Nhận dạng tiếng nói trên cơ sở mạng Nơron nhân tạo

Hồ Văn Hương

Người hướng dẫn : PGS.TS. Nguyễn Quang Hoan

MỞ ĐẦU

Nhận dạng tiếng nói là mong ước của khoa học và con người. Những người máy có thể hiểu được tiếng người nói và thực thi nhiệm vụ.

Hiện nay, nhận dạng tiếng nói chưa thực sự đáp ứng đầy đủ các yêu cầu thực tế, song những hệ thống nhận dạng tiếng nói đã có bước phát triển đáng kể.

Trên thế giới, một số hệ thống nhận dạng tiếng nói cỡ lớn có độ chính xác tương đối cao. Các hệ thống này chủ yếu được phát triển trên nền công nghệ hiện đại với những máy tính lớn, những vi mạch xử lý tiếng nói chuyên dụng và sử dụng cơ sở dữ liệu tiếng nói khá hoàn chỉnh, nhưng phần lớn vẫn là xử lý cho tiếng Anh.

Ở Việt Nam, việc tìm hiểu, nghiên cứu và phát triển các hệ thống nhận dạng tiếng nói còn đang bước đầu có kết quả. Do có những đặc thù riêng của tiếng Việt, nên việc chọn lựa phương pháp tiếp cận bài toán nhận dạng phù hợp với tiếng Việt là một vấn đề tương đối khó khăn.

Những năm gần đây, cũng có khá nhiều đề tài nghiên cứu về nhận dạng tiếng nói tiếng Việt. Các hệ thống nhận dạng tiếng nói thành công nhất chủ yếu dựa trên khuynh hướng nhận dạng mẫu. Các kỹ thuật nhận dạng mẫu đơn giản như lượng tử hoá véctơ, hiệu chỉnh thời gian động…, đã được áp dụng khá thành công vào các chương trình nhận dạng tiếng nói tiếng Việt phát âm rời rạc với số lượng từ vựng hạn chế.

Tuy nhiên, mục tiêu của nhận dạng tiếng nói tự động bằng máy là phải tiến tới hệ thống nhận dạng tiếng nói liên tục, kích thước từ điển lớn, không phụ thuộc vào người nói. Vì vậy, các hệ thống nhận dạng tiếng nói hiện nay thường xây dựng trên cơ sở áp dụng các kỹ thuật nhận dạng mẫu phức tạp hơn, đó là mô hình Markov ẩn và mạng nơron nhân tạo đã cho một số thành công nhất định.

Xuất phát từ nhận thức trên, đề tài luận văn Thạc sỹ của em là tìm hiểu, đưa ra phương pháp và xây dựng một ứng dụng nhận dạng tiếng nói tiếng Việt. Với những khả năng của mạng

nơron nhân tạo trong ứng dụng, đã cho nhiều thành công đáng khích lệ. Vì vậy, em đã chọn mạng nơron làm cơ sở cho việc nghiên cứu nhận dạng tiếng nói tiếng Việt.

Nội dung luận văn gồm 5 chương như sau:

- **Chương 1:** Tìm hiểu lịch sử việc nghiên cứu nhận dạng tiếng nói, tổng quan về tiếng nói và nhận dạng tiếng nói.
- **Chương 2:** Trình bày một số tính chất của tiếng nói như: cơ chế tạo ra tiếng nói, cơ chế thu tiếng nói, các đặc trưng tiếng nói. Ngoài ra, chương này cũng đề cập đến kỹ thuật tính hệ số MFCC, là một phương pháp trích chọn đặc trưng tín hiệu tiếng nói khá phổ biến đã được áp dụng hiệu quả trong các hệ thống nhận dạng.
- **Chương 3:** Tìm hiểu tổng quan về mạng nơron, những khái niệm, cấu trúc, các luật học. Chương này cũng đề cập đến những ứng dụng của mạng nơron trong nhận dạng và phân lớp.
- **Chương 4:** Nghiên cứu về mạng nơron lan truyền ngược gồm: cấu trúc, phương pháp huấn luyện mạng. Chương này cũng đề cập đến cấu trúc cụ thể của mạng áp dụng cho bài toán nhận dạng tiếng nói tiếng Việt và đồng thời đánh giá các tham số của hệ thống nhận dạng.
- **Chương 5:** Xây dựng hệ thống nhận dạng, giao diện chương trình, các kết quả thực nghiệm.

Cuối cùng là kết luận và định hướng phát triển của đề tài.

CHƯƠNG 1

TỔNG QUAN VỀ TIẾNG NÓI VÀ BÀI TOÁN NHẬN DẠNG

1.1 Lịch sử phát triển của nhận dạng tiếng nói

Nhận dạng tiếng nói tự động đã phát triển khoảng 40 năm nay. Những nhân tố quan trọng giúp cho sự phát triển của công nghệ nhận dạng này có thể kể đến như sự phát triển của hệ thống phân tích phổ âm thanh (năm 1946) cho phép thể hiện trực quan các tín hiệu âm, lý thuyết tạo âm thanh tiếng nói của người (năm 1948), sự xuất hiện và phát triển mạnh mẽ của các hệ thống máy tính số thương mại đầu tiên trên thế giới (năm 1958).

Các hệ thống nhận dạng đầu tiên có khả năng nhận dạng từ rời rạc và phụ thuộc người nói. Để phân tích và nhận dạng các chữ số hoặc các từ đơn âm sử dụng đặc tính trong miền thời gian và các ngân hàng bộ lọc tương tự. Tương tự như vậy, với phương pháp âm học, hệ thống nhận dạng âm vị phụ thuộc người nói và không phụ thuộc người nói được thiết kế mặc dù mới cho được kết quả rất khiêm tốn.

Trong thập kỷ 70, với sự phát triển của các thuật toán phân tích tín hiệu như mô hình dự đoán tuyến tính, so sánh mẫu theo thời gian…công nghệ nhận dạng tiếng nói tiếp tục có những bước phát triển mạnh mẽ. Với các phương pháp này những hệ thống nhận dạng với số lượng từ khá lớn được thiết kế.

Trong những năm 60 của thế kỷ 20, nhiều phòng thí nghiệm của nhiều hãng lớn đã được đầu tư nghiên cứu phát triển các hệ thống nhận dạng tiếng nói các ngôn ngữ khác nhau. Đến đầu những năm 80, khả năng về kỹ thuật đã cho phép các nhà nghiên cứu xây dựng các hệ thống nhận dạng được hàng trăm từ rời rạc. Gần đây công nghệ nhận dạng đã có những bước phát triển vô cùng nhanh chóng.

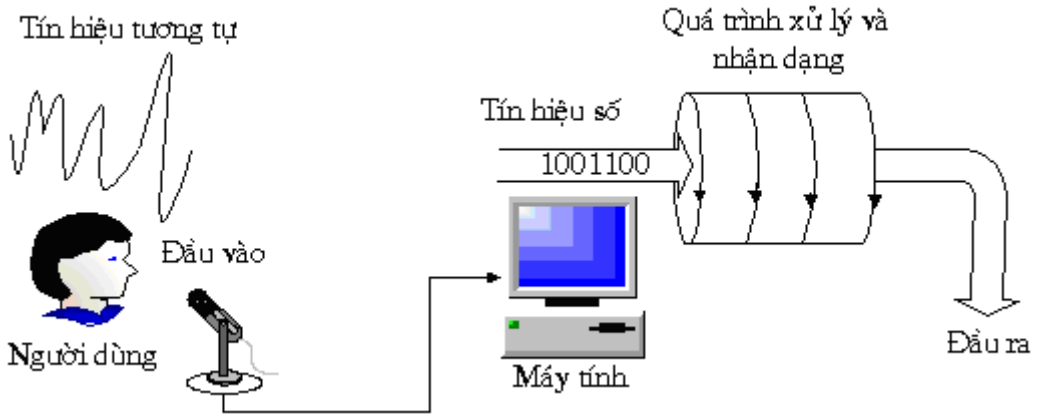
1.2 Tổng quan về bài toán nhận dạng tiếng nói

Nhận dạng tiếng nói là làm cho máy hiểu, nhận biết được ngữ nghĩa của lời nói. Đây là quá trình biến đổi tín hiệu âm thanh thu được qua micro, qua các thiết bị thu thanh khác… thành một chuỗi các từ, sau đó được nhận dạng để sử dụng trong các ứng dụng điều khiển thiết bị, nhập dữ liệu hoặc soạn thảo văn bản bằng lời… hoặc đưa đến một quá trình xử lý ngôn ngữ ở mức cao hơn.

Tiếng nói là công cụ truyền đạt thông tin quan trọng của người. Bình thường, chúng ta không để ý quá trình nhận dạng tiếng nói diễn ra như thế nào? tại sao chúng ta hiểu được các từ, các câu một cách đơn giản như vậy?

Trên thực tế, quá trình nhận dạng tiếng nói của người là một quá trình phức tạp. Hiện nay, các nhà nghiên cứu cố gắng tìm hiểu, phân tích và mô phỏng quá trình nhận dạng tiếng nói của người dưới dạng các chương trình máy tính. Nhưng đây là vấn đề rất rộng, có liên quan tới nhiều ngành nghiên cứu như sinh học, hoá học, vật lý ... Do vậy, việc mô phỏng tiếng nói cũng gặp nhiều khó khăn.

Chúng ta có thể thấy được một cách trực quan bài toán nhận dạng tiếng nói qua hình 1.1.



Hình 1.1 *Mô hình nhận dạng tiếng nói*

Nhận dạng tiếng nói là quá trình phức tạp bao gồm nhiều khâu biến đổi. Tín hiệu mà người phát ra là tín hiệu tương tự, qua quá trình lấy mẫu, lượng tử hoá và mã hoá để thu được các mẫu tín hiệu dạng số (tín hiệu mà máy tính có thể hiểu và xử lý được). Các mẫu tín hiệu này được trích chọn đặc trưng. Những đặc trưng này sẽ là đầu vào cho quá trình nhận dạng. Sau khi nhận dạng tín hiệu người dùng phát âm, hệ thống sẽ đưa ra kết quả nhận dạng. Tuỳ thuộc vào mô hình ứng dụng mà cho chúng ta các dạng đầu ra khác nhau.

Do tính chất của tiếng nói phụ thuộc vào nhiều yếu tố nên việc thu nhận, phân tích các đặc trưng của tiếng nói là việc không dễ. Ở đây, chúng ta có thể nêu ra một số yếu tố khó khăn cho bài toán nhận dạng tiếng nói:

- Khi phát âm, người nói thường nói nhanh, chậm khác nhau.
- Các từ được nói thường dài ngắn khác nhau.

- Một người cùng nói một từ, nhưng ở hai lần phát âm khác nhau. Kết quả phân tích khác nhau.
- Mỗi người có một chất giọng riêng được thể hiện thông qua độ cao, độ to, cường độ của âm và âm sắc.
- Những yếu tố như nhiễu của môi trường, nhiễu của thiết bị thu…ảnh hưởng không nhỏ tới hiệu quả nhận dạng.

Có thể thấy nhận dạng tiếng nói là một lĩnh vực nghiên cứu có nhiều ứng dụng trong thực tế. Các hệ thống nhận dạng góp phần rất lớn trong việc thúc đẩy phát triển nhiều ngành. Tuy là lĩnh vực mang ý nghĩa to lớn đó, nhưng việc phát triển các hệ thống nhận dạng cũng gặp không ít những khó khăn, nhất là ở Việt Nam khi các kết quả nghiên cứu về nhận dạng tiếng Việt chưa nhiều, cũng như cơ sở hạ tầng cho việc nghiên cứu còn ít.

1.3 Một số hệ thống nhận dạng tiếng nói

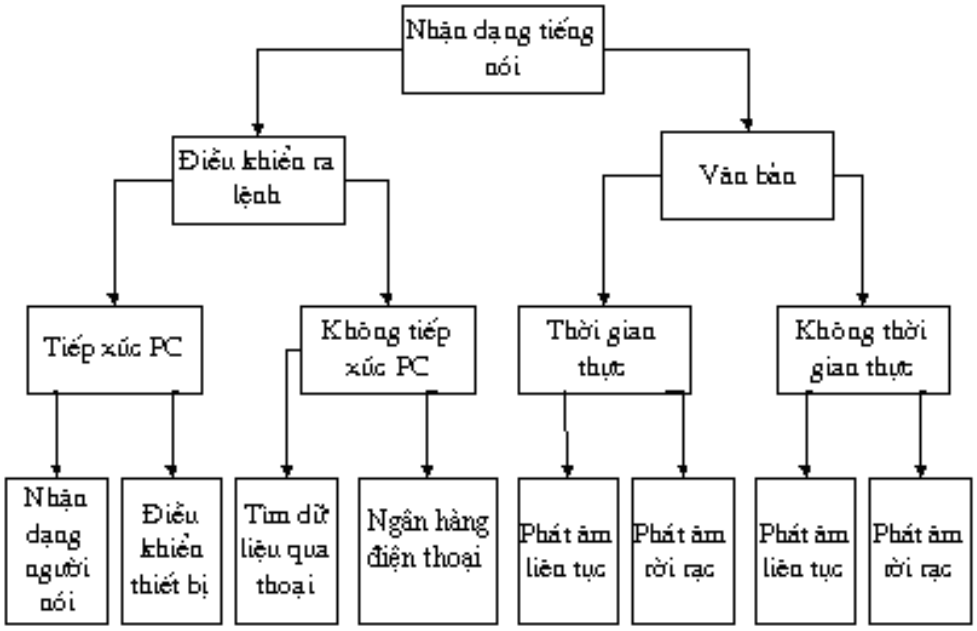
Nhận dạng tiếng nói là vấn đề đã được chia thành hai nhóm riêng biệt dựa trên mục đích sử dụng khác nhau.

- Một nhóm được sử dụng với mục đích điều khiển thiết bị thông qua giọng nói.
- Một nhóm sử dụng nhằm xử lý từ tiếng nói sang văn bản.

Phân loại các hệ thống nhận dạng tiếng nói sẽ giúp chúng ta có một cái nhìn trực quan hơn về bài toán. Các hệ thống nhận dạng được phân loại như hình vẽ 1.2.

Có 3 phương pháp phổ biến được sử dụng trong nhận dạng tiếng nói [10]:

- Phương pháp âm học - ngữ âm học.
- Phương pháp nhận dạng mẫu.
- Phương pháp sử dụng các kết quả của lĩnh vực trí tuệ nhân tạo.



Hình 1.2 *Sơ đồ phân loại các hệ thống nhận dạng tiếng nói*

1.4 Tổng quan về tiếng nói

1.4.1 Âm thanh và tiếng nói

Âm thanh thực chất là sự nén và dản một cách tuần hoàn không khí, tạo ra một sóng đàn hồi dọc. Sóng trong không khí truyền đến tai, tác động vào màng nhĩ, làm cho màng nhĩ dao động với cùng tần số (dao động cưỡng bức), có khả năng tạo ra cảm giác âm thanh trong tai khi tần số sóng đạt tới một độ lớn nhất định. Tai người chỉ có thể cảm nhận được âm thanh trong một khoảng tần số từ 20Hz đến 20000Hz. Những sóng này gọi là sóng âm hay âm thanh.

Tiếng nói là âm thanh do người phát ra. Khi phát âm, nguồn không khí từ phổi sẽ kích hoạt bộ phát âm làm căng các dây thanh quản và khi không khí đi qua làm cho các dây thanh quản này dao động tạo nên âm thanh tiếng nói. Tiếng nói của người có năng lượng tập trung nằm trong khoảng tần số từ 1000Hz đến 4000Hz.

1.4.2 Cao độ của âm (pitch)

Cao độ của âm thanh là độ cao hay thấp của âm thanh được quyết định bởi sự

rung dây thanh. Dây thanh rung với tần số nhanh sẽ cho những âm cao, dây thanh rung chậm sẽ cho những âm thấp. Đơn vị đo cao độ ở đây được dùng là Hz, đo số chu kỳ dao động thực hiện được trong 1 giây, *gọi là tần số*. Những âm có tần số khác nhau gây cho ta những cảm giác âm khác nhau. Độ cao của âm mang đặc tính sinh lý của âm. Nó dựa vào đặc tính của âm là tần số. Do cấu tạo của dây thanh khác nhau, mà tần số tạo ra phụ thuộc vào giới tính và lứa tuổi của người phát âm (phụ nữ và trẻ em thường có tần số cao hơn nam giới và người lớn tuổi).

1.4.3 Cường độ (volume) và mức cường độ âm

Cũng như các sóng cơ học khác, sóng âm mang năng lượng tỷ lệ với bình phương biên độ sóng. Năng lượng đó truyền đi từ nguồn âm đến tai ta.

Cường độ âm (I): là lượng năng lượng được sóng âm truyền trong một đơn vị thời gian qua một đơn vị diện tích đặt vuông góc với phương truyền âm, đơn vị đo là (W/m^2).

Đối với tai người, giá trị tuyệt đối của cường độ âm I không quan trọng bằng giá trị tỷ số của I so với một giá trị I_o nào đó được chọn làm chuẩn. Người ta định nghĩa mức cường độ âm L là logarith thập phân của tỉ số I/I_o:

$$L(B) = \lg(I/I_o) \text{ hoặc } L(dB) = 10\lg(I/I_o) \text{ với } 1B = 10dB \qquad (1.1)$$

Thực tế, người ta thường dùng đơn vị dB (deciben) hơn là B (ben). Khi L = 1dB, thì $I/I_o = 10^{1/10}$. Đây là mức cường độ nhỏ nhất mà tai ta có thể phân biệt được.

1.4.4 Độ to của âm

Muốn gây cảm giác âm, cường độ âm phải lớn hơn một giá trị cực tiểu nào đó gọi là ngưỡng nghe. Do đặc điểm sinh lý của tai người, ngưỡng nghe thay đổi tùy theo tần số âm. Với các tần số 1000Hz - 5000Hz, ngưỡng nghe khoãng $10^{-12}W/m^2$. Với tần số 50Hz, ngưỡng nghe lớn gấp 10^5 lần.

Nếu cường độ âm lên tới $10W/m^2$ thì sóng âm gây ra một cảm giác nhức nhối. Giá trị cực đại này gọi là ngưỡng đau. Miền nằm giữa ngưỡng đau và ngưỡng nghe gọi là miền nghe được. Khi xác định cường độ âm, người ta lấy I_o là ngưỡng nghe của âm có tần số 1000Hz gọi là tần số âm chuẩn.

Tai người nghe thính nhất với các âm trong miền tần số 1000Hz - 4000Hz, và nghe âm cao thính hơn nghe âm trầm.

1.4.5 Âm sắc (phonetics)

Âm sắc là sắc thái của âm thanh. Hầu hết các âm thanh trong tự nhiên cũng như âm thanh trong lời nói đều phức hợp, được tạo thành từ các âm cơ bản, các họa âm bậc cao về cao độ và cường độ.

Âm sắc là một đặc tính sinh lý của âm, được hình thành trên cơ sở các đặc tính vật lý của âm là tần số và biên độ. Thực nghiệm chứng tỏ rằng khi một nhạc cụ hoặc một người phát ra một âm có tần số f_1 thì đồng thời cũng phát ra các âm có tần số

$f_2 = 2f_1$; $f_3 = 3f_1$; $f_4 = 4f_1$ …

Âm có tần số f_1 gọi là âm cơ bản (hay họa âm thứ nhất), các âm có tần số f_2, f_3, f_4… gọi là các hoạ âm thứ hai, thứ ba, thứ tư… Tùy theo cấu trúc từng loại nhạc cụ, hoặc cấu trúc khoang miệng và cổ họng từng người mà trong số các hòa âm cái nào có biên độ lớn, biên độ nhỏ và cái nào chóng bị tắt đi. Do hiện tượng đó, âm phát ra không còn là đường sin, mà trở thành một đường phức tạp có chu kỳ. Sự tương quan giữa âm cơ bản và các hoạ âm mà tạo nên âm sắc khác nhau. Sự khác nhau về âm sắc là do sự phân bố vị trí môi, lưỡi, vòm miệng của từng người.

1.5 Mục tiêu của đề tài

Cho đến nay, các hệ thống nhận dạng tiếng nói tiếng Việt khá thành công chủ yếu là dựa trên khuynh hướng nhận dạng mẫu đơn giản. Trong khi đó, phương pháp sử dụng trí tuệ nhân tạo vào nhận dạng tiếng nói còn chưa nhiều, mặc dù mạng nơron là một công cụ rất mạnh và hứa hẹn nhiều ứng dụng mới. Đặc biệt là ở Việt Nam, việc áp dụng mạng nơron vào các lĩnh vực ứng dụng là rất ít và mới chỉ cho những kết quả ban đầu. Vì vậy, mục tiêu của đề tài là thử nghiệm mạng nơron lan truyền ngược để nhận dạng mười từ số đếm tiếng Việt từ: không đến chín.

Cụ thể là:

- Thiết kế mạng nơron lan truyền ngược để nhận dạng các từ tiếng Việt đơn âm tiết là số đếm và phân tích các tham số của mạng.
- Thử nghiệm nhận dạng với một số người nói.

Tài liệu tham khảo

Tiếng Việt

1. Mai Ngọc Chừ, Vũ Đức Nghệu, Hoàng Trọng Phiến (2000), *Cơ sở ngôn ngữ học và Tiếng Việt*, Nhà xuất bản Giáo dục.

2. Bùi Công Cường, Nguyễn Doãn Phước (2002*), Hệ mờ mạng nơron và ứng dụng*, Nhà xuất bản Khoa học kỹ thuật.
3. Đặng Ngọc Đức, Lương Chi Mai (3/2004), Tăng cường độ chính xác của hệ thống mạng nơron nhận dạng tiếng Việt, *Tạp chí Bưu chính Viễn thông, số 11*.
4. Nguyễn Quang Hoan (1996), *Ổn định mạng nơron Hopfield và khả năng ứng dụng trong điều khiển Robot*, Luận án Tiến sỹ.
5. Nguyễn Quang Hoan, Nguyễn Mạnh Tùng, Phạm Thượng Hàn (2002), Ứng dụng mạng nơron tương tác bậc cao cho bài toán phân lớp có giới hạn, tr.126-131, *Tuyển tập báo cáo khoa học, Hội nghị toàn Quốc lần thứ năm về tự động hoá.*
6. Ngô Hoàng Huy, Lương Chi Mai, Bùi Quang Trung, Nguyễn Thị Thanh Mai, Vũ Kim Bảng, Vũ Thị Thanh Hà (2003), Thiết kế các hệ thống nhận dạng tiếng Việt trong thời gian thực, *Kỷ yếu hội thảo Fair*.
7. Nguyễn Thanh Phúc (2000), *Một phương pháp nhận dạng lời Việt: Áp dụng phương pháp kết hợp mạng nơron với mô hình Markov ẩn cho các hệ thống nhận dạng lời Việt*, Luận Văn Tiến Sỹ kỹ thuật, Đại học Bách khoa Hà nội.

Tiếng Anh

8. Cart G. Looney (1997), *Parttern Recognition Using Neural Network*, Oxfoxd University Press.
9. Chin – Teng Lin, C. S. George Lee (1996), *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice-Hall International, Inc.
10. Claudio Becchetti and Lucio Prina Ricotti (1999), *Speech Recognition Theory and C++ Implementation*, Printed and Bound Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire.
11. Hong–Goo Kang (2003), *Speech Signal Processing*, Yonsei University.
12. Hunt, K. J and Others (1992), Neural Networks for Control System – A Survey, Automatica.. Vol. 28, No.6, pp. 1080-1120.
13. José C. Principe, Neil R. Euliano, W. Curt Lefebvre (1999), *Neural and Adaptive Systems: Fundamentals through Simulations*, John Wiley and Sons, Inc.
14. L.R Rabiner, R.W.Sharfer (1978), *Digital Processing of Speech Signals*, Prentice-Hall.

15. Ravi P. Ramachandra, Kevin R. Farell, Roopashri Ramachandra, Richard J. Mammone (2002), *Speaker Recognition - General Classifier Approaches and Data Fusion Methods*
16. Qifeng Zhu and Abeer Alwan (2003), *Non-linear Feature Extraction for Robust Speech Recognition in Stationary and Non-Stationary Noise,* Q. Zhu, A. Alwan/Computer Speech and Language .
17. Sadaoki Furui (2001), *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker.
18. Simon Haykin (1999), *Neural Networks A Comprehensive Foundation*, Prentice Hall International, Inc.
19. Veronique Stouten, Huge Van Hamme, Kris Demuynck, Patrick Wambacq (2003), *Robust Speech Recognition Using Model-Based Feature Enhancement, Center for Processing Speech and Images (PSI)* Dept of Electrical Engeneering–ESAT Katholieke Universiteit Leuven, Belgium.
20. Wu Chou and Biing Hwang Juang (2003), *Parttern Recognition in Speech and Language Processing*, CRC Press LLC.