Đại học CNTT và TT, Đại học Thái Nguyên
Thái Nguyên, Việt Nam
Email: ptnghia@ictu.edu.vn

***Tóm tắt*—Ngôn ngữ là một công cụ giao tiếp mạnh mẽ, con người dễ dàng giao tiếp thông qua việc sử dụng các ngôn ngữ của nhau. Trong tình hình hội nhập và giao lưu quốc tế hiện nay, việc trao đổi thông tin giữa các quốc gia vẫn vấp phải rào cản về ngôn ngữ. Hiện nay, trên thế giới việc tổng hợp tiếng nói đã đạt được những tiến bộ đáng kể ở nhiều loại ngôn ngữ với chất lượng tốt và được ứng dụng rộng rãi. Đối với tiếng Việt, đã có nhiều công trình nghiên cứu khác nhau, nhưng chất lượng âm thanh và ngữ điệu của tiếng nói tổng hợp vẫn còn nhiều hạn chế, việc ứng dụng vào các lĩnh vực đời sống xã hội vẫn còn khiêm tốn.**

Từ khóa—tổng hợp tiếng nói tiếng Việt, formant, ghép nối, mô hình Markov ẩn, lai ghép

I. TỔNG HỢP TIẾNG NÓI

A. Định nghĩa

Tổng hợp tiếng nói (Speech Synthesis, viết tắt là SS) là quá trình tạo ra tiếng nói của con người một cách nhân tạo. Tổng hợp tiếng nói từ văn bản (Text-To-Speech, viết tắt là TTS) là quá trình chuyển đổi tự động một văn bản có nội dung bất kỳ thành lời nói. Hệ thống được sử dụng cho mục đích này gọi là hệ thống tổng hợp tiếng nói và có thể cài đặt bằng phần mềm hoặc trong sản phẩm phần cứng [6]. Một hệ thống TTS gồm hai thành phần cơ bản: phần xử lý ngôn ngữ tự nhiên (Natural Language Processing, viết tắt là NLP) và phần xử lý tổng hợp tiếng nói (Speech Synthesis Processing, viết tắt là SSP) [6]. Vì vậy, SS là thành phần cốt lõi của TTS (xem Hình 1).

B. Ứng dụng tổng hợp tiếng nói

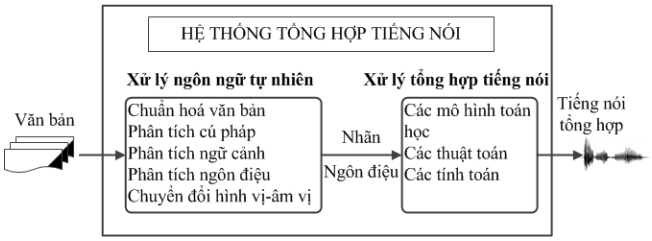
Tổng hợp tiếng nói được ứng dụng trong nhiều lĩnh vực khác nhau của đời sống con người, chẳng hạn như các ứng dụng cho người mù [9], [15], các ứng dụng cho người điếc và người gặp khó khăn về phát âm (câm, ngọng) [1], ứng dụng giáo dục, dạy ngoại ngữ [15], dịch tiếng nói [10], [17] và các trung tâm hỗ trợ khách hàng. Về nguyên tắc, tổng hợp tiếng nói có thể được sử dụng trong tất cả các hệ thống tương tác người-máy. Tùy thuộc vào từng ứng dụng cụ thể mà áp dụng các phương pháp và triển khai các hệ thống tổng hợp tiếng nói khác nhau.

Ngày nay, tổng hợp tiếng nói là một trong những lĩnh vực ngày càng được đầu tư nghiên cứu và ứng dụng phổ biến trong cuộc sống. Tổng hợp tiếng nói hiện đang được ứng dụng để đọc thông tin cập nhật hàng ngày cho người khiếm thị, kết hợp với ngôn ngữ diễn tả bằng ký hiệu cho người câm điếc, sử dụng trong giảng dạy chính tả và cách phát âm ngoại ngữ. Tổng hợp tiếng nói là thành phần lõi của hệ thống dịch tiếng nói, đó sẽ là công cụ giao tiếp phổ dụng để kết nối mọi người không nói cùng một ngôn ngữ trên thế giới. Đặc biệt, TTS hiện tại không chỉ đọc văn bản với chất lượng dễ hiểu, tính tự nhiên cao, mà còn có thể tổng hợp tiếng nói mang yếu tố tình cảm hay trạng thái cảm xúc, thậm chí có thể tổng hợp giọng hát. Một tính năng nữa của các hệ thống TTS hiện tại là có thể tổng hợp nhiều giọng nói mang đặc trưng âm học riêng biệt của người nói thay vì sử dụng một giọng nói chuẩn chung.

II. HỆ THỐNG TỔNG HỢP TIẾNG NÓI

A. Cấu trúc một hệ thống tổng hợp tiếng nói

Nếu đầu vào của một hệ thống tổng hợp tiếng nói là văn bản, thì hệ thống này được gọi là tổng hợp tiếng nói từ văn bản (TTS), minh họa trong Hình 1. Trong trường hợp các hệ thống tổng hợp tiếng nói với bộ từ vựng hạn chế, chẳng hạn như các máy trò chơi, các hệ thống trả lời tự động với các mẫu âm thanh thu âm trước, đôi khi có thể coi đó là một hệ thống TTS hạn chế cho một bài toán cụ thể, có giới hạn đầu vào.



Hình 1. Sơ đồ chức năng tổng quát của một hệ thống TTS

Sơ đồ chức năng tổng quát của một hệ thống TTS được minh hoạ trong Hình 1. Một hệ thống tổng hợp tiếng nói về cơ bản bao gồm hai khối chức năng: (1) khối phân tích xử lý ngôn ngữ tự nhiên (NLP) hay còn gọi là khối tổng hợp mức cao; và (2) khối xử lý tổng hợp tiếng nói (SSP) có nhiệm vụ tổng hợp tiếng nói hay còn gọi là khối tổng hợp mức thấp.

Tổng hợp mức cao có nhiệm vụ chuyển đổi chuỗi các ký tự văn bản đầu vào thành một dạng chuỗi các nhãn 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 cũng 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 ra 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, âm tiết). Sau đó, tiếng nói ở dạng sóng tín hiệu sẽ được tạo ra bằng một kỹ thuật tổng hợp.

B. Khối xử lý ngôn ngữ tự nhiên

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. Việc xử lý ngôn ngữ tự nhiên bao gồm: chuẩn hóa văn bản, phân tích cú pháp, phân tích ngữ cảnh và ngữ nghĩa, chuyển đổi hình vị sang âm vị, dự đoán và phát sinh thông tin ngữ âm và ngữ điệu.

Khối xử lý ngôn ngữ tự nhiên được chia thành ba phần chính:

- Thành phần phân tích văn bản.
- Thành phần chuyển đổi hình vị sang âm vị.
- Thành phần dự đoán và sinh ngôn điệu cho văn bản.

C. Khối xử lý tổng hợp tín hiệu tiếng nói

Khối xử lý tổng hợp tín hiệu tiếng nói đảm nhiệm việc thực hiện 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 đề cập đến nội dung của tiếng nói tổng hợp có thể hiểu được dễ dàng 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 đa hai tính chất này. Có nhiều phương pháp tổng hợp tiếng nói khác nhau được áp dụng, một số thiên về mức độ dễ hiểu hơn hoặc mức độ tự nhiên hơn, tùy thuộc vào mục đích mà các phương pháp tổng hợp được lựa chọn. Nhưng mục đích cơ bản của bất kỳ phương pháp tổng hợp là tạo ra tiếng nói với chất lượng dễ hiểu nội dung. Hiện nay, có ba phương pháp chính thường được dùng là tổng hợp mô hình hoá 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 [19].

III. CÁC PHƯƠNG PHÁP 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 ứng dụng tuỳ theo các tiêu chí cụ thể. Việc phân loại các phương pháp tổng hợp cơ bản tuỳ thuộc vào tiếng nói tổng hợp được tạo ra từ các tham số nhân tạo (các tần số formant), hay từ các mẫu tiếng nói thu âm trước (kho ngữ liệu) [27].

A. Tổng hợp mô phỏng hệ thống phát âm

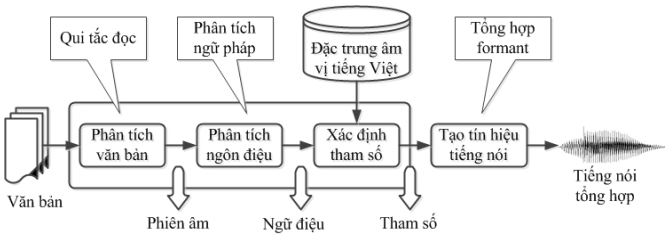
Tổng hợp mô phỏng hệ thống phát âm là phương pháp mà con người cố gắng mô phỏng quá trình tạo ra tiếng nói sao cho càng giống cơ chế phát âm của của con người càng tốt. Vì vậy, về mặt lý thuyết, đây được xem là phương pháp cơ bản nhất để tổng hợp tiếng nói, nhưng cũng vì thế mà phương pháp này khó thực hiện và tính toán nhất, và khó có thể tổng hợp được tiếng nói chất lượng cao [4][18]. Do những hạn chế trong vấn đề mô phỏng các tham số tiếng nói và năng lực tính toán, mà tổng hợp mô phỏng hệ thống phát âm đã không đạt được nhiều thành công mong đợi như phương pháp tổng hợp tiếng nói khác. Tuy nhiên, nó có rất nhiều ứng dụng hữu ích trong nghiên cứu cơ bản về quá trình tạo tiếng nói, và hiện nay phương pháp này đang được đầu tư nghiên cứu và phát triển trở lại. Sự phát triển của khoa học tính toán, giảm giá thành thiết bị và khả năng, tài nguyên dành cho tính toán ngày càng tăng khiến cho việc mô phỏng cơ chế phát âm hiệu quả hơn [27].

B. Tổng hợp tần số formant

Tổng hợp tần số formant, hay còn 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 [2], [11], [12]. 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 [9].

Hiện nay, với những công cụ thích hợp chúng ta hoàn toàn có thể xác định tần số formant cho các âm vị của tiếng Việt [3], [11], [12]. Đi theo hướng này có ưu điểm là tiết kiệm được bộ nhớ, có khả năng điều khiển mềm dẻo các tham số âm học của tiếng nói. Nhược điểm của phương pháp này là khó xây dựng, cần nghiên cứu sâu sắc về ngữ âm của ngôn ngữ, phức tạp trong việc xác định các tham số điều khiển bộ tổng hợp, hạn chế về tính tự nhiên, độ giống tiếng người của tiếng nói tạo ra, chất lượng tiếng nói không tự nhiên (nói nghe như tiếng robot, khác hoàn toàn giọng nói con người) và phụ thuộc nhiều vào chất lượng của quá trình phân tích tiếng nói của từng ngôn ngữ. Ngoài ra, tổng hợp formant yêu cầu chuẩn bị trước các tham số chính xác trước khi tiến hành tổng hợp tiếng nói, khiến cho quá trình tổng hợp thiếu linh hoạt.

Tại Việt Nam, phương pháp tổng hợp formant cũng đã có vài công trình nghiên cứu và đã có các kết quả đưa vào ứng dụng thực tế. Chẳng hạn, phần mềm "đọc văn bản tiếng Việt", năm 2004 [11]; Phần mềm tổng hợp tiếng nói tiếng Việt VnSpeech (xem Hình 2), năm 2009 [12], tổng hợp tiếng nói theo hướng tiếp cận này. Hệ thống tổng hợp formant có thể đọc được hầu hết các âm tiết tiếng Việt ở mức nghe rõ, tuy vậy, nó có nhược điểm là mức độ tự nhiên không cao.



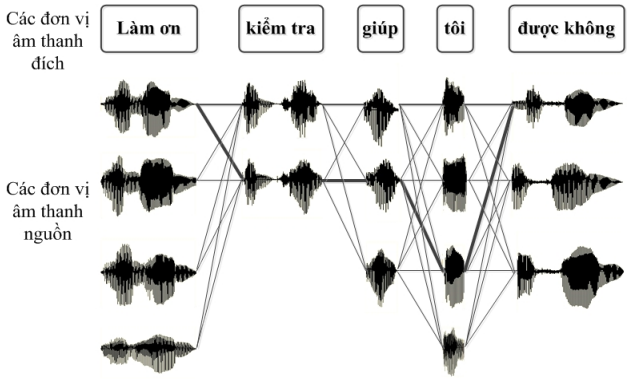
Hình 2. Mô hình VnSpeech tổng hợp tiếng Việt dựa vào formant

C. Tổng hợp dựa trên ghép nối

Tổng hợp ghép nối (hay còn gọi là lựa chọn đơn vị âm) là một trong số các phương pháp tổng hợp mới phát triển sau này, kết hợp (ghép nối) các mẫu tiếng nói tự nhiên thu âm sẵn lại với nhau để tạo ra câu nói tổng hợp **[7]**. Đơn vị âm (unit) phổ biến là âm vị, âm tiết, bán âm tiết, âm đôi, âm ba, từ, cụm từ. Do các 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, 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.

Ngoài ra, tập các đơn vị âm luôn bị hạn chế về số lượng cũng như nội dung. Điều này dẫn đến tiếng nói tổng hợp nghe "thô ráp", các đơn vị âm ghép nối với nhau thường không phù hợp ngữ cảnh. 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 [16]. Hạn chế này khiến tính linh hoạt của tổng hợp ghép nối bị ảnh hưởng và phương pháp này chỉ có thể "bắt chước" một giọng người nói cụ thể trong tập dữ liệu đơn vị âm rất lớn của người đó.

Do hạn chế về chất lượng của tiếng nói tổng hợp dựa vào formant, nên phương pháp tổng hợp ghép nối được tập trung đầu tư, nghiên cứu. Trước đây, đã có phần mềm V-Talk của Viện Khoa học kỹ thuật Bưu điện [24], phát triển dựa trên tổng hợp ghép nối diphone (phụ âm đầu và phần vần). Hiện nay, có các phần mềm VnVoice (Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam) theo hướng ghép nối bán âm tiết; Phần mềm nhu liệu đọc tiếng Việt VietVoice và một số sản phẩm tổng hợp tiếng Việt bằng cách ghép âm tiết như phần mềm đọc tiếng Việt Sao Mai; Phần mềm VietSound do Đại học Bách Khoa TP Hồ Chí Minh phát triển, phần mềm này kết hợp sử dụng phương pháp ghép nối diphone và phương pháp tổng hợp formant. Trung tâm MICA (Đại học Bách khoa Hà Nội) hiện nay cũng đang có các nghiên cứu về tổng hợp tiếng nói dựa trên ghép nối các đơn vị âm không đồng nhất [5]. Hệ thống tổng hợp tiếng nói "Tiếng nói phương Nam" (VoS) của Phòng thí nghiệm Trí tuệ nhân tạo AILab (Đại học Khoa học tự nhiên TP HCM) được phát triển theo hướng kết hợp ghép nối âm tiết và cụm từ [28] (xem Hình 3).



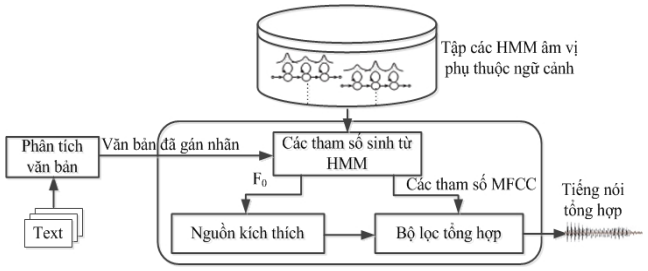
Hình 3. Mô hình VoS tổng hợp ghép nối âm tiết và cụm từ

D. Tổng hợp dùng tham số thống kê

Một phương pháp khác được nghiên cứu rộng rãi hiện nay trong tổng hợp tiếng nói là sử dụng các HMM [8], [23], [25], [26]. Ở đây, HMM là mô hình thống kê, sử dụng để mô hình hoá các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể, được trích rút đồng thời từ cơ sở dữ liệu tiếng nói. Nhờ tập các HMM này, hệ thống sau đó có thể phát sinh ra các tham số tiếng nói, tuỳ thuộc vào nội dung văn bản đầu vào, để tạo ra tiếng nói dưới dạng sóng nhờ các tham số được phát xạ này.

Hệ thống tổng hợp tiếng nói dựa trên HMM, cũng có thể xem là một phát triển của kỹ thuật tổng hợp ghép nối mà đơn vị âm là âm vị, là một hệ thống có khả năng tạo ra tiếng nói mang các phong cách nói khác nhau, với đặc trưng của nhiều người nói khác nhau, thậm chí mang cả 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 dựa trên ghép nối và có thể điều chỉnh tham số để thay đổi ngữ điệu, thay đổi các đặc trưng người nói. Tuy nhiên, mức độ tự nhiên trong tiếng nói tổng hợp của các hệ thống TTS dựa trên HMM thường bị suy giảm so với tổng hợp tiếng nói dựa trên ghép nối.

Mặc dù có nhiều ưu điểm, nhưng hệ thống tổng hợp tiếng nói dựa trên HMM vẫn còn những tồn tại. Trong hệ thống này, phổ tín hiệu và tần số cơ bản được ước lượng từ các giá trị xấp xỉ trung bình của phổ và tần số cơ bản, phát xạ từ các HMM được huấn luyện từ nhiều dữ liệu khác nhau. Các đặc trưng ngôn điệu của tiếng nói thu âm gốc có thể bị thay thế bởi các đặc trưng "trung bình" này, khiến cho tiếng nói tổng hợp nghe có vẻ "đều đều", quá "mịn" hay quá "ổn định". Đặc điểm quá "mịn" của tiếng nói tổng hợp dựa trên HMM vẫn có thể chấp nhận được khi chỉ chú ý đến tính chất nghe hiểu. Nhưng chính những hạn chế này khiến cho tiếng nói tổng hợp dựa trên HMM nghe như bị "nghẹt mũi" và làm giảm ngôn điệu, sắc thái cảm xúc hay phong cách nói trong câu nói.



Hình 4. Mô hình hệ thống TTS dựa trên mô hình Markov ẩn

Ở Việt Nam hiện nay, tổng hợp tiếng nói dựa trên HMM (xem Hình 4) là hướng nghiên cứu mới đang được triển khai ứng dụng cho hệ thống tổng hợp tiếng Việt. Trước đây, năm 2009 đã có đề tài nghiên cứu "Phát triển Engine tổng hợp tiếng Việt (VieTalk) cho người khiếm thị", sử dụng phương pháp này [25]. Hiện nay, phương pháp này cũng được nghiên cứu, cải tiến, nâng cao chất lượng tiếng nói tổng hợp, và là một phần của đề tài cấp nhà nước "Nghiên cứu phát triển hệ thống dịch tiếng nói hai chiều Việt - Anh, Anh - Việt có định hướng lĩnh vực" của Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Từ nửa cuối năm 2013, Viện nghiên cứu Quốc tế MICA và Phòng thí nghiệm Trí tuệ nhân tạo AILab cũng đang bắt đầu có những nghiên cứu, phát triển hệ thống tổng hợp tiếng Việt tham số thống kê dựa trên HMM.

E. Hướng tiếp cận tổng hợp bằng phương pháp lai ghép

Gần đây, hướng tiếp cận tổng hợp bằng phương pháp lai ghép giữa tổng hợp lựa chọn đơn vị dựa trên ghép nối và tổng hợp dựa trên HMM đang được nghiên cứu áp dụng, nhằm tận dụng ưu thế của từng phương pháp trong hệ thống mới.

Một cách tiếp cận là sử dụng các mô hình HMM để làm mịn các điểm ghép nối của phương pháp tổng hợp lựa chọn đơn vị [14]. Mặc dù cách tiếp cận này có thể cải thiện sự gián đoạn tại vị trí ghép nối, nhưng nó lại tạo ra thành phần không mong muốn khi có sự nhầm lẫn giữa các hệ số làm mịn và tín hiệu nguồn kích thích. Một hình thức lai ghép khác là sử dụng các tham số phổ, tần số cơ bản và thời gian trạng thái sinh ra từ các HMM để tính toán chi phí mục tiêu và chi phí ghép nối cho quá trình ghép nối lựa chọn đơn vị [13], [16]. Phương pháp lai ghép này có thể cải thiện chất lượng và tính ổn định của tiếng nói tổng hợp và vẫn bảo toàn tính ưu việt của hệ thống TTS dựa trên HMM là thích nghi, thay đổi đặc trưng người nói trong điều kiện dữ liệu huấn luyện hạn chế.

IV. SO SÁNH CÁC KẾT QUẢ TỔNG HỢP TIẾNG NÓI

Kết quả tổng hợp tiếng Việt được thực hiện bằng các phương pháp khác nhau: tổng hợp formant [11], [12], tổng hợp ghép nối đơn vị âm thanh [24], [28], tổng hợp dựa trên HMM [25], [20], [21] và tổng hợp theo phương pháp lai ghép [13], [22]. So sánh, đánh giá kết quả tổng hợp từ các phương pháp khác nhau được thực hiện dựa trên tiêu chí chất lượng nghe rõ nội dung và tính tự nhiên của tiếng nói tổng hợp.

Để đánh giá chất lượng nghe rõ và tính tự nhiên của tiếng nói tổng hợp, ngoài các đánh giá khách quan dựa trên so sánh sự biến dạng của cepstral tần số thang Mel (Mel-Frequency Cepstral Coefficients Distortion, MFCD), sai lệch căn bậc hai trung bình bình phương (Root-Mean-Square Error, RMSE) của logF0 và so sánh trực quan trên ảnh phổ, trên đường bao cao độ của tiếng nói tổng hợp và thu âm gốc, thì cũng cần có các kiểm tra chủ quan dựa trên tiêu chí điểm đánh giá ý kiến trung bình (Mean Opinion Score, MOS) của người nghe và các đánh giá khác. So sánh đánh giá được thực hiện trên 10 câu tổng hợp chọn ngẫu nhiên cho mỗi một phương pháp trong tập dữ liệu đánh giá (xem Bảng 1).

Đánh giá MOS được thực hiện thông qua nghe và cho điểm theo thang điểm 5 (1: tồi, 2:hơi tồi, 3:tạm được, 4:khá tốt, 5:tốt), tuỳ theo mức độ cảm nhận của người nghe, dựa trên hai tiêu chí: mức độ nghe rõ nội dung và mức độ giống tiếng nói tự nhiên. Số lượng người tham gia nghe và đánh giá là 50 người.

BẢNG 1. DẠNG BẢNG

Tiêu chí đánh giá	**VnSpeech**	**VietVoice**	**VoS**	**HMM**	**Hybrid**
Mức độ nghe hiểu	2.65	*3.86*	**4.08**	***4.02***	*4.10*
Mức độ tự nhiên	2.26	*2.95*	**3.78**	***3.93***	*3.75*

V. KẾT LUẬN

Bài báo đã trình bày khái quát về tình hình nghiên cứu tổng hợp tiếng nói tại Việt Nam từ trước đến nay. Kết quả thực nghiệm cho thấy chất lượng tiếng nói tổng hợp theo phương pháp ghép nối đơn vị âm và tổng hợp sử dụng tham số thống kê dựa trên HMM cho kết quả có chất lượng tốt nhất. Ngoài ra trong trong các nghiên cứu [20] và [21], ngôn điệu của tiếng nói tổng hợp được cải tiến rõ rệt so với [25]. Cách tiếp cận mà chúng tôi đề xuất trong [13] và [22] thực hiện tốt hơn hẳn so với các phương pháp trước, ngoại trừ phương pháp ghép nối.

Trong thời gian tới, bên cạnh những kết quả đã đạt được, chúng tôi sẽ tiếp tục có những nghiên cứu về ngôn điệu, chất giọng, phong cách nói để tăng tính tự nhiên cũng như tính mềm dẻo cho hệ thống tổng hợp tiếng Việt.

TÀI LIỆU THAM KHẢO

[1] Abadjieva, E.; Murray, I. và Arnott, J., "Applying Analysis of Human Emotion Speech to Enhance Synthetic Speech," *Proc. in Eurospeech*, Berlin, Germany, tr. 909-912, 1993.

[2] Bạch Hưng Khang và các cộng sự, *Nghiên cứu phát triển công nghệ nhận dạng, tổng hợp và xử lý ngôn ngữ tiếng Việt*, Đề tài cấp nhà nước KC.01.03, 2004.

[3] Bùi Tiến Lên, *Xây dựng hệ tổng hợp tiếng Việt dựa trên luật*, Luận văn thạc sĩ ngành công nghệ thông tin, Đại học KHTN, Đại học Quốc gia Tp Hồ Chí Minh, 2001.

[4] Dang, J. và Honda, K., "Construction and control of a physiological articulatory model," *Journal of Acoustical Society of America*, Vol.115(2), tr. 853-870, 2004.

[5] DO Van Thao, TRAN Do Dat, NGUYEN Thi Thu Trang, "Non-uniform unit selection in Vietnamese Speech Synthesis," *Proceedings of the 2nd SoICT 2011*, tr. 165-171, 2011.

[6] Dutoit, Thierry, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Netherlands, 1997.

[7] Hunt, A.; Black,A. và Alan,W., "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. in ICASSP*, Vol.1, tr. 373-376, 1996.

[8] Kim, Sang-Jin, *HMM-Based Korean Speech Synthesizer with Two-Band Mixed Excitation Model for Embedded Applications*, Doctoral Dissertation, Information and Communications University, Korea, 2007.

[9] Klatt, D., "Review of Text-to-Speech Conversion for English," *Journal of the Ac oustic al Society of America*, Vol. 82 (3), tr. 737-793, 1987.

[10] Liang, Hui và Dines John, "Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation," *Proc. in InterSpeech*, Florence, Italy, tr. 1825-1828, 2011.

[11] Lê Hồng Minh, "Một số kết quả nghiên cứu và phát triển hệ phần mềm chuyển văn bản thành tiếng nói cho tiếng Việt bằng tổng hợp formant," *Kỷ yếu Hội thảo Khoa học Quốc gia lần thứ nhất - Nghiên cứu Phát triển và Ứng dụng Công nghệ Thông tin và Truyền thông (ICT.rda'03),* Hà Nội, tr. 292-301, 2003.

[12] Nguyễn Hữu Minh, *Xác định khoảng ngừng giữa các âm tiết, cường độ và trường độ của âm tiết cho bộ phát âm tiếng Việt*, Luận văn thạc sĩ ngành tin học, Đại học KHTN, Đại học Quốc gia Tp Hồ Chí Minh, 2009.

[13] Phung, Trung-Nghia; Luong, Chi-Mai và Masato, Akagi, "A Hybrid TTS between Unit Selection and HMM-based TTS under limited data conditions," *Proc. in 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013.

[14] Plumpe, M. và các cộng sự, "HMM-based smoothing for concatenative speech synthesis," *Proc. in ICSLP*, tr. 2751-2754, 1998.

[15] Portele, T. và Kramer, J., "Adapting a TTS System to a Reading Machine for the Blind," *Proc. in ICSLP 96*, Philadenphia, USA, tr. 184-187, 1996.

[16] Qian, Yao và các cộng sự, "A fast table lookup based, statistical model driven non-uniform unit selection TTS," *Proc. In ICASSP2013*, Vancouver, Canada, 2013.

[17] Sakti, Sakriani và các cộng sự, "The Asian Network-based Speech-to-Speech Translation System," *Proc. in Automatic Speech Recognition & Understanding (ASRU)*, Merano, Italy, tr. 507-512, 2009.

[18] Sondhi, M. M. và Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol.35(7), tr. 955-967, 1987.

[19] Taylor, Paul, *Text-to-Speech Synthesis*, University of Cambridge, Cambridge, UK, 2009.

[20] Thanh-Son PHAN, Anh-Tuan DINH, Tat-Thang VU and Chi-Mai LUONG, "An improvement of prosodic characteristics in Vietnamese Text to Speech System," *Proc. in The Fifth International Conference on Knowledge and Systems Engineering (KSE),* Hanoi, Vietnam, 2013.

[21] Thanh-Son PHAN, Tu-Cuong DUONG, Anh-Tuan DINH, Tat-Thang VU, Chi-Mai LUONG, "Improvement of Naturalness for an HMM-based Vietnamese Speech Synthesis using the Prosodic information," *The 10th IEEE RIVF International Conference on Computing and Communication Technologies*, Hanoi, Vietnam, 2013.

[22] Thanh-Son PHAN, Dang-Hung PHAN, Tu-Cuong DUONG, "A Study on Hybrid Speech Synthesis System between Concatenation TTS and Statistical TTS based on HMM," *Hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông,"* Đại học Duy Tân, Đà Nẵng, Việt Nam, 2013.

[23] Tokuda, K.; Zen H. và Black, Alan W., "An HMM-based speech synthesis system applied to English," *Proc. in IEEE Speech Synthesis Workshop*, Santa Monica, USA, 2002.

[24] Trịnh Anh Tuấn, "Một số phương pháp nâng cao chất lượng hệ thống tổng hợp tiếng Việt V-TALK," *Tạp chí Bưu chính Viễn thông*, Số 3, Hà Nội, tr. 19-23, 2000.

[25] Vu, Thang Tat; Luong, Mai Chi và Satoshi, Nakamura, "An HMM-based Vietnamese Speech Synthesis System," *Proc. in Oriental COCOSDA*, Urumqi, China, tr. 116-121, 2009.

[26] Yamagishi, J., *An Introduction to HMM-Based Speech Synthesis*, Technical Report, Tokyo Institute of Technology, Japan, 2006.

[27] Youcef, T. và Mohamed, B., "Speech synthesis techniques. A survey," *7th International Workshop on Systems*, *Signal Processing and their Applications (WOSSPA)*, Tipaza, Algeria, tr. 67-70, 2011.

[28] Vũ Hải Quân và Cao Xuân Nam, "Tổng hợp tiếng nói tiếng Việt theo phương pháp ghép nối cụm từ," *Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT*, Tạp chí CNTT và TT, Tập V-1(1), tr. 70-76, 2009.