

Mô hình Ngôn điệu trong Tổng hợp Tiếng nói

Trịnh Quốc Sơn

Trường Đại học Công nghệ Thông tin
sontq@uit.edu.vn

Tóm tắt. Báo cáo trình bày tổng thể về vai trò của mô hình ngôn điệu trong hệ thống tổng hợp tiếng nói. Điểm mạnh, điểm yếu của từng phương pháp mô hình hóa cũng được trình bày, qua đó thấy được các góc nhìn khác nhau về các phương pháp khi nghiên cứu mô hình hóa cho một ứng dụng, hệ thống tổng hợp tiếng nói cụ thể. Các thách thức trong quá trình xây dựng mô hình ngôn điệu để tạo cho tiếng nói nhân tạo tự nhiên như cách con người giao tiếp luôn là một vấn đề không nhỏ, tuy nhiên, những thách thức này có thể được giải quyết trên cơ sở sự hỗ trợ mạnh mẽ của các chuyên gia ngôn ngữ và nỗ lực của chuyên gia kỹ thuật, trong kết hợp để tạo ra các mô hình phù hợp.

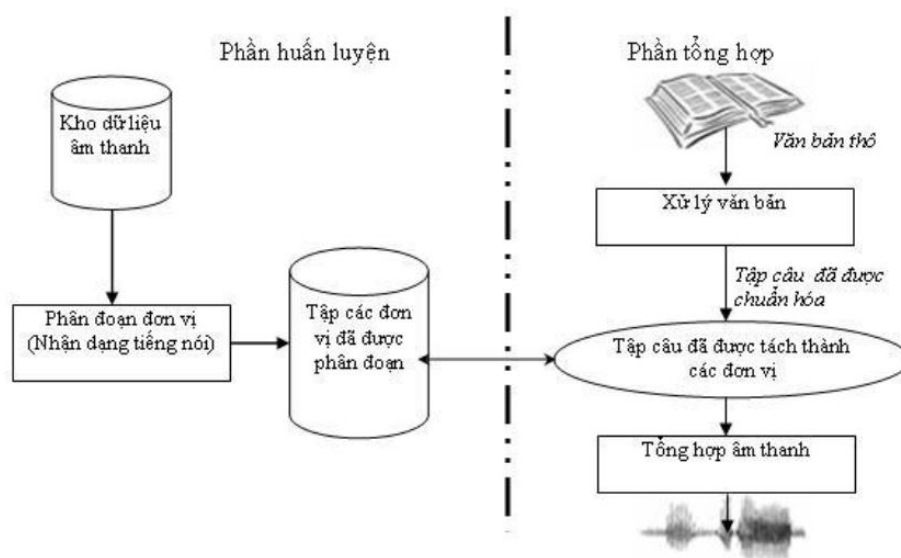
Từ khóa: Tổng hợp tiếng nói, mô hình ngôn điệu, phương pháp tiếp cận.

1 Giới thiệu

Trong những năm gần đây, các nghiên cứu về tổng hợp tiếng nói đã được thực hiện để hướng đến một hệ thống tổng hợp tiếng nói có khả năng đọc văn bản một cách tự nhiên và dễ hiểu. Có thể nói, quá trình chuyển đổi từ văn bản sang tiếng nói có thể có thể đạt được nhanh chóng. Tuy nhiên, tiếng nói được tạo ra đó được bao hàm các thông tin về ngôn điệu văn bản là một yêu cầu quan trọng trong việc hướng đến một hệ thống tổng hợp tiếng nói đạt chất lượng tự nhiên như tiếng nói của con người, qua đó giúp tạo ra những tương tác giữa người và máy.

Có thể nói trong bất kỳ hệ thống tổng hợp tiếng nói nào cũng bao gồm hai giai đoạn thực hiện để tạo ra giọng nói tổng hợp. Một là phân tích văn bản và hai là tạo ra giọng nói tổng hợp. Nhiệm vụ của giai đoạn phân tích văn bản là tiền xử lý văn bản, quá trình này trải qua nhiều bước thực hiện, bao gồm: chuẩn hóa văn bản, sửa lỗi văn bản, tách câu, phân đoạn,... và nhiệm vụ của giai đoạn hai là tạo ra tiếng nói tổng hợp [27,28].

Hai tính chất quan trọng của chất lượng hệ thống tổng hợp tiếng nói là mức độ tự nhiên và mức độ dễ nghe. Mức độ tự nhiên của giọng nói tổng hợp chỉ đến sự giống nhau giữa giọng tổng hợp và giọng nói tự nhiên của người thật. Mức độ dễ nghe chỉ đến việc câu phát âm có thể hiểu được dễ dàng không. Một máy tổng hợp giọng nói lý tưởng cần vừa tự nhiên vừa dễ nghe, do đó mục tiêu xây dựng máy tổng hợp giọng nói là làm gia tăng đến mức tối đa hai tính chất này. Một số hệ thống thiên về mức độ dễ nghe hơn, hoặc mức độ tự nhiên hơn; tùy thuộc vào mục đích mà công nghệ được lựa chọn.

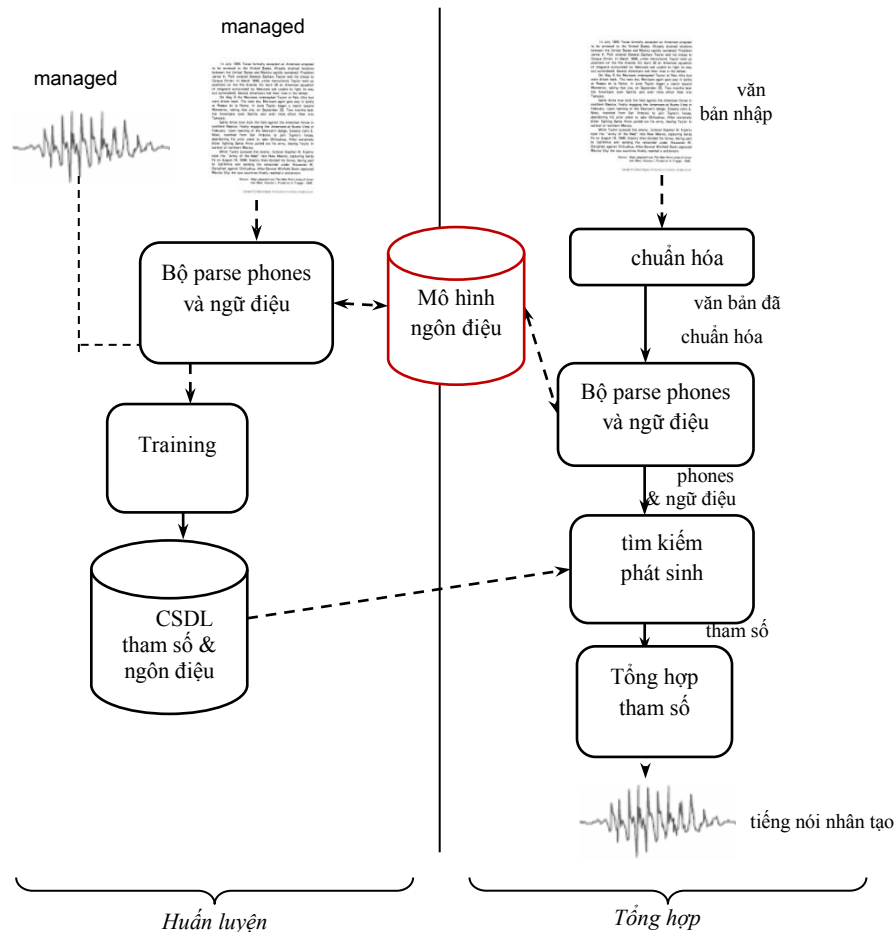


Hình 1. Các thành phần của hệ thống tổng hợp tiếng nói

Để đạt được chất lượng tiếng nói tổng hợp có kết quả tiệm cận theo hướng lý tưởng, việc tạo ra tiếng nói tổng hợp có điệu tính phát âm tự nhiên, rõ ràng và phù hợp là một yêu cầu quan trọng. Vì các thông tin điệu tính khi phát âm tự nhiên đóng vai trò quan trọng trong một hệ thống tổng hợp tiếng nói chất lượng cao được mong đợi. Do đó, trong các nghiên cứu về tổng hợp tiếng nói hiện nay, nghiên cứu tạo các mô hình ngôn điệu trong hệ thống tổng hợp tiếng nói là yêu cầu quan trọng, cấp thiết.

2 Mô hình hóa Ngôn điệu

Ngôn điệu là khía cạnh quan trọng trong tiếng nói giúp duy trì biểu cảm và dễ hiểu trong các hệ thống tổng hợp giọng nói. Mô hình hóa ngôn điệu là quá trình xây dựng mô hình tính toán để tạo ra các biến thể điệu tính tự động trong tiếng nói tổng hợp.



Hình 2. Hệ thống tổng hợp tiếng nói và vai trò của mô hình ngôn điệu

Tác vụ của mô hình ngôn điệu được thực hiện cả trong khâu huấn luyện và khâu tổng hợp trong một hệ thống tổng hợp tiếng nói. Cụ thể là:

- Ở bước huấn luyện, mô hình ngôn điệu được dùng làm tham chiếu cho việc định nghĩa ngôn điệu tiềm ẩn trong tập ngữ liệu âm thanh dùng cho huấn luyện. Kết quả của bước huấn luyện sẽ là cơ sở dữ liệu ngữ âm được đồng bộ theo ngôn điệu đã qui định.
- Ở bước tổng hợp, văn bản đầu vào sau khi chuẩn hóa sẽ được chuyển qua xử lý trong bộ phân tích phones và ngữ điệu. Tại đây, văn bản nhập sẽ được chia ra thành các câu, cụm từ, từ, âm tiết, âm vị. Mỗi đơn vị sau đó sẽ được gán nhãn ngữ điệu căn cứ theo mô hình ngôn điệu. Cuối cùng, bộ tổng hợp sẽ thực hiện các phép tìm kiếm trong cơ sở dữ liệu ngữ âm/ngữ điệu và tạo ra tiếng nói nhân tạo tương ứng.

2.1 Các Cấp độ trong Mô hình Ngôn điệu

Tùy theo phương pháp tổng hợp tiếng nói mà mô hình ngôn ngữ điệu sẽ được xây dựng theo cách riêng tương ứng. Có 8 cấp độ ngôn điệu trong một hệ thống tổng hợp tiếng nói.

- Cấp độ 1 - ngữ cảnh văn bản: các đơn vị văn bản sẽ được bổ sung thông tin về ngữ cảnh lân cận theo hệ số lẻ. Thông tin ngữ cảnh này giúp cho việc lựa chọn các đơn vị ngữ âm đồng bộ ngữ điệu hơn.
- Cấp độ 2 - nhãn từ loại: đơn vị văn bản ở mức từ đã được gán thêm thông tin về từ loại (part-of-speech tag). Thông tin này góp phần đồng bộ giữa ngữ âm và văn bản theo cấp độ từ loại.
- Cấp độ 3 - định lượng: thông tin ngữ điệu ở cấp độ này có được từ việc định lượng các đơn vị văn bản từ âm vị, âm tiết, từ, cho đến cụm từ, câu... Định lượng cả về số lượng và trật tự.
- Cấp độ 4 - thanh điệu: đặc trưng của từng ngôn ngữ (ví dụ tiếng Việt). Ngữ điệu của tiếng nói sẽ không thể tự nhiên suông mượt nếu không xét đến yếu tố này.
- Cấp độ 5 – quy tắc phát âm: thông tin ngữ điệu có được ở cấp này chủ yếu dựa theo cơ chế phát âm của ngôn ngữ đang xét. Các âm chập, âm đóng, âm trượt... sẽ có trường độ, cao độ khác nhau rất nhiều. Nếu không xét đến yếu tố này, giọng đọc tổng hợp sẽ không tự nhiên tốt được.
- Cấp độ 6 – nhịp: ngưng nghỉ trong tiếng nói là tất yếu. Đặc biệt là trong tiếng Việt, nếu ngưng nghỉ sai chỗ, ngữ nghĩa câu có thể thay đổi. Tuy nhiên, ý nghĩa của nhịp ngưng nghỉ còn quan trọng trong ngữ điệu tiếng nói. Nhịp hay là nhịp tự nhiên theo cách đọc của con người.
- Cấp độ 7 – cảm xúc: điều không thể thiếu trong ngữ điệu tiếng nói chính là cảm xúc. Cảm xúc là nhấn mạnh, lên giọng, xuống giọng, đều giọng, hay cảm thán theo loại câu, hoặc ở cấp độ cao là theo đúng nghĩa đen của từ cảm xúc như người đọc.
- Cấp độ 8 – thể loại văn bản: đọc truyện sẽ khác đọc thơ và đọc tin tức. Mỗi thể loại văn bản thường sẽ có một cách đọc phù hợp với nó. Việc xác định thể loại văn bản là một điều không dễ. Tuy nhiên, nếu khai thác được thông tin này, ngữ điệu của tiếng nói tổng hợp sẽ được cải thiện nhiều.

2.2 Các Phương pháp Tiếp cận

Về phương pháp tiếp cận, phương pháp tiếp cận xây dựng dựa trên luật (quy tắc) và dựa trên ngữ liệu là hai phương pháp tiếp cận chính cho mô hình ngôn điệu.

- Phương pháp dựa trên luật (qui tắc): các chuyên gia ngôn ngữ trích rút được một tập hợp phức tạp các quy tắc để mô hình ngôn điệu có thể biến thể điệu tính bằng cách quan sát ngôn luận tự nhiên. Phương pháp này thực hiện phân

tích các phân đoạn tiếng nói bằng tay để làm cơ sở nền tảng cho bước xử lý tổng hợp tiếng nói, tuy nhiên phương pháp này không thể thực hiện được khi dữ liệu lớn. Có thể thấy phương pháp này phụ thuộc vào ngôn ngữ học, ngữ âm và các yếu tố ảnh hưởng đến thời gian của các đơn vị âm thanh như phân đoạn, âm tiết hoặc âm vị. Nhìn chung các phương pháp dựa trên luật khó khi triển khai nghiên cứu, do sự tương tác phức tạp giữa các tính năng ngôn ngữ ở các cấp độ khác nhau.

- Với phương pháp tiếp cận dựa trên ngữ liệu, tập ngữ liệu được tạo ra là một tập ngữ liệu đặc biệt mà trong đó thông tin về các cấp độ ngữ điệu được chú thích với mức độ khác nhau của thông tin điệu tính được sử dụng. Như thế, trong tiếp cận này ngữ liệu được phân tích tự động để tạo ra các mô hình ngôn điệu và sau đó được đánh giá trên các dữ liệu thử nghiệm. Căn cứ vào hiệu suất của dữ liệu thử nghiệm, các mô hình được xem xét, đánh giá và cải thiện. Phương pháp có ưu điểm so với phương pháp dựa trên luật. Phương pháp này hiệu quả khi có các đơn vị ngữ âm đủ lớn (ví dụ như câu, cụm từ, từ), phong phú có độ phủ trong tập ngữ liệu. Phương pháp này được dựa trên một trong hai mô hình tham số hoặc không tham số sử dụng chức năng xác suất hoặc tối ưu hóa khả năng kết hợp các tham số.
- Phương pháp lai: phương pháp này tiếp cận sử dụng cách kết hợp của cả hai phương pháp dựa trên luật và thống kê.

Trong công trình [6], tác giả Krishna cho rằng, thách thức trong mô hình hóa ngôn điệu là việc xem xét nhiều tham số khác nhau và có tính kết hợp với nhau (ví dụ như âm tiết) cho mô hình theo thời gian và từ công trình [7] như âm vị cho mô hình ngôn điệu.

Mô hình Ngôn điệu Dựa trên Luật

Mô hình ngôn điệu dựa trên luật khá tự nhiên được thực hiện trên cơ sở các tri thức tiềm ẩn hoặc rõ ràng được rút trích từ ngữ liệu.

Trong công trình [10], tác giả Ovidiu Buza và cộng sự trình bày rằng các quy tắc cần phải được quan tâm ở các giai đoạn khác nhau trong quá trình tạo ra tiếng nói tổng hợp như giai đoạn tiền xử lý văn bản đầu vào, giai đoạn xử lý tín hiệu số và các quy tắc của ngôn ngữ đang xử lý trong hệ thống tổng hợp tiếng nói. Cụ thể trong giai đoạn tiền xử lý văn bản, các công việc cần thực hiện là xác định các qui tắc về ngữ âm để xác định âm, thông tin điệu tính và chuẩn hóa văn bản. Trong giai đoạn xử lý tín hiệu số, phân đoạn âm thanh tiếng nói là một tác vụ quan trọng phải được thực hiện. Tác vụ phân đoạn này dựa trên việc sử dụng các luật kết hợp đặc biệt để nhận diện nhóm các đơn vị âm thanh kết hợp với nhau từ trong ngữ liệu dựa theo đặc điểm của ngôn ngữ.

Có thể nhận thấy rằng, trong phương pháp này mặc dù bộ luật (quy tắc) đã được định nghĩa trong hệ thống tổng hợp tiếng nói, tuy nhiên phạm vi bao phủ của âm tiết là rất hạn chế. Do đó, các bộ luật (quy tắc) là không hoàn chỉnh, vì không phát hiện chính xác 100% âm tiết và thường chỉ chính xác 98% âm tiết chính xác được xác định. Phương pháp này được thiết kế dựa trên cách tiếp cận tôn trọng quy tắc ngôn

ngữ và các tính năng của ngôn ngữ có liên quan. Tuy nhiên, cách tiếp cận này không bảo đảm tính khả thi cho hệ thống đa ngôn ngữ nói.

Mô hình Ngôn điệu Dựa trên Thống kê

Về mặt tổng quát hóa, mô hình ngôn điệu dựa trên ghép nối trong các hệ thống tổng hợp tiếng nói là dự đoán giá trị cao độ, thời gian và các cách kết hợp tường minh hay tiềm ẩn của các đơn vị âm thanh ứng viên trong kho ngữ liệu âm thanh (có ngôn điệu khác nhau) tương ứng với nội dung văn bản cần tổng hợp để tạo thành tiếng nói tổng hợp. Các đơn vị âm thanh này có thể là cụm từ, từ hoặc âm tiết trong câu. Các mô hình ngôn điệu dựa trên xác suất được xây dựng trong hệ thống tổng hợp tiếng nói để dự đoán xác suất hoặc khả năng kết hợp tối ưu của các đơn vị ghép nối. Các chi phí tính toán bao gồm chi phí mục tiêu và chi phí chuyển đổi.

Trong công trình [4], sáu mô hình cụ thể đã được tác giả xây dựng nhằm hướng đến xác định chi phí và xác suất cho việc kết ghép. Tất cả các mô hình huấn luyện được thực hiện dựa trên cây quyết định bối cảnh phụ thuộc và dữ liệu được nhóm lại với nhau và được biểu diễn trên các nút lá của cây quyết định theo mô hình xác suất Gaussian (GMM - Gaussian Mixture Model).

Cũng tương tự như cách tiếp cận trên, trong công trình [5], tác giả trình bày cách thực hiện theo hướng sử dụng một cây quyết định T và tiến hành duyệt cây T theo bối cảnh phụ thuộc của các nút tương ứng theo mô hình Gaussian M, do đó chi phí được xác định như sau cho một ứng viên x theo một ngữ cảnh cho trước như sau :

$$C(x_i | M_i^1) = -\log P(x_i | M_i^1) \quad (1)$$

Trong cách thực hiện này, có thể thấy rằng :

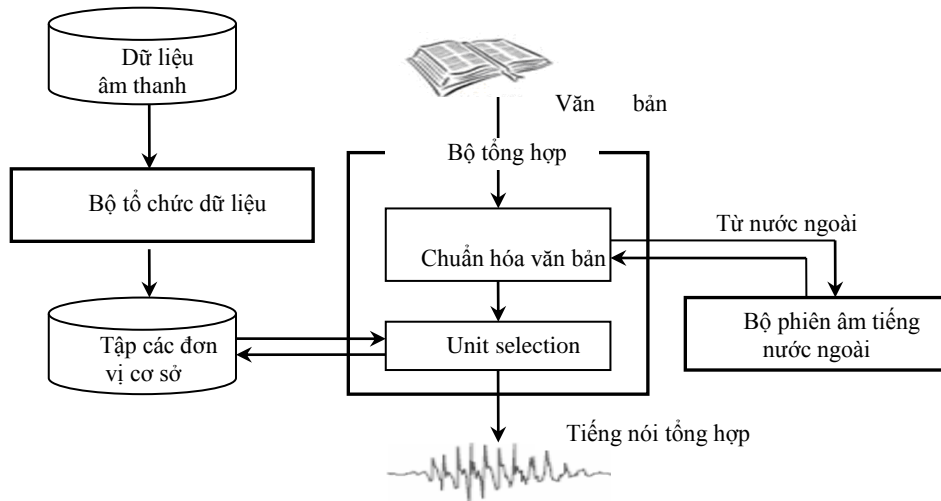
- Chi phí có thể nhỏ hơn không.
- Mỗi GMM mục tiêu là để tối ưu hóa đầu ra tại mức địa phương nhưng không phải ở mức toàn cục.
- Trọng số cho các mô hình khác nhau là khác nhau.
- Đặc điểm và các xử lý về ngôn ngữ cụ thể có thể được tích hợp để điều chỉnh trọng số tính toán
- Phương pháp này thiếu tối ưu hóa.

Trong [5], xác suất có điều kiện được định nghĩa như sau :

$$P(M_i^1 | x_i) = \frac{P(x_i | M_i^1) P(M_i^1)}{\sum_{j=1}^N P(x_i | M_j^1) P(M_j^1)} \quad (2)$$

$$C_2 = -\log P(M_i^1 | x_i) = -\log \frac{P(x_i | M_i^1) P(M_i^1)}{\sum_{j=1}^N P(x_i | M_j^1) P(M_j^1)} \quad (3)$$

Vì vậy, mô hình xác suất tính chi phí C_2 (công thức 3) có thể đạt được tối ưu hóa toàn cục tốt hơn so với cách tính chi phí ở trên. Rõ ràng mô hình tính xác suất theo như cách này chỉ được thực hiện trong phạm vi dữ liệu phù hợp.



Hình 3. Mô hình hệ thống tổng hợp tiếng nói dựa trên ghép nối

Mô hình Ngôn điệu Lai

Mô hình lai là sự kết hợp của hai mô hình dựa trên luật và thống kê. CART là một mô hình lai được sử dụng rộng rãi cho mô hình ngôn điệu. Các nghiên cứu trước đó thực hiện tiền xử lý gom nhóm các âm tiết dựa trên vị trí của âm tiết trong từ.

Trong công trình [8], tác giả Ashwin Bellur (2011) đã thực hiện gom nhóm các âm tiết cùng loại, quan tâm các thông tin điệu tính như cao độ và các đặc trưng về ngữ âm. Mô hình CART đã sử dụng theo hướng tiếp cận trong [8], đồng thời CART định nghĩa hàm đo khoảng cách giữa các âm tiết để phân biệt giữa các âm tiết. Cụ thể, đầu tiên tập các âm tiết được xác định và sau đó các đặc trưng được lựa chọn. Việc lựa chọn các đặc trưng phải được thực hiện theo cách dựa trên tất cả các âm của các âm tiết thu được.

Cũng trong công trình [8], tác giả đã xây dựng cây quyết định CART cho hệ thống. Dựa trên cây quyết định, sẽ cho dự đoán biên của các cụm từ sau các (cụm) từ trước. Một đặc trưng mới được sử dụng để dự đoán biên giữa các cụm từ (morpheme tag). Như thế có hai cách thức tổng hợp tiếng nói, một là thực hiện bằng tay thao tác đánh dấu biên giữa các cụm từ, và hai là thực hiện thao tác đánh dấu tự động bằng cách sử dụng cây quyết định như trình bày ở trên. Các kết quả thử nghiệm được tiến hành và quan sát thấy rằng kết quả tổng hợp theo cách tự động cho kết quả khá tốt và cho kết quả tốt hơn so với các cách thực hiện dự đoán biên giữa các cụm từ trước đó.

Trong công trình [18], tác giả thực hiện một mô hình ngôn điệu gồm 3 thành phần F0, cường độ và phân đoạn âm thanh dựa trên cách tiếp cận bằng cách sử dụng CART và thử nghiệm cho ngôn ngữ Séc, kết quả cho thấy hệ thống có được ngữ điệu tốt hơn. Tuy nhiên, khi áp dụng mô hình này bằng cách áp dụng tiếp cận CART riêng biệt cho mỗi âm vị thì điều này không thể thực hiện cho ngữ liệu lớn.

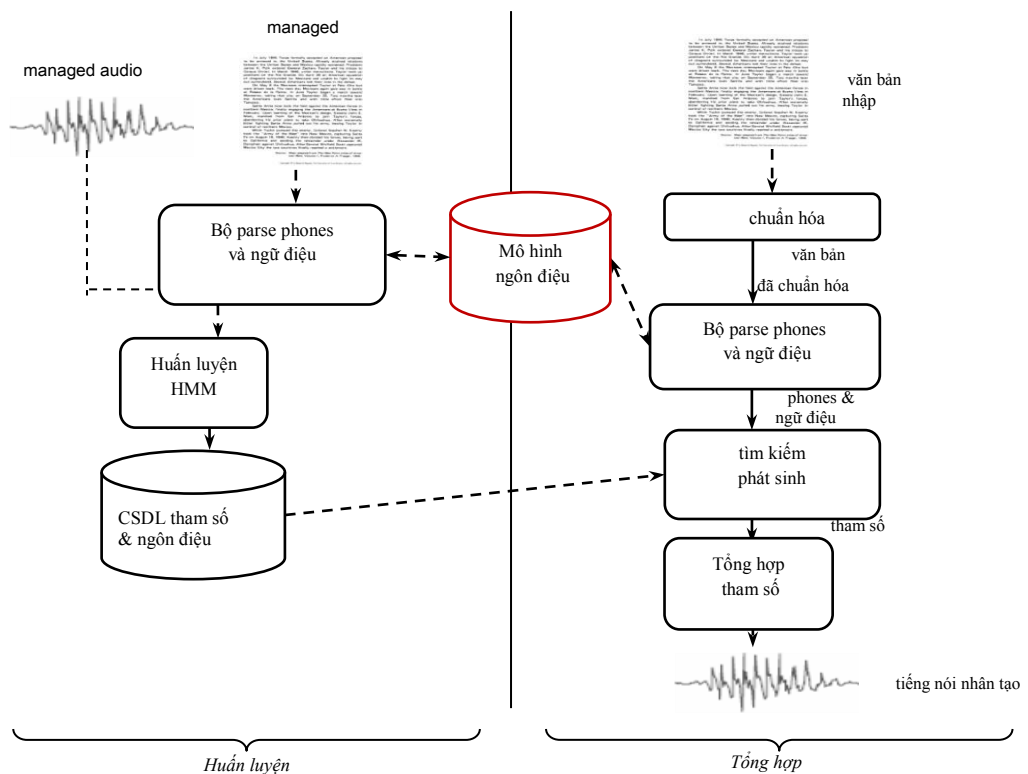
Trong cách thực hiện này, các ký hiệu đánh dấu ngắt (morpheme tag) cần được liệt kê riêng biệt cho mỗi ngôn ngữ, đặc biệt là cho các ngôn ngữ thiếu dấu ngắt câu. Mô hình lai có lợi thế của cả hai phương pháp tiếp cận dựa trên quy tắc dựa và phương pháp tiếp cận thống kê, nhưng cần được tiếp tục phân tích về tính hiệu quả khi áp dụng cho các hệ thống tổng hợp tiếng nói với nhiều ngôn ngữ khác nhau.

Mô hình Ngôn điệu và Phương pháp Tổng hợp Tiếng nói Dựa trên HMM

Mô hình HMM (Hidden Markov Model) là một trong những mô hình tốt nhất hiện nay, sử dụng cho hầu hết các hệ thống tổng hợp giọng nói.

Trong công trình [17], tác giả trình bày vấn đề hạn chế trong mô hình HMM là các biến thể trong các tham số về điệu tính. Để khắc phục nhược điểm này, một cải thiện là hướng đến khai thác các đặc trưng ở các cấp độ khác nhau của ngôn ngữ được trích rút từ trong văn bản cần tổng hợp, qua đó gia tăng chất lượng của tiếng nói tổng hợp.

Trong công trình [20], một lần nữa cách kết hợp giữa HMM và đặc trưng ở các cấp độ khác nhau của ngôn ngữ được trích rút từ trong văn bản cần tổng hợp được sử dụng và khẳng định có thể mang đến chất lượng tốt cho hệ thống tổng hợp tiếng nói. Thông tin ngôn điệu tốt hơn, có thể để đạt được tiếng nói tổng hợp dễ hiểu.



Hình 4. Mô hình hệ thống tổng hợp tiếng nói dựa trên HMM

3 Các Mô hình Ngôn điệu Khác

Trong công trình [12], tác giả đã phát triển một phương pháp mô hình hóa và tạo ra các thành phần điệu tính, cường độ sử dụng mô hình HMM. Phương pháp này sử dụng S-CART để dự đoán điểm ngắt điệu tính và U-CART để tạo các đường cao độ.

Trong công trình [13], mô hình HMM được kết hợp với các mô hình ANN (Artificial Neural Network) được đề xuất bởi GU Hung-Yan để gia tăng chất lượng đồng thời ở khía cạnh ngôn điệu và khía cạnh tạo ra âm thanh tổng hợp có chất lượng lưu loát.

Trong công trình [14], mô hình khung gán nhãn tự động ngôn điệu dựa trên mô hình cực đại Entropy đã được sử dụng cho cả hai khía cạnh là thông tin về ngôn ngữ và tiếng nói.

Ở công trình [15], một hệ thống tổng hợp tiếng nói được xây dựng, trong đó mô hình ngôn điệu được xây dựng để nhận diện tiếng địa phương của ngôn ngữ Tamil thông qua việc giám sát giá trị của các tham số về thời gian phát âm, F0, và các giá trị quan trọng khác như phạm vi và cao độ lên xuống. Điều quan trọng là xem xét tất cả các phương pháp, mô hình và các tham số có liên quan trong khi thiết kế mô hình ngôn điệu để tổng hợp tiếng nói cho bất kỳ ngôn ngữ cụ thể nào.

Với công trình [16], một mô hình ngôn điệu đa cấp phụ thuộc bối cảnh được định nghĩa để ước lượng mức độ các đơn vị ngôn ngữ có tác động đến sự biến thiên của các tham số điệu tính trên mỗi mức độ độc lập. Bằng việc áp dụng phương pháp này hiệu suất được cải thiện trong cả hai khía cạnh, một là dự đoán khoảng thời gian phát âm tốt hơn và hai là dự đoán lỗi.

Trong một công trình khác [17], mô hình HMM được cải thiện để khắc phục hạn chế điểm hạn chế hiện nay dựa trên HMM đó là thiếu biến thể các tham số điệu tính.

Với công trình [20], một mô hình thời gian mở rộng được sử dụng để phân tích ba cách tiếp cận khác nhau để cải thiện chất lượng của tiếng nói tổng hợp dựa trên mô hình HMM. Ba cách tiếp cận khác nhau là mô hình ED (Explicit Duration), ID (Implicit Duration) và mô hình lai khi kết hợp giữa ED và ID. Kết quả thực nghiệm cho thấy ED cho kết quả tốt hơn khi ước lượng thời gian phát âm của một âm tiết. Qua kết quả cũng cho thấy ID không tốt bằng ED. Mô hình lai thực hiện theo hướng tận dụng các ưu điểm của ED và ID, trong đó đẩy mạnh ở bộ phận xử lý ngôn ngữ khi đề xuất rút trích các tham số đặc trưng ngôn ngữ ở mức độ cao để cải thiện chất lượng tiếng nói tổng hợp.

Tại công trình [19] của tác giả Yu-Lun Chou khảo sát ý nghĩa về thông tin điệu tính của tiếng nói tổng hợp qua việc mô hình hóa và gán nhãn ngôn điệu cho các ứng dụng tiếng nói.

Bảng dưới đây so sánh điểm mạnh và điểm yếu trong mỗi hướng tiếp cận xây dựng mô hình ngôn điệu.

Bảng 1. So sánh điểm mạnh, điểm yếu trong mỗi hướng tiếp cận xây dựng mô hình ngôn điệu.

Hướng tiếp cận	Điểm mạnh	Điểm yếu
Tiếp cận dựa trên luật	Yêu cầu ít tài nguyên.	Cách tiếp cận tự nhiên. Không làm việc với lượng dữ liệu lớn
Tiếp cận dựa trên thống kê	Yêu cầu lượng lớn dữ liệu để thực hiện	Ít phù hợp cho bộ dữ liệu thực tế. Không tối ưu
Tiếp cận lai	Kết hợp lợi thế của cả hai phương pháp tiếp cận dựa trên luật và thống kê.	Nếu ngôn ngữ thiếu dấu chấm câu cần phải bổ sung.
Mô hình phụ thuộc ngữ cảnh	Các hình thức ngôn điệu có thể được phối hợp quan sát và mỗi mức điệu tính có thể được mô hình hóa và kiểm soát độc lập với nhau	Có sai số tương đối
Mô hình hóa và gán nhãn	Có khả năng có được thông tin ngôn điệu phong phú	Thích hợp nhất trong ngữ liệu thoại (giao tiếp)
Các mô hình ngôn điệu dựa trên HMM	Thông tin ngôn điệu tốt hơn, có thể để đạt được tiếng nói tổng hợp dễ hiểu	Phải làm cho tiếng nói tổng hợp được tự nhiên

4 Mô hình Ngôn điệu trong Hệ thống Tổng hợp Tiếng nói Đặc biệt

Mô hình hóa ngôn điệu còn được nghiên cứu triển khai trong hệ thống tổng hợp tiếng nói đặc biệt khác, như hệ thống tổng hợp tiếng nói có cảm xúc (Emotional Speech), ...Thách thức lớn khi xử lý các dữ liệu phức tạp loại này là phải hướng đến đọc dữ liệu dựa trên một mô hình ngôn điệu đã được mô hình hóa trước đó [21].

Tương tự như thế, trong các nghiên cứu hướng về nghiên cứu tạo ngôn điệu có cảm xúc như trong công trình [22], đường cao độ được phân cấp thành các cấu trúc phân cấp câu, điệu tính của từ và âm tiết. Trong công trình [23], tác giả đã trình bày rõ tầm quan trọng của mô hình hóa ngôn điệu cho bài toán xây dựng hệ thống tổng hợp theo cảm xúc bằng cách xem xét xem liệu các tính năng điệu tính độc lập có thể đạt được sự phù hợp (biểu hiện cảm xúc phù hợp với nội dung bằng lời nói) và hiệu quả (cảm xúc biểu hiện liên quan với thái độ của người nói). Kết quả thu được cho thấy rằng các đặc trưng ngôn điệu có tác động để đạt được kết quả có ý nghĩa trong việc tạo ra tiếng nói có cảm xúc, tuy nhiên không cần thiết phải sử dụng một tập ngữ liệu đặc biệt trong đó có dữ liệu mang tính cảm xúc.

Tổng hợp tiếng nói cho các dữ liệu dạng bảng có cấu trúc (table-to-speech) cũng là một hệ thống tổng hợp tiếng nói đặc biệt. Trong hệ thống này cần phải thực hiện theo hướng có sự hiểu biết về cấu trúc ngữ nghĩa của dữ liệu. Đối với điều này, một tập hợp các tham số về ngôn điệu phải có trước và được phân tích trong mối tương quan với âm giọng của người nói, với cụm từ và thời gian ngừng nghỉ sao cho đảm bảo không vi phạm tính nhất quán trong nội dung và tính trực quan của cấu trúc dữ liệu.

5 Kết luận

Báo cáo trình bày tổng thể về vai trò của mô hình ngôn điệu trong hệ thống tổng hợp tiếng nói. Điểm mạnh, điểm yếu của từng phương pháp mô hình hóa cũng được trình bày, qua đó thấy được các góc nhìn khác nhau về các phương pháp khi nghiên cứu mô hình hóa cho một ứng dụng, hệ thống tổng hợp tiếng nói cụ thể. Các thách thức trong quá trình xây dựng mô hình ngôn điệu để tạo cho tiếng nói nhân tạo tự nhiên như cách con người giao tiếp luôn là một vấn đề không nhỏ, tuy nhiên, những thách thức này có thể được giải quyết trên cơ sở sự hỗ trợ mạnh mẽ của các chuyên gia ngôn ngữ và nỗ lực của chuyên gia kỹ thuật, trong kết hợp để tạo ra các mô hình phù hợp. Mô hình ngôn điệu có tầm quan trọng không chỉ trong các hệ thống tổng hợp tiếng nói hiện nay mà còn trong các hệ thống chuyên biệt, đặc thù khác nhau và tổng hợp tiếng nói nhấn mạnh cảm xúc. Do đó, các nghiên cứu trong tương lai sẽ được tập trung trong việc phát triển mô hình ngôn điệu để nâng cao chất lượng giọng nói tổng hợp theo hướng tiệm cận gần hơn với giao tiếp của con người. Tuy nhiên, đó là một con đường dài cần phải tiếp tục chinh phục trong thời gian tới.

Tài liệu tham khảo

1. M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, "Text-to-speech synthesis using syllable like units," in National Conference on Communication, Kharagpur, India, Jan 2005, pp 277-280
2. G.L.Jayavardhana Rama, A G Ramakrishnan, R. Muralishankar and Vijay Venkatesh" Thirukkural – A text to speech synthesis system". Proc. Tamil Internet 2001, Kuala Lumpur 2001, 92-97.
3. Vinodh M Vishwanath, Ashwin Bellur, Badri Narayan K, Deepali M Thakare, Anila Susan, Suthakar N M and Hema A Murthy, "Using Polysyllabic units for Text to Speech Synthesis in Indian languages," Proceedings of National Conference on Communication, pp.1-5, 29-31, Jan. 2010.
4. X.J. Ma, W. Zhang, W.B. Zhu, Q. Shi and L. Jin, "Probability Based Prosody Model for Unit Selection", ICASSP 2004, Montreal, Canada
5. Wei Zhang, Liang Gu and Yuqing Gao "Recent improvements of probability based prosody model for unit selection in concatenative Text to Speech", in the proceedings of ICASSP 2009, pp 3777-3780
6. N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, A. G. Ramakrishnan, "Duration Modeling for Hindi Text to Speech Synthesis System", in Proc. ICSLP 2004, South Korea, 2004.

7. A. S. Madhukumar, S. Rajendran and B. Yegnanarayana, "Intonation component of a Text to Speech system for Hindi", *Proceedings of International journal of Computer Speech and Language*, 1993, Volume7, pp 283-301
8. Ashwin Bellur, K Badri Narayan, Raghava Krishnan K, Hema A Murthy, "Prosody modeling for syllable based concatenative speech synthesis of Hindi and Tamil", in *National conference on Communications*, Jan 2011, pp 28-30.
9. Samuel Thomas, M. Nageshwara Rao, Hema A.Murthy and C.S. Ramalingam, "Natural sounding TTS based on syllable-like units," in the proceedings of the 14th European Signal Processing Conference, Florence, Italy, Sep 2006.
10. Ovidiu Buza, Gavril Todorean, Jozsef Domokos, "A rule based approach to build a Text to speech system for Romanian", in *proceedings of international Conference on communications*, June 2010, pp. 33-36.
11. G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and R Prathibha, "A Complete Text-To- Speech Synthesis System in Tamil", in 0-7803-7395-2/02, *IEEE proceedings of ICASSP*, 2002.
12. Chi-Chun Hsia, Chung-Hsien Wu, and Jung-Yun Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM based speech synthesis", in *International journal of Audio, Speech and Language processing*, Nov 2010, Volume 18, pp,1994-2003.
13. Hung-Yan GU, Ming-Yen LAI and Sung-Feng TSAI, "Combining HMM spectru models and ANN prosody models for speech synthesis of syllable prominent languages", in *International journal of Audio, Speech and Language processing*, 2010, pp,451-454.
14. Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework", in *International Journal of Audio, Speech and Language processing*, May 2008, Volume 16, pp,797-811.
15. Raja Mohamed S, Raviraj P, "Prosodic Feature Extraction for Regional Tamil dialects", in *International Conference on emerging Trends in electrical and Computer Technology*, March 2011, pp 922-925.
16. Nicolas Obin, Xavier Rodet and Anne Lacheret Dujour, "A multi-level context-dependent prosodic model applied to duration modeling", in the tenth annual conference, *Inerspeech*, France, 2009.
17. Nicolas Obin, Pierre Lanchantin, Mathieu Avanzi, Anne Lacheret-Dujour and Xavier Rodet, "Towards improved HMM-based speech synthesis using high-level syntactical features", in the fifth International Conference on Speech Prosody, Chicago, 2010.
18. Jan Romportl and Jiri Kala, "Prosody Modeling in Czech Text-to-Speech Synthesis", in the proceeding of Sixth International workshop on speech synthesis, 2007.
19. Yu-Lun Chou, Chen-Yu- Chiang, Yih-Ru Wang, Hsui-Min Yu and Sin-Horng Chen, "Prosody labeling and modeling for Mandarin spontaneous Speech", in the International Conference on Speech Prosody, Chicago, 2010.
20. Javier Latorre, Sabine Buchholz, Masami kamine, "Usages of an external duration model for HMM-based speech synthesis", in fifth International conference on Speech Prosody, Chicago, 2010
21. Dimitris Spiliotopoulos, Gerasimos Xydias, and Georgios Kouroupetroglou, "Diction Based Prosody Modeling in Table-to-Speech Synthesis", in *LNAI 3658*, pp. 294-301, 2005
22. Chung-Hsien Wu, Chi-Chun Hsia, Chung-Han Lee, and Mai-Chun Lin, "Hierarchical Prosody Conversion using Regression-Based Clustering for Emotional Speech Synthesis", in *IEEE Transactions on Audio, Speech and Language Processing*, Vol.18, No.6, August 2010.
23. Dan-ning Jiang, Wei Zhang, Li-qin Shen and Lian-Hong Cai, "Prosody Analysis and Modeling for Emotional Speech Synthesis", in *IEEE proceedings of ICASSP*, 0-7803-8874-7/05, pp 281-284, 2005.

24. Marc Schröder , Foundations of Language Science and Technology Speech synthesis, book chapter, 2005.
25. Matoušek Jindřich, Acoustic speech synthesis, <http://musslap.zcu.cz/en/acoustic-speech-synthesis/>, 2005.
26. Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, Keiichi Tokuda, The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6, pp.294-299, Bonn, Germany, Aug. 2007.
27. Tu Trong Do and Tomio Takara, “Vietnamese Text-To-Speech system with precise tone generation”, Acoust. Sci. & Tech. 25, 5 (2004), pp. 247-353
28. Cao Nam, Ha Nguyen, Quan VU, “Phrase-Based Concatenation for Vietnamese TTS”, Tạp chí Công nghệ thông tin và truyền thông. Chuyên san: "Các công trình nghiên cứu, phát triển và ứng dụng công nghệ thông tin và truyền thông", 2009.