ISSN 2354-0575

TỔNG QUAN ĐỊNH DANH NGÔN NGỮ TỰ ĐỘNG

Lê Trung Hiếu, Chu Bá Thành
Trường Đại học Sư phạm Kỹ thuật Hưng Yên

Ngày nhận: 09/2/2016
Ngày xét duyệt: 15/3/2016

Tóm tắt:

Trong bài báo này chúng tôi sẽ trình bày tổng quan về việc định danh ngôn ngữ tự động (LID – Language Identification). Việc định danh ngôn ngữ sẽ dựa trên các đặc trưng của tiếng nói như âm học, ngữ âm, ràng buộc âm vị, điệu tính, hình vị học, cú pháp và các hệ thống định danh phổ biến như hệ thống định danh ngôn ngữ tường minh và hệ thống là hệ thống định danh ngôn ngữ ẩn. Dựa vào các đặc trưng ngôn ngữ và các hệ thống định danh bài báo tiếp tục trình bày các vấn đề đặt ra cho một hệ thống định danh ngôn ngữ tự động cần phải giải quyết.

Từ khoá: *Định danh ngôn ngữ tự động, LID.*

1. Giới thiệu

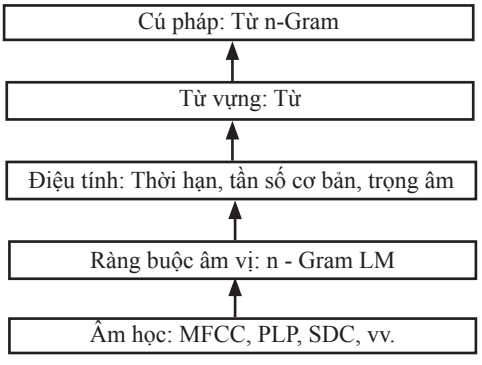
Con người được coi là những hệ thống định danh ngôn ngữ tự động nổi tiếng nhất thế giới hiện nay. Đơn giản, khi nghe một hoặc hai giây tiếng nói của một ngôn ngữ quen thuộc, họ có thể dễ dàng trích xuất các dấu hiệu cụ thể để xác định ngôn ngữ đó. Con người sử dụng kiến thức như: từ vựng, cú pháp, ngữ pháp và cấu trúc câu để xác định ngôn ngữ.

Tuy con người là những hệ thống LID nổi tiếng nhưng họ lại muốn thiết kế các hệ thống LID bằng máy móc nhằm tạo ra các hệ thống tương tác người - máy phục vụ nhiều hơn trong công việc và cuộc sống. Các hệ thống LID này cũng có khả năng xác định tiếng nói trong một thời gian ngắn mà tín hiệu tiếng nói phát ra bởi người nói. Một hệ LID tốt là hệ thống đảm bảo các tính năng chính sau đây của một hệ thống nhận dạng ngôn ngữ:

- Thời gian định danh tiếng nói là nhỏ.
- Hệ thống không phân biệt với bất kỳ ngôn ngữ hoặc nhóm ngôn ngữ nào.
- Hệ thống luôn đáp ứng với sự thay đổi người nói, biến đổi giọng, sự biến đổi kênh, môi trường…
- Hệ thống phải đơn giản và việc đưa thêm ngôn ngữ mới vào hệ thống phải được thực hiện một cách dễ dàng.

2. Các đặc trưng của tiếng nói

Trên thực tế có một loạt các thông tin mà con người và máy móc có thể sử dụng để phân biệt ngôn ngữ. Ở mức độ thấp, các đặc trưng giọng nói như thông tin âm học (acoustic), ngữ âm (phonetic), ràng buộc âm vị (phonotactic) và điệu tính (prosodic) được sử dụng rộng rãi trong các hệ thống LID. Ở một mức độ 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 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. Khi so sánh với các đặc trưng tiếng nói ở mức độ cao hơn, đặc trưng âm thanh ở mức độ thấp hơn là dễ thu được, nhưng dễ bị thay đổi bởi vì việc thay đổi người nói hoặc kênh đều có thể xảy ra. Ở các cấp độ đặc trưng cao hơn, như những đặc trưng cú pháp (syntactic features), được cho là mang nhiều thông tin ngôn ngữ tách biệt [1], nhưng những thông tin này được sử dụng bởi những hệ thống nhận dạng có vốn từ vựng lớn và do đó là khó để có được nó.



Hình 1. *Các mức đặc trưng của hệ thống nhận dạng ngôn ngữ*

2.1. Âm học-Ngữ âm

Thông tin âm học thường được coi là mức phân tích đầu tiên về quá trình tạo tiếng nói. Tiếng nói khác nhau có thể được phân biệt ở một mức độ tùy theo biên độ âm thanh và thành phần tần số của sóng âm [2]. Thông tin âm học là một trong những hình thức đơn giản nhất của thông tin có thể tham số hóa được trong quá trình nói. Ngoài ra, thông tin cấp cao hơn như thông tin về ràng buộc âm vị

(phonotactic) và âm tiết có thể được chiết xuất từ các thông tin âm thanh. Các phương pháp được sử dụng rộng rãi nhất là Linear Prediction, Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) và Linear Prediction Cepstral Coefficient (LPCC) [3, 4].

2.2. Ràng buộc âm vị

Âm vị học (phonology) là nghiên cứu về hệ thống âm thanh của một ngôn ngữ cụ thể hoặc trong ngôn ngữ nói chung và ràng buộc âm vị (phonotactics) là một nhánh của âm vị học mà ở đó các liên kết âm của các ngôn ngữ khác nhau là khác nhau. Những kết hợp cho phép của âm bao gồm các cụm phụ âm và nguyên âm được liên kết theo một quy luật nào đó [5]. Ràng buộc âm vị là sự chi phối một cách khác nhau về âm vị, được kết hợp từ các âm tiết hoặc các từ ngữ không giống nhau giữa các ngôn ngữ. Một số cụm âm vị hoặc âm tiết phổ biến trong một ngôn ngữ này có thể không có trong ngôn ngữ khác, ví dụ các cụm âm vị / st / là rất phổ biến trong tiếng Anh, trái lại nó không được cho phép ở tiếng Nhật,… Do đó, thông tin ràng buộc âm vị mang nhiều thông tin ngôn ngữ rõ ràng hơn những âm vịcủa chính ngôn ngữ đó và nó thích hợp cho việc khai thác các đặc thù của ngôn ngữ.

2.3. Điệu tính

Điệu tính (prosody) là một trong những thành phần quan trọng trong việc nhận thức bằng thính giác của con người. Giai điệu, trọng âm, thời hạn, cường độ và nhịp điệu là các mặt chính của điệu tính và nó thay đổi khác nhau từ ngôn ngữ này sang ngôn ngữ khác. Thông thường tần số cơ bản (fundamental frequency) được sử dụng để biểu diễn các giai điệu của âm, cường độ được sử dụng để chỉ ra trọng âm và chuỗi thời hạn được sử dụng để đại diện cho nhịp điệu. Một số âm vị được dùng trên các ngôn ngữ khác nhau và đặc tính thời gian của nó sẽ phụ thuộc vào các ràng buộc âm vị của ngôn ngữ. Các biểu hiện của điệu tính ràng buộc về ngôn luận, truyền tải một vài thông tin quan trọng liên quan tới ngôn ngữ.

2.4. Hình vị học

Hình vị (morpheme) là đơn vị nhỏ nhất về mặt ngữ pháp của một ngôn ngữ và là đơn vị nhỏ nhất có nghĩa của ngôn ngữ đó. Lĩnh vực dành cho nghiên cứu hình vị được gọi là hình vị học (morphology) [6]. Hình vị không hoàn toàn giống như một từ. Sự khác nhau giữa hình vị và từ là hình vị có thể hoặc không đứng riêng còn từ thì đứng tùy ý. Khi đứng riêng, hình vị được xem như là gốc từ (root) vì có nghĩa của riêng nó (chẳng hạn hình vị *cat* trong tiếng Anh) còn khi hình vị phụ thuộc vào hình vị khác để biểu diễn một ý nào đó, nó trở thành phụ tố (affix) vì lúc đó có chức năng ngữ pháp (chẳng hạn, -*s* trong *cats* để cho biết số nhiều). Mỗi từ có thể bao gồm một hoặc nhiều hình vị. Như vậy hệ thống nhận dạng ngôn ngữ tự động có thể được thực hiện ở cấp độ từ bằng cách kiểm tra đặc điểm riêng của hình thức từ.

2.5. Cú pháp

Trong ngôn ngữ học, cú pháp (syntax) là việc nghiên cứu về các nguyên tắc và quy tắc ảnh hưởng, cách mà các từ ghép với nhau trong một câu. Các mẫu câu khác nhau qua các ngôn ngữ khác nhau. Ngay cả trường hợp một từ đơn được chia sẻ bởi hai ngôn ngữ khác nhau, nhưng trong văn cảnh (ví dụ như các từ đứng trước và các từ theo sau) có thể khác nhau giữa các ngôn ngữ [7]. Việc tích hợp từ vựng và ngữ pháp, bằng cách khai thác thông tin về hình vị học và cú pháp, dẫn đến cải thiện trong các hệ thống nhận dạng tiếng nói và việc đưa các thông tin này vào hệ thống LID đã đạt được một số thành công nhất định. Tuy nhiên, việc xây dựng các từ điển và ngữ pháp dựa trên từ cho các hệ thống LID cần một nỗ lực đáng kể so với việc chỉ dừng ở mức ngữ âm. Các hệ thống sử dụng các thông tin về hình vị học và cú pháp hiện nay không phải là phổ biến.

3. Các hệ thống định danh ngôn ngữ

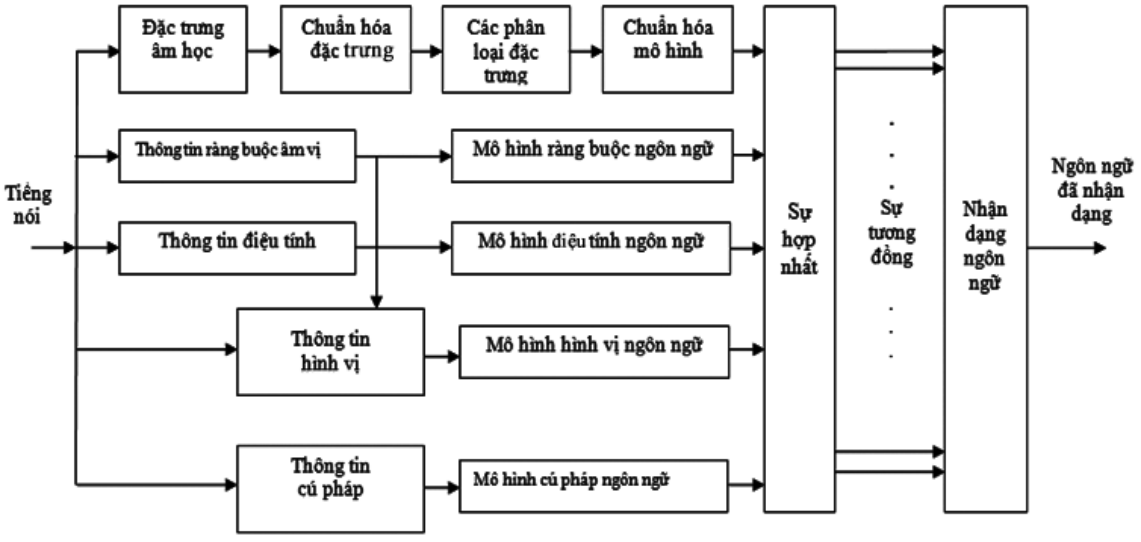
Các hệ thống LID điển hình bao gồm các hệ thống con sử dụng một số hoặc tất cả các loại thông tin đã nêu trên để đánh giá sự giống nhau nào đó của các ngôn ngữ khác nhau và việc đánh giá từ các hệ thống con này được kết hợp để đưa ra quyết định cuối cùng về ngôn ngữ cần định danh.

Hình 2 là sơ đồ khối tổng quan của hệ thống LID sử dụng với tất cả các mức thông tin. Tuy nhiên, không cần thiết cho một hệ thống LID phải làm như vậy, và thực sự các hệ thống LID cũng không làm như vậy. Các phương pháp phổ biến nhất là sử dụng thông tin âm học (acoustic) và ràng buộc âm vị.

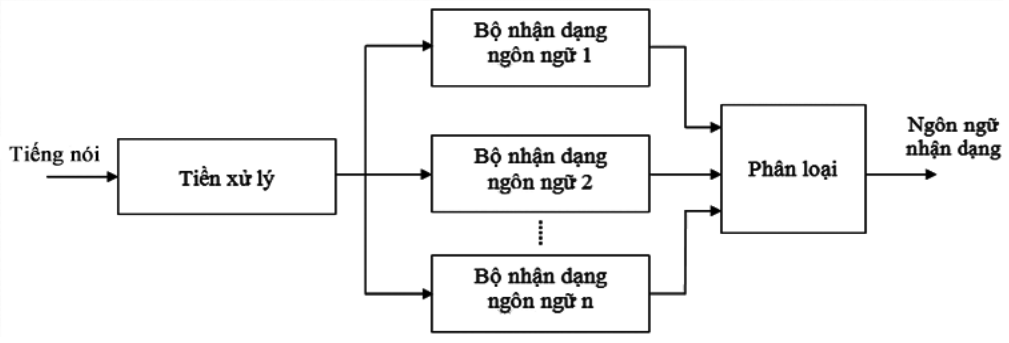
Trên thực tế các hệ thống định danh ngôn ngữ tự động có thể được chia thành hai loại đó là hệ thống định danh ngôn ngữ tường minh và hệ thống định danh ngôn ngữ ẩn.

3.1. Hệ thống định danh ngôn ngữ tường minh

Hệ thống định danh ngôn ngữ tường minh được thể hiện trong Hình 3. Nguyên tắc hoạt động của hệ thống là ban đầu dữ liệu tiếng nói sẽ được đưa vào bộ tiền xử lý, sau đó dữ liệu của các ngôn ngữ khác nhau đã được xác định sẽ được đưa vào các bộ nhận dạng ngôn ngữ 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.



Hình 2. *Mô hình tổng quan của hệ thống định danh ngôn ngữ*



Hình 3. *Hệ thống định danh ngôn ngữ tường minh*

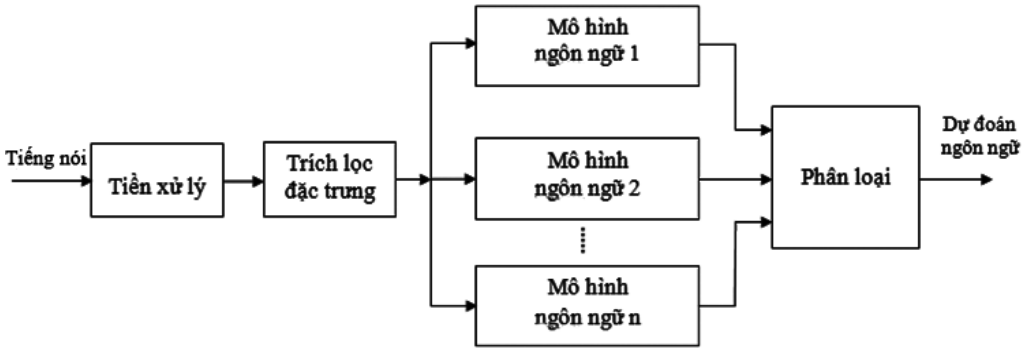
Nhiều kết quả nghiên cứu ứng dụng hệ thống định danh ngôn ngữ tường minh đã được công bố trên thế giới. Lamel và Gauvain [8, 9] đã sử dụng bộ nhận dạng âm vị như là bước xử lý đầu tiên để thực hiện nhiệm vụ định danh. Bộ nhận dạng âm vị cho tiếng Pháp và tiếng Anh đã được xây dựng và sử dụng song song. Tín hiệu tiếng nói của bất kỳ ngôn ngữ nào trong số hai ngôn ngữ này được hai bộ nhận dạng âm vị xử lý song song. Ngôn ngữ gắn với mô hình có tính tương đồng cao nhất sẽ được xem là ngôn ngữ của tín hiệu tiếng nói ở đầu vào. Berking và cộng sự [10] đã xét một tập hợp cha các âm vị của 3 ngôn ngữ khác nhau như tiếng Anh, tiếng Nhật và tiếng Đức. Họ đã khai thác tìm kiếm và sử dụng chỉ các âm vị này để phân biệt tốt nhất từng cặp ngôn ngữ. Hazen và Zue [11] đã theo đuổi việc sử dụng chỉ một bộ nhận dạng âm vị ở đầu vào cho nhận dạng đa ngôn ngữ thay cho việc sử dụng bộ nhận dạng âm vị phụ thuộc ngôn ngữ và đã kết hợp các thông tin điệu tính, âm học, ngữ âm suy diễn từ tiếng nói trong khuôn khổ thống kê.

3.2. Hệ thống định danh ngôn ngữ ẩn

Hệ thống định danh ngôn ngữ ẩn được mô tả trong Hình 4.

Nguyên lý hoạt động của hệ thống là ban đầu dữ liệu tiếng nói được đưa vào bộ tiền xử lý, sau đó dữ liệu đưa vào bộ trích lọc đặc trưng để lấy ra đặc trưng của các ngôn ngữ, tiếp theo dữ liệu được đưa vào mô hình ngôn ngữ khác nhau (các mô hình ngôn ngữ khác nhau sẽ xử lý và cho biết các đặc trưng của từng ngôn ngữ khác nhau). Tiếp theo thông tin sẽ được đưa ra bộ phân loại và cho ra kết quả ngôn ngữ được định danh.

Các kết quả nghiên cứu ứng dụng hệ thống định danh ngôn ngữ ẩn đã được công bố trên thế giới như: Carrasquillo PAT [12] hay Wong E [13] đã sử dụng mô hình hỗn hợp Gaussian trong hệ thống định danh ngôn ngữ. Campbell et al. [14], Zhai et al. [15] and Castaldo et al. [16] đã ứng dụng SVMs (Support Vector Machine) cho nhiệm vụ định danh ngôn ngữ và đã cho kết quả cải thiện hơn so với phương pháp tiếp cận dùng GMM (Gaussian Mixture Model). Hay Chung-Hsien [17] và cộng sự đã thực hiện phân đoạn tự động và nhận dạng giọng nói của hỗn hợp ngôn ngữ sử dụng delta-BIC (delta - Bayesian Information Criterion và GMMs LSA (Latent Semantic Analysis).



Hình 4. *Hệ thống định danh ngôn ngữ ẩn*

4. Một số vấn đề đặt ra cho hệ thống định danh ngôn ngữ

Việc định danh một ngôn ngữ mà không có sự hiểu biết về ngôn ngữ đó là một thách thức rất lớn. Trong lĩnh vực định danh ngôn ngữ, nên giả thiết rằng không có phổ hoặc bất kỳ kiểu thông tin nào khác của người nói đã hiện diện trong tập tham chiếu. Việc so sánh giữa mẫu cần nhận dạng và các mẫu tham chiếu luôn xuất phát từ các phát ngôn không bị ràng buộc của hai người nói khác nhau. Như vậy, giữa hai phát ngôn đó luôn có sự khác biệt như nội dung phát ngôn, người nói, môi trường ghi âm và ngôn ngữ. Vì thế, để định danh các ngôn ngữ khác nhau, ngoài nội dung nói, người nói và môi trường ghi âm khác nhau sẽ là những vấn đề quan trọng. Có thể nêu chi tiết về những vấn đề này như sau.

• Biến đổi về đặc tính của người nói. Người nói khác nhau sẽ có sắc thái nói khác nhau, điều này làm tăng tính biến đổi hay biến thiên đặc tính người nói đối với các ràng buộc đặt ra ngay trong cùng một ngôn ngữ. Vì vậy cần vô hiệu hóa sự biến đổi này khi mô hình hóa ngôn ngữ.

• Biến đổi về ngữ điệu. Ngữ điệu liên quan chủ yếu đến phát âm. Từ ngữ điệu, ta có thể nhận ra người nói có giọng tự nhiên bản xứ hay không. Tuy nhiên, sẽ gặp phải khó khăn khi mô tả sự khác biệt về ngữ điệu.

• Biến đổi về môi trường và các đặc tính của kênh truyền dẫn. Các đặc tính của tín hiệu tiếng nói chịu ảnh hưởng nhiều của điều kiện môi trường trong đó dữ liệu được thu thập hoặc được truyền dẫn. Các yếu tố này có ảnh hưởng đáng kể đến các đặc trưng được trích xuất từ phân tích phổ ngắn hạn. Do đó, cần phải có các đặc trưng ít chịu ảnh hưởng của mô trường và kênh truyền dẫn để có một hệ thống nhận dạng tốt ngôn ngữ.

• Biến đổi về phương ngữ. Phương ngữ là sự đa dạng của ngôn ngữ theo khu vực và theo tập thể cư dân được phân biệt theo cách phát âm, ngữ pháp, từ vựng và đặc biệt là sự đa dạng của tiếng nói khác với ngôn ngữ văn học chuẩn hoặc nguyên mẫu tiếng nói của nền văn hóa mà phương ngữ đó tồn tại.

• Tính tương tự của các ngôn ngữ. Có nhiều sự tương tự giữa các ngôn ngữ. Chẳng hạn phần lớn các ngôn ngữ Ấn Độ có chung tập gốc từ và cũng theo cấu trúc ngữ pháp tương tự.

• Việc trích chọn và biểu diễn điệu tính đặc trưng cho ngôn ngữ. Các đặc trưng về tính điệu như thanh điệu, thời hạn, cường độ, trọng âm, nhịp điệu là thay đổi đối với các ngôn ngữ khác nhau. Nhưng bản chất của các đặc tính này không được định nghĩa rõ ràng. Chẳng hạn, nhịp điệu của một ngôn ngữ nào đấy có thể được cảm nhận do sự kế tiếp của các âm tiết, nguyên âm, biến thiên biên độ đột ngột, thanh điệu đi lên hoặc đi xuống song thực sự vẫn chưa hiểu rõ chúng. Hơn nữa, không có sẵn các ký thuật thích hợp xử lý tiếng nói nhằm biểu diễn tri thức nguồn ở mức cao giống như điệu tính. Do vậy, việc trích rút và biểu diễn điệu tính chuyên biệt cho ngôn ngữ hãy còn là điều khó khăn.

Có thể thấy rằng, việc định danh một ngôn ngữ sẽ thuận lợi hơn nếu các ngôn ngữ cần định danh rất khác biệt nhau (tức là tập các âm vị là hoàn toàn khác cho mỗi ngôn ngữ). Mặc dù vậy, tất cả các ngôn ngữ chia sẻ một tập là chung của các âm vị vì phần lớn các ngôn ngữ có chung một gốc.

5. Kết luận và hướng phát triển

Bài báo đã trình bày các đặc trưng của tiếng nói và các đặc điểm của từng đặc trưng; mô hình tổng quan định danh ngôn ngữ dựa vào các đặc trưng khác nhau của tiếng nói; hai hệ thống định danh ngôn ngữ được được sử dụng rộng rãi trên thực tế đó là: hệ thống định danh ngôn ngữ tường minh và hệ thống danh ngôn ngữ ẩn. Dựa vào các kết quả nghiên cứu về định danh ngôn ngữ của các tác giả khác nhau trên thế giới chúng tôi đã đưa ra một số vấn đề đặt ra cho hệ thống định danh ngôn ngữ cần phải xử lý như:vấn đề về biến đổi đặc tính của người nói, ngữ điệu, môi trường, đặc các tính của kênh truyền dẫn, phương ngữ, tính tương tự

của ngôn ngữ… Từ đây giúp người đọc có cái nhìn tổng quan về định danh ngôn ngữ tự động và các vấn đề cần giải quyết.Trên cơ sở các nghiên cứu đã đạt được chúng tôi sẽ phát triển hệ thống định danh ngôn ngữ tự động với các ngôn ngữ khác nhau đặc biệt là việc định danh các ngôn ngữ khác cùng với tiếng Việt.

Tài liệu tham khảo

[1]. Schultz T, Rogina I, Waibel A (1996), *LVCSR-Based Language Identification*, In: Proceedings of IEEE International Conference Acoustics, Speech, And Signal Processing (ICASSP-96), Vol 2, PP 781–784.

[2]. Laver J (1994), *Principles of Phonetics*, Cambridge University Press, Cambridge.

[3]. Jurafsky D, Martin J (2008), *Speech And Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2 edn. Prentice Hall, New Jersey.

[4]. Rabiner L, Juang B (1993), *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey.

[5]. Schultz T, Kirchhoff K (2006), *Multilingual Speech Processing*, Academic, New York.

[6]. Bauer L (2003), *Introducing Linguistic Morphology*, Georgetown University Press, Washington D.C.

[7]. Zissman MA (1996), *Comparison of Four Approaches to Automatic Language Identification of Telephone Speech*, IEEE Trans Speech Audio Process 4:31–44.

[8]. Lamel LF, Gauvain JL (1993), *Cross Lingual Experiments with Phone Recognition*, In: Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing, PP 507–510, April 1993.

[9]. Lamel LF, Gauvain JL (1994), *Language Identification Using Phonebased Acoustic Likelihoods*, In: Proceedings of IEEE International Conference On Acoustics, Speech, And Signal Processing, Vol 1, PP 293–296, April 1994.

[10]. Berkling KM, Arai T, Bernard E (1994), *Analysis of Phoneme Based Features for Language Identification*, In: Proceedings Of IEEE International Conference On Acoustics, Speech, And signal Processing, PP 289–292, April 1994.

[11]. Hazen TJ, Zue VW (1994), *Recent Improvements in An Approach to Segement-Based Automatic Language Identification*, In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, PP 1883–1886, Sept 1994.

[12]. Carrasquillo PAT, Reynolds DA, Deller JR (2002), *Language Identification Using Gaussian Mixture Model Tokenization*, In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol I, PP 757–760, 2002.

[13]. Wong E, Sridharan S (2002), *Gaussian Mixture Model Based Language Identification System*, In: Proceedings International Conference Spoken Language Processing (ICSLP-2002), PP 93–96, 2002.

[14]. Campbell W, Singera E, Torres-Carrasquillo P, Reynolds D (2004), *Language Recognition With Support Vector Machines*, In Proceedings of ODYSSEY- 2004:2004.

[15]. Lu-Feng Z, Man-hung S, Xi Y, Gish H (2006), *Discriminatively Trained Language Models Using Support Vector Machines for Language Identification*, In: Proceedings of Speaker and Language Recognition Workshop, 2006. IEEE Odyssey, PP1–6.

[16]. Castaldo F, Dalmasso E, Laface P, Colibro D, Vair C (2007), *Language Identification Using Acoustic Models and Speaker Compensated Cepstral-Time Matrices*, In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), pp IV-1013IV-1016, 2007.

[17]. Wu C-H, Chiu Y-H, Shia C-J, Lin C-Y (2006), *Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs*, IEEE Trans Audio Speech Lang Process 14:266–276.

ISSN 2354-0575

AN OVERVIEW OF AUTOMATIC LANGUAGE IDENTIFICATION

Abstract:

In this article, we will present an overview of automatic language identification (LID – Language Identification). The language identification will base on the speech feature such as acoustic, phonetics, pholotactics, prosody, morphology, systax and the popular identification systems such as the explicit language identification system and the implicit language identification system. Relying on the feature languges and the identification systems, the article will continue to present the issues that it is had got to solve for the automatic spoken language identification system.

Keywords: *Language Identification, LID.*