Tạp chí Tin học và Điều khiển học, T.29, S.1 (2013), 55–65

TRÍCH CHỌN CÁC THAM SỐ ĐẶC TRƯNG TIẾNG NÓI CHO HỆ THỐNG TỔNG HỢP TIẾNG VIỆT DỰA VÀO MÔ HÌNH MARKOV ẨN

PHAN THANH SƠN, DƯƠNG TỬ CƯỜNG

Học viện Kỹ thuật Quân sự; sonphan.hts@gmail.com

Tóm tắt. Phương pháp tổng hợp tiếng nói dựa trên mô hình Markov ẩn (HMM) chỉ cần một kho ngữ liệu tiếng nói thu âm sẵn đủ lớn (bao hàm tất cả các âm vị của một ngôn ngữ) để phục vụ cho mục đích huấn luyện. Trong phương pháp này, mô hình thống kê được sử dụng để mô hình hóa sự phân bố của các véctơ âm thanh phụ thuộc ngữ cảnh, các véctơ này được trích rút từ tín hiệu tiếng nói, mỗi véctơ là một tham số đặc trưng cho khung tín hiệu và các qui tắc ngữ âm tiếng Việt, phục vụ cho quá trình tổng hợp tiếng nói. Hiệu quả của hệ thống bị hạn chế bởi mức độ chính xác khi tham số hóa các đặc trưng tiếng nói và phương pháp tái tạo tín hiệu tiếng nói từ những tham số này. Bài báo này giới thiệu một phương pháp trích chọn các tham số MFCC, F0 và tái tạo tín hiệu tiếng nói chất lượng cao sử dụng bộ lọc MLSA. Phương pháp này thích hợp cho tổng hợp tiếng nói dựa trên HMM và kết quả của nó được đánh giá qua thực tế là khá tốt so với một số phương pháp khác.

Từ khóa. Tổng hợp tiếng Việt, tham số hóa tiếng nói, tổng hợp tiếng nói tham số thống kê, mô hình Markov ẩn, hệ số phổ tần số thang Mel, tần số cơ bản.

Abstract. Recently, the statistical framework based on Hidden Markov Models (HMMs) plays an important role in the speech synthesis method. The system can be built without requiring a very large speech corpus for training the system. In this method, statistical modeling is applied to learn distributions of context-dependent acoustic vectors extracted from speech signals, each vector contains a suitable parametric representation of one speech frame and Vietnamese phonetic rules to synthesize the speech. The overall performance of the systems is often limited by the accuracy of the underlying speech parameterization and reconstruction method. The method proposed in this paper allows accurate MFCC, F0 and tone extraction and high-quality reconstruction of speech signals assuming Mel Log Spectral Approximation filter. Its suitability for high-quality HMM-based speech synthesis is shown through evaluations subjectively.

Key words. Vietnamese speech synthesis, context-dependent, speech parameterization, statistical parametric speech synthesis, Hidden Markov Models, mel-frequency cepstral coefficient, fundamental frequency.

1. GIỚI THIỆU

Các phương pháp tổng hợp tiếng nói ở mức thấp có thể kể đến là: mô phỏng bộ máy phát âm, tổng hợp format, ghép nối và tổng hợp các tham số thống kê dựa trên các mô hình Markov ẩn. Về mặt lý thuyết, phương pháp tổng hợp bộ máy phát âm cho chất lượng tiếng nói chính xác nhất bởi vì phương pháp này mô phỏng hệ thống tạo tiếng nói con người một

cách trực tiếp, nhưng nhược điểm phương pháp này khó tiếp cận. Tổng hợp format dựa trên việc mô hình hóa sự cộng hưởng của các dây thanh khi phát âm, đây là phương pháp tiếp cận tổng hợp tiếng nói phổ biến nhất trong một vài thập niên qua. Tổng hợp ghép nối là phương pháp dựa trên sự ghép nối một lượng lớn các mẫu thu âm sẵn để tạo ra tiếng nói với chất lượng tự nhiên nhất. Phương pháp này đang được ứng dụng phổ biến trong các hệ thống tổng hợp tiếng nói có sử dụng server (chẳng hạn như các hệ thống giải đáp, trả lời tự động, hệ thống dịch tiếng nói), nhưng nhược điểm của hệ thống này là thụ động, không linh hoạt (phụ thuộc vào server), không ổn định, thời gian đáp ứng (phụ thuộc vào đường truyền), đặc biệt là khi chúng ta cần khả năng tổng hợp tiếng nói với nhiều đặc trưng giọng nói và ngữ điệu khác nhau. Một lý do xuất phát từ thực tế, đó là khó có thể chuẩn bị, tổ chức và lưu trữ một số lượng lớn các dữ liệu tiếng nói của nhiều người khác nhau với các cách nói khác nhau. Hệ thống tổng hợp tiếng nói tham số thống kê dựa trên HMM (HTS) đã được nghiên cứu và phát triển phổ biến trong vài năm gần đây để khắc phục nhược điểm này của tổng hợp theo phương pháp ghép nối. Bên cạnh đó, các nghiên cứu, cải tiến thuật toán nhằm nâng cao chất lượng tín hiệu tiếng nói tổng hợp từ các tham số tham số thống kê, dựa trên mô hình Markov ẩn, đang là chủ đề được quan tâm hiện nay [1].

HTS đòi hỏi các tín hiệu đầu vào phải được dịch thành tập các véc tơ dễ xử lý với những đặc trưng tốt. Do đó, các hệ số Mel-frequency Cepstral Coefficients - MFCC (sử dụng trong nhiều lĩnh vực của xử lý tiếng nói) được sử dụng để mô hình hóa phổ tiếng nói trong các hệ thống tổng hợp và chuyển đổi tiếng nói [1]. Ngoài khả năng mô hình hóa phổ, MFCCs còn có một ưu điểm nổi bật là chúng cho phép sử dụng các ma trận hiệp phương sai chéo hóa, vì các thành phần riêng biệt trong mỗi véc tơ ít tương quan với nhau.

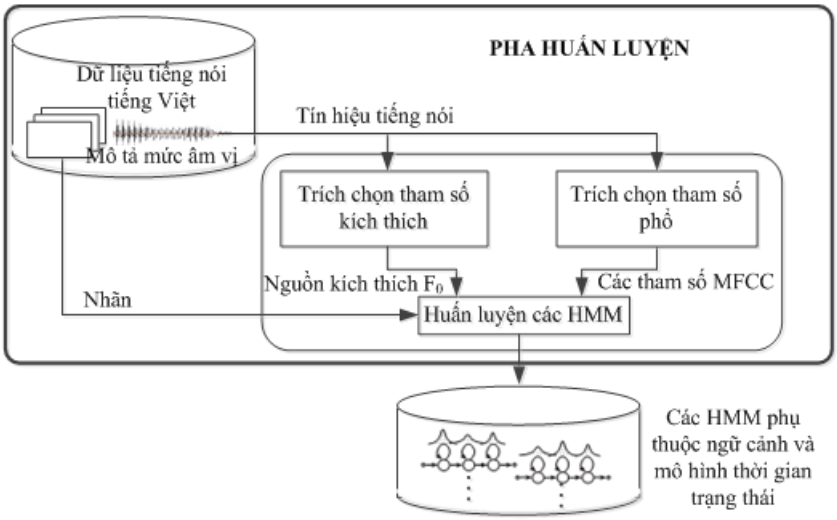
Đặc trưng của HTS là một hệ thống có khả năng huấn luyện các mô hình và tổng hợp tiếng nói không phụ thuộc ngôn ngữ và chỉ cần một kho ngữ liệu thu âm đủ lớn (chứa đủ các âm vị cần có của một ngôn ngữ). Vì vậy, chúng tôi chọn HTS để cải tiến và làm công cụ tổng hợp tiếng Việt (là ngôn ngữ đơn lập âm tiết tính và có thanh điệu). Đồng thời chúng tôi cũng tiến hành thu âm và xây dựng tập dữ liệu tiếng nói tiếng Việt, phục vụ cho việc thử nghiệm tổng hợp và so sánh, đánh giá kết quả. Tín hiệu tiếng nói dạng sóng trong cơ sở dữ liệu được phân đoạn và gán nhãn với các thông tin ngữ cảnh như thanh điệu, âm tiết, từ, cụm từ và câu nói để làm đầu vào cho quá trình huấn luyện các mô hình và tái tạo tiếng nói từ các mô hình này [2].

Bố cục bài báo gồm: Mục 1 giới thiệu tổng quan, Mục 2 mô tả sơ lược hệ thống tổng hợp tiếng nói áp dụng cho tiếng Việt dựa trên HTS. Các kết quả thử nghiệm tổng hợp tiếng Việt được đề cập đến trong Mục 3, chất lượng tiếng nói tổng hợp được so sánh, đánh giá trong Mục 4, và cuối cùng là kết luận và định hướng nghiên cứu.

2. HỆ THỐNG TỔNG HỢP TIẾNG NÓI THAM SỐ THỐNG KÊ DỰA TRÊN HMM

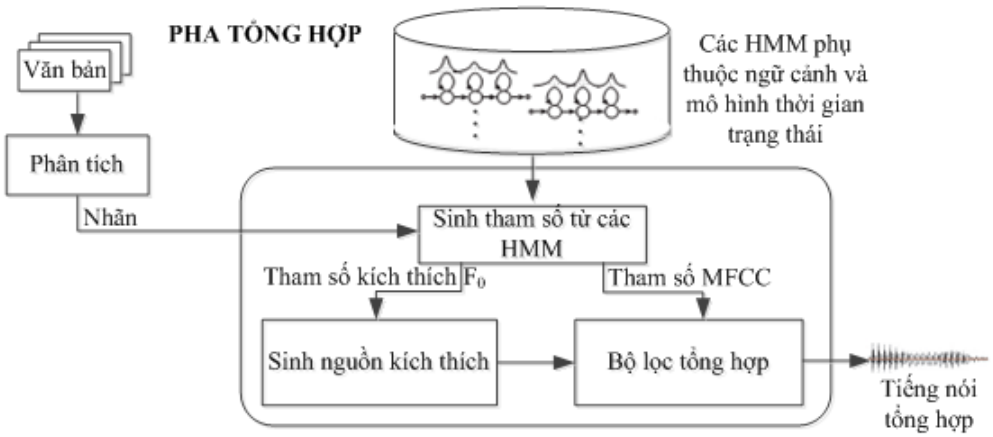
Về mặt lý thuyết, tín hiệu tiếng nói có thể được tổng hợp từ các vectơ đặc trưng. Trong HTS, các vectơ đặc trưng bao gồm các tham số phổ (các MFCC như thanh điệu, trường độ, các tần số khác) và các tham số nguồn kích thích (tần số cơ bản F0).

Hình 1 mô tả pha huấn luyện của hệ thống tổng hợp tiếng nói tiếng Việt, trong phần này,



Hình 1. Pha huấn luyện của hệ thống tổng hợp tiếng nói dựa trên HMM

các tham số phổ (MFCC) và tham số nguồn kích thích (tần số cơ bản) được trích chọn từ cơ sở dữ liệu tiếng nói, sau đó chúng được mô hình bởi các HMM phụ thuộc ngữ cảnh.



Hình 2. Pha tổng hợp của hệ thống tổng hợp tiếng nói dựa trên HMM

Hình 2 minh họa pha tổng hợp của hệ thống tổng hợp tiếng Việt, tại pha này, từ chuỗi các nhãn phụ thuộc ngữ cảnh của văn bản cần tổng hợp mà chuỗi các HMM phụ thuộc ngữ cảnh tương ứng chúng được chọn từ cơ sở dữ liệu các HMM. Sau đó, các tham số phổ, tham số trường độ và tham số kích thích sẽ được sinh ra từ các chuỗi HMM bằng cách sử dụng thuật toán sinh tham số [5]. Cuối cùng, thông qua một bộ lọc tổng hợp, các tham số này được tổng hợp thành tín hiệu tiếng nói ở dạng sóng [6]. Tham số phổ, tham số trường độ và nguồn kích thích là các tham số cần thiết cho mọi bộ lọc tổng hợp, do vậy các tham số này đều phải được mô hình đồng thời bởi các HMM. Chi tiết các phần huấn luyện và tổng hợp áp dụng cho tổng hợp tiếng nói tiếng Việt được miêu tả như sau:

A. Pha huấn luyện

Trong phần huấn luyện, đầu vào là các câu nói được thu âm sẵn và các mô tả mức âm vị của chúng, tiếp đó các HMM phụ thuộc ngữ cảnh của từng âm vị được huấn luyện từ các tham số phổ và nguồn kích thích cùng với các đặc trưng động của chúng. Các tham số phổ được mô hình thông qua việc sử dụng các HMM phân bố liên tục [7], trong khi đó các tham số kích thích lại được mô hình bằng cách sử dụng các HMM phân bố xác suất đa không gian (Multi-Space probability Distribution HMMs, MSD-HMM) để khắc phục sự đan xen của các âm hữu thanh và vô thanh [8]. Đồng thời các mật độ thời gian trạng thái cũng được mô hình bởi các phân bố Gaussian đơn [4].

Quá trình huấn luyện các HMM âm vị sử dụng đồng thời các tham số phổ, tham số trường độ và tham số kích thích trong cùng một cơ chế thống nhất thông qua việc sử dụng các MSD-HMM và các phân bố Gauss đa chiều [8]. Trong khi đó, quá trình huấn luyện các HMM phụ thuộc ngữ cảnh sử dụng đồng thời tần số cơ bản F0 và MFCC. Quá trình phân cụm phụ thuộc ngữ cảnh của các phân bố Gauss được thực hiện độc lập với phổ, tần số cơ bản và thời gian trạng thái do hệ số phâm cụm khác nhau.

1) Mô hình hóa phổ tín hiệu

Trong cách tiếp cận của bài báo này, các MFCC gồm các tham số thanh điệu, thời gian trạng thái và các hệ số delta và delta-delta tương ứng của chúng được sử dụng như là các tham số phổ. Các hệ số delta và delta-delta tương ứng với các tham số thanh điệu, thời gian trạng thái được tính toán nhằm phản ánh sự biến thiên tiếng nói theo thời gian. Các giá trị delta được tính toán dựa trên các giá trị MFCC của các khung tín hiệu lân cận. Ngoài ra giá trị delta của delta (hay còn gọi là acceleration) cũng có thể được tính toán từ các giá trị delta tính toán ở trên.

Các chuỗi vectơ MFCC (trích chọn từ cơ sở dữ liệu tiếng nói), được mô hình bởi các HMM mật độ liên tục. Kỹ thuật phân tích cho phép tổng hợp tiếng nói từ các MFCC nhờ sử dụng bộ lọc Mel Log Spectral Approximation (MLSA) [10]. Các MFCC được trích chọn thông qua phân tích Mel-cepstral bậc 24 (giá trị tối ưu rút ra từ thực nghiệm với nhiều ngôn ngữ khác nhau), sử dụng cửa sổ Hamming 40 ms, độ dịch khung là 8 ms. Các xác suất đầu ra của các MFCC tương ứng với các phân bố Gauss đa biến [2].

2) Mô hình hóa nguồn kích thích

Các tham số nguồn kích thích bao gồm các logarit của tần số cơ bản ($\log F0$) và các hệ số delta và delta-delta tương ứng của chúng. Chuỗi tham số $\log F0$ của các vùng âm vô thanh được mô hình bởi một HMM dựa trên phân bố xác suất đa không gian [8].

3) Mô hình hóa thời gian trạng thái

Mật độ thời gian trạng thái được mô hình thông qua phân bố Gauss đơn [4]. Chiều của các mật độ này chính là số trạng thái của HMM, và chiều thứ n của mật độ thời gian trạng thái tương ứng với trạng thái thứ n của HMM. Cấu trúc các HMM bao gồm các trạng thái trái sang phải, không bỏ qua trạng thái.

Hiện nay, có nhiều kỹ thuật huấn luyện HMM sử dụng mật độ thời gian trạng thái đồng thời. Tuy nhiên, những kỹ thuật này đòi hỏi không gian lưu trữ lớn và khả năng tính toán của hệ thống. Trong bài báo này, mật độ thời gian trạng thái được ước lượng bằng cách sử dụng các xác suất xuất hiện trạng thái nhận được tại lần lặp cuối cùng của quá trình tái ước lượng nhúng [4].

4) Các yếu tố ngữ cảnh phụ thuộc ngôn ngữ

Có nhiều yếu tố ngữ cảnh (ví dụ như: nhận dạng âm tố, trọng âm, phương ngữ, thanh điệu) có ảnh hưởng đến phổ, cao độ và thời gian trạng thái. Chú ý là mỗi HMM phụ thuộc ngữ cảnh tương ứng với một âm vị.

Các yếu tố ngữ cảnh phụ thuộc ngôn ngữ sử dụng trong HTS chính là các nhãn ngữ cảnh và các yếu tố phân cụm ngữ cảnh. Do tiếng Việt là ngôn ngữ có thanh điệu, nên cần có một tập phát âm phụ thuộc thanh điệu và tập ngữ âm và yếu tố điệu tính tương ứng để xây dựng cây quyết định. Vấn đề phân cụm ngữ cảnh dựa vào cây được thiết kế để có được thanh điệu chính xác là vấn đề rất quan trọng trong bài toán tổng hợp các ngôn ngữ thanh điệu, trong đó có tiếng Việt [11, 12].

Một số thông tin ngữ cảnh cần thiết cho quá trình gán nhãn trong dữ liệu tiếng nói tiếng Việt có thể kể đến là [2]:

a) Mức âm vị:

- Âm vị trước, âm vị hiện tại, hai âm vị phía sau;
- Vị trí hiện tại của âm vị trong âm tiết (tính từ đầu và từ cuối âm tiết);

b) Mức âm tiết:

- Thanh điệu của âm tiết trước, âm tiết hiện tại, âm tiết phía sau;
- Số lượng âm vị trong âm vị trước, âm vị hiện tại, âm vị sau;
- Vị trí của âm tiết trong từ hiện tại (tính từ đầu và từ cuối từ);
- Mức độ trọng âm (thể hiện điệu tính);
- Khoảng cách đến âm tiết có trọng âm trước và đến âm tiết có trọng âm sau;

c) Mức từ:

- Loại từ (Part-of-speech) của từ trước, từ hiện tại và từ phía sau;
- Số lượng âm tiết trong từ trước, từ hiện tại và từ phía sau;
- Vị trí của từ trong cụm từ;
- Số lượng từ trong nhóm từ {trước, sau} tính từ vị trí hiện tại;
- Khoảng cách đến từ trước và từ sau tính từ vị trí hiện tại;

d) Mức cụm từ:

- Số lượng âm tiết, từ trong cụm từ trước, cụm từ hiện tại và cụm từ phía sau;
- Vị trí của cụm từ hiện tại trong câu nói;

e) Mức câu nói:

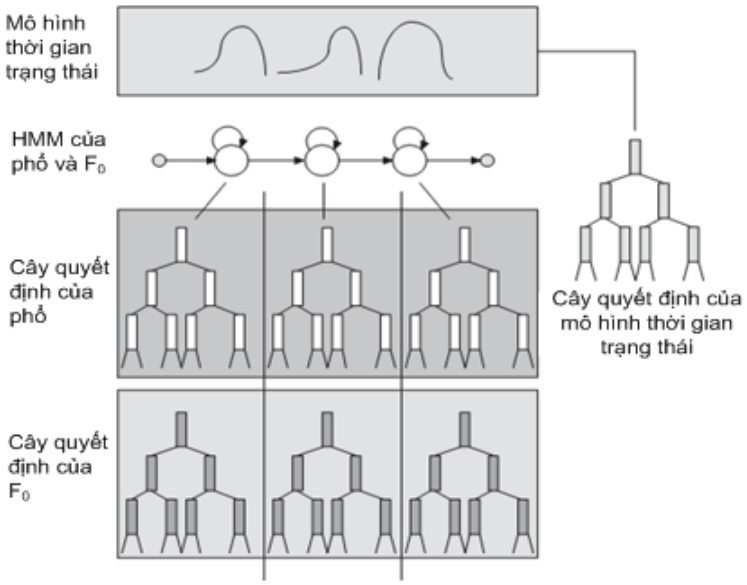
- Số lượng âm tiết, từ, cụm từ trong câu nói;

5) Phân cụm ngữ cảnh dựa vào cây quyết định

Trong một số trường hợp, dữ liệu tiếng nói không có đủ số mẫu ngữ cảnh hoặc sinh ra nhãn ngữ cảnh không tương ứng với HMM trong tập mô hình huấn luyện. Vì vậy, để khắc phục vấn đề này, kỹ thuật phân cụm ngữ cảnh dựa vào cây quyết định được áp dụng vào các phân bố của các tham số phổ, tần số cơ bản và thời gian trạng thái.

Để thực hiện phân cụm ngữ cảnh dựa trên cây quyết định, một số yếu tố quyết định cần phải được xây dựng và tuân theo để phân cụm các âm vị. Sau đó, những yếu tố quyết định này được mở rộng dần để bao hàm tất cả thông tin ngữ cảnh, chẳng hạn như là thanh điệu, âm tiết, từ, cụm từ và câu nói. Các yếu tố quyết định trong pha huấn luyện của HTS được phân chia theo đặc tính ngữ âm của các thanh điệu, nguyên âm, bán nguyên âm, âm đôi và

phụ âm. Các âm vị và thanh điệu được phân lớp để xây dựng các yếu tố quyết định và áp dụng vào quá trình sinh ra các cây quyết định.



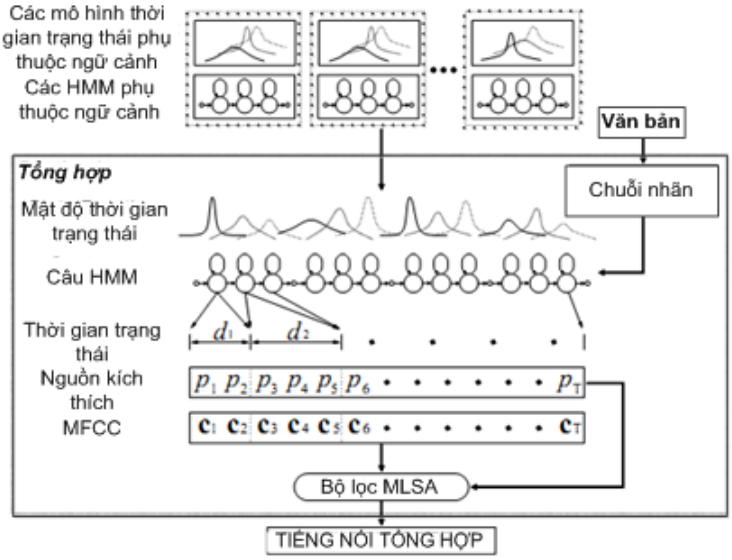
Hình 3. Phân cụm ngữ cảnh dựa vào cây quyết định

Hình 3 minh họa các cây quyết định của phổ, F0 và thời gian trạng thái trước khi chúng được sử dụng trong pha tổng hợp.

B. Pha tổng hợp

Trong pha tổng hợp, các tham số tiếng nói được sinh ra từ tập các HMM phụ thuộc ngữ cảnh thứ tự theo chuỗi nhãn ngữ cảnh tương ứng với phát âm của văn bản cần tổng hợp. Các tham số kích thích và Mel-cepstral sinh ra được sử dụng để tạo ra tín hiệu tiếng nói dạng sóng thông qua một mô hình nguồn lọc (bộ lọc tổng hợp). Ưu điểm của phương pháp tiếp cận này là trích rút được những đặc trưng âm thanh của các phát âm phụ thuộc ngữ cảnh trong kho ngữ liệu tiếng nói. Các đặc tính của tiếng nói tổng hợp có thể dễ dàng thay đổi bằng cách điều chỉnh các tham số HMM và hệ thống cũng hoàn toàn có thể áp dụng cho một ngôn ngữ khác.

Pha tổng hợp của HTS được mô tả trong Hình 4. Trong phần này, một đoạn văn bản tùy ý được phân tích và chuyển đổi thành chuỗi các nhãn phụ thuộc ngữ cảnh. Sau đó, tùy thuộc vào chuỗi nhãn này mà một câu HMM sẽ được sinh ra bằng cách ghép nối các HMM phụ thuộc ngữ cảnh lại với nhau. Các mô hình thời gian trạng thái của câu HMM được xác định để cực đại hóa lân cận các mật độ thời gian trạng thái [6]. Tùy thuộc vào các thời gian trạng thái mà chuỗi các MFCC và các giá trị tham số kích thích (bao gồm âm hữu thanh và vô thanh) được tạo ra từ câu HMM bằng cách sử dụng thuật toán sinh tham số tiếng nói [5]. Cuối cùng, tiếng nói được tổng hợp trực tiếp từ các MFCC và các giá trị tham số kích thích thông qua bộ lọc MLSA [10].



Hình 4. Phần tổng hợp của hệ thống

3. THỬ NGHIỆM

Ở đây, ta sử dụng hai bộ ngữ liệu tiếng Việt của Phòng Nhận dạng và công nghệ tri thức để tiến hành thử nghiệm và đánh giá kết quả của hệ thống tổng hợp thống kê dựa trên HMM. Tất cả dữ liệu tiếng nói thu âm đều được lấy mẫu ở 48 kHz, kênh đơn (mono chanel) và mã hóa ở định dạng PCM 16 bit, sau đó tín hiệu tiếng nói được chuyển đổi về tần số lấy mẫu ở 16 kHz, định khung 40 ms với cửa sổ Hamming và độ dịch khung là 8ms trước khi đưa vào hệ thống để huấn luyện. Hai bộ dữ liệu tiếng Việt: 500 câu giọng nam (trong 568 câu giọng miền Nam) và 500 câu giọng nữ (trong 567 câu giọng miền Bắc) được sử dụng riêng biệt cho quá trình huấn luyện các HMM phụ thuộc ngữ cảnh. Các MFCC và $F0$ được tính toán cho từng câu nói thu âm nhờ sử dụng bộ công cụ SPTK [14]. Các vectơ đặc trưng như phổ, thanh điệu và các vectơ tham số cao độ (F0) bao gồm các MFCC bậc 24 (giá trị này được cho là hiệu quả nhất với các tín hiệu lấy mẫu ở tần số 16 kHz thông qua rất nhiều thực nghiệm), các giá trị logarit của F0 (mục đích để chuyển các giá trị F0 sang một miền khác mà các giá trị tương ứng của chúng dễ biểu diễn và đồng thời các phép tính cũng được chuyển từ phép nhân sang phép cộng), các hệ số delta và delta-delta của chúng. Qua nhiều thực nghiệm có thay đổi tham số và tham khảo từ các công trình tương tự của các tác giả trên thế giới, cuối cùng chúng tôi chọn sử dụng hình trạng các HMM 5 trạng thái trái sang phải với các phân bố Gauss đơn, huấn luyện nhúng sử dụng thuật toán cực đại hóa kỳ vọng (EM – expectation maximization, là phương pháp lặp để tìm khả năng cực đại các ước lượng hậu nghiệm, MAP, cực đại) được lặp 20 lần để tạo ra các tham số tiếng nói, phạm vi tần số trích chọn tham số F0 trong khoảng từ 80-450 Hz (bao hàm cả giọng nam và giọng nữ). Các nhãn phụ thuộc ngữ cảnh của hai bộ dữ liệu tiếng nói tiếng Việt được sinh ra tự động từ các văn bản tương ứng nhờ sử dụng bộ phân tích văn bản tiếng Việt [2]. Ngoài ra, chúng tôi sử dụng kỹ thuật phân cụm ngữ cảnh dựa trên cây quyết định để huấn luyện các HMM phụ thuộc ngữ cảnh tương

ứng với từng tham số phổ, F0 và các thành phần tuần hoàn khác.

Trong phần tổng hợp và đánh giá, chúng tôi sử dụng phần dữ liệu còn lại (68 câu giọng nam và 67 câu giọng nữ) trong bộ ngữ liệu đã nói ở trên. Quá trình tổng hợp được thực hiện ở cả 4 trường hợp:

a) Tổng hợp giọng nam trên phần dữ liệu giọng nam

Các HMM phụ thuộc ngữ cảnh thu được sau quá trình huấn luyện 500 câu nói giọng nam (giọng miền Nam), sau đó các HMM này được kết hợp với 68 chuỗi văn bản cần tổng hợp đã gán nhãn (phân tích văn bản và gán nhãn được thực hiện theo [2] và [13]). Từ các mô hình này, sử dụng thuật toán sinh tham số để tạo ra các tham số vectơ phổ MFCC và các tham số nguồn kích thích (F0 hay cao độ). Cuối cùng, các tham số này được tổng hợp thành tiếng nói dưới dạng sóng thông qua một bộ lọc tổng hợp (MLSA).

b) Tổng hợp giọng nữ trên phần dữ liệu giọng nữ

Tương tự như đối với quá trình huấn luyện và tổng hợp giọng nam ở trên, 500 câu nói giọng nữ (miền Bắc) được sử dụng để huấn luyện các HMM phụ thuộc ngữ cảnh, sau đó kết hợp với chuỗi nhãn của 67 câu còn lại để sinh ra các vectơ tham số cần thiết cho quá trình tổng hợp tiếng nói.

c) Tổng hợp giọng nam trên phần dữ liệu giọng nữ Một mở rộng của phần tổng hợp và đánh giá kết quả khác là chúng tôi sử dụng các HMM của giọng nam đã huấn luyện để tổng hợp 67 câu văn bản đã gán nhãn trong phần còn lại của bộ dữ liệu giọng nữ. Kết quả này cũng được so sánh, đánh giá với kết quả của phần b và các câu nói thu âm gốc trong bộ dữ liệu.

d) Tổng hợp giọng nữ trên phần dữ liệu giọng nam

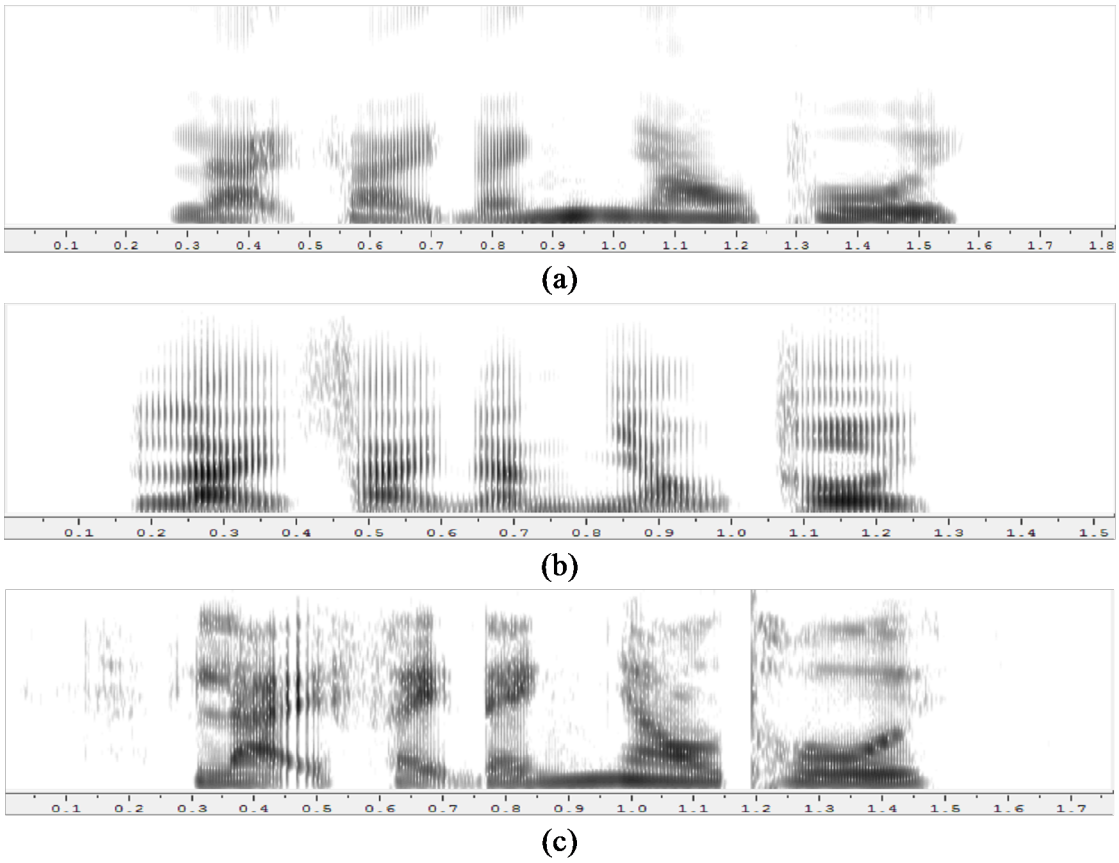
Tương tự như phần c), các HMM thu được sau khi huấn luyện giọng nữ được sử dụng để tổng hợp 68 câu văn bản đã gán nhãn còn lại của bộ dữ liệu giọng nam. Sau đó, so sánh, đánh giá kết quả này với kết quả của phần a) và dữ liệu gốc.

4. ĐÁNH GIÁ KẾT QUẢ

Trong phần này sẽ thực hiện so sánh, đánh giá khách quan về chất lượng tiếng nói tổng hợp sử dụng phương pháp thống kê trên cơ sở HMM. Đánh giá chủ quan được thực hiện thông qua phương pháp so sánh sự tương đồng giữa ảnh phổ (spectrogram) và đường bao cao độ của các kết quả tổng hợp và dữ liệu gốc.

Do quá trình sinh tham số sử dụng các giá trị trung bình của các mô hình thời gian trạng thái, nên trường độ (khoảng thời gian nghỉ giữa các âm tiết) của các câu nói tổng hợp có thể khác với trường độ trong câu nói trong dữ liệu gốc. Trong phần thử nghiệm, sử dụng chuỗi các trạng thái (thu được từ quá trình force-align) cùng với các mô hình phổ và cao độ, để sinh tham số tiếng nói. Vì thế, có thể đánh giá kết quả thử nghiệm thông qua so sánh tín hiệu tiếng nói tổng hợp và tiếng nói thu âm gốc mà không quan tâm đến đặc trưng trường độ trong câu kết quả.

Hình 5 so sánh ảnh phổ của các câu nói: (a) tổng hợp từ mô hình giọng nữ miền Bắc, (b) tổng hợp từ giọng nam miền Nam và (c) thu âm gốc của văn bản "Lại phải đánh nhau thôi" (trích trong truyện đọc "Dế mèn phiêu lưu ký của nhà văn Tô Hoài", thu âm giọng nữ). Chú ý trục thời gian, ta sẽ thấy cho sự khác nhau về trường độ của các kết quả tổng hợp và câu nói thu âm gốc.



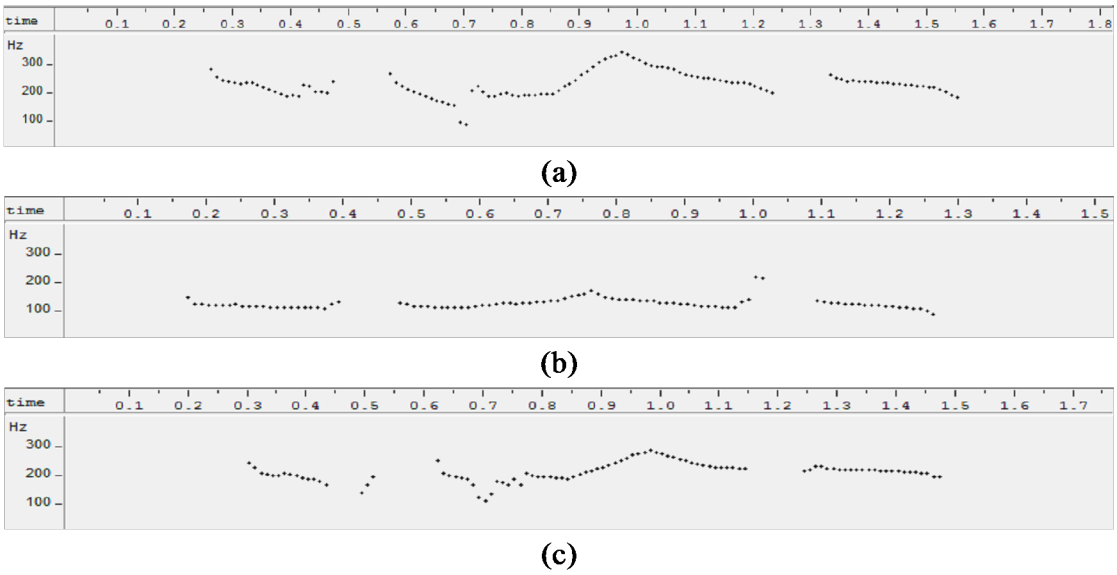
Hình 5. Ảnh phổ của các câu nói: (a) tổng hợp từ mô hình giọng nữ miền Bắc, (b) tổng hợp từ mô hình giọng nam miền Nam và (c) thu âm gốc của văn bản "Lại phải đánh nhau thôi"

Hình 6 minh họa sự tương đồng của đường bao cao độ của các câu nói: (a) tổng hợp từ mô hình giọng nữ miền Bắc, (b) tổng hợp từ mô hình giọng nam miền Nam và (c) thu âm gốc của văn bản "Lại phải đánh nhau thôi".

Trong hình 6, có thể nhận thấy có sự đồng dạng tương đối về đường bao cao độ của các kết quả và dữ liệu gốc. Chú ý trục tần số, ta sẽ thấy có sự khác nhau về tần số cơ bản (F0) giữa giọng nam và giọng nữ.

5. KẾT LUẬN

Bài báo đã đề xuất một hệ thống tổng hợp tiếng nói thống kê dựa trên HMM, phát triển cho tổng hợp tiếng Việt. Trong đó, tập trung trích chọn các tham số đặc trưng phổ, thanh điệu, thời gian trạng thái và tần số cơ bản để mô hình hóa đồng thời sử dụng HMM. Thông tin ngữ cảnh và các lựa chọn cho việc phân cụm ngữ cảnh trên cây quyết định, sử dụng để huấn luyện các HMM, được xây dựng dựa vào tập các âm có thanh điệu, kết hợp với tập các lựa chọn ngữ âm và ngữ điệu trong các cây quyết định tương ứng. Hệ thống tổng hợp tiếng nói dựa trên HMM được thử nghiệm trên hai bộ dữ liệu được huấn luyện với thời gian hơn 5 tiếng. Kết quả tiếng nói được hệ thống tổng hợp được tiến hành đánh giá sơ bộ dựa trên



Hình 6. Đường bao cao độ của các câu nói: (a) tổng hợp từ mô hình giọng nữ miền Bắc, (b) tổng hợp từ mô hình giọng nam miền Nam và (c) thu âm gốc của văn bản "Lại phải đánh nhau thôi"

đánh giá và cảm nhận của người nghe, và mang tính chất chủ quan, dựa trên việc so sánh các ảnh phổ và đường bao cao độ (thực chất là F0). Kết quả đánh giá cho thấy rằng hệ thống đề xuất ở đây có thể tổng hợp tiếng nói tiếng Việt với chất lượng khá gần với tiếng nói tự nhiên.

Tóm lại, với hệ thống này, có thể tổng hợp được tiếng nói với các đặc điểm giọng nói khác nhau, ví dụ như cảm xúc, trọng âm, bằng phương pháp thích nghi người nói hoặc kỹ thuật nội suy người nói. Trong tương lai, việc tập trung nghiên cứu, áp dụng các yếu tố ngữ cảnh và điều kiện phân cụm ngữ cảnh, cải tiến quá trình xử lý văn bản và đánh giá tiếng nói tổng hợp để đạt được mục tiêu chất lượng tiếng nói tổng hợp tốt hơn và tổng hợp tiếng nói với các đặc tính âm học khác nhau.

TÀI LIỆU THAM KHẢO

[1] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis, *Speech Communication* **51** (11) (2009) 1039–1064.

[2] Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura, An HMM-based Vietnamese speech synthesis system, *Proc. Oriental COCOSDA,* Urumqi, China, 2009.

[3] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, *Proc. of ICASSP,* Phoenix, Arizona, USA, 1999.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, Duration modeling in HMM-based speech synthesis system, *Proc. of ICSLP,* tập 2, Sydney, Australia, 1998 (29—32).

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, *Proc.ICASSP 2000*, Orlando, Florida, USA, June 2000 (1315—1318).

[6] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems", Doctoral Dissertation, Nagoya Institute of Technology, January 2002.

[7] K. Tokuda, H. Zen, and A. Black, An HMM-based speech synthesis system applied to English, *IEEE Speech Synthesis Workshop*, Santa Monica, USA, 2002.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, Multi-space probability distribution HMM, *IEICE* **85-d** (3) (2002).

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, An adaptive algorithm for Mel-cepstral analysis of speech, *Proc. of ICASSP*, tập 1, San Francisco, California, 1992 (137—140).

[10] S. Imai, Cepstral analysis synthesis on the mel frequency scale, *Proc. of ICASSP*, Boston Massachusetts, 1983 (93—96).

[11] T.T Vu, T.K. Nguyen, H.S. Le, C.M. Luong, Vietnamese tone recognition based on MLP neural network, *Proc. Oriental COCOSDA,* Kyoto, Japan, 2008.

[12] H. Mixdorff, H. B. Nguyen, H. Fujisaki, C. M. Luong, Quantitative analysis and synthesis of syllabic tones in Vietnamese, *Proc. EUROSPEECH*, Geneva, 2003 (177-180).

[13] Phan Thanh Son, Vu Tat Thang, HMM-based Speech Synthesis for Vietnamese language, *Kỷ yếu Hội nghị Khoa học kỷ niệm 45 năm thành lập trường Đại học Điện lực*, Hà Nội, 10-2011.

[14] Department of Computer Science, Nagoya Institute of Technology, "Speech Signal Processing Toolkit, SPTK 3.0. Reference manual", http://ktlab.ics.nitech.ac.jp/t̃okuda/SPTK/, Japan, 12-2003. [cập nhật 28-4-2011].

Ngày nhận bài 17 - 8 - 2012
Ngày lại sau sửa ngày 13 - 3 - 2013