LỜI CẢM ƠN

Em xin chân thành cảm ơn Thầy giáo, Thạc sĩ Võ Văn Tùng – Công tác tại Cục kỹ thuật nghiệp vụ I, Bộ công an, người đã trực tiếp hướng dẫn tận tình chỉ bảo em trong suốt quá trình làm tốt nghiệp.

Em xin chân thành cảm ơn tất cả các thầy cô giáo trong khoa Công nghệ thông tin - Trường ĐHDL Hải Phòng, những người đã nhiệt tình giảng dạy và truyền đạt những kiến thức cần thiết trong suốt thời gian em học tập tại trường, để em hoàn thành tốt đề tài này.

Em cũng xin chân thành cảm ơn Ban lãnh đạo, tất cả các cô chú, các anh chị tại Công ty Cổ phần Thiết bị Bưu điện, đã giúp đỡ và tạo mọi điều kiện tốt cho em trong thời gian thực tập và làm tốt nghiệp tại Trung tâm.

Trong quá trình làm tốt nghiệp tuy có nhiều cố gắng nhưng không thể tránh khỏi những thiếu sót, em rất mong nhận được sự góp ý quý báu của tất cả các thầy cô giáo, của hội đồng phản biện và của tất cả các bạn.

Em xin chân thành cảm ơn!

Hải Phòng, ngày .....tháng 7 năm 2009

Sinh viên

Trương Ngọc Sơn.

MỤC LỤC

MỞ ĐẦU

Ngày nay, cùng với sự phát triển nhanh chóng của công nghệ thông tin, trong đó có công nghệ xử lý âm thanh. Đặc biệt trong lĩnh vực xử lý âm thanh trong nhận dạng tiếng Việt có một ý nghĩa quan trọng mang lại nhiều ứng dụng thiết thực cho xã hội, mang lại những thay đổi mang tính cách mạng trong nhiều lĩnh vực, phát thanh, truyền hình, viễn thông... Trong vài thập kỷ gần đây, nhận dạng là một vấn đề cuốn hút nhiều nhà khoa học ở các lĩnh vực khác nhau : Toán học, điều khiển, điện tử, sinh học ... Trước sự phát triển mạnh mẽ của công nghệ thông tin, vấn đề nhận dạng càng được quan tâm nhiều hơn nhằm nâng cao hiệu quả giao tiếp người - máy.

Trên thế giới, các ngôn ngữ phổ biến như Anh, Pháp... đã có nhiều phần mềm nhận dạng rất hiệu quả. Ở Việt Nam đã có nhiều công trình nghiên cứu về lĩnh vực nhận dạng tiếng nói (Speech recognition) trên cơ sở lý thuyết các hệ thống thông minh nhân tạo, nhiều kết quả đã trở thành sản phẩm thương mại như ViaVoice, Dragon..., các hệ thống bảo mật thông qua nhận dạng tiếng nói các hệ quay số điện thoại bằng giọng nói... Triển khai những công trình nghiên cứu và đưa vào thực tế ứng dụng vấn đề này là một việc làm hết sức có ý nghĩa đặc biệt trong giai đoạn công nghiệp hoá hiện đại hoá hiện nay của nước ta.

Mục đích của đề tài là nghiên cứu xây dựng một chương trình nhận dạng tiếng nói tiếng Việt trong môi trường có nhiễu với đầu vào là tập từ hạn chế là tiếng việt sau đó so sánh với các mẫu có sẵn để đưa ra kết quả. Ngoài phần mở đầu và kết luận đồ án gồm 4 chương:

Chương 1 : Tín hiệu – Cơ sở xử lý Tín hiệu

Chương 2 : Đặc trưng Tiếng Việt

Chương 3 : Bài toán nhận dạng Tiếng nói

Chương 4: Chương trình Demo

CHƯƠNG 1: TÍN HIỆU - CƠ SỞ XỬ LÝ TÍN HIỆU

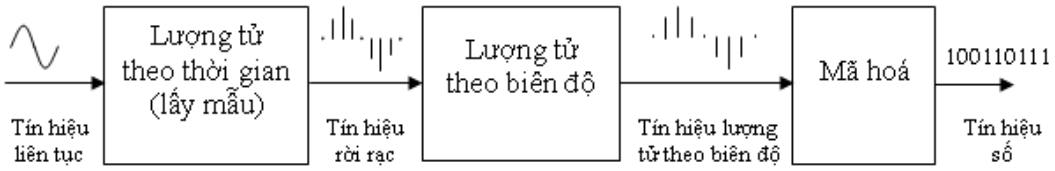
Cơ sở của xử lý tín hiệu chính là bước đầu của quá trình nhận dạng tiếng nói, khi bạn nói một từ máy sẽ thu giọng của bạn, tiếng nói sẽ được biểu diễn dưới dạng tín hiệu, qua quá trình xử lý tín hiệu, tiếng nói đầu vào sẽ được đối chiếu với tập mẫu mà máy đã được học sẵn để đưa ra kết quả. Dưới đây chính là một sô cách nhìn tổng quan về tín hiệu.

1.1. Tín hiệu

Tín hiệu về mặt toán học là hàm biểu diễn trạng thái vật lý của thông tin. Nói chung, tín hiệu là một hàm phức tạp của nhiều thông số. Để đơn giản chúng ta coi tín hiệu là hàm của biến thời gian - tín hiệu có 3 dạng cơ bản:

- Tín hiệu liên tục (tương tự).
- Tín hiệu rời rạc (lấy mẫu).
- Tín hiệu số.

Ba loại tín hiệu này có mặt ở các vị trí của sơ đồ hình 1.1



Hình 1.1. Sơ đồ biến đổi tín hiệu liên tục thành tín hiệu số

Tín hiệu liên tục là tín hiệu được biểu diễn bằng hàm số có biến số thời gian độc lập (*hình 1.2a*).

Tín hiệu rời rạc (*còn gọi là tín hiệu trích mẫu*) là dãy giá trị tín hiệu liên tục ở từng thời điểm rời rạc và tín hiệu đó được biểu diễn dưới dạng một dãy số (*hình 1.2b*). Tín hiệu rời rạc gặp ở đầu ra mạch lượng tử theo thời gian (*mạch trích mẫu*).

Tín hiệu rời rạc lượng tử theo biên độ là tín hiệu được lượng tử theo biên độ, thực chất là dãy giá trị mẫu được quy tròn theo các mức lượng tử biên độ (*hình 1.2c*). Tín hiệu này gặp ở đầu ra bộ lượng tử biên độ.

Tín hiệu số là tín hiệu lượng tử theo biên độ và mã hoá (*hình 1.2d*). Các dạng tín hiệu vừa nêu trên được mô tả trên hình 1.2.

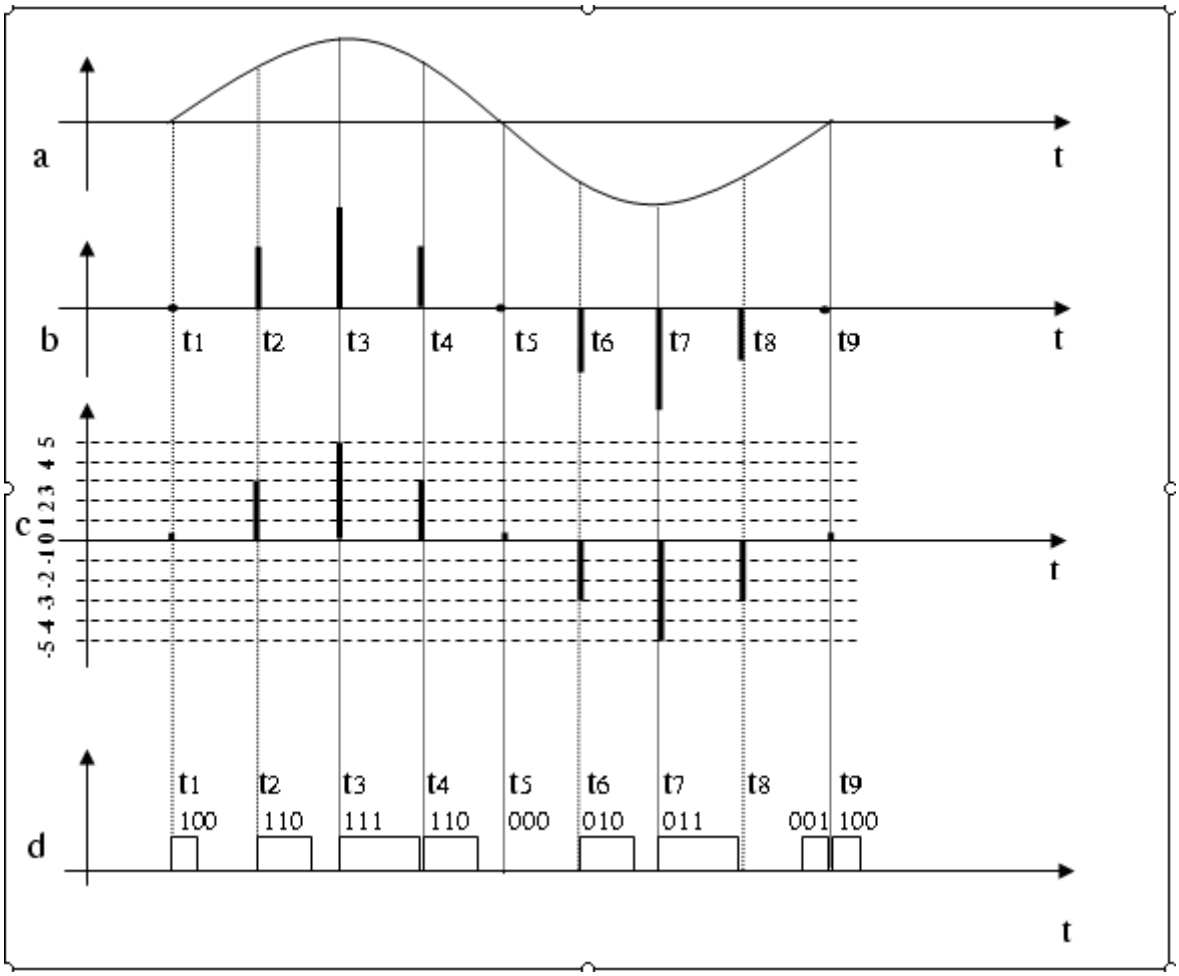
a. Tín hiệu tương tự.

b. Tín hiệu rời rạc (lấy mẫu).

c. Tín hiệu rời rạc lượng tử theo biên độ (lượng tử hoá).

d. Tín hiệu số (gán các bít cơ 2 cho các mẫu đã làm tròn).

Các kiểu tín hiệu này được biểu diễn trong hình 1. 2



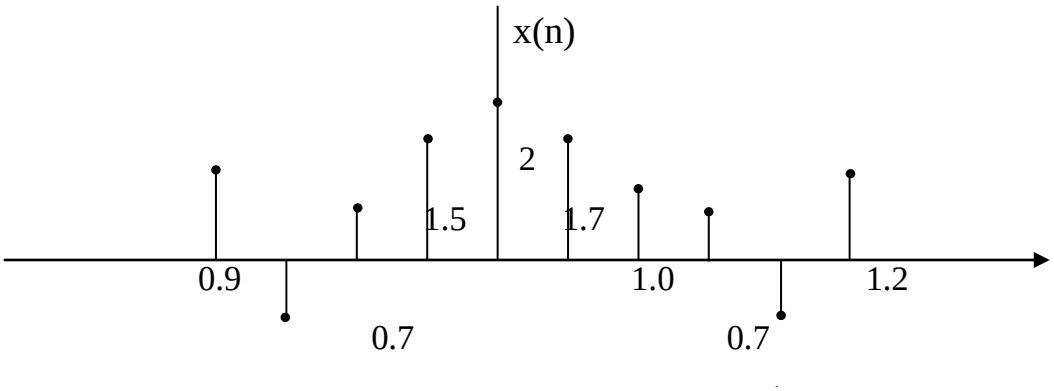
Hình 1.2. mô tả các dạng tín hiệu

1.2. Các tín hiệu rời rạc theo thời gian

1.2.1 Các phương pháp biểu diễn tín hiệu rời rạc

Như ta đã biết, tín hiệu rời rạc theo thời gian *x(n)* thực chất là hàm của biến độc lập có kiểu số nguyên. tín hiệu *x(n)* chỉ được định nghĩa đối với các giá trị nguyên của *n*. Trong khi nghiên cứu, chúng ta giả sử rằng tín hiệu rời rạc theo thời gian được định nghĩa đối với giá trị nguyên của *n* thuộc khoảng $-\infty < n < \infty$. Theo qui ước xem *x(n)* như là "mẫu thứ *n*" của tín hiệu, Nếu cho rằng *x(n)* là tín hiệu nhận được do quá trình lấy mẫu của tín hiệu tương tự *xa(t)* thì $x(n) \equiv x(nT)$, trong đó *T* là chu kỳ lấy mẫu (*thời gian giữa hai lần lấy mẫu liên tiếp nhau*)

Trong tài liệu khi viết *x(n)* như là cách viết đơn giản của *x(nT)* hoặc sẽ hiểu là T=1.



Hình 1.3. Biểu diễn đồ thị của tín hiệu rời rạc theo thời gian.

Ngoài phương pháp sử dụng đồ thị như mô tả trên, còn có một số phương pháp khác tương đối thuận tiện được sử dụng để biểu diễn tín hiệu (*hoặc dãy*) rời rạc theo thời gian.

a. Biểu diễn bằng hàm

Ví dụ: $x(n) = \begin{cases} 1, & \text{với n = 1,3} \\ 4, & \text{với n = 2} \\ 0, & \text{với các giá trị còn lại} \end{cases}$

b. Biểu diễn bằng bảng

Ví dụ:

n	… -2`	-1	0	1	2	3	4	5	…
x(n	*.... 0*	*0*	*0*	*1*	*4*	*1*	*0*	*0*	*...*

c. Biểu diễn qua dãy số

Tín hiệu hoặc dãy vô tận được mô tả qua ví dụ dưới đây.

$$x(n) = \{\ldots 0,0\ \underset{\uparrow}{1},4,1,0,0\ldots\}$$

trong ký hiệu ↑ dùng để ↑ chỉ thời điểm gốc *(n = 0).*

Dãy *x(n)* có giá trị bằng 0 với *n < 0* được biểu diễn bằng cách sau:

$$x(n) = \{\underset{\uparrow}{0},1,4,1,0,0\ldots\}$$

ở đây thời điểm gốc với dãy *x(n)* có giá trị bằng 0 nếu *n<0* được hiểu như là điểm bên trái nhất của dãy.

Dãy hữu hạn có thể được biểu diễn bằng cách:

$$x(n) = \{3,-1,\underset{\uparrow}{-2},5,0,4,-1\}$$

Nếu dãy hữu hạn thoả mãn điều kiện *x(n)* = 0 với *n<0* thì dãy có thể được biểu diễn theo cách sau:

$$x(n) = \{\underset{\uparrow}{0},1,4,1\}$$

1.2.2 Một vài tín hiệu rời rạc cơ bản

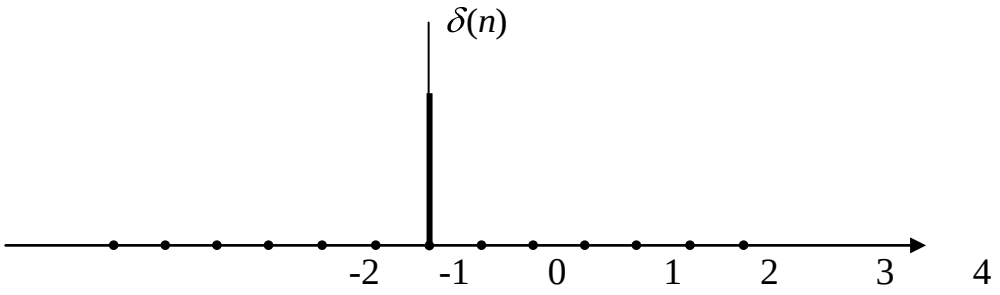
a. Dãy mẫu đơn vị

Tín hiệu này còn được gọi là dãy xung đơn vị và được định nghĩa như sau:

$$\delta(n) \equiv \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}$$

Như vậy, dãy mẫu đơn vị là tín hiệu chỉ có một giá trị duy nhất bằng đơn vị tại thời điểm *n = 0* trong khi tất cả các giá trị còn lại đều bằng 0.

Tín hiệu dãy xung đơn vị được mô tả bằng đồ thị sau:



1.4 Biểu diễn đồ thị của tín hiệu mẫu đơn vị

b. Dãy nhảy bậc đơn vị

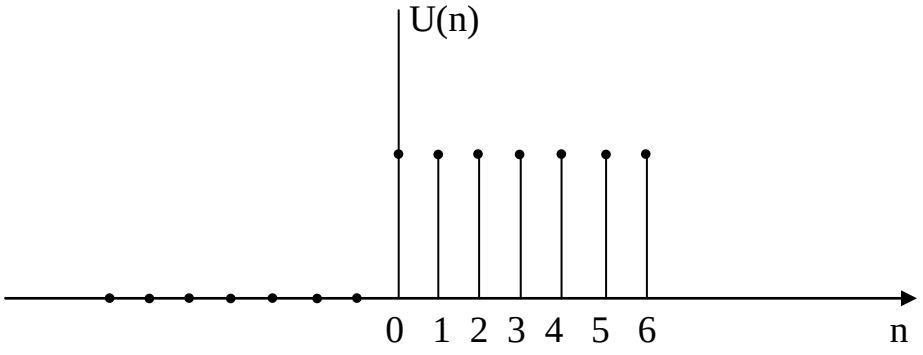
Dãy này còn được gọi là tín hiệu nhảy bậc đơn vị hay hàm bậc thang và được định nghĩa qua hàm sau:

$$u(n) \equiv \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

Giữa tín hiệu nhẩy bậc đơn vị và tín hiệu xung đơn vị có mối quan hệ:

$u(n) = \sum_{k=0}^{\infty} \delta(n-k)$ và $\delta(n) = u(n) - u(n-1)$

Tín hiệu nhảy bậc đơn vị được mô tả trên hình sau:



1.5 Biểu diễn bằng đồ thị của tín hiệu nhãy bậc đơn vị

1.2.3 Phân loại các tín hiệu rời rạc

Các phương pháp toán học được dùng trong việc phân tích tín hiệu và hệ thống rời rạc theo thời gian hoàn toàn phụ thuộc vào đặc thù của tín hiệu.

a. Tín hiệu năng lượng và tín hiệu công suất

Năng lượng E của tín hiệu *x(n)* được định nghĩa bằng công thức:

$$E \equiv \sum_{n=-\infty}^{\infty} |x(n)|^2 ,$$

ở đây $|x(n)|$ là modul của tín hiệu. Với cách định nghĩa này thì công thức trên có thể được sử dụng để tính năng lượng của tín hiệu phức cũng như của tín hiệu thực.

Năng lượng của tín hiệu có thể là hữu hạn hoặc vô hạn. Nếu E là hữu hạn ($0 < E < \infty$) thì *x(n)* được gọi là tín hiệu năng lượng. Để phân biệt năng lượng của tín hiệu rời rạc, thông thường người ta sử dụng thêm chỉ số *x* đối với *E* và biết là *Ex.*

Rất nhiều tín hiệu với năng lượng vô hạn lại có công suất hữu hạn. Công suất trung bình của tín hiệu rời rạc theo thời gian *x(n)* được định nghĩa bằng biểu thức:

$$P = \lim_{n \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \left| x(n)^2 \right|$$

Nếu định nghĩa năng lượng tín hiệu của dãy *x(n)* trong khoảng hữu hạn $-N \leq n \leq N$ *là:*

$$EN \equiv \sum_{n=-N}^{N} |x(n)|^2$$

thì có thể xác định năng lượng tín hiệu *E* qua biểu thức

$$E \equiv \lim_{N \to \infty} EN$$

và công suất trung bình của tín hiệu *x(n):*

$$P \equiv \lim_{N \to \infty} \frac{1}{2N+1} E_N$$

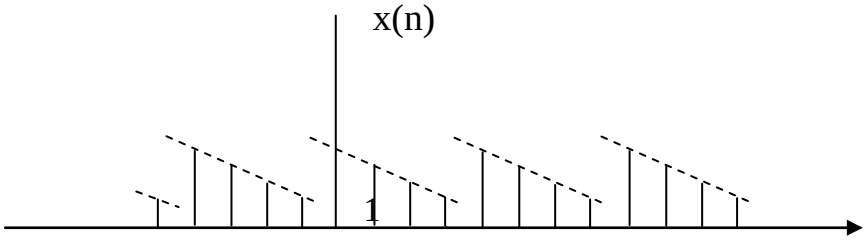
Rõ ràng rằng nếu *E* là hữu hạn thì *P=0.* Trong khi đó nếu *E* là vô hạn thì công suất trung bình *P* có thể là hữu hạn hoặc vô hạn. Nếu *P* là hữu hạn (và khác 0) tín hiệu sẽ được gọi là tín hiệu công suất.

b. Tín hiệu tuần hoàn và không tuần hoàn

Như đã định nghĩa trong phần 1.3 tín hiệu *x(n)* được gọi là tuần hoàn với chu kỳ *N(N>0)* khi và chỉ khi:

x(n + N) = x(n) với mọi *n*

Giá trị nhỏ nhất của N thoả mãn biểu thức trên được gọi là chu kỳ cơ bản. Nếu không có bất cứ một giá trị nào của N để bt trên là đúng thì tín hiệu được gọi là không tuần hoàn. Hình dưới là một ví dụ về tín hiệu tuần hoàn.



Mô tả bằng đồ thị tín hiệu tuần hoàn

Khi khảo sát tín hiệu hình sin ta nhận thấy rằng tín hiệu.

$x(n) = A\sin 2\pi f_0 n$

là tín hiệu tuần hoàn nếu f_0 là một số hữu tỷ, hay nói cách khác f_0 có thể được biểu diễn qua biểu thức:

$$f_0 = \frac{k}{N}$$

trong đó k và N là những số nguyên.

Năng lượng của tín hiệu tuần hoàn x*(n)* trong một chu kỳ hay trong một khoảng $0 \le n \le N-1$ là hữu hạn nếu x*(n)* nhận các giá trị hữu hạn trong một chu kỳ. Tuy vậy, năng lượng của tín hiệu tuần hoàn với $-\infty \le n \le \infty$ là vô hạn. Mặt khác, công suất trung bình của tín hiệu tuần hoàn là hữu hạn và bằng công suất trung bình trong một chu kỳ. Như vậy, nếu x*(n)* là tín hiệu tuần hoàn với tần số cơ bản N và có các giá trị hữu hạn thì công suất của nó được xác định qua biểu thức:

$$P = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2$$

Suy ra rằng tín hiệu tuần hoàn là tín hiệu công suất.

c. Tín hiệu đối xứng (chẵn) và tín hiệu không đối xứng (lẻ)

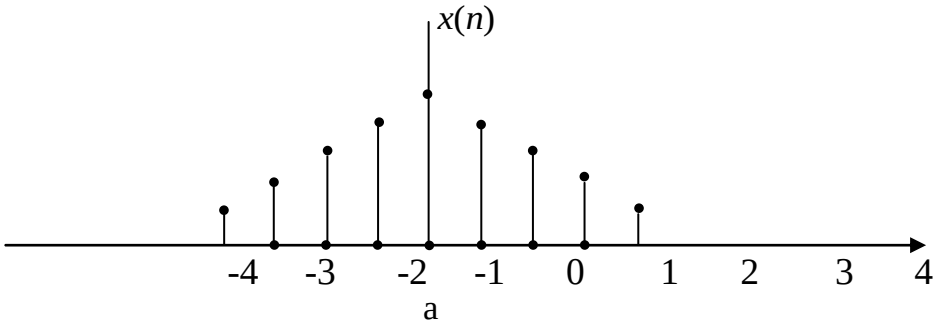
Tín hiệu có giá trị thực x*(n)* được gọi là đối xứng (chẵn) nếu:

$$x(-n) = x(n)$$

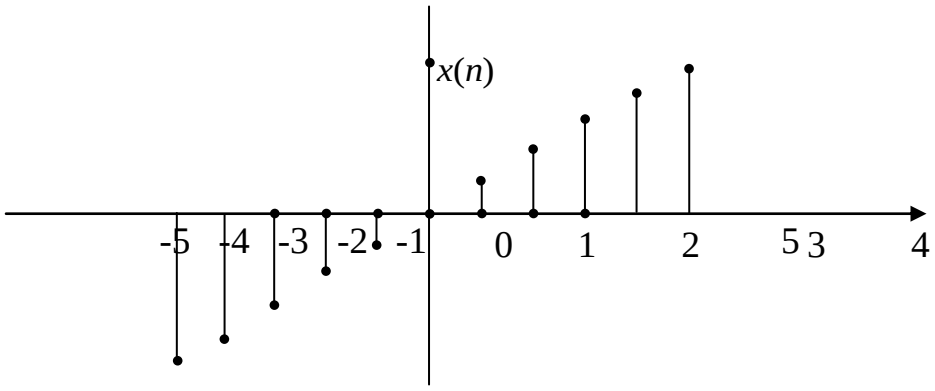
và được gọi là phản đối xứng (lẻ) nếu:

$$x(-n) = -x(n)$$

Có thể nhận thấy rằng nếu x*(n)* là lẻ thì x*(0)* =0



a



Tín hiệu chẵn có thể được biểu diễn qua công thức:

$$x_e(n) = \frac{1}{2}[x(n) + x(-n)]$$

Tín hiệu lẻ có thể được biểu diễn qua công thức

$$x_0(n) = \frac{1}{2}[x(n) - x(-n)]$$

Như vậy nếu x*(n)* là tín hiệu bất kỳ thì ta có thể biểu diễn x*(n)* dưới dạng sau:

$$x(n) = \frac{1}{2}[x(n) + x(n) + x(-n) - x(-n)]$$

$$= \frac{1}{2}[x(n) + x(-n)] + \frac{1}{2}[x(n) - x(-n)$$

$$= \mathrm{x}\, e(n) + xo(n)$$

Như vậy một tín hiệu bất kỳ có thể được biểu diễn dưới dạng tổng của tín hiệu chẵn và tín hiệu lẻ

1.2.4 Các thao tác xử lý đơn giản trên tín hiệu rời rạc theo thời gian.

Trong phần này ta sẽ xem xét một vài xử lý đơn giản liên quan đến các biến độc lập và biên độ của tín hiệu.

a.Phép dịch các biến độc lập.

Tín hiệu x(n) có thể được dịch chuyển theo thời gian bằng cách thay thế biến độc lập n bởi n- k trong đó k là số nguyên. Nếu k là số nguyên dương thì kết quả của sự dịch chuyển về thời gian sẽ là sự trễ của tín hiệu với k đơn vị của thời gian. Nếu k là số âm thì kết quả của sự dịch chuyển theo thời gian là sự vượt trước của tín hiệu với k đơn vị thời gian.

b. Phép nhân, cộng và phép lấy tỷ lệ.

Việc thay đổi của biên độ tín hiệu rời rạc theo thời gian có thể được thực hiện qua các phép toán (thao tác) *cộng, nhân, lấy tỷ lệ*.

Lấy tỷ lệ còn được gọi là phép nhân của dãy với hằng số và thực hiện bằng cách nhân giá trị của mỗi mẫu với chính hằng số đó. Giả sử rằng số được ký hiệu là A, khi đó ta có thể viết:

$$y(n) = Ax(n), -\infty \le n \le \infty$$

Tổng của hai tín hiệu *x1(n)* và *x2(n)* là một tín hiệu *y(n)* với giá trị ở mỗi thời điểm bằng tổng các giá trị *x1(n)* và *x2(n)* tương ứng ở thời điểm đó và như vậy:

$$y(n) = x1(n) + x2(n), -\infty \le n \le \infty$$

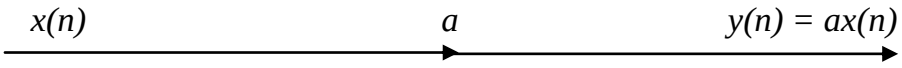
Tích của hai tín hiệu là một tín hiệu khác với giá trị ở mỗi thời điểm bằng tích các giá trị của hai tín hiệu ở thời điểm tương ứng, hay:

$$y(n) = x1(n).x2(n), -\infty \leq n \leq \infty$$

1.2.5 Biểu diễn hệ thống rời rạc theo thời gian bằng sơ đồ khối

a. Bộ nhân với hằng số (constant muLTIplier)

Phép toán này được mô tả trên hình dưới và biểu diễn một phép lấy tỷ lệ của tín hiệu đầu vào *x(n).*

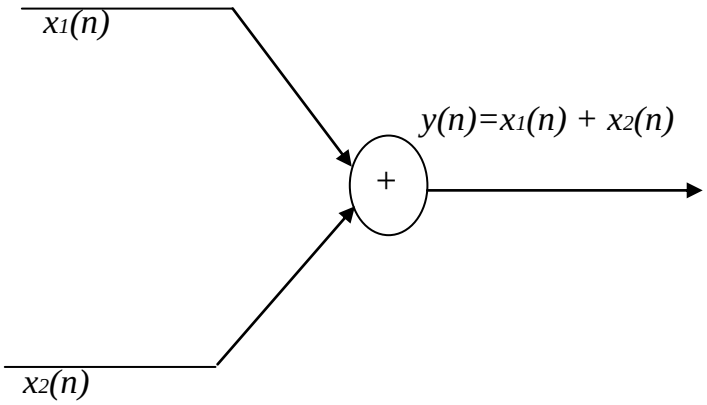


Biểu diễn sơ đồ của hệ nhân với hằng số.

b. Bộ cộng (Adder)

Hình dưới mô tả một hệ thống (bộ cộng) thực hiện cộng hai dãy tín hiệu với kết quả là một dãy khác - dãy *y(n)* (dãy tổng).

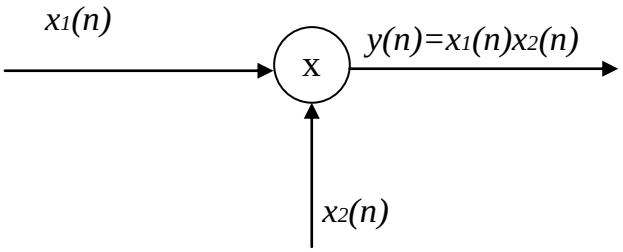
Trong quá trình thực hiện thao tác cộng ta không cần phải lưu trữ bất cứ một giá trị trung gian nào bởi vì phép cộng được thực hiện tức thì không nhớ.



Biểu diễn qua sơ đồ của bộ cộng.

c. Bộ nhân tín hiệu (signal muLTIplier)

biểu diễn một bộ nhân của hai dãy tín hiệu với kết quả là một dãy tích *y(n)*. Cũng giống như hai trường hợp trước, ở đây phép nhân cũng là phép toán không nhớ.

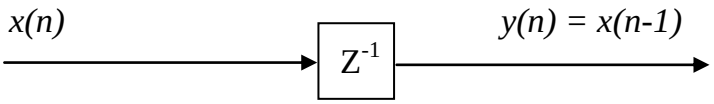


Biểu diễn qua sơ đồ của hệ nhân.

d. Phần tử trễ đơn vị

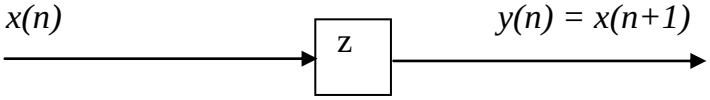
Phần tử trễ đơn vị (unit delay element) là hệ thống đặc biệt có tác dụng làm trễ tín hiệu đi qua với thời gian bằng một đơn vị. hệ thống này là hệ thống có nhớ

Trong miền Z, phần tử này được ký hiệu bởi z^{-1}. sơ đồ biểu diễn



e. Phần tử vượt trước đơn vị (Unit advance element)

Trái ngược với hệ trễ đơn vị, hệ vượt trước đơn vị sẽ chuyển đầu vào x*(n)* dịch về trước một mẫu theo thời gian để có thể nhận được ở đầu ra tín hiệu y*(n)* = x*(n+1)*.



Biểu diễn qua sơ đồ của phần tử vượt trước.

1.2.6 Phân loại các hệ thống rời rạc theo thời gian

a. Hệ nhớ và không nhớ

Hệ thống rời rạc theo thời gian được gọi là không nhớ (*memoryless*) hoặc tĩnh (*static*) nếu tín hiệu ra của nó ở mọi thời điểm chỉ phụ thuộc vào tín hiệu đầu vào ở cùng một thời điểm mà không phụ thuộc vào các giá trị mẫu của tín hiệu đầu vào trong quá khứ hoặc trong tương lai. Trong trường hợp ngược lại, hệ thống được gọi là có nhớ hoặc biến đổi (*dynamic*). Nếu đầu ra của hệ thống ở thời điểm n có thể được xác định một cách hoàn toàn bởi các mẫu đầu vào trong khoảng từ n-N đến n ($N \geq 0$) thì hệ thống được gọi là có nhớ trong khoảng N. Nếu $N = 0$ thì hệ sẽ là hệ không nhớ. Nếu $0 < N < \infty$ hệ thống được gọi là hệ nhớ hữu hạn, ngược lại nếu $N = \infty$ thì hệ được gọi là hệ nhớ vô hạn.

b. Hệ thống bất biến và không bất biến theo thời gian

Một hệ được gọi là bất biến theo thời gian nếu như đặc trưng vào/ra của nó không thay đổi theo thời gian

Định lý. Một hệ thống relaxed được gọi là bất biến theo thời gian khi và chỉ khi:

$x(n) \xrightarrow{T} y(n)$

suy ra $x(n-k) \xrightarrow{T} y(n-k)$

đối với mọi tín hiệu đầu vào $x(n)$ và mọi thời gian dịch chuyển k.

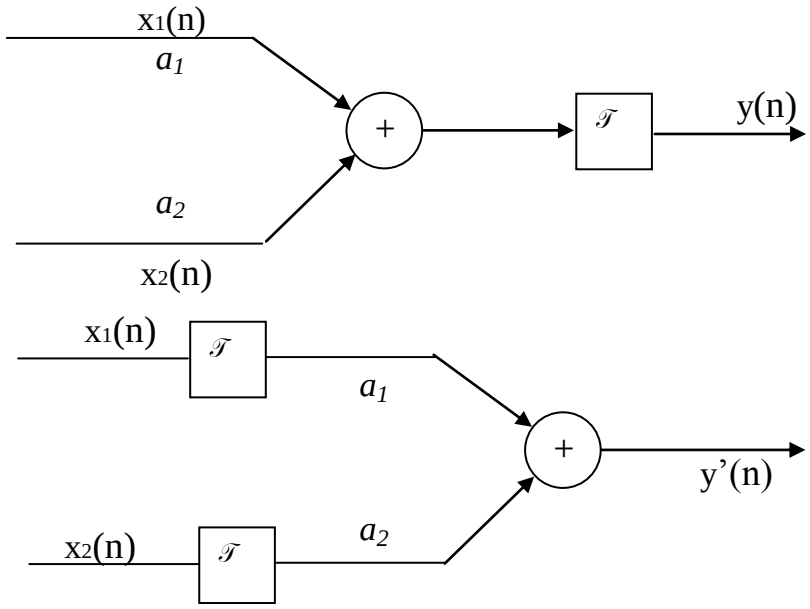
c. Hệ tuyến tính và không tuyến tính

Các hệ thống có thể được chia làm hai loại tuyến tính và không tuyến tính. Hệ thống được gọi là tuyến tính nếu nó thỏa mãn nguyên lý xếp chồng

Định lý : Hệ thống được xem là tuyến tính khi và chỉ khi:

$$T[a_1x_1(n) + a_2x_2(n)] = a_1\,T[x_1(n)] + a_2T[x_2(n)]$$

đối với mọi dãy tín hiệu đầu vào $x_1(n)$, $x_2(n)$ và các hằng số a1, a2



Biểu diễn đồ hoạ của nguyên tắc xếp chồng

CHƯƠNG 2: ĐẶC TRƯNG TIẾNG VIỆT

2.1. Đặc điểm của Tiếng Việt

Tiếng nói thường xuất hiện dưới nhiều hình thức mà ta gọi là đàm thoại , việc đàm thoại thể hiện kinh nghiệm của con người.Những người có điều kiện thể chất và tình thần bình thường thì rất dễ diễn đạt tiếng nói của mình do đó tiếng nói là phương tiện giao tiếp chính trong lúc đàm thoại.

Tiếng nói là âm thanh mang mục đích diễn đạt thông tin,là công cụ tư duy và trí tuệ,tiếng nói mang tính đặc trưng của loài người.

Tiếng Việt thuộc ngôn ngữ đơn lập, tức là mỗi một tiếng (âm tiết) được phát âm tách rời nhau và được thể hiện bằng một chữ viết. Đặc điểm này thể hiện rõ rệt ở tất cả các mặt ngữ âm, từ vựng, ngữ pháp.

2.2. Đặc điểm ngữ âm

Trong tiếng Việt có một loại đơn vị đặc biệt gọi là *"tiếng"*. Về mặt ngữ âm, một tiếng là một âm tiết. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối, tạo ra tiềm năng của ngữ âm tiếng Việt trong việc thể hiện các đơn vị có nghĩa. Nhiều từ tượng hình, tượng thanh có giá trị gợi tả đặc sắc. Khi tạo câu, tạo lời, người Việt rất chú ý đến sự hài hoà về ngữ âm, đến ngữ điệu của câu văn.

2.3. Đặc điểm từ vựng

Mỗi tiếng, nói chung, là một yếu tố có nghĩa. Tiếng là đơn vị cơ sở của hệ thống các đơn vị có nghĩa của tiếng Việt. Từ tiếng, người ta tạo ra các đơn vị từ vựng khác để định dạng sự vật, hiện tượng..., chủ yếu nhờ phương thức ghép và phương thức láy.

Việc tạo ra các đơn vị từ vựng ở phương thức ghép luôn chịu sự chi phối của quy luật kết hợp ngữ nghĩa. Theo phương thức này, tiếng Việt triệt để sử dụng các yếu tố cấu tạo từ thuần Việt hay vay mượn từ các ngôn ngữ khác để tạo ra các từ, ngữ mới, ví dụ: *tiếp thị, karaoke, thư điện tử* (e-mail),

thư thoại (voice mail), *phiên bản* (version), *xa lộ thông tin, liên kết siêu văn bản, truy cập ngẫu…*

Việc tạo ra các đơn vị từ vựng ở phương thức láy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng ví dụ *chôm chỉa, chỏng chơ, đỏng đà đỏng đảnh, thơ thẩn, lúng la lúng liếng,* v.v.

2.4. Đặc điểm ngữ pháp

Từ của tiếng Việt không biến đổi hình thái. Đặc điểm này sẽ chi phối các đặc điểm ngữ pháp khác. Khi từ kết hợp từ sẽ trở thành các kết cấu như *ngữ, câu.* Trong tiếng Việt khi nói "Anh ta lại đến" là khác với "Lại đến anh ta", Nhờ trật tự kết hợp của từ mà *"củ cải"* khác với *"cải củ"*, *"tình cảm"* khác với *"cảm tình"*. Trật tự chủ ngữ đứng trước, vị ngữ đứng sau là trật tự phổ biến của kết cấu câu tiếng Việt

Tiếng Việt rất coi trọng phương thức trật tự từ và hư từ ngoài ra trong tiếng Việt còn dùng phương thức là ngữ điệu.

Phương thức hư từ cũng là phương thức ngữ pháp chủ yếu của tiếng Việt. Nhờ hư từ mà tổ hợp *"anh của em"* khác với tổ hợp *"anh và em"*, *"anh vì em*". Hư từ cùng với trật tự từ cho phép tiếng Việt tạo ra nhiều câu cùng có nội dung thông báo cơ bản như nhau nhưng khác nhau về sắc thái biểu cảm. Ví dụ, so sánh các câu sau đây:

- *Ông ấy không hút thuốc*
- *Thuốc, ông ấy không hút*

Ngữ điệu giữ vai trò trong việc biểu hiện quan hệ cú pháp của các yếu tố trong câu, nhờ đó nhằm đưa ra nội dung muốn thông báo. Trên văn bản, ngữ điệu thường được biểu hiện bằng dấu câu. Chúng ta thử so sánh hai câu sau để thấy sự khác nhau trong nội dung thông báo:

- *Đêm hôm qua, cầu gãy.*
- *Đêm hôm, qua cầu gãy.*

Qua một số đặc điểm nổi bật vừa nêu trên đây, chúng ta có thể hình dung được phần nào bản sắc và tiềm năng của tiếng Việt.

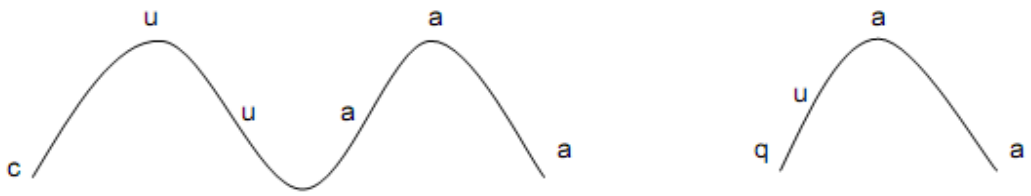
2.5. Âm tiết trong tiếng Việt

Âm tiết là âm vị nhỏ nhất khi nói. Dù phát âm có thật chậm,thật tách bạch thì những âm thanh của phát ngôn cũng không thể chia nhỏ được nữa. Mỗi âm tiết tiếng Việt là một khối hoàn chỉnh trong phát âm, nhưng không phải là một khối bất biến mà có cấu tạo lắp ghép. Khối lắp ghép ấy có thể tháo rời từng bộ phận của âm tiết này để hoán vị với bộ phận tương ứng của các âm tiết khác.

Ví dụ:

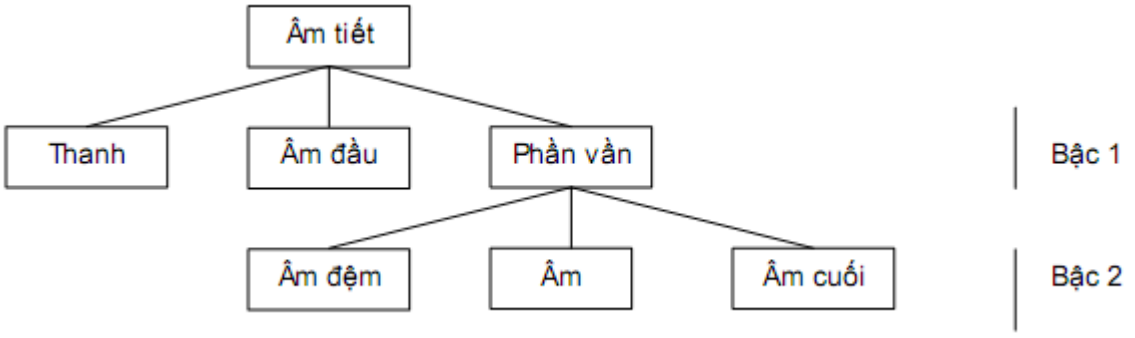
tiền đâu → đầu tiên	đảo tật tự âm tiết và hoán vị thanh điệu “”
hiện đại → hại điện	hoán vị phần sau “iên” cho “ai”
nhỉ đay → nhảy đi	thanh điệu giữ nguyên vị trí cùng với phần đầu “nh” và “đ”

Âm tiết vì thế có tính toàn vẹn được phát âm bằng một đợt căng của bộ máy phát âm.Các đợt căng của cơ nối tiếp nhau làm thành một chuỗi âm tiết và có thể hình dung bằng một chuỗi đường cong hình sin .



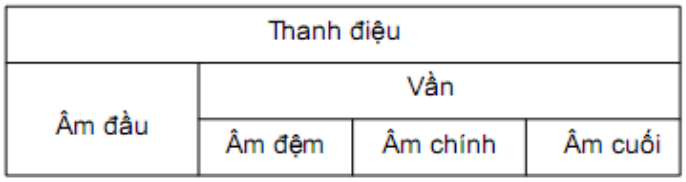
Trong sơ đồ trên là hai cách phát âm “cụ ạ” và ”quạ”.Trong phát âm thứ nhất có 2 âm tiết,âm [u] nằm ở đỉnh âm tiết đầu.Trong phát âm thứ hai có một âm tiết và âm [u] nằm ở sườn của âm tiết.

Cấu trúc tổng quát của một âm tiết trong tiếng Việt là



Còn đây là cấu trúc chặt chẽ của một âm tiết trong tiếng Việt

Mỗi âm tiết tiếng Việt ở dạng đầy đủ có 5 phần như hình:



Có thể hình dung về cấu tạo âm tiết tiếng việt trong một mô hình như sau:

Thanh điệu: không (zero), huyền (`), hỏi (?), ngã (~) Sắc ('), nặng (.)			
t Âm đầu	Vần		
	o	a	n
	Âm đệm	Âm chính	Âm cuối

Âm đầu: thường là phụ âm, được gọi là phụ âm đầu,nó có chức năng tạo ra âm sắc cho âm tiết lúc mở đầu.Âm đầu có thể vắng mặt trong một số trường hợp như khi ta nói an,ấm…

Âm đệm: Âm đệm là yếu tố đứng ở vị trí thứ hai, sau âm đầu. Nó tạo nên sự đối lập tròn môi *(voan)* và không tròn môi *(van)*, có chức năng làm thay đổi âm sắc của âm tiết lúc khởi đầu và làm khu biệt âm tiết này với âm tiết khác.ví dụ như ” tán” và ”toán”.Âm đệm có thể vắng mặt trong một số trường hợp khi có âm “u” và ”o”.

Âm chính : Âm chính đứng ở vị trí thứ ba trong âm tiết, là hạt nhân, là đỉnh của âm tiết, nó mang âm sắc chủ yếu của âm tiết. Âm chính luôn luôn có mặt trong mọi âm tiết có chức năng quy định âm sắc chủ yếu của âm tiết .Âm chính luôn luôn là nguyên âm.

Âm cuối : có thể là phụ âm hoặc là bán nguyên âm (tiếng việt có 2 bán nguyên âm là i và u). âm cuối có vị trí cuối cùng của âm tiết và có chức năng kết thúc âm tiết,do vậy khi có am cuối thì âm tiết ko có khả năng kết hợp với âm khác,vd như “cúi”…một số âm cuối vẫn có khả năng kết hợp với âm khác

như “quý”có thể thành “quýt” hay “quýnh” thì “y” vẫn được coi là âm cuối vì sau đó là có mặt của một âm cuối gọi là âm cuối “zezo”.

Thanh điệu : luôn có mặt trong âm tiết và có ý nghĩa quyết định âm tiết về độ cao. Tiếng Việt có 6 thanh điệu: thanh ngang (không dấu, tiếng Anh: zero /level), huyền (falling), ngã (broken), hỏi (curve), sắc (rising), nặng (drop).Có nhiều ý kiến khác nhau về vị trí của thanh điệu trong âm tiết. Nhưng ý kiến cho rằng thanh điệu nằm trong cả quá trình phát âm của âm tiết (nằm trên toàn bộ âm tiết) là đáng tin cậy nhất về vị trí của thanh điệu.

CHƯƠNG 3: BÀI TOÁN NHẬN DẠNG TIẾNG NÓI

Khái quát về nhận dạng

Hiện nay chưa có một định nghĩa chung nào về nhận dạng, nhưng về bản chất của quá trình nhận dạng một đối tượng chưa biết nào đó là sắp xếp đưa đối tượng chưa biết về lớp các đối tượng đã biết. Thực hiện việc so sánh để đưa ra kết luận đối tượng cần nhận dạng thuộc lớp đối tượng nào đã biết.

Những yếu tố cần quan tâm trong bài toán nhận dạng

Không gian biểu diễn quan sát: Là tập hợp các ký hiệu, số liệu miêu tả đối tượng sau quá trình cảm nhận.

Không gian đặc tính: là tập hợp các miêu tả đặc tính sau quá trình trích chọn đặc tính.

Không gian diễn dịch: là tập hợp các tên của các đối tượng hoặc tên của các lớp đối tượng cho biết đối tượng quan sát thuộc về lớp nào.

Các vấn đề của hệ thống nhận dạng

Biểu diễn hoặc miêu tả đối tượng nhận dạng

Trích chọn đặc tính: Quá trình trích chọn đặc tính, các đặc trưng cơ bản phải đảm bảo các tiêu chí sau:

. Giảm được thứ nguyên không gian biểu diễn

. Đảm bảo được đủ lượng thông tin để phân biệt đối tượng này với đối tượng khác

.Cô đọng các đặc tính chính

Quá trình học: quá trình học thực chất là quá trình nhóm các lớp có cùng một số đặc tính chính, có một số phương pháp học sau:

. Học có mẫu: là sự học được bắt đầu bởi đã tồn tại sự phân lớp đối với một số đối tượng mẫu hoặc đã biết đặc tính của các lớp đối tượng, nói cách khác là xác định được biên giới giữa các lớp để sao cho để có thể nhận biết được đối tượng thuộc lớp nào.

. Học không có mẫu: quá trình học không có mẫu bắt đầu khi sự phân lớp chưa hình thành, và không có mẫu. Quá trình học nhằm tiến hành

nhóm dần dần trên cơ sở các đối tượng đã quan sát có độ tương tự gần nhau để tiến hành sự phân lớp.

Quá trình ra quyết định : Quá trình ra quyết định là tìm ra 1 luật dựa trên cơ sở đã biết sự phân lớp các đối tượng cũng như đặc trưng của các lớp để quyết định một đối tượng quan sát thuộc 1 lớp nào hoặc đồng nhất với một phần tử nào đó.

Khái quát về nhận dạng tiếng nói

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với mục đích là phân lớp (classify) thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ trong bộ nhớ. Các mẫu là các đơn vị nhận dạng, chúng có thể là các từ, hoặc các âm vị.

Nhận dạng tiếng nói là một kỹ thuật có thể ứng dụng trong rất nhiều lĩnh vực của cuộc sống : trong việc điều khiển (điều khiển robot, động cơ, điều khiển xe lăn cho người tàn tật…), an ninh quốc phòng…

Các nghiên cứu về nhận dạng tiếng nói dựa trên ba nguyên tắc cơ bản:

+) Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn (short-term amplitude spectrum). Nhờ vậy ta có thể trích ra các đặc điểm tiếng nói từ những khoảng thời gian ngắn và dùng các đặc điểm này làm dữ liệu để nhận dạng tiếng nói.

+) Nội dung của tiếng nói được biểu diễn dưới dạng chữ viết, là một dãy các ký hiệu ngữ âm. Do đó ý nghĩa của một phát âm được bảo toàn khi chúng ta

phiên âm phát âm thành dãy các ký hiệu ngữ âm.

+) Nhận dạng tiếng nói là một quá trình nhận thức. Thông tin về ngữ nghĩa (semantics) và suy đoán (pragmatics) có giá trị trong quá trình nhận dạng tiếng nói, nhất là khi thông tin về âm học là không rõ ràng.

Người ta chia các dạng bài toán nhận dạng tiếng nói theo một số tiêu chí sau:

- Nhận dạng tiếng nói phụ thuộc người nói/ độc lập người nói

- Kiểu lời nói: liên tục hay rời rạc
- Kích thước từ điển: nhỏ, trung bình hoặc lớn
- Nhận dạng trong môi trường có nhiễu hay không có nhiễu

Dựa vào kích thước từ điển, các hệ thống nhận dạng tiếng nói còn được chia thành 3 loại chính sau :

- Các hệ thống từ điển nhỏ: thường từ 20- 200 từ.
- Các hệ thống từ điển trung bình: thường từ 201- 1000 từ.
- Các hệ thống từ điển cỡ lớn: có từ trên 1000 từ.

3.1. Một số khái niệm cơ bản về âm thanh và tiếng nói.

3.1.1 Âm thanh

+ sóng âm và cảm giác âm

- Khi một vật giao động về một phía nào đó, lớp không khí liền trước nó bị nén lại và lớp không khí liền sau nó bị dãn ra. Sự dãn và nén của các lớp không khí lặp đi lặp lại tạo ra trong không khí một sóng dọc đàn hồi với tần số nào đó. Sóng không khí truyền tới tai người làm cho màng nhĩ dao động theo tần số đó, khi tần số sóng đạt đến một mức độ nhất định thì tạo ra cảm giác âm thanh trong tai người

- Màng nhĩ tai người nói chung thu đươc sóng có tần số từ 16hz đến 20.000hz. Trong khoảng tần số đó dao động được gọi là dao động âm thanh hay âm thanh.

+ Độ cao của âm

- Những âm thanh có tần số khác nhau gây cho ta những cảm giác âm khác nhau, âm có tần số lớn gọi là âm cao còn âm có tần số nhỏ gọi là âm thấp hay âm trầm.

+ Năng lượng của âm

- Cũng như các sóng cơ học khác, sóng âm mang năng lượng tỷ lệ với bình phương biên độ sóng. Năng lượng đó sẽ truyền từ nguồn âm tới tai người.

+ Cường độ âm

- Là năng lượng được sóng âm truyền trong một đơn vị thời gian qua một đơn vị diện tích đặt vuông góc với phương truyền (w/m^2). Đối với tai người, cường độ âm (I) là tham số không quan trọng bằng giá trị tỉ số I/I_0 với (I_0 là cường độ chuẩn). Người ta định nghĩa độ ồn của âm thanh L qua biểu thức sau:

$$L=lg(I/I_0)$$

Thứ nguyên của L là Ben (kí hiệu: B). Như vậy khi L=1,2,3… có nghĩa là cường độ âm I lớn hơn 10, 102, 103…lần cường độ âm chuẩn I_0

Sau đây là một số mức âm lượng

- Tiếng ồn trong phòng: khoảng 30 dB
- Tiếng ồn ào ngoài đường phố: khoảng 90 dB
- Ngưỡng đau tai: khoảng 130 dB

+ Độ to của âm

Độ to của âm (âm lượng) đối với tai người không trùng với cường độ âm. Tai người nghe thính nhất đối với các âm trong miền tần số 1000-5000Hz và nghe âm cao thính hơn âm trầm.

+ Âm sắc

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 dao động âm thanh phát ra một âm có tần số f_0 thì đồng thời cũng phát ra các âm có tần số $f_1=2f_0$, $f_3=3f_0$…

Âm có tần số f_0 gọi là âm cơ bản hay hoạ âm thứ nhất, các âm có tần số cao hơn gọi là hoạ âm thứ 2, thứ 3,…Âm cơ bản bao giờ cũng mạnh nhất, các hoạ âm có tác dụng quyết định âm sắc của âm cơ bản. Tuỳ theo cấu trúc khoang miệng, cổ họng và khoang mũi của từng người mà có các hoạ âm khác nhau.

3.1.2 Các đặc trưng của Tiếng nói

Năng lượng và độ lớn trung bình thời gian ngắn

Năng lượng thời gian ngắn được định nghĩa theo công thức sau:

$$E = \sum_{m=-\infty}^{+\infty} [x(m)\, w(n-m)]^2 \qquad (3.1.1)$$

Do tính năng lượng có phép tính bình phương nên kết quả thường có giá trị rất lớn. Người ta thay thế bằng một đại lượng khác là độ lớn trung bình.

$$M_n = \sum_{m=-\infty}^{+\infty} |x(m)|\, w(n-m) \qquad (3.1.2)$$

Trong đó *w(n-m)* là khung cửa sổ lấy tín hiệu âm thanh.

Căn cứ vào các giá trị năng lượng hoặc độ lớn thời gian ngắn có thể phân biệt được các đoạn hữu thanh – vô thanh hoặc các đoạn tín hiệu nhiễu nền.

Tần số cắt không trung bình thời gian ngắn

Các tín hiệu rời rạc theo thời gian, khái niệm tần số cắt không có nghĩa là số lần tín hiệu đổi dấu. Đây là một đại lượng tần số đơn giản của tín hiệu. Ví dụ tín hiệu hình sin có tần số F_0 , tần số lấy mẫu Fs có F_s/F_0 mẫu trong một chu kỳ sóng sin, trong khi đó mỗi chu kỳ có hai lần cắt không, do đó tần số cắt không trung bình thời gian dài là $Z = 2F_0/F_s$ số lần cắt trên mẫu. Như vậy tần số cắt không trung bình cũng là một cách để xác định tần số của sóng hình sin. Tín hiệu tiếng nói là tín hiệu băng rộng nên thường xác định tần số cắt không trong đoạn thời gian ngắn, công thức chung như sau:

$$Z_n = \sum_{m=-\infty}^{\infty} |\,\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\,|\, w(m-n) \qquad (3.1.3)$$

Trong đó :

sgn[x(n)] = 1 khi x(n) ≥ 0

= -1 khi x(n) < 0

w(n) : cửa sổ lấy tín hiệu

w(n) = 1 nếu 0 ≤ n ≤ N-1= 0 trường hợp còn lại.

Năng lượng, độ lớn và tần số cắt không thời gian ngắn là cách đơn giản và hiệu quả để xác định phần nhiễu nền và tín hiệu, phần tín hiệu vô thanh và hữu thanh. Bằng thực nghiệm quan sát trực quan ta thấy : Phần có tín hiệu âm thanh thì biên độ sóng âm lớn hơn phần nhiễu nền. Mặt khác giá trị trung bình biên độ sóng âm của hai đoạn âm thanh có tín hiệu và nhiễu nền đều xấp xỉ không.

Khi cần phân biệt phần nhiễu nền và tín hiệu, phần tín hiệu vô thanh và hữu thanh, thường ta chỉ cần một chỉ tiêu trên cũng đủ để phân biệt. Nhưng đôi khi trường hợp phức tạp hơn trong phân biệt âm xát và nhiễu nền ta cần phải sử dụng đến cả hai chỉ tiêu năng lượng và tần số cắt không. Ngoài ra các chỉ tiêu trên còn được sử dụng để thiết lập chu kỳ Pitch(tần số cơ bản của tiếng nói).

Hàm sai khác độ lớn trung bình thời gian ngắn

Dưới đây sẽ trình bày một phương pháp rất hữu dụng để trích ra được tần số Pitch(tần số cơ bản của tiếng nói). Hàm sai khác độ lớn trung bình thời gian ngắn được định nghĩa như sau :

$$AMDF(P) = \frac{1}{N}\sum_{i=ko+1}^{ko+N} | y_i - y_{i-P} | \qquad (3.1.4)$$

Giả sử chuỗi $\{y_n\}$ tuần hoàn với chu kỳ P_0 thì hàm AMDF sẽ đạt giá trị cực tiểu tại P_0 . Như vậy việc xác định chu kỳ Pitch của tiếng nói sẽ thông qua xác định chỉ số P_0 mà tại đó hàm AMDF đại giá trị cực tiểu. Trong thực tế chu kỳ Pitch tiếng nói của một người nằm trong một miền giới hạn, vì vậy không cần thiết phải tính toán cho mọi giá trị P của hàm AMDF. Qua thực nghiệm âm thanh tiếng nói con người, chu kỳ Pitch nằm trong khoảng 2.5 mili giây đến 19.5 mili giây. Với tốc độ lấy mẫu thực hiện trong đồ án là 11025 mẫu trên giây thì chu kỳ Pitch nằm trong khoảng 30 đến 220.

3.2. Một số phương pháp nhận dạng tiếng nói

3.2.1 Một số khuynh hướng nghiên cứu nhận dạng tiếng nói

Hiện nay trên thế giới có 4 khuynh hướng nghiên cứu nhận dạng tiếng nói, gồm :

- Hướng tiếp cận âm học – ngữ âm học.
- Hướng tiếp cận nhận dạng theo mẫu thống kê.
- Hướng tiếp cận trí tuệ nhân tạo.
- Hướng tiếp cận sử dụng mạng nơron.

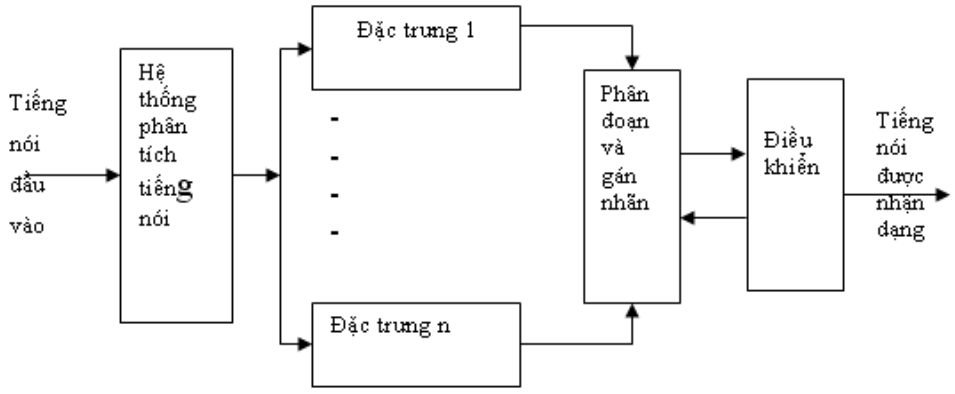
3.2.1.1 *Hướng tiếp cận âm học – ngữ âm học để nhận dạng tiếng nói*

Khuynh hướng âm học – ngữ âm học dựa trên lý thuyết về ngữ âm học. Lý thuyết này cho rằng tồn tại các đơn vị ngữ âm xác định, có tính phân biệt trong lời nói và các đơn vị ngữ âm đó được đặc trưng bởi một tập các đặc tính tín hiệu tiếng nói . Mặc dù các đặc tính âm học của các đơn vị ngữ âm biến thiên rất lớn theo cả giọng người nói lẫn tác động của các đơn vị ngữ âm xung quanh (còn gọi là hiện tượng đồng phát âm), nhưng vẫn tồn tại các qui tắc cho phép giải quyết những vấn đề như vậy

Bước đầu tiên trong hướng tiếp cận âm học – ngữ âm học để nhận dạng tiếng nói là phân đoạn và gán nhãn. Bước này chia tín hiệu tiếng nói thành các đoạn có các đặc tính âm học đặc trưng cho một (hoặc vài) đơn vị ngữ âm (hoặc lớp), đồng thời gán cho mỗi đoạn âm thanh đó một hay nhiều nhãn ngữ âm phù hợp.

Bước thứ hai xác định một từ (hoặc một chuỗi từ) đúng trong số chuỗi các nhãn ngữ âm được tạo ra sau bước một và phải tuân thủ một số điều kiện ràng buộc (tức là các từ được chọn ra trong từ điển cho trước phải phù hợp nguyên tắc ngữ pháp và có nghĩa)

Sơ đồ khối của hệ thống nhận dạng tiếng nói theo hướng âm học – ngữ âm học thể hiện trên Hình 1.1



Hình 1.1: Sơ đồ khối của hệ thống nhận dạng tiếng nói theo hướng âm học – ngữ âm học

Hệ thống nhận dạng tiếng nói theo khuynh hướng này gặp phải khá nhiều vấn đề khó khăn do đó nó chưa được áp dụng nhiều trong thực tế. Khuynh hướng này đòi hỏi sự hiểu biết sâu sắc về các tính chất âm học của các đơn vị ngữ âm. Nguồn kiến thức này khó có thể đầy đủ được nên nhận dạng tiếng nói theo khuynh hướng này vẫn còn là chủ đề nghiên cứu thú vị nhưng cần được nghiên cứu và tìm hiểu sâu sắc hơn để có thể áp dụng thành công vào các hệ thống nhận dạng tiếng nói thực tế.

3.2.1.2 ***Hướng tiếp cận nhận dạng theo mẫu thống kê***

Nhận dạng tiếng nói theo khuynh hướng này là sử dụng trực tiếp các mẫu tín hiệu tiếng nói mà không phải xác định rõ ràng các đặc tính âm học (so với khuynh hướng âm học – ngữ âm học) và không phải phân đoạn tiếng nói. Các hệ thống nhận dạng tiếng nói theo khuynh hướng này được thực hiện theo hai bước:

Bước thứ nhất: Sử dụng tập mẫu tiếng nói (cơ sở dữ liệu tiếng nói) để huấn luyện hệ thống, "tri thức" về tiếng nói của hệ thống nhận dạng tiếng nói được tích luỹ thông qua quá trình huấn luyện

Bước thứ hai: Nhận dạng, thực hiện so sánh tiếng nói chưa biết với các mẫu đã được huấn luyện.

Nguyên tắc cơ bản của hướng này là nếu cơ sở dữ liệu dùng cho huấn luyện có đủ các phiên bản của mẫu cần nhận dạng thì quá trình nhận dạng có

thể xác định được các đặc tính âm học của mẫu (mẫu có thể là âm vị, từ hoặc cụm từ…).

Hướng tiếp cận theo mẫu thống kê có cácchức năng chủ yếu sau:

- Phân tích và xác định các tham số: Tín hiệu tiếng nói được phân tích thành một chuỗi các đặc trưng để xác định các mẫu nhận dạng. Đối với tín hiệu tiếng nói, các đặc trưng này thường là kết quả của một số kỹ thuật phân tích phổ như ngân hàng bộ lọc, phân tích mã hoá dự báo tuyến tính (LPC), biến đổi Fourier rời rạc (DFT)…
- Huấn luyện mẫu: Một số mẫu tương ứng với các đơn vị âm thanh cùng loại được sử dụng để học, trích chọn ra các đặc trưng của mẫu đó.
- Khối phân lớp mẫu: Mẫu đầu vào chưa biết được so sánh với mẫu đại diện của một loại âm thanh nào đó và đo khoảng cách (còn gọi là "độ" giống nhau) giữa mẫu đầu vào và mẫu chuẩn.
- Khối nguyên tắc chọn: Các chỉ số về điểm giống nhau giữa các mẫu tín hiệu tiếng nói đầu vào và mẫu chuẩn được sử dụng để quyết định mẫu chuẩn nào phù hợp nhất với mẫu đầu vào chưa biết.

  Việc chọn hướng tiếp cận này có những ưu và nhược điểm sau:

  - Tính đơn giản và dễ hiểu trong việc áp dụng thuật toán
  - Tính bất biến trong thuật toán so sánh mẫu và quyết định đối với mọi từ vựng, mọi người sử dụng
  - Sự thực hiện của hệ thống rất nhạy cảm với số lượng dữ liệu huấn luyện có thể cung cấp cho lớp các mẫu chuẩn. Nói chung, huấn luyện càng nhiều thì hiệu suất thực hiện của hệ thống càng cao.
  - Không có kiến thức tiếng nói đặc biệt dùng để xác định hệ thống vì vậy phương pháp này không nhạy cảm với việc chọn từ vựng, cú pháp và ngữ nghĩa.

- Sự tính toán cho huấn luyện mẫu và phân lớp mẫu nói chung là tuyến tính đối với số mẫu huấn luyện hoặc nhận dạng, vì vậy khi số lớp lớn thì số phép tính tăng lên càng nhanh.
- Tương đối dễ ràng buộc trực tiếp các thành phần ngữ pháp (và cả ngữ nghĩa) vào cấu trúc nhận dạng mẫu, do đó cải thiện được tính chính xác nhận dạng và giảm được sự tính toán

3.2.1.3 ***Hướng tiếp cận trí tuệ nhân tạo cho nhận dạng tiếng nói***

Nhận dạng tiếng nói theo hướng trí tuệ nhân tạo là sự kết hợp giữa khuynh hướng âm học với khuynh hướng nhận dạng mẫu vì nó khai thác các ý tưởng của hai khuynh hướng đó. Nhận dạng tiếng nói theo khuynh hướng này là cố gắng tự động hoá thủ tục nhận dạng theo cách mà con người áp dụng trí tuệ của mình để hình dung, phân tích và cuối cùng đưa ra quyết định trên các đặc trưng âm học đo được. Trong thực tế, các kỹ thuật nhận dạng tiếng nói theo khuynh hướng này là sự sử dụng hệ chuyên gia cho sự phân đoạn và gán nhãn, như thế bước cốt yếu và khó khăn nhất này có thể được thực hiện không chỉ nhờ các thông tin âm học (ý tưởng nhận dạng theo khuynh hướng âm học) mà còn phân biệt các mẫu âm thanh (ý tưởng của nhận dạng mẫu).

Ý tưởng cơ bản của hướng tiếp cận trí tuệ nhân tạo vào nhận dạng tiếng nói là thu thập kiến thức từ các nguồn tri thức khác nhau để giải quyết các vấn đề đang đặt ra, ví dụ tiếp cận trí tuệ nhân tạo cho việc phân đoạn và gán nhãn tiếng nói cần có sự tổng hợp các kiến thức về âm học, kiến thức từ vựng, kiến thức ngữ pháp, kiến thức ngữ nghĩa và thậm chí cả kiến thức thực tế.

3.2.1.4 ***Hướng tiếp cận sử dụng mạng nơron***

Xét về khía cạnh mô phỏng trí tuệ con người thì mạng nơron nhân tạo có thể coi là phương pháp tiếp cận trí tuệ nhân tạo, tuy nhiên có thể coi đây là một phương pháp riêng.

Phương pháp này thực chất có cơ sở là phương pháp nhận dạng mẫu thống kê. Khác cơ bản là cách thức lưu trữ mẫu. Phương pháp này chỉ lưu trữ

vectơ số liệu thể hiện tham số đặc trưng thông qua trọng số liên kết và hệ số hiệu chỉnh.

3.2.2 Các đơn vị xử lý tiếng nói

3.2.2.1 *Tần số lấy mẫu*

Quá trình lấy mẫu tạo ra tín hiệu rời rạc hoặc tín hiệu số từ tín hiệu tương tự. Tần số lấy mẫu là số lần lấy mẫu được tính trong một đơn vị thời gian, thông thường là giây. Tần số lấy mẫu ký hiệu là Fs.

Khoảng thời gian mà quá trình lấy mẫu được lặp lại gọi là chu kỳ lấy mẫu.

Ví dụ: Fs = 11025 Hz

1s thu được 11025 mẫu

1ms thu được 11025/1000 ≈ 11 mẫu.

Số bit lưu một mẫu có thể là 8 hoặc 16 bit.

+ 8 bit/1 mẫu: $x(n) \in (0, 2^8 - 1)$

Ngưỡng lặng tuyệt đối là 128

+ 16 bit/1 mẫu: $x(n) \in (2^{-15}, 2^{15-1})$

Ngưỡng lặng tuyệt đối là 0

3.2.2.2 ***Tần số cơ bản***

Một âm thanh có thể là tổ hợp của nhiều tần số, tần số chính bao trùm trong âm được gọi là tần số cơ bản. Trong tiếng nói, tần số cơ bản là đáp ứng của sự rung động các dây thanh âm, tần số cơ bản thường được ký hiệu là F_0.

Tần số cơ bản có giá trị phụ thuộc vào tần số lấy mẫu và khoảng cách a, là khoảng cách giữa hai đỉnh của các sóng âm tuần hoàn.

Đơn vị của tần số là Hertz, ký hiệu là Hz. Mỗi Hz bằng 1 dao động/1s. và 1KHz sẽ bằng 1000 Hz.

3. 2. 2. 3 *Nhiễu*

Nhiễu đối với hệ thống là loại âm thanh ngoài mong muốn hoặc không phải tiếng nói sinh ra trong môi trường xung quanh. Ngay cả bộ phát âm của con người đôi khi cũng sinh ra nhiễu, chẳng hạn như tiếng thở, tiếng bật lưỡi,

tiếng chép miệng cả khi môi chạm vào micro... Không dễ gì có thể lọc được mọi thứ nhiễu, ta chỉ tìm cách tối thiểu hoá chúng để có thể nâng cao chất lượng của hệ thống nhận dạng.

Với tín hiệu tiếng nói là s*n*, tín hiệu nhận được sau quá trình thu sẽ được ký hiệu là $\tilde{s}_n$. Như vậy:

$\tilde{s}_n$ - s*n* chính là tín hiệu nền. Độ nhiễu của tín hiệu được xác định thông qua năng lượng đo được của tín hiệu: $E = 10\log_{10}\left(\frac{\left(\sum_{n=0}^{N} s_n^2\right)}{\left(\sum_{n=0}^{N}\left(\tilde{s}_n - s_n\right)^2\right)}\right)$

(Đơn vị năng lượng tính bằng dB)

Như vậy, nếu năng lượng E càng lớn thì $\tilde{s}_n$ càng gần với s*n*, tín hiệu nền có giá trị gần về 0. Nếu $E \to \infty$ thì tín hiệu thu được là tín hiệu sạch, không có nhiễu.

3.2.2.4 ***Thông số độ ồn nhiễu.***

Cách xác định: Thông báo yêu cầu người sử dụng dừng nói trong 3 giây và thu tín hiệu trong thời gian đó để lấy tiếng ồn nhiễu của môi trường xung quanh. Ngưỡng im lặng được xác định là năng lượng cao nhất của các frame. Ngoài ra có thể dùng biến đổi Fourier để tính ra các tần số nhiễu phục vụ cho việc lọc nhiễu.

3.2.2.5 ***Lọc nhiễu***

Hiện tại, việc lọc nhiễu của hệ thống được thực hiện theo phương pháp kinh điển là dùng phép biến đổi Fourier với thuật toán FFT. Dùng biến đổi Fourier thuận xác định được các tần số tham gia và loại đi tất cả tần số không thuộc phạm vi tiếng nói (*nếu biết được phạm vi tần số đúng của người sử dụng thì kết quả lọc sẽ càng cao*) bằng cách cho các hệ số tương ứng giá trị zero sau đó biến đổi ngược lại.

3.2.3 Một số kỹ thuật khử nhiễu

1. Kỹ thuật CMS

Đây là một kỹ thuật thông dụng để khử nhiễu trong các hệ thống nhận dạng, được dùng kết hợp trong quá trình tính toán các đặc tính phổ của tiếng nói. Phương pháp này dựa trên giả thiết là các đặc tính tần số của môi trường là thường xuyên cố định hoặc biến đổi chậm. Các tham số cepstral của một phát âm được trừ đi giá trị trung bình của các tham số trong một khoảng thời gian nào đó và làm cho các giá trị này ít bị ảnh hưởng bởi môi trường

$$\hat{O}(\tau) = O(\tau) - \frac{1}{T}\sum_{t=1}^{\tau} O(\tau) \qquad (3.2.1)$$

Trong đó, T là độ dài của vùng lấy giá trị trung bình, thường là độ dài của cả phát âm.

2. Kỹ thuật RASTA

RASTA là kỹ thuật lọc dựa trên giả thiết rằng các tính chất thời gian của các nhiễu là khác so với các tính chất thời gian của giọng nói. Tốc độ thay đổi của các thànhphần không phải tiếng nói thường xuyên nằm ngoài tốc độ hoạt động của bộ máy phát âm con người. Bằng cách dùng bộ lọc số, kỹ thuật RASTA có thể loại bỏ được một phần các nhiễu của môi trường và các nhiễu bổ sung bất thường khác. Bộ lọc dùng trong RASTA là:

$$H(z) = \frac{0{,}2 + 0{,}1z^{-1} - 0{,}2z^{-2} - 0{,}1z^{-3}}{1 - 0{,}94z^{-1}} \qquad (3.2.2)$$

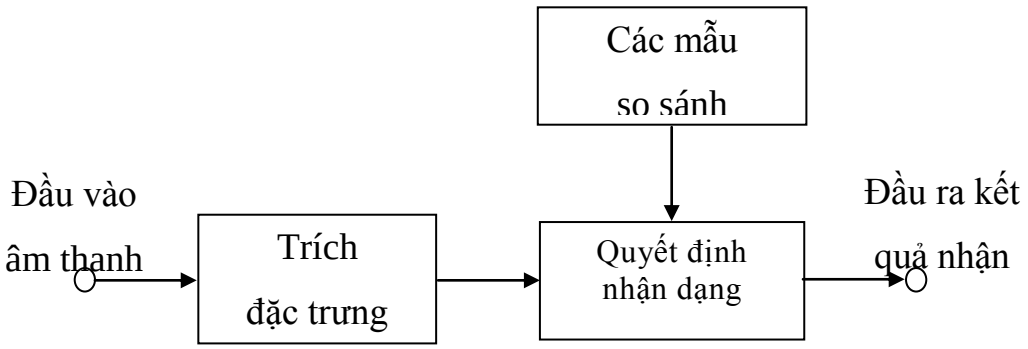
Các kỹ thuật khử nhiễu thường yêu cầu một đoạn tiếng nói đủ lớn để phân tích, thống kê. Vì vậy, khi áp dụng các kỹ thuật khử nhiễu vào nhận dạng tiếng nói, cần lưu ý đến tốc độ xử lý và bảo tồn các đặc trưng âm học của phụ âm, đặc biệt là các phụ âm vô thanh. Để đảm bảo thực hiện được trong thời gian thực, hiện nay, người ta áp dụng mô hình tham số thích nghi với nhiễu. Cụ thể như sau: Khi huấn luyện tham số, người ta lấy một mẫu sạch, không bị nhiễu, để huấn luyện, sau đó, người ta lấy các mẫu sạch này trộn với các loại nhiễu sinh bởi các mô hình toán học khác nhau và tham số mô hình sẽ được biến đổi bởi mẫu nhiễu nhờ các công cụ mô hình như mạng Nơron. Do đó,

trong giai đoạn nhận dạng, khi tín hiệu thực được đưa vào hệ thống, người ta sẽ tính thẳng các đặc trưng và quyết định từ chính tín hiệu chứ không cần lọc

3.2.4 Một số phương pháp nhận dạng tiếng nói

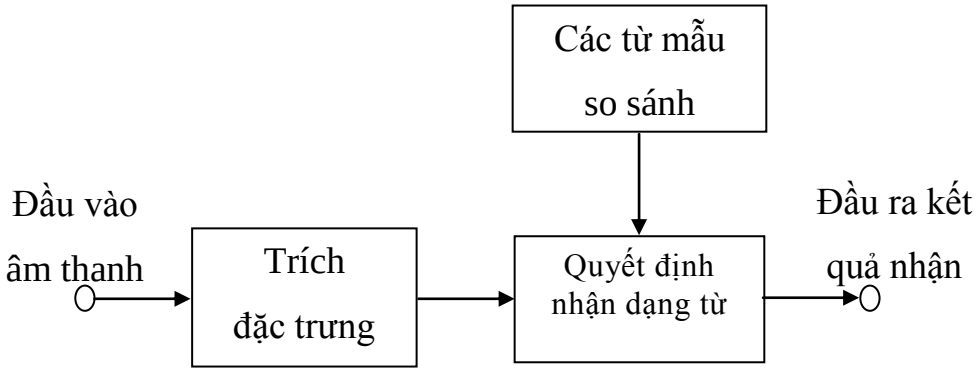
3.2.4.1 Sơ đồ khối hệ thống nhận dạng tiếng nói

Quá trình nhận dạng tiếng nói có thể mô tả sơ lược như sau: Âm thanh mới đưa vào máy sẽ được phân tích để trích ra các đặc trưng của tiếng nói, ngôn ngữ. Sau đó đem so sánh với các mẫu ta đã thực hiện phân tích từ trước. Cuối cùng là đưa ra quyết định nhận dạng đối với âm thanh mới. So sánh mẫu thực hiện tính toán và quyết định cho nhận dạng tiếng nói.

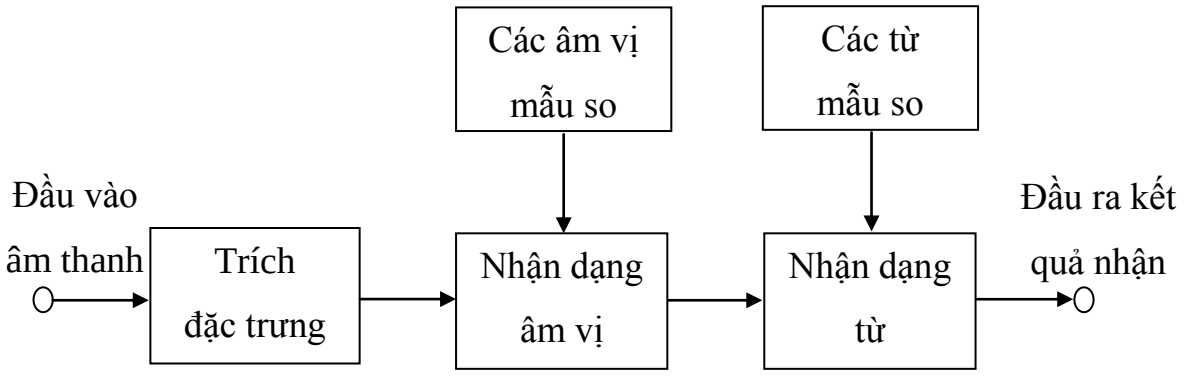


Hệ thống so sánh mẫu

Với hệ thống so sánh mẫu, đây chỉ là mô tả một cách khái quát nhất cho một ứng dụng nhận dạng tiếng nói. Dưới đây là sơ đồ khối cho hệ thống nhận dạng tiếng nói theo từng từ riêng biệt.



Hệ thống này có nhiều khả năng thực hiện. Nhưng đối với mỗi một ngôn ngữ số lượng từ thường là rất lớn. Việc thiết lập và lưu trữ hệ thống dữ liệu cho tất cả các mẫu từ trong một ngôn ngữ gặp nhiều khó khăn về dung lượng bộ nhớ cũng như tốc độ xử lý thực. Để nâng cao tính khả thi của hệ thống này, người ta bổ xung thêm một khâu nhận dạng đơn giản : nhận dạng âm vị. Một từ điển âm vị được xây dựng thực hiện nhận dạng âm vị của từ. Sau đó dựa trên khả năng âm vị sẽ sinh ra từ nào và tiếp tục nhận dạng từ đó. Điểm khác biệt là ở chỗ trong từ điển mỗi từ là một chuỗi các âm vị cấu thành lên từ đó. Nhờ đó kích thước của từ điển từ nhỏ hơn, giảm được dung lượng bộ nhớ dành cho từ điển từ.



Nhận dạng từ dựa trên nhận dạng âm vị

3.2.4.2 So sánh độ tương đồng giữa các mẫu

Ngày nay nhận dạng tiếng nói đã trở nên phát triển và được ứng dụng rất nhiều trong cuộc sống, đã có nhiều phương pháp nghiên cứu về nhận dạng tiếng nói như phương pháp mô hình markow ẩn (HMM), phương pháp dùng mạng noron , hay phương pháp LPC-10…

Trong vấn đề nhận dạng dù là nhận dạng ảnh, âm thanh hoặc một kiểu dữ liệu nào khác, hay mô hình nhận dạng nào thì việc ra quyết định đều dựa trên phương pháp chung nhất là so sánh. Đối chiếu giữa mẫu mới và các lớp mẫu sau đó dựa trên các quy tắc để quyết định về mẫu đó. Mẫu nào có độ tương đồng đối với 1 lớp mẫu là lớn nhất thì quyết định mẫu thuộc về lớp mẫu đó.

3.2.4.2.1 Định nghĩa

Độ tương đồng(hay giống nhau) giữa các mẫu được xác định bằng công thức toán học. Nó cho phép ta khẳng định sự giống nhau của 2 mẫu.

Giả sử ta có 2 vector mẫu đầu và A_i , A_j n chiều. Độ giống nhau $L(A_i , A_j)$ phải thoả mãn các điều kiện sau:

- Số đo độ giống nhau phải dương : $L(A_i , A_j) >= 0$;
- Phải có tính đối xứng $L(A_i , A_j) = L(A_j , A_i)$
- Số đo độ giống nhau phải có giá trị cực đại khi ước lượng sự giống nhau giữa một ảnh nào đó với chính nó

$L(A_i , A_j) = \max L(A_i , A_j)$ (với mọi i <> j)

3.2.4.2.2 Hàm khoảng cách

Hàm khoảng cách của hai mẫu có thể cho ta đánh giá được dễ dàng độ giống nhau giữa hai mẫu. Theo trực quan chúng ta khẳng định độ giống nhau giữa hai mẫu càng lớn khi và chỉ khi khoảng cách giữa chúng càng nhỏ.

Giả sử hai mẫu A_i , A_j n chiều :

$A_i = \{a_{i1} , a_{i2} , ... , a_{in}\}$

$A_j = \{a_{j1} , a_{j2} , ... , a_{jn}\}$

Dưới đây là một số hàm khoảng cách phổ biến.

+ Khoảng cách Ơclid

$$d_1(A_i , A_j) = \{\sum_{k=1}^{N} (a_{ik} - a_{jk})^2\}^{1/2} \qquad (3.2.3)$$

+ Khoảng cách Manhattan

$$d_2(A_i , A_j) = \sum_{k=1}^{N} \left| a_{ik} - a_{jk} \right| \qquad (3.2.4)$$

+ Khoảng cách Trebusep

$$d_3(A_i , A_j) = \max_k (a_{ik} - a_{jk}) \qquad (3.2.5)$$

+ Khoảng cách Micowskiedo

$$d_4(A_i , A_j) = \left(\sum_{k=1}^{N} \left(a_{ik} - a_{jk} \right)^{1/\lambda} \right)^{\lambda} \qquad (3.2.6)$$

+ Khoảng cách Korelasi

$$d_5(A_i , A_j) = \frac{\sum_{k=1}^{N} \left(a_{ik} - \overline{a_i} \right)\left(a_{jk} - \overline{a_j} \right)}{\left(\sum_{k=1}^{N} \left(a_{ik} - \overline{a_i} \right)^2 . \sum_{k=1}^{N} \left(a_{jk} - \overline{a_j} \right)^2 \right)^{1/2}} \qquad (3.2.7)$$

ở đây:

$$\overline{a_i} = \frac{1}{N} \sum_{k=1}^{N} x_{ik} \qquad \overline{a_j} = \frac{1}{N} \sum_{k=1}^{N} x_{jk}$$

+ Khoảng cách Cambera:

$$d_6(A_i , A_j) = \sum_{k=1}^{N} \frac{\left| a_{ik} - a_{jk} \right|}{\left| a_{ik} + a_{jk} \right|} \qquad (3.2.8)$$

Các hàm tính khoảng cách Trebusep, Manhattan và Ơclid có khối lượng tính toán tăng dần. Các hàm còn lại đều đòi hỏi thời gian tính nhiều hơn. Trong ba khoảng cách có khối lượng tính toán ít nhất thì khoảng cách Ơclid đảm bảo khắc phục được đặc tính biến động của mẫu (mà dữ liệu tín hiệu âm thanh có sự biến động rất lớn vả về biên độ và thời gian). Lựa chọn hàm tính khoảng cách Ơclid giảm tối thiểu tính biến động của dữ liệu âm thanh mà vẫn đảm bảo tốc độ tính toán cho hệ thống.

3.2.4.2.3 Nhận dạng trên cơ sở tương đồng của các đối tượng

Giả sử ta có S = $\{S_j\}$ là tập các mẫu nhận dạng N chiều. Việc nhận dạng sẽ được thực hiện thông qua các số đo về độ tương đồng giữa mẫu mới với các mẫu đã có.

A_i thuộc S_j nếu $L(A_i, S_j) = \max (L(A_i, S_j))$ theo j

Theo quan điểm về khoảng cách :

A_i thuộc S_j nếu $d(A_i, S_j) = \min (L(A_i, S_j))$ theo j

Như vậy ta có thể vận dụng nhận dạng mẫu theo nguyên tắc :

+ Tính khoảng cách giữa mẫu mới với từng lớp mẫu khác nhau.

+ Khẳng định mẫu đó thuộc lớp nào có trị khoảng cách là nhỏ nhất.

3.2.4.3 Đối sánh mẫu dựa trên phương pháp LPC-10

Phương pháp này được áp dụng để nhận dạng các từ đơn lẻ. Nó có các ưu và nhược điểm sau:

Ưu điểm:

- Tiêu hao tài nguyên hệ thống ít

- Có thể sử dụng cho một từ đơn hoặc một nhóm từ mà không cần thay đổi thuật toán nhiều.

- Dễ cài đặt và triển khai. Với số lượng từ không lớn, độ chính xác đạt cao.

Nhược điểm:

- Độ chính xác của hệ phụ thuộc vào số lượng mẫu học. Các mẫu tham khảo sẽ bị ảnh hưởng bởi môi trường lúc tạo mẫu như tiếng ồn, thiết bị, tâm lý...

- Thời gian tính toán phụ thuộc vào số lượng mẫu học và số lượng từ cần nhận dạng

3.2.4.3.1 Phân tích dự báo tuyến tính

Phân tích dự báo tuyến tính là một trong những kỹ thuật phân tích tiếng nói được sử dụng rộng rãi. Nó có thể tính toán hiệu quả các tham số như hàm diện tích của tuyến âm và lưu trữ hoặc truyền thông tiếng nói với tỷ lệ lưu trữ nhỏ.

Phân tích dự báo tuyến tính dựa trên cơ sở mẫu tín hiệu y_n sẽ được dự báo bằng p mẫu tín hiệu trước nó.

$$y_n \approx \sum_{i=1}^{p} \alpha_i y_{n-i} + G \cdot \varepsilon_n$$

Trong đó : α_i (i=1..p) là các hệ số dự báo

$\{y_{n-i}\}$(i=1..p) là dãy p tín hiệu ngay trước của tín hiệu y_n

G : hệ số lọc lặp

ε_n : sai số dự báo (hay còn gọi là nhiễu)

Sai số dự báo ε_n chính là sai số phân tích tiếng nói. Yêu cầu đặt ra cho hệ thống là phải giảm tối thiểu sai số dự báo ε_n . Ở đây ta thực hiện đạo hàm riêng phần ε_n^2 cho từng biến α_i (i=1..p), tính giá trị hệ số dự báo α_i mà với giá trị đó ε_n^2 đạt cực tiểu.

$$\frac{\partial}{\partial \alpha_j} E\left[\left(y_n - \sum_{i=1}^{P} \alpha_i . y_{n-i} - G.\varepsilon_n\right)^2\right] = 0$$

$$=> \quad -2E\left[\left(y_n - \sum_{i=1}^{P} \alpha_i . y_{n-i} - G.\varepsilon_n\right) y_{n-j}\right] = 0$$

$$=> \quad \sum_{i=1}^{P} \alpha_i E\left[y_{n-i}\, y_{n-j}\right] = E\left[y_n y_{n-j}\right]$$

Bài toán đưa về giải hệ phương trình P ẩn để tìm α_i . Để giải hệ trên ta cần tính được các E [y_{n-i} . y_{n-j}]. Có hai phương pháp cho phép ta tính các E [y_{n-i} . y_{n-j}] là phương pháp tự tương quan (autocorrelation) và phương pháp hiệp biến (autocovariance). Trong phương pháp tự tương quan ta được :

$$E\left[\, y_{n-i} \cdot y_{n-j}\right] = R_{yy}(|\, i - j \,|)$$

Giả sử dãy tín hiệu $\{y_n\}$ bằng 0 ngoài đoạn tín hiệu ta cần tính hệ số dự báo. Khi đó :

$$R_{YY}(k) = \sum_{n=k+1}^{N} y_n \cdot y_{n-k}$$

Bây giờ ta có thể mô tả hệ phương trình dạng ma trận như sau :

$$\begin{bmatrix} R_{yy}(0) & R_{yy}(1) & ... & R_{yy}(P-1) \\ R_{yy}(1) & R_{yy}(0) & ... & R_{yy}(P-2) \\ ... & ... & ... & ... \\ R_{yy}(P-1) & R_{yy}(P-2) & ... & R_{yy}(0) \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_P \end{bmatrix} = \begin{bmatrix} R_{yy}(1) \\ R_{yy}(2) \\ ... \\ R_{yy}(P) \end{bmatrix}$$

Hệ phương trình trên có thể giải bằng phương pháp nghịch đảo ma trận bởi vì ma trận R_{yy} là ma trận Toeflitz (ma trận đối xứng qua đường chéo chính và các đường chéo song song với đường chéo chính có các phần tử giống nhau). Ma trận Toeflitz luôn có định thức khác 0, cũng có nghĩa là luôn tìm được ma trận nghịch đảo cho ma trận Toeflitz. Nhưng giải hệ phương trình bằng phương pháp ma trận nghịch đảo không hiệu quả, hơn nữa độ sai số rất lớn. Dưới đây trình bày thuật toán Levinson – Durbin cho phép ta tính hệ số dự báo α_i (i=1..P) mà không cần giải hệ phương trình trên.

** Thuật toán Levinson – Durbin :*

1. **Đặt $E_0 = R_{yy}(0)$, i = 0**
2. **i = 1**
3. **$k_i = \left(\sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_{yy}(i-j+1) - R_{yy}(i)\right) / E_{i-1}$)**
4. **Đặt $\alpha_i^{(i)} = k_i$**
5. **$\alpha_j^{(i)} = \alpha_j^{(i-1)} + k_i \cdot \alpha_j^{(i-1)}$ (với j = 1..i-1)**
6. **$E_i = (1 - k_i^2) E_{i-1}$**
7. **Nếu i < P thì quay lại 2.**

3.2.4.3.2 Nhận dạng tiếng nói bằng phương pháp LPC-10

LPC là viết tắt của Linear Predictive Coder (mã hoá dự báo tuyến tính). Chỉ số 10 có nghĩa là hệ thống dự báo tuyến tính có số lượng hệ số từ 10 trở lên là tốt nhất, hệ thống phải có tối thiểu 10 hệ số dự báo mới đảm bảo mức chính xác của kết quả dự báo. Số lượng hệ số càng cao thì hiệu quả dự báo càng cao. Nhưng ngược lại thao tác tính toán càng phức tạp và tốn nhiều thời gian. Chương trình nhận dạng từ tiếng Việt chọn số lượng hệ số LPC-10 là 10 hệ số. Qua thực nghiệm, tiêu chuẩn LPC-10 có được các hệ số tốt nhất nếu ta lấy kích thước frame từ 10 – 20 ms(dạng file WAVE 11.025kHz, mono, 8 bits, thì kích thước theo mẫu tín hiệu của 1 frame từ 110 đến 220 mẫu). Kỹ thuật nhận dạng tiếng nói bằng phương pháp LPC-10 là thực hiện tính toán các hệ số dự báo tuyến tính sau đó so sánh với bộ mẫu là các bộ hệ số dự báo đã được tính toán trước đó.

Xử lý tín hiệu âm thanh bằng phương pháp dự báo tuyến tính (LPC) rất phổ biến. Nó đáp ứng được các yêu cầu đặt ra về xử lý âm thanh: Tổng hợp tiếng nói, nhận dạng tiếng nói ... Nhận dạng tiếng nói dựa trên tiêu chuẩn LPC-10 chính là vận dụng kỹ thuật dự báo tuyến tính nhằm tăng hiệu quả nhận dạng chương trình.

CHƯƠNG 4: CHƯƠNG TRÌNH DEMO

4.1. Thiết kế các chức năng chính

Với nhiệm vụ đồ án là nghiên cứu và xây dựng chương trình nhận dạng từ trong tiếng Việt. Trước hết, chức năng chính của chương trình là mô phỏng được công việc nhận dạng các từ đơn của tiếng Việt. Nó là cơ sở cho việc nhận dạng các đơn vị tiếng Việt lớn hơn như: từ ghép, cụm từ, câu… Chương trình gồm 2 chức năng chính:

+ Huấn luyện hệ thống: Chức năng này nhằm mục đích tạo và cập nhật vào cơ sở dữ liệu các đặc trưng cơ bản nhất của từ, tức là cho máy học để trích rút các đặc trưng của từ đó đối với nhiều người nói, phục vụ nhu cầu nhận dạng từ cho nhiều người khác nhau. Mỗi người thực hiện cho máy học một số từ tiếng Việt và sau đó sẽ ghi âm từ đó ở những lần nói khác rồi cho hệ thống nhận dạng.

+ Nhận dạng từ đơn (từ chỉ có một tiếng) của tiếng Việt từ file nguồn: Một từ chỉ được nhận dạng sau khi đã cho máy học về từ đó, chức năng này nhằm nhận dạng các từ từ file âm thanh. Nếu chưa có ta phải ghi âm từ cần nhận dạng bằng trình SoundRecorder của Window đã tích hợp sẵn trong hệ thống rồi ghi vào các file Wave, sau đó với nhận dạng các file âm thanh này bằng chức năng nhận dạng của chương trình. Hoặc để kiểm tra khả năng nhận dạng chương trình ta sử dụng các từ đã ghi âm sẵn trong thư mục Data-for-NhanDang, do các từ này đã có một tập hợp các mẫu được học trước đó và đã lưu các đặc trưng của các từ đó trong cơ sở dữ liệu.

+ Nhận dạng trực tiếp qua Microphone: Trên cơ sở từ đó đã được học rồi. Hệ thống sẽ thường trực chờ người nói nói vào Micro và hiển thị thông tin nhận được ở dạng text lên màn hình. Đồng thời sóng âm được hiển thị trực quan trong hộp ảnh.

+ Ngoài ra còn có các chức năng khác như:

- Ghi âm: để ghi âm các từ mẫu để học và các từ để nhận dạng.

- Hiển thị thông tin về file Wave đang đọc.

- Hiển thị sóng âm thanh khi đọc từ tệp.
- Hiển thị sóng âm thanh sau khi đã được xử lý.
- Đưa ra loa dữ liệu âm thanh đang xử lý (để kiểm tra).

4.2. Lựa chọn ngôn ngữ lập trình

Trong thiết kế chương trình nhận dạng từ tiếng Việt, chương trình phải đọc dữ liệu âm thanh vào mảng. Sau đó phải thực hiện xử lý dữ liệu âm thanh thu được qua nhiều công đoạn để đưa về dạng chuẩn hoá và tính toán đưa ra bộ tham số đặc trưng. Tiếp đó mở cơ sở dữ liệu và so sánh với tất cả các mẫu trong đó rồi đưa ra kết luận nhận dạng, cuối cùng là hiển thị từ nhận dạng được. Để nhận dạng được một từ phải xử lý rất nhiều thao tác, đặc biệt khi số lượng từ trong cơ sở dữ liệu lớn.

Do sự phức tạp của hệ thống và yêu cầu của đồ án, tôi lựa chọn ngôn ngữ Visual Basic với hệ quản trị cơ sở dữ liệu Access. Ngôn ngữ lập trình này tuy có tốc độ xử lý không cao lắm nhưng lại hỗ trợ người lập trình tốt trên cơ sở dữ liệu và có giao diện thân thiện, dễ sử dụng. Đó là ngôn ngữ có khả năng đáp ứng được yêu cầu của hệ thống.

4.3. Xây dựng bộ mẫu nhận dạng

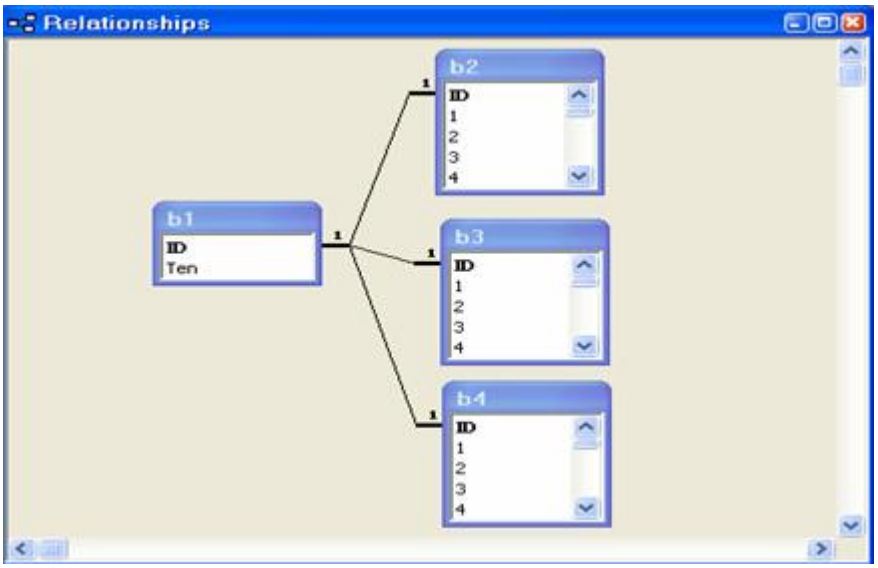
Mô hình nhận dạng từ tiếng Việt dựa trên phương pháp dự báo tuyến tính LPC-10. Mỗi một mẫu từ sẽ được chia thành các frame nhỏ. Sau đó thực hiện tính toán hệ số LPC-10 cho từng frame, cho tất cả các frame, bộ tham số này sẽ được lưu trữ trong cơ sở dữ liệu.

Ta có thể hình dung thao tác tạo dữ liệu từ 1 mẫu như sau :

+ File âm thanh được cắt để trích lấy phần dữ liệu âm thanh có tiếng nói.
+ Chuẩn hoá thời gian
+ Chuẩn hoá biên động

+ Chia file âm thanh ra thành 30 frame nhỏ (kích thước mỗi frame 110 byte). Tính hệ số LPC-10 cho mỗi một frame. Sau đó lưu trữ bộ hệ số này trong cơ sở dữ liệu.

4.4. Một số hình ảnh của chương trình



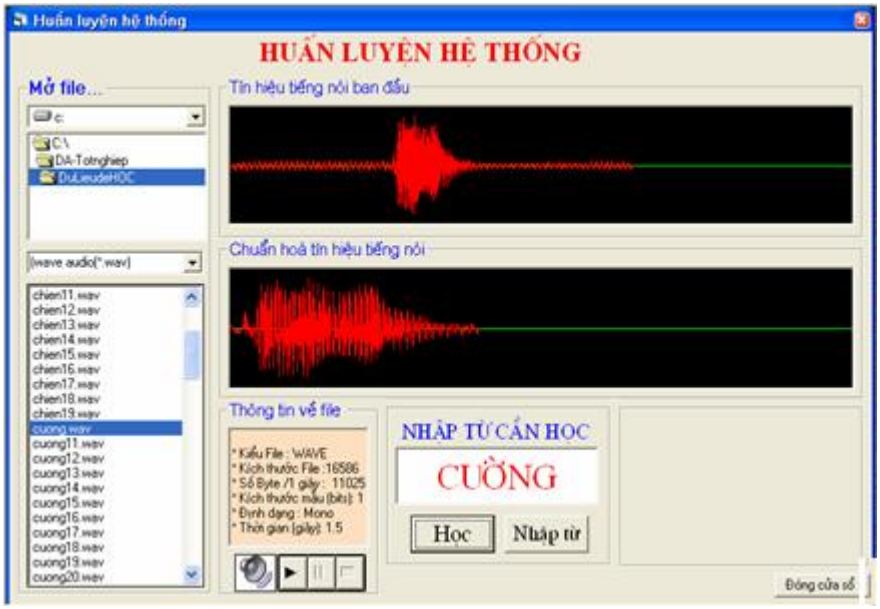
Lược đồ quan hệ cơ sở dữ liệu của chương trình

Dữ liệu được tổ chức gồm 4 bảng:

+ Bảng 1 là bảng chính: gồm 2 trường, trường thứ nhất là khóa ID autonumber. Trường thứ 2 có dạng text để lưu mẫu ký tự của từ được học.

+ 3 bảng còn lại, mỗi bảng gồm một trường khóa ID và 100 trường dạng số double để lưu 30 bộ hệ số LPC-10 (mỗi bộ hệ số LPC-10 gồm 10 số đặc trưng, 30 bộ hệ số là 300 con số tương ứng với 300 trường của tổng 3 bảng).

Các trường ID của cả 4 bảng được liên kết với nhau. Quan hệ giữa các bảng là 1-1. Mỗi mẫu âm thanh được học (1 từ được học) được lưu vào CSDL là 1 bản ghi với chỉ số ID, tên, và 300 con số đặc trưng. Do số trường lưu trữ là rất lớn nên ta tách ra thành 4 bảng



Huấn luyện hệ thống học các từ tiếng việt

Giao diện gồm hộp ảnh thứ nhất, vẽ sóng âm thanh của file âm thanh được mở để học. Hộp ảnh bên dưới để hiển thị sóng âm thanh sau khi đã cắt lấy phần chỉ có tiếng nói. Mục đích trực quan hoá dữ liệu cắt được. Một hộp TextBox để nhập từ cần học. Ngoài ra còn hiển thị thông tin về file âm thanh, phát ra loa tín hiệu âm thanh khi một file âm thanh được mở.

Nhận dạng từ tiếng việt từ file nguồn

Nhận dạng từ tiếng Việt từ Microphone

ĐÁNH GIÁ KẾT QUẢ VÀ KẾT LUẬN

Với đề tài được giao, sau thời gian thực hiện đồ án tốt nghiệp, vận dụng những kiến thức cơ bản đã được học cùng với nỗ lực bản thân, sự chỉ bảo tận tình của giáo viên hướng dẫn - Thạc Sĩ Võ Văn Tùng đồ án " **Nhận dạng tập từ hạn chế Tiếng Việt trong môi trường nhiễu**" đã hoàn thành. Chương trình đã đáp ứng được cơ bản các yêu cầu đặt ra.

Những vấn đề đạt được:

+ Đã đáp ứng được tên đề tài yêu cầu là nhận dạng từ tiếng việt

+ Khi số lượng mẫu huấn luyện lớn thì kết quả nhận dạng đạt chất lượng

+ Thử nghiệm hệ thống cho kết quả nhận dạng tốt khi mà số lượng từ không lớn (hơn 20 từ).

+ Hệ thống nhận dạng tốt với cùng người nói và những người tham gia huấn luyện mẫu.

+ Khi người nói không tham gia huấn luyện mẫu thì kết quả nhận dạng chưa được khả quan.

Các yêu cầu của để tăng chất lượng hệ thống

+ Chọn mẫu huấn luyện phải là các mẫu chuẩn, ít nhiễu

+ Tăng số lượng mẫu học

+ Kiểm tra, nghe thử trước khi cập nhật vào CSDL

Hướng phát triển của đồ án

+ Làm cơ sở để thiết kế hệ thống nhận dạng cụm từ và câu

+ Phát triển chương trình để giao tiếp với máy tính trực tiếp qua Microphone thực hiện một số câu lệnh cơ bản.

TÀI LIỆU THAM KHẢO

+ Visual Basic 6 Certification Exam Guide – Chaper 1- Dan mezick & Scot Hillier – Mcgraw- Hill – 1998.

+ "Digital Signal Processing: Principles, Algorithms, and Applications"- Prentice Hall. John G. Proakis, Dimitris G. Manolakis

+ Xử lý tín hiệu và lọc số - Nguyễn Quốc Trung.

+ Visual Basic - Lập trình cơ sở dữ liệu- Nxb Lao động xã hội-2004- Nguyễn Thị Ngọc Mai.

+ "Digital Signal Processing: A Computer-Based Approach"- McGraw-Hill. Sanjit K. Mitra

+ Xử lý tín hiệu số- Nguyễn Hữu Phương.

+ Tài liệu tham khảo môn học Xử lý tiếng nói [Lê Bá Dũng- khoa CNTT- ĐH Hàng Hải Việt Nam].

+ Voice Processing - Gordon E. Pelton năm 1993.