*Tạp chí Tin học và Điều khiển học, T.29, S.4 (2013), 379–388*

# ÁP DỤNG BOTTLE NECK FEATURE CHO NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

NGUYỄN VĂN HUY[1], LƯƠNG CHI MAI[2], VŨ TẤT THẮNG[2]

[1]*Khoa Điện tử, trường ĐH Kỹ thuật Công nghiệp – Thái Nguyên; huynguyen@tnut.edu.vn*

[2]*Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học & Công nghệ Việt Nam; lcmai,vtthang@ioit.ac.vn*

**Tóm tắt.** Bài báo trình bày việc áp dụng Bottle Neck Feature(BNF) - một dạng đặc trưng của tín hiệu tiếng nói được trích chọn thông qua mạng neural (Neural Network) - cho nhận dạng tiếng nói tiếng Việt. Nghiên cứu sử dụng mạng Multilayer Perceptron(MLP) năm lớp với kích thước của lớp ẩn thứ nhất khác nhau để trích chọn đặc trưng BNF từ hai loại dữ liệu đầu vào là Perceptual Linear Prediction(PLP) và Mel Frequency Cepstral Coeffcient(MFCC), nhằm đánh giá hiệu quả của mỗi loại đặc trưng sau khi được áp dụng BNF. Kết quả thử nghiệm chứng tỏ BNF hiệu quả với tiếng nói tiếng Việt, kết quả nhận dạng trên đặc trưng BNF tốt hơn so với hệ thống cơ sở (baseline system) trong khoảng từ 6% đến 7%, và đặc trưng MFCC cho kết quả tốt hơn PLP.

**Từ khóa.** Bottle Neck Feature, nhận dạng tiếng Việt, mô hình makov ẩn.

**Abstract.** In the paper, the basic idea of Bottle Neck Feature(BNF) and the process how to extract BNF are presented. In this study, we apply BNF for Vietnamese speech recognition with five layers MLP network of different sizes for the first hidden layer. Input features to extract BNF feature are Perceptual Linear Prediction(PLP) and Mel Frequency Cepstral Coeffcient(MFCC). The experiments are carried out on a data set of VOV(Voice of Vietnam).The results show that using BNF for Vietnamese speech recognition, a WER(Word Error Rate)is improved up to 6-7% comparing to the baseline system, and MFCC feature gives a better result than PLP feature.

**Key words.** BNF, Bottle Neck Feature, Vietnamese speech recognition, HMM-GMM.

## 1. GIỚI THIỆU

Phương pháp trích chọn các đặc trưng của tiếng nói sử dụng mạng neural đang trở thành một phần quan trọng trong hệ thống nhận dạng tiếng nói [1], phương pháp này nhằm tận dụng ưu điểm phân lớp của mạng neural đồng thời khắc phục một trong các nhược điểm của mô hình Markov ẩn (HMM– Hidden Markov Model), mô hình HMM không mô hình hóa được các đặc tính phụ thuộc thời gian của tín hiệu tiếng nói do HMM giả thiết rằng mỗi trạng thái hiện tại chỉ phụ thuộc vào trạng thái ngay trước nó [2, 3]. Đã có nhiều phương pháp được đưa ra nhằm khắc phục nhược điểm trên của HMM, một phương pháp được sử dụng phổ biến là bổ sung thêm các vector đặc trưng lân cận với vector đặc trưng đang xét tại thời điểm $t$, tức là tổ hợp nhiều hơn một khung dữ liệu (frame) để đưa vào huấn luyện HMM tại

thời điểm $t$, ta gọi đó là cửa sổ khung (frame windows). Ví dụ tại thời điểm $t$ nếu ta dùng frame window là 7 thì một quan sát $o_t$ sẽ là tổ hợp của 7 frames liên tiếp từ $(t-3)$ đến $(t+3)$, $o_t = \{x_{t-3}, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, x_{t+3}\}$ với $x_t$ là vector đặc trưng của tiếng nói tại thời điểm $t$. Tuy nhiên phương pháp này làm gia tăng kích thước của $o_t$, dẫn đến làm tăng kích thước của hệ thống nhận dạng. Một phương pháp vừa làm giảm kích thước của $o_t$ đồng thời phân lớp lại $o_t$ để thu được một đặc trưng mới tốt hơn đó là sử dụng mạng neural. Bằng cách này với đầu vào $o_t$ được đưa qua một mạng MLP đặc biệt đã được huấn luyện để tách những thông tin quan trọng và nén các thông tin này tạo ra một đặc trưng mới $o'_t$ ở lớp ra (output) của mạng, khi đó kích thước của $o'_t$ có thể thiết lập được thông qua việc thiết lập kích thước lớp output, kích thước của $o'_t$ thường nhỏ hơn rất nhiều so với $o_t$ mà vẫn đảm bảo chứa đủ các thông tin quan trọng bao gồm cả thông tin về ngữ cảnh thời gian. Qua nhiều nghiên cứu đã cho thấy việc sử dụng $o'_t$ như là một đầu vào cho mô hình HMM làm tăng đáng kể chất lượng nhận dạng của hệ thống, $o'_t$ còn được gọi là đặc trưng xác suất hoặc đặc trưng MLP (probability feature, MLP feature).

Đã có nhiều nghiên cứu nhằm tìm ra phương pháp trích chọn các đặc trưng của tiếng nói thông qua mạng neural sao cho có thể sử dụng nó trực tiếp như một đầu vào cho việc huấn luyện các mô hình HMM, phương pháp "Bottle Neck Feature" hiện nay là một trong các phương pháp được sử dụng rộng rãi và hiệu quả nhất [4, 5]. Ý tưởng chính của phương pháp là sử dụng một mạng neural đa lớp MLP đã được huấn luyện để tách và phân lớp lại các đặc trưng đầu ra từ các vector đặc trưng đầu vào. Các đặc trưng thu được thực chất là các giá trị kích hoạt (activation) tại các nút mạng ở lớp output, tuy nhiên điều đặc biệt là lớp output được chọn để lấy các giá trị đặc trưng là một trong các lớp ẩn và có kích thước nhỏ, hàm kích hoạt được sử dụng khi tách đặc trưng thường là hàm tuyến tính (linear function) thay vì hàm phi tuyến (non-linear function) [6, 7] như lúc huấn luyện mạng, khi đó lớp ẩn được chọn để tách các đặc trưng gọi là lớp Bottle Neck(BN), và đặc trưng thu được qua lớp BN gọi là Bottle Neck Feature (BNF).

Trong nghiên cứu này sẽ tiến hành thử nghiệm cài đặt BNF cho nhận dạng tiếng nói tiếng Việt (gọi tắt là nhận dạng tiếng Việt), nghiên cứu nhằm giải quyết hai câu hỏi chính, một là hiệu quả của BNF với nhận dạng tiếng Việt. Hai là ảnh hưởng của một số cấu trúc mạng (topology) MLP và loại đặc trưng (feature) đầu vào khác nhau tới kết quả nhận dạng. Phần 2 của bài báo giới thiệu về BNF và cách áp dụng nó trong mô hình HMM. Mô hình này được trình bày trong Phần 3, là cách tiếp cận chính trong nhận dạng tiếng nói. Phần 4 mô tả các thử nghiệm BNF với các cấu trúc MLP khác nhau, và kết quả nhận dạng của các hệ thống HMM sử dụng BNF có so sánh với hệ thống cơ sở. Kết luận và hướng nghiên cứu tiếp theo được trình bày trong Phần 5.

## 2. BOTTLE NECK FEATURE

Mạng MLP là một trong các cấu trúc mạng nơron nhận tạo ANN (Artificial Neural Network) thông dụng, cấu trúc mạng bao gồm 1 lớp đầu vào (input), 1 lớp đầu ra (output) và 1 hoặc nhiều lớp ẩn (hidden). Các giá trị của vector đầu vào sẽ được lan truyền thẳng (feed-forward) từ lớp input qua các lớp ẩn tới lớp output. Tại các nút mạng (notes) hàm kích hoạt (activation function) có thể là hàm tuyến tính hoặc là hàm phi tuyến, và có thể khác nhau giữa các notes hoặc khác nhau giữa lớp ẩn với lớp output. Một trong các ưu điểm nổi bật của ANN là khả năng phân lớp, vì thế ANN được ứng dụng rộng rãi trong các hệ thống nhận dạng nói chung và nhận dạng tiếng nói nói riêng. Có hai cách tiếp cận thông dụng trong việc ứng dụng ANN cho nhận dạng tiếng nói. Cách thứ nhất là dùng mạng ANN đã được huấn

luyện như một hệ thống nhận dạng độc lập, khi đó đầu vào input là mẫu nhận dạng, đầu ra output là kết quả nhận dạng. Cách tiếp cận thứ hai là lai ghép giữa ANN và HMM, đối với phương pháp này ANN tham gia vào việc tính hàm xác suất phát tán (2), trong nghiên cứu [8] chúng tôi cũng đã đề xuất cách tiếp cận này cho nhận dạng tiếng Việt. Bài báo này sẽ áp dụng cách tiếp cận thứ ba cho nhận dạng tiếng Việt và để trả lời hai câu hỏi: một là hiệu quả của BNF với nhận dạng tiếng Việt. Hai là ảnh hưởng của một số cấu trúc mạng (topology) MLP và loại đặc trưng (feature) đầu vào khác nhau tới kết quả nhận dạng. Một mạng MLP có cấu trúc 5 lớp dạng cổ chai (Bottle Neck - BN) như Hình 1 sẽ được sử dụng để trích chọn đặc trưng cho tiếng nói, sau đó đặc trưng này sẽ được sử dụng trực tiếp như là đầu vào cho mô hình HMM, khi trích chọn đặc trưng chỉ sử dụng 3 lớp đầu tiên (lớp input, lớp ẩn thứ nhất, lớp BN) của mạng MLP.

Như vậy đặc trưng BNF là một dạng đặc trưng của tiếng nói được trích chọn thông qua một mạng MLP có cấu trúc dạng cổ chai, để tăng tính hiệu quả của đặc trưng này cần tìm ra một cấu trúc mạng MLP tốt nhất để tối ưu khả năng phân lớp của mạng khi áp dụng cho một ngôn ngữ hay trên một tập dữ liệu cụ thể, bốn tham số cơ bản của mạng MLP cần xác định trong trường hợp này là số lớp ẩn, kích thước của các lớp ẩn, kích thước và vị trí của lớp BN.

Trong các nghiên cứu [6] và [9] các tác giả đã làm các thử nghiệm khác nhau nhằm tìm ra cấu trúc MLP và vị trí cho lớp BN tốt nhất áp dụng cho tiếng Anh. Cụ thể trong nghiên cứu [9], tác giả thử nghiệm với hai loại cấu trúc mạng MLP bốn lớp và năm lớp (một lớp input, một lớp output và hai hoặc ba lớp ẩn), kết quả các thử nghiệm cho thấy cấu trúc mạng MLP năm lớp cho kết quả tốt hơn cấu trúc mạng MLP bốn lớp khoảng 1% WER. Cũng theo các kết quả đó thì vị trí của lớp BN là lớp ẩn thứ hai sẽ cho kết quả tốt nhất, với vị trí này lớp BN sẽ tận dụng được khả năng phân lớp qua lớp ẩn thứ nhất [6]. Kích thước của lớp BN nằm trong khoảng từ 25-65, kết của của nghiên cứu [6] và [7] đạt kết quả tốt nhất với kích thước BN là 39. Việc chọn kích thước của lớp BN lớn hơn có thể làm giảm WER trên một bộ dữ liệu cụ thể, tuy nhiên việc giảm là rất nhỏ trong khi đó nó sẽ làm tăng đáng kể thời gian huấn luyện mạng MLP và đồng thời cũng làm tăng kích thước của vector đặc trưng BNF dẫn đến làm tăng kích thước của mô hình HMM.

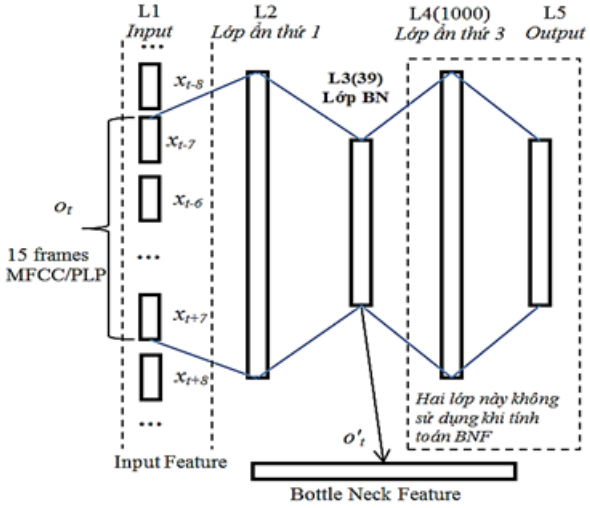
Vấn đề tiếp theo là lựa chọn kích thước của lớp ẩn thứ nhất và lớp ẩn thứ ba, trong các nghiên cứu [10] và [11] gần đây nhất của nhóm "Reseach group 3-01, KIT" nhóm đã làm các thử nghiệm trên tiếng Anh và tiếng Tây Ban Nha, các tác giả cũng sử dụng cấu trúc mạng MLP năm lớp trong đó lớp BN là lớp ẩn thứ hai có kích thước là 42. Kết quả tối ưu nhất họ thu được trên cấu trúc MLP này tương ứng với kích thước lớp ẩn thứ nhất và thứ ba lần lượt là 4000, 4000.

Dựa trên các kết quả nghiên cứu ở trên, dẫn đến quyết định cài đặt BNF cho tiếng Việt với cấu trúc mạng MLP năm lớp dạng L1-L2-L3-L4-L5. Trong đó: L1 là lớp input, kích thước của L1 phụ thuộc vào kích thước của đặc trưng đầu vào. L2 là lớp ẩn thứ nhất, trong phạm vi nghiên cứu này sẽ làm các thử nghiệm với L2 có kích thước khác nhau để tìm ra kích thước tối ưu nhất. L3 là lớp BN với kích thước 39. L4 là lớp ẩn thứ ba với kích thước định sẵn là 1000. L5 là lớp ouput, kích thước của L5 phục thuộc vào số lớp(classes) đầu ra mà mạng MLP cần phân lớp. Cấu trúc mạng MLP này được mô tả ở Hình 1.

Mạng MLP này sau đó được huấn luyện bằng phương pháp học lan truyền ngược (back propagation) có giám sát trên tập dữ liệu huấn luyện như quá trình huấn luyện một mạng MLP thông thường. Ở nghiên cứu này ta sử dụng hàm kích hoạt ở các lớp ẩn là hàm Sigmoid, tại lớp output là hàm Softmax khi huấn luyện mạng. Sự khác biệt chỉ ở bước sử dụng mạng này để trích chọn đặc trưng, toàn bộ dữ liệu huấn luyện được sử dụng lại như là đầu vào để

trích chọn đặc trưng ở đầu ra. Các vector dữ liệu này lần lượt được lan truyền thẳng từ đầu vào mạng qua lớp ẩn thứ nhất và dừng lại ở lớp BN, tại lớp BN các giá trị kích hoạt được tính toán trên các notes sử dụng hàm kích hoạt tuyến tính như công thức (1) sẽ được dùng như các đặc trưng thu được từ đầu vào, lớp ẩn thứ ba và lớp output của mạng không được sử dụng tại bước này.

Một trong các ưu điểm của phương pháp này là kích thước của vector đặc trưng BNF thu được không thay đổi (trong trường hợp này là 39) dẫn đến cấu trúc của mô hình HMM sử dụng đặc trưng này cũng không thay đổi cho dù ta muốn thay đổi kích thước của frame window để tăng thông tin về ngữ cảnh thời gian, số trạng thái cần phân lớp ở lớp output hay kích thước của các lớp ẩn.




*Hình 1.* Cấu trúc mạng MLP năm lớp với lớp Bottle Neck là lớp ẩn thứ 2

$$|BNF_i = \sum_{j=1}^{n} |r_j * |W_i + h_i, \text{ với } i = 1, ..., 39 \tag{1}$$

trong đó:

$BNF_i$ là giá trị của thành phần thứ $i$ trong vector 39 chiều BNF thu được.

$n$ là kích thước của lớp ẩn thứ nhất.

$r_j$ là giá trị kích hoạt tại note thứ $j$ ở lớp ẩn thứ nhất.

$W_i$ trọng số của note thứ $i$ tại lớp BN.

$h_i$ là hệ số Bias của note thứ $i$ tại lớp BN.

## 3. MÔ HÌNH HMM

### 3.1. Định nghĩa HMM

HMM là mô hình xác suất dựa trên lý thuyết về chuỗi Markov [13] bao gồm các đặc trưng sau:

$O = \{o_1, o_2, \ldots, o_T\}$ là tập các vector quan sát.

$S = \{s_1, s_2, ..., s_N\}$ là tập hữu hạn các trạng thái $s$ gồm $N$ phần tử.

$A = \{a_{11}, a_{12}, ..., a_{NN}\}$ là ma trận hai chiều trong đó $a_{ij}$ thể hiện xác suất để trạng thái $s_i$ chuyển sang trạng thái $s_j$, với $a_{ij} \geq 0$ và $\sum_{j=k} a_{ij} = 1, |\forall i$.

$B = \{b_{2t}, b_{it}, ..., b_{(N-1)t}\}$ là tập các hàm xác suất phát tán của các trạng thái từ $s_2$ đến $s_{N-1}$, trong đó $b_{it}$ thể hiện xác suất để quan sát $o_t$ thu được từ trạng thái $s_i$ tại thời điểm $t$. Trong nhận dạng tiếng nói hàm $b_{it}$ thường được sử dụng là hàm Gaussian với nhiều thành phần trộn (mixture) có dạng như công thức (2), trong trường hợp này ta gọi là mô hình kết hợp Hidden Markov Model và Gaussian Mixtrue Model(HMM-GMM)

$$b_i(o_t) = \sum_{k=1}^{M} c_{ik} \mathcal{N}(o_t; \mu_{ik,ik}) \tag{2}$$

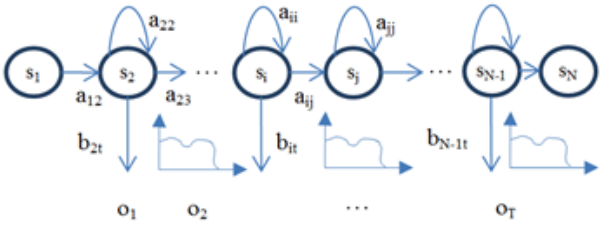
trong đó, $o_t$ là vector quan sát tại thời điểm $t$, $M$ là số thành phần trộn của hàm Gaussian, $c_{ik}, \mu_{ik,ik}$ theo thứ tự là trọng số, vector trung bình và ma trận phương sai (covariance matrix) của thành phần trộn thứ $k$ của trạng thái $s_i$.

$\Pi = \{\pi_i\}$ là tập xác suất trạng thái đầu, với $\pi_i = P(q_1 = s_i)$ với $i = 1..N$ là xác suất để trạng thái $s_i$ là trạng thái đầu $q_1$.

Như vậy một cách tổng quát một mô hình HMM $\lambda$ có thể được biểu diễn bởi $\lambda = (A, B, \Pi)$. Trong lĩnh vực nhận dạng các mô hình HMM được áp dụng với hai giả thiết sau:

+ Một là giả thiết về tính độc lập, tức không có mỗi liên hệ nào giữa hai quan sát lân cận nhau $o_i$ và $o_{i+1}$, khi đó xác suất của một chuỗi các quan sát $O = \{o_i\}$ có thể được xác định thông qua xác suất của từng quan sát $o_i$ như sau $P(O) = \prod_{i=t}^{T} P(o_i)$.

+ Hai là giả thiết Markov, xác suất chuyển thành trạng thái $s_t$ chỉ phụ thuộc vào trạng thái trước nó $s_{t-1}$. Trong nghiên cứu này sử dụng mô hình HMM-GMM có cấu trúc dạng Left-Right liên kết không đầy đủ được minh hoạ như Hình 2.




*Hình 2.* Mô hình HMM-GMM Left-Right với $N$ trạng thái

### 3.2. Áp dụng mô hình HMM trong nhận dạng tiếng nói

Trong nhận dạng tiếng nói, mô hình HMM-GMM có thể được sử dụng để mô hình hoá cho các đơn vị tiếng nói như Âm vị (phoneme), Từ (word) hoặc Câu (sentence). Khi đó tập quan sát $O = \{o_t\}$ sẽ tương ứng với mỗi một phát âm(utterance) trong đó $o_t$ là tập các vector đặc trưng (feature vector) tiếng nói đầu vào thu được tại thời điểm $t$. Có nhiều cấu trúc HMM khác nhau, tuy nhiên trong thực thế, cấu trúc của HMM-GMM thường được sử dụng có 5 hoặc 7 trạng thái theo cấu trúc Left-Right được mô tả ở Hình 2. Các hệ thống nhận dạng tiếng nói sử dụng HMM-GMM thường chia ra làm hai quá trình:

*a. Huấn luyện(training)*

Đối với từng ngôn ngữ, dữ liệu và mục đích cụ thể ta sẽ dùng HMM-GMM để mô hình cho các đơn vị nhận dạng là Âm vị, Từ hoặc Câu. Khi đó một hệ thống sẽ bao gồm một tập các mô hình HMM-GMM $\lambda = \{\lambda_i\}$. Đối với mỗi phát âm $O = \{o_t\}$ được mô hình bởi một chuỗi các trạng thái $Q = \{q_t\}$ với $q_t \in S$ từ một hoặc nhiều mô hình $\lambda_i$. Quá trình huấn luyện là quá trình ước lượng các tham số sao cho xác suất $P(Q \mid O, \lambda)$ là lớn nhất, $P(Q \mid O, \lambda)$ [13] được tính theo công thức (3), $P(Q \mid O, \lambda)$ được gọi là xác suất mô hình ngữ âm (acoustic model).

$$P(Q \mid O, \lambda) = \sum_{q_t}^{Q} \pi_{tk} a_{t_{k-1} t_k} b_{tk}(o_t),\ k = 1..N. \tag{3}$$

*b. Nhận dạng(decoding).* Nhận dạng là quá trình xác định chuỗi trạng thái $\{q_i\} = Q,\ q_i \in S$ từ các mô hình HMM $\{\lambda_i\} = \lambda$ đã được huấn luyện tương ứng với một chuỗi đầu vào $\{o_t\} = O$ sao cho xác suất $P(O, Q \mid \lambda)$ là lớn nhất, với

$$P(O, Q \mid \lambda) = \max(P(q_1, q_2, ..., q_t = i, o_1, o_2, ..., o_t \mid \lambda)).$$

## 4. CÀI ĐẶT THỬ NGHIỆM

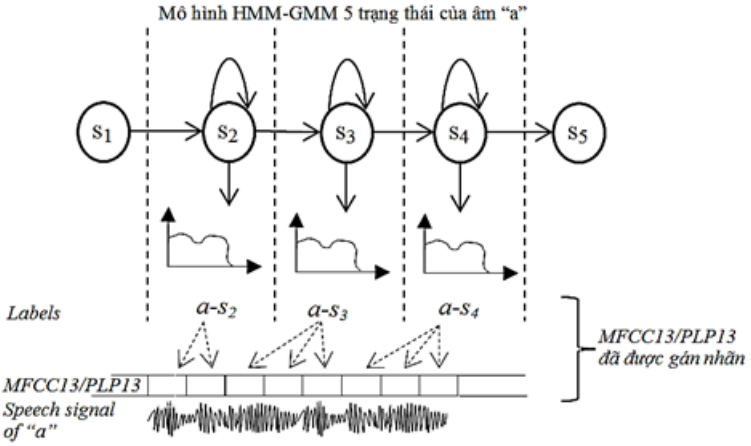
### 4.1. Dữ liệu thử nghiệm

Nghiên cứu cài đặt thử nghiệm trên bộ dữ liệu Vietnam BroadcastNews(VOV) của Viện Công nghệ thông tin-Viện Hàn Lâm Khoa học và Công nghệ Việt Nam. Tổng thời lượng khoảng 19 giờ thu âm lấy từ các mục Dọc truyện, Tin tức, Phỏng vấn của đài tiếng nói Việt Nam. Có tổng cộng 23424 câu phát âm (utterances), 30 người nói (speaker) gồm cả nam và nữ giọng miền Bắc. Từ điển được xây dựng trên tập gồm 46 âm vị không có thanh điệu (đã bao gồm cả 2 âm silence, và short-pause), bộ từ vựng có 4923 âm tiết, bao gồm hầu hết các âm tiết thường sử dụng. Trong thử nghiệm này sử dụng 17 giờ cho huấn luyện(training) và 2 giờ cho nhận dạng (decoding). Tại bước nhận dạng sử dụng mô hình ngôn ngữ (language model) mức tri-gram, được tạo ra từ toàn bộ dữ liệu phiên âm tương ứng với phần dữ liệu huấn luyện.

### 4.2. Dữ liệu huấn luyện mạng BN-MLP

Nghiên cứu đã tiến hành thử nghiệm trên hai loại đặc trưng là PLP và MFCC với số chiều là 13, ký hiệu lần lượt là PLP13, MFCC13. Đầu tiên một hệ thống nhận dạng được huấn luyện với 4000 mô hình HMM-GMM 5 trạng thái cho các âm phụ thuộc ngữ cảnh (tied-state triphone) được tạo ra từ 46 âm vị đơn, mỗi trạng thái sử dụng 16 thành phần trộn với đặc trưng đầu vào PLP13. Sau đó hệ thống này được dùng để phân đoạn (segmentation) và gán nhãn (force alignment) lại cho các đặc trưng PLP13 và MFCC13 thu được ở trên. Dữ liệu được gán nhãn ở mức trạng thái âm đơn (monophone), do hệ thống được xây dựng trên tập 46 âm vị đơn, mỗi âm vị đơn này được mô hình hoá bởi một mô hình HMM-GMM 5 trạng thái, như vậy không xét hai trạng thái đầu vào và đầu ra, mỗi âm vị sẽ có ba nhãn tương ứng với ba trạng thái (phoneme-state) $s_2, s_3$ và $s_4$ của mô hình HMM-GMM. Hình 3 là một ví dụ về mô hình HMM-GMM 5 trạng thái của âm "a", mô hình này sau khi huấn luyện được dùng để phân đoạn và gán nhãn lại cho các vector đặc trưng MFCC13/PLP13 thu được từ

tín hiệu tiếng nói tương ứng với âm "a". Trong thử nghiệm này mô hình HMM-GMM của âm short-pause chỉ có một trạng thái được cấu hình dựa trên HMM-GMM của âm silence, vì vậy ta có $45 * 3 + 1 = 136$ nhãn. Đây cũng chính là kích thước của lớp đầu ra của mạng MLP cần sử dụng. Dữ liệu sau khi được phân đoạn và gán nhãn sẽ được dùng để huấn luyện các mạng MLP.




*Hình 3.* Ví dụ mô hình HMM-GMM 5 trạng thái của âm "a" được sử dụng để phân đoạn và gán nhãn cho đặc trưng đầu vào tương ứng của âm "a"

### 4.3. Huấn luyện MLP

*Bảng 1.* Kết quả huấn luyện MLP

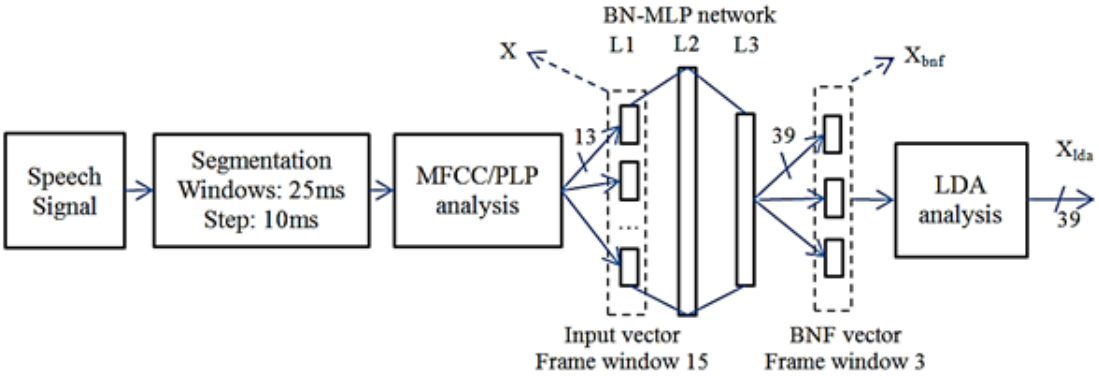
| **Feature** | **Denote** (*Input feature-Size of L2*) | **Topology** | **Cross Validation Accuracy(%) (CV)** |
|---|---|---|---|
| MFCC13 | MFCC13-1000 | 195-1000-39-1000-136 | 56.06 |
| | MFCC13-2000 | 195-2000-39-1000-136 | **56.92** |
| | MFCC13-3000` | 195-3000-39-1000-136 | 54.81 |
| PLP13 | PLP13-1000 | 195-1000-39-1000-136 | 61.41 |
| | PLP13-2000 | 195-2000-39-1000-136 | 62.94 |
| | PLP13-3000 | 195-3000-39-1000-136 | **63.54** |

Mạng MLP được sử dụng có cấu trúc như đã trình bày ở Phần 2, sử dụng frame window là 15, mỗi frame có kích thước là 13 tương ứng với số chiều của PLP13 và MFCC13, như vậy một vector đầu vào cho mạng có kích thước là $15 * 13 = 195$. Mạng có ba lớp ẩn trong đó lớp BN là lớp ẩn thứ hai L3 có kích thước cố định là 39, kích thước của lớp ẩn thứ nhất L2 sẽ được thay đổi, lớp ẩn thứ ba L4 có kích thước cố định là 1000, lớp ra output L5 có kích thước 136 tương ứng với 136 nhãn của dữ liệu huấn luyện đã được gán nhãn ở Phần 4.2. Hàm Sigmoid được sử dụng làm hàm kích hoạt ở các lớp ẩn, hàm phân lớp ở lớp đầu ra là Sorftmax. Các mạng MLP được huấn luyện với công cụ Quicknet [16], các vòng lặp huấn luyện được thực hiện cho tới khi độ lệch CV giữa hai vòng lặp liên tiếp nhỏ hơn 0.5. Để đánh giá được sự ảnh hưởng cũng như hiệu quả của các cấu trúc khác nhau lên kết quả nhận dạng tiến hành thử nghiệm với ba loại cấu trúc khác nhau trên mỗi loại đặc trưng. Sự khác biệt giữa các cấu trúc

chỉ là kích thước của lớp ẩn thứ nhất L2, ba cấu trúc áp dụng tương ứng với kích thước của L2 lần lượt là 1000, 2000, và 3000. Kết quả huấn luyện mạng MLP được trình bày ở Bảng 1.

### 4.4. Trích chọn đặc trưng BNF

Quy trình trích chọn đặc trưng BNF được mô tả ở Hình 4. Toàn bộ dữ liệu huấn luyện đã được phân đoạn ở Phần 4.2 sẽ được sử dụng như là đầu vào để trích chọn đặc trưng BNF. Tín hiệu tiếng nói sau khi được phân đoạn sử dụng cửa sổ có độ dài 25ms với tốc độ 10ms sẽ được đưa qua module phân tích để thu được đặc trưng PLP13 hoặc MFCC13, sau đó mỗi 15 khung liên tiếp sẽ được tổ hợp để tạo ra một vector đầu vào cho MLP, ta gọi đầu vào này là $X$. Như đã trình bày ở Phần 2, tại bước trích chọn đặc trưng này ta chỉ sử dụng ba lớp đầu tiên của mạng MLP (L1,L2,L3) để tính toán BNF. Sử dụng hàm lan truyền (forward function) của công cụ Quicknet với hàm đầu ra là hàm tuyến tính được trình bày ở công thức (1) để tính toán đặc trưng BNF, đặc trưng BNF này kí hiệu là $X_{bnf}$. $X_{bnf}$ được phân lớp một lần nữa sử dụng phương pháp Linear Discriminant Analysis (LDA) [15] với frame window là 3, đầu ra của bước này kí hiệu là $X_{lda}$ có kích thước 39, sau đó $X_{lda}$ sẽ được sử dụng như đầu vào cho việc huấn luyện cũng như nhận dạng với mô hình HMM-GMM.




*Hình 4.* Sơ đồ các bước trích chọn đặc trưng BNF

### 4.5. Huấn luyện mô hình HMM-GMM

Các mô hình HMM-GMM của các âm ba (triphones) được huấn luyện sử dụng công cụ Sphinx [17] trên bộ dữ liệu VOV với 4000 trạng thái buộc (tied-state triphone), mỗi trạng thái sử dụng 16 thành phần trộn. Xây dựng 6 hệ thống từ 6 đặc trưng BNF khác nhau, các đặc trưng này thu được bằng cách sử dụng 6 mạng MLP đã được huấn luyện ở Phần 4.3 tương ứng với hai đặc trưng PLP13 và MFCC13. Ở cột "Feature" của Bảng 2 các đặc trưng này được ký hiệu là MFCC13/PLP13-xxx, thể hiện đặc trưng thu được thông qua việc sử dụng mạng MLP có ký hiệu tương ứng trong bảng 2, trong đó MFCC13 hoặc PLP13 thể hiện đặc trưng đầu vào được đưa vào mạng MLP để trích chọn BNF, xxx thể hiện kích thước của lớp ẩn thứ nhất L2 của mạng được sử dụng. Để so sánh hiệu quả của BNF ta xây dựng thêm 2 hệ thống cơ sở (baseline) cùng tham số nhưng không sử dụng BNF dựa trên hai đặc trưng là PLP13 và MFCC13.

### 4.6. Kết quả thử nghiệm

Các kết quả thử nghiệm được đánh giá trên tham số WER (Word Error Rate), với các kết quả cụ thể như ở Bảng 2. Qua kết quả thử nghiệm ta dễ nhận thấy việc áp dụng BNF cho kết quả tốt hơn các hệ thống cơ sở với WER thấp hơn trung bình là 6-7%. Các kết quả thử nhiệm cũng cho thấy đặc trưng PLP cho kết quả tốt hơn MFCC 2.01% trên hệ thống cơ sở, nhưng trên hệ thống dùng BNF thì MFCC lại cho kết tốt hơn PLP 0.11%. Các kết quả cũng chỉ ra rằng kết quả huấn luyện MLP ảnh hưởng trực tiếp đến kết quả của hệ thống nhận dạng, độ chính xác đánh giá chéo (Cross Validation-CV) tỉ lệ nghịch với WER. Đối với đặc trưng PLP cho kết quả tốt nhất trên cấu hình mạng MLP có kích thước L2=3000, đối với đầu vào MFCC là L2=2000.

*Bảng 2.* Kết quả thử nghiệm

| System | Feature | WER (%) |
|---|---|---|
| Baseline | MFCC13 | 22.10 |
| BNF System | MFCC13-1000 | 15.50 |
| | MFCC13-2000 | **14.09** |
| | MFCC13-3000 | 15.80 |
| Baseline | PLP13 | 20.09 |
| BNF System | PLP13-1000 | 14.70 |
| | PLP13-2000 | 14.60 |
| | PLP13-3000 | **14.20** |

## 5. KẾT LUẬN

Nghiên cứu đã trình bày phương pháp cài đặt BNF cho nhận dạng tiếng Việt trên bộ dữ liệu có kích thước trung bình, kết quả cho thấy BNF có hiệu quả đối với tiếng Việt, kết quả trung bình tốt hơn so với hệ thống cơ sở là 6-7%, kết quả thử nghiệm tốt nhất với đặc trưng MFCC sử dụng cấu trúc là 195-2000-39-1000-136. Theo kết quả ở Bảng 1, ta có thể thấy độ chính xác đánh giá chéo CV tỉ lệ thuận với kích thước của lớp ẩn thứ nhất L2, việc làm tăng CV có thể làm giảm WER. Tuy nhiên việc tăng kích thước L2 làm tăng đáng kể thời gian huấn luyện hệ thống, trong khi giá trị giảm trên WER không thực sự lớn. Và trong thực tế việc tăng kích thước L2 không phải lúc nào cũng làm tăng giá trị CV, từ Bảng 1 ta dễ nhận thấy với đặc trưng MFCC13 kết quả CV với L2=3000 thấp hơn cấu trúc có L2=2000 là 2.1%.

So sánh các kết quả của nghiên cứu này với các nghiên cứu [6, 9, 10, 11] cho thấy để tìm ra một cấu trúc mạng MLP tối ưu nhất sẽ phụ thuộc vào từng ngôn ngữ và đặc tính của tập dữ liệu huấn luyện cụ thể. Tuy nhiên hầu hết các thí nghiệm đều đạt kết quả tốt với kích thước của L2 và L4 trong khoảng từ 1000-4000. Trong các nghiên cứu tiếp theo để hoàn thiện chúng tôi sẽ tiếp tục thử nghiệm các cấu trúc, hàm kích hoạt ở lớp BN trên các loại đặc trưng khác, nhằm đánh giá và tìm ra các tham số cho một hệ thống BNF tốt nhất cho nhận dạng tiếng Việt. Về đặc trưng thanh điệu của tiếng Việt, sẽ tiến hành nghiên cứu mô hình MSD (Multi Space Distribution) đã được nghiên cứu thành công cho tiếng Mandarin, tiếng Thái và tích hợp vào hệ thống.

## TÀI LIỆU THAM KHẢO

[1] A. Janin et al., The ICSI-SRI Spring 2006 meeting recognition system, *Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science*, vol.4299, Springer, 2006 (444–456).

[2] B. H. Juang, L. R. Rabiner, Hidden markov models for speech recognition, *Technometrics* **33** (3) (Aug. 1991) 251–272.

[3] M. Gales, S. Young, The application of hidden markov models in speech recognition, *Signal Processing* **1** (3) (2007) 195–304.

[4] Hynek Hermansky, Daniel P.W. Ellis, Sangita Sharma, Tandem connectionist feature extraction for conventional HMM systems, *Proc. ICASSP-2000*, Turkey, 2000.

[5] Christian Plahl, Ralf Schluter and Hermann Ney, Improved Acoustic Feature Combination for LVCSR by Neural Networks, in INTERSPEECH, August 2011.

[6] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocký, Probabilistic and Bottle-Neck Features for LVCSR of meetings, *Proc. ICASSP-2007, Vol.4*, Honolulu-Hawaii, 2007 (757–760).

[7] K. Vesely, M. Karafiat, F. Grezl, Convolutive Bottleneck Network features for LVCSR, *Automatic Speech Recognition and Understanding Workshop*, Haiwaii, December 2011 (42–47).

[8] Dang Ngoc Duc, John-Paul Hosom, Luong Chi Mai, HMM/ANN system for Vietnamese continuous digit recognition, *16th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Loughborough-UK, 2003, (481–486).

[9] Frantisek Grézl, Petr Fousek, Optimizing Bottel-neck features for LVCSR, *Proc. ICASSP-2008*, Las Vegas, 2008 (4729–4732).

[10] S. Stuker, K. Kilgour, C. Saam, and A. Waibel, The 2011 kit english asr system for the iwslt evaluation, *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, December, 8-9, 2011.

[11] K. Kilgour, C. Saam, C. Mohr, S. Stuker, and A. Waibel, The 2011 KIT Quaero Speech-to-text system for Spanish, *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011*, San Francisco, December, 2011 (199–205).

[12] Christian Plahl , Ralf Schlüter , Hermann Ney, Hierarchical Bottle Neck Features for LVCSR, *Proc. INTERSPEECH,* Makuhari, Japan, 2010.

[13] L. Rabiner, B. Juang, An introdution to Hidden Markov Models, *IEEE* **77** (2) (1989) 257–286.

[14] BhupinderSingh, Neha Kapur, PuneetKaur, Speech recognition with Hidden Markov Model: A review, *International Journal of Advanced Research in Computer Science and Software Engineering* **2** (3) (March 2012).

[15] M. Sakai, N. Kitaoka, S. Nakagawa, Generalization of Linear Discriminant Analysis used in Segmental Unit Input HMM for Speech Recognition, *Proc. ICASSP-2007, Vol.4*, Honolulu-Hawaii, 2007 (333–336).

[16] International computer science institute, MLP toolkit Quicknet, online: http://www1.icsi.berkeley.edu/Speech/qn.html.

[17] Carnegie Mellon University, Open Source Toolkit For Speech Recognition, CMUSphinx, online: http://cmusphinx.sourceforge.net/wiki/tutorial.

*Ngày nhận bài 14 - 7 - 2013*
*Nhận lại sau sửa ngày 22 - 11 - 2013*