(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **2-0002326**

(51)[7] **G10L 15/00**

(13) **Y**

(21) 2-2010-00201

(22) 23/09/2010

(45) 25/06/2020 387

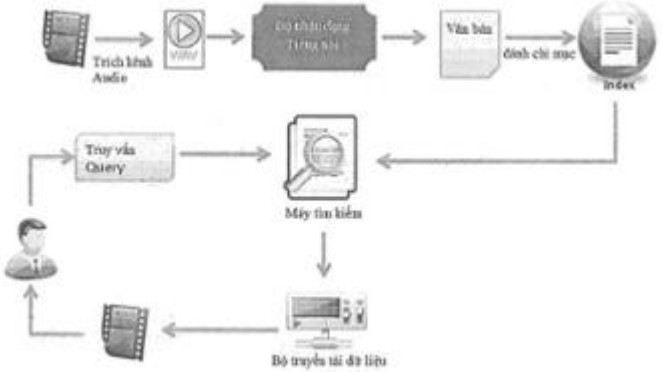
(43)

(73) Đại học Quốc Gia thành phố Hồ Chí Minh (VN)
Khu phố 6, phường Linh Trung, quận Thủ Đức, thành phố Hồ Chí Minh

(72) Vũ Hải Quân (VN); Dương Anh Đức (VN).

(54) HỆ THỐNG TRUY VẤN VIĐEO HƯỚNG NGỮ NGHĨA DỰA TRÊN CÔNG NGHỆ NHẬN DẠNG TIẾNG NÓI

(57) Giải pháp hữu ích đồ xuất hệ thống truy vấn viđeo theo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói, giải quyết ba vấn đề chính: dữ liệu nhiễu, sự xuất hiện của các từ nước ngoài trong lời thoại, các trạng thái cảm xúc khác nhau trong tiếng nói ảnh hưởng rất lớn đến độ chính xác của bộ nhận dạng tiếng nói, dẫn đến hiệu suất truy vấn cũng không cao. Hệ thống truy vấn viđeo theo hướng ngữ nghĩa theo giải pháp hữu ích gồm bộ nhận dạng tiếng nói thực hiện chuyển dữ liệu auđio/tiếng nói sang nội dung văn bản tương ứng; bộ thu thập dữ liệu và đánh chỉ mục thực hiện thu thập các viđeo, rút trích kênh tiếng để chuyển giao cho bộ nhận dạng, và đánh chỉ mục văn bản kết quả trả về từ bộ nhận dạng; máy tìm kiếm có nhiệm vụ thực hiện tìm kiếm trong kho chỉ mục dựa trên câu truy vấn của người sử dụng và trả về các kết quả liên quan; bộ truyền tải dữ liệu thực hiện điều phối dữ liệu trên kênh truyền từ server đến người sử dụng.



Lĩnh vực kỹ thuật được đề cập

Giải pháp này thuộc lĩnh vực truy vấn viđeo hướng ngữ nghĩa. Đây là một trong những hướng nghiên cứu phổ biến nhất hiện nay. Một hệ thống truy vấn viđeo hướng ngữ nghĩa phải cho phép truy vấn các clip, các cảnh, hoặc các sự kiện trong viđeo dựa trên nội dung hình ảnh và âm thanh của viđeo đó; đồng thời, câu truy vấn có thể gồm nhiều dạng khác nhau như: dạng văn bản, dạng âm thanh, dạng viđeo. Các hệ thống như vậy đòi hỏi những kỹ thuật học mẫu, nhận dạng đối tượng, phát hiện sự kiện, nhận dạng tiếng nói, các kỹ thuật so khớp, đánh chỉ mục đa năng,v.v..

Khác với truy vấn dựa trên thông tin văn bản, truy vấn viđeo dựa trên nội dung ngữ nghĩa của nó đưa ra rất nhiều thách thức lớn, cụ thể là các bài toán trong hai lĩnh vực nghiên cứu chính: thị giác máy tính và nhận dạng tiếng nói. Lĩnh vực nghiên cứu về thị giác máy tính đi sâu tìm cách khai thác thông tin trong kênh ảnh của viđeo, còn lĩnh vực nhận dạng tiếng nói tập trung vào việc chuyển các lời thoại trong kênh tiếng thành văn bản.

Giải pháp hữu ích này đề xuất xây dựng hệ thống truy vấn viđeo hướng ngữ nghĩa theo nhánh của nhận dạng tiếng nói, nghĩa là dựa trên việc khai thác thông tin lời thoại và ngôn ngữ nói của viđeo.

Tình trạng kỹ thuật của giải pháp hữu ích

Viđeo là một dạng dữ liệu giàu thông tin ngữ nghĩa, hơn hẳn âm thanh và văn bản. Các nghiên cứu trên lĩnh vực truy vấn viđeo đã được tiến hành rất nhiều trong những năm gần đây, và cũng đã gặt hái được những kết quả khá

khả quan, ví dụ: hệ thống LODEM truy vấn viđeo bài giảng của Nhật, hệ thống SportsVBR truy vấn viđeo thể thao của Trung Quốc…

Một hệ thống truy vấn viđeo lý tưởng phải kết hợp cả nội dung hình ảnh lẫn thông tin thoại của lời nói. Tuy nhiên do những hạn chế trong lĩnh vực xử lý ảnh và thị giác máy tính, việc xây dựng hoàn chỉnh một hệ thống truy vấn lý tưởng vẫn còn gặp nhiều khó khăn. Do đó, các nghiên cứu về truy vấn viđeo hướng ngữ nghĩa thường tập trung theo ba hướng chính: truy vấn viđeo dựa trên thông tin thị giác/hình ảnh, truy vấn viđeo dựa trên thông tin lời thoại của tiếng nói và các phương pháp tổng hợp kết quả truy vấn trên hình ảnh và tiếng nói.

Đối với truy vấn viđeo dựa trên tiếng nói, lời thoại trong viđeo sẽ được chuyển sang văn bản thông qua bộ nhận dạng tiếng nói. Các văn bản này sẽ được đánh chỉ mục và việc truy vấn được thực hiện trên văn bản như trong các hệ thống truy vấn dựa trên văn bản. Tuy nhiên, do tiếng nói trong các đoạn viđeo bài giảng, thể thao…, rất đa dạng về chất lượng âm thanh và nội dung, nên bộ nhận dạng tiếng nói phải đối mặt với ba thách thức lớn là: dữ liệu nhiễu, sự xuất hiện của các từ nước ngoài và các trạng thái cảm xúc khác nhau trong tiếng nói.

Tìm ra giải pháp cho các vấn đề này sẽ giúp nâng cao hiệu suất truy vấn của hệ thống.

Bản chất kỹ thuật của giải pháp hữu ích

Giải pháp hữu ích được thực hiện nhằm giải quyết những vấn đề kỹ thuật nêu trên. Cụ thể là xây dựng hệ thống truy vấn viđeo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói. Thông qua việc khai thác các thông tin của lời thoại được nhúng trong viđeo, các sự kiện sẽ được truy vấn dựa

trên bản kết quả của bộ nhận dạng tiếng nói. Tuy nhiên, do dữ liệu tiếng nói trong viđeo, ví dụ thể thao, rất đa dạng cả về chất lượng âm thanh lẫn nội dung, nên bộ nhận dạng tiếng nói cần giải quyết ba vấn đề chính: dữ liệu nhiễu, sự xuất hiện của các từ nước ngoài (tên cầu thủ, tên quốc gia…) trong lời thoại, và các trạng thái cảm xúc khác nhau trong tiếng nói.

Ba vấn đề này ảnh hưởng rất lớn đến độ chính xác của bộ nhận dạng tiếng nói, dẫn đến hiệu suất truy vấn cũng không cao. Do đó, ba hướng giải quyết tương ứng cho các vấn đề này đã được đề xuất, nghiên cứu thử nghiệm, và ứng dụng vào hệ thống gồm:

- Phương pháp khử nhiễu nền.
- Mô hình phiên âm các từ nước ngoài sang cách đọc Việt.
- Mô hình ngữ âm cho tiếng nói có cảm xúc.

Theo khía cạnh thứ nhất của giải pháp, phương pháp khử nhiễu nền giúp làm giảm thiểu các âm thanh/nhiễu nền đã được trộn vào chung với lời thoại trong quá trình thu âm, việc loại bỏ/giảm thiểu chúng sẽ giúp làm nâng cao chất lượng tiếng nói, từ đó cải thiện đáng kể độ chính xác của bộ nhận dạng và hiệu suất truy vấn của cả hệ thống.

Theo khía cạnh thứ hai của giải pháp, bộ nhận dạng tiếng nói phải có khả năng xử lý một từ tiếng nước ngoài bất kỳ, nghĩa là khi trong lời thoại có xuất hiện một từ nước ngoài, hệ thống phải nhận ra từ đó dưới dạng cách đọc Việt rồi chuyển sang từ gốc tiếng nước ngoài tương ứng.

Theo khía cạnh thứ ba của giải pháp, việc cải tiến mô hình ngữ âm cho bộ nhận dạng sẽ giúp nâng cao hiệu suất nhận dạng đối với tiếng nói có cảm xúc như các lời bình, lời thoại trong viđeo thể thao.

Mô tả vắn tắt các hình vẽ

Fig.1 là hình vẽ minh họa kiến trúc tổng quát của hệ thống truy vấn viđeo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói.

Fig.2 là hình vẽ minh họa bộ nhận dạng tiếng nói của hệ thống truy vấn viđeo hướng ngữ nghĩa.

Fig.3 là hình vẽ thể hiện các bước rút trích đặc trưng MFCC từ tín hiệu âm thanh/tiếng nói.

Fig.4 là hình vẽ thể hiện tiến trình nhận dạng thích nghi của bộ nhận dạng tiếng nói.

Fig.5 là hình vẽ thể hiện thành phần phiên âm từ nước ngoài của bộ nhận dạng tiếng nói.

Fig.6 là hình vẽ minh họa phương pháp khử nhiễu nền.

Fig.7 là hình vẽ minh họa mô hình ngữ âm cho tiếng nói có cảm xúc.

Mô tả chi tiết giải pháp hữu ích

Kiến trúc tổng quát của hệ thống truy vấn viđeo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói được thể hiện trên Fig.1, gồm các thành phần:

- Bộ nhận dạng tiếng nói: thực hiện nhận dạng lời thoại trong kênh tiếng của viđeo, nghĩa là chuyển dữ liệu auđio/tiếng nói sang nội dung văn bản tương ứng.

- Bộ thu thập dữ liệu và đánh chỉ mục: thực hiện thu thập các đoạn viđeo, rút trích kênh tiếng chuyển giao cho bộ nhận dạng, và đánh chỉ mục văn bản kết quả trả về từ bộ nhận dạng.

- Máy tìm kiếm: có nhiệm vụ thực hiện tìm kiếm trong kho chỉ mục dựa trên câu truy vấn của người sử dụng, và trả về kết quả tương ứng.

- Bộ truyền tải dữ liệu: thực hiện điều phối dữ liệu kết quả trên kênh truyền từ server đến người sử dụng.

Fig.1 còn thể hiện tiến trình hoạt động của hệ thống truy vấn. Dữ liệu audio trích được từ video sẽ được tự động chuyển sang văn bản nhờ bộ nhận dạng tiếng nói, sau đó chúng được đánh chỉ mục để tạo cơ sở dữ liệu chỉ mục cho máy tìm kiếm. Máy tìm kiếm nhận câu truy vấn của người dùng dưới dạng văn bản, phân tích chúng thành các từ khóa, đối chiếu với cơ sở dữ liệu chỉ mục có sẵn trong hệ thống. Sau đó máy tìm kiếm sẽ xác định các tài liệu liên quan và trả kết quả về cho người dùng.

Các phần tiếp theo sau đây sẽ mô tả chi tiết từng thành phần của hệ thống truy vấn.

Bộ nhận dạng tiếng nói

Bộ nhận dạng tiếng nói nhận đầu vào là các tín hiệu tiếng nói dưới dạng các tệp/luồng âm thanh. Tín hiệu âm thanh này sẽ được đưa qua bước rút trích đặc trưng (thể hiện trên Fig.3), rồi thực hiện tiến trình nhận dạng và trả về kết quả ở dạng văn bản. Nói nôm na, bộ này làm nhiệm vụ chuyển tiếng nói thành văn bản.

Cách tiếp cận của bộ nhận dạng tiếng nói (ASR) là dựa vào thống kê, cụ thể là các mô hình Markov ẩn, mô hình ngôn ngữ N-gram. Ở mức tổng quát nhất, bộ ASR được thể hiện trên Fig.2, bao gồm các thành phần:

- *Mô hình ngữ âm*: liên quan đến việc biểu diễn tri thức cho tín hiệu ngữ âm, âm vị, ngữ điệu…

- *Mô hình ngôn ngữ*: liên quan đến việc biểu diễn tri thức của các từ, chuỗi từ, hình thành nên câu.
- *Tiến trình tìm kiếm trên đồ thị nhận dạng*: chọn lựa chuỗi từ ứng với tín hiệu ngữ âm. Về mặt bản chất đây là việc tìm kiếm tối ưu trên đồ thị được xây dựng bằng cách kết-ghép các mô hình ngôn ngữ, mô hình ngữ âm và từ điển phát âm.

Tuy nhiên, khác biệt giữa dữ liệu huấn luyện và dữ liệu tiếng nói của viđeo sẽ gây suy giảm đáng kể độ chính xác của bộ nhận dạng. Do đó, bộ nhận dạng của hệ thống truy vấn này áp dụng tiến trình nhận dạng thích nghi như thể hiện trên Fig.4 để giảm thiểu sai biệt giữa huấn luyện và nhận dạng.

Ba vấn đề quan trọng nhất mà bộ nhận dạng tiếng nói của một hệ thống truy vấn viđeo cần giải quyết là: nhiễu nền xen vào dữ liệu, sự xuất hiện của các từ nước ngoài trong lời thoại, và sự đa dạng cảm xúc trong tiếng nói. Các vấn đề này ảnh hưởng rất lớn đến hiệu suất nhận dạng, do đó bộ nhận dạng được trang bị thêm: phương pháp khử nhiễu nền, mô hình phiên âm từ nước ngoài, và một mô hình ngữ âm mới.

Mô hình phiên âm từ nước ngoài

Lời thoại trong viđeo, ví dụ thể thao, rất hay xuất hiện từ nước ngoài, đặc biệt là tên riêng (như tên cầu thủ, tên quốc gia…). Để có thể nhận dạng được các từ này thì phương pháp đơn giản nhất là tiến hành xây dựng một từ điển phiên âm từ nước ngoài. Tuy nhiên cách này không khả thi vì khó có thể xây dựng được một bộ từ điển từ nước ngoài đầy đủ, và nếu thực hiện thì phải thường xuyên bổ sung dữ liệu, cho nên rất tốn chi phí.

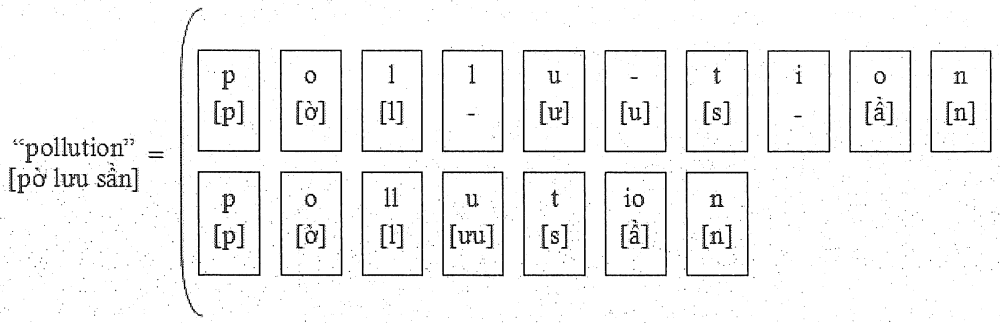
Để có khả năng nhận dạng một từ nước ngoài bất kỳ, bộ nhận dạng tiếng nói được trang bị thêm thành phần phiên âm được thể hiện trên Fig.5. Thành phần này sử dụng phương pháp dựa trên sự tương đồng về cách phát âm, thực hiện chuyển từ nước ngoài sang cách đọc Việt, quá trình này gọi là quá trình phiên âm chuyển ngữ. Cụ thể phương pháp được sử dụng là phương pháp chuyển ngữ dựa trên mô hình joint-sequence và khái niệm đơn vị liên hợp của nó. Một đơn vị liên hợp gồm hai thành phần: thành phần ký tự nước ngoài và thành phần âm vị Việt.

Ý tưởng chính của mô hình âm vị liên hợp là mối quan hệ giữa chuỗi đầu vào và chuỗi đầu ra có thể được phát sinh từ một chuỗi các đơn vị kết nối mà mang cả ký hiệu đầu vào và đầu ra. Trong trường hợp đơn giản nhất, mỗi đơn vị có thể mang 0 hoặc 1 ký hiệu đầu vào và mang 0 hoặc 1 ký hiệu đầu ra. Điều này tương ứng với định nghĩa của Finite State Transducer (FST). Khi các đơn vị kết nối có thể mang nhiều ký hiệu đầu vào và đầu ra, thuật ngữ graphone được sử dụng. Một graphone là một cặp $q = (g, \varphi) \in Q \subseteq G^* \times \phi^*$ của chuỗi ký tự và chuỗi âm vị với chiều dài có thể khác nhau. Các ký hiệu g_q và φ_q được sử dụng để chỉ thành phần thứ nhất và thứ hai của q. Kho graphone Q có thể được suy ra tự động từ dữ liệu huấn luyện hoặc nó có thể được đặc tả bằng tay. Trong mô hình này, giả sử rằng mỗi từ với dạng chính tả của tiếng nước ngoài và cách phát âm của nó trong tiếng Việt được phát sinh bằng chuỗi graphone thông dụng. Ví dụ, các phát âm của từ "pollution" có thể được tách ra thành chuỗi bảy graphone như sau:

"pollution" [pờ lưu sần]	=	p [p]	o [ờ]	ll [l]	u [ưu]	t [s]	io [ần]	n [n]

Chuỗi ký tự và âm vị được nhóm vào trong cùng số lượng phân đoạn. Mỗi nhóm như vậy được gọi là một joint segmentation hay co-segmentation.

Thuật ngữ tổng quát hơn "căn chỉnh" có thể được sử dụng thay thế. Với mỗi cặp chuỗi đầu vào và đầu ra, sự phân đoạn thành các graphone có thể không là duy nhất. Ví dụ, từ "pollution" có thể được phân đoạn thành các chuỗi graphone như sau đều là hợp lệ:



Bởi vì sự nhập nhằng này, xác suất kết $p(g,\varphi)$ được xác định bằng tổng của tất cả chuỗi graphone được so khớp $p(g,\varphi) = \sum_{q \in S(g,\varphi)} p(q)$ trong đó $q \in Q^*$ là một chuỗi graphone và $S(g,\varphi)$ là tập tất cả co-segmentation của g và φ.

Như vậy, phân phối xác suất kết $p(g,\varphi)$ được tinh giảm thành phân phối xác suất $p(q)$ trên chuỗi graphone $q = q_1,...,q_K$ mà có thể được mô hình hoá sử dụng một xấp xỉ N-gram chuẩn :

$$p(q_1^K) \cong \prod_{i=1}^{K+1} p(q_i \mid q_{i-1},...,q_{i-M+1})$$

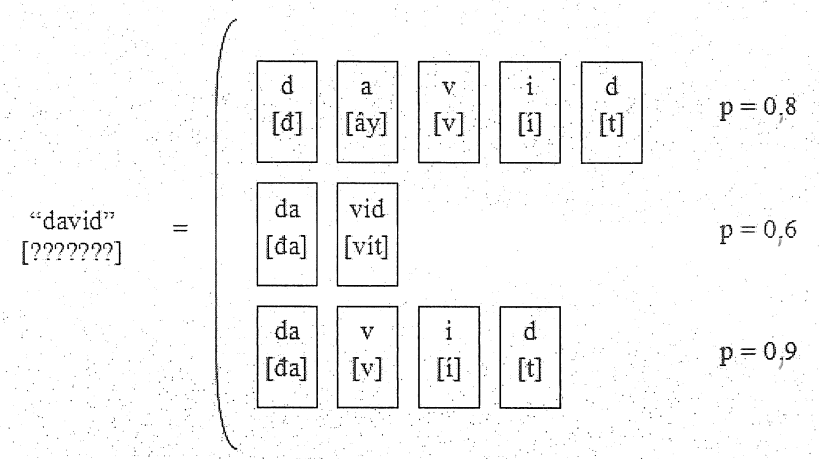
Các tham số N-gram này có thể được huấn luyện bằng thuật toán Discounted EM. Quá trình huấn luyện và giải mã của mô hình joint-sequence được thể hiện trên Fig.5.

Sử dụng mô hình này, thành phần phiên âm từ nước ngoài của bộ nhận dạng tiếng nói có thể thực hiện chuyển trực tiếp từ chuỗi ký tự tiếng nước ngoài sang chuỗi âm vị tiếng Việt. Quá trình chuyển ngữ được thực hiện thông qua việc tìm kiếm một chuỗi graphone có xác suất N-gram lớn nhất mà

thành phần ký tự của nó là từ nước ngoài; khi đó thành phần âm vị chính là kết quả phiên âm cần tìm:

$$\varphi(g) = \varphi(\arg\max_{q \in Q^* | g(q)=g} p(q))$$

Ví dụ: với từ nước ngoài DAVID, tìm một chuỗi graphone có xác suất N-gram lớn nhất sao cho thành phần kí tự trong chuỗi khớp với từ nước ngoài đầu vào, thì khi đó thành phần âm vị của chuỗi đó sẽ là cách chuyển ngữ của từ nước ngoài trên:



Phương pháp khử nhiễu nền

Nhiễu âm thanh là sự trộn lẫn những tín hiệu âm thanh không mong muốn trong quá trình thu âm. Nhiễu âm thanh được phân thành hai loại: nhiễu cộng hưởng (additive noise) và nhiễu chập (convolutional noise).Nhiễu cộng hưởng là loại tín hiệu nhiễu sinh ra do tiếng động cơ, tiếng gió, tiếng các loại nhạc cụ (kèn, trống…) cộng thêm vào tín hiệu thu âm. Nhiễu chập là loại tín hiệu nhiễu sinh ra do sự chồng chéo tiếng vang của microphone trong phòng thu và tiếng người phát thanh, do sự chồng chéo mà tín hiệu nhiễu thay đổi tại từng thời điểm khác nhau, loại nhiễu này bị ảnh hưởng bởi đặc điểm của microphone và phòng thu. Theo đó, nhiễu trong kênh âm thanh của viđeo thể

thao được xem là loại nhiễu cộng hưởng do có sự cộng hưởng của rất nhiều tiếng ồn khác nhau trên sân vận động như: tiếng hò hét, tiếng kèn, tiếng trống vào tiếng nói của bình luận viên.

Để giảm thiểu ảnh hưởng của nhiễu cộng hưởng/nhiễu nền, dữ liệu âm thanh của viđeo trước khi đưa vào nhận dạng sẽ được áp dụng tiến trình khử nhiễu như trên Fig 6:

- Đầu tiên, tín hiệu âm thanh bị nhiễu sẽ được chia frame và chuyển sang miền tần số.
- Phổ của âm thanh nhiễu sau đó sẽ được trừ đi một khoảng năng lượng phổ của tín hiệu toàn nhiễu.
- Kết quả của phép trừ phổ cùng với pha ban đầu sẽ được dùng để tái cấu trúc lại tín hiệu âm thanh sau khi khử nhiễu.

Mô hình ngữ âm cho tiếng nói có cảm xúc

Mô hình ngữ âm truyền thống trong các hệ thống nhận dạng tiếng nói tiếng Việt trước đây xem dấu như là các phone riêng. Cơ chế này sẽ gặp nhiều khó khăn trong việc mô hình hóa tiếng nói có cảm xúc, do việc biểu đạt cảm xúc được dồn vào âm vị chính của âm tiết – âm vị có mang dấu/thanh điệu.

Do đó, để tăng khả năng mô hình hóa ngữ âm biểu cảm, bộ nhận dạng tiếng nói của hệ thống truy vấn viđeo sử dụng một mô hình ngữ âm mới. Trong đó, dấu không được xem là các phone riêng, và tập phone chính là tập đầy đủ các âm vị trong tiếng Việt như thể hiện trên Fig.7.

Để có thể dễ hình dung hơn, xét một ví dụ về âm tiết "vào" trong đó âm vị "a" mang thanh điệu "huyền." Mô hình ngữ âm truyền thống biểu diễn "à" bằng 2 đơn vị HMM là "a" và "\" trong khi mô hình ngữ âm mới chỉ sử

dụng 1 HMM kết dấu là "à." Các thông tin về cảm xúc trong tiếng nói luôn được hàm chứa trong thanh điệu (kể cả thanh ngang/không dấu) và trường độ, cao độ của nguyên âm mang thanh điệu đó. Vì vậy, dạng thể hiện HMM đơn (như "à") của mô hình ngữ âm mới sẽ biểu diễn thông tin cảm xúc chính xác hơn, cũng như cho phép tùy biến trường độ, cao độ của âm mang cảm xúc.

Bộ thu thập dữ liệu và đánh chỉ mục

Mục tiêu của bộ này là thực hiện thu thập viđeo thể thao, trích ra kênh auđio và chuyển sang bộ nhận dạng. Văn bản được tạo ra từ bộ nhận dạng sẽ được bộ này đánh chỉ mục và đưa vào cơ sỡ dữ liệu tìm kiếm.

Việc đánh chỉ mục trước tiên sẽ loại ra các từ khóa không xác định: mạo từ, giới từ, liên từ… sau đó thực hiện đánh chỉ mục dựa trên đặc trưng của dữ liệu. Với dữ liệu bóng đá, thì các thông tin như: "tên cầu thủ," "tên đội bóng," "tỉ số bàn thắng," "cầu thủ ghi bàn," "phút ghi bàn," "thẻ vàng," "thẻ đỏ" được đánh giá như là những thông tin quan trọng mà người dùng quan tâm nhất sẽ được thực hiện đánh chỉ mục.

Danh sách các từ khóa được dùng để đánh chỉ mục được cập nhật liên tục qua việc thống kê các từ khóa trong truy vấn của người dùng. Dữ liệu chỉ mục được lưu dưới dạng inverted file, một cấu trúc thường được dùng trong việc xây dựng các bộ máy tìm kiếm.

Máy tìm kiếm

Máy tìm kiếm có nhiệm vụ tiếp nhận từ khóa truy vấn của người dùng, tiến hành truy vấn cơ sở dữ liệu chỉ mục trong hệ thống và trả về danh sách các kết quả tìm kiếm mong muốn.

Trước tiên câu truy vấn của người dùng sẽ qua bước tiền xử lý nhằm loại bỏ các từ khóa không cần thiết, chuẩn hóa các từ khóa theo quy ước của

hệ thống (như chuyển các con số thành chữ số, chuẩn hóa các tên riêng, tên nước ngoài, xử lý lỗi chính tả) và mở rộng từ khóa tìm kiếm nhằm cung cấp khả năng tìm các từ đồng nghĩa, các nhóm từ xuất hiện trong cùng một ngữ cảnh.

Sau khi từ khóa tìm kiếm được chuẩn hóa và mở rộng, hệ thống sẽ sử dụng một hàm đo mức độ liên quan của câu truy vấn với tất cả các tài liệu trong hệ thống. Dựa trên kết quả của độ đo này, hệ thống trả về danh sách các tài liệu liên quan đến câu truy vấn. Một cách hình thức, nếu biểu diễn câu truy vấn q và tài liệu d dưới dạng vectơ :

$$q = [q1q2\ldots qn]$$

$$d = [d1d2\ldots dn]$$

trong đó, giá trị thành phần thứ i của vector là giá trị của từ khóa thứ i trong câu truy vấn q / tài liệu d được tính bằng công thức:

$$q_i = \left(\log\left(tf_{q_i}\right) + 1.0\right).\log\left(\frac{N+1}{n_i}\right)$$

$$d_i = \log\left(tf_{d_i}\right) + 1.0$$

Với tf_{qi} là tần số của từ khóa thứ i trong vectơ truy vấn q.

tf_{di} là tần số của từ khóa thứ i trong vectơ tài liệu d.

N: tổng số tài liệu.

n_i: số tập tin chứa từ khóa thứ i.

Giá trị hàm đo mức độ liên quan của câu truy vấn q với tài liệu d được tính bằng công thức:

$$score(q,d) = \frac{q.d}{\|q\|.\|d\|}$$

Hay nói cách khác, giá trị này là cosin của góc giữa hai vector q và d. Giá trị hàm độ đo càng lớn thì vector d càng gần vector q, nghĩa là tài liệu d có mức độ liên quan càng cao với câu truy vấn q.

Bộ truyền tải dữ liệu

Trong trường hợp hệ thống truy vấn được ứng dụng vào dịch vụ lưu trữ/cung cấp viđeo, bộ truyền tải dữ liệu sẽ được tích hợp thêm vào. Từ kết quả trả về của máy tìm kiếm – danh sách các tài liệu (các tập tin viđeo) có chứa nội dung tương ứng với từ khóa tìm kiếm, hệ thống tiếp tục xử lý trả về nội dung tập tin viđeo theo thời gian thực nếu người dùng gửi yêu cầu xem tập tin này.

Theo đó, kết quả trả về từ máy tìm kiếm phải bao gồm hai thông tin : tên tập tin viđeo chứa từ khóa tìm kiếm và khoảng thời gian xuất hiện từ khóa tìm kiếm trong tập tin viđeo (khoảng thời gian này đã được tính toán và lưu trữ lại trong quá trình chia đoạn âm thanh, chính là khoảng thời gian bắt đầu của mỗi đoạn âm thanh được chia nhỏ).

Việc xử lý trả kết quả tập tin viđeo theo thời gian thực dựa trên giao thức RTMP (Real Time Messaging Protocol), là giao thức được phát triển bởi Adobe, hỗ trợ việc truyền dữ liệu viđeo trên môi trường mạng theo thời gian thực. Việc xử lý streaming viđeo dựa trên giao thức RTMP gồm hai phần: client và server. Client được tích hợp một trình flash player có khả năng trình diễn dữ liệu viđeo được truyền về từ server và một server hỗ trợ việc truyền dữ liệu theo giao thức RTMP. Việc truyền nhận dữ liệu theo giao thức này diễn ra như sau :

- Dữ liệu viđeo gồm hình ảnh và auđio sẽ được phân chia thành các đoạn gọi là các đoạn (fragment) trước khi được truyền đi, kích thước của một

fragment phụ thuộc vào băng thông đường truyền, thường là 128 bytes cho dữ liệu viđeo và 64 bytes cho dữ liệu auđio.

- Các fragment có thể được truyền đi theo cơ chế ghép kênh và cơ chế kiểm soát lỗi dữ liệu trên đường truyền. Một số kênh truyền được sử dụng trong giao thức RTMP bao gồm: kênh đảm nhiệm RPC (Remote procedure call) sử dụng Active Message Format, kênh đảm nhiệm truyền nhận dữ liệu viđeo, kênh đảm nhiệm truyền nhận dữ liệu auđio, kênh đảm nhiệm truyền nhận thông điệp quá tải đường truyền.

- Dữ liệu đến client sẽ được thực hiện tách kênh và trình diễn thông qua flash player.

Ví dụ thực hiện giải pháp

Hệ thống truy vấn viđeo thể thao theo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói được triển khai ứng dụng trong nhiều lĩnh vực khác nhau như:

- Dịch vụ truy vấn dữ liệu thể thao của Đài truyền hình.
- Search engine cho các công ty thể thao.

YÊU CẦU BẢO HỘ

1. Hệ thống truy vấn viđeo hướng ngữ nghĩa dựa trên công nghệ nhận dạng tiếng nói bao gồm:

bộ nhận dạng tiếng nói thực hiện chuyển dữ liệu auđio/tiếng nói sang nội dung văn bản tương ứng;

bộ thu thập dữ liệu và đánh chỉ mục thực hiện thu thập các viđeo, rút trích kênh tiếng để chuyển giao cho bộ nhận dạng, và đánh chỉ mục văn bản kết quả trả về từ bộ nhận dạng;

máy tìm kiếm có nhiệm vụ thực hiện tìm kiếm trong kho chỉ mục dựa trên câu truy vấn của người sử dụng và trả về các kết quả liên quan;

bộ truyền tải dữ liệu thực hiện điều phối dữ liệu trên kênh truyền từ server đến người sử dụng; trong đó khác biệt ở chỗ:

bộ nhận dạng tiếng nói gồm:

môđun rút trích đặc trưng thực hiện biến đổi tín hiệu âm thanh thành các vectơ đặc trưng MFCC làm đầu vào cho nhận dạng;

môđun phiên âm từ nước ngoài thực hiện phiên âm một từ nước ngoài bất kỳ sang cách phát âm Việt hoặc ngược lại, được xây dựng theo hướng tiếp cận của mô hình thống kê;

môđun khử nhiễu nền/nhiễu cộng hưởng thực hiện giảm thiểu nhiễu nền/nhiễu cộng hưởng trong tín hiệu âm thanh, dựa trên phương pháp trừ phổ (Spectral Subtraction);

bộ mô hình hóa ngữ âm cho tiếng nói có cảm xúc thực hiện mô hình hóa thông tin ngữ âm có cảm xúc, dựa trên phương pháp tích hợp thanh

điệu/ngữ điệu vào âm vị (sử dụng HMM kết dấu) cùng với cơ chế tùy biến cao độ, trường độ phát âm.

2. Hệ thống theo điểm 1, trong đó môđun phiên âm từ nước ngoài thực hiện phiên âm một từ nước ngoài bất kỳ sang cách phát âm Việt hoặc ngược lại thông qua việc tìm kiếm một chuỗi graphone có xác suất N-gram lớn nhất mà thành phần ký tự của nó là từ nước ngoài; khi đó thành phần âm vị chính là kết quả phiên âm cần tìm:

$$\varphi(g) = \varphi(\arg\max_{q \in Q^* | g(q)=g} p(q))$$

với $p(q)$ là phân phối xác suất trên chuỗi graphone xấp xỉ N-gram chuẩn $p(q_1^K) \cong \prod_{i=1}^{K+1} p(q_i \mid q_{i-1}, \ldots, q_{i-M+1})$.

3. Hệ thống theo điểm 1 hoặc 2, trong đó môđun phiên âm (chuyển ngữ) theo phương pháp chuyển ngữ dựa trên mô hình Joint-Sequence và đơn vị liên hợp của nó.

4. Hệ thống theo điểm 3, trong đó đơn vị liên hợp bao gồm thành phần ký tự nước ngoài và thành phần âm vị Việt.

5. Hệ thống theo điểm 1, trong đó môđun khử nhiễu nền/nhiễu cộng hưởng dựa trên phương pháp trừ phổ được thực hiện như sau:

i) chia khung (frame) tín hiệu âm thanh bị nhiễu và chuyển sang miền tần số;

ii) trừ đi một khoảng năng lượng phổ của tín hiệu toàn nhiễu phổ của âm thanh nhiễu;

iii) tái cấu trúc tín hiệu âm thanh sau khi khử nhiễu từ kết quả của phép trừ phổ cùng với pha ban đầu.

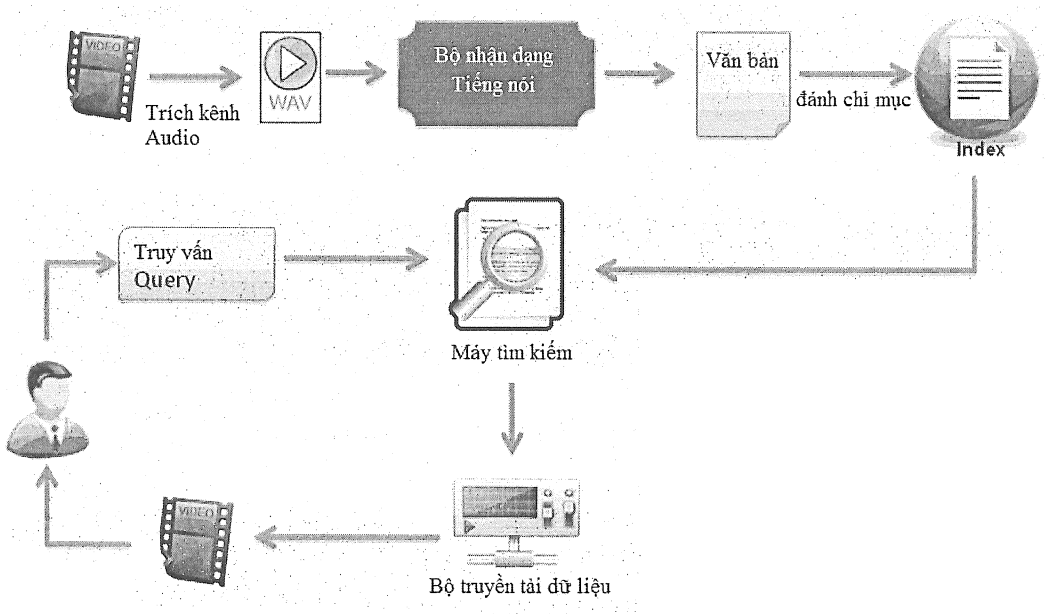

Trích kênh
Audio
WAV
Bộ nhận dạng
Tiếng nói
Văn bản
đánh chỉ mục
Index
Truy vấn
Query
Máy tìm kiếm
Bộ truyền tải dữ liệu

Fig 1.

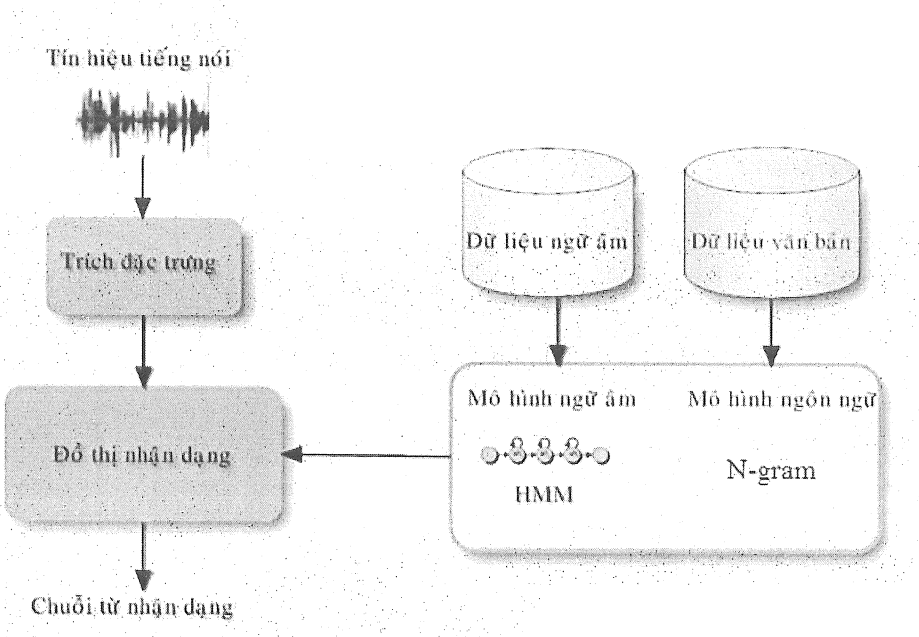

Tín hiệu tiếng nói
Trích đặc trưng
Đồ thị nhận dạng
Chuỗi từ nhận dạng
Dữ liệu ngữ âm
Dữ liệu văn bản
Mô hình ngữ âm
HMM
Mô hình ngôn ngữ
N-gram

Fig 2

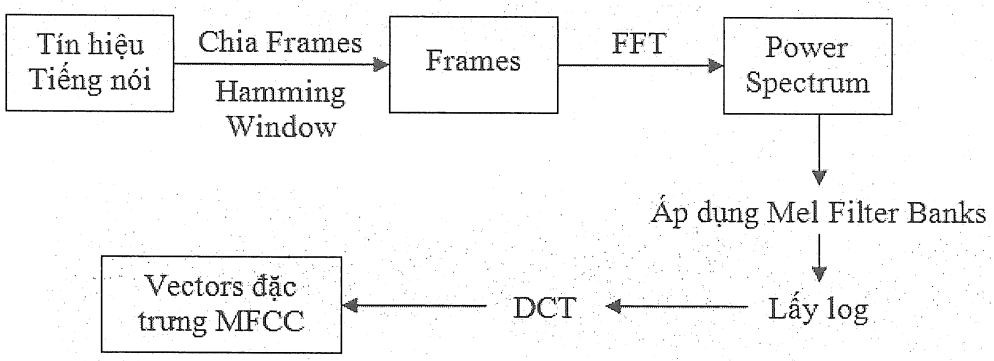

Tín hiệu Tiếng nói
Chia Frames
Hamming Window
Frames
FFT
Power Spectrum
Áp dụng Mel Filter Banks
Lấy log
DCT
Vectors đặc trưng MFCC

Fig 3

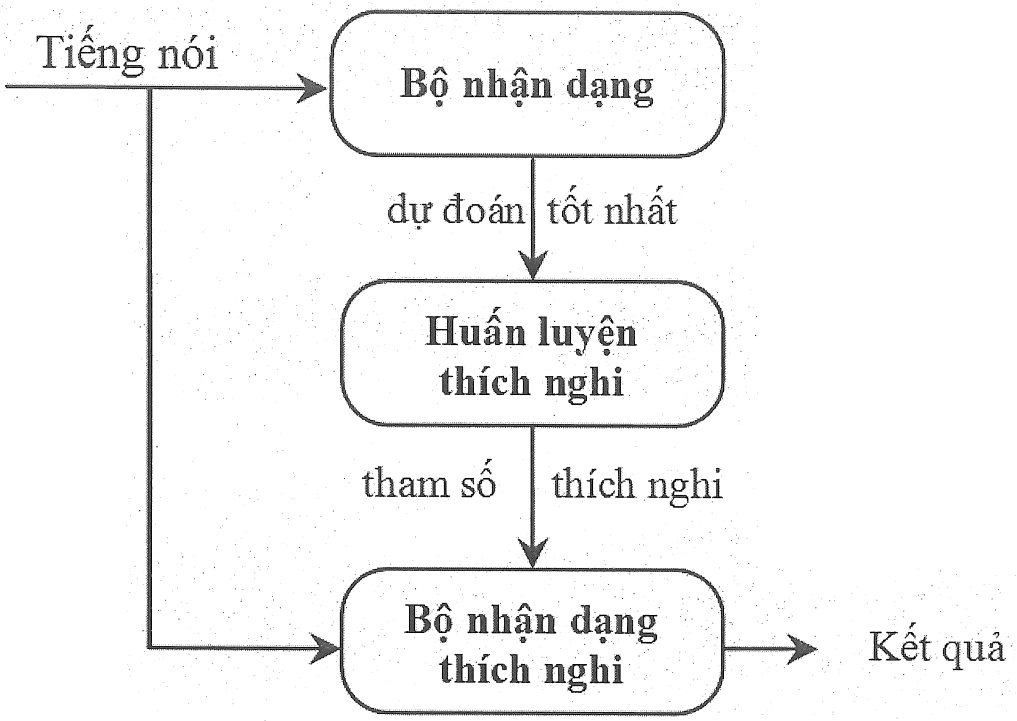

Tiếng nói
Bộ nhận dạng
dự đoán tốt nhất
Huấn luyện thích nghi
tham số thích nghi
Bộ nhận dạng thích nghi
Kết quả

Fig 4

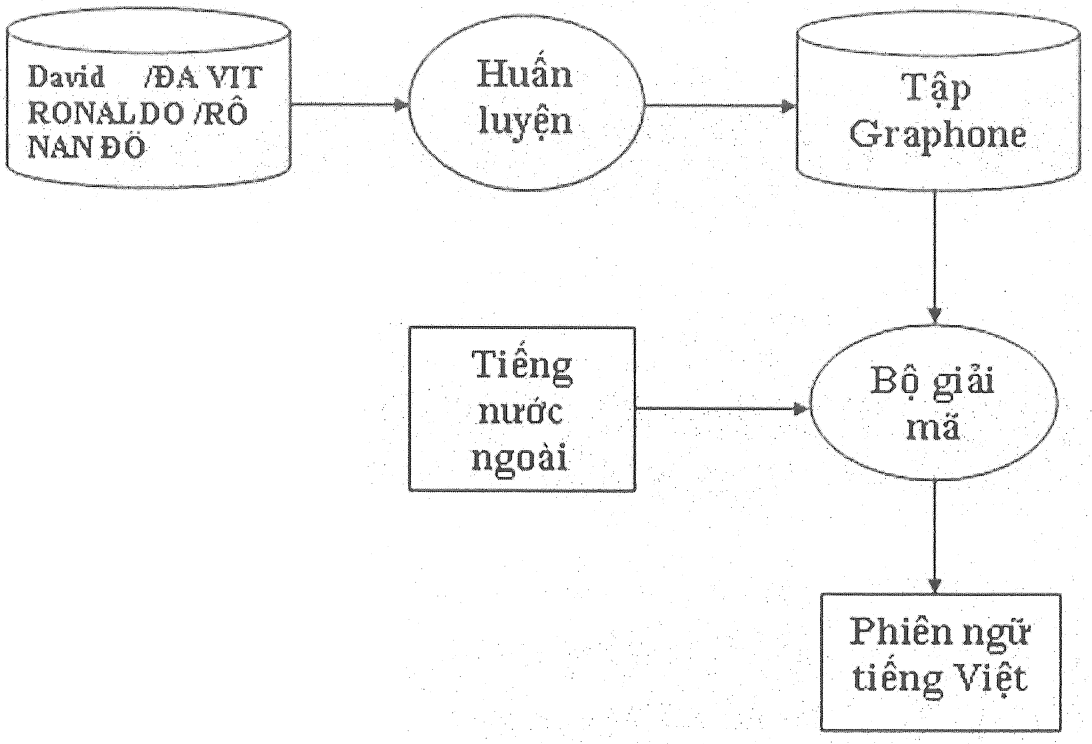

David /ĐA VIT
RONALDO /RÔ
NAN ĐÔ
Huấn luyện
Tập Graphone
Tiếng nước ngoài
Bộ giải mã
Phiên ngữ tiếng Việt

Fig 5

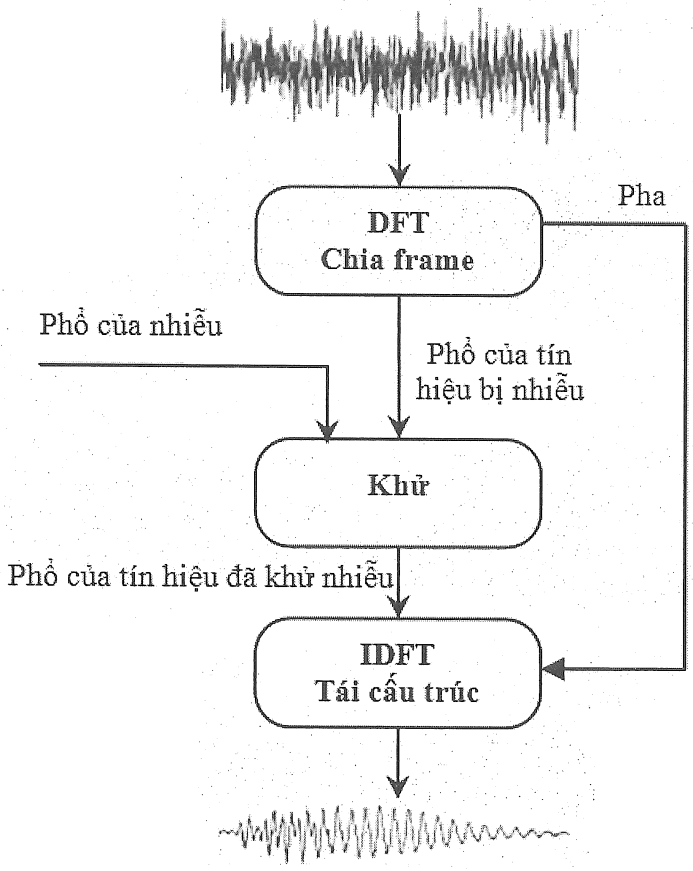

DFT
Chia frame
Pha
Phổ của nhiễu
Phổ của tín
hiệu bị nhiễu
Khử
Phổ của tín hiệu đã khử nhiễu
IDFT
Tái cấu trúc

Fig 6

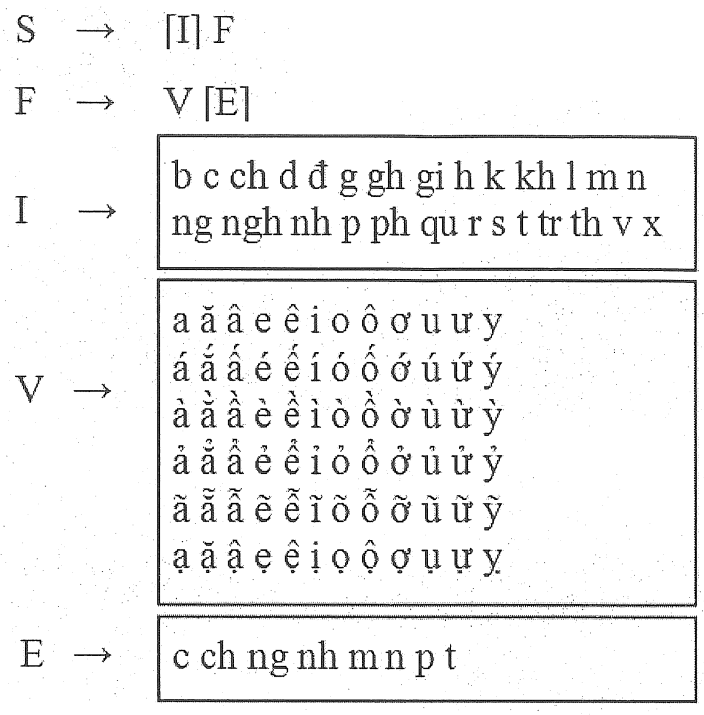

S → [I] F
F → V [E]
I →
b c ch d đ g gh gi h k kh l m n
ng ngh nh p ph qu r s t tr th v x
V →
a ă â e ê i o ô ơ u ư y
á ắ ấ é ế í ó ố ớ ú ứ ý
à ằ ầ è ề ì ò ồ ờ ù ừ ỳ
ả ẳ ẩ ẻ ể ỉ ỏ ổ ở ủ ử ỷ
ã ẵ ẫ ẽ ễ ĩ õ ỗ ỡ ũ ữ ỹ
ạ ặ ậ ẹ ệ ị ọ ộ ợ ụ ự ỵ
E →
c ch ng nh m n p t

Fig 7