

VIỆN CÔNG NGHỆ THÔNG TIN - TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI

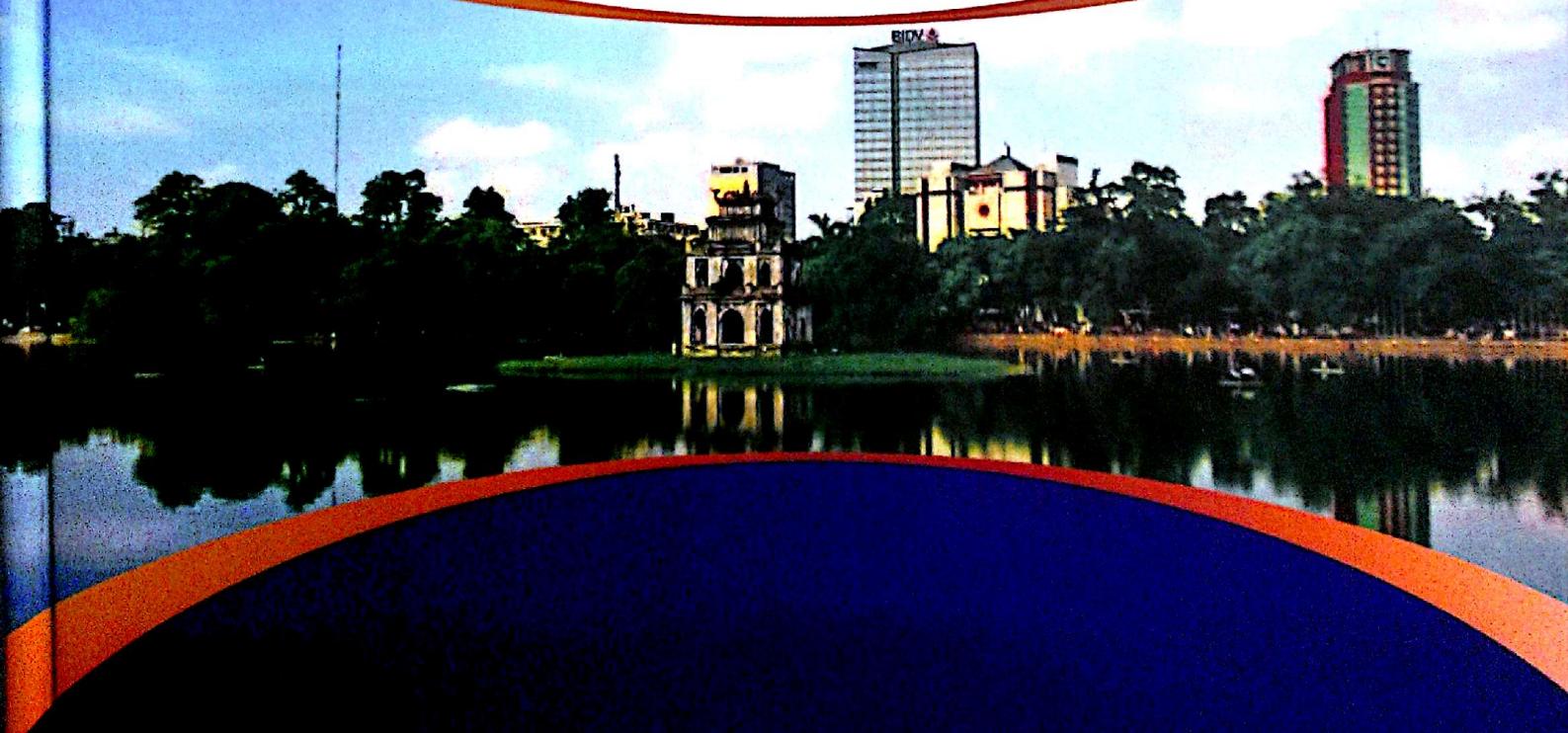
HỘI THẢO QUỐC GIA LẦN THỨ XIX

HÀ NỘI, NGÀY 1-2 THÁNG 10 NĂM 2016



MỘT SỐ VẤN ĐỀ CHỌN LỌC CỦA CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

CHỦ ĐỀ: GIÁO DỤC ĐIỆN TỬ
TOÀN VĂN CÁC BÁO CÁO



NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT

VIỆN CÔNG NGHỆ THÔNG TIN - TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI

HỘI THẢO QUỐC GIA LẦN THỨ XIX

HÀ NỘI, NGÀY 1-2 THÁNG 10 NĂM 2016



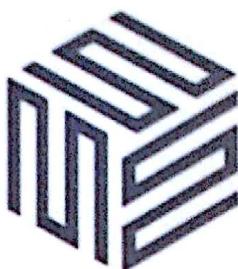
MỘT SỐ VĂN ĐỀ CHỌN LỌC CỦA CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Chủ đề: Giáo dục điện tử

TOÀN VĂN CÁC BÁO CÁO



NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT



**Hội thảo quốc gia lần thứ XIX
MỘT SÓ VẤN ĐỀ CHỌN LỌC CỦA
CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

Chủ đề: Giáo dục điện tử

Hà Nội, ngày 1-2/10/2016

BAN TỔ CHỨC

Đồng trưởng ban

GS. TS. Nguyễn Văn Minh
TS. Nguyễn Trường Thắng

Đại học Sư phạm Hà Nội
Viện Công nghệ thông tin

Thành viên

PGS. TS. Nguyễn Minh Thùy
PGS. TS. Nguyễn Văn Hiền
TS. Đặng Thành Trung
ThS. Nguyễn Thu Anh
TS. Phạm Thanh Giang
TS. Nguyễn Như Sơn
TS. Bùi Thị Thanh Quyên

Đại học Sư phạm Hà Nội
Đại học Sư phạm Hà Nội
Đại học Sư phạm Hà Nội
Viện Công nghệ thông tin
Viện Công nghệ thông tin
Viện Công nghệ thông tin
Viện Công nghệ thông tin

So sánh một số phương pháp phân lớp dùng cho định danh ba ngôn ngữ Việt - Anh - Pháp

Lê Trung Hiếu

Bộ môn Kỹ thuật máy tính,
Đại học Bách Khoa Hà Nội
Hà Nội, Việt Nam
Email: hieult.ktmt@gmail.com

Nguyễn Kim Khánh

Bộ môn Kỹ thuật máy tính,
Đại học Bách Khoa Hà Nội
Hà Nội, Việt Nam
Email: khanhnk @soict.hust.edu.vn

Trịnh Văn Loan

Bộ môn Kỹ thuật máy tính,
Đại học Bách Khoa Hà Nội
Hà Nội, Việt Nam
Email: loantv@soict.hust.edu.vn

Tóm tắt: Có nhiều phương pháp và mô hình khác nhau đã được nghiên cứu và áp dụng cho nhận dạng ngôn ngữ như mô hình GMM, HMM, SVM, ANN... Bài báo trình bày kết quả thử nghiệm nhận dạng ba ngôn ngữ Việt, Anh, Pháp sử dụng các bộ phân lớp SMO (Sequential Minimal Optimization), iBK, Multilayer Perceptron của Weka với các đặc trưng được OpenSMILE trích chọn. Số lượng các đặc trưng gồm 384 hệ số. Kết quả thử nghiệm cho thấy tỷ lệ nhận dạng tiếng Việt là cao nhất đạt 98.75 % với bộ phân lớp SMO, tiếng Pháp đạt cao nhất 93,5% với bộ phân lớp SMO và Multilayer Perceptron còn tiếng Anh đạt cao nhất 94,75% với bộ phân lớp Multilayer Perceptron.

Từ khóa: Định danh ngôn ngữ, tiếng Việt, tiếng Anh, tiếng Pháp, SVM, SMO, iBK, Multilayer Perceptron, Weka, tần số cơ bản

I. GIỚI THIỆU

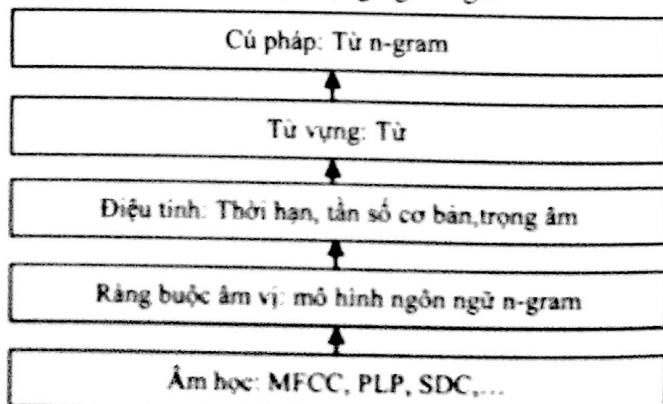
Định danh ngôn ngữ đóng vai trò quan trọng trong các hệ thống dịch, nhận dạng tự động. Bài báo sẽ trình bày các thử nghiệm sử dụng SVM (Support-Vector Machines) có so sánh với một số phương pháp phân lớp khác để định danh các ngôn ngữ Việt, Anh, Pháp theo phương thức phát âm mà không phụ thuộc vào nội dung. SVM là một phương pháp máy học tiên tiến đã được áp dụng khá phổ biến không chỉ trong các lĩnh vực khai phá dữ liệu mà còn trong lĩnh vực nhận dạng cho phép hệ thống đạt hiệu năng cao [1], [2], [3], [4], [5]. Các phần tiếp theo của bài báo được tổ chức như sau: phần II giới thiệu tổng quan về định danh ngôn ngữ, phần III trình bày các thử nghiệm nhận dạng với bộ công cụ Weka cho ba ngôn ngữ Việt, Anh, Pháp. Cuối cùng phần IV là kết luận và hướng phát triển.

II. TỔNG QUAN VỀ ĐỊNH DANH NGÔN NGỮ

Các ngôn ngữ khác nhau trên thế giới có các đặc trưng phân biệt và nhờ các đặc trưng này có thể tiến hành định danh các ngôn ngữ đó.

A. Đặc trưng ngôn ngữ

Con người là hệ thống định danh ngôn ngữ hoàn thiện nhất [6]. Trên thực tế, có một loạt các thông tin mà con người và máy móc có thể sử dụng để phân biệt các ngôn ngữ khác. Ở mức thấp, các đặc trưng của tiếng nói như thông tin âm học (acoustic), ngữ âm (phonetic), ràng buộc âm vị (phonotactic) và ngôn điệu (prosodic) được sử dụng rộng rãi trong các hệ thống nhận dạng ngôn ngữ tự động. Ở một mức cao hơn, sự khác biệt giữa các ngôn ngữ có thể được khai thác dựa trên hình vị học (morphology) và cú pháp câu (sentence syntax). Hình 1 [6] mô tả các mức khác biệt giữa các đặc trưng khác nhau của tiếng nói từ các đặc trưng ở mức thấp đến các đặc trưng ở mức cao để nhận dạng ngôn ngữ.



Hình 1. Các mức đặc trưng của ngôn ngữ

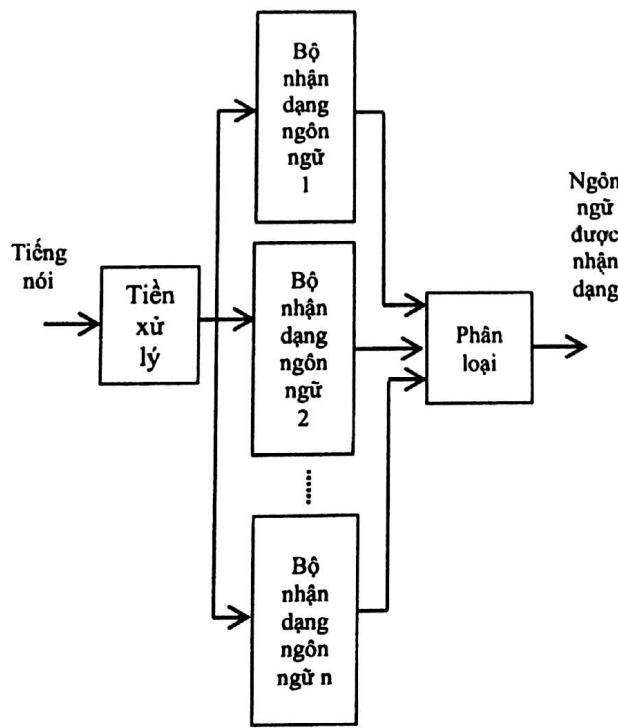
Về mặt âm học, có thể sử dụng các đặc trưng như MFCC (Mel-Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction), SDC (Shifted Delta Cepstrum). Về mặt ràng buộc âm vị có thể sử dụng mô hình ngôn ngữ n-gram [7] với n-gram là dãy gồm n phần tử đi với nhau của văn bản hoặc tiếng nói, phần tử có thể là âm vị, âm tiết, chữ hoặc từ. Với $n=1$ ta có unigram, $n=2$ có bigram, và $n=3$ là trigram.

B. Mô hình định danh ngôn ngữ

Các mô hình định danh ngôn ngữ có thể được phân loại theo hai trường hợp: mô hình định danh ngôn ngữ tường minh và mô hình định danh ngôn ngữ ẩn.

1) Mô hình định danh ngôn ngữ ngôn ngữ tường minh

Mô hình định danh ngôn ngữ tường minh được thể hiện trên hình 2 [6]. Nguyên tắc hoạt động của mô hình là dữ liệu tiếng nói ban đầu sẽ được đưa qua bộ tiền xử lý, sau đó dữ liệu của các ngôn ngữ khác nhau đã được xác định sẽ đưa vào các bộ nhận dạng cụ thể. Tại các bộ nhận dạng ngôn ngữ, thông tin sẽ được xử lý và đưa ra bộ phân loại. Cuối cùng hệ thống sẽ đưa ra kết quả ngôn ngữ được nhận dạng.

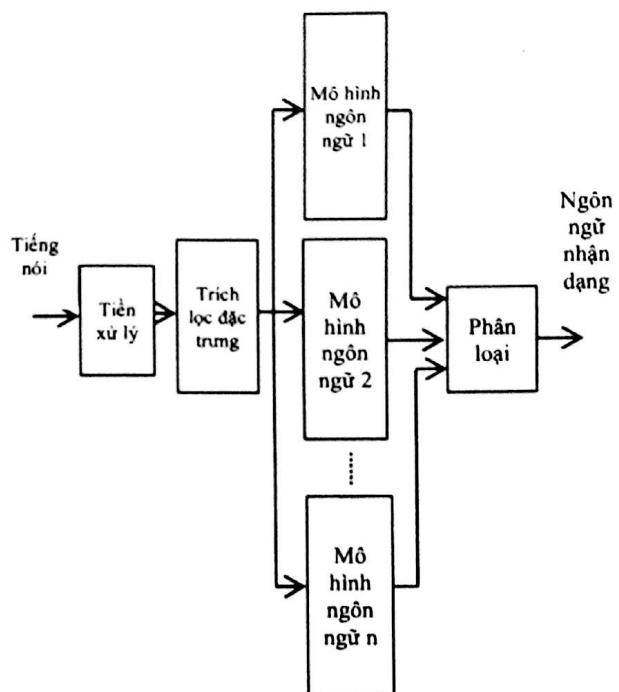


Đã có nhiều kết quả nghiên cứu ứng dụng mô hình định danh ngôn ngữ tường minh được công bố như [8], [9], [10], [11], [12].

2) Mô hình định danh ngôn ngữ ẩn

Mô hình định danh ngôn ngữ ẩn được trình bày

trên hình 3 [6]. Với mô hình này, dữ liệu tiếng nói ban đầu sẽ được đưa qua bộ tiền xử lý và đưa vào bộ trích chọn đặc trưng để lấy ra các đặc trưng của từng ngôn ngữ. Sau đó, các mô hình ngôn ngữ khác nhau sẽ nhận dạng để đưa ra kết quả ngôn ngữ được định danh.



Hình 3. Mô hình định danh ngôn ngữ ẩn

Các kết quả nghiên cứu ứng dụng mô hình định danh ngôn ngữ ẩn được công bố tại [13], [14], [15], [16].

Khác biệt giữa hai mô hình là: với mô hình định danh ngôn ngữ tường minh, việc trích chọn đặc trưng được thực hiện riêng cho từng ngôn ngữ, trong khi đó, mô hình định danh ngôn ngữ ẩn lại thực hiện trích chọn đặc trưng chung cho các ngôn ngữ.

C. Một số đặc trưng về mặt ngữ âm của tiếng Việt, Anh, Pháp

Các ngôn ngữ nói chung, ngôn ngữ Việt, Anh, Pháp nói riêng đều có các đặc trưng khác nhau về âm học, ràng buộc âm vị, từ vựng, ngữ pháp... để nhận biết và phân biệt các ngôn ngữ đó. Có thể đưa ra một số đặc trưng khác nhau nổi bật giữa ba ngôn ngữ tiếng Việt, tiếng Anh và tiếng Pháp như sau:

- Tiếng Anh và tiếng Pháp là các ngôn ngữ đa âm tiết trong khi đó tiếng Việt là ngôn ngữ đơn âm tiết.
- Tiếng Việt là ngôn ngữ có thanh điệu còn tiếng Anh và tiếng Pháp là ngôn ngữ không có thanh điệu. Vì vậy, đặc tính biến thiên tần số cơ bản là rất khác nhau giữa tiếng Việt với tiếng Anh và tiếng Pháp. Đây là một đặc trưng rất quan trọng

để có thể nhận biết tiếng Việt so với hai thứ tiếng còn lại.

- Tiếng Pháp có bốn nguyên âm mũi trong khi tiếng Anh không có nguyên âm mũi mà chỉ có ba phụ âm hưu thanh là các phụ âm mũi [13]. Tiếng Việt cũng không có nguyên âm mũi nhưng lại có bốn phụ âm hưu thanh là các phụ âm mũi [14].
- Về mặt đặc trưng âm vị, một số cụm âm vị phổ biến ở ngôn ngữ này lại không được sử dụng ở ngôn ngữ khác. Ví dụ, trong tiếng Anh, cụm âm vị /st/ là rất phổ biến, âm vị /i/ là đối lập với /i:/, trong khi đó với tiếng Việt và tiếng Pháp hai âm này chỉ là hai cách phát âm khác nhau của cùng âm vị /i/.

D. Tổng quan về định danh ba ngôn ngữ Việt, Anh, Pháp

Đã có nhiều nghiên cứu về định danh ngôn ngữ nói chung. Tuy nhiên, nghiên cứu định danh ngôn ngữ trong đó có tiếng Việt, Anh, Pháp nói riêng hay còn ở mức khiêm tốn. Zissman [15] đã dùng mô hình HMM và GMM để định danh ngôn ngữ. Bộ ngữ liệu được sử dụng là ngữ liệu OGI (Oregon Graduate Institute) [16] thu qua điện thoại cho 11 thứ tiếng: Anh, Pháp, Việt, Đức, Án Độ, Nhật, Hàn Quốc, Tây Ban Nha, Hindi, Tamil, Farsi. Kết quả định danh với tiếng Việt trung bình đạt 77,7% số câu nhận dạng đúng trên tổng số câu, tiếng Pháp trung bình đạt 74,37%, tiếng Anh trung bình đạt 71, 25%. Cùng với ngữ liệu OGI, Manchala và cộng sự [17] đã sử dụng GMM với MFCC và formant để nhận dạng. Kết quả trung bình đạt được khi dùng 8 thành phần Gauss: đối với tiếng Việt đạt 81,67%, tiếng Anh đạt 77,33%, tiếng Pháp đạt 76,67%; khi dùng 16 thành phần Gauss: tiếng Việt đạt 83%, tiếng Anh đạt 78,33%, tiếng Pháp đạt 78%; khi sử dụng 32 thành phần Gauss tỷ lệ nhận dạng tiếng Việt vẫn cao nhất (83%) so với tiếng Anh (79,67%) và tiếng Pháp (80%). Bằng cách dùng DNNs (Deep Neural Networks) với ngữ liệu NIST [18] lấy từ VOV cho 23 thứ tiếng trong đó có tiếng Việt, Anh, Pháp, Luciana Ferrer và các cộng sự [19] đã cải thiện kết quả nhận dạng từ 40% đến 70% so với GMM. Trong [20], Ana Montalvo và các cộng sự tiến hành nhận dạng 5 thứ tiếng: Anh, Pháp, Trung Quốc, Nga và Tây Ban Nha bằng cách dùng spectrogram, phổ Fourier và các thuộc tính của phổ để phát hiện tính tuần hoàn. Tỷ lệ lỗi trung bình lớn nhất đạt 16,8%. Để định danh tiếng Việt và tiếng Pháp, các tác giả [21] đã dùng mạng nơ-ron lan truyền ngược để phân lớp với tham số đặc trưng chỉ gồm thông tin về tần số cơ bản. Kết quả nhận dạng đúng đạt được là 90%. Có thể nói, phần lớn các nghiên cứu định danh ngôn ngữ trong đó có tiếng Việt, tiếng Anh, tiếng Pháp đã nêu trên chủ yếu do

các tác giả người nước ngoài thực hiện. Trong khi đó, nghiên cứu định danh tự động tiếng Việt, tiếng Anh, tiếng Pháp hầu như còn rất ít tác giả người Việt Nam thực hiện và công bố kết quả.

III. THỰC NGHIỆM NHẬN DẠNG VỚI BỘ CÔNG CỤ WEKA

Trong bài báo này, Weka là bộ công cụ đã được dùng thử nghiệm để nhận dạng ba ngôn ngữ Việt, Anh, Pháp. Bộ công cụ này gồm tập hợp các thuật giải học máy dùng cho khai phá dữ liệu do Đại học Waikato, New Zealand phát triển [22]. Weka hỗ trợ nhiều định dạng dữ liệu đầu vào dùng cho huấn luyện và thử nghiệm trong đó có file các tham số đặc trưng theo định dạng ARFF (Attribute-Relation File Format) [22] hoặc CSV được xuất bởi OpenSMILE. Với Weka, có thể sử dụng SVM để nhận dạng hoặc lựa chọn các phương pháp phân lớp khác nhau như SMO, iBK, Multilayer Perceptron.

A. Bộ ngữ liệu dùng cho định danh ba ngôn ngữ Việt, Anh, Pháp

Bộ ngữ liệu dùng để định danh ba ngôn ngữ Việt, Anh, Pháp được thu thập từ những người nói khác nhau gồm 50 giọng nam, 50 giọng nữ cho mỗi ngôn ngữ với tần số lấy mẫu là 16000 Hz, 16 bit cho một mẫu. Tổng thời lượng cho mỗi ngôn ngữ là 30 phút. Số liệu thống kê về bộ ngữ liệu này được trình bày ở bảng 1.

Bảng 1. SỐ LIỆU THÔNG KÊ NGỮ LIỆU

Ngôn ngữ	Số người nói	Số file (wav)	Tổng thời gian (phút)
Việt	25 nam	200	30
	25 nữ	200	
Anh	25 nam	200	30
	25 nữ	200	
Pháp	25 nam	200	30
	25 nữ	200	

B. Bộ công cụ thử nghiệm

Như trên đã nói, tiếng Việt, tiếng Anh, tiếng Pháp có những đặc trưng khác biệt về mặt ngữ âm. Về mặt tín hiệu, các đặc trưng về mặt ngữ âm này được thể hiện thông qua các thuộc tính của tín hiệu như phổ, tần số cơ bản, xác suất âm hưu thanh... Để thử nghiệm, các đặc trưng thông dụng nhất và quan trọng mang thông tin về ngôn điệu, phổ và chất lượng âm hưu thanh theo đề xuất trong [23] đã được sử dụng. Các đặc trưng này bao gồm 12 hệ số MFCC, tỷ lệ biến thiên qua trực không, cao độ, tỷ lệ hài trên nhiều. Tiếp theo, các đặc trưng kể trên lại được bổ sung thêm các hệ số delta và 12 đại lượng sau: trung bình, độ lệch chuẩn, mô men bậc 3, mô

men bậc 4, giá trị cực đại và cực tiểu, vị trí tương đối, dài giá trị và 2 hệ số hồi quy tuyến tính cùng với sai số trung bình bình phương tương ứng. Tổng cộng sẽ gồm có 384 hệ số.

Thử nghiệm nhận dạng ba ngôn ngữ Việt, Anh, Pháp sử dụng phương pháp đánh giá chéo với tỷ lệ dữ liệu huấn luyện và thử nghiệm là 90% và 10%. Người nói trong ngữ liệu huấn luyện khác với người nói trong ngữ liệu dùng cho nhận dạng. Bài báo sẽ trình bày kết quả thử nghiệm định danh ba ngôn ngữ Việt, Anh, Pháp bằng cách sử dụng SVM với thuật giải SMO, các bộ phân lớp iBK và Multilayer Perceptron. Đây là một trong các bộ phân lớp mà các nghiên cứu khác đã nêu ở mục II.D hầu như chưa sử dụng để định danh các ngôn ngữ trong đó có tiếng Anh, tiếng Pháp và tiếng Việt. Mặt khác, các bộ phân lớp dùng mạng nơ-ron nói chung tỏ ra có hiệu quả như kết quả của [19] đã so sánh với phân lớp dùng GMM.

1) Thử nghiệm định danh ba ngôn ngữ Việt, Anh, Pháp sử dụng SMO

SMO là thuật giải tối thiểu tuần tự. Đây là thuật giải cải tiến của SVM được tác giả John Platt đưa ra vào năm 1998, chạy nhanh hơn và dễ dàng mở rộng hơn so với thuật giải huấn luyện chuẩn SVM [24].

a) Thử nghiệm sử dụng SMO với các tham số đặc trưng đầu vào đầy đủ 384 hệ số

Kết quả thử nghiệm với số file tiếng nói nhận dạng đúng ngôn ngữ và nhận dạng nhầm ngôn ngữ được thể hiện ở ma trận sai nhầm trên bảng 2.

Bảng 2. MA TRẬN SAI NHẦM VỚI THỬ NGHIỆM DÙNG SMO ĐỊNH DANH 3 NGÔN NGỮ BAO GỒM ĐẦY ĐỦ CÁC HỆ SỐ

Ngôn ngữ	Việt	Anh	Pháp
Việt	395	3	2
Anh	5	371	24
Pháp	7	19	374

Bảng 2 và các bảng 3, 4, 5, 6 sau đây cho kết quả thử nghiệm nhận dạng đúng cho các ngôn ngữ với tổng cộng 10 lần thử nghiệm, mỗi lần có 40 file. Với bảng 2, tỷ lệ nhận dạng đúng với tiếng Việt đạt 98,75%, tiếng Anh đạt 92,75%, tiếng Pháp đạt 93,5%. Tỷ lệ nhận dạng đúng trung bình của phương pháp này là 95%.

b) Thử nghiệm sử dụng SMO với trường hợp không có thông tin liên quan tới tần số cơ bản (F0)

Với thử nghiệm này, trong tập tham số đặc trưng ban đầu ta loại bỏ toàn bộ các hệ số liên quan trực tiếp tới F0.

Kết quả thử nghiệm với các file tiếng nói nhận dạng đúng ngôn ngữ và nhầm ngôn ngữ được thể hiện trên bảng 3.

Bảng 3. MA TRẬN SAI NHẦM VỚI THỬ NGHIỆM DÙNG SMO ĐỊNH DANH 3 NGÔN NGỮ KHÔNG SỬ DỤNG F0

Ngôn ngữ	Việt	Anh	Pháp
Việt	390	4	6
Anh	3	371	24
Pháp	9	18	373

Với bảng 3, tỷ lệ nhận dạng đúng của tiếng Việt đạt 97,5%, tiếng Anh đạt 92,75% và tiếng Pháp đạt 93,25%. So với trường hợp trên, tỷ lệ nhận dạng đúng đổi với tiếng Việt giảm nhiều nhất là 1,25%, với tiếng Pháp giảm 0,25% còn với tiếng Anh tỷ lệ này không thay đổi.

c) Thử nghiệm sử dụng SMO với trường hợp chỉ có F0

Trong thử nghiệm này chỉ để lại các hệ số liên quan trực tiếp tới F0, các hệ số khác sẽ được loại bỏ.

Kết quả nhận dạng đúng và sai ngôn ngữ được thể hiện ở bảng 4.

Bảng 4. MA TRẬN SAI NHẦM VỚI THỬ NGHIỆM DÙNG SMO ĐỊNH DANH 3 NGÔN NGỮ CHỈ SỬ DỤNG F0

Ngôn ngữ	Việt	Anh	Pháp
Việt	309	42	49
Anh	55	223	112
Pháp	63	124	213

Bảng 4 cho thấy kết quả nhận dạng đúng của cả ba ngôn ngữ đều giảm mạnh, đặc biệt là tiếng Anh và tiếng Pháp. Tỷ lệ nhận dạng đúng của tiếng Việt còn 77,25%, tiếng Anh còn 55,75%, và tiếng Pháp còn 53,25%.

2) Thử nghiệm định danh ba ngôn ngữ sử dụng iBK với các tham số đặc trưng đầu vào đầy đủ 384 hệ số

iBK là bộ phân lớp k lâng giềng gần nhất (Lazy k-nearest-neighbor classifier) [22]. Kết quả thử nghiệm với phương pháp này được cho ở ma trận sai nhầm trên bảng 5.

Bảng 5. MA TRẬN SAI NHẦM VỚI THỬ NGHIỆM DÙNG iBK ĐỊNH DANH 3 NGÔN NGỮ BAO GỒM ĐẦY ĐỦ CÁC HỆ SỐ

Ngôn ngữ	Việt	Anh	Pháp
Việt	371	4	25
Anh	5	349	46
Pháp	10	23	367

Bảng 5 cho thấy kết quả thử nghiệm nhận dạng đúng cao nhất đối với tiếng Việt là 92,75%, thấp nhất là tiếng Anh với 87,25% và tiếng Pháp là 91,75%. Trung bình tỷ lệ nhận dạng đúng cho cả ba ngôn ngữ là 90,58, giảm 4,42% so với phương pháp SMO (sử dụng đầy 384 hệ số) đã nêu trên.

3) Thử nghiệm định danh ba ngôn ngữ sử dụng Multilayer Perceptron

Multilayer Perceptron là mạng nơ-ron nạp trước (feedforward artificial neural network) trong đó sử dụng thuật giải lan truyền ngược (backpropagation) để phân lớp. Với thử nghiệm dùng bộ phân lớp này, toàn bộ các đặc trưng đã được trích chọn đều được sử dụng, kết quả được trình bày trên bảng 6.

Bảng 6. MA TRẬN SAI NHÀM VỚI THỬ NGHIỆM DÙNG MULTILAYER PERCEPTRON

Ngôn ngữ	Việt	Anh	Pháp
Việt	393	2	5
Anh	2	379	19
Pháp	3	23	374

Bảng 6 cho thấy kết quả thử nghiệm nhận dạng đúng đối với tiếng Việt đạt 98,25%, tiếng Anh là 94,75% và tiếng Pháp đạt 93,5%. Trung bình tỷ lệ nhận dạng đúng cho cả ba ngôn ngữ là cao nhất so với các phương pháp đã thử nghiệm ở trên, tỷ lệ này đạt 95,5% tăng 0,5% so với phương pháp SM0 và tăng 4,92% so với phương pháp iBK.

C. Tổng hợp kết quả thử nghiệm

Bảng 7 là kết quả nhận dạng tiếng Việt, Anh, Pháp với các phương pháp khác nhau đã được nêu

Bảng 7. MA TRẬN SAI NHÀM TỔNG HỢP KẾT QUẢ THỬ NGHIỆM

Phương pháp	Tỷ lệ nhận dạng đúng cho từng ngôn ngữ			Tỷ lệ nhận dạng đúng trung bình
	Việt	Anh	Pháp	
Multilayer Perceptron	98,25%	94,75%	93,5%	95,5%
SM0	98,75%	92,75%	93,5%	95%
iBK	92,75	87,25	91,75%	90,58%

Nhận xét: bảng 7 là bảng tổng hợp kết quả định danh cho ba ngôn ngữ Việt, Anh, Pháp theo cả ba phương pháp với tỷ lệ nhận dạng trung bình từ cao xuống thấp.

- Nhìn chung cả ba phương pháp đã thử nghiệm cho định danh đều đạt kết quả trung bình nhận dạng đúng là trên 90% và cao nhất là phương pháp MultilayerPerceptron (đạt 95,5%). Điều này cho thấy các phương pháp đã thử nghiệm đều khá quan cho định danh ngôn ngữ.
- Xét riêng đối với từng ngôn ngữ: tiếng Việt được nhận dạng đúng với tỷ lệ cao nhất khi dùng phương pháp SM0 (98,75%), phương pháp MultilayerPerceptron cho tỷ lệ nhận dạng cao nhất đối với tiếng Anh (94,75%). Trong khi đó, đối với tiếng Pháp, hai phương pháp SM0 và MultilayerPerceptron cho tỷ lệ nhận dạng tương đương nhau (93,5%).

Thử nghiệm cũng chỉ ra vai trò của tần số cơ bản đối với tiếng Việt. Bảng 3 cho thấy, khi không sử dụng F0 thì tỷ lệ nhận dạng đúng của tiếng Việt bị giảm xuống còn 97,5% trong khi với tiếng Anh và tiếng Pháp tỷ lệ nhận dạng đúng hầu như không thay đổi. Việc chỉ sử dụng F0 vào nhận dạng với kết quả ở bảng IV cho thấy tiếng Việt đạt tỷ lệ nhận dạng đúng cũng khá cao (77,25%) trong khi tiếng Anh và tiếng Pháp chỉ đạt ở mức 55,75% và 53,25%.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã trình bày các kết quả thử nghiệm định danh tiếng Việt, tiếng Anh, tiếng Pháp bằng cách sử dụng bộ công cụ Weka với các phương pháp phân lớp khác nhau. Tỷ lệ trung bình định danh ba ngôn ngữ đạt cao nhất khi sử dụng bộ phân lớp Multilayer Perceptron và thấp nhất là phương pháp iBK. Ảnh hưởng của tần số cơ bản đến kết quả định danh của ba ngôn ngữ cũng đã được khảo sát. Khi loại bỏ các tham số liên quan trực tiếp đến tần số cơ bản, kết quả định danh đúng tiếng Việt giảm nhiều nhất. Trong trường hợp chỉ sử dụng các tham số liên quan đến tần số cơ bản, tiếng Việt lại được định danh đúng với tỷ lệ cao nhất. Điều này cho thấy, đối với các ngôn ngữ có thanh điệu trong đó có tiếng Việt, cần lưu ý đến vai trò của tần số cơ bản trong các hệ thống nhận dạng tiếng nói nói chung và định danh ngôn ngữ nói riêng.

Hướng nghiên cứu tiếp theo của chúng tôi là sử dụng các mô hình định danh khác như GMM hoặc học sâu (Deep Learning) có kết hợp với các bộ phân lớp có hiệu quả nhằm nâng cao hiệu năng định danh ngôn ngữ.

TÀI LIỆU THAM KHẢO

- William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, and Pedro Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210-229, Apr. 2006.
- Shigeo Abe, *Support Vector Machines for Pattern Classification*, 2nd ed. London: Springer, 2010.
- Shady Y. EL-Mashed, Mohammed I. Sharway, and Hala H. Zayed, "Speaker independent Arabic speech recognition using support vector machine," in *Department of Electrical Engineering, Shoubra Faculty of Engineering, Benha University, Cairo, Egypt*, 2009.
- Juc Hou, Yi Liu, Thomas Fang Zheng, Jesper Olsen, and Jilei Tian, "Multi-layered features with SVM for Chinese accent identification," in *Audio Language and Image Processing*, 2010, pp. 25-30.
- Fred Richardson and William M. Campbell, "Discriminative Keyword Selection Using Support Vector Machines," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 209-216.
- K. Sreenivasa Rao, V. Ramu Reddy, and Sudhamay Maity, "Language Identification Using Spectral and Prosodic

- Features.: Springer International Publishing, 2015, ch. 1, pp. 2-7.
- [7] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai, "Class-Based n-gram Models of Natural," *Computational Linguistics*, vol. 18, no. 4, pp. 467-479 , Dec. 1992.
 - [8] Haizhou Li, Bin Ma, and Chin Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271-284, Jan. 2007.
 - [9] Khe Chai Sim and Haizhou Li, "On Acoustic Diversification Front-End for Spoken Language Identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 1029 - 1037, July 2008.
 - [10] Rong Tong, Bin Ma, Haizhou Li, and Eng Siong Chng, "Target-Oriented Phone Selection from Universal Phone Set for Spoken Language Recognition," in *Interspeech* , 2008.
 - [11] Jia Li You, Yi Ning Chen, Min Chu, Frank K. Soong, and Jin Lin Wang, "Identifying Language Origin of Named Entity With Multiple Information Sources," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, pp. 1077 - 1086.
 - [12] Gerrit Reinier Botha and Etienne Barnard, "Factors that affect the accuracy of text-based language identification," *Computer Speech & Language*, vol. 26, no. 5, pp. 307-320, Oct. 2012.
 - [13] Marc Picard, *An Introduction to the Comparative Phonetics.*: John Benjamins Publishing Company, Amsterdam/Philadelphia, 1987.
 - [14] Nguyễn Hữu Quỳnh, *Tiếng Việt hiện đại (Ngữ âm, ngữ pháp, phong cách)*: Trung tâm biên soạn từ điển bách khoa Việt Nam, Hà Nội, 1994.
 - [15] Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-* 93 1993 IEEE International Conference on, 1993, pp. 399-402.
 - [16] Muthusamy , Yeshwant K , Ronald A , Cole , and Beatrice T. Oshika, "The OGI multi-language telephone speech corpus," *ICSLP*, vol. 92, pp. 895-898, Oct. 1992.
 - [17] Manchala, V. Kamakshi Prasad, and V. Janaki, "GMM based language identification system using robust features," *International Journal of Speech Technology*, vol. 17, no. 2, pp. 99–105, June 2014.
 - [18] Martin , Alvin F, and Craig S. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey*, 2010.
 - [19] Luciana Ferrer, Yun Lei, Mitchell McLaren, and Nicolas Scheffer, "Study of Sennic-Based Deep Neural Network Approaches for Spoken Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 105 - 116, Jan. 2016.
 - [20] Ana Montalvo, Yandré M. G. Costa, and José Ramón Calvo, "Language Identification Using Spectrogram Texture," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.*: Springer International Publishing, 2015, pp. 543-550.
 - [21] Hà Hải Nam, Trịnh Văn Loan, "Một hướng tiếp cận dựa trên tần số cơ bản để định danh tự động ngôn ngữ có thanh điệu và không có thanh điệu," *Kỷ yếu Hội thảo khoa học Quốc gia lần thứ hai về nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và truyền thông ICT.rda*, Hà Nội, 2004, pp. 211-215.
 - [22] Ian H.Witten, Eibe Frank, and Mark A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Korean : Morgan Kaufmann, 2011.
 - [23] Schuller , Björn , Stefan Steidl, and Anton Batliner, "The InterSpeech 2009 Emotion Challenge," in *INTERSPEECH*, 2009, pp. 312-315.
 - [24] John Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *technical report msr-tr-98-14. Microsoft Research*, vol. 112, Apr 1998.

53. Quy hoạch động và thuật toán rời rạc hoá thuộc tính liên tục
Đỗ Thị Bích Lê, Nguyễn Tiến Đạt, Nguyễn Quốc Hùng, Nguyễn Đăng Cao 321
54. Scyther: Công cụ kiểm chứng và phân tích giao thức bảo mật
Bùi Thị Thư, Nguyễn Trường Thắng, Trần Mạnh Đông, Nguyễn Thị Ánh Phương 326
55. So sánh các chính sách lưu trữ dữ liệu trong mạng hướng nội dung
Lê Phong Dũ, Lê Tuấn Anh, Nguyễn Đức Thái 332
56. So sánh một số phương pháp phân lớp dùng cho định danh ba ngôn ngữ Việt - Anh - Pháp
Lê Trung Hiếu, Nguyễn Kim Khánh, Trịnh Văn Loan 336
57. Statistics-theoretical approach for evaluating the similarity of fuzzy objects in fuzzy object-oriented databases
Nguyễn Tân Thuận, Trần Thị Thuý Trinh, Đoàn Văn Ban, Trương Ngọc Châu 342
58. Tăng cường thuật toán AdaBoost bằng cách sử dụng tập Pareto trong ứng dụng tra cứu ảnh dựa vào nội dung
Vũ Văn Hiệu, Nguyễn Trường Thắng, Ngô Quốc Tạo, Nguyễn Hữu Quỳnh 348
59. Thiết kế hệ điều khiển tự động gia nhiệt bằng hơi nước cho tháp chưng cất tinh dầu
Vũ Thị Quyên, Phạm Ngọc Minh, Vũ Chấn Hưng, Nguyễn Hà Phương, Vương Huy Hoàng, Dương Đức Hùng 355
60. Tương quan giữa phụ thuộc hàm xấp xỉ và phụ thuộc Boole dương tổng quát trong cơ sở dữ liệu quan hệ
Nguyễn Xuân Huy, Trương Thị Thu Hà, Nguyễn Thị Vân 361
61. Ứng dụng thuật toán Naïve Bayes trong phát hiện các trang web giả mạo
Nguyễn Ngọc Cường, Nguyễn Thị Huyền, Trần Đức Thắng 366
62. Về một phương pháp xây dựng độ phân hạt mờ mở rộng dựa trên khoảng cách mờ
Nguyễn Văn Thiện, Nguyễn Như Sơn, Nguyễn Long Giang, Cao Chính Nghĩa 371