Một kỹ thuật biến đổi giọng người nói hiệu quả sử dụng kỹ thuật phân rã tiếng nói theo thời gian

An Efficient Approach for Voice Transformation using Temporal Decomposition

Phùng Trung Nghĩa

Abstract: *Voice transformation is an important issue in speech synthesis when we need to synthesize multiple output voices but do not want to rebuid the synthesis system. Speech transformed by the conventional method using Gaussian Mixture Model (GMM) is not high-quality due to the oversmoothness of GMM. Therefore, a number of methods have been proposed to overcome the disadvantages of the conventional method using GMM. Among them, Hidden Markov Model Trajectory Tiling (HTT) and Temporal Decomposition – GMM (TD-GMM) improve the effectiveness of voice transformation. However, they still have drawbacks. In this paper, a voice transformation method using the modified restricted TD (MRTD) is proposed. The experimental results with Vietnamese and English corpus confirm the effectiveness of the proposed method compared with HTT and TD-GMM.*

Keyword: *Voice transformation, voice conversion, speech synthesis, temporal decomposition.*

I. GIỚI THIỆU

Hầu hết các hệ thống xử lý tiếng nói truyền thống tập trung vào xử lý các thông tin ngôn ngữ để đảm bảo tiếng nói sau xử lý có thể hiểu được [1]. Tuy nhiên để các ứng dụng xử lý tiếng nói trong máy tính có thể được áp dụng rộng rãi trong thực tế, tính tự nhiên của tiếng nói được xử lý cũng cần được quan tâm [2]. Để đảm bảo tiếng nói sau xử lý (như tiếng nói được tổng hợp) được tự nhiên, một trong những vấn đề quan trọng cần đảm bảo là thông tin về người nói, bao gồm cả các thông tin chung về người nói như giới tính, độ tuổi,…, đến các thông tin chi tiết như thông tin nhận danh chính xác người nói [3-7]. Các hệ thống tổng hợp tiếng nói nhân tạo thường chỉ có thể tổng hợp ra tiếng nói của một số giọng nói đã được thu sẵn và huấn luyện trước cho máy tính. Để có thể tổng hợp ra nhiều giọng nói đầu ra mà không cần xây dựng lại hệ thống tổng hợp tiếng nói cần đến các hệ thống biến đổi giọng người nói [3-6].

Trên thế giới đã có nhiều nghiên cứu về biến đổi giọng người nói trong tiếng nói [3-6]. Phương pháp truyền thống là phương pháp sử dụng học máy thống kê dùng mô hình Gaussian hỗn hơn GMM [3]. Do chất lượng tiếng nói tổng hợp / tái tạo bằng các mô hình thống kê như GMM có xu hướng bị trung bình hóa, quá trơn và chất lượng không cao, nhiều nghiên cứu đã đề xuất các phương pháp biến đổi giọng người nói khác khắc phục các nhược điểm của phương pháp GMM truyền thống. Trong số đó hai phương pháp có kết quả nổi bật là phương pháp lai giữa GMM và kỹ thuật phân rã tiếng nói theo thời gian TD có tên gọi TD-GMM [4], và phương pháp ghép nối / thay thế khung có tên gọi HTT [5].

Nghiên cứu này đề xuất phương pháp biến đổi giọng người nói trong tiếng nói lai giữa hai phương pháp TD-GMM [4] và phương pháp thay thế khung HTT [5], sử dụng kỹ thuật phân rã tiếng nói theo thời gian cải tiến MRTD [8]. Phương pháp đề xuất cũng như hai phương pháp TD-GMM và HTT được cài đặt và đánh giá thực nghiệm với cơ sở dữ liệu tiếng nói tiếng Anh và tiếng Việt.

II. PHƯƠNG PHÁP BIẾN ĐỔI TD-GMM

Phương pháp biến đổi giọng người nói kinh điển là phương pháp sử dụng mô hình GMM để huấn luyện cặp người nói nguồn – đích với tập dữ liệu huấn luyện song song kích cỡ nhỏ, sau đó sử dụng hàm biến đổi đã được huấn luyện để biến đổi tiếng nói giọng nguồn thành tiếng nói giọng đích [3].

Mặc dù phương pháp GMM đã chứng tỏ được hiệu quả trong nhiều nghiên cứu, đặc biệt có ưu điểm chỉ sử dụng một lượng nhỏ dữ liệu huấn luyện, nó vẫn có nhiều hạn chế. Do cấu trúc phổ được ước lượng bởi mô hình GMM ứng với phổ trung bình của tất cả dữ liệu trong tập dữ liệu huấn luyện (do mô hình GMM sử dụng vector kỳ vọng trung bình làm cơ sở), nên tiếng nói được biến đổi bằng mô hình GMM thường quá trung bình, hay quá trơn (over-smooth). Việc tiếng nói bị biến đổi quá trơn sẽ làm những đặc trưng chi tiết của tiếng nói vốn mang nhiều thông tin người nói sẽ bị mất đi trong quá trình biến đổi.

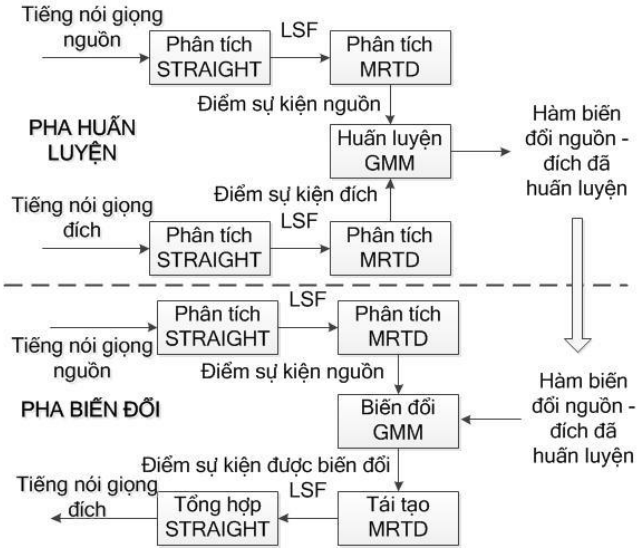
Trong [4] đã sử dụng kỹ thuật phân rã tiếng nói theo thời gian TD kết hợp với mô hình GMM dựa trên dữ liệu đã gán nhãn ở mức âm vị trong phương pháp tên gọi TD-GMM để khắc phục hạn chế biến đổi tiếng nói quá trơn và bị mất thông tin người nói của phương pháp biến đổi giọng người nói bằng GMM.

TD được sử dụng để phân tích tiếng nói thành hai thành phần độc lập, thành phần *"động"*- hàm sự kiện (event functions) để đảm bảo cho tiếng nói có độ trơn cần thiết còn thành phần *"tĩnh"*- điểm sự kiện (event targets) giúp tiếng nói vẫn giữ được thông tin chi tiết để tiếng nói tái tạo từ hai thành phần này có mức độ trơn phù hợp, không bị quá trơn [4].

Một số nghiên cứu cũng đã chỉ ra rằng, hàm sự kiện TD mang các thông tin ngôn ngữ vốn quan trọng để hiểu tiếng nói, còn các điểm sự kiện mang thông tin phi ngôn ngữ như thông tin người nói hay cảm xúc nói [4, 8].

Do vậy, trong phương pháp TD-GMM, chỉ thành phần điểm sự kiện được huấn luyện và biến đổi như trong Hình 1, trong khi thành phần hàm sự kiện được giữ nguyên, khác với việc biến đổi tất cả các khung như trong phương pháp biến đổi GMM truyền thống với mong muốn biến đổi được các giọng người nói một cách hiệu quả trong khi tiếng nói được biến đổi vẫn có độ trơn phù hợp. Các kết quả thực nghiệm cho thấy TD-GMM cho kết quả tốt hơn phương pháp GMM truyền thống về mặt chất lượng tiếng nói biến đổi [4].

Mặc dù cho kết quả tốt hơn mô hình biến đổi GMM truyền thống, việc vẫn sử dụng mô hình GMM để huấn luyện và biến đổi dẫn tới tiếng nói biến đổi bằng TD-GMM vẫn có xu hướng hơi quá trơn so với tiếng nói tự nhiên, dẫn tới chất lượng tiếng nói được biến đổi chưa cao so với tiếng nói tự nhiên [4].



Hình 1. Phương pháp biến đổi TD-GMM [4].

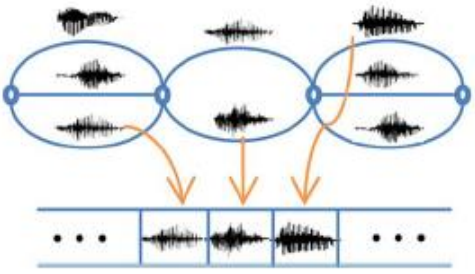
III. PHƯƠNG PHÁP BIẾN ĐỔI GIỌNG NGƯỜI NÓI DỰA VÀO THAY THẾ KHUNG

Để khắc phục yếu điểm biến đổi tiếng nói quá trơn (quá trung bình) trong các phương pháp sử dụng mô hình GMM, bao gồm cả phương pháp GMM kinh điển [3] và phương pháp TD-GMM [4], một số phương pháp đã được đề xuất. Nổi bật nhất trong số đó là phương pháp biến đổi giọng người nói lai giữa tổng hợp tiếng nói dùng mô hình Markov ẩn (HMM) và thay thế mẫu / ghép nối HTT được tác giả Yao Qian và cộng sự đề xuất năm 2013 [5].

Trong phương pháp HTT, ở bước thứ nhất tiếng nói tổng hợp bằng mô hình HMM với giọng nguồn. Tiếp theo ở bước thứ hai, tiếng nói đã tổng hợp được biến đổi thành tiếng nói giọng đích dựa trên kỹ thuật lựa chọn và thay thế các khung nguồn có độ dài rất ngắn 5ms bằng các khung đích phù hợp như mô tả trong Hình 2.

Nếu bỏ qua vấn đề tổng hợp giọng nguồn bằng HMM, bản chất của phương pháp biến đổi giọng người nói HTT là các khung của tiếng nói giọng nguồn được thay thế bằng các khung vật lý giống nhất của giọng đích trong cùng âm vị. Mặc dù việc lựa chọn và thay thế mẫu tiếng nói giọng nguồn bằng mẫu tiếng nói giọng đích đã được đề xuất trước đó [7], hiệu quả biến đổi giọng người nói trong HTT là vượt trội so với các phương pháp thay thế mẫu khác do việc sử dụng các khung tiếng nói rất ngắn thay thế các mẫu tiếng nói dài như âm vị [7] sẽ tối ưu việc tìm được khung/mẫu tiếng nói đích phù hợp nhất.

Các kết quả thực nghiệm cho thấy phương pháp thay thế khung HTT cho chất lượng và hiệu quả biến đổi giọng người nói rất cao [5]. HTT đã được thực nghiệm trên tiếng Anh, tiếng Trung và đã đạt thứ hạng cao trong cuộc thi về tổng hợp tiếng nói và chuyển đổi giọng nói quốc tế Blizzard Challenge 2013 [5]. Tuy nhiên các phương pháp lựa chọn / thay thế khung như HTT kế thừa tất cả các nhược điểm của tổng hợp ghép nối như đòi hỏi dữ liệu lớn, tốc độ thực thi khó đảm bảo thời gian thực, dữ liệu cần lưu trữ online lớn.



Hình 2. Lựa chọn khung đích phù hợp và thay thế khung nguồn [5]

IV. PHƯƠNG PHÁP BIẾN ĐỔI GIỌNG NGƯỜI NÓI SỬ DỤNG KỸ THUẬT TD ĐỀ XUẤT

IV.1. Đặt vấn đề

Do cả hai phương pháp biến đổi giọng người nói TD-GMM và HTT đều có ưu và nhược điểm, nghiên cứu này đề xuất phương pháp tận dụng các ưu điểm và hạn chế các yếu điểm của cả hai.

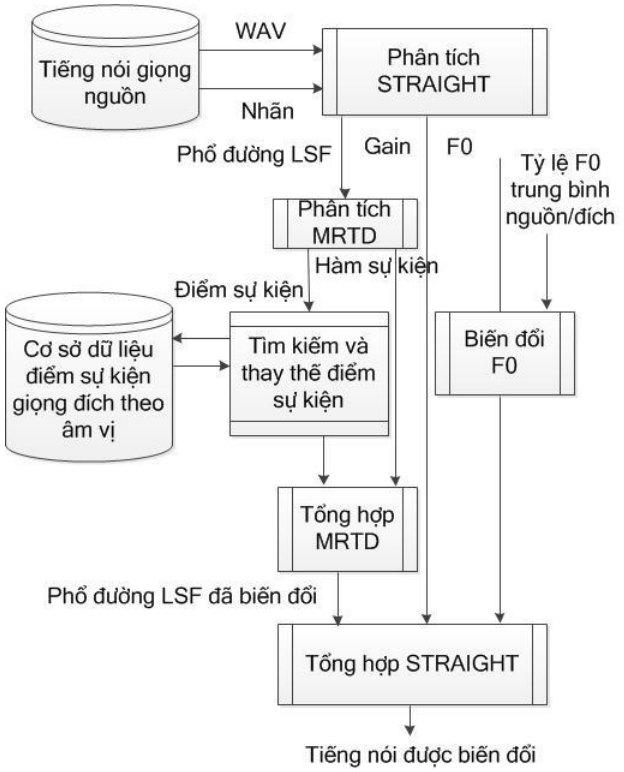
Điểm mạnh của phương pháp TD-GMM là kỹ thuật TD cho phép biến đổi thông tin người nói hiệu quả với việc dùng biến đổi điểm sự kiện thay thế cho biến đổi các khung tiếng nói. Trong khi điểm yếu của phương pháp này là việc mô hình hóa bằng GMM vẫn khiến tiếng nói được biến đổi có xu hướng quá trơn.

Điểm mạnh của phương pháp HTT là chất lượng cao do quá trình lựa chọn và thay thế trực tiếp mẫu tiếng nói đích bằng mẫu tiếng nói nguồn theo khoảng cách vật lý gần nhất. Trong khi điểm yếu của phương pháp này là việc tìm kiếm và thay thế tất cả các khung tiếng nói ngắn đòi hỏi dữ liệu đích để tìm kiếm lớn, tốc độ thực thi khó đảm bảo thời gian thực, dữ liệu đích cần lưu trữ online cũng lớn.

Do vậy, ý tưởng kết hợp của phương pháp đề xuất trong nghiên cứu này là sử dụng kỹ thuật TD để phân rã tiếng nói thành các hàm sự kiện và điểm sự kiện. Hàm sự kiện sẽ được giữ nguyên như trong TD-GMM. Việc huấn huyện và biến đổi điểm sự kiện giọng nguồn thành điểm sự kiện giọng đích sử dụng học máy thống kê GMM sẽ được thay bằng việc tìm kiếm và lựa chọn, thay thế trực tiếp điểm sự kiện giọng nguồn bằng điểm sự kiện giọng đích gần nhất về mặt vật lý (giống nhất). Quá trình lựa chọn và thay thế điểm sự kiện trong phương pháp đề xuất sẽ tương tự quá trình lựa chọn và thay thế khung trong phương pháp HTT. Tuy nhiên việc lựa chọn thay thế điểm sự kiện thưa thay vì tất cả các khung ngắn như trong HTT sẽ khắc phục được yếu điểm của HTT về không gian tìm kiếm lớn, thời gian thay thế và ghép nối lâu.

IV.2. Mô hình phương pháp đề xuất

Mô hình tổng thể của phương pháp đề xuất được thể hiện trên Hình 3.



Hình 3. Mô hình biến đổi giọng người nói đề xuất

Tiếng nói giọng nguồn được phân tích thành các đặc trưng như tần số cơ bản (F0), hệ số độ lợi ứng với năng lượng tiếng nói, và phổ đường (LSF) sử dụng bộ phân tích / tái tạo tiếng nói chất lượng cao STRAIGHT [9]. Đặc trưng F0 của giọng nguồn được biến đổi thành giống giọng đích mà không thay đổi tính chất thanh điệu, ngữ điệu (*thể hiện qua đường vận động F0*) bằng cách biến đổi mức F0 trung bình. Đặc trưng phổ đường LSF là đặc trưng vector nhiều chiều và cũng là đặc trưng mang thông tin người nói quan trọng nhất được phân tích bằng kỹ thuật MRTD, một kỹ thuật TD cải tiến, đơn giản hóa [8]. MRTD có nhiều ưu điểm so với kỹ thuật TD cổ điển như có độ phức tạp tính toán thấp, lỗi tái tạo nhỏ, các hàm sự kiện trơn và linh hoạt, dễ dàng biến đổi như đã chứng tỏ trong nhiều nghiên cứu trước đây [4, 8].

Giả sử vector phổ đường giọng nguồn LSF là $y(n)$, MRTD phân rã $y(n)$ thành K hàm sự kiện động ϕ_k và K điểm sự kiện tĩnh a_k với k = 1..*K*, như trong công thức (1). Ở đây $\hat{y}(n)$ là vector xấp xỉ của $y(n)$ được tái tạo từ các hàm sự kiện ϕ_k và điểm sự kiện a_k.

Có tổng số K điểm sự kiện trong tổng số N khung với $K << N$, khi đó MRTD (hay TD nói chung) là một biểu diễn thưa của tiếng nói. Các hàm sự kiện là các hàm nội suy biểu diễn sự chuyển dịch trên miền thời gian của các sự kiện thưa.

$$\hat{y}(n) = \sum_{k=1}^{K} a_k \phi_k(n), 1 \le n \le N \qquad (1)$$

Công thức (1) có thể viết lại dưới dạng ma trận như công thức (2) với P là số chiều của tham số đặc trưng tiếng nói đang phân tích (ở đây là phổ đường LSF).

$$\hat{Y}_{P\times N} = A_{P\times K}\Phi_{K\times N} \qquad (2)$$

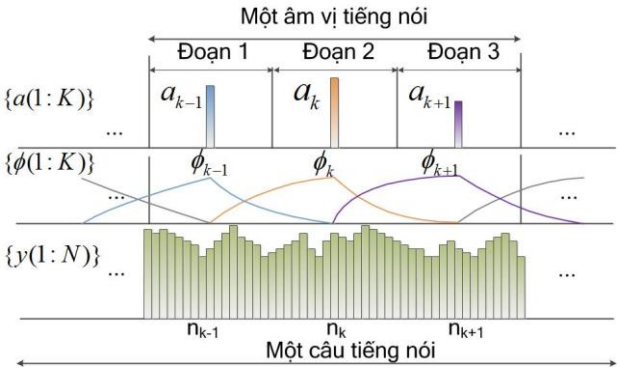
Hình 4 vẽ một ví dụ của MRTD khi phân tích vector $y(1:N)$, các điểm sự kiện $a(1:K)$, và các hàm sự kiện $\phi(1:K)$.

Điểm sự kiện *a* và hàm sự kiện ϕ là chưa biết trong công thức (1), (2) và cần được ước lượng bằng các kỹ thuật tối ưu hóa để tối thiểu lỗi tái tạo.

Trong bước đầu tiên của quá trình tối ưu trong MRTD, các điểm sự kiện được đặt bằng vector đặc trưng tại khung tiếng nói cùng vị trí như trong công thức (3).

$$a_k = y(n_k) \qquad (3)$$

Ở đây, n_k là vị trí của điểm sự kiện a_k.



Hình 4. Ví dụ phân tích / tái tạo tiếng nói bằng MRTD với N khung và K điểm sự kiện

Trong bước 2 của quá trình tối ưu, các hàm sự kiện trong MRTD được ước lượng như trong công thức (4) và (5). Ở đây $\langle \ldots \rangle$ và $\|.\|$ ứng với tích trong của 2 vector và chuẩn của 1 vector.

$$\phi_k(n)=\begin{cases} 1-\phi_{k-1}(n), \text{if } n_{k-1}<n<n_k \\ 1, \text{ if } n=n_k \\ \min(\phi_k(n-1), \max(0, \hat{\phi}_k(n))), \\ \text{if } n_k<n<n_{k-1} \\ 0, \text{ khác} \end{cases} \quad (4)$$

$$\hat{\phi}_k(n)=\frac{\langle (y(n)-a_{k+1}), (a_k-a_{k+1})\rangle}{\| a_k-a_{k+1}\|^2} \quad (5)$$

Sử dụng công thức (4) và (5), mỗi hàm sự kiện $\phi_k(n)$ đều trơn, chỉ có một đỉnh, hai hàm chồng lấp có tổng là 1 như mô tả trong Hình 4 và được giải thích tường minh tại [8]. Các tính chất này của hàm sự kiện dẫn tới sự chuyển dịch từ từ của các vector phổ $\hat{y}(n)$ phù hợp với sự biến đổi chậm tự nhiên của tiếng nói. Sự thay đổi các giá trị điểm sự kiện thưa a_k trực tiếp sẽ ảnh hưởng dần dần đến tất cả các khung tiếng nói trong khoảng mà hàm sự kiện $\phi_k \neq 0$. Do đó, tiếng nói có thể được biến đổi một cách linh hoạt quanh vị trí các điểm sự kiện cụ thể trên miền thời gian bằng cách biến đổi các điểm sự kiện MRTD a như trong [4].

Sau khi các hàm sự kiện được ước lượng, các điểm sự kiện được ước lượng lại ở bước cuối cùng của quá trình tối ưu như trong công thức (6) để tối thiểu lỗi nội suy, ở đây ^{T} là phép chuyển vị ma trận.

$$A = Y\Phi^T(\Phi\Phi^T)^{-1} \quad (6)$$

Công thức (6) có ý nghĩa là mỗi điểm sự kiện được ước lượng lại bởi chính giá trị khởi tạo của nó, là giá trị vector đặc trưng khung tiếng nói tại cùng vị trí, và các hàm sự kiện khác 0 được ước lượng ở cùng vị trí với điều kiện hội tụ tối thiểu lỗi tái tạo và đảm bảo tính chất thứ tự của phổ đường LSF.

Sau khi được phân tích bằng MRTD, các hàm sự kiện được giữ nguyên để đảm bảo tiếng nói sau khi biến đổi giữ được độ trơn cần thiết cũng như để giữ nguyên các đặc trưng ngôn ngữ không bị biến đổi. Trong khi đó các điểm sự kiện nguồn được thay thế bằng các điểm sự kiện đích gần nhất tìm thấy từ cơ sở dữ liệu giọng đích ứng với nhãn tiếng nói tương ứng.

Cuối cùng, bộ phân tích / tái tạo tiếng nói STRAIGHT được sử dụng để tổng hợp lại tiếng nói từ các đặc trưng F0, phổ đã được biến đổi.

IV.3. Thủ tục tìm kiếm và thay thế điểm sự kiện

Các điểm sự kiện được thay đổi trong phương pháp đề xuất bằng cách thay thế chúng với các điểm sự kiện giống nhất ở tiếng nói đích trong cùng một đơn vị tiếng nói như âm vị. Do vậy cần một thủ tục căn lề trên miền thời gian phù hợp. Ở đây, kỹ thuật cố định số lượng điểm sự kiện trong mỗi âm vị và đặt các điểm sự kiện cách đều nhau trong mỗi âm vị đã được đề xuất và chứng tỏ hiệu quả trong phương pháp biến đổi TD-GMM [4]. Đây là một kỹ thuật biến đổi song song với mỗi âm vị khi các điểm sự kiện theo thứ tự của âm vị nguồn được thay thế bằng các điểm sự kiện có thứ tự tương ứng ở âm vị đích. Phát triển từ kỹ thuật này, mỗi âm vị trong phương pháp đề xuất ở đây được chia thành 3 khoảng con đều nhau, mỗi điểm sự kiện được đặt ở trung tâm của mỗi khoảng con như trong Hình 4. Trong các thử nghiệm của chúng tôi khi tăng số lượng điểm sự kiện trong mỗi âm vị lớn hơn 3 không làm tăng chất lượng tiếng nói được tái tạo, nhưng lại làm tăng kích thước dữ liệu đích phải lưu trữ cho quá trình tìm kiếm / thay thế. Trong khi nếu số lượng điểm sự kiện nhỏ hơn 3 sẽ làm giảm chất lượng của tiếng nói được tái tạo.

Điểm sự kiện đích gần nhất với điểm sự kiện nguồn được tìm kiếm bằng thuật toán tìm láng giềng gần nhất NNS (Nearest Neighbor Search) với hàm khoảng cách *d* giữa điểm sự kiện nguồn a_s và điểm sự kiện đích a_t với vector phổ đường LSF có số chiều P được định nghĩa trong công thức (7).

$$d=\sqrt{\frac{1}{P}\sum_{i=1}^{P}(a_t^i-a_s^i)^2} \quad (7)$$

$$N(d) = \frac{d - \mu_d}{\sigma_d} \quad (8)$$

Hàm chi phí được chuẩn hóa theo công thức (8) bằng phân bố chuẩn với μ_d, σ_d là giá trị kỳ vọng trung bình và độ lệch chuẩn của các khoảng cách của các mẫu.

Trong phần cài đặt, quá trình lựa chọn điểm sự kiện đích để thay thế được giám sát bằng nhãn dữ liệu tiếng nói trong từng âm vị để đảm bảo độ chính xác và giảm thời gian tìm kiếm, trong đó mỗi điểm sự kiện với thứ tự xác định trong một âm vị được thay thế bằng điểm sự kiện đích có cùng thứ tự trong cùng âm vị của giọng đích.

Trong pha offline, cơ sở dữ liệu tiếng nói với giọng đích được chuẩn bị trước với hai bước. Trong bước thứ nhất, tất cả các câu tiếng nói đã gán nhãn mức âm vị được phân tích bằng MRTD. Trong bước thứ hai, các điểm sự kiện của các câu tiếng nói đã phân tích được trích xuất và lưu trữ theo từng âm vị riêng để tăng tốc độ tìm kiếm trong pha online.

V. ĐÁNH GIÁ VÀ THẢO LUẬN

V.1. Tiêu chí đánh giá

V.1.1. Đánh giá khách quan

Phương pháp đánh giá khách quan được sử dụng phổ biến trong các hệ thống biến đổi giọng người nói là phương pháp chỉ số hiệu năng PI (Performance Index) [4]. PI với tham số phổ đường LSF được tính bằng công thức (9).

$$PI_{LSF} = 1 - \frac{E_{LSF}(t(n), \hat{t}(n))}{E_{LSF}(t(n), s(n))} \quad (9)$$

Trong đó, *t(n)* biểu diễn mẫu tiếng nói giọng đích, *s(n)* biểu diễn mẫu tiếng nói giọng nguồn, $\hat{t}(n)$ biểu diễn mẫu tiếng nói được chuyển đổi từ nguồn thành đích. E_{LSF} là sai số LSF trung bình được tính bằng công thức (10).

$$E_{LSF}(A,B) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{P}\sum_{i=1}^{P}(LSF_A^{l,i} - LSF_B^{l,i})^2} \quad (10)$$

Với *L* là tổng số khung tiếng nói (sau khi đã căn thời gian để tổng số khung trùng khớp), *P* là số hệ số LSF.

PI_{LSF} = 0 chỉ ra rằng hệ thống chuyển đổi không giống hệ thống đích chút nào còn PI_{LSF} = 1 chỉ ra rằng hệ thống chuyển đổi hoàn toàn giống hệ thống đích.

V.1.2. Đánh giá chủ quan

Trong các phương pháp đánh giá chủ quan, phương pháp được áp dụng rộng rãi trong các hệ thống biến đổi giọng nói là phương pháp ABX [4]. Trong đó A là tiếng nói với giọng người nói nguồn, B là tiếng nói với giọng người nói đích, X là tiếng nói với giọng chuyển đổi từ A thành B. Người nghe sẽ được nghe thử tiếng nói với giọng nguồn A và giọng đích B trước. Sau đó khi đánh giá sẽ nghe các mẫu đã biến đổi giọng X xem giống A hay giống B theo thang điểm trung bình MOS (Mean Opinion Score) từ 1 đến 5. Điểm là 1 tức là giọng biến đổi rất giống giọng nguồn A, điểm là 5 tức là giọng biến đổi rất giống giọng đích B.

V.2. Cơ sở dữ liệu đánh giá

Với tiếng Việt, chưa có cơ sở dữ liệu nhiều người nói với kịch bản giống nhau được gán nhãn. Do vậy, chúng tôi đã sử dụng bộ cơ sở dữ liệu DEMEN567 (còn gọi là cơ sở dữ liệu VNSpeech) có kích cỡ trung bình gồm 567 câu, người nữ nói, làm cơ sở dữ liệu giọng đích [10]. DEMEN567 được gán nhãn ở mức âm vị và bao phủ gần như 100% các âm vị tiếng Việt. Cơ sở dữ liệu giọng nguồn được chúng tôi tổng hợp nhân tạo bằng phương pháp HMM [11] với kịch bản nói giống như DEMEN567 sử dụng dữ liệu huấn luyện là cơ sở dữ liệu VOV [12], người nữ nói, kết hợp trích xuất nhãn ở mức âm vị tự động.

Với tiếng Anh, chúng tôi sử dụng 460 câu trong bộ cơ sở dữ liệu MOCHA-TIMIT [13] gồm nhiều người nói với các kịch bản giống nhau và chọn một người nói nữ nguồn và một người nói nữ đích. MOCHA-TIMIT chưa phải là cơ sở dữ liệu lớn như cơ sở dữ liệu sử dụng với HTT trong [5], đây là bộ cơ sở dữ liệu có kích cỡ trung bình, được gán nhãn ở mức âm vị

và bao phủ gần như toàn bộ các âm tiết tiếng Anh [13].

Do các phương pháp TD-GMM, HTT và phương pháp đề xuất đều tập trung vào biến đổi đặc trưng phổ thay vì đặc trưng F0, chúng tôi chọn lựa trước giọng nguồn và giọng đích có mức cao độ trung bình tương đương để dễ dàng phân biệt sự thay đổi về đặc trưng phổ trong quá trình biến đổi.

V.3. Thực nghiệm các phương pháp

Phương pháp đề xuất được thực nghiệm và so sánh với phương pháp HTT và TD-GMM. Các tham số thực nghiệm sử dụng trong các phương pháp được cho trong Bảng 1.

Bảng 1. Các tham số thực nghiệm

Tần số lấy mẫu DEMEN và VOV-HMM được lấy mẫu lại	11025 Hz
Tần số lấy mẫu MOCHA-TIMIT	16000 Hz
Chiều dài khung	5 ms
Độ dịch khung	1 ms
Số chiều LSF	20
Số thành phần GMM	20
Số điểm sự kiện / âm vị	3

Khi thực nghiệm cả ba phương pháp với cơ sở dữ liệu tiếng Việt (DEMEN/VOV-HMM) và tiếng Anh (MOCHA-TIMIT), 400/567 cặp câu tiếng Việt và 400/460 cặp câu tiếng Anh được sử dụng để huấn luyện (với TD-GMM) và tìm kiếm / thay thế (với HTT và phương pháp đề xuất). 30 cặp câu không có trong tập dữ liệu huấn luyện và tập dữ liệu để tìm kiếm / thay thế được sử dụng để đánh giá. Phân tích mức độ bao phủ về mặt âm vị giữa các câu trong tập huấn luyện và các câu trong tập đánh giá cho thấy 100% các âm vị trong tập đánh giá (30 câu) nằm trong tập âm vị của tập dữ liệu huấn luyện cũng như tập dữ liệu tìm kiếm / thay thế (400 câu tiếng Việt, 400 câu tiếng Anh).

Phương pháp đánh giá khách quan PI được tính tự động theo công thức (9). Phương pháp đánh giá chủ quan được thực hiện với 05 người đánh giá người Việt là các sinh viên độ tuổi 18 đến 20, có khả năng nghe bình thường. Do mục đích của phần đánh giá chủ quan ABX là đánh giá giọng nói X giống với người nguồn A hay người đích B là vấn đề độc lập ngôn ngữ, không cần người đánh giá phải hiểu được ngữ nghĩa của các mẫu tiếng nói đánh giá. Chính vì vậy, 05 sinh viên người Việt được lựa chọn để thực hiện đánh giá ABX với cả phần dữ liệu tiếng Việt và tiếng Anh. Điểm MOS đánh giá là điểm ABX trung bình của tất cả các mẫu đánh giá.

V.4. Kết quả đánh giá

Bảng 2. Kết quả đánh giá khách quan với tiếng Anh

Phương pháp	PI_{LSF}
Thay thế khung HTT	0.714
TD-GMM	0.525
Phương pháp đề xuất	0.706

Bảng 3. Kết quả đánh giá khách quan với tiếng Việt

Phương pháp	PI_{LSF}
Thay thế khung HTT	0.663
TD-GMM	0.468
Phương pháp đề xuất	0.612

Bảng 4. Kết quả đánh giá chủ quan ABX với tiếng Anh

Phương pháp	MOS
Thay thế khung HTT	4.0
TD-GMM	3.2
Phương pháp đề xuất	4.0

Bảng 5. Kết quả đánh giá chủ quan ABX với tiếng Việt

Phương pháp	MOS
Thay thế khung HTT	3.8
TD-GMM	3.2
Phương pháp đề xuất	3.8

Kết quả đánh giá trong các Bảng 2, 3, 4, 5 cho thấy hiệu quả biến đổi giọng người nói của phương pháp đề xuất cao hơn phương pháp TD-GMM và gần như tương đương với HTT (đặc biệt với đánh giá chủ quan) với các cơ sở dữ liệu kích cỡ trung bình tiếng Anh và tiếng Việt được thử nghiệm.

V.5. Thảo luận

Phương pháp biến đổi giọng người nói đề xuất đã cố gắng tận dụng ưu điểm của 2 phương pháp HTT và TD-GMM.

So với TD-GMM, phương pháp đề xuất có chất lượng tiếng nói chuyển đổi cao hơn hẳn đối với các cơ sở dữ liệu vừa phải được lựa chọn để đánh giá thực nghiệm do thay thế phương pháp huấn luyện / biến đổi thống kê với GMM bằng phương pháp thay thế vật lý trực tiếp. Cả TD-GMM và phương pháp đề xuất đều sử dụng cơ sở dữ liệu tiếng nói đích đã gán nhãn ở mức âm vị và yêu cầu cơ sở dữ liệu đích bao phủ hết các âm vị.

So với HTT, mặc dù chỉ tương đương về hiệu quả chuyển đổi giọng nói, phương pháp đề xuất đã thể hiện 03 ưu điểm nổi bật sau.

Thứ nhất, HTT yêu cầu một bộ dữ liệu đích phải rất lớn mới đảm bảo độ trơn của tiếng nói sau khi thay thế và ghép nối. Trong khi đó, độ trơn của tiếng nói sau thay thế trong phương pháp đề xuất được đảm bảo do các hàm sự kiện nguồn vốn đã trơn được giữ nguyên, không thay đổi trong quá trình thay thế. Do đó, yêu cầu về độ lớn bộ dữ liệu đích với phương pháp đề xuất nhỏ hơn HTT.

Thứ hai, do chỉ yêu cầu cơ sở dữ liệu người nói đích vừa phải và các điểm sự kiện là một vector thưa với độ dài ngắn hơn rất nhiều so với vector khung tiếng nói ($K<<N$ như mô tả trong phần 4.2), nên kích thước của dữ liệu đích phải lưu trữ trong phương pháp đề xuất là nhỏ hơn rất nhiều so với HTT.

Thứ ba, thời gian tìm kiếm các khung ngắn 5ms trong toàn bộ cơ sở dữ liệu đích lớn trong HTT là rất lớn so với thời gian tìm kiếm các điểm sự kiện với số lượng ít hơn trong một cơ sở dữ liệu đích nhỏ hơn trong phương pháp đề xuất.

Nói tóm lại, trong điều kiện cơ sở dữ liệu người đích có gán nhãn ở mức âm vị, phương pháp đề xuất đã chứng tỏ sự hiệu quả so với hai phương pháp TD-GMM và HTT nếu xét tổng hợp trên nhiều phương diện: hiệu quả chuyển đổi, mức độ yêu cầu về dữ liệu đích, kích cỡ dữ liệu lưu trữ online, thời gian tìm kiếm mẫu. Điểm yếu của phương pháp đề xuất cũng như cả TD-GMM và HTT nói chung là khi chỉ có cơ sở dữ liệu đích nhỏ thì không sử dụng được. Trong trường hợp này, phương pháp GMM kinh điển [3] vẫn sẽ là một lựa chọn chấp nhận được. Khi có cơ sở dữ liệu đích rất lớn như trong [5], mặc dù nghiên cứu này chưa có điều kiện thực nghiệm, có thể khẳng định HTT sẽ cho chất lượng chuyển đổi giọng nói vượt trội hơn phương pháp đề xuất do việc sử dụng kỹ thuật TD luôn đi kèm với lỗi nội suy và lỗi tái tạo trong khi HTT sẽ luôn lựa chọn được những khung thay thế hoàn hảo để ghép nối trực tiếp với dữ liệu đích lớn mà không cần sử dụng bộ tổng hợp/tái tạo tiếng nói nào. Tuy nhiên yêu cầu có bộ cơ sở dữ liệu đích lớn như trong [5] về cơ bản là không khả thi trong thực tế.

VI. KẾT LUẬN

Để đảm bảo tiếng nói sau xử lý (như tiếng nói được tổng hợp) được tự nhiên, một trong những vấn đề quan trọng cần đảm bảo là thông tin về người nói. Trong bài báo này, chúng tôi đề xuất một phương pháp biến đổi giọng người nói dùng kỹ thuật phân rã tiếng nói theo thời gian cải tiến MRTD. Các phân tích lý thuyết và các kết quả đánh giá thực nghiệm trên cả tiếng Anh và tiếng Việt cho thấy phương pháp đề xuất có hiệu quả hơn hai phương pháp TD-GMM và HTT, là hai phương pháp được nhiều nhà nghiên cứu trên thế giới sử dụng, trong điều kiện bộ cơ sở dữ liệu người nói đích có kích cỡ trung bình được gán nhãn ở mức âm vị và bao phủ tất cả các âm vị tiếng nói, xét trên tập các tiêu chí chất lượng biến đổi giọng, thời gian tìm kiếm, và kích cỡ dữ liệu đích phải lưu trữ online.

Do điều kiện thực tế không có các bộ cơ sở dữ liệu nhiều người nói cùng kịch bản nói lớn, các kết quả đánh giá thực nghiệm trong nghiên cứu này mới chỉ dừng lại với hai bộ cơ sở dữ liệu trung bình vừa đủ bao phủ tập các âm vị tiếng Anh và tiếng Việt. Trong các nghiên cứu tiếp theo, chúng tôi cũng sẽ tiếp tục so sánh thực nghiệm phương pháp đề xuất với một số phương pháp chuyển đổi giọng nói khác. Khi có các bộ cơ sở dữ liệu lớn hơn để thực nghiệm, chúng tôi sẽ

đánh giá với tập dữ liệu đánh giá lớn hơn, chia cặp dữ liệu huấn luyện / đánh giá theo từng mức dựa trên phân tích chi tiết về mật độ âm vị giữa các mức để đảm bảo kết quả đánh giá thực nghiệm được tin cậy và khách quan hơn.

TÀI LIỆU THAM KHẢO

[1] JURAFSKY. DANIEL, JAMES H. MARTIN. *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics and Speech Recognition, 1st Edition, 577-583, 2000.

[2] AKAGI MASATO, "*Analysis of Production and Perception Characteristics of Non-linguistic Information in Speech and Its Application to Inter-language Communications*", Proceedings APSIPA ASC 2009.

[3] KAIN ALEXANDER, MICHAEL W. MACON, "*Spectral voice conversion for text-to-speech synthesis*", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.

[4] PHU NGUYEN BINH, MASATO AKAGI, "*Phoneme-based spectral voice conversion using temporal decomposition and Gaussian mixture model*", Second IEEE International Conference Communications and Electronics, ICCE 2008, 2008.

[5] QIAN YAO, FRANK K. SOONG, ZHI-JIE YAN, "*A unified trajectory tiling approach to high quality speech rendering*", IEEE Transactions on Audio, Speech, and Language Processing, 21.2, 280-290, 2013.

[6] FUJII KEI, JUN OKAWA, KAORI SUIGETSU, "*High individuality voice conversion based on concatenative speech synthesis*", World Academy of Science, Engineering and Technology, 2.1, 2007.

[7] NGHIA PHUNG TRUNG, et al., "*A robust wavelet-based text-independent speaker identification*", International Conference on Conference on Computational Intelligence and Multimedia Applications, Vol. 2, 2007.

[8] NGUYEN PHU CHIEN, OCHI TAKAO, AND MASATO AKAGI, "*Modified restricted temporal decomposition and its application to low rate speech coding*", IEICE Transactions on Information and Systems 86.3, 397-405, 2003.

[9] KAWAHARA HIDEKI, "*STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds*", Acoustical science and technology 27.6 , 349-353, 2006.

[10] L.C. MAI, D.N. DUC, "*Design of Vietnamese speech corpus and current status*", Proc. ISCSLP-06, pp. 748-758, 2006.

[11] TT. VU, MC. LUONG, S. NAKAMURA, "*An HMM-based Vietnamese speech synthesis system, Speech Database and Assessments*", Proc. COCOSDA-2009, pp. 116-121, 2009.

[12] BẠCH HƯNG KHANG, *Báo cáo tổng kết khoa học và kỹ thuật đề tài nghiên cứu phát triển công nghệ nhận dạng, tổng hợp và xử lý ngôn ngữ tiếng Việt KC01-03*, trang 26, 2004.

[13] A. WRENCH, "The MOCHA-TIMIT articulatory database," Queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.

Nhận bài ngày: 03/10/2015

SƠ LƯỢC VỀ TÁC GIẢ

PHÙNG TRUNG NGHĨA

Sinh năm 1980.

Tốt nghiệp Trường ĐH Bách Khoa Hà Nội năm 2002. Nhận bằng thạc sĩ năm 2007 tại ĐH Quốc Gia Hà Nội. Nhận bằng tiến sĩ năm 2013 tại Viện KHCN tiên tiến Nhật Bản (JAIST).

Hiện công tác tại Trường ĐH CNTT và Truyền thông, Đại học Thái Nguyên.

Lĩnh vực nghiên cứu bao gồm Xử lý tín hiệu (âm thanh, tiếng nói, y sinh), Học máy trong xử lý tín hiệu.

Email: ptnghia@ictu.edu.vn