NHẬN DẠNG TIẾNG VIỆT NÓI SỬ DỤNG BỘ CÔNG CỤ KALDI

Nguyễn Thị Thanh[1], Nguyễn Hồng Quang[1], Trịnh Văn Loan[1], Phạm Ngọc Hưng[2]
[1]Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội
[2] Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên

Tóm tắt: Nhận dạng tiếng nói ngày càng được ứng dụng trong nhiều lĩnh vực như tổng đài tự động; an ninh bảo mật; tìm kiếm bằng giọng nói…, tuy nhiên chất lượng nhận dạng đang là vấn đề đáng quan tâm nhất. Kaldi là một bộ công cụ mới được phát triển năm 2009. Kaldi được giới thiệu tại hội thảo diễn ra ở trường Đại học Johns Hopkins University với tiêu đề "Phát triển hệ thống nhận dạng tiếng nói chi phí thấp, chất lượng cao cho các miền và các ngôn ngữ mới" ("Low Development Cost, High Quality Speech Recognition for New Languages and Domains"). Trong bài báo này mô tả hệ thống nhận dạng tiếng Việt nói được xây dựng dựa trên bộ công cụ Kaldi. Bài báo cũng đánh giá chất lượng của hệ thống dựa trên việc đánh giá tỷ số WER của các mô hình âm học. Hệ thống đã cho ra kết quả vượt trội so với các bộ công cụ trước đó với tiếng Việt.

Từ khóa: Nhận dạng tiếng nói; tiếng Việt nói; bộ công cụ nhận dạng Kaldi; mô hình ngôn ngữ; mô hình âm học; từ điển phát âm.

I. GIỚI THIỆU

Nhận dạng tiếng nói và đặc biệt cho tiếng Việt là một lĩnh vực nghiên cứu phát triển mạnh trong những năm gần đây. Năm 2003, Đặng Ngọc Đức [1] đã sử dụng mạng nơ ron và mô hình Markov ẩn cho nhận dạng tiếng Việt nói. Năm 2004, Bạch Hưng Khang [2] đã phân tích các đặc điểm của tiếng Việt bao gồm ngữ âm, thanh điệu,… trong bài toán nhận dạng và tổng hợp tiếng Việt nói. Một đặc điểm rất quan trọng của tiếng Việt là thanh điệu tính, nghĩa là tiếng Việt bao gồm một hệ thống sáu thanh điệu khác nhau. Năm 2001, Nguyễn Quốc Cường và cộng sự [3] đã sử dụng tần số cơ bản F0 làm tham số sử dụng cho mô hình Markov ẩn để nhận dạng thanh điệu của từ phát âm rời rạc với độ chính xác 94%. Năm 2008, Vũ Tất Thắng và cộng sự [4] đề xuất phương pháp nhận dạng thanh điệu sử dụng mạng nơ ron perceptron. Bài toán phức tạp nhất đó là nhận dạng tự động tiếng Việt nói từ vựng lớn. Năm 2005, Vũ Tất Thắng và cộng sự [5] đã thử nghiệm với tập các âm vị không bao gồm thông tin thanh điệu, các bộ tham số MFCC (Mel Frequency Cepstral Coeff cient) và PLP (Perceptual Linear Prediction) được sử dụng để mô hình hóa mô hình âm học của các âm vị với độ chính xác nhận dạng đạt được 86,06%. Năm 2010, TS. Nguyễn Hồng Quang và cộng sự [6] đã tích hợp thông tin thanh điệu cho các âm vị và kết quả nhận dạng đạt được là rất khả quan.

Các nghiên cứu trên chưa đề cập đến ứng dụng các mô hình âm học tiên tiến cũng như ảnh hưởng của trọng số mô hình ngôn ngữ đến kết quả nhận dạng tiếng Việt nói. Trong bài báo này, bộ công cụ Kaldi được chọn vì hỗ trợ hiệu quả những vấn đề trên. Và hơn thế nữa, Kaldi cho chất lượng nhận dạng cao hơn các bộ công cụ nhận dạng tiếng nói khác như HTK, Sphinx hay Alize… Christian Gaida và cộng sự [7] đã đánh giá trên quy mô lớn các bộ công cụ nhận dạng tiếng nói mã nguồn mở bao gồm bộ công cụ HTK (bộ giải mã HDecode), Julius, PocketSphinx, Sphinx-4 và Kaldi. Họ điều chỉnh các hệ thống và chạy thử nghiệm trên tiếng Đức và tiếng Anh. Kết quả thí nghiệm cho thấy Kaldi chạy nhanh hơn so với tất cả các bộ công cụ nhận dạng khác. Kaldi huấn luyện và giải mã theo kỹ thuật đường ống bao gồm các kỹ thuật cao cấp nhất, điều này cho phép hệ thống đạt kết quả tốt nhất trong thời gian ngắn. Kết quả chạy thử nghiệm được mô tả ở hình 1.

Tác giả liên hệ: Nguyễn Thị Thanh,
email: nguyenthanhktdt@gmail.com
Đến tòa soạn: 11/10/2016, chỉnh sửa: 01/01/2016, chấp nhận đăng: 09/01/2017.

Thời gian cho việc thiết lập, chuẩn bị, chạy và tối ưu hóa cho các bộ công cụ lớn nhất với HTK, ít hơn là Sphinx và ít nhất là Kaldi. Bộ công cụ nhận dạng họ Sphinx (PocketSphinx và Sphinx-4) không bao gồm tất cả các kỹ thuật tích hợp trong một như Kaldi, do đó dẫn đến độ chính xác thấp hơn. HTK là bộ công cụ khó nhất, mặc dù các kết quả thu được tương tự với Sphinx, tuy nhiên thiết lập hệ thống cần tốn thời gian. So sánh với các bộ nhận dạng khác, hiệu năng vượt trội của Kaldi được xem như là cuộc cách mạng trong công nghệ nhận dạng tiếng nói mã nguồn mở.

recognizer	VM1	WSJ1
HDecode v3.4.1	22.9	19.8
Julius v4.3	27.2	23.1
pocketsphinx v0.8	23.9	21.4
Sphinx-4	26.9	22.7
Kaldi	12.7	6.5

Hình 1. Tỷ lệ lỗi nhận dạng từ WER trên tập kiểm thử VM1 (tiếng Đức) và tập WSJ1 November '93 (tiếng Anh)

Hiện tại đã có một số nghiên cứu về nhận dạng tiếng Việt nói, tuy nhiên đa phần mới chỉ sử dụng bộ công cụ HTK [6]. Do vậy mục tiêu nghiên cứu của bài báo là xây dựng bộ công cụ nhận dạng tiếng Việt nói sử dụng bộ công cụ Kaldi, thử nghiệm các kỹ thuật tiên tiến trong Kaldi để đánh giá khả năng của Kaldi với tiếng Việt.

Phần tiếp theo của bài báo sẽ giới thiệu bộ công cụ nhận dạng tiếng nói Kaldi, phần III mô tả phương pháp xây dựng bộ nhận dạng tiếng Việt nói sử dụng bộ công cụ Kaldi và các giải pháp tối ưu cho hệ thống. Phần IV là kết luận và hướng phát triển tiếp theo.

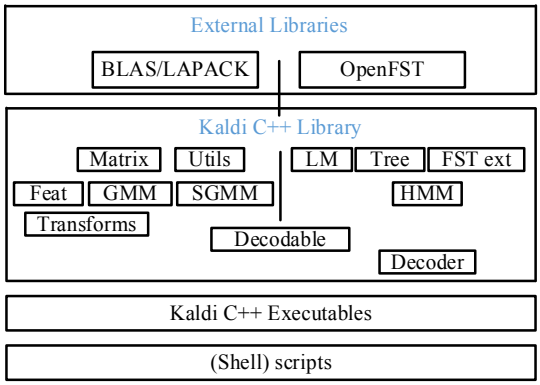
II. GIỚI THIỆU BỘ CÔNG CỤ NHẬN DẠNG TIẾNG NÓI KALDI

A. Giới thiệu bộ công cụ Kaldi

Kaldi là bộ công cụ nhận dạng tiếng nói được viết bằng C++, được cấp phép theo giấy phép Apache 2.0 [8]. Kaldi được thiết kế cho các nhà nghiên cứu nhận dạng tiếng nói. So với các bộ công cụ nhận dạng tiếng nói khác, Kaldi tương tự như HTK về mục đích và phạm vi. Mục đích là để có mã nguồn hiện đại và linh hoạt được viết bằng C++, có thể dễ dàng sửa đổi và mở rộng. Kaldi có các tính năng quan trọng: hỗ trợ số học tuyến tính mở rộng gồm một thư viện ma trận với gói BLAS và các chương trình con LAPACK; thiết kế mở rộng, bộ giải mã có thể làm việc với các mô hình khác, chẳng hạn như mạng nơ ron; giấy phép mở cho phép sử dụng thuận tiện.

B. Cấu trúc bộ công cụ Kaldi

Kaldi gồm một thư viện, các bộ chương trình dòng lệnh và kịch bản cho các mô hình âm học. Kaldi triển khai nhiều bộ giải mã để đánh giá các mô hình âm học, sử dụng huấn luyện Viterbi cho việc ước lượng mô hình âm học. Chỉ trong trường hợp đặc biệt của huấn luyện discriminative thích nghi người nói thì được mở rộng sử dụng thuật toán Baum-Welsh. Các kiến trúc của bộ công cụ Kaldi có thể được tách thành các thư viện Kaldi và các kịch bản huấn luyện. Các kịch bản này truy cập vào các hàm của thư viện Kaldi qua các chương trình dòng lệnh. Thư viện Kaldi C++ được xây dựng dựa trên thư viện OpenFST [9]. Các hàm này có liên quan đến nhau và thường được nhóm trong một tên miền trong mã nguồn C++ mà tương ứng với một thư mục trên một hệ thống tập tin. Các ví dụ của tên miền và các thư mục được thể hiện trong hình 2.



Hình 2. Kiến trúc bộ công cụ Kaldi

Các mô-đun thư viện có thể được nhóm lại thành hai nửa riêng biệt, mỗi nửa phụ thuộc vào một trong các thư viện bên ngoài. Mô-đun *DecodableInterface* là cầu nối hai nửa này.

Kaldi thực thi bằng cách tải đầu vào từ các tập tin và lưu trữ kết quả tới các tập tin một lần nữa. Ngoài

ra, đầu ra của một chương trình Kaldi có thể được đưa vào lệnh kế tiếp sử dụng hệ thống đường ống (pipe). Thường có nhiều lựa chọn thay thế cho mỗi tác vụ nhận dạng tiếng nói sẽ được thể hiện trong danh sách các tập tin thực thi như sau:

- Tham số hóa tiếng nói
  - *apply-mfcc*
  - *compute-mfcc-feats*
  - *compute-plp-feats*
  - ...
- Biến đổi các tham số
  - *apply-cmvn*
  - *compute-cmvn-stats*
  - *fmpe-apply-transform [10]*
  - ...
- Các bộ giải mã
  - *gmm-latgen-faster*
  - *gmm- latgen-faster-parallel*
  - *gmm-latgen-biglm-faster*
  - ...
- Đánh giá và các tiện ích
  - *compute-wer*
  - *show-alignments*
  - ...

Ngoài ra Kaldi còn cung cấp kịch bản chuẩn hoặc các hàm thêm mới tiện ích. Các kịch bản được đặt tại thư mục */utils* và /steps được sử dụng trong kịch bản huấn luyện các công thức cho các dữ liệu khác nhau.

Bài báo này mô tả công thức huấn luyện sử dụng Kaldi cho tiếng Việt. Phần tiếp theo sẽ mô tả chi tiết quá trình này.

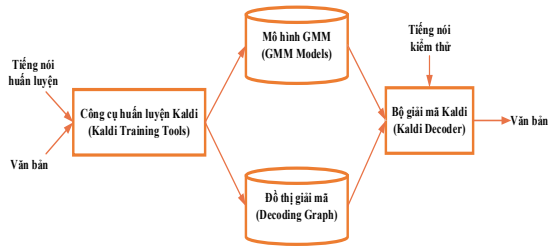
III. NHẬN DẠNG TIẾNG VIỆT NÓI SỬ DỤNG BỘ CÔNG CỤ KALDI

A. Mô hình hệ thống nhận dạng tiếng Việt nói với bộ công cụ Kaldi

Sơ đồ tổng quan của hệ thống nhận dạng tiếng Việt nói với bộ công cụ Kaldi được mô tả ở hình 3. Trong mô hình này, mô hình âm học (AM) là trái tim của nhận dạng tiếng nói. Các AM cho phép ước lượng xác suất P*(a/w;θ)*, giá trị này được sử dụng trong bộ nhận dạng tiếng nói theo phương trình (1).

$$W^* = argmax_w \left\{ \frac{P(a|W) * P(w)}{P(a)} \right\} = argmax_w \left\{ P(a|w) * P(w) \right\} \quad (1)$$

trong đó P(a) là xác suất của chuỗi âm học, nó được cố định cho một cách phát âm và không có vai trò xác định chuỗi từ; là xác suất chuỗi từ; P(a|W) là xác suất chuỗi âm thanh cho biết chuỗi từ W. Từ công thức (1) ta có thể phân chia việc giải mã từ thành các thành phần ngôn ngữ học riêng biệt và thực hiện song song, thành phần đầu tiên là mô hình âm học của tiếng nói, thành phần thứ hai là các mô hình ngôn ngữ.



Hình 3. Mô hình nhận dạng tiếng nói với bộ công cụ Kaldi

Mô hình âm học chỉ có một phần thông tin có sẵn cho tham số huấn luyện mô hình âm học θ vì các văn bản phiên âm tương ứng không có liên kết về mặt thời gian. Thông tin ẩn của từ (thời gian) liên kết trong một cách phát âm tạo ra mô hình huấn luyện âm học với nhiều thách thức. Bộ công cụ nhận dạng tiếng nói hiện đại sử dụng mô hình Markov ẩn cho mô hình bất định giữa các tham số âm học phiên âm tương ứng.

B. Cơ sở dữ liệu tiếng Việt nói

Hiện nay đã có một số nghiên cứu xây dựng cơ sở dữ liệu tiếng Việt nói [3][4], tuy nhiên những cơ sở dữ liệu này lại cho phép truy cập miễn phí. Vì vậy chúng tôi đã tiến hành xây dựng một cơ sở dữ liệu tiếng nói mới. Cơ sở dữ liệu thu âm bởi 35 người (16 nam và 19 nữ) có độ tuổi từ 17 - 29 tuổi,

trong đó giọng nói miền Bắc gồm: giọng Hà Nội, Hà Tây, Hưng Yên, Hải Dương.

Dữ liệu được ghi về các chủ đề gồm: đời sống, kinh doanh, khoa học, ô tô - xe máy, pháp luật. Tiếng nói được ghi âm ở dạng đọc, được thu trong môi trường phòng làm việc bình thường, thu âm ở tần số lấy mẫu 16kHz, 16 bits cho một mẫu, ở chế độ mono. Dữ liệu được ghi vào f le WAV. Dữ liệu được chia thành hai phần: một phần để huấn luyện mô hình và một phần để thử nghiệm. Thông tin chi tiết về dữ liệu được mô tả ở bảng 1.

Bảng 1. Cơ sở dữ liệu tiếng Việt nói

Tập dữ liệu	Giới tính người nói		Bản ghi âm (giờ)	Tổng số câu
	Nam	Nữ		
Huấn luyện	12	15	6	3.375
Kiểm thử	4	4	2	1.000
Tổng	16	19	8	4.375

C. Dữ liệu văn bản

Dữ liệu văn bản được sử dụng để tạo mô hình ngôn ngữ thống kê bao gồm 4 triệu câu với 90 triệu âm tiết thu thập từ các tài liệu điện tử tiếng Việt [6]. Các ký tự được chuyển đổi sang mã văn bản BKTC (Bach Khoa Text Code). Độ hỗn loạn thông tin (perplexity) của mô hình ngôn ngữ (LM) bigram và trigram là 108.57 và 62.43. Sử dụng bộ công cụ SRILM trên dữ liệu văn bản để tạo ra mô hình ngôn ngữ trong định dạng ARPA. Mô hình ngôn ngữ bigram chứa 8925 unigrams và 3,742,980 bigram. Mô hình ngôn ngữ trigram có tất cả gram trong mô hình bigram và 11,593,319 trigram. Các tập tin được sử dụng để tạo mô hình ngôn ngữ trong định dạng FST.

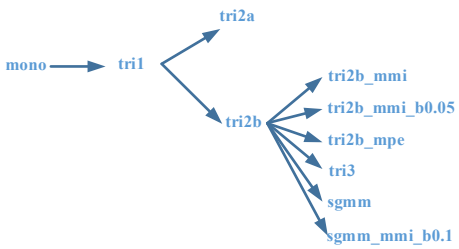
D. Kịch bản mô hình âm học

Các bản ghi và phiên âm của chúng từ tập dữ liệu huấn luyện được sử dụng cho mô hình âm học. Các mô hình âm học được đánh giá trên tập kiểm thử. Việc giải mã tiếng nói trong tập kiểm thử luôn được thực hiện với các tham số tương tự nhau, do đó các mô hình âm học khác nhau có thể so sánh với nhau. Bảng 2 liệt kê các mô hình âm học được huấn luyện trong kịch bản. Một AM cao cấp thường được khởi tạo bởi các liên kết âm thanh (tương ứng với các liên kết tham số âm học) sử dụng một AM đơn giản hơn.

Các phương pháp được sử dụng được liệt kê trong hình 4 với hệ thống phân cấp của chúng. Các hệ phân cấp cho thấy một phương pháp cao cấp điển hình là tái sử dụng giá trị huấn luyện ban đầu từ các AM đơn giản.

Đầu tiên, một mô hình mono-phone được huấn luyện sử dụng bộ tham số MFCC (Mel Frequency Cepstral Coeff cient) cùng với các tham số Δ và ΔΔ. Các vectơ tham số được xếp vào các trạng thái HMM sử dụng các phiên âm của tiếng nói. Sau đó, chúng huấn luyện lại mô hình triphone (tri1a). Một phần của quá trình kết thúc bởi huấn luyện mô hình MFCC + Δ + ΔΔ triphone (tri2a). Một phần khác, phần thứ hai của quá trình thay vì chuyển hóa Δ + ΔΔ, sử dụng LDA + MLTT để huấn luyện mô hình âm học (tri2b). Sử dụng mô hình thứ ba tri2b được huấn luyện Discriminative (hay còn gọi là mô hình có điều kiện) và sử dụng LDA + MLTT + SAT để huấn luyện mô hình tri3b sử dụng các phương pháp:

- MMI (Maximum Mutual Information). [11]
- BMMI (Boosted Maximum Mutual Information). [12]
- MPE (Minimum Phone Error). [13]
- SAT (speaker adaptive training). [14]



Hình 4. Hệ thống phân cấp các mô hình âm học được huấn luyện

Bảng 2. Các phương thức huấn luyện của hệ thống

Phương thức huấn luyện	Mô tả
Monophone	Mono
Triphone	Tri1
Δ + ΔΔ	Tri2a
LDA + MLLT	Tri2b
LDA + MLLT + MMI	Tri2b_mmi
LDA + MLLT + bMMI	Tri2b_mmi_b0.05
MPE	Tri2b_mpe
LDA + MLLT + SAT	Tri3
SGMM	Sgmm
SGMM + bMMI	Sgmm_mmi_b0.1

E. Mô hình GMM

Kaldi hỗ trợ GMM [15] với cấu trúc hiệp phương sai chéo và đầy đủ. Thay vì thể hiện các hàm mật độ Gauss riêng biệt, Kaldi thực hiện trực tiếp một lớp GMM được tham số hóa bởi các tham số tự nhiên. Các lớp GMM cũng được lưu trữ các số hạng không đổi trong tính toán xác suất, bao gồm các số hạng không phụ thuộc vào các vectơ dữ liệu. Việc thực thi như vậy là phù hợp cho hiệu quả tính toán tích vô hướng đơn giản (dot-product).

Một mô hình GMM biểu diễn các tham số như tổng các trọng số của nhiều Gauss phân tán. Mỗi trạng thái Gauss có: Mean (μi), hiệp phương sai (Σi), trọng số (Wi).Trong quá trình huấn luyện, hệ thống học về những dữ liệu mà nó sử dụng để đưa ra quyết định. Một tập hợp các tham số được thu thập từ một người nói (hoặc ngôn ngữ hoặc phương ngữ).

Thay vì huấn luyện mô hình người nói chỉ dựa trên dữ liệu người nói, mô hình GMM điều chỉnh mô hình phổ nền UBM (Universal Background Model) với người nói, tận dụng lợi thế của tất cả các dữ liệu, thích ứng MAP (Maximum A Posteriori): mỗi một Gaussian là một trọng số kết hợp của UBM và người nói. Trọng số người nói nhiều nếu ta có nhiều dữ liệu hơn: $\mu_i = \alpha E_i(x) + (1-\alpha)\mu_i$,với $\alpha = n/(n+16)$.

Các tham số thông thường MFCC có thể sử dụng nhiều chiều hơn (20 + delta). Mô hình phổ nền UBM: 512-2048 mixture, GMM của người nói: 64-256 mixture, thường được kết hợp đặc biệt với các phân lớp khác trong một mixture-of-experts.

F. Xây dựng đồ thị giải mã

Một đồ thị giải mã được biểu diễn như là một đối tượng OpenFst. Nó lưu giữ tất cả các thông tin mô hình ngôn ngữ và một phần thông tin của mô hình âm học. Đồ thị giải mã là cần thiết cho công việc giải mã với các bộ giải mã Kaldi [16]. Trong bài báo xây dựng đồ thị HCLG sử dụng chuẩn OpenFst được thực thi trong các tiện ích Kaldi. Ta thiết kế kịch bản để chúng tự động cập nhật các mô hình ngôn ngữ, mô hình âm học và tạo ra tất cả các tập tin cần thiết cho việc giải mã.

Các kịch bản yêu cầu để xây dựng HCLG:

- Mô hình ngôn ngữ (LM);
- Mô hình âm học (AM);
- Các cây quyết định âm vị âm học;
- Từ điển phiên âm.

Ngoài việc xây dựng HCLG, kịch bản cũng sao chép các tập tin cần thiết cho việc giải mã từ mô hình âm học và đồ thị HCLG đến một thư mục. Các tập tin sau là cần thiết cho việc giải mã:

- Đồ thị giải mã HCLG;
- Mô hình âm học;
- Một ma trận định nghĩa các tham số biến đổi;
- Một tập tin cấu hình cho các tham số tiếng nói và các tham số biến đổi với các thiết lập tương tự được sử dụng cho huấn luyện AM (mô hình âm học);
- Một bảng ký tự các từ (WST - Word Symbol Table). Bảng này là một tập tin chứa ánh xạ giữa các nhãn (label) với các số nguyên.

G. Bộ giải mã Kaldi

Trong bộ công cụ Kaldi [17] không có bộ giải mã "chuẩn" đơn lẻ, hoặc một giao diện cố định. Hiện tại có hai bộ giải mã có sẵn: SimpleDecoder, FastDecder và cũng có các phiên bản lattice-generating. "Decoder" có nghĩa là các mã bên trong của bộ giải mã, có các dòng lệnh chương

trình, các gói bộ giải mã có thể giải mã các loại mô hình cụ thể (ví dụ GMM) hoặc với các điều kiện cụ thể đặc biệt (ví dụ đa lớp fMLLR). Ví dụ về các chương trình dòng giải mã: gmm-decode-simple, gmm-decode-faster, gmm-decode-kaldi và gmm-decode-faster-fmllr.

H. Thiết lập các tham số giải mã

Đầu tiên, Δ + ΔΔ gấp 3 lần của 13 tham số MFCC bằng cách tính đạo hàm lần 1 và lần 2 từ hệ số MFCC. Việc tính toán hệ số MFCC với việc xử lý đạo hàm 39 tham số trên một khung.

Thứ hai, sự kết hợp của LDA và MLLT được tính toán từ 9 khung ghép gồm 13 tham số MFCC. Phạm vi cửa sổ mặc định của 9 khung lấy 1 khung hiện tại, 4 khung bên trái và 4 khung bên phải. Các phép biến đổi tham số LDA và MLLT đạt được cải thiện đáng kể so với biến đổi Δ + ΔΔ.

Sử dụng mô hình âm học được huấn luyện được mô tả ở trên cho giải mã các phiên âm từ tập dữ liệu kiểm thử. Đối với mỗi mô hình âm học, sử dụng cùng một phương thức tham số hóa tiếng nói và phép biến đổi các tham số cho việc huấn luyện mô hình âm học, bài viết thử nghiệm với tất cả các mô hình âm học được huấn luyện với cả mô hình ngôn ngữ zerogram và bigram.

Mô hình ngôn ngữ zerogram và bigram mặc định được xây dựng từ các phép biến đổi trực giao. Mô hình ngôn ngữ bigram được ước lượng từ các phép biến đổi dữ liệu huấn luyện. Do đó, trong tập kiểm thử xuất hiện các từ chưa biết, được gọi là "Out of Vocabulary Word - OOV". Các zerogram được trích chọn từ các phép biến đổi tập kiểm thử. Zerogram là một danh sách các từ với xác suất phân bố đều, vì vậy nó giúp giải mã chỉ bằng việc giới hạn kích thước bộ từ vựng. Các mô hình ngôn ngữ bigram chứa 1075 unigram và 3517 bigram cho tiếng Việt. Mô hình ngôn ngữ zerogram được giới hạn 1076 từ tiếng Việt.

Các tham số nhận dạng tiếng nói được thiết lập giá trị mặc định; các trường hợp ngoại lệ là các tham số giải mã: beam=12.0, lattice-beam=6.0, max-active-states=14000 và LMW (các trọng số mô hình ngôn ngữ - Language Model Weight). Tham số LMW thiết lập trọng số của LM, tức là nó quy định có bao nhiêu LM (mô hình ngôn ngữ) được sử dụng cho mô hình âm học trong việc giải mã. Giá trị LMW được ước tính trên tập phát triển và các giá trị tốt nhất được sử dụng cho giải mã trên tập dữ liệu kiểm thử.

Các bộ giải mã GMM-latgen-faster được sử dụng cho việc đánh giá dữ liệu thử nghiệm. Nó tạo ra một mạng liên kết các cấp độ từ cho mỗi phiên âm và một giả thuyết tốt nhất được trích chọn từ các mạng được giải mã và được đánh giá bởi WER (Word Error Rate) và SER (Sentence Error Rate).

IV. KẾT QUẢ THỬ NGHIỆM

Mô hình âm học mono, tri1, tri2a, tri2b, được huấn luyện generative. Mô hình tri2b_mmi, tri2b_mmi_b0.05, tri2b_mpe, tri3, sgmm, sgmm_mmi_b0.1 được huấn luyện discriminatively trong bốn vòng lặp. Các mô hình discriminative mang lại kết quả tốt hơn mô hình generative thể hiện trong hình 4.

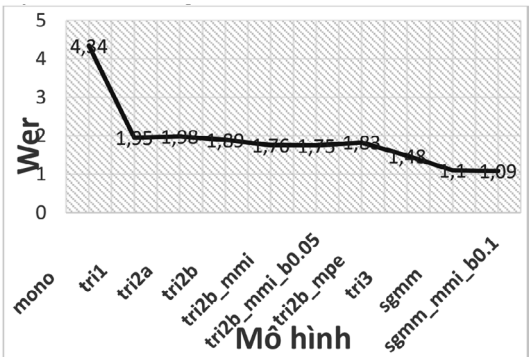
A. Kết quả thực hiện với các mô hình huấn luyện

Phần này trình bày các kết quả thử nghiệm hệ thống nhận dạng tiếng Việt nói với phương pháp huấn luyện âm học khác nhau. Bảng 3 biểu diễn kết quả các mô hình âm học.

Bảng 3. WER và SER cho các phương pháp huấn luyện

Model	% WER	% SER
mono	4.34	53.4
tri1	1.95	37.4
tri2a	1.98	37.6
tri2b	1.89	36.2
tri2b_mmi	1.76	34
tri2b_mmi_b0.05	1.75	33.8
tri2b_mpe	1.83	35.5
tri3	1.48	30.4
sgmm	1.1	23.7
sgmm_mmi_b0.1	1.09	23.5

Biểu đồ WER qua các mô hình huấn luyện thể hiện trong hình 5.



Hình 5. Biểu đồ WER thể hiện qua các mô hình huấn luyện

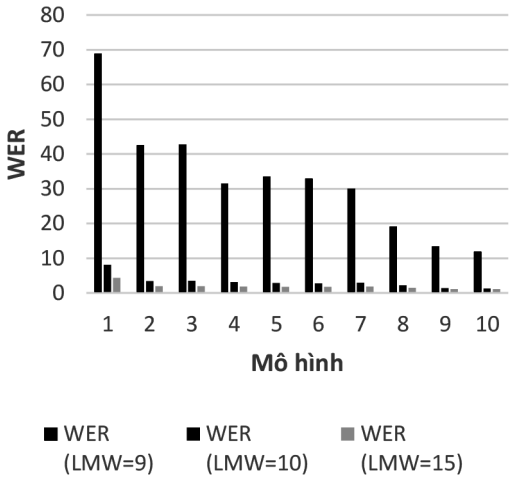
Kết quả cho thấy các phương pháp huấn luyện discriminative vượt trội so với các mô hình âm học generative, tham số LDA + MLTT cũng hiệu quả hơn việc sử dụng tham số Δ + ΔΔ. Mặt khác, có những sự khác biệt tinh tế giữa 3 mô hình âm học (tri3, sgmm, sgmm_mmi_b0.1) được huấn luyện discriminative về hiệu suất.

B. Kết quả thực hiện với các trọng số mô hình ngôn ngữ khác nhau

Thử nghiệm với LMW lần lượt bằng 9, 10 và 15. Kết quả được mô tả ở bảng 4 và hình 6.

Bảng 4: Bảng kết quả với các trọng số mô hình ngôn ngữ khác nhau

Mô hình	WER LMW=9	WER LMW=10	WER LMW=15
mono	68.84	8.09	4.34
tri1	42.49	3.42	1.95
tri2a	42.76	3.55	1.98
tri2b	31.55	3.14	1.89
tri2b_mmi	33.51	2.87	1.76
tri2b_mmi_b0.05	32.92	2.81	1.75
tri2b_mpe	30.1	2.96	1.83
tri3	19.07	2.22	1.48
sgmm2	13.4	1.44	1.16
sgmm2_mmi_b0.1	11.94	1.35	1.15



Hình 6. Biểu đồ WER với các tham số LMW khác nhau

Kết quả cho thấy với tham số LMW = 15 cho kết quả vượt trội so với LMW = 9. Như vậy, việc chọn lựa một trọng số phù hợp cho mô hình ngôn ngữ cũng là một trong các tham số quan trọng của hệ thống nhận dạng tiếng Việt nói.

V. KẾT LUẬN

Bài báo này đã mô tả phương pháp xây dựng hệ thống nhận dạng tiếng Việt nói sử dụng bộ công cụ Kaldi. Chúng tôi đã thử nghiệm các phương pháp huấn luyện mô hình âm học khác nhau được hỗ trợ bởi Kaldi. Các trọng số của mô hình ngôn ngữ cũng được xem xét và đánh giá. Các thử nghiệm cho thấy bộ công cụ Kaldi cho kết quả nhận dạng rất tốt với tiếng Việt nói. Ngoài ra trọng số của mô hình ngôn ngữ là một tham số quan trọng khi xây dựng hệ thống.

TÀI LIỆU THAM KHẢO

[1] Đặng Ngọc Đức, "Mạng nơ ron và mô hình Markov ẩn trong nhận dạng tiếng Việt nói", Luận văn tiến sĩ, Đại học Quốc Gia Hà Nội, 2003.

[2] Bạch Hưng Khang, "Tổng hợp và nhận dạng tiếng Việt", Viện Công nghệ thông tin, Viện Hàn lâm và Khoa học Việt Nam, 2004.

[3] Nguyen Quoc Cuong, Pham Thi Ngoc and Castelli, E. "Shape vector characterization of Vietnamese tones and application to automatic recognition". Automatic Speech Recognition and Understanding (ASRU), Italy, 2001. 437-440.

[4] Vu, Tat Thang, Khanh Nguyen and Le, Son Hai and Luong, Mai Chi. "Vietnamese tone recognition based on multi-layer perceptron network." Conference of Oriental Chapter of the International Coordinating Committee on Speech Database and Speech I/O System. Kyoto, 2008. 253-256.

[5] Vu, Thang Tat and Nguyen, Dung Tien and Luong, Mai Chi and Hosom, John Paul. "Vietnamese large vocabulary continuous speech recognition" INTERSPEECH. Lisbon, 2005. 1172-1175.

[6] Nguyen Hong Quang, Trinh Van Loan, Le The Dat, Automatic Speech Recognition for Vietnamese using HTK system, IEEE-RIVF 2010, Ha noi, November, 2010.

[7] Christian Gaida, Patrick Lange, Rico Petrick, Patrick Proba, Ahmed Malatawy, and David Suendermann-Oeft, "Comparing Open-Source Speech Recognition Toolkits".

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlcek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely, "*The Kaldi Speech Recognition Toolkit*".

[9] KyleGorman, http://www.openfst.org/twiki/bin/view/FST/WebHome, 2016.

[10] Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau & Geoffrey Zweig, "fMPE: Discriminatively Trained Features for Speech Recognition," ICASSP 2005.

[11] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang and Sanjeev Khudanpur "*Purely sequence-trained neural networks for ASR based on lattice-free MMI"*, Interspeech 2016.

[12] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon & Karthik Visweswariah, *"Boosted MMI for Model and Feature Space Discriminative Training"*, ICASSP 2008.

[13] Daniel Povey & Brian Kingsbury, "Evaluation of Proposed Modif cations to MPE for Large Scale Discriminative Training", ICASSP 2007.

[14] Yajie Miao, Hao Zhang, Florian Metze Language Technologies Institute, "*Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models*", School of Computer Science, Carnegie Mellon University Pittsburgh, PA, USA.

[15] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondˇrej Glembek, Nagendra Kumar Goel, Martin Karaf´at, Ariya Rastrow, Richard C. Rose, Petr Schwarz, Samuel Thomas**,** *"Subspace gaussian mixture models for speech recognition"*.

[16] Daniel Povey and Partner "http://kaldi-asr.org/doc/graph.html" Generated on Wed Aug 10 2016 for Kaldi by Doxygen 1.8.1.2 .

[17] Daniel Povey and Partner http://kaldi-asr.org/doc/decoders.html Generated on Wed Aug 10 2016 for Kaldi by Doxygen 1.8.1.2 .

[18] Tuan, Nguyen and Hai Quan, Vu. "Advances in Acoustic Modeling for Vietnamese LVCSR" Asian Language Processing. Singapore: IEEE, 2009. 280-284.

THE VIETNAMESE SPEECH RECOGINITION USING KALDI TOOLKIT

Abstract: Speech recognition has been increasingly applied in various f elds such as automatic switchboards, security, searching by voice… however the quality of recognition is the problem of utmost concern. The Kaldi toolkit is a new tool developed in 2009. Kaldi was introduced at a workshop held at Johns Hopkins University with the title "Low Development Cost, High Quality Speech Recognition for New Languages and Domains". This paper describes the Vietnamese speech recognition system built on Kaldi toolkit. The paper also evaluates quality of the system based on the evaluation the ratio of the WER on AMs (Acoustic models). The system has superior results compared the previous toolkit to Vietnamese speech.

Keywords: Speech recognition, the Vietnamese speech, Kaldi toolkit, Language models, Acoustic models, Pronounce dictionary.

Nguyễn Thị Thanh tốt nghiệp đại học năm 2013, tại Học viện Công nghệ Bưu chính Viễn thông. Hiện là học viên tại Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội. Lĩnh vực nghiên cứu: Xử lý tiếng nói.

Nguyễn Hồng Quang nhận học vị Tiến sĩ năm 2008. Hiện công tác tại Bộ môn Kỹ thuật máy tính, Viện Công nghệ thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội. Lĩnh vực nghiên cứu: Học máy, xử lý ảnh, âm thanh và tiếng nói.

Trịnh Văn Loan nhận học vị Phó Giáo sư năm 2009. Hiện công tác tại Bộ môn Kỹ thuật máy tính, Viện Công nghệ thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội. Lĩnh vực nghiên cứu: Tổng hợp, nhận dạng tiếng nói, Cải thiện chất lượng tín hiệu tiếng nói; Lượng giá và đánh giá chất lượng tiếng nói; Hệ nhúng.

Phạm Ngọc Hưng nhận bằng Thạc sĩ năm 2010. Hiện công tác tại Bộ môn Kỹ thuật máy tính, Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên. Lĩnh vực nghiên cứu: Nhận dạng tiếng nói, hệ thống nhúng.