TẬP ĐOÀN BƯU CHÍNH VIỄN THÔNG VIỆT NAM
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
****

BÀI GIẢNG

XỬ LÝ TIẾNG NÓI

BIÊN SOẠN:

PHẠM VĂN SỰ

LÊ XUÂN THÀNH

HÀ NỘI - 2010

Lời nói đầu

Tiếng nói là một phương tiện trao đổi thông tin tiện ích vốn có của con người. Ước mơ về những "máy nói", "máy hiểu tiếng nói" đã không chỉ xuất hiện từ những câu truyện khoa học viễn tưởng xa xưa mà nó còn là động lực thôi thúc của nhiều nhà nghiên cứu, nhóm nghiên cứu trên thế giới. Hoạt động nghiên cứu và xử lý tiếng nói đã trải qua gần một thế kỷ cùng với nhiều thành tựu to lớn trong việc xây dựng phát triển các kỹ thuật công nghệ xử lý tiếng nói đã đạt được. Tuy vậy, việc có được một "máy nói" mang tính tự nhiên (về giọng điệu, phát âm...) cũng như một "máy hiểu tiếng nói" thực thụ vẫn còn khá xa vời. Xu thế phát triển của công nghệ hội tụ ở thế kỷ 21 càng thôi thúc việc hoàn thiện hơn nữa công nghệ để có thể đạt được mục tiêu của con người về lĩnh vực xử lý tiếng nói. Chính vì thế, việc nắm bắt được các kỹ thuật cơ bản cũng như các công nghệ tiến tiến cho việc xử lý tiếng nói là thực sự cần thiết cho sinh viên chuyên ngành Xử lý Tín hiệu và Truyền thông nói riêng, sinh viên chuyên ngành Kỹ thuật Điện - Điện tử nói chung. Với mục đích đó, bài giảng môn học Xử lý tiếng nói được biên soạn nhằm trang bị cho sinh viên các khái niệm cơ bản quan trọng và cần thiết cũng như nhằm giới thiệu cho sinh viên các công nghệ tiên tiến, xu thế nghiên cứu và phát triển của lĩnh vực xử lý tiếng nói. Cuốn sách được chia làm 5 chương:

1. Một số khái niệm cơ bản.
2. Biểu diễn số của tín hiệu tiếng nói.
3. Phân tích tiếng nói.
4. Tổng hợp tiếng nói.
5. Nhận dạng tiếng nói.

Các chương 1 và 2 do giảng viên Lê Xuân Thành biên soạn, các chương còn lại do giảng viên Phạm Văn Sự biên soạn. Trong thời gian gấp rút hoàn thành cuốn bài giảng này, mặc dù với sự cố gắng nỗ lực hết sức, như do kinh nghiệm còn nhiều hạn chế, nhóm tác giả không tránh khỏi những sai sót và nhầm lẫn. Nhóm tác giả chân thành mong muốn nhận được những đóng góp từ đồng nghiệp và các em sinh viên để hoàn thiện hơn trong phiên bản sau.

Mọi góp ý xin gửi về: Bộ môn Lý thuyết mạch, Khoa Kỹ thuật Điện tử I, Học viện Công nghệ Bưu chính Viễn thông, Km10 Đường Nguyễn Trãi, Hà Đông, Hà Nội hoặc gửi email về địa chỉ xulytiengnoi@gmail.com.

Hà Nội, ngày 02 tháng 05 năm 2010

Nhóm biên soạn

Danh mục các từ viết tắt

ADC	Analog Digital Converter	Bộ chuyển đổi tương tự - số
ADM	Adaptive Delta Modulation	Điều chế Delta thích nghi
ADPCM	Adaptive Differential PCM	Điều xung mã vi sai thích nghi
CSR	Continuous Speech Recognition	Nhận dạng tiếng nói liên tục
DCT	Discrete Cosine Transform	Biến đổi Cosine rời rạc
DFT	Discrete Fourier Transform	Biến đổi Fourier rời rạc
DM	Delta Modulation	Điều chế Delta
DTFT	Discrete Time FT	Biến đổi Fourier với thời gian rời rạc
DPCM	Differential PCM	Điều chế xung mã vi sai
FFT	Fast FT	Biến đổi Fourier nhanh
FIR	Finite Impulse Response	Bộ lọc đáp ứng hữu hạn
FT	Fourier Transform	Biến đổi Fourier
HMM	Hidden Markov Model	Mô hình Markov ẩn
IDFT	Inverse Discrete FT	Biến đổi Fourier rời rạc ngược
IDTFT	Inverse DTFT	Biến đổi Fourier với thời gian rời rạc ngược
IFT	Inverse FT	Biến đổi Fourier ngược
LMS	Least Mean Square	Bình phương trung bình tối thiểu
LPC	Linear Predictive Coding	Mã hóa dự đoán tuyến tính
LTI	Linear Time-Invariant	Bộ lọc tuyến tính không thay đổi theo thời gian
MFCC	Mel frequency cepstral coefficient	Các hệ số cepstral tần số Mel
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
PAM	Pulse Amplitude Modulation	Điều chế biên độ xung mã
SNR	Signal to Noise Ratio	Tỷ số tín hiệu trên nhiễu
ST	Short-time Transform	Biến đổi ngắn hạn
STFT	Short-time FT	Biến đổi Fourier ngắn hạn
TDNN	Time delay Neural Network	Mạng nơ-ron với thời gian trễ
TD-PSOLA	Time-domain PSOLA	Phương pháp chồng lấn đồng bộ pitch trong miền thời gian

Mục lục

Chương 1: Một số khái niệm cơ bản

1.1. Mở đầu

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. Đàm thoại là một quá trình gồm nhiều người, có sự hiểu hiết chung và một nghi thức luân phiên nhau nói. Những người có điều kiện thể chất và tinh 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 có rất nhiều yếu tố khác hỗ trợ nhằm giúp người nghe hiểu được ý cần diễn đạt như biểu hiện trên gương mặt, cử chỉ, điệu bộ. Vì có đặc tính tác động qua lại, nên tiếng nói được sử dụng trong nhu cầu giao tiếp nhanh chóng. Trong khi đó, chữ viết lại có khoảng cách về không gian lẫn thời gian giữa tác giả và người đọc. Sự biểu đạt của tiếng nói hỗ trợ mạnh mẽ cho việc ra đời các hệ thống máy tính có sử dụng tiếng nói, ví dụ như lưu trữ tiếng nói như là một loại dữ liệu, hay dùng tiếng nói làm phương tiện giao tiếp qua lại. Nếu chúng ta có thể phân tích quá trình giao tiếp qua nhiều lớp, thì lớp thấp nhất chính là âm thanh và lớp cuối cùng là tiếng nói diễn tả ý nghĩa muốn nói.

1.1.1 Nguồn gốc của tiếng nói

Âm thanh của lời nói cũng như âm thanh trong thế giới tự nhiên xung quanh ta, về bản chất đều là những sóng âm được lan truyền trong một môi trường nhất định (thường là không khí). Khi chúng ta nói dây thanh trong hầu bị chấn động, tạo nên những sóng âm, sóng truyền trong không khí đến màng nhĩ – một màng mỏng rất nhạy cảm của tai ta – làm cho màng nhĩ cũng dao động, các dây thần kinh của màng nhĩ sẽ nhận được cảm giác âm khi tần số dao động của sóng đạt đến một độ lớn nhất định. Tai con người chỉ cảm thụ được những dao động có tần số từ khoảng 16Hz đến khoảng 20000Hz. Những dao động trong miền tần số này gọi là dao động âm hay âm thanh, và các sóng tương ứng gọi là sóng âm. Những sóng có tần số nhỏ hơn 16Hz gọi là sóng hạ âm, những sóng có tần số lớn hơn 20000Hz gọi là sóng siêu âm, con người không cảm nhận được (ví dụ loài dơi có thể nghe được tiếng siêu âm). Sóng âm, sóng siêu âm và hạ âm không chỉ truyền trong không khí mà còn có thể lan truyền tốt ở những môi trường rắn, lỏng, do đó cũng được sử dụng rất nhiều trong các thiết bị máy móc hiện nay.

1.1.2 Phân loại tiếng nói

Tiếng nói là âm thanh mang mục đích diễn đạt thông tin, rất uyển chuyển và đặc biệt. Là công cụ của tư duy và trí tuệ, tiếng nói mang tính đặc trưng của loài người. Nó không thể tách riêng khi nhìn vào toàn thể nhân loại, và nhờ có ngôn ngữ tiếng nói mà loài người sống và phát triển xã hội tiến bộ, có văn hóa, văn minh như ngày nay. Trong quá trình giao tiếp người nói, có nhiều câu nói, mỗi câu gồm nhiều từ, mỗi từ lại có thể gồm 1 hay nhiều âm tiết. Ở tiếng Việt, số âm tiết được sử dụng vào khoảng 6700. Khi chúng ta phát ra một tiếng thì có rất nhiều bộ phận như lưỡi, thanh môn, môi, họng, thanh quản,… kết hợp với nhau để tạo thành âm thanh. Âm thanh phát ra được lan truyền trong không khí để đến tai người nhận. Vì âm thanh phát ra từ sự kết hợp của rất nhiều bộ phận, do đó âm thanh ở mỗi lần nói khác nhau hầu như khác nhau dẫn đến khá khó khăn khi ta muốn phân chia tiếng nói theo những đặc tính riêng. Người ta chỉ chia tiếng nói thành 3 loại cơ bản như sau:

- **Âm hữu thanh**: Là âm khi phát ra thì có thanh, ví dụ như chúng ta nói “i”, “a”, hay “o” chẳng hạn. Thực ra âm hữu thanh được tạo ra là do việc không khí qua thanh môn

(thanh môn tạo ra sự khép mở của dây thanh dưới sự điều khiển của hai sụn chóp) với một độ căng của dây thanh sao cho chúng tạo nên dao động.

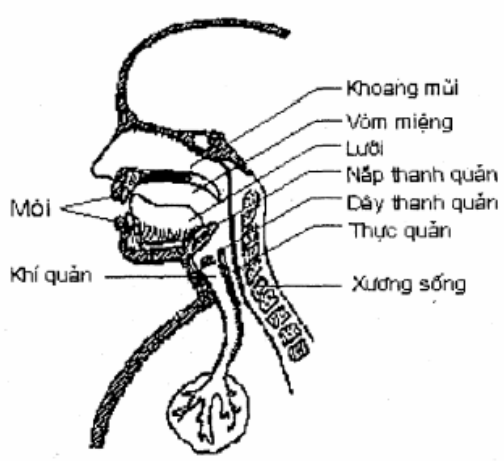
- **Âm vô thanh**: Là âm khi tạo ra tiếng thì dây thanh không rung hoặc rung đôi chút tạo ra giọng như giọng thở, ví dụ "h", "p" hay "th".
- **Âm bật**: Để phát ra âm bật, đầu tiên bộ máy phát âm phải đóng kín, tạo nên một áp suất, sau đó không khí được giải phóng một cách đột ngột, ví dụ "ch", "t".

1.2. Quá trình tạo tiếng nói

1.2.1 Cấu tạo của hệ thống cấu âm

Lời nói là kết quả của sự hoạt động với mối liên kết giữa các bộ phận hô hấp và nhai. Hành động này diễn ra dưới sự kiểm soát của hệ thần kinh trung ương, bộ phận này thường xuyên nhận được thông tin bằng những tác động ngược của các bộ phận thính giác và cảm giác bản thể. Bộ máy hô hấp cung cấp lực cần thiết khi khí được thở ra bằng khí quản. Ở đỉnh khí quản là thanh quản nơi áp suất khí được điều biến trước khi đến tuyến âm kéo dài từ hầu đến môi (hình 1.1).

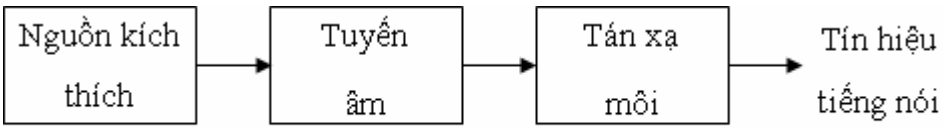
Thanh quản là tập hợp các cơ và sụn động bao quanh một khoang nằm ở phần trên của khí quản. Các dây thanh giống như là một đôi môi đối xứng nằm ngang thanh quản, hai môi này có thể khép hoàn toàn thanh quản và khi mở ra chúng có thể tạo ra độ mở hình tam giác gọi là thanh môn. Không khí qua thanh quản một cách tự do trong quá trình thở và cả trong quá trình cấu âm của những âm điếc hay âm vô thanh. Còn các âm hữu thanh thì lại là kết quả của sự rung động tuần hoàn của những dây thanh. Và như vậy những rung động liên tiếp sẽ đến được tuyến âm. Tuyến âm là tập hợp những khoang nằm giữa thanh môn và môi, trên hình ta có thể phân biệt được khoang hầu (họng), khoang miệng và khoang mũi.



Hình 1.1 Hệ thống phát âm của con người

Khi nói, lồng ngực mở rộng và thu hẹp, không khí được đẩy từ phổi vào khí quản, đi qua thanh môn do các dây thanh tạo thành. Luồng khí này được gọi là tín hiệu kích cho tuyến âm vì sau đó nó được đẩy qua tuyến âm và cuối cùng tán xạ ra ở môi. Tuyến âm có thể được coi như một ống âm học (gồm các đoạn ống với độ dài bằng nhau và thiết diện các mặt cắt khác nhau mắc nối tiếp) với đầu vào là các dây thanh (hay thanh môn) và đầu ra là môi. Như vậy tuyến âm có dạng thay đổi như một hàm theo thời gian. Các mặt cắt của tuyến âm được xác định bằng vị trí của lưỡi, môi, hàm, vòm miệng và thiết diện của những mặt cắt này thay đổi từ $0cm^2$ (khi ngậm môi) đến khoảng $20cm^2$ (khi hở môi). Tuyến mũi tạo thành tuyến âm học

phụ trợ cho truyền âm thanh, nó bắt đầu từ vòm miệng và kết thúc ở các lỗ mũi. Khi vòm miệng hạ thấp, tuyến mũi được nối với tuyến âm về mặt âm học và tạo nên tiếng nói âm mũi. Các âm của tiếng nói được tạo trong hệ thống này theo ba cách phụ thuộc vào tín hiệu kích. âm hữu thanh như âm /i/ được tạo nên khi kích tuyến âm bằng chuỗi xung (hay chu kỳ dao động của đôi dây thanh) xác định chu kỳ pitch T và đại lượng nghịch đảo của nó là tần số cơ bản F_0. Đối với ngôn ngữ có thanh điệu thì kiểu thay đổi này còn phụ thuộc vào thanh điệu. Âm vô thanh như âm /s/ được tạo nên khi các dây thanh không dao động, xung kích được coi như các tạp ngẫu nhiên, kích bởi các dòng khí xoáy qua các chỗ hẹp của tuyến âm (thường là phía khoang miệng). Âm nổ như âm /p/ được tạo ra bằng cách đóng hoàn toàn tuyến âm, gây nên áp suất bên cạnh vị trí đóng, rồi nhanh chóng giải phóng âm này. Vì tuyến âm và tuyến mũi bao gồm các ống âm học có mặt cắt khác nhau nên khi âm truyền trong ống, phổ tần số thay đổi theo tính chọn lọc tần số của ống. Trong phạm vi tạo tiếng nói, những tần số cộng hưởng của tuyến âm được gọi là tần số formant hay đơn giản là formant. Những tần số này phụ thuộc vào dạng và kích thước của tuyến âm, do đó mỗi dạng tuyến âm được đặc trưng bằng một tổ hợp tần số formant. Các âm khác nhau được tạo bởi sự thay đổi dạng của tuyến âm. Như vậy tính chất phổ của tín hiệu tiếng nói thay đổi theo thời gian giống với sự thay đổi dạng của tuyến âm. Quá trình truyền âm qua tuyến âm làm mạnh lên ở một vùng tần số nào đó bằng cộng hưởng và tạo cho mỗi âm những tính chất riêng biệt gọi là quá trình phát âm. Âm được phát có nghĩa nó đã mang thông tin về âm vị được tán xạ ra ngoài từ môi. Trong một vài trường hợp, đối với những âm mũi (như /m/, /n/ trong tiếng Anh), tuyến mũi cũng tham gia vào quá trình phát âm và âm được tán xạ ra từ mũi. Tóm lại, sóng tín hiệu được chế tạo bằng ba động tác: tạo nguồn âm (hữu thanh và vô thanh), phát âm khi truyền qua tuyến âm và tán xạ âm từ môi hoặc từ mũi, như hình 1.2 sau đây:



Hình 1.2 Quá trình cơ bản tạo tín hiệu tiếng nói

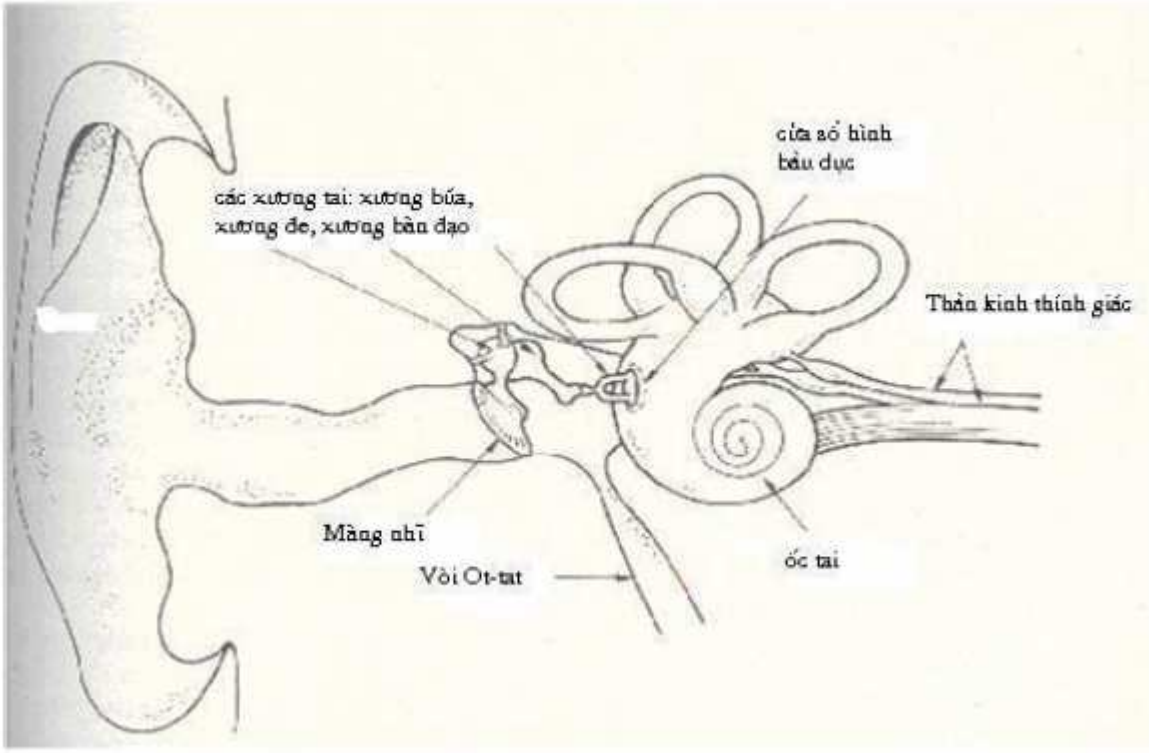
1.2.2 Cấu tạo của hệ thống tiếp âm

Không giống như các cơ quan tham gia vào quá trình tạo ra tiếng nói khi thực hiện các chức năng khác trong cơ thể như: thở, ăn, ngửi. Tai chỉ sử dụng cho chức năng nghe. Tai đặc biệt nhạy cảm với những tần số trong tín hiệu tiếng nói chứa thông tin phù hợp nhất với việc liên lạc (những tần số xấp xỉ 200 – 5600Hz). Người nghe có thể phân biệt được những sự khác biệt nhỏ trong thời gian và tần số của những âm thanh nằm trong vùng tần số này.

Tai gồm có ba phần: tai ngoài, tai giữa và tai trong. Tai ngoài dẫn hướng những thay đổi áp xuất tiếng nói vào trong màng nhĩ, ở đó tai giữa sẽ chuyển đổi áp xuất này thành chuyển động cơ học. Tai trong chuyển đổi những rung động cơ học này thành những luồng điện trong nơron thính giác dẫn đến não.

Tai ngoài: bao gồm LOA TAI (pina) hay TÂM NHĨ (aurical) và LỖ (meatus) thính giác hay ống tai ngoài. Loa tai có tham gia rất ít hoặc hầu như không vào độ thính của tai, nhưng

có chức năng bảo vệ lối vào ống tai và dường như cũng tham gia vào khả năng khu biệt các âm, đặc biệt là ở những tần số cao hơn. Loa tai nối với ống tai ngoài, một ống ngắn có hình dáng thay đổi có chiều dài khoảng từ 25 đến 53 cm làm đường cho các tín hiệu âm học đến tai giữa. Lỗ tai có hai chức năng chính. Chức năng thứ nhất là bảo vệ các cấu trúc phức tạp và không có tính chất cơ học lắm của tai giữa. Chức năng thứ hai là đóng vai trò như một bộ máy cộng hưởng hình ống vốn ưu tiên cho việc truyền các âm có tần số cao giữa 2000 Hz và 4000Hz. Chức năng này là quan trọng đối với việc tiếp nhận lời nói và đặc biệt trợ giúp cho việc tiếp nhận các âm xát, vì đặc điểm của chúng thường được lập mã trong nguồn năng lượng không có chu kì trong khu vực ảnh phổ âm học này. Sự cộng hưởng trong lỗ thính giác cũng tham gia vào độ thính chung của chúng ta giữa 500Hz và 4000Hz, vốn là một dải tần có chứa nhiều dấu hiệu chính đối với cấu trúc âm vị học.

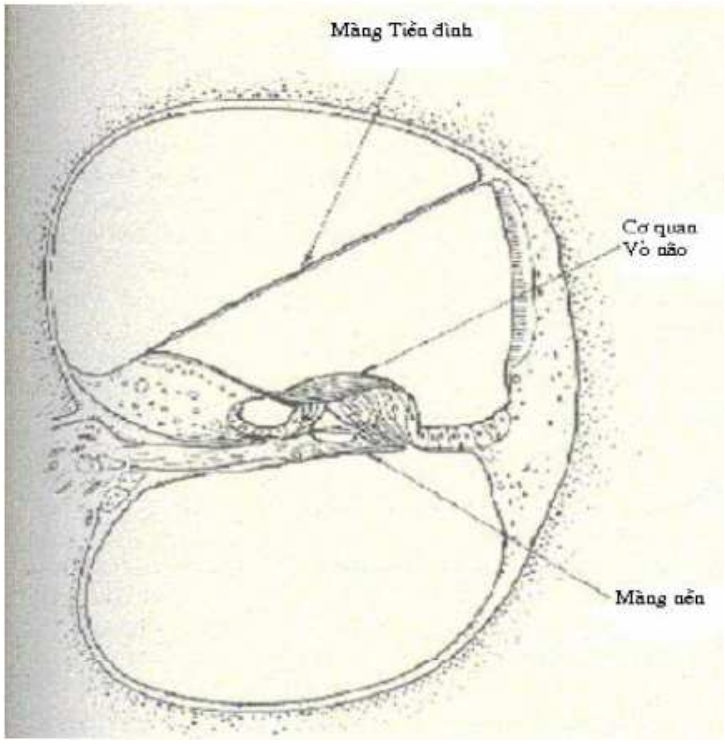


Hình 1.3 Cấu trúc hệ thính giác ngoài

Tai giữa bao gồm một khoang nằm trong cấu trúc hộp sọ có chứa màng nhĩ (eardrum) - màng ở đầu trong của ống tai ngoài , một bộ ba khúc xương liên kết với nhau, được gọi là xương vồ (mallet), xương đe (anvil) và xương bàn đạp (stirrup) (cũng có thuật ngữ là xương tai (auditory ossicle)) và cấu trúc cơ liên kết. Mục đích của tai giữa là truyền những biến đổi áp suất âm trong không khí đến tai ngoài vào những dịch chuyển cơ khí tương ứng. Quá trình truyền này bắt đầu ở màng nhĩ, bị làm lệch đi bởi những biến đổi áp suất khí truyền đến nó qua lỗ tai. Sự dịch chuyển này được truyền đến các xương tai, vốn đóng vai trò như một hệ thống đòn bẩy cơ học khéo léo để chuyển tải những dịch chuyển này đến cửa hình bầu dục ở giao diện đến tai trong và chất dịch trong lỗ tai ở trên.

Hoạt động làm đòn bẩy của các xương tai, và sự thực là màng nhĩ có vùng bề mặt lớn hơn nhiều so với cửa hình bầu dục, đảm bảo cho việc truyền hiệu ứng của năng lượng âm học giữa 500Hz và 4000Hz, làm tăng đến mức tối đa khả năng thính của tai ở vùng tần số này. Hệ cơ gắn với các xương tai cũng hoạt động để bảo vệ tai chống lại những âm lớn do hoạt động cơ

chế phản xạ âm học. Cơ chế này đi vào hoạt động khi các âm có biên độ khoảng 90dB và lớn hơn truyền đến tai: hệ cơ kết hợp và sắp xếp lại các xương tai để làm giảm hiệu quả truyền âm đến cửa hình bầu dục (Borden và Harris 1980, Moore 1989). Tai giữa được nối với họng bằng một ống hẹp gọi là vòi ốc tai (eustachian tube). Điều này hình thành một đường khí và con đường này sẽ mở ra khi cần cân bằng những thay đổi áp suất khí nền giữa cấu trúc tai giữa và tai ngoài. Tai trong là một cấu trúc phức tạp bọc trong hộp sọ, ốc tai (cochlea) có trách nhiệm biến đổi sự chuyển dịch cơ khí thành các tín hiệu thần kinh: sự dịch chuyển cơ khí được truyền đến cửa hình bầu dục bằng các ốc tai được chuyển thành các tín hiệu thần kinh và các tín hiệu thần kinh này được truyền đến hệ thống thần kinh trung ương. Về cơ bản, ốc tai là một cấu trúc hình xoắn tận hết bằng một cửa sổ có một màng linh hoạt ở mỗi đầu. Ở bên trong, ốc tai chia thành hai màng, một trong số đó, màng nền (basilar membrane) là cực kì quan trọng đối với hoạt động nghe. Khi những dịch chuyển (do các rung động âm gây ra) diễn ra tại cửa sổ hình bầu dục, chúng được truyền qua chất dịch trong ốc tai và gây ra sự dịch chuyển (displacement) của màng nền. Ở một đầu màng nền cứng hơn so với ở đầu kia, và điều này có nghĩa là cách thức mà trong đó nó được dịch chuyển phụ thuộc vào tần số của âm tác động vào. Các âm có tần số cao sẽ gây ra sự dịch chuyển lớn hơn ở đầu cứng; với tần số giảm dần, sự dịch chuyển cực đại sẽ di chuyển liên tục về phía đầu ít cứng hơn. Gắn dọc với màng nền là cơ quan vỏ não (organ of corti), một cấu trúc phức tạp chứa nhiều tế bào tóc. Nó là sự dịch chuyển và sự kích thích của các tế bào tóc này vốn biến sự dịch chuyển của màng nền thành các tín hiệu thần kinh. Vì màng nền được dịch chuyển ở nhiều vị trí khác nhau phụ thuộc vào tần số, cho nên ốc tai và các cấu trúc bên trong của nó có thể biến tần số và cường độ của âm thành các tín hiệu thần kinh. Nhưng cần phải nhấn mạnh rằng sự tái hiện có tính thần kinh cuối cùng của thông tin tần số không phụ thuộc vào vị trí của chỉ riêng sự dịch chuyển màng nền không, và hiểu biết của chúng ta về cách thức tần số được lập mã thông qua hệ thống thính giác là chưa hoàn thiện.



Hình 1.4 Mặt cắt ngang của ốc tai

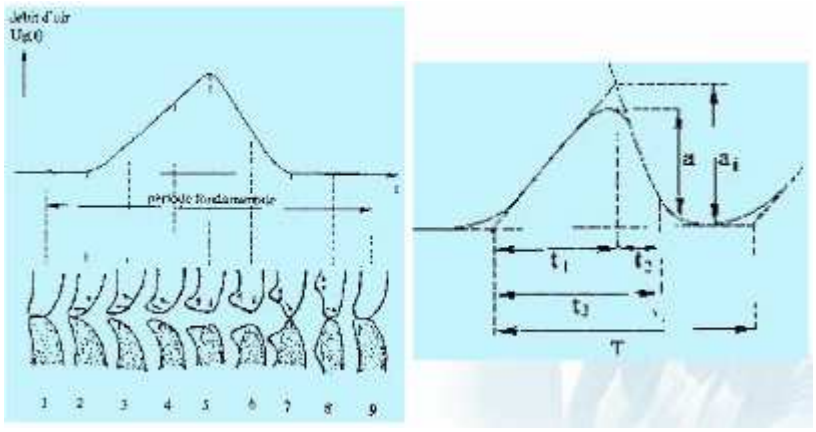
Nghiên cứu đầu tiên về thẩm nhận lời nói chỉ tính đến rất ít các thuộc tính thẩm nhận cơ bản của tai. Hơn nữa, nó đã cố gắng gắn kết các thuộc tính thẩm nhận của tín hiệu lời nói với kiểu tái hiện phổ thay đổi theo thời gian tuyến tính. Đến khoảng năm 1980 nhiều nhà nghiên cứu đã nhận ra rằng cần phải hiểu những hiệu ứng có tính chất phân tích của hệ thính giác người về các tín hiệu lời nói và thật là sai lầm khi cho rằng người nghe chỉ đang xử lí thông tin theo cách giống như chiếc máy ghi phổ bình thường mà thôi.

1.3. Các đặc tính cơ bản của tiếng nói

1.3.1 Tần số cơ bản và phổ tần

Thông lượng: thể tích không khí vận chuyển qua thanh môn trong một đơn vị thời gian (khoảng $1cm^3$ /s).

Chu kỳ cơ bản T_0: khi dây thanh rung với chu kỳ T_0 thì thông lượng cũng biến đổi tuần hoàn theo chu kỳ này và ta gọi T_0 là chu kỳ cơ bản.



Hình 1.5 Tần số cơ bản

Giá trị nghịch đảo của T_0 là F_0=1/ T_0 được gọi là tần số cơ bản của tiếng nói. F_0 phụ thuộc vào giới tính và lứa tuổi của người phát âm; F_0 thay đổi theo thanh điệu và F_0 cũng ảnh hưởng đến ngữ điệu của câu nói.

1.3.2 Biểu diễn tín hiệu tiếng nói

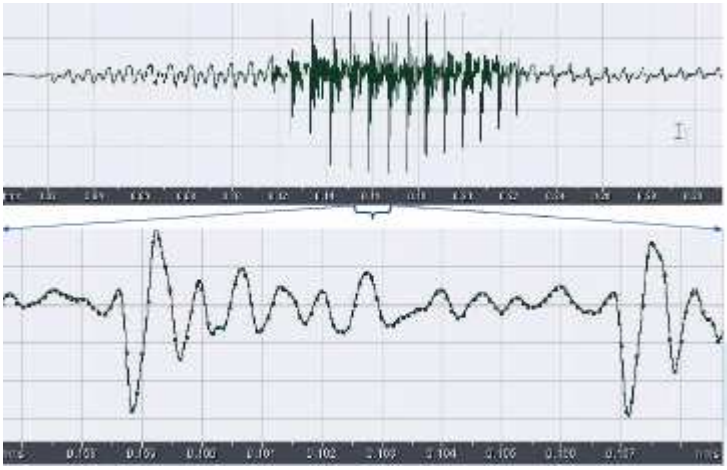
Có 3 phương pháp biểu diễn tín hiệu tiếng nói cơ bản là:

- Biểu diến dưới dạng sóng theo thời gian.
- Biểu diến trong miền tần số: phổ của tín hiệu tiếng nói.
- Biểu diễn trong không gian 3 chiều (Sonagram)

a) Dạng sóng theo thời gian

Phần tín hiệu ứng với âm vô thanh là không tuần hoàn, ngẫu nhiên và có biên độ hay năng lượng nhỏ hơn của nguyên âm (cỡ khoảng 1/3).

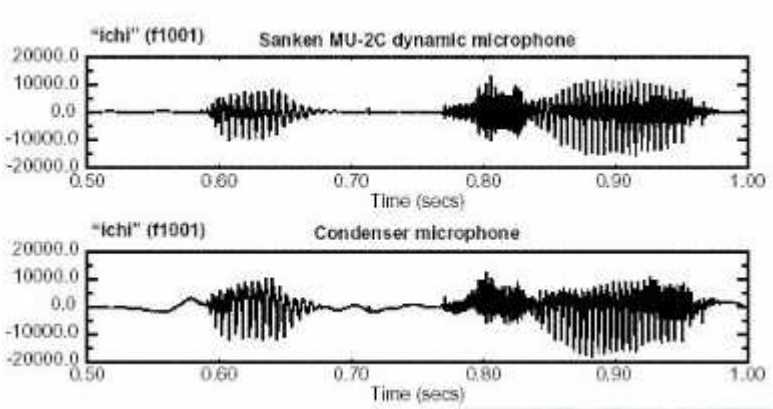
Ranh giới giữa các từ: là các khoảng lặng (Silent). Ta cần phân biệt rõ các khoảng lặng với âm vô thanh.



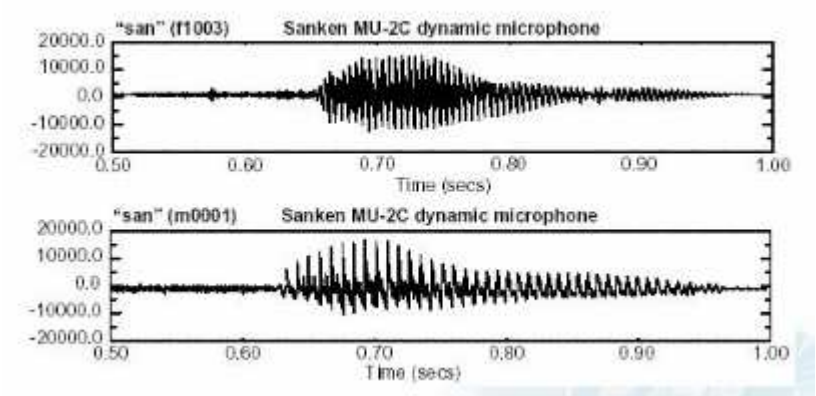
Hình 1.6 Dạng sóng theo thời gian

Âm thanh dưới dạng sóng được lưu trữ theo định dạng thông dụng trong máy tính là *.WAV với các tần số lấy mẫu thường gặp là: 8000Hz, 10000Hz, 11025Hz, 16000Hz, 22050Hz, 32000Hz, 44100Hz,…; độ phân giải hay còn gọi là số bít/mẫu là 8 hoặc 16 bít và số kênh là 1 (Mono) hoặc 2 (Stereo).

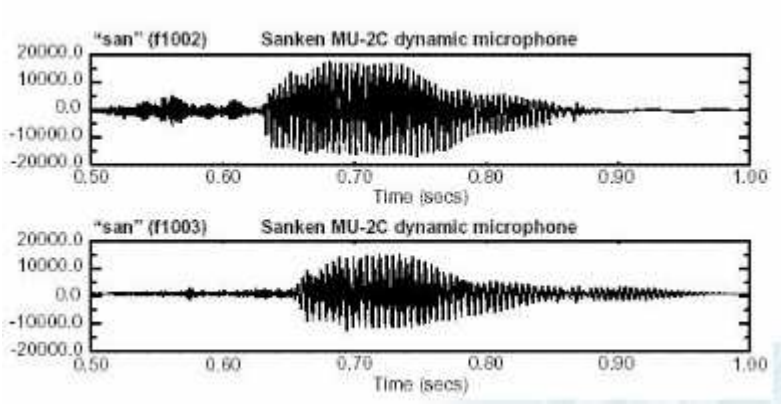
Như vậy, dữ liệu lưu trữ của tín hiệu âm thanh sẽ khác nhau tuỳ theo máy thu thanh, thời điểm phát âm hay người phát âm, điều này được thể hiện rõ nét trong các hình vẽ sau:



Hình 1.7 Âm thanh được thu bằng 2 micro khác nhau



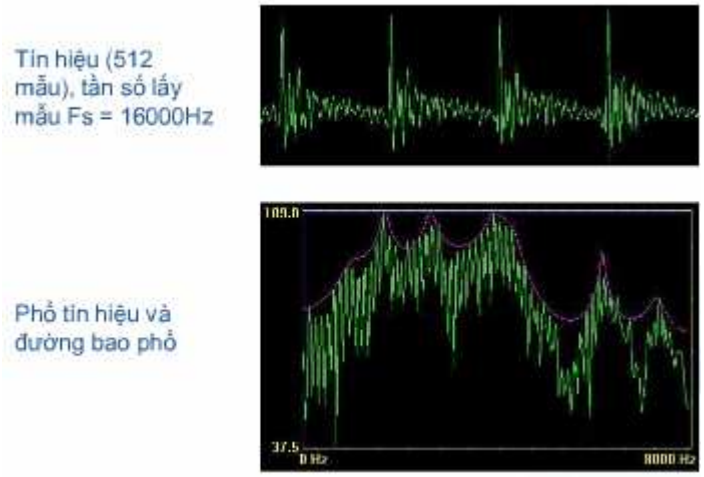
Hình 1.8 Âm thanh do hai ng ười khác nhau phát ra



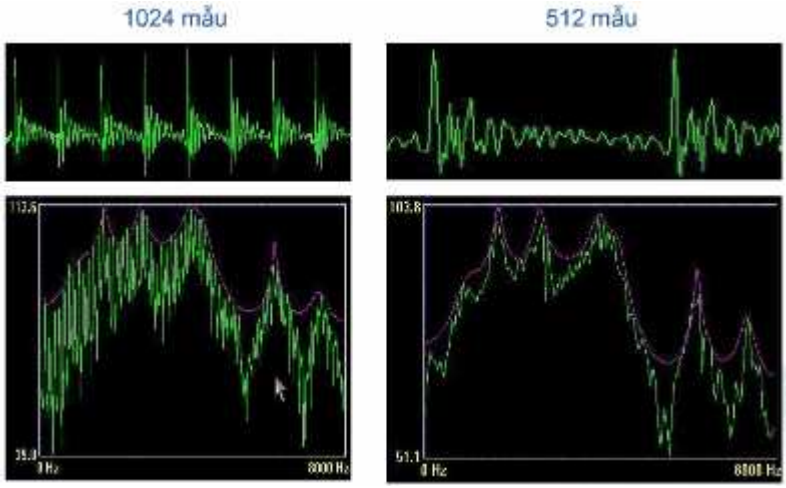
Hình 1.9 Âm thanh do một người phát ra ở hai thời điểm khác nhau

b) Phổ tín hiệu tiếng nói

Ở phần trên ta đã biết rằng dải tần số của tín hiệu âm thanh là khoảng từ 0Hz đến 20KHz, tuy nhiên phần lớn công suất nằm trong dải tần số từ 0,3KHz đến 3,4KHz. Dưới đây là một số hình ảnh của phổ tín hiệu tiếng nói:



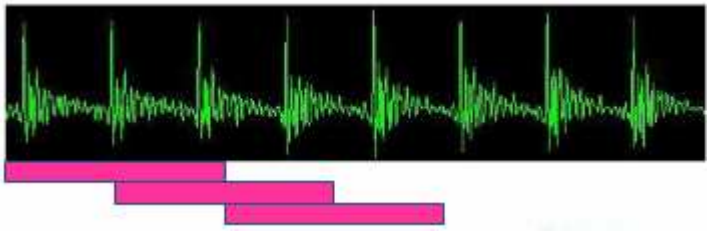
Hình 1.10 Phổ tín hiệu tiếng nói và đường bao phổ



Hình 1.11 Phổ tín hiệu tiếng nói với số mẫu khác nhau

c) Biểu diễn tín hiệu tiếng nói trong không gian ba chiều (Sonagram)

Để biểu diễn trong không gian 3 chiều người ta chia tín hiệu thành các khung cửa sổ (frame) ứng với các ô quan sát như hình vẽ 1.12.



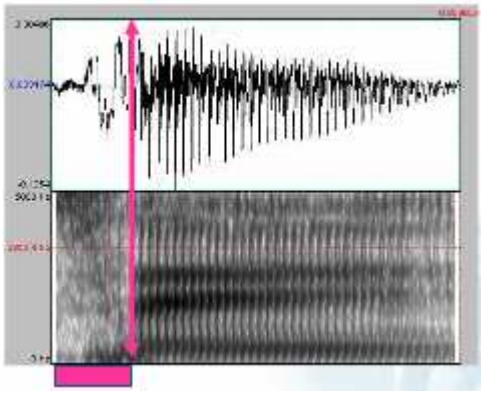
Hình 1.12 Chia tín hiệu thành các khung cửa sổ

Độ dài một cửa sổ tương ứng là 10ms.

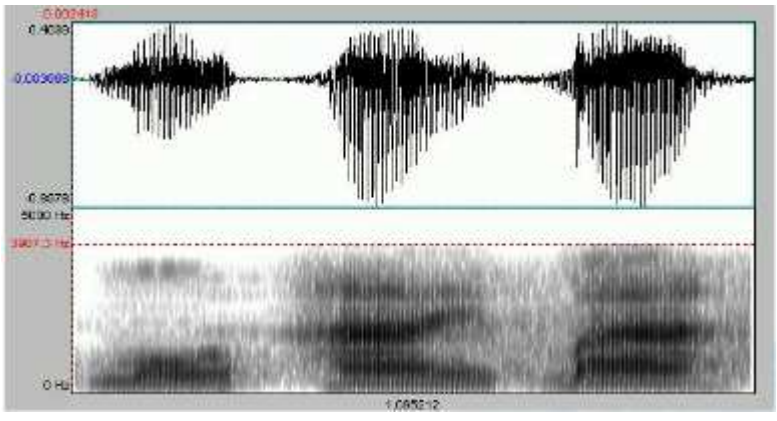
Vậy, nếu tần số F_s = 16000Hz thì ta có 160 mẫu trên một cửa sổ.

Các cửa số có đoạn chồng lẫn lên nhau (khoảng 1/2 cửa sổ).

Tiếp theo ta vẽ phổ của khung tín hiệu trên trục thẳng đứng, biên độ phổ biểu diễn bằng độ đậm, nhạt của màu sắc. Sau đó ta vẽ theo trục thời gian bằng cách chuyển sang cửa số tiếp theo.



Hình 1.13 Phổ của một khung cửa sổ



Hình 1.14 Các khung cửa sổ liền nhau và spectrogram tương ứng

Biểu diễn tín hiệu tiếng nói theo không gian 3 chiều là một công cụ rất mạnh để quan sát và phân tích tín hiệu. Ví dụ : theo phương thức biểu diễn này ta có thể dễ dàng phân biệt âm vô thanh và âm hữu thanh dựa theo các đặc điểm sau:

+Âm vô thanh:

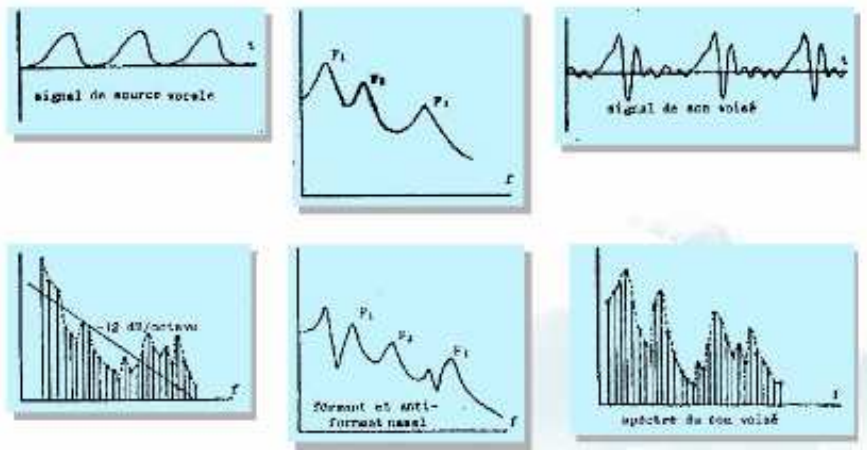
- Năng lượng tập trung ở tần số cao.

- Các tần số phân bố khá đồng đều trong 2 miền tần số cao và tần số thấp.

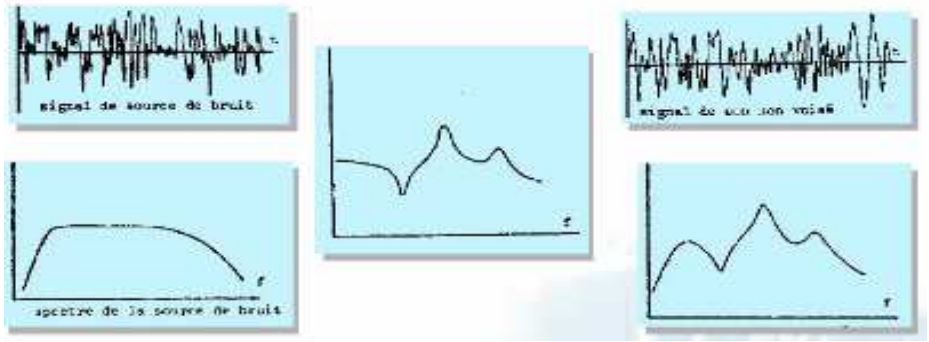
+ Âm hữu thanh:

- Năng lượng tập không đồng đều.

- Có những vạch cực trị.



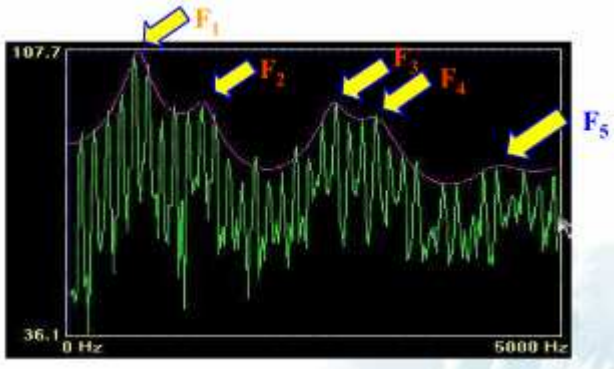
Hình 1.15 Âm hữu thanh



Hình 1.16 Âm vô thanh

d) Formant và Antiformant

Tuyến âm được coi như một hốc cộng hưởng có tác dụng tăng cường một tần số nào đó. Những tần số được tăng cường lên được gọi là các Formant. Nếu khoang miệng được coi là tuyến âm thì khoang mũi cũng được coi như là một hốc cộng hưởng. Khoang mũi và khoang miệng được mắc song song nên sẽ làm suy giảm một tần số nào đó và những tần số bị suy giảm này được gọi là các AntiFormant.



Hình 1.17 Đường bao phổ và các Formant

Dựa trên hình 1.17 ta thấy có thể tính đến Formant thứ 5 (F5) nhưng quan trọng nhất cần chú ý ở đây là các F1 và F2. Cùng một người phát âm nhưng Formant có thể khác nhau. Nếu ta chỉ căn cứ vào giá trị của Formant để đặc trưng cho âm hữu thanh thì chưa chính xác mà phải dựa vào phân bố tương đối giữa các Formant. Ngoài ra, nếu xác định Formant trực tiếp từ phổ thì không chính xác mà phải dựa vào đường bao phổ, đây cũng chính là đáp ứng tần số của tuyến âm.

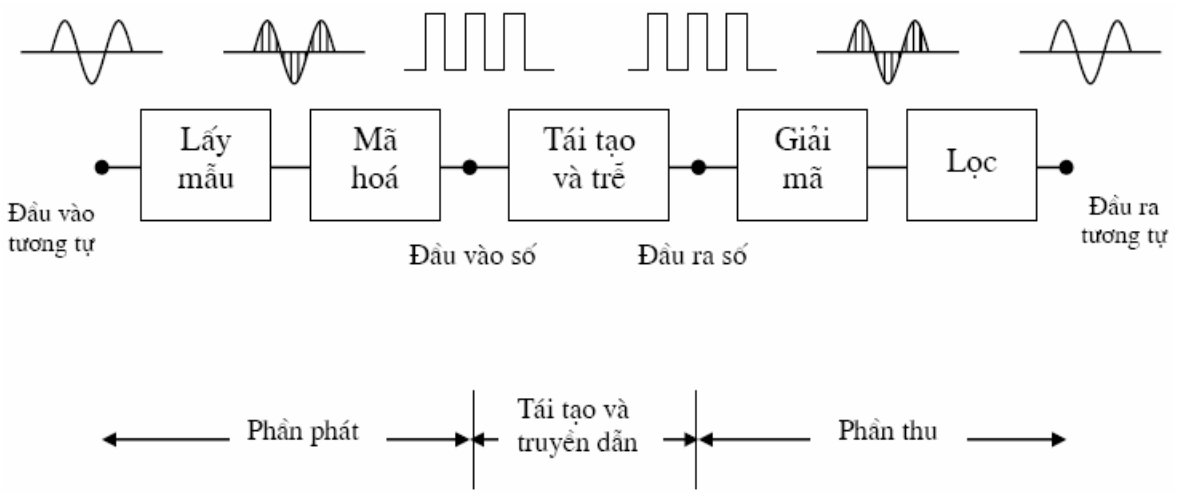
Chương 2: Biểu diễn số của tín hiệu tiếng nói

2.1. Mở đầu

Mã hoá là quá trình biến đổi các giá trị rời rạc thành các mã tương ứng. Nhìn chung, việc lấy mẫu liên quan tới quá trình biến đổi các tín hiệu liên tục thành các tín hiệu rời rạc của trường thời gian gọi là PAM (điều chế biên độ xung mã). Việc mã hoá là quá trình lượng tử hoá các giá trị mẫu này thành các giá trị rời rạc của trường biên độ và sau đó biến đổi chúng thành mã nhị phân hay các mã ghép kênh. Khi truyền thông tin mã, nhiều xung được yêu cầu cho mỗi giá trị lấy mẫu và vì thế độ rộng dải tần số cần thiết cho truyền dẫn phải được mở rộng. Đồng thời xuyên âm, tạp âm nhiệt, biến dạng mẫu, mất xung mẫu, biến dạng nén, tạp âm mã hoá, tạp âm san bằng được sinh ra trong lúc tiến hành lấy mẫu và mã hoá. Việc giải mã là quá trình khôi phục các tín hiệu đã mã hoá thành các tín hiệu PAM được lượng tử hoá. Quá trình này tiến hành theo thứ tự đảo đúng như quá trình mã hoá. Mặt khác quá trình lượng tử hoá, nén và mã hoá các tín hiệu PAM được gọi là quá trình mã hoá và quá trình chuyển đổi các tín hiệu PCM thành D/A, sau đó, lọc chúng sau khi giãn để đưa về tiếng nói ban đầu gọi là quá trình giải mã. Cấu hình cơ sở của hệ thống truyền dẫn PCM đối với việc thay đổi các tín hiệu tương tự thành các tín hiệu xung mã để truyền dẫn được thể hiện ở hình (pcm1). Trước tiên các tín hiệu đầu vào được lẫy mẫu một cách tuần tự, sau đó được lượng tử hoá thành các giá trị rời rạc trên trục biên độ. Các giá trị lượng tử hoá đặc trưng bởi các mã nhị phân. Các mã nhị phân này được mã hoá thành các dạng mã thích hợp tuỳ theo đặc tính của đường truyền dẫn.

Thiết bị đầu cuối mã hoá chuyển đổi các tín hiệu thông tin như tiếng nói thành các tín hiệu số như PCM. Khi các tín hiệu thông tin là các tín hiệu tương tự, việc chuyển đổi A/D được tiến hành và việc chuyển đổi D/D đợc tiến hành ở trường hợp của các tín hiệu số. Đôi khi, quá trình nén và mã hoá băng tần rộng được tiến hành bằng cách triệt sự dư thừa trong quá trình tiến hành chuyển đổi A/D hoặc D/D).

Các quy luật đối với PCM vi phân thích ứng 32Kbps có nén giãn như mã hoá dự đoán của các tín hiệu tiếng được chỉ rõ trong các khuyến nghị G712 của ITU. Phương pháp ADPCM 32 Kbps được chấp nhận vào tháng 10 năm 1984 được dùng để chuyển đổi các tín hiệu PCM 64 Kbps theo luật A hay luật μ hiện nay sang các tín hiệu ADPCM. Phương pháp 32 Kbps ADPCM có khả năng chuyển một lượng tiếng nói lớn gấp hai lần thậm trí còn nhiều hơn phương pháp qui ước 64 Kbps PCM, được chấp nhận một cách rộng rãi bởi bộ chuyển mã hoặc các thiết bị đầu cuối mã hoá với hiệu quả cao. Hiện nay các nước tiên tiến trên thế giới đang tiến hành nghiên cứu một cách ráo riết về công nghệ mã hoá tốc độ không những cho thoại mà cả truyền hình. Cụ thể sẽ bàn đến tiếp ở các phần tiếp theo.



Hình 2.1 Cấu hình hệ thông truyền và xử lý thông tin cơ bản

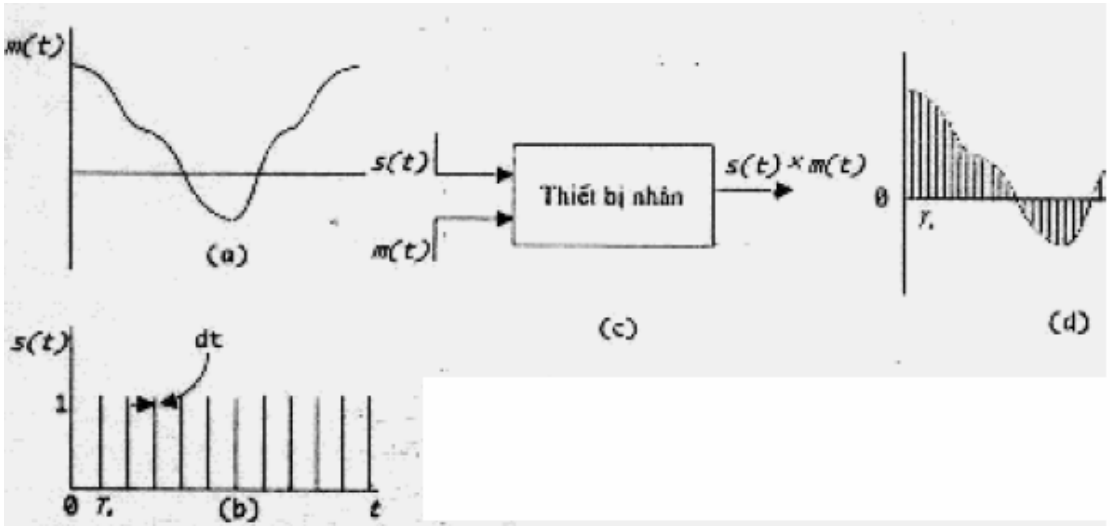
2.2. Lấy mẫu tín hiệu tiếng nói

Nguyên tắc cơ bản của điều xung mã là quá trình chuyển đổi các tín hiệu liên tục như tiếng nói thành tín hiệu số rời rạc và sau đó tái tạo chúng lại thành thông tin ban đầu. Để tiến hành việc này, các phần tử thông tin được rút ra từ các tín hiệu tương tự một cách tuần tự. Quá trình này được gọi là công việc lấy mẫu.

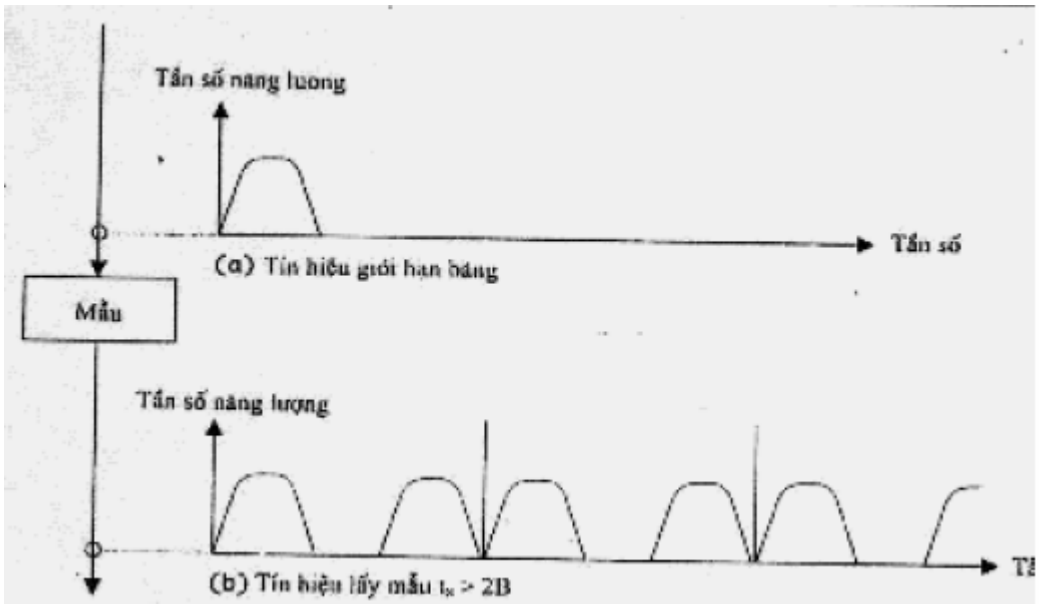
- Tín hiệu tiếng nói m(t).
- Xung lấy mẫu s(t).
- Chức danh lấy mẫu.
- Tín hiệu PAM đã lấy mẫu.

Theo thuyết lấy mẫu của Shannon, các tín hiệu ban đầu có thể được khôi phục khi tiến hành công việc lấy mẫu trên các phần tử tín hiệu được truyền đi lớn hơn hoặc bằng hai lần tần số cao nhất. Các tín hiệu xung lấy mẫu là tín hiệu dạng sóng chu k, là tổng các tín hiệu sóng hài có đường bao hàm số sin đối với các tần số. Vì thế, phổ tín hiệu tiếng nói tạo ra sau khi đã qua quá trình lấy mẫu thể hiện ở hình 2.3.

Có hai kiểu lấy mẫu tuỳ theo dạng của đỉnh độ rộng xung, lấy mẫu tự nhiên và lấy mẫu đỉnh bằng phẳng. Lấy mẫu tự nhiên được tiến hành một cách lý tưởng khi phổ tần số sau khi lấy mẫu trùng với phổ của các tín hiệu ban đầu. Tuy nhiên trong các hệ thống thực tế, điều này không thể có được. Khi tiến hành lấy mẫu đỉnh bằng phẳng, một sự nén gọi là hiệu ứng biên độ lấy mẫu làm xuất hiện méo. Ngoài ra, nếu các phần tử tín hiệu đầu vào vượt quá độ rộng dải tần 4 KHz, xuất hiện sự nén quá nếp gấp. Vì vậy, việc lọc băng rộng các tín hiệu đầu vào phải được tiến hành trước khi lấy mẫu.



Hình 2.2 Quá trình lấy mẫu

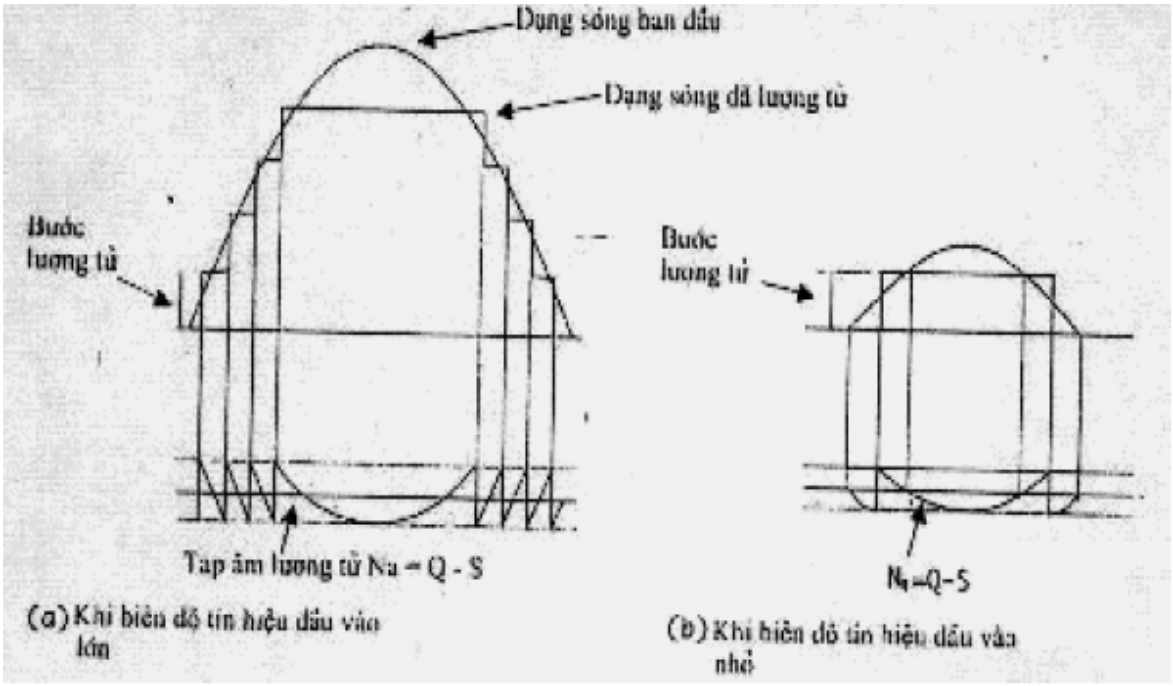


Hình 2.3 Phổ tín hiệu trước và sau lấy mẫu

2.3. Lượng tử hóa

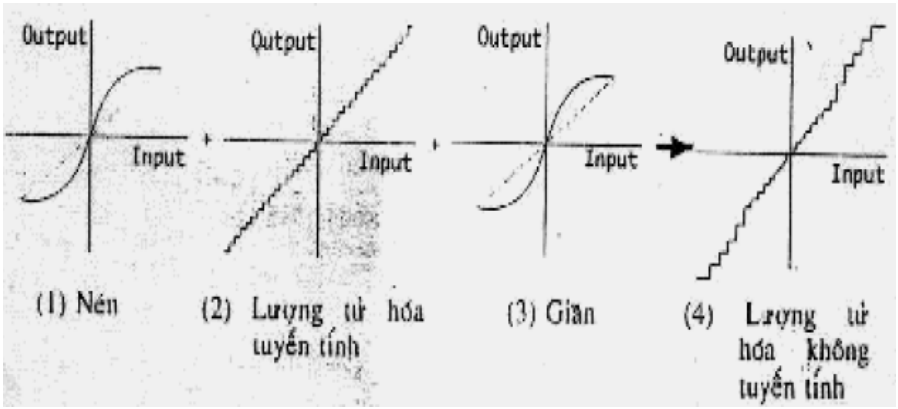
PAM với biên độ tương tự chuyển đổi thành các tín hiệu số là các tín hiệu rời rạc sau khi đi qua quá trình lượng tử hoá. Khi chỉ thị biên độ của tiếng nói liên tục với số lượng hạn chế, nó được đặc trưng với dạng sóng xấp xỉ của bước. Tạp âm lượng tử NQ = Q ư S tồn tại giữa dạng sóng ban đầu (S) và dạng sóng đã lượng tử (Q); nếu bước nhỏ tạp âm lượng tử được giảm đi nhưng số lượng bước đầu cần thiết cho lượng tử toàn bộ dải tín hiệu đầu vào trở nên rộng hơn. Vì thế số lượng các dãy số mã hoá tăng lên.

Tạp âm tạo ra khi biên độ của các tín hiệu đầu vào vượt quá dãy lượng tử gọi là tạp âm quá tải hay tạp âm bão hoà. S/NQ được sử dụng như một đơn vị để đánh giá những ưu điểm và nhược điểm của phương pháp PCM. Khi số lượng các dãy số mã hoá trên mỗi mẫu tăng lên 1 bit, S/NQ được mở rộng thêm 6 dB.

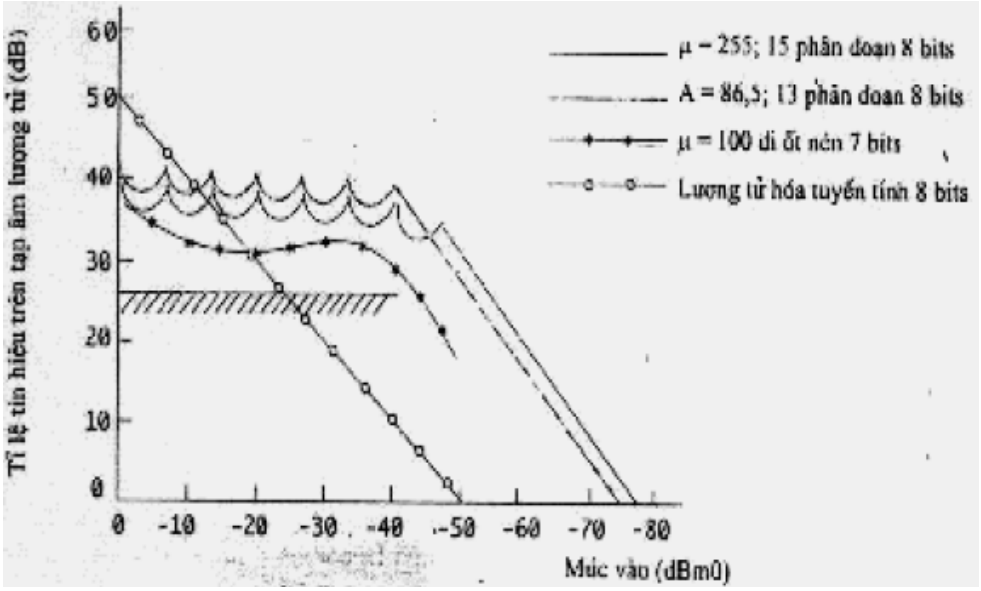


Hình 2.4 Tạp âm lượng tử theo biên độ của tín hiệu đầu vào

Như phương pháp tiến hành mã hoá hoặc giải mã, mã đường, mã không phải mã đường và mã đánh giá có thể được lựa chọn theo các kiểu của nguồn thông tin. Mã đường là một quá trình triệt số lượng tạp âm lượng tử sinh ra trên thông tin được gửi đi bất chấp mức đầu vào. Nó được sử dụng trong một hệ thống ở đó giá trị tuyệt đối của số lượng tạp âm là tới hạn hơn S/NQ. Mã không phải là mã đường được sử dụng rộng dãi trong một hệ thống ở đó S/N của hệ thống thu được quan trọng hơn số lượng tuyệt đối của tạp âm như tiếng nói. Khi bước lượng tử là một hằng số, S/N thay đổi theo mức tín hiệu. Chất lượng gọi trở nên xấu hơn khi mức tín hiệu thấp. Vì thế đối với các tín hiệu mức thấp, bước lượng tử được giảm và đối với các tín hiệu mức cao nó được tăng để ít hoặc nhiều cân bằng S/N với mức tín hiệu đầu vào. Những vấn đề trên được tiến hành bằng cách nén biên độ. Một cách lý tưởng, đối với các tín hiệu mức thấp đường cong nén và giãn là truyến tính. Đối với các tín hiệu mức cao chúng đặc trưng bởi đường cong đại số. Hiện nay, ITU-T khuyến nghị luật μ (μ =255) là phương pháp 15 đoạn (các hệ thống của Hoa Kỳ và Nhật) và luật (A= 87,6) (các hệ thống của châu âu, trong đó có Việt nam) là phương pháp 13 đoạn như là phương pháp nén đoạn mà các hàm đại số được biểu diễn gần đúng với một vài đường tuyến tính.



Hình 2.5 Lượng tử hoá tuyến tính và phi tuyến



Hình 2.6 Các đặc tính S/NQ của các phương pháp lượng tử

Cả hai phương pháp mã hoá và phương pháp nén là đồng thời được tiến hành qua bước nén số ư số hoặc tự mã hoá mà không thêm những mạch riêng rẽ khác bởi sử dụng tính chất tuyến tính của phương pháp nén đoạn trong số. Một bảng giá trị với phương pháp mã hoá và cách nén mã μ =255 được chỉ ra trên bảng 2.1.

Bảng mã hoá μ=255		**Bảng giải mã μ= 255**
Mã vào hướng tuyến tính	**Mã nén**	**Mã ra hướng tuyến tính**
0 0 0 0 0 0 0 1 w x y z a	0 0 w x y z	0 0 0 0 0 0 0 1 w x y z 1
0 0 0 0 0 0 1 w x y z a b	0 0 w x y z	0 0 0 0 0 0 1 w x y z 1 0
0 0 0 0 0 1 w x y z a b c	0 1 w x y z	0 0 0 0 0 1 w x y z 1 0 0
0 0 0 0 1 w x y z a b c d	0 1 w x y z	0 0 0 0 1 w x y z 1 0 0 0
0 0 0 1 w x y z a b c d e	1 0 w x y z	0 0 0 1 w x y z 1 0 0 0 0
0 0 1 w x y z a b c d e f	1 0 w x y z	0 0 1 w x y z 1 0 0 0 0 0
0 1 w x y z a b c d e f g	1 1 w x y z	0 1 w x y z 1 0 0 0 0 0 0
1 w x y z a b c d e f g h	1 1 w x y z	1 w x y z 1 0 0 0 0 0 0 0

Bảng 2.1 Bảng mã hoá và giải mã với μ =255

2.4. Mã hóa và giải mã

Mã hoá là một quá trình so các giá trị rời rạc nhận được bởi quá trình lượng tử hoá với các xung mã. Thông thường các mã nhị phân được sử dụng cho việc mã hoá là các mã nhị phân tự nhiên, các mã Gray (các mã nhị phân phản xạ), và các mã nhị phân kép. Phần lớn các kí hiệu mã so sánh các tín hiệu vào với điện áp chuyển để đánh giá xem có các tín hiệu nào không. Như vậy, một bộ phận chuyển đổi D/A hoặc bộ giải mã là cần thiết cho việc tạo ra điện áp

chuẩn. Trong liên lạc công cộng PCM, tiếng nói được biểu diễn với 8 bits. Tuy nhiên trong trường hợp của luật μ, các từ PCM được lập nên như sau (8 bits).

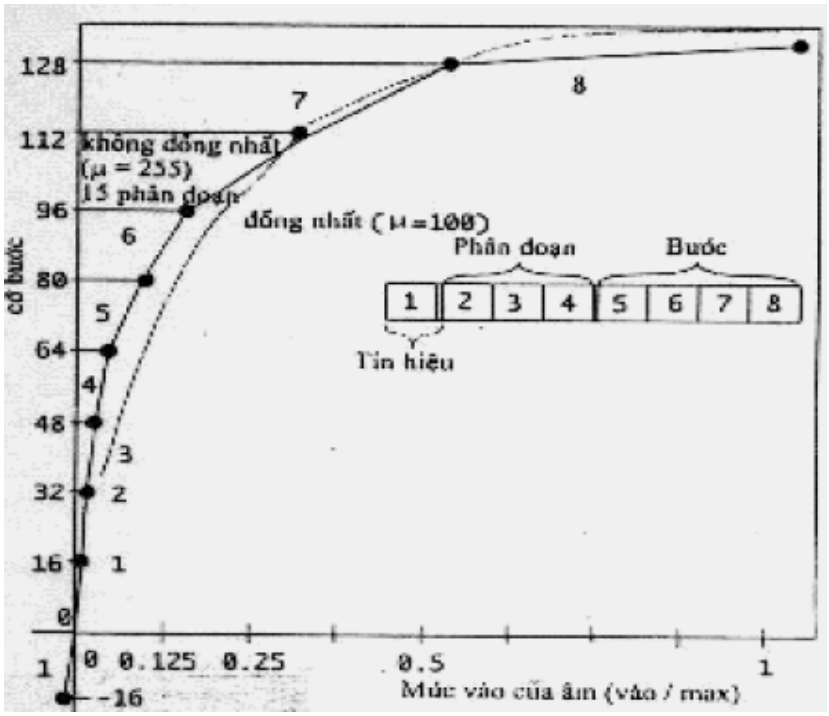
Bit phân cực = {0,1}.

Bit phân đoạn = { 000, 001,…, 111}.

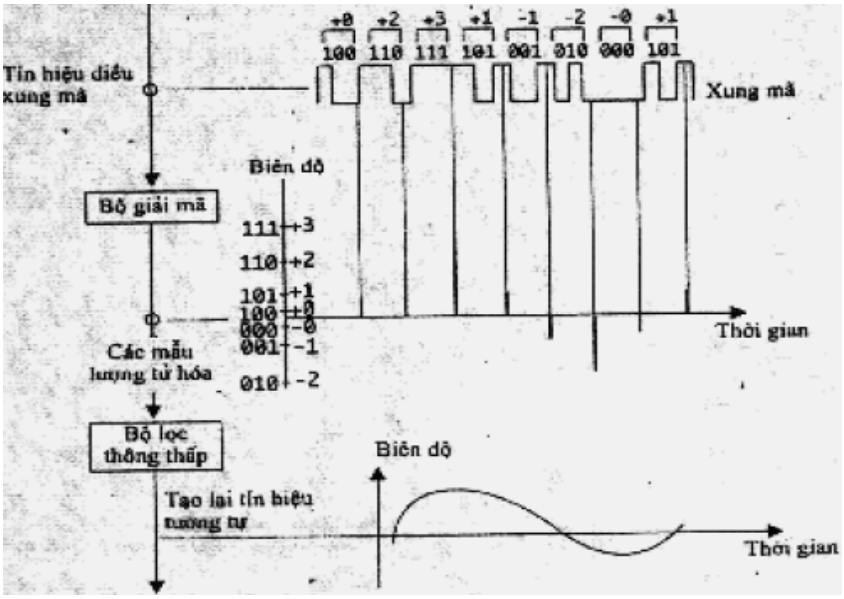
Bit phân bước = {0000, 0001,… , 1111}.

Từ đoạn thứ nhất của tín hiệu "+" và tín hiệu "ư" là các đường thẳng, có 15 phân đoạn. Cực "+" của dạng sóng tín hiệu tương ứng với bit phân cực 0 và cực "ư", với "1".

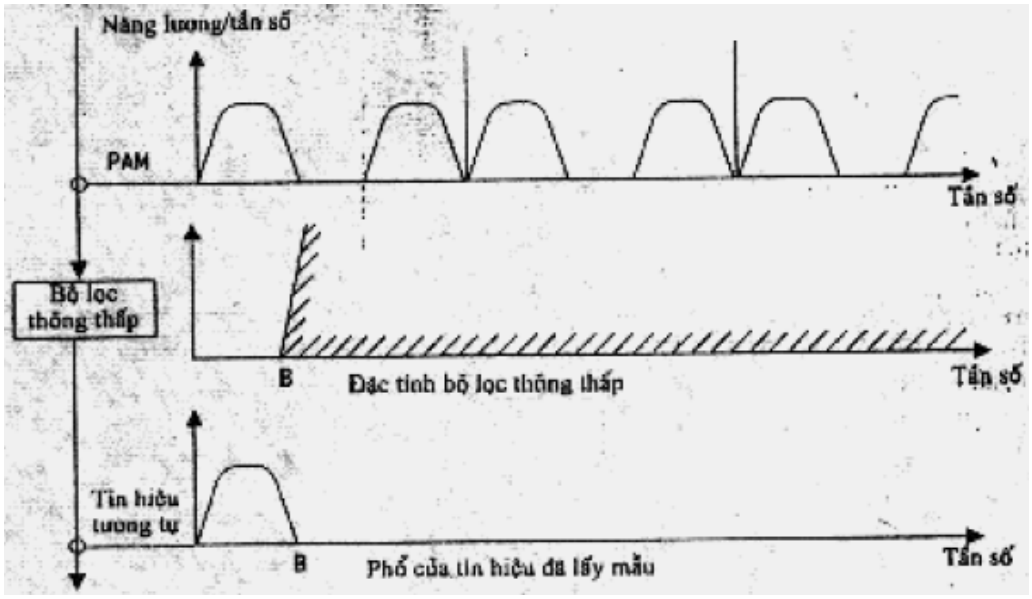
Việc báo hiệu được thực hiện sau khi thay đổi "0" của từ PCM sang "1" và "1" sang "0" và vì thế, một lượng lớn số 1 đã được thu thập chung quanh mức 0 và sự tách các tín hiệu thời gian trong khi thu nhận có thể dễ dàng thực hiện. B8 là bít thứ 8 của từ PCM, đôi khi được dùng như là một bit báo hiệu. B7 (hoặc B8) chuyển đổi sang "1" khi mọi từ của PCM là "0". Như vậy, trong các tín hiệu PCM được gửi đi, các số "0" liên tục luôn luôn ít hơn 16. Mặt khác, khi sử dụng phương pháp Bắc Mỹ, bit B2 của mọi kênh được thay đổi thành "0" nhằm chuyển đi thông tin cảnh báo cho đối phương. ở Nhật Bản, bit "S" đó là một phần của khung các bit chỉ định được dùng thay thế cho mục đích này. Các từ PCM nhận được, được chuyển đổi thành các tín hiệu PAM bởi bộ giải mã. ở phía thu, các xung tương ứng với mỗi kênh được chọn lọc từ các dẫy xung ghép kênh để tạo ra các tín hiệu PAM. Rồi, các tín hiệu tiếng nói được phục hồi bằng một bộ lọc thông thấp.



Hình 2.7 Mã hoá từ PCM



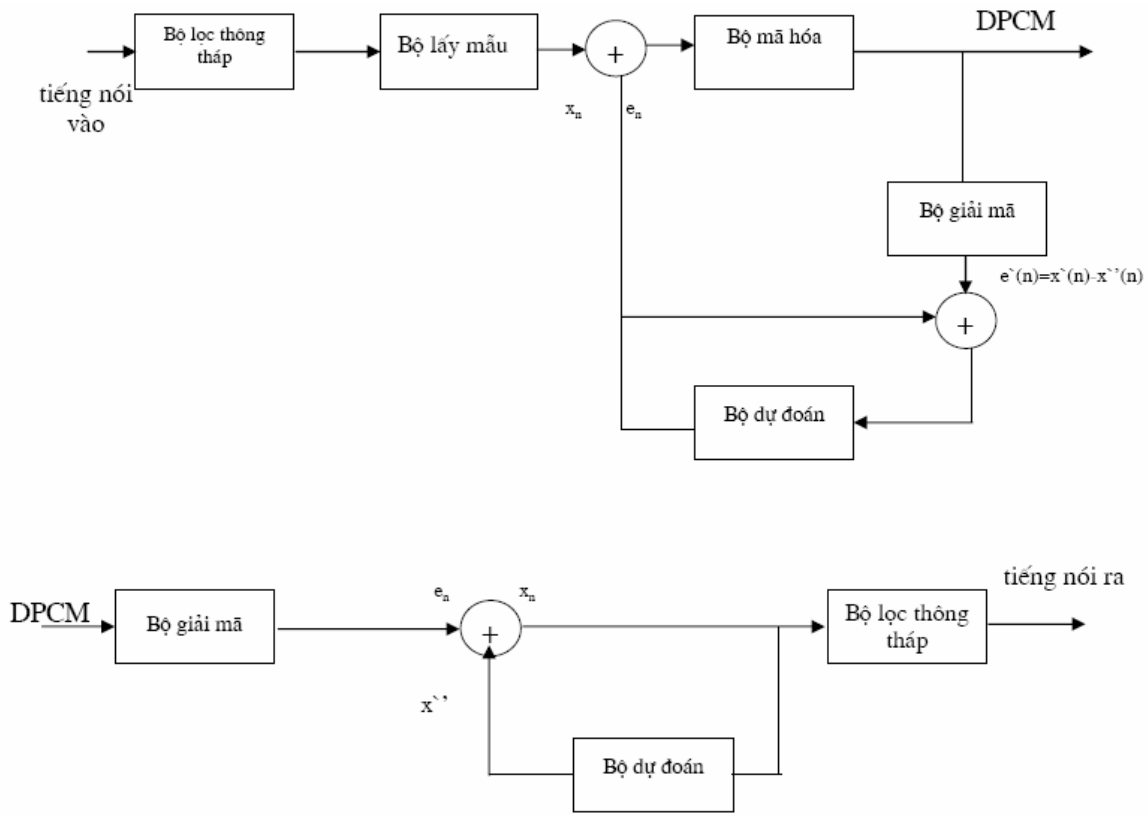
Hình 2.8 Quá trình giải mã



Hình 2.9 Quá trình giải mã và phổ

2.5. Điều chế xung mã vi sai DPCM

Đây là phương pháp dựa trên tính chất tương quan của tín hiệu tiếng nói, chỉ truyền đi độ chênh lệch giữa các mẫu cạnh nhau của tín hiệu tiếng nói:



Hình 2.10 Sơ đồ mã hoá và giải mã DPCM

Tín hiệu tiếng nói tương tự vào qua bộ lọc thông thấp, hạn chế băng tần của tín hiệu vào (thường là một nửa tần số lấy mẫu), máy phát lượng tử và mã hoá lượng tử trênh lệch giữa xung lấy mẫu tương tự x_n và tín hiệu dự đoán x_n lấy từ đầu ra bộ dự đoán x'_n. Giá trị dự đoán của mẫu tiếp theo có được nhờ ngoại suy từ p giá trị mẫu cho trước:

$$x'(n) = \sum_{i=1}^{p} a_i x'_{n-i} \tag{2.1}$$

a_i là hệ số của các bộ dự đoán, độ chênh lệch giữa xung lấy mẫu đầu vào và tín hiệu ra lấy mẫu là:

$$e_n = x_n - x'(n) \tag{2.2}$$

Đây chính là giá trị dùng để lượng tử hoá và truyền đi, ở phía thu sẽ tiến hành hồi phục lại tín hiệu sai số này và tích phân lại công với tín hiệu đã hồi phục trước đó, tuy nhiên để giảm lỗi cộng lại của nhiều lần ta dùng phia thu một bộ dự đoán giống với phía phát. Việc sử dụng vòng phản hồi giúp cho bộ lượng tử hạn chế độ chênh lệch giữa sai số e_n và s_i số được lượng tử e'_n (e'_n-e_n). Nếu giá trị này càng nhỏ thì chất lượng tiếng nói càng tốt, theo các tính toán thì phương pháp này có độ rộng băng tần đi một nửa.

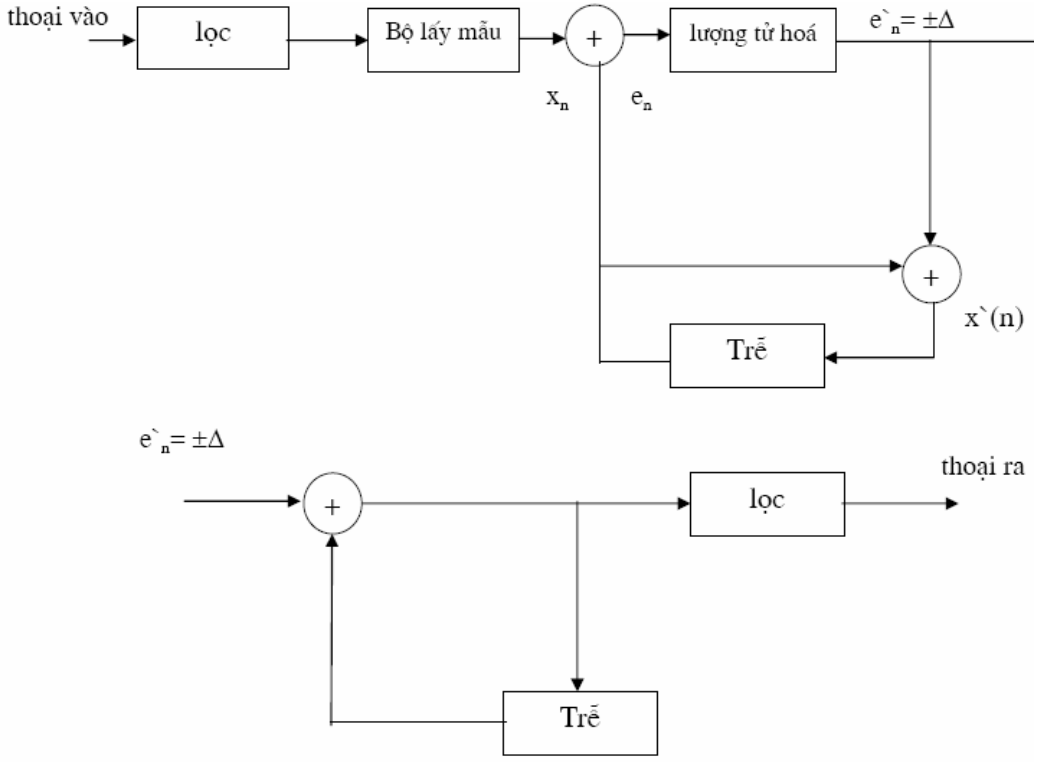
2.6. Điều chế Delta (DM)

Điều chế DM là một loại điều chế DPCM trong đó mỗi từ mã chỉ có một bít nhị phân, có ưu điểm mạch điện dễ dàng chế tạo (hình dưới). Tín hiệu thoại sau khi được lọc băng tần 0,3-3,4Khz được rời rạc hoá tạo thành tín hiệu PAM x_n, so sánh tín hiệu này với tín hiệu dự đoán x'_n, độ lệch giữa hai giá trị này (e_n) được lượng tử thành một trong hai giá trị -Δ, hoặc +Δ. Phía ra bộ lượng tử hoạ sẽ truyền đi một bit nhị phân cho mỗi xung lấy mẫu. Tại phía thu các giá trị ±Δ được cộng với các giá trị dự đoán tức thời phía ra bộ giải mã khôi phục lại tiếng

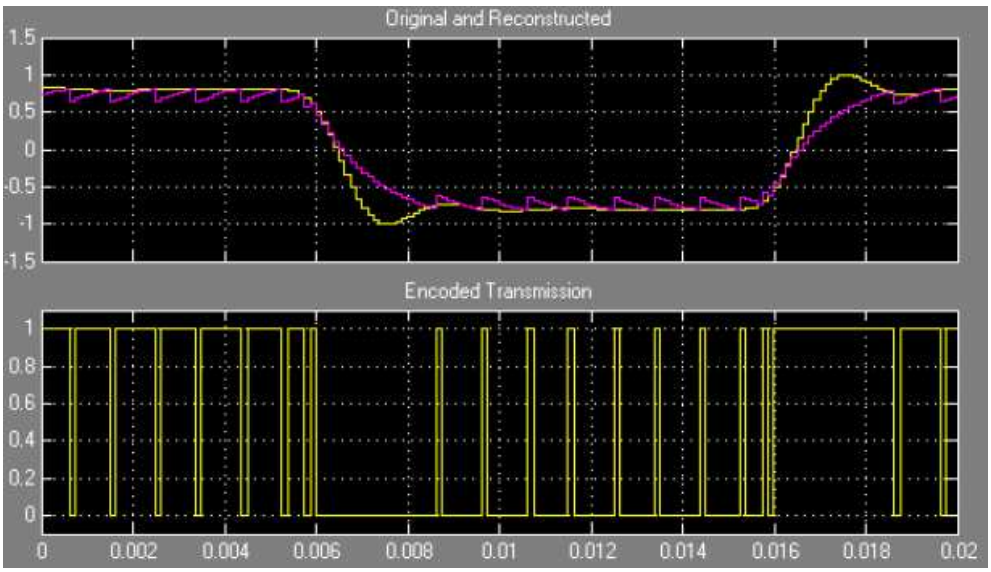
nói ban đầu. Tốc độ bit của điều chế delta bằng tốc độ của tần số lấy mẫu, tức là 8 kbps. Phương pháp này như đã nói là khá đơn giản, đạt được tốc độ mã hoá rất thấp, nó là phương pháp duy nhất của phương pháp mã hoá dạng sóng có thể so sánh về tốc độ với phương pháp tham số nguồn về tốc độ, song chất lượng tín hiệu mã hoá không cao, không đảm bảo được phạm vi động của hệ thống PCM.

2.7. Điều chế Delta thích nghi (ADM)

Phương pháp này còn gọi là phương pháp điều chế delta có độ dốc thay đổi liên tục. Phương pháp này khắc phục cho điều chế delta về khả năng dải động, phương pháp này dựa trên phương pháp thay đổi động hệ số khuyếch đại của bộ tích phân phù hợp với mức công suất trung bình của tín hiệu vào.

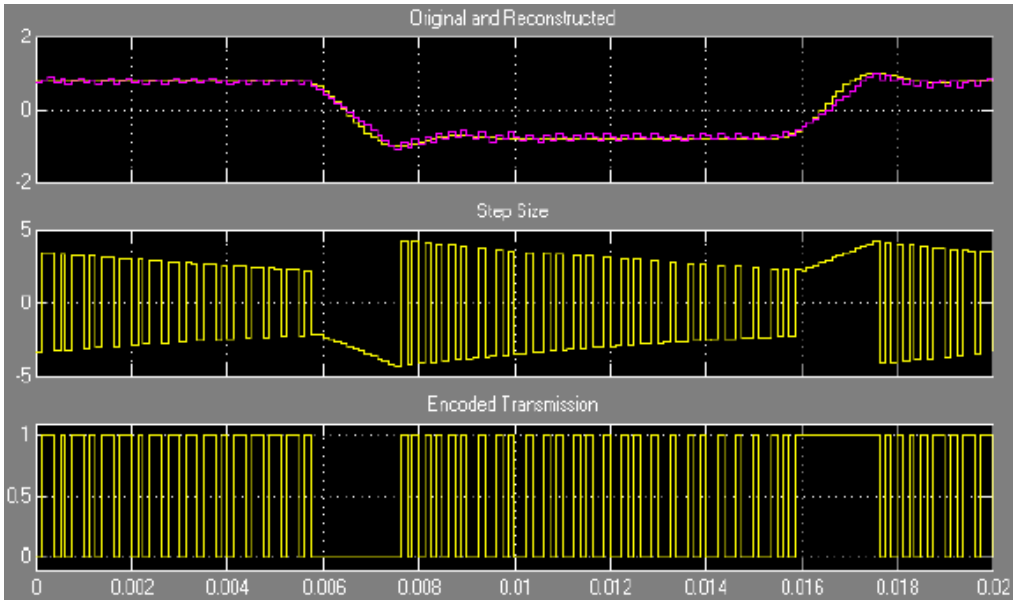


Hình 2.11 Sơ đồ mã hoá và giải mã Delta

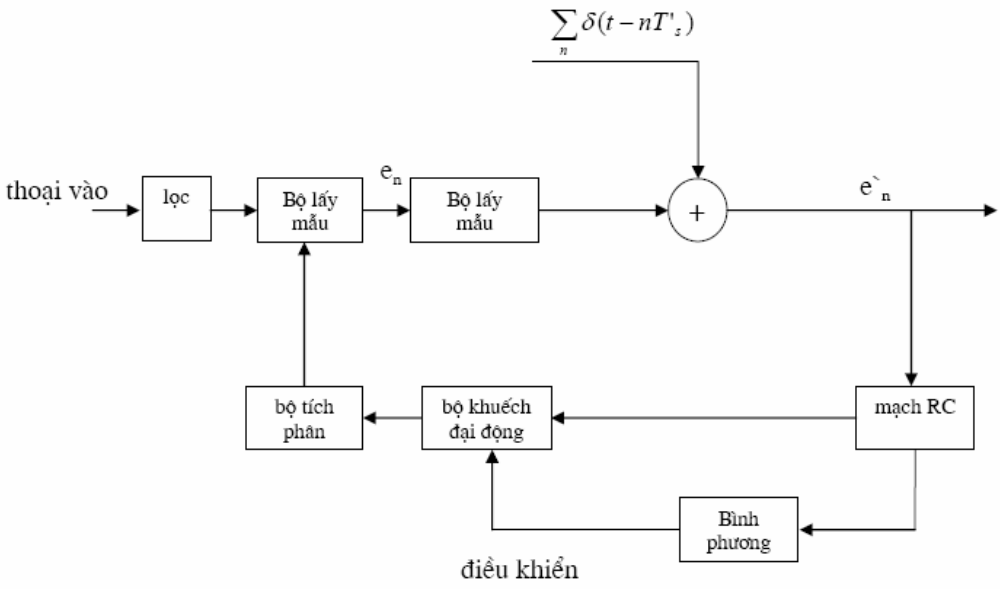


Hình 2.12 Dạng sóng tín hiệu của điều chế DM

Cỡ của bước lượng tử thay đổi nhờ thay đổi hệ số khuyếch đại của bộ tích phân nhờ mạch RC và mạch bình phương, khi tín hiệu vào là hằng số hoặc thay đổi chậm theo thời gian thì bộ điều chế này sẽ tìm kiếm và đưa ra một dãy xung có cực tính xen kẽ, mạch RC lấy trung bình các dãy này, khi nó đưa ra gía trị bằng zero. Có nghĩa là tín hiệu điều khiển làm hệ số khuyếch đại của bộ khuyếch đại thay đổi rất ít. Đầu ra bộ khuyếch đại có bước Δ kích thước nhỏ, khi tín hiệu vào có sườn dốc thì hàm bậc thang được tạo ra để kịp độ dốc của tín hiệu vào. Lúc đó sẽ tạo ra một loạt xung âm mạch RC lấy trung bình loạt xung này và đưa ra điện áp điều khiển lớn, tức là cỡ của bước tăng lên, nhờ mạch bình phương nên điện điều khiển bộ khuyếch đại luôn luôn dương, mà không phụ thuộc cực tính của xung thế nào phương pháp này có khả năng giảm méo do quá tải sườn và tạp âm hạt.



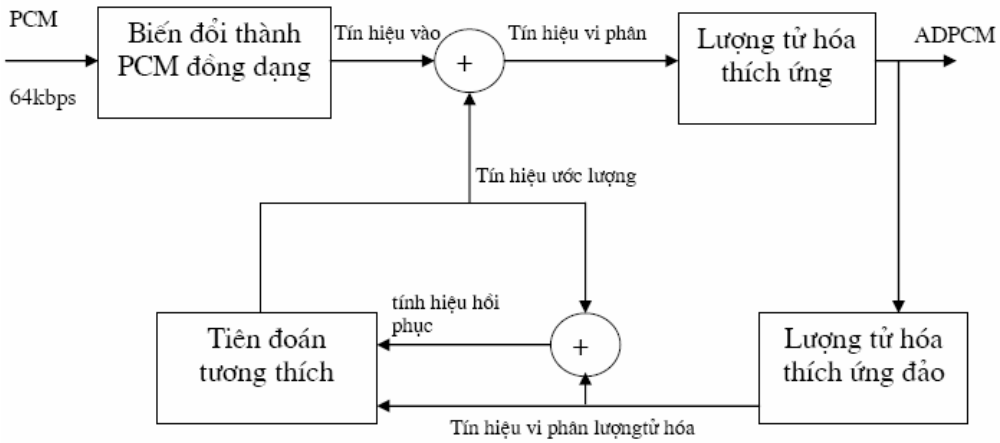
Hình 2.13 Dạng sóng tín hiệu trong ADM



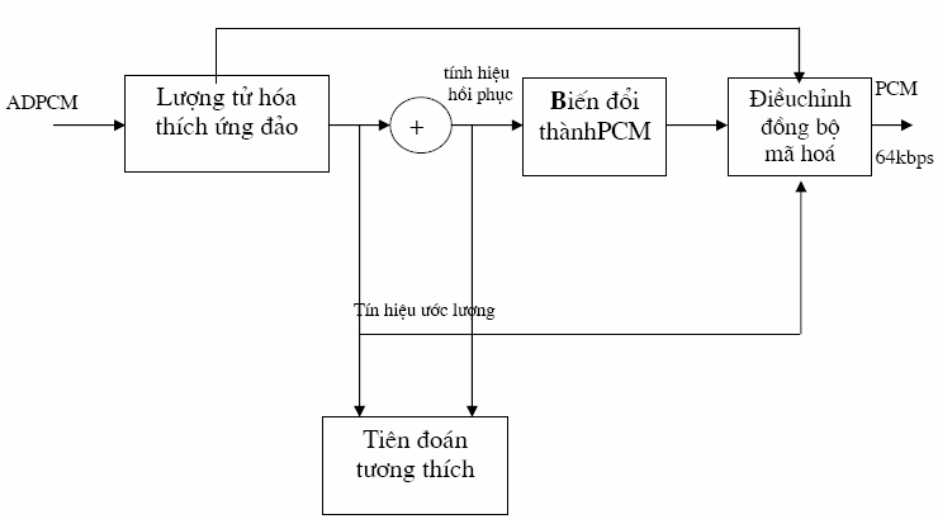
Hình 2.14 Sơ đồ mã hoá và giải mã ADM

2.8. Điều chế xung mã vi sai thích nghi (ADPCM)

Đây là phương pháp mã hoá khá quan trọng, tập hợp được những ưu điểm của các phương pháp trên và đã được ITU-T tiêu chuẩn hoá trong khuyến nghị G721, và đã có nhiều ứng dụng trong thực tế như hệ thống di động CT2 của Hàn Quốc, DECT của Mỹ. Vì vậy ta sẽ nghiên cứu sâu phương pháp. Các tốc độ được tiêu chuẩn là 40, 32, 24, 26 kbps. Phương pháp này dựa trên tính chất thay đổi chậm của phương sai và hàm tự tương quan, với phương pháp PCM ta dùng bộ lượng tử đều có công suất tạp âm là $\Delta_2/12$, phương pháp ADPCM và các phương pháp dự đoán tuyến tính nói chung là thay đổi Δ hay còn gọi là phương pháp dùng bộ lượng tử hoá tự thích nghi. Các thuật toán được phát triển cho hệ thống điều xung mã vi sai khi khi mã hoá tín hiệu tiếng nói bằng cách sử dụng bộ lượng tử hoá và bộ dự đoán thích nghi, có thông số thay đổi theo chu kỳ để phản ánh tính thông kê của tín hiệu tiếng nói.



Hình 2.15 Sơ đồ mã hoá ADPCM



Hình 2.16 Sơ đồ giải mã ADPCM

2.9. Bài thực hành các phương pháp biểu diễn số tín hiệu tiếng nói

Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) thực hiện các công việc sau:

Ghi âm một đoạn tín hiệu tiếng nói bất kỳ. Lưu tệp ở định dạng thô (*.wav).

Sử dụng Matlab hoặc các ngôn ngữ lập trình khác đọc và hiển thị tín hiệu theo dạng sóng ở miền thời gian.

Biểu diễn phổ của một phân đoạn tín hiệu với các dạng hàm cửa sổ khác nhau.

Sử dụng một trong các phương pháp biến đổi đã học trong chương này cho đoạn tín hiệu. Kết quả thu được được kiểm tra theo các tiêu chí: dung lượng tệp, chất lượng âm thanh cảm thụ,…

Chương 3: Phân tích tiếng nó

3.1. Mở đầu

Trong chương này chúng ta sẽ xem xét các phương pháp phân tích tín hiệu tiếng nói. Phân tích tiếng nói thực hiện giải quyết các vấn đề tìm ra một dạng thức tối ưu biểu diễn được tiếng nói một các hiệu quả. Nó là cơ sở cho việc phát triển các kỹ thuật, công nghệ tổng hợp, nhận dạng và nâng cao chất lượng tín hiệu tiếng nói. Phân tích tiếng nói thường thực hiện việc trích chọn hoặc chuyển đổi tín hiệu tiếng nói sang một dạng thức biểu diễn khác sao cho có thể biểu diễn thông tin tiếng nói tốt hơn theo cách mà chúng ta cần. Một cách tổng quát, hầu hết các phương pháp phân tích tín hiệu tiếng nói tập trung vào một trong ba vấn đề chính. Thứ nhất là tìm cách loại bỏ ảnh hưởng của pha, thành phần không đóng vai trong quan trọng trong việc truyền tải thông tin tiếng nói. Thứ hai, thực hiện việc chia tách nguồn âm và mạch lọc (mô hình tuyến âm) sao cho chúng ta có thể nghiên cứu biên phổ của tín hiệu một cách độc lập. Cuối cùng là chuyển đổi tín hiệu hoặc biên phổ tín hiệu sang một dạng biểu diễn khác hiệu quả hơn.

3.2. Mô hình phân tích tiếng nói

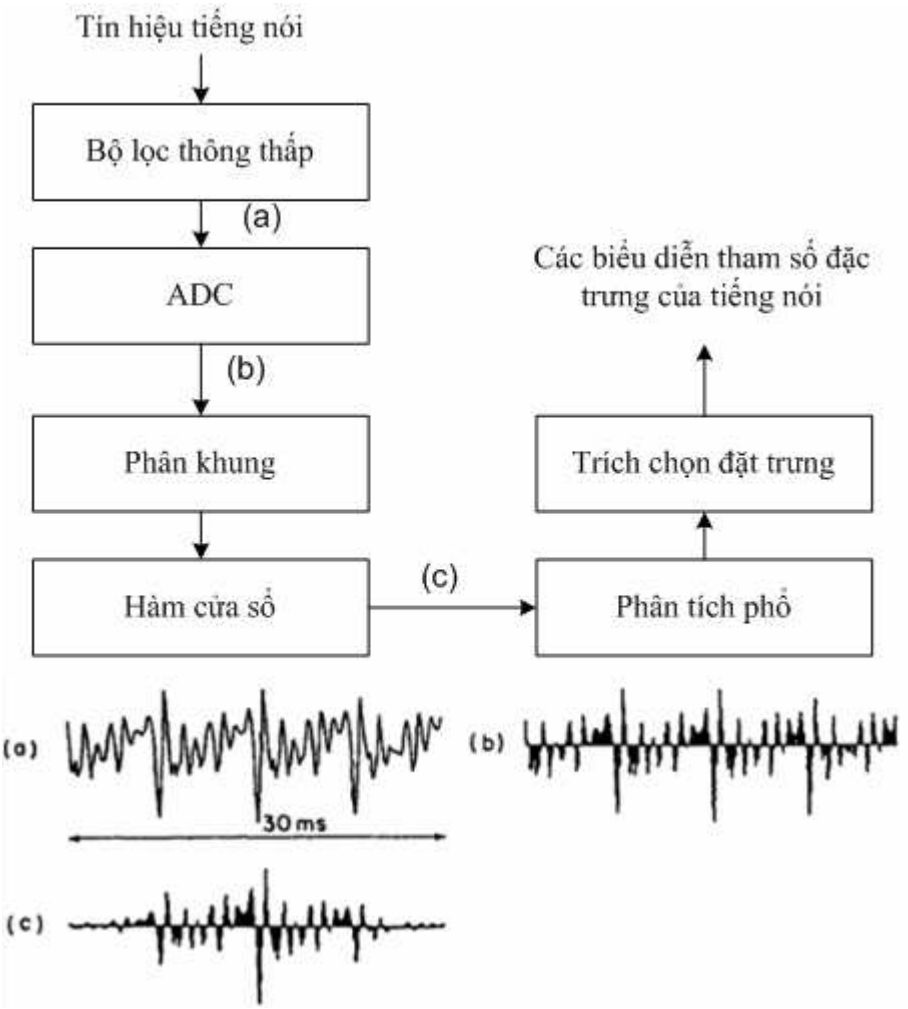
Mô hình tổng quát cho việc phân tích tiếng nói được trình bày trong hình 3.1. Các dạng tín hiệu tại các bước cũng được trình bày kèm theo trong minh họa.

Tín hiệu tiếng nói được tiền xử lý bằng cách cho qua một bộ lọc thông thấp với tần số cắt khoảng 8kHz. Tín hiệu thu được sau đó được thực hiện quá trình biến đổi sang dạng tín hiệu số nhờ bộ biến đổi ADC. Thông thường, tần số lấy mẫu bằng 16kHz với tốc độ bít lượng từ hóa là 16bit.

Tín hiệu tiếng nói dạng số được phân khung với chiều dài khung thường khoảng 30ms và khoảng lệch các khung thường bằng 10ms. Khung phân tích tín hiệu sau đó được chỉnh biên bằng cách lấy cửa sổ với các hàm cửa sổ phổ biến như Hamming, Hanning.... Tín hiệu thu được sau khi lấy cửa sổ được đưa vào phân tích với các phương pháp phân tích phổ (chẳng hạn như STFT, LPC,...). Hoặc sau khi phân tích phổ cơ bản, tiếp tục được đưa đến các khối để trích chọn các đặc trưng.

3.3. Phân tích tiếng nói ngắn hạn

Trong lý thuyết phân tích, chúng ta thường không để ý đến một điểm quan trọng là các phân tích phải được tiến hành trong một khoảng thời gian giới hạn. Chẳng hạn, chúng ta biết rằng biến đổi Fourier theo thời gian liên tục là một công cụ vô cùng hữu ích cho việc phân tích tín hiệu. Tuy nhiên, nó yêu cầu phải biết được tín hiệu trong mọi khoảng thời gian. Hơn nữa, các tính chất hay đặc trưng của tín hiệu mà chúng ta cần tìm hiểu phải là các đại lượng không đổi theo thời gian. Điều này trong thực tế phân tích tín hiệu khó mà đạt được vì việc phân tích tín hiệu đáp ứng các ứng dụng thực tế có thời gian hữu hạn. Hầu hết các tín hiệu, đặc biệt là tín hiệu tiếng nói, không phải là tín hiệu không đổi theo thời gian.



Hình 3.1 Mô hình tổng quát của việc xử lý tín hiệu tiếng nói

Về mặt nguyên lý, chúng ta có thể áp dụng các kỹ thuật phân tích đã biết vào phân tích tín hiệu trong ngắn hạn. Tuy nhiên vì tín hiệu tiếng nói là một quá trình mang thông tin động nên chúng ta không thể chỉ đơn thuần xem xét phân tích ngắn hạn trong chỉ một khung thời gian đơn lẻ.

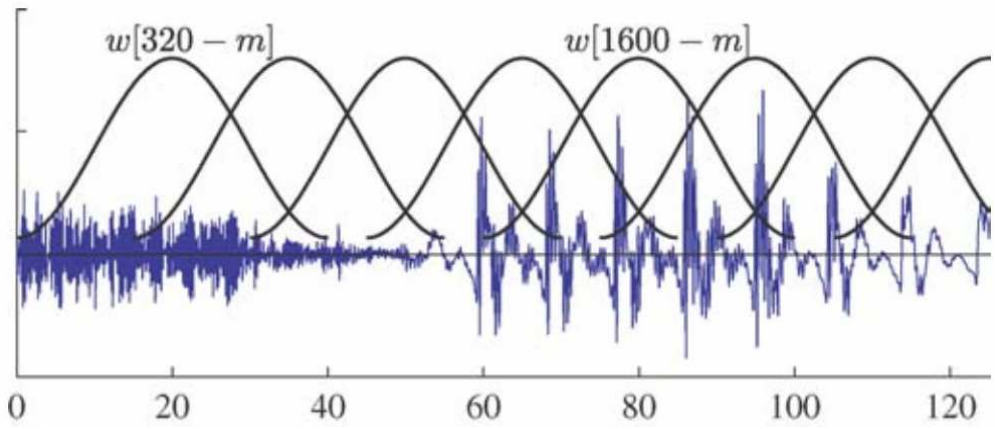
Tín hiệu tiếng nói như đã đề cập là tín hiệu thay đổi theo thời gian. Nó có các đặc trưng cơ bản như nguồn kích thích (excitation), cường độ (pitch), biên độ (amplitude), ... Các tham số thay đổi theo thời gian của tín hiệu tiếng nói có thể kể đến là tần số cơ bản (fundamental frequency - pitch), loại âm (âm hữu thanh - voiced, vô thanh - unvoiced, tắc - fricative hay khoảng lặng - silence), các tần số cộng hưởng chính (formant), hàm diện tích của tuyến âm (vocal tract area), ...

Việc thực hiện phân tích ngắn hạn tức là xem xét tín hiệu trong một khoảng nhỏ thời gian xung quanh thời điểm đang xét n nào đó. Các khoảng này thường khoảng từ 10-30ms. Điều này cho phép chúng ta giả thiết rằng trong khoảng thời gian đó các tính chất của dạng sóng tín hiệu tiếng nói là tương đối ổn định. Khoảng nhỏ tín hiệu dùng để phân tích thường được gọi là một khung (frame), hay một đoạn (segment). Một khung tín hiệu được xác định là tích của một hàm cửa sổ dịch w(m) và dãy tín hiệu s(n):

$$s_n(m) = s(m)\,w(n-m) \tag{3.1}$$

Một khung tín hiệu có thể được hiểu như một đoạn tín hiệu được cắt gọt bởi một hàm cửa sổ để tạo thành một dãy mới mà các giá trị của nó bằng không bên ngoài khoảng n∈[m-N+1,m]. Từ công thức (3.1) chúng ta thấy rằng khung tín hiệu này phụ thuộc vào khoảng thời gian kết thúc m. Trong khung tín hiệu nhỏ vừa được định nghĩa, dễ dàng thấy rằng các phép xử lý ngắn hạn cũng có ý nghĩa tương đương các phép xử lý dài hạn.

Như đã đề cập, việc phân tích tín hiệu tiếng nói không thể đơn giản chỉ bằng phân tích một khung tín hiệu đơn lẻ mà phải bằng các phân tích của các khung tín hiệu liên tiếp. Thực tế, để tránh mất thông tin, các khung tín hiệu thường được lấy bao trùm nhau. Nói một các khác, hai khung cạnh nhau có chung ít nhất M>0 mẫu. Hình 3.2 minh họa việc phân chia khung với hàm cửa sổ.



Hình 3.2 Phân tích tín hiệu trên các khung bao trùm nhau

Một phép phân tích ngắn hạn tổng quát có thể biểu diễn là:

$$X_n(m) = \sum_{m=-\infty}^{\infty} T\{s(m)\,w(n-m)\} \tag{3.2}$$

trong đó, X_n biểu diễn tham số phân tích (hoặc véc-tơ các tham số phân tích) tại thời điểm phân tích n. Toán tử T{} định nghĩa một hàm phân tích ngắn hạn. Tổng (3.2) được tính với giới hạn vô cùng được hiểu là phép lấy tổng được thực hiện với tất cả các thành phần khác không của khung tín hiệu là kết quả của phép lấy cửa sổ. Nói cách khác, tổng được thực hiện với mọi giá trị của m trong tập xác định (support) của hàm cửa sổ.

Một số hàm cửa sổ phổ biến thường hay được sử dụng là: hàm cửa sổ chữ nhật (rectangular window), hàm cửa sổ Hanning, và hàm cửa sổ Hamming.

3.4. Phân tích tiếng nói trong miền thời gian

Việc phân tích tiếng nói trong miền thời gian tức là phân tích trực tiếp trên dạng sóng tín hiệu sau khi thực hiện việc lấy cửa sổ trong miền thời gian. Như đã đề cập trong phần trước, chúng ta chỉ xem xét các phân tích ngắn hạn của tín hiệu. Vì vậy, để đơn giản trong trình bày chúng ta mặc định các công thức xây dựng là các phân tích ngắn hạn. Trong trường hợp nếu các phân tích không phải là ngắn hạn thì chúng sẽ được chú thích rõ ràng.

a) Năng lượng trung bình

Tham số đầu tiên chúng ta cần quan tâm trong phân tích tín hiệu tiếng nói trong miền thời gian đó là *năng lượng trung bình*. Năng lượng trung bình của tín hiệu tiếng nói được xác định như sau:

$$E_n = \sum_{m=-\infty}^{\infty} \left(s_n(m)\right)^2 = \sum_{m=-\infty}^{\infty} \left(s(m)\,\mathrm{w}(n-m)\right)^2 \tag{3.3}$$

Việc xác định năng lượng trung bình của tín hiệu rất hữu ích trong việc ước lượng các tính chất của các hàm kích thích trong mô hình mô phỏng bộ máy phát âm hay các mô hình tổng hợp tín hiệu tiếng nói. Ngoài ra, nó cung cấp cho chúng ta một công cụ hữu ích để phát hiện một tín hiệu âm là của âm hữu thanh, vô thanh hay một khoảng lặng. Điều này là bởi vì biên độ tín hiệu âm vô thanh thường rất nhỏ hơn so với biên độ tín hiệu âm hữu thanh.

Cần chú ý rằng độ dài cửa sổ phân tích phải được chọn thích hợp. Nó phải đủ dài để sự thay đổi của năng lượng tín hiệu trong một khung có thể được làm mịn. Tuy nhiên cũng không được quá dài dẫn đến luật thay đổi năng lượng tín hiệu từ một đoạn này sang một đoạn tín hiệu khác bị hiểu lầm.

Một nhược điểm của việc sử dụng năng lượng trung bình của tín hiệu là với các mức tín hiệu lớn, chúng có xu thế làm lệch một cách đáng kể giá trị ước lượng năng lượng toàn khung.

b) Độ lớn biên độ trung bình

Như đã đề cập trong phần trên, năng lượng trung bình tín hiệu khá nhạy cảm với độ lớn của tín hiệu. Do đó, người ta thường hay sử dụng một đại lượng thay thế là *độ lớn biên độ trung bình*, được xác định bởi:

$$M_n = \sum_{m=-\infty}^{\infty} |\,s(m)\,|\,\mathrm{w}(n-m) \tag{3.4}$$

c) Tốc độ trở về không

Một tham số khác cũng thường được quan tâm trong các phép phân tích tín hiệu tiếng nói trong miền thời gian đó là *tốc độ trở về không* (zero-crossing rate). Sự kiện trở về không xảy ra khi tín dạng sóng tín hiệu cắt trục hoành hay nói cách khác khi các mẫu liên tục nhau có dấu khác nhau. Về mặt toán học, tốc độ trở về không được xác định như sau:

$$Z_n = \sum_{m=-\infty}^{\infty} 0{,}5\left|\mathrm{sgn}\{\mathrm{s}(m)\} - \mathrm{sgn}\{\mathrm{s}(m-1)\}\right|\mathrm{w}(n-m) \tag{3.5}$$

Trong đó hàm sgn(a) là hàm dấu: bằng 1 nếu a≥0; bằng -1 nếu a<0. Dễ thấy 0,5|sgn{s(m)}-sgn{s(m-1)}| bằng 1 nếu s(m) và s(m-1) khác dấu nhau và bằng 0 nếu chúng cùng dấu. Điều này nghĩa là Z_n là tổng trọng số của tất cả các thay đổi dấu của các mẫu trong vùng xác định (support) của cửa sổ dịch w(n-m). Tốc độ trở về không có thể xem như là một đo lường của tần số. Mặc dù tốc độ trở về không thay đổi khá lớn theo thời gian và loại tín hiệu, nhưng nó biểu hiện sự khác biệt rõ rệt với tín hiệu âm vô thanh và hữu thanh. Các tín hiệu âm hữu thanh có sự suy giảm lớn ở vùng tần cao do đặc tính tự nhiên thông thấp của các xung dây thanh (glottal pulse), trong khi các tín hiệu âm vô thanh có năng lượng lớn ở vùng tần cao. Do vậy, cũng như đại lượng năng lượng trung bình tín hiệu, tốc độ trở về không cũng là các tham số quan trọng để phát hiện xem một tín hiệu là tín hiệu của âm vô thanh, hữu thanh hay khoảng lặng.

d) Hàm tự tương quan

Hàm tự tương quan thường được sử dụng như một công cụ để xác định tính chu kỳ của tín hiệu và nó cũng là cơ sở cho nhiều phương pháp phân tích phổ khác. Hàm tự tương quan được định nghĩa tương tự như hàm tự tương quan thông thường:

$$\begin{aligned}\Phi_n(k) &= \sum_{m=-\infty}^{\infty} s_n(m)s_n(m+k) \\ &= \sum_{m=-\infty}^{\infty} s(m)\mathrm{w}(n-m)s(m+k)\mathrm{w}(n-k-m) \\ &= \sum_{m=-\infty}^{\infty} s(m)s(n-m)\tilde{\mathrm{w}}_n(n-m)\end{aligned} \tag{3.6}$$

Trong công thức (3.6) chúng ta đã sử dụng tính chất của hàm tự tương quan là một hàm chẵn, đối xứng và $\tilde{\mathrm{w}}_k(m) = \mathrm{w}(m)\mathrm{w}(m+k)$.

Cũng tương tự như hàm tự tương quan tín hiệu chúng ta đã biết, có một mối quan hệ giữa hàm tự tương quan và năng lượng trung bình tín hiệu như sau:

$$E_n = \sum_{m=-\infty}^{\infty} \left(s(m)\mathrm{w}(n-m)\right)^2 = \Phi_n(0) \tag{3.7}$$

e) Hàm vi phân biên độ trung bình

Hàm vi phân biên độ trung bình được định nghĩa như sau:

$$\Delta M_n = \sum_{m=-\infty}^{\infty} |s(m)-s(m-\eta)|\,\mathrm{w}(n-m) \tag{3.8}$$

Công thức (3.8) cho thấy giá trị hàm vi phân biên độ trung bình, với tham số về sự khác nhau về thời gian η sẽ rất nhỏ khi η tiến đến chu kỳ (nếu có) của tín hiệu s(n). Do đó hàm vi phân biên độ trung bình là một trong các công cụ hữu ích cho việc xác định tần số cơ bản của tín hiệu tiếng nói.

3.5. Phân tích tiếng nói trong miền tần số

3.5.1 Cấu trúc phổ của tín hiệu tiếng nói

Trong phân tích tín hiệu tiếng nói, thay vì sử dụng trực tiếp tín hiệu tiếng nói trong miền thời gian, người ta thường hay sử dụng các đặc trưng phổ của tiếng nói. Điều này xuất phát từ quan điểm rằng tín hiệu tiếng nói cũng giống như các tín hiệu xác định khác có thể xem như là tổng của các tín hiệu hình sin với biên độ và pha thay đổi chậm. Hơn nữa, một nguyên nhân quan trọng không kém đó là việc cảm nhận tiếng nói của con người liên quan trực tiếp đến thông tin phổ của tín hiệu tiếng nói nhiều hơn trong khi các thông tin về pha của tín hiệu tiếng nói không có vai trò quyết định.

Phổ biên độ phức của tín hiệu tiếng nói được định nghĩa là biến đổi Fourier (FT) của khung tín hiệu với khoảng thời gian phân tích n cố định:

$$S_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} s(m)\mathrm{w}(n-m)e^{j\omega m} \tag{3.9}$$

Biểu thức (3.9) có thể viết lại như sau:

$$S_n\left(e^{j\omega}\right)=\left(s\left(\tilde{n}\right)e^{-j\omega\tilde{n}}\right)*\mathrm{w}\left(\tilde{n}\right)|_{\tilde{n}=n} \tag{3.10}$$

Biểu thức (3.10) được gọi là một cách diễn dịch phép biến đổi Fourier rời rạc theo khía cạnh mạch lọc. Tín hiệu điều biên $s(\tilde{n})e^{-j\omega \tilde{n}}$ dịch phổ của $s(\tilde{n})$ xuống ω lần và kết quả thu được sẽ được lựa chọn bởi một bộ lọc cửa sổ thông dải với tần số trung tâm bằng không.

Mặt khác công thức (3.9) cũng có thể viết là:

$$S_n\left(e^{j\omega}\right)=\left(s\left(\tilde{n}\right)*\left(\mathrm{w}\left(\tilde{n}\right)e^{j\omega\tilde{n}}\right)\right)*e^{-j\omega\tilde{n}}\,|_{\tilde{n}=n} \tag{3.11}$$

Công thức (3.11) có thể diễn giải như sau. Tín hiệu $s\left(\tilde{n}\right)$ được đưa qua bộ lọc thông dải có tần số trung tâm ω và đáp ứng xung $\mathrm{w}\left(\tilde{n}\right)e^{j\omega\tilde{n}}$. Kết quả thu được được dịch tần xuống bằng cách điều chế biên độ với $e^{j\omega\tilde{n}}$ để tạo ra tín hiệu băng tần thấp.

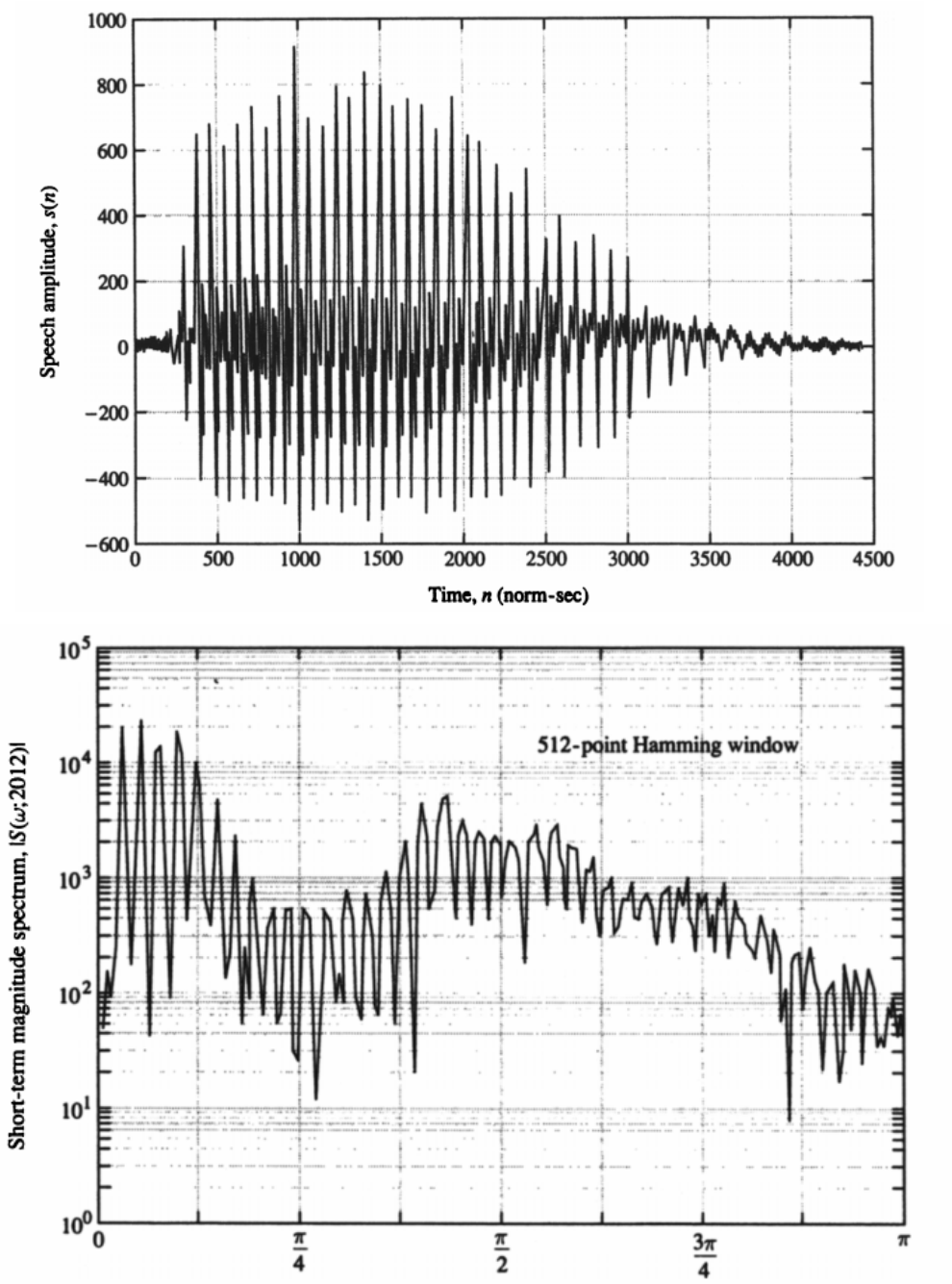
Hình 3.3 minh họa một khung tín hiệu và phổ tương ứng.

Mật độ phổ công suất trong một khoảng thời gian ngắn, tức là phổ ngắn hạn của tín hiệu tiếng nói, có thể được xem như là tích của hai thành phần: thành phần thứ nhất là đường biên phổ thay đổi một cách chậm chạp theo tần số; thành phần thứ hai là cấu trúc phổ mịn (spectral fine structure) thay đổi rất nhanh theo tần số. Đối với các âm hữu thanh thì cấu trúc phổ mịn tạo thành các mẫu tuần hoàn, còn đối với các âm vô thanh thì không. Biên phổ, hay cũng chính là đặc trưng phổ tổng quát (overall), mô tả không chỉ các đặc tính (characteristics) cộng hưởng và phản cộng hưởng (anti-resonance) của các cơ quan phát âm (articulatory organs) mà còn mô tả các đặc trưng tổng quát của phát xạ (radiation) và phổ nguồn glottal ở môi và khoang mũi. Trong khi đó, cấu trúc phổ mịn mô tả tính tuần hoàn của nguồn âm.

Công thức (3.9) là một hàm của tần số phân tích liên tục ω. Do đó để FT trở thành một công cụ hữu ích trong các phân tích thực tế chúng ta cần tính toán nó với tập tần số rời rạc và hàm cửa sổ có bề rộng hữu hạn với mỗi bước dịch chuyển R>1. Khi đó chúng ta có:

$$S_{rR}\left(k\right)=\sum_{m=rR-L+1}^{rR}s\left(m\right)\mathrm{w}\left(rR-m\right)e^{-j\frac{2\pi k}{N}m}\quad\left(k=0,1,...,N-1\right) \tag{3.12}$$

N là số các tần số cách đều nhau trong khoảng 0≤ ω ≤ 2π, L là độ dài hàm cửa sổ (đo lường bằng số mẫu). Vì chúng ta giả thiết hàm cửa sổ w(n) là hàm có tính nhân quả và có giá trị khác không chỉ trong khoảng 0≤ m ≤ L-1 do đó phần tín hiệu lấy qua cửa sổ s(m)w(rR-m) sẽ có giá trị khác không trên khoảng rR-L+1≤ m ≤ rR.



Hình 3.3 Khung tín hiệu và phổ tương ứng

3.5.2 Spectrogram

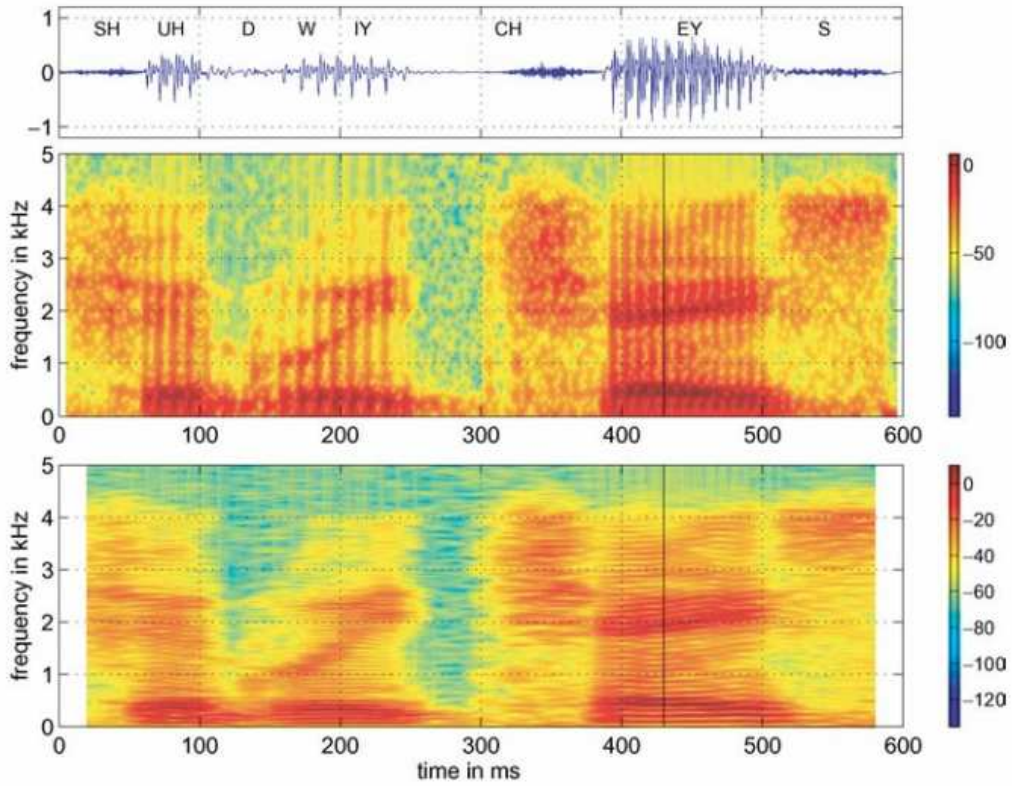
Spectrogram là một trong những công cụ cơ bản của phân tích phổ tín hiệu tiếng nói, trong đó nó chuyển đổi dạng sóng tín hiệu tiếng nói hai chiều thanh cấu trúc ba chiều (biên độ/tần số/thời gian). Trong đồ spectrogram, thời gian và tần số tương ứng là các trục ngang và dọc, còn biên độ được biểu diễn bởi độ đậm nhạt. Các đỉnh của phổ tín hiệu xuất hiện là các dải nằm ngang màu đậm. Tần số trung tâm của các dải thường được coi là các formant. Các âm hữu thanh tạo ra các mảng dọc trong biểu đồ spectrogram bởi vì có một sự tăng cường biên độ tín hiệu tiếng nói mỗi khi thanh quản đóng lại. Nhiễu trong các âm vô thanh tạo ra các cấu trúc đậm hình chữ nhật và kết thúc ngẫu nhiên với nhiều đốm nhạt do sự thay đổi tức thì của năng lượng tín hiệu. Lược đồ spectrogram chỉ diễn tả biên độ phổ của tín hiệu mà bỏ qua các

thông tin về pha bởi vì các thông tin về pha được cho rằng không có vai trò quan trọng trong hầu hết các ứng dụng liên quan đến tiếng nói.

Để xây dựng lược đồ spectrogram, người ta thực hiện việc biểu diễn biên độ của biến đổi Fourier ngắn hạn (STFT) $|S_n(e^{j\omega})|$ theo thời gian trên trục nằm ngang, đồng thời theo tần số ω (từ 0 đến π) trên trục thẳng đứng (tức là từ 0 đến $F_s/2$, với F_s là tần số lấy mẫu), đồng thời độ lớn biên độ bằng độ đậm nhạt (thường theo thang tỷ lệ lô-ga-rít)

$$\tilde{S}(t_r, f_k)_n = 20\log_{10}|S_{rR}(k)| \tag{3.13}$$

Trong đó t_r=rRT và f_k=k/(NT) và T là chu kỳ lấy mẫu của tín hiệu. Hình 3.4 minh họa spectrogram của tín hiệu tiếng nói cùng với dạng sóng tín hiệu tương ứng.



Hình 3.4 Lược đồ spectrogram của tín hiệu tiếng nói "Should we chase"

Hai lược đồ spectrogram được xây dựng với các hàm cửa sổ có độ dài khác nhau.Lược đồ spectrogram phía trên là kế quả khi sử dụng cửa sổ có chiều dài 101 mẫu tương ứng với 10ms. Chiều dài của cửa sổ phân tích này xấp xỉ bằng chu kỳ của dạng sóng trong các khoảng tín hiệu âm hữu thanh. Kết quả là trong các khoảng tín hiệu âm hữu thanh, spectrogram biểu hiện các vằn định hướng thẳng đứng tương ứng với thực tế rằng cửa sổ trượt lúc gom hầu hết các mẫu có biên độ lớn, lúc gom hầu hết các mẫu có biên độ nhỏ. Nói một cách khác, khi cửa sổ phân tích có độ dài ngắn, mỗi chu kỳ pitch riêng rẽ được hiển thị rõ nét theo thời gian, trong khi độ phân giải theo tần số thì rất kém. Cũng chính vì lý do này, nếu chiều dài cửa sổ phân tích mà ngắn, thì lược đồ spectrogram thu được gọi là lược đồ spectrogram băng rộng. Ngược lại, nếu chiều dài cửa sổ phân tích lớn, thì lược đồ spectrogram thu được gọi là lược đồ spectrogram băng hẹp. Lược đồ spectrogram băng hẹp có độ phân giải theo tần số cao nhưng theo thời gian thì nhỏ. Minh họa phía dưới của hình 3.4 là kết quả của việc sử dụng cửa sổ phân tích có độ dài 401 mẫu, tương ứng với 40ms, bằng khoảng vài chu kỳ tín hiệu. Và như

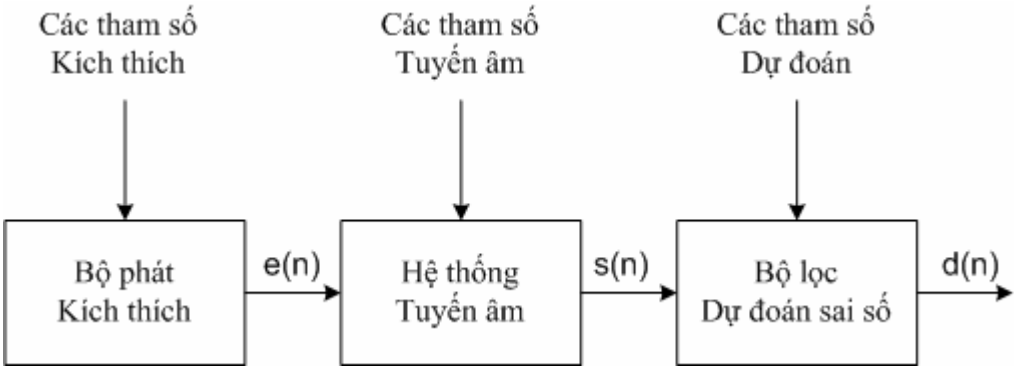
chúng ta thấy, lược đồ spectrogram tương ứng không còn nhạy với sự thay đổi về thời gian nữa.

3.6. Phương pháp phân tích mã hóa dự đoán tuyến tính (LPC)

Phương pháp phân tích dự đoán tuyến tính là một trong các phương pháp phân tích tín hiệu tiếng nói mạnh nhất và được sử dụng phổ biến. Điểm quan trọng của phương pháp này nằm ở khả năng nó có thể cung cấp các ước lượng chính xác của các tham số tín hiệu tiếng nói và khả năng thực hiện tính toán tương đối nhanh.

Mô hình của phương pháp phân tích tín hiệu tiếng nói dựa trên mã dự đoán tuyến tính (LPC- Linear Predictive Coding) được trình bày trong hình vẽ 3.5. Phương pháp phân tích LPC thực hiện việc phân tích phổ trên các khung (khối - block) tín hiệu hay còn gọi là các khung tín hiệu (speech frames) bằng việc sử dụng một mô hình hóa toàn điểm cực. Điều này có nghĩa là kết quả biểu diễn phổ thu được $X_n(e^{j\omega})$ được giới hạn trong dạng $\delta/A(e^{j\omega})$, trong đó $A(e^{j\omega})$ là một đa thức bậc p tương ứng khi thực hiện phép biến đổi z:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_p z^{-p} \tag{3.14}$$



Hình 3.5 Mô hình phân tích LPC cho tín hiệu tiếng nói

Bậc của đa thức, p, còn được gọi là bậc phân tích LPC. Kết quả thu được từ khối phân tích phổ LPC là một véc-tơ các hệ số (còn gọi là các tham số LPC) cụ thể hóa (specify) phổ của một mô hình toàn điểm cực mà phù hợp nhất với phổ tín hiệu gốc trên toàn khoảng thời gian xem xét các mẫu tín hiệu.

Ý tưởng đằng sau việc sử dụng mô hình LPC là ở việc có thể xấp xỉ một mẫu tín hiệu tiếng nói ở thời điểm n bất kỳ, $s(n)$, như là một tổ hợp tuyến tính của p mẫu trước đó. Nói cách khác:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p) \tag{3.15}$$

Các hệ số a_1, a_2, …, a_p được giả thiết là không đổi trong khung phân tích tín hiệu. Biểu thức (3.15) có thể được viết lại thành đẳng thức nếu ta thêm vào một thành phần kích thích (excitation term) Gu(n), ta được:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{3.16}$$

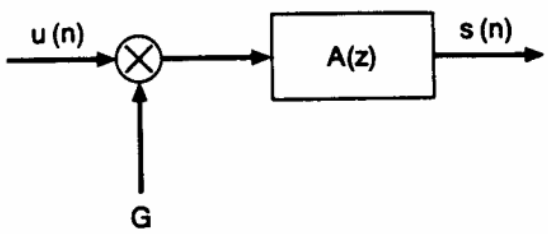
Trong công thức (3.16), u(n) là thành phần kích thích chuẩn và G là hệ số khuếch đại của thành phần kích thích. Nếu xem xét biểu thức (316) trong miền z chúng ta có biểu thức:

$$S(z)=\sum_{i=1}^{p} a_i z^{-i} S(z)+GU(z) \tag{3.17}$$

Hay hàm truyền đạt tương ứng là:

$$H(z)=\frac{S(z)}{GU(z)}=\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}}=\frac{1}{A(z)} \tag{3.18}$$

Hàm truyền đạt (3.18) có thể được thực hiện bởi sơ đồ khối trong hình 3.6. Sơ đồ khối đó có thể được giải thích như sau. Nguồn kích thích chuẩn hóa u(n) được nhân với hệ số khuếch đại G trở thành đầu vào của một hệ thống toàn điểm cực H(z)=1/A(z) để tạo ra tín hiệu tiếng nói s(n). Chúng ta biết rằng hàm kích thích thực của tín hiệu tiếng nói là dãy xung bán tuần hoàn đối với tín hiệu âm hữu thanh và là nguồn nhiễu ngẫu nhiên đối với tín hiệu âm vô thanh. Từ thực tế này, dễ dàng xây dựng được mạch tổng hợp tín hiệu tiếng nói dựa vào mô hình phân tích LPC như trong hình 3.7. Trong sơ đồ tổng hợp tiếng nói sử dụng mô hình phân tích LPC, nguồn kích thích được chọn tương ứng phù hợp với tín hiệu âm hữu thanh hay vô thanh nhờ một chuyển mạch. Hệ số khuếch đại G của tín hiệu được ước lượng từ tín hiệu tiếng nói. Mạch lọc số H(z) được điều khiển bởi các tham số của bộ máy phát âm tương ứng với tín hiệu tiếng nói được tạo ra. Nói một cách cụ thể, các tham số của mô hình tổng hợp này là các phân loại (classification) âm hữu thanh hay vô thanh, khoảng chu kỳ pitch (pitch period) của tín hiệu, tham số độ khuếch đại, các hệ số của bộ lọc a_k. Tất cả các tham số này thay đổi chậm theo thời gian.



Hình 3.6 Mô hình dự đoán mô phỏng tiếng nói

Giả sử rằng tổ hợp tuyến tính của các mẫu trước thời điểm xem xét là một ước lượng của tín hiệu, kí hiệu là $\tilde{s}(n)$:

$$\tilde{s}(n)=\sum_{k=1}^{p} a_k s(n-k) \tag{3.19}$$

Khi đó, sai số dự tính e(n) sẽ được tính là:

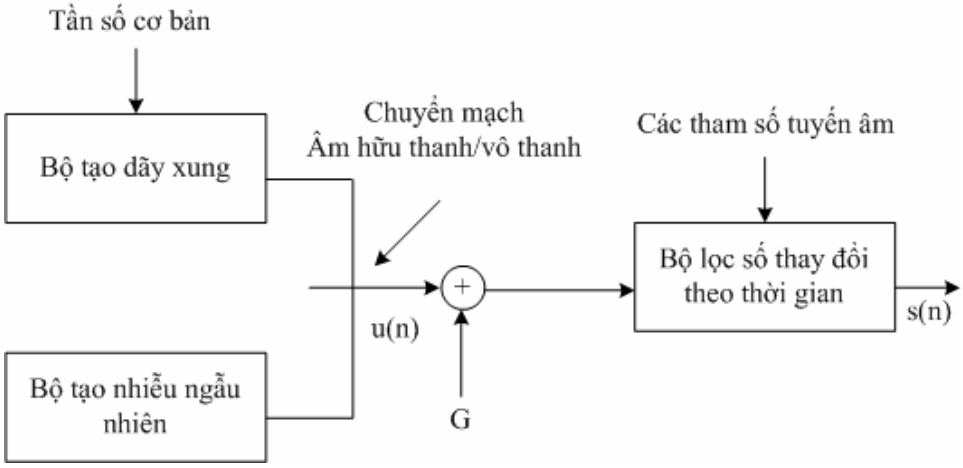
$$e(n)=s(n)-\tilde{s}(n)=s(n)-\sum_{k=1}^{p} a_k s(n-k) \tag{3.20}$$

Hay nói cách khác, hàm truyền đạt sai số tương ứng là:

$$A(z)=\frac{E(z)}{S(z)}=1-\sum_{k=1}^{p} a_k z^{-k} \tag{3.21}$$

Từ đây ta thấy rằng, nếu tín hiệu tiếng nói được tạo ra từ sơ đồ mạch 3.6 thì sai số dự đoán e(n) sẽ bằng tín hiệu kích thích Gu(n).

Vấn đề đặt ra đối với phương pháp phân tích LPC là xác định được tập các hệ số a_k một cách trực tiếp từ tín hiệu tiếng nói sao cho tính chất phổ của mạch lọc trong sơ đồ 3.7 tương đồng với phổ của tín hiệu tiếng nói trong khoảng cửa sổ phân tích. Vì đặc tính phổ của tín hiệu tiếng nói luôn thay đổi theo thời gian, các hệ số dự đoán ở thời điểm n xác định phải là những giá trị được ước lượng từ các đoạn ngắn hạn của tín hiệu tiếng nói xung quanh thời điểm n. Từ đây chúng ta thấy phương pháp tiếp cận cơ bản là tìm được một tập các hệ số dự đoán (predictor coefficients) sao cho chúng làm tối thiểu hóa sai số dự đoán trung bình bình phương trên toàn đoạn ngắn hạn của tín hiệu phân tích. Thường thì phương pháp phân tích phổ theo cách này được thực hiện trên các khung tín hiệu liên tiếp mà khoảng cách giữa các khung vào khoảng bậc của 10ms.



Hình 3.7 Mô hình tổng hợp tiếng nói dùng LPC

Để xây dựng biểu thức và từ đó tìm ra được các hệ số dự đoán thích hợp, chúng ta định nghĩa các khung tín hiệu ngắn hạn và tương ứng là các sai số ngắn hạn:

$$s_n(m) = s(n+m) \tag{3.22}$$

$$e_n(n) = e(n+m) \tag{3.23}$$

Chúng ta cần tối thiểu hóa tín hiệu sai số trung bình bình phương ở thời điểm n:

$$\varepsilon_n = \sum_m e_n^2(m) \tag{3.24}$$

Biểu thức (3.24) có thể được viết lại bằng cách sử dụng các định nghĩa $e_n(m)$ và $s_n(m)$ như sau:

$$\varepsilon_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2 \tag{3.25}$$

Để tìm cực tiểu của (3.25), chúng ta lấy đạo hàm lần lượt theo các hệ số a_k và cho chúng bằng không:

$$\frac{\partial \varepsilon_n}{\partial a_k} = 0 \ \ (k = 1, 2, ..., p) \tag{3.26}$$

Khi đó chúng ta có:

$$\sum_m s_n(m-i)\, s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m-i)\, s_n(m-k) \tag{3.27}$$

Chúng ta biết rằng hệ số có dạng $\sum s_n(m-i)\, s_n(m-k)$ là các thành phần của covariance ngắn hạn của $s_n(m)$. Nói cách khác:

$$\Psi_n(i,k) = \sum_m s_n(m-i)\, s_n(m-k) \tag{3.28}$$

Chúng ta có thể thu gọn biểu thức (3.27) như sau:

$$\Psi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \Psi_n(i,k) \tag{3.29}$$

Biểu thức (3.29) biểu diễn hệ thống gồm p biểu thức của p biến số. Dễ có giá trị sai số trung bình bình phương tối thiểu, $\hat{\varepsilon}_n$ được tính như sau:

$$\begin{aligned} \hat{\varepsilon}_n &= \sum_m s_n^2(m) - \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m)\, s_n(m-k) \\ &= \Psi_n(0,0) - \sum_{k=1}^{p} \hat{a}_k \Psi_n(0,k) \end{aligned} \tag{3.30}$$

Chúng ta thấy rằng, giá trị sai số trung bình bình phương tối thiểu có chứa một thành phần cố định $\Psi_n(0,0)$ và các thành phần khác phụ thuộc vào các hệ số dự đoán.

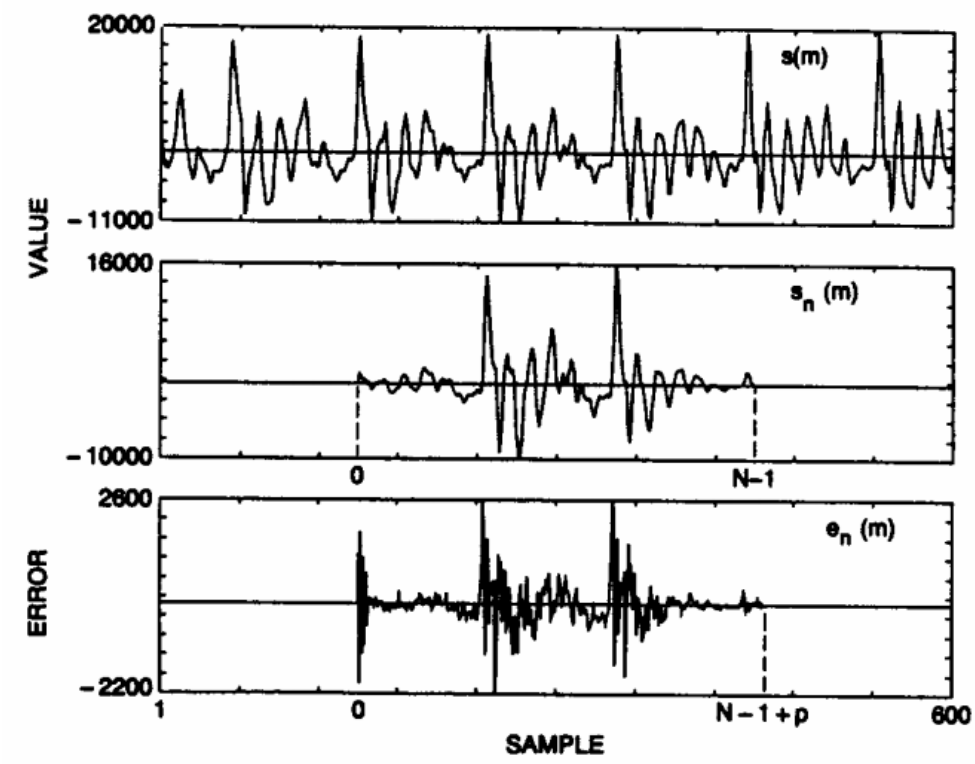
Để tìm các hệ số dự đoán tối ưu $\hat{a}_k$ trước hết chúng ta phải tính $\Psi_n(i,k)$ ($1 \le i \le p$ và $0 \le k \le p$) và sau đó giải hệ (3.29) đồng thời của p biểu thức. Trong thực tế, việc giải hệ và tính toán các thành phần Ψ phụ thuộc rất nhiều vào khoảng thời gian m được sử dụng để định ra khung tín hiệu phân tích và vùng mà trên đó sai số trung bình bình phương được ước lượng. Có hai phương pháp chuẩn để định ra khoảng thích hợp cho tín hiệu tiếng nói: phương pháp sử dụng sự tự tương quan; và phương pháp sử dụng covariance.

Phương pháp sử dụng hàm tự tương quan xuất phát trực tiếp từ việc định ra khoảng giới hạn m trong tổ hợp tuyến tính sao cho đoạn tín hiệu tiếng nói $s_n(m)$ bằng 0 ở ngoài khoảng $0 \le m \le N-1$. Điều này tương đương với việc giả thiết tín hiệu tiếng nói $s(n+m)$ được nhân với hàm của sổ $w(m)$ hữu hạn có giá trị bằng 0 ở ngoài khoảng $0 \le m \le N-1$. Nói một cách khác, mẫu tín hiệu tiếng nói để làm tối thiểu hóa sai số trung bình bình phương có thể biểu diễn dưới dạng:

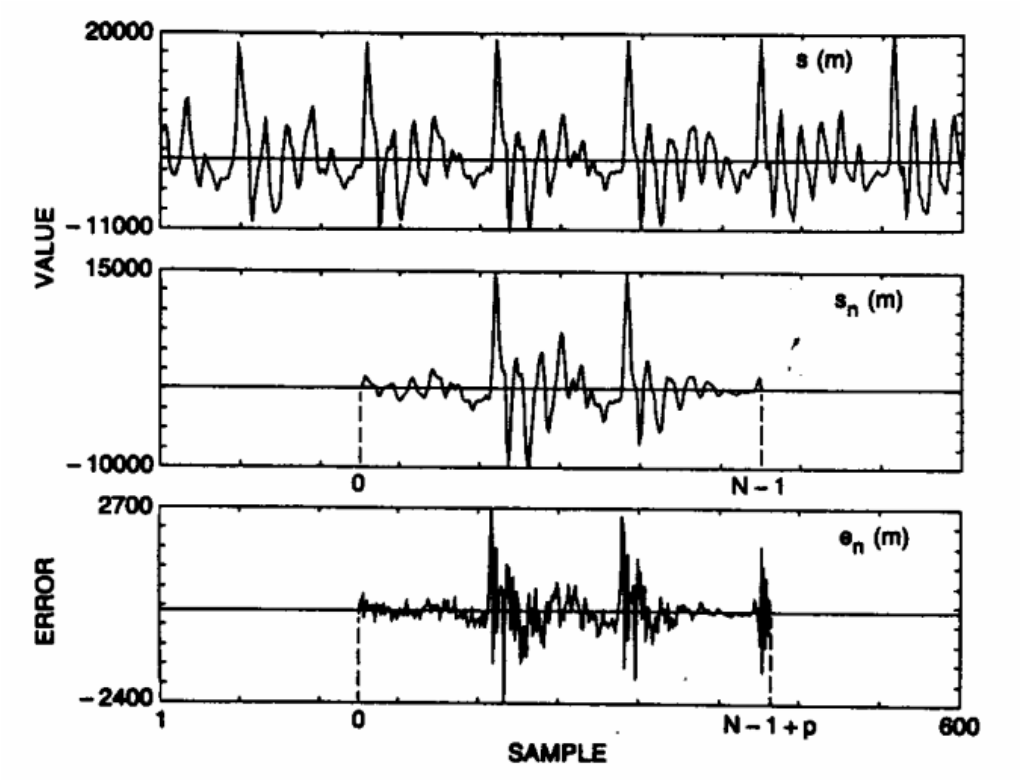
$$s_n(m) = \begin{cases} s(n+m)\, w(m) & 0 \le m \le N-1 \\ 0 & m \notin [0, N-1] \end{cases} \tag{3.31}$$

Từ công thức (3.31), khi $m<0$ tín hiệu sai số $e_n(m)$ bằng 0 vì khi đó $s_n(m)=0$. Mặt khác, cũng tương tự khi $m>N-1+p$ sẽ không có sai số dự đoán bởi vì khi đó ta cũng có $s_n(m)=0$. Tuy nhiên trong vùng $m=0$ (tức là từ $m=0$ đến $m=p-1$) tín hiệu thu được sau khi thực hiện việc lấy cửa sổ có thể được dự đoán từ các mẫu trước đó, mà một số trong chúng có thể bằng 0. Và

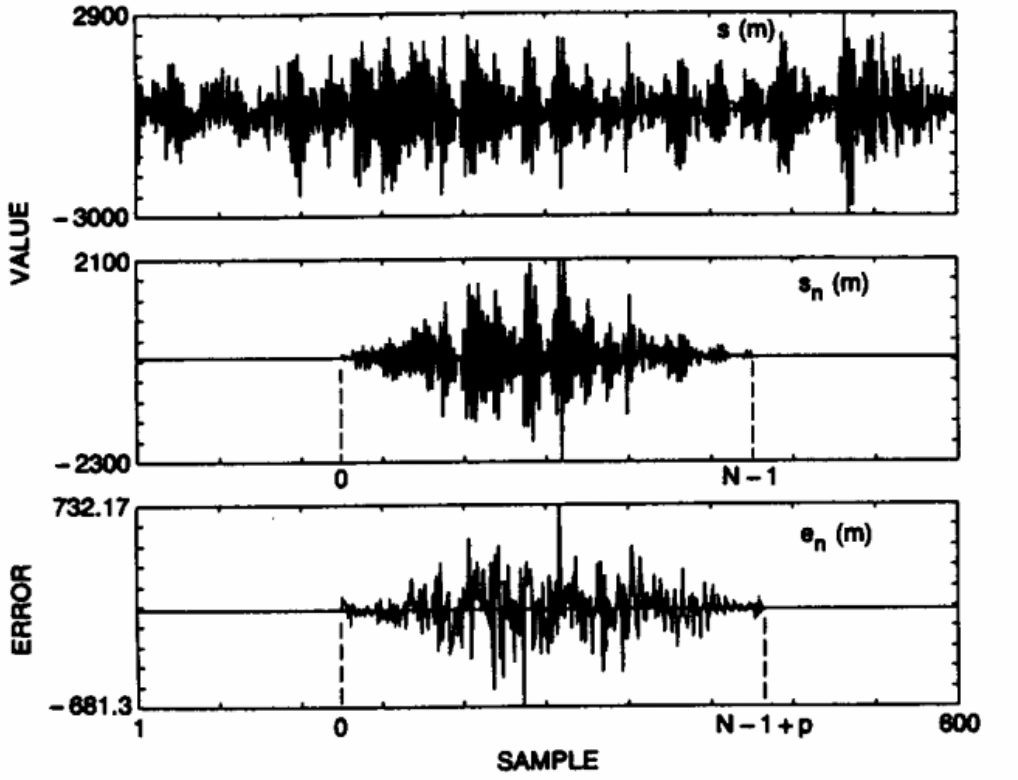
như vậy, khả năng sai số dự đoán tương đối lớn có thể tồn tại trong vùng này. Tại vùng m=N-1 (tức là từ m=N-1 đến m=N-1+p) khả năng có thể tồn tại sai số dự đoán lớn cũng có thể tồn tại bởi vì các tín hiệu thu được từ quá trình lấy của sổ bằng 0 được dự đoán từ một vài mẫu cuối cùng khác không của tín hiêu. Với tín hiệu âm hữu thanh,các hiệu ứng tiềm năng tồn tại sai số dự đoán lớn ở đầu hoặc cuối khung tín hiệu thể hiện rõ ràng khi bắt đầu chu kỳ của pitch hoặc rất gần với các điểm m=0 hoặc m=N-1. Đối với tín hiệu âm vô thanh thì hiện tượng này gần như được loại bỏ bởi vì không có phần tín hiệu nào nhạy cảm (position sensitive). Các hiện tượng này cùng với tín hiệu cửa sổ được minh họa trong các hình 3.8-3.10.



Hình 3.8 Minh họa trường hợp sai số dự đoán lớn ở đầu khung với tín hiệu âm hữu thanh



Hình 3.9 Minh họa trường hợp sai số dự đoán lớn ở cuối khung với tín hiệu âm hữu thanh



Hình 3.10 Minh họa trường hợp sai số dự đoan lớn với tín hiệu âm vô thanh

Mục đích của việc lấy của sổ là nhằm chỉnh (taper) tín hiệu ở gần các điểm m=0 và m=N-1 để làm tối thiểu hóa các sai số ở các vùng biên này.

Với việc định nghĩa khoảng tín hiệu sau phép lấy qua cửa sổ, chúng ta có thể viết biểu thức tính sai số trung bình bình phương như sau:

$$\varepsilon_n = \sum_{m=0}^{N-1+p} e_n^2(n) \tag{3.32}$$

Khi đó Ψ_n (i,k) có thể được viết lại là:

$$\Psi_n(i,k) = \sum_{m=0}^{N-1+p} s_n(m-i)\, s_n(m-k) \quad (1 \le i \le p, 0 \le k \le p) \tag{3.33}$$

Bằng cách thay chỉ số biểu thức trên có thể được viết dưới dạng:

$$\Psi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)\, s_n(m+i-k) \quad (1 \le i \le p, 0 \le k \le p) \tag{3.34}$$

Ta thấy biểu thức (3.34) là một hàm chỉ phụ thuộc vào hiệu i-k chứ không phải phụ thuộc hai biến số độc lập i và k. Do đó, hàm covariance Ψ_n(i,k) trở thành hàm tự tương quan:

$$\begin{aligned} \Psi_n(i,k) &= \Phi_n(i-k) \\ &= \sum_{m=0}^{N-1-(i-k)} s_n(m)\, s_n(m+i-k) \quad (1 \le i \le p, 0 \le k \le p) \end{aligned} \tag{3.35}$$

Do hàm tự tương quan là hàm đối xứng, tức là $\Phi_n(-k) = \Phi_n(k)$, biểu thức tương ứng của LPC có thể được biểu diễn là:

$$\sum_{k=1}^{p} \Phi_n(|i-k|)\, \hat{a}_k = \Phi_n(i) \quad (1 \le i \le p) \tag{3.36}$$

Nếu biểu diễn dưới dạng ma trận chúng ta có:

$$\begin{bmatrix} \Phi_n(0) & \Phi_n(1) & \Phi_n(2) & \cdots & \Phi_n(p-1) \\ \Phi_n(1) & \Phi_n(0) & \Phi_n(1) & \cdots & \Phi_n(p-2) \\ \Phi_n(2) & \Phi_n(1) & \Phi_n(0) & \cdots & \Phi_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_n(p-1) & \Phi_n(p-2) & \Phi_n(p-3) & \cdots & \Phi_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \Phi_n(1) \\ \Phi_n(2) \\ \Phi_n(3) \\ \vdots \\ \Phi_n(p) \end{bmatrix} \tag{3.37}$$

Trong công thức trên, ma trận các thành phần tự tương quan là một ma trận Toeplitz (ma trận đối xứng với các thành phần đường chéo chính bằng nhau), do đó việc giải hệ phương trình trên dễ dàng thực hiện được bằng việc áp dụng các thuật toán tính toán hiệu quả đã biết.

Phương pháp sử dụng covariance là một phương pháp khác với phương pháp sử dụng hàm tự tương quan đã đề cập ở trên. Phương pháp này cố định khoảng mà trên đó sai số trung bình bình phương được tính trong khoảng 0≤ m ≤N-1 và sử dụng khung tín hiệu trong khoảng đó một cách trực tiếp mà không thực hiện phép lấy của sổ.

Sai số trung bình bình phương khi đó được tính là:

$$\varepsilon_n = \sum_{m=0}^{N-1} e_n^2(m) \tag{3.38}$$

Và covariance được tính bởi:

$$\Psi_n(i,k) = \sum_{m=0}^{N-1} s_n(m-i)\, s_n(m-k) \quad (1 \le i \le p,\ 0 \le k \le p) \tag{3.39}$$

Hoặc bằng cách đổi chỉ số:

$$\Psi_n(i,k) = \sum_{m=0}^{N-i-1} s_n(m)\, s_n(m+i-k) \quad (1 \le i \le p, 0 \le k \le p) \tag{3.40}$$

Để ý thấy rằng việc tính toán theo biểu thức (3.40) liên quan đến các mẫu tín hiệu $s_n(m)$ từ thời điểm m=-p đến m=N-1-p khi i=p, và liên quan đến các mẫu $s_n(m+i-k)$ từ thời điểm 0 đến thời điểm N-1. Do đó, khoảng tín hiệu cần thiết để có thể tính toán hoàn thiện là từ $s_n(-p)$ đến $s_n(N-1)$. Nói một cách khác, việc tính toàn cần đến các mẫu bên ngoài khoảng tối thiểu sai số gồm $s_n(-p)$, $s_n(-p+1)$, …, $s_n(-1)$.

Bằng việc sử dụng khoảng tín hiệu mở rộng để tính toán các giá trị covariance $\Psi_n(i,k)$, biểu thức phân tích LPC dạng ma trận được biểu diễn như sau:

$$\begin{bmatrix} \Psi_n(1,1) & \Psi_n(1,2) & \Psi_n(1,3) & \cdots & \Psi_n(1,p) \\ \Psi_n(2,1) & \Psi_n(2,2) & \Psi_n(2,3) & \cdots & \Psi_n(2,p) \\ \Psi_n(3,1) & \Psi_n(3,2) & \Psi_n(3,3) & \cdots & \Psi_n(3,4) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Psi_n(p,1) & \Psi_n(p,2) & \Psi_n(p,3) & \cdots & \Psi_n(p,p) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \Psi_n(1,0) \\ \Psi_n(2,0) \\ \Psi_n(3,0) \\ \vdots \\ \Psi_n(p,0) \end{bmatrix} \tag{3.41}$$

Ma trận các hệ số covariance là một ma trận đối xứng (vì $\Psi_n(i,k)= \Psi_n(k,i)$) tuy nhiên không phải ma trận Toeplitz. Việc giải hệ phương trình trên có thể thực hiện bằng việc sử dụng thuật toán phân tích Cholesky. Trong thực tế, mô hình phân tích LPC biểu diễn dạng covariance đầy đủ thường không được sử dụng trong các hệ thống nhận dạng tín hiệu tiếng nói.

3.7. Phương pháp phân tích cepstral

Khái niệm cepstrum được đưa ra bởi Bogert, Healy và Tukey. Cepstrum được định nghĩa là biến Fourier ngược (IFT) của lô-ga-rít độ lớn biên độ phổ của tín hiệu. Nói các khác, cepstrum của một tín hiệu với thời gian rời rạc được cho bởi công thức:

$$c_n(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left|S_n\left(e^{j\omega}\right)\right| e^{j\omega} d\omega \tag{3.42}$$

Ở đây, $\log|S_n(e^{j\omega})|$ là lô-ga-rít của độ lớn biên độ (magnitude) của FT tín hiệu. Khái niệm (3.42) có thể được mở rộng thành cepstrum phức như sau:

$$\hat{c}_n(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\{S_n\left(e^{j\omega}\right)\} e^{j\omega m} d\omega \tag{3.43}$$

Trong công thức (3.43), $\log\{S_n(e^{j\omega})\}$ là lô-ga-rít phức của $S_n(e^{j\omega})$ và được định nghĩa như sau:

$$\hat{S}_n\left(e^{j\omega}\right) = \log\{S_n\left(e^{j\omega}\right)\} = \log\left|S_n\left(e^{j\omega}\right)\right| + j \arg\left[S_n\left(e^{j\omega}\right)\right] \tag{3.44}$$

Giả sử $s(n)=s_1(n)*s_2(n)$, với định nghĩa cepstrum dễ dàng thấy rằng $\hat{c}(n) = \hat{c}_1(n) + \hat{c}_2(n)$. Như vậy phép toán với cepstrum đã chuyển tích chập thành phép cộng. Chính điều này đã làm cho phép phân tích cepstrum trở thành một công cụ hữu ích cho việc phân tích tín hiệu tiếng nói.

Tuy nhiên các công thức (3.42)-(3.44) là các định nghĩa dựa trên các công thức toán học. Để công thức có ý nghĩa trong các phân tích thực tế, chúng ta phải xây dựng các công thức mà

việc tính toán có thể dễ dàng thực hiện được. Vì biến đổi Fourier rời rạc (DFT) là phiên bản lấy mẫu của biến đổi Fourier với thời gian rời rạc (DTFT) của một dãy chiều dài cố định (tức là $S(k)=S(e^{j2\pi k/N})$), do đó IDFT và DFT có thể được thay thế tương ứng bằng IDTFT và DTFT.

$$S(k)=\sum_{n=0}^{N-1}s(n)e^{-j2\pi kn/N} \tag{3.45}$$

$$\hat{X}(k)=\log\left|S(k)\right|+j\arg\left[S(k)\right] \tag{3.46}$$

$$\tilde{s}(n)=\frac{1}{N}\sum_{n=0}^{N-1}\hat{X}(k)e^{j2\pi kn/N} \tag{3.47}$$

3.8. Một số phương pháp xác định tần số Formant

Formant của tín hiệu tiếng nói là một trong các tham số quan trọng và hữu ích có ứng dụng rộng rãi trong nhiều lĩnh vực chẳng hạn như trong việc xử lý, tổng hợp và nhận dạng tiếng nói. Các formant là các tần số cộng hưởng của tuyến âm (vocal tract), nó thường được thể hiện trong các biểu diễn phổ chẳng hạn như trong biểu diễn spectrogram như là một vùng có năng lượng cao, và chúng biến đổi chậm theo thời gian theo hoạt động của bộ máy phát âm. Sở dĩ formant có vai trò quan trọng và là một tham số hữu ích trong các nghiên cứu xử lý tiếng nói là vì các formant có thể miêu tả được các khía cạnh quan trọng nhất của tiếng nói bằng việc sử dụng một tập rất hạn chế các đặc trưng. Chẳng hạn trong mã hóa tiếng nói, nếu sử dụng các tham số formant để biểu diễn cấu hình của bộ máy phát âm và một vài tham số phụ trợ biểu diễn nguồn kích thích, chúng ta có thể đạt được tốc độ mã hóa thấp đến 2,4kbps.

Nhiều nghiên cứu về xử lý và nhận dạng tiếng nói đã chỉ ra rằng các tham số formant là ứng cử viên tốt nhất cho việc biểu diễn phổ của bộ máy phát âm một cách hiệu quả. Tuy nhiên việc xác định các formant không đơn giản chỉ là việc xác định các đỉnh trong phổ biên độ bởi vì các đỉnh phổ của tín hiệu ra của bộ máy phát âm phụ thuộc một cách phức tạp vào nhiều yếu chẳng hạn như cấu hình bộ máy phát âm, các nguồn kích thích, ...

Các phương pháp xác định formant liên quan đến việc tìm kiếm các đỉnh trong các biểu diễn phổ, thường là từ kết quả phân tích phổ theo phương pháp STFT hoặc mã hóa dự đoán tuyến tính (LPC).

a) Xác định formant từ phân tích STFT

Các phân tích STFT tương tự và rời rạc đã trở thành một công cụ cơ bản cho nhiều phát triển trong phân tích và tổng hợp tín hiệu tiếng nói.

Dễ dàng thấy STFT trực tiếp chứa các thông tin về formant ngay trong biên độ phổ. Do đó, nó trở thành một cơ sở cho việc phân tích các tần số formant của tín hiệu tiếng nói.

b) Xác định formant từ phân tích LPC

Các tần số formant có thể được ước lượng từ các tham số dự đoán theo một trong hai cách. Cách thứ nhất là xác định trực tiếp bằng cách phân tích nhân tử đa thức dự đoán và dựa trên các nghiệm thu được để quyết định xem nghiệm nào tương ứng với formant. Cách thứ hai là sử dụng phân tích phổ và chọn các formant tương ứng với các đỉnh nhọn bằng một trong các thuật toán chọn đỉnh đã biết.

Một lợi điểm khi sử dụng phương pháp phân tích LPC để phân tích formant là tần số trung tâm của các formant và băng tần của chúng có thể xác định được một cách chính xác thông qua việc phân tích nhân tử đa thức dự đoán. Một phép phân tích LPC bậc p được chọn

trước, thì số khả năng lớn nhất có thể có các điểm cực liên hợp phức là p/2. Do đó, việc gán nhãn trong quá trình xác định xem điểm cực nào tương ứng với các formant đơn giản hơn các phương pháp khác. Ngoài ra, với các điểm cực bên ngoài thường có thể dễ dàng phân tách trong phân tích LPC vì băng tần của chúng thường rất lớn so với băng tần thông thường của các formant tín hiệu tiếng nói.

3.9. Một số phương pháp xác định tần số cơ bản

Tần số cơ bản F_0 là tần số giao động của dây thanh. Tần số này phụ thuộc vào giới tính và độ tuổi. F_0 của nữ thường cao hơn của nam, F_0 của người trẻ thường cao hơn của người già. Thường với giọng của nam, F_0 nằm trong khoảng từ 80-250Hz, với giọng của nữ, F_0 trong khoảng 150-500Hz. Sự biến đổi của F_0 có tính quyết định đến thanh điệu của từ cũng như ngữ điệu của câu. Câu hỏi đặt ra là làm thế nào để xác định tần cố cơ bản (fundamental frequency). Một số phương pháp xác định tần số cơ bản có thể kể đến là: Phương pháp sử dụng hàm tự tương quan, phương pháp sử dụng hàm vi sai biên độ trung bình; Phương pháp sử dụng bộ lọc đảo và hàm tự tương quan; Phương pháp xử lý đồng hình (homomophic).

a) Sử dụng hàm tự tương quan

Hàm tự tương quan $\Phi_n(k)$ sẽ đạt các giá trị cực khi tương ứng tại các điểm là bội của chu kỳ cơ bản của tín hiệu. Khi đó các tần số cơ bản là tần số xuất hiện của các đỉnh của $\Phi_n(t)$. Bài toán trở thành bài toán xác định chu kỳ hàm tự tương quan.

b) Sử dụng hàm vi sai biên độ trung bình (AMDF)

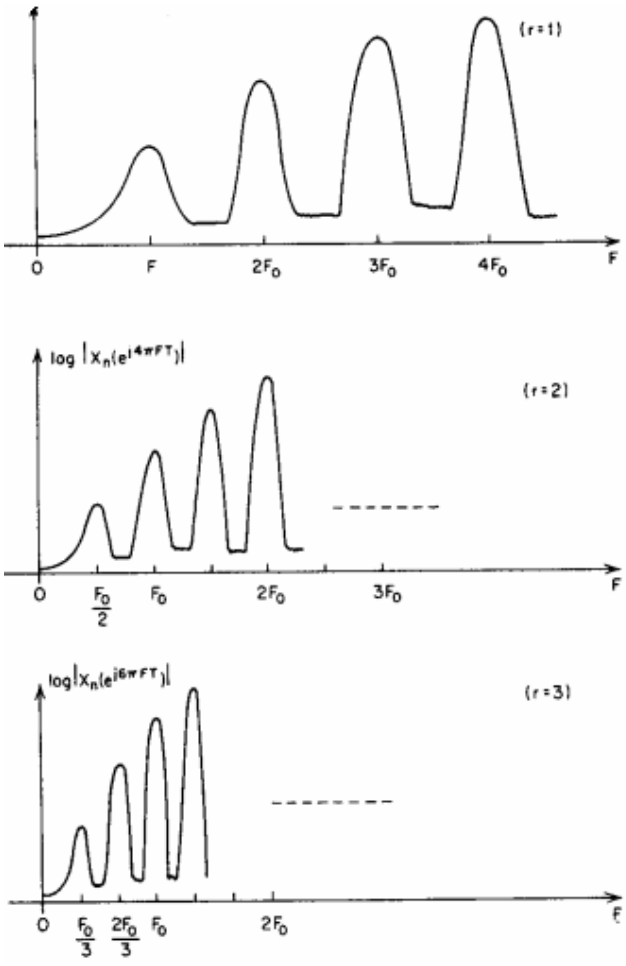
Như đã đề cập nếu dãy s(n) tuần hoàn với chu kỳ T thì hàm AMDF ΔM_n sẽ triệt tiêu tại các giá trị t là bội của số T. Do đó, chúng ta chỉ cần xác định hai điểm cực tiểu gần nhau nhất và từ đó có thể xác định được chu kỳ của dãy và từ đó suy ra tần số cơ bản.

c) Sử dụng tốc độ trở về không - zero crossing rate

Khi xem xét các tín hiệu với thời gian rời rạc, một lần qua điểm không của tín hiệu xảy ra khi các mẫu cạnh nhau có dấu khác nhau. Do vậy, tốc độ qua điểm không của tín hiệu là một đo lường đơn giản của tần số của tín hiệu. Lấy ví dụ, một tín hiệu hình sin có tần số F_0 được lấy mẫu với tần số F_s sẽ có F_s/F_0 mẫu trong một chu kỳ. Vì mỗi chu kỳ có hai lần qua điểm không nên tốc độ trung bình qua điểm không là $Z_n=2F_0/F_s$. Như vậy, tốc độ qua điểm không trung bình cho là một cách đánh giá tương đối về tần số của sóng sin.

d) Phương pháp sử dụng STFT

Từ kết quả phần biểu diễn Fourier của tín hiệu tiếng nói, dễ thấy rằng nguồn kích thích của tín hiệu âm hữu thanh được tăng cường ở những đỉnh nhọn và các đỉnh này xảy ra ở các điểm là bội số của tần số cơ bản. Đây chính là nguyên lý cơ bản của một trong các phương pháp xác định tần số cơ bản.



Hình 3.11 Sự nén tần số

Xét biểu thức phổ tích các hài (harmonic) như sau:

$$P_n\left(e^{j\omega}\right) = \prod_{r=1}^{K}\left|S_n\left(e^{j\omega r}\right)\right| \tag{3.48}$$

Nếu lấy lô-ga-rít của biểu thức (3.48), thu được phổ tích các hài trong thang lô-ga-rít:

$$\hat{P}_n\left(e^{j\omega}\right) = 2\sum_{r=1}^{K}\log\left|S_n\left(e^{j\omega r}\right)\right| \tag{3.49}$$

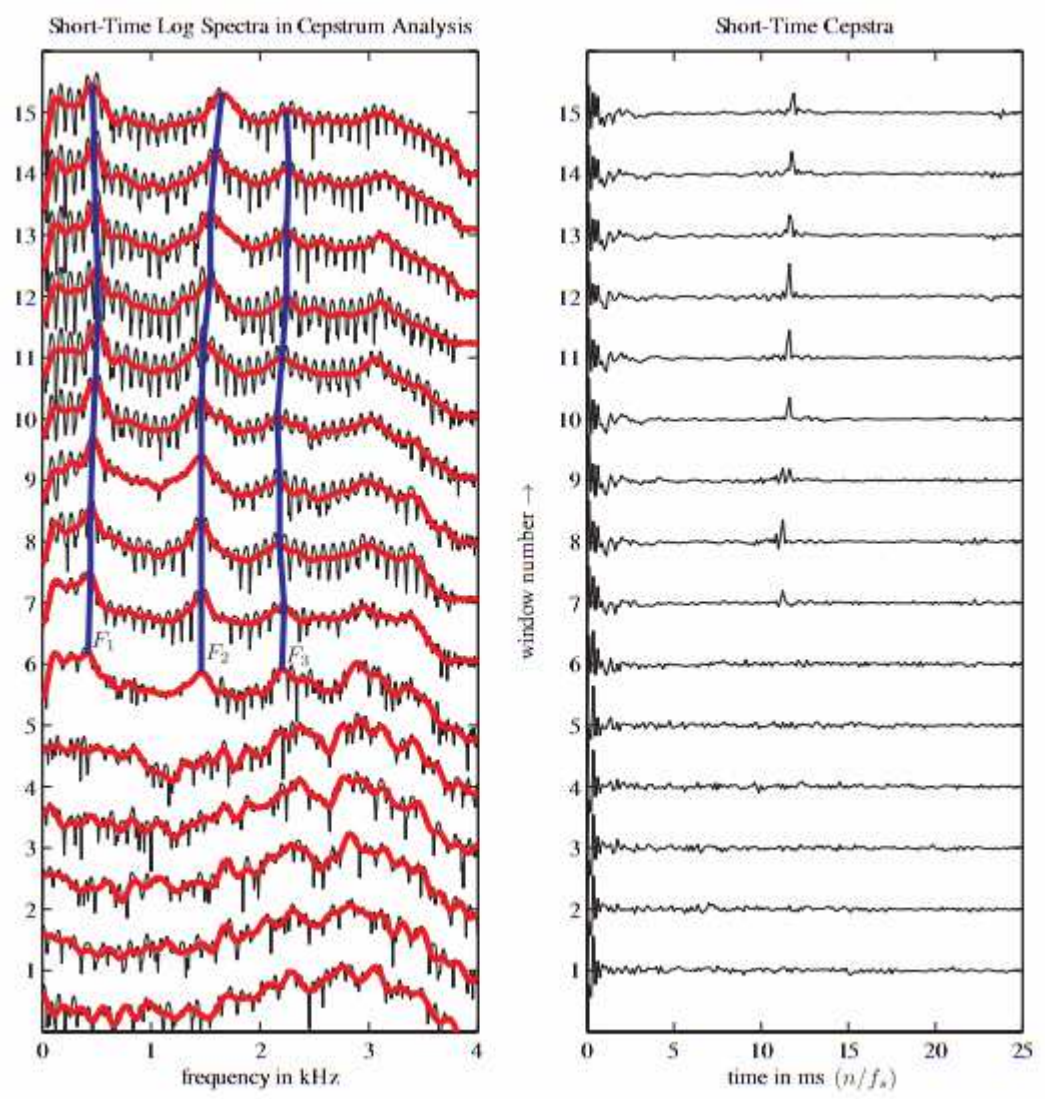
Hàm $\hat{P}_n\left(e^{j\omega}\right)$ trong công thức (3.49) là một tổng của K phổ nén tần số của $|S_n(e^{j\omega})|$. Việc sử dụng hàm trong công thức (3.49) xuất phát từ nhận xét rằng với tín hiệu âm hữu thanh, việc nén tần số bởi các hệ số nguyên sẽ làm các hài của tần số cơ bản trùng với tần số cơ bản. Ở vùng tần số giữa các hài, có một hài của các số tần số khác cũng bị nén trùng nhau, tuy nhiên chỉ tại tần số cơ bản là được củng cố. Hình 3.11 minh họa nhận xét vừa nêu.

e) Sử dụng phân tích Cepstral

Trong phân tích cepstral người ta quan sát thấy rằng, với tín hiệu âm hữu thanh, có một đỉnh nhọn tại chu kỳ cơ bản của tín hiệu. Tuy nhiên với tín hiệu âm vô thanh thì đỉnh nhọn này không xuất hiện. Do đó, phân tích cepstral có thể được sử dụng như một công cụ cơ bản dùng để xác định xem một đoạn tín hiệu tiếng nói là tín hiệu âm vô thanh hay hữu thanh, và để xác định chu kỳ cơ bản của tín hiệu âm hữu thanh. Phương pháp sử dụng phân tích cepstral để ước lượng tần số cơ bản khá đơn giản. Trước hết các cepstrum được tính toán và tìm kiếm

đỉnh nhọn trong một khoảng lân cận của chu kỳ phỏng đoán. Nếu đỉnh cepstrum tại đó lớn hơn một ngưỡng định trước thì tín hiệu tiếng nói đưa vào có khả năng lớn là tín hiệu âm hữu thanh và vị trí đỉnh đó là một ước lượng chu kỳ tín hiệu cơ bản (cũng tức là xác định được tần số cơ bản).

Hình 3.12 minh họa việc sử dụng phương pháp phân tích cepstral để xác định tín hiệu âm vô thanh và hữu thanh cùng với xác định tần số cơ bản của âm hữu thanh. Phía bên trái là dãy các lô-ga phổ ngắn hạn (các đường thay đổi rất nhanh theo thời gian), phía bên phải là các dãy cepstra tương ứng được tính toán từ các lô-ga phổ phía bên tai trái. Các dãy lô-ga phổ và cepstra tương ứng là các đoạn liên tiếp chiều dài 50ms thu được từ hàm cửa sổ dịch 12,5ms mỗi bước (nghĩa là dịch khoảng 100 mẫu ở tần số lấy mẫu 800mẫu/giây). Từ hình vẽ, chúng ta thấy các dãy 1-5, cửa sổ tín hiệu chỉ bao gồm tín hiệu âm vô thanh (không xuất hiện đỉnh, sự thay đổi phổ rất nhanh và xảy ra ngẫu nhiên không có cấu trúc chu kỳ) trong khi các dãy 6 và 7 bao gồm cả tín hiệu âm vô thanh và hữu thanh. Các dãy 8-15 chỉ bao gồm tín hiệu âm hữu thanh. Dễ dàng thấy đỉnh cepstrum tại tần số ứng với 11-12ms tín hiệu âm hữu thanh. Và như vậy, tần số của đỉnh là một ước lượng chính xác tần số cơ bản trong khoảng tín hiệu hữu thanh.



Hình 3.12 Lô-ga-rít các thành phần hài trong phổ tín hiệu

3.10. Bài thực hành phân tích tiếng nói

Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) thực hiện các công việc sau:

Với cùng một nội dung thông tin, các thành viên trong nhóm lần lượt phát âm (đọc/nói) và ghi âm. Lưu tệp ở định dạng thô (*.wav).

Sử dụng phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) và kiến thức đã học trong chương này:

Xác định tần số cơ bản

Xác định tần số của Formant đầu tiên của mỗi thành viên

Lập bản đồ phân bố của các nguyên âm trong tiếng Việt.

Chương 4: Tổng hợp tiếng nói

4.1. Mở đầu

Trước đây khái niệm "tổng hợp tiếng nói" thường được dùng để chỉ quá trình tạo âm thanh tiếng nói một cách nhân tạo từ máy dựa theo nguyên lý mô phỏng cơ quan phát âm của người. Tuy nhiên ngày nay, cùng với sự phát triển của khoa học công nghệ, khái niệm này đã được mở rộng bao gồm cả quá trình cung cấp các thông tin dạng tiếng nói từ máy trong đó các bản tin được tạo dựng một cách linh động để phù hợp cho nhu cầu nào đó. Các ứng dụng của các hệ thống tổng hợp tiếng nói ngày nay rất rộng rãi, từ việc cung cấp các thông tin dạng tiếng nói, các máy đọc cho người mù, những thiết bị hỗ trợ cho người gặp khó khăn trong việc giao tiếp,...

4.2. Các phương pháp tổng hợp tiếng nói

4.2.1 Tổng hợp trực tiếp

Một phương pháp đơn giản thực hiện việc tổng hợp các bản tin là phương pháp tổng hợp trực tiếp trong đó các phần của bản tin được chắp nối bởi các phần (fragment) đơn vị của tiếng nói con người. Các đơn vị tiếng nói thường là các từ hoặc các cụm từ được lưu trữ và bản tin tiếng nói mong muốn được tổng hợp bằng cách lựa chọn và chắp nối các đơn vị thích hợp. Có nhiều kỹ thuật trong việc tổng hợp trực tiếp tiếng nói và các kỹ thuật này được phân loại theo kích thước của các đơn vị dùng để chắp nối cũng như những loại biểu diễn tín hiệu dùng để chắp nối. Các phương pháp phổ biến có thể kêt đến là: phương pháp chắp nối từ, chắp nối các đơn vị từ con (âm vị sub-word unit), chắp nối các phân đoạn dạng sóng tín hiệu.

a) Phương pháp tổng hợp trực tiếp đơn giản

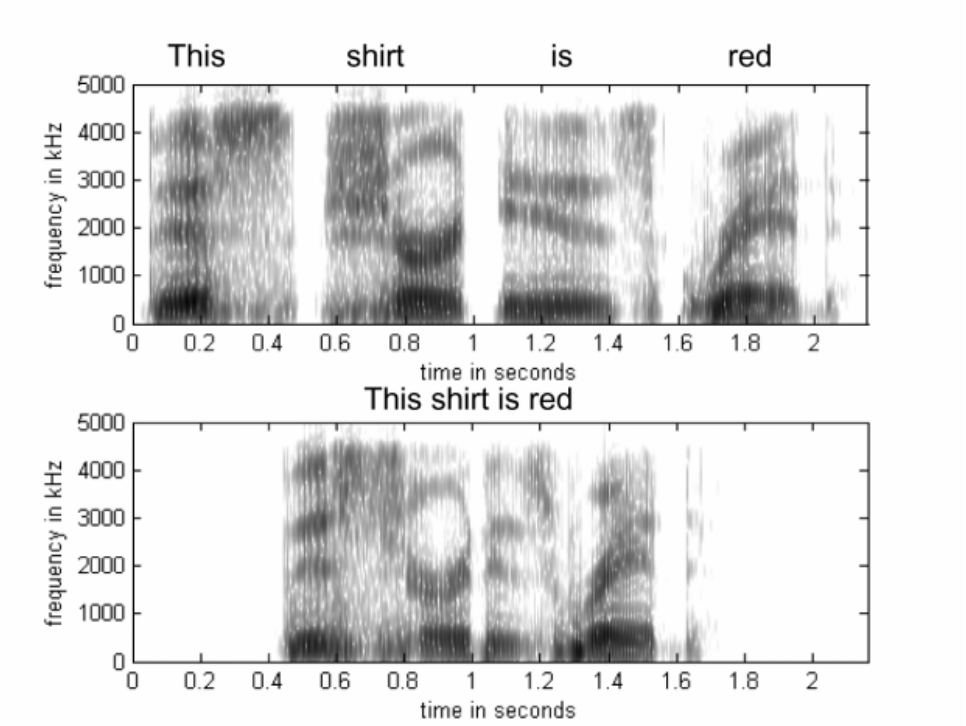
Phương pháp đơn giản nhất để tạo các bản tin tiếng nói là ghi và lưu trữ tiếng nói của con người theo các đơn vị từ riêng lẻ khác nhau và sau đó chọn phát lại các từ theo thứ tự mong muốn nào đó. Phương pháp này được đưa vào sử dụng trong hệ thống điện thoại của nước Anh từ những năm 36 của thế kỷ trước, từ những năm 60 của thế kỷ trước thường được dùng trong một số hệ thống thông báo công cộng, và ngày nay vẫn còn có mặt ở nhiều hệ thống quản lý điện thoại trên thế giới. Hệ thống phải lưu trữ đầy đủ các thành phần của các bản tin cần thiếtt phải tái tạo và lưu trong một bộ nhớ. Bộ tổng hợp chỉ làm nhiệm vụ kết nối các đơn vị yêu cầu cấu thành bản tin lại với nhau theo một thứ tự nào đó mà không phải thay đổi hay biến đổi các thành phần riêng rẽ.

Chất lượng của bản tin tiếng nói được tổng hợp theo phương pháp này bị ảnh hưởng bởi chất lượng của tính liên tục của các đặc trưng âm học (biên phổ, biên độ, tần số cơ bản, tốc độ nói) của các đơn vị được chắp nối. Phương pháp tổng hợp này tỏ ra hiệu quả khi các bản tin có dạng một danh sách chẳng hạn như một dãy số cơ bản, hoặc các khối bản tin thường xuất hiện ở một vị trí nhất định trong câu. Điều này dễ hiểu bởi vì điều đó cho phép dễ dàng đảm bảo rằng bản tin được phát ra có tính tự nhiên về mặt thời gian và cao độ. Khi có yêu cầu một cấu trúc câu đặc biệt nào đó mà trong đó các từ thay thế ở những vị trí nhất định trong câu thì các từ đó phải được ghi lại đúng như thứ tự của nó ở trong câu nếu không nó sẽ không phù hợp với ngữ điệu của câu. Chẳng hạn với các dãy số cơ bản cũng cần thiết phải ghi lại chúng ở hai dạng: một tương ứng với vị trí cuối câu và một dạng không. Điều này là vì cấu trúc pitch của mỗi đơn vị tiếng nói thay đổi tùy theo vị trí của từ trong câu. Như vậy, quá trình biên soạn

là một quá trình rất tốn thời gian và công sức. Ngoài ra việc chắp nối trực tiếp các đơn vị tiếng nói gặp rất nhiều khó khăn trong việc diễn tả sự ảnh hưởng tự nhiên giữa các từ, cũng như ngữ điệu và nhịp điệu của câu. Một hạn chế nữa phải kể đến là kích thước của bộ nhớ cho các ứng dụng với số lượng các bản tin lớn là rất lớn.

Yêu cầu bộ nhớ lưu trữ lớn có thể được phần nào giải quyết bằng việc sử dụng phương pháp mã hóa tốc độ thấp cho các đơn vị tiếng nói trước khi thực hiện việc lưu trữ. Tuy nhiên cả phương pháp sử dụng lưu trữ trực tiếp hoặc mã hóa của các đơn vị lớn (từ, cụm từ) của tiếng nói, số lượng bản tin có thể tổng hợp được rất hạn chế. Để tăng số lượng bản tin có thể tổng hợp được, các đơn vị từ có thể được chia nhỏ hơn thành đơn vị từ con, diphone, demisyllable, syllable... được ghi và lưu trữ. Tuy nhiên khi đơn vị tiếng nói càng được chia nhỏ thì chất lượng bản tin tổng hợp được chất lượng càng bị giảm.

Hình 4.1 minh họa sự so sánh spectrogram của câu tổng hợp được theo phương pháp tổng hợp trực tiếp đơn giản và bản tin nguyên thủy.



Hình 4.1 So sánh kết quả từ bản tin tổng hợp trực tiếp và bản tin nguyên thủy

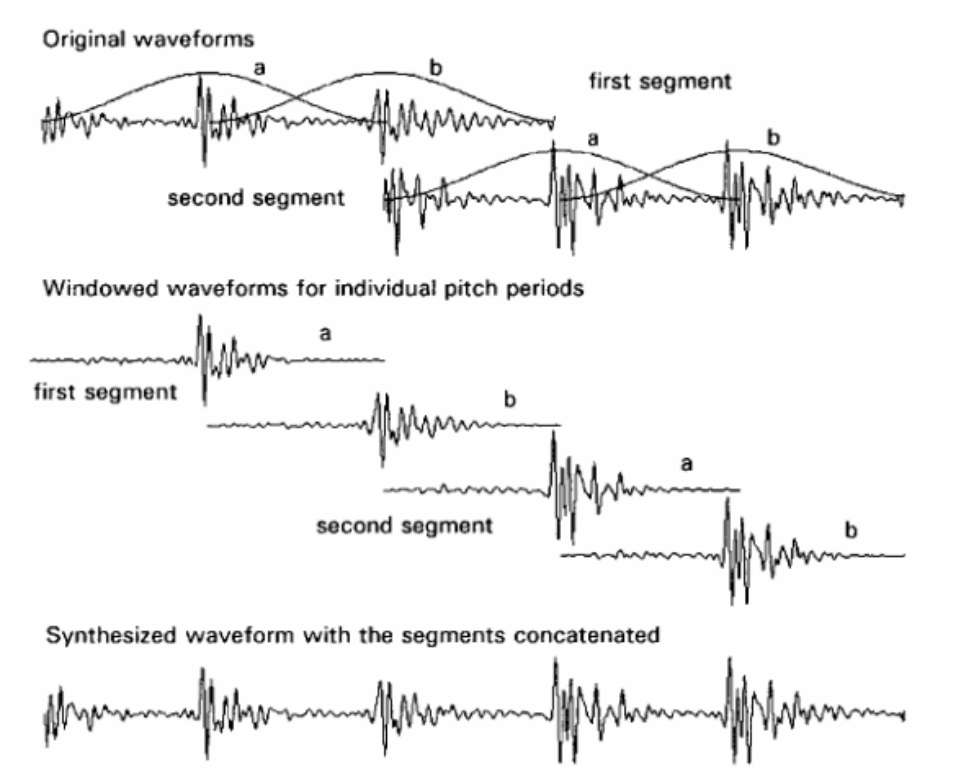
b) Phương pháp tổng hợp trực tiếp từ các phân đoạn dạng sóng

Như đã đề cập phần trên, phương pháp tổng hợp trực tiếp đơn giản gặp phải hạn chế trong việc khôi phục tốc độ và tính tự nhiên (nhấn, nhịp, ngữ điệu) của bản tin được tổng hợp. Vấn đề này có thể được giải quyết bằng cách sử dụng phương pháp tổng hợp từ các phân đoạn dạng sóng hay còn gọi là phương pháp tổng hợp chồng và thêm các đoạn sóng theo độ dài pitch. Xem xét bài toán chắp nối hai phân đoạn của dạng sóng của tín hiệu của nguyên âm. Chúng ta thấy rằng sự không liên tục trong dạng sóng tổng hợp sẽ được giảm nhỏ tối thiểu nếu việc chắp nối xảy ra ở cùng vị trí của một chu kỳ glottal của cả hai phân đoạn. Vị trí này thường là vị trí tương ứng với vùng có biên độ tín hiệu nhỏ nhất khi đáp ứng tuyến âm với xung glottal hiện tại có sự suy giảm lớn và chỉ ngay trước một xung tiếp theo. Nói cách khác, hai phân đoạn tín hiệu được chắp nối theo kiểu đồng bộ pitch (pitch-synchronous manner).

Phương pháp phổ biến thực hiện việc này là phương pháp TD-PSOLA (Time domain Pitch Synchronous Overlap Add).

TD-PSOLA thực hiện việc đánh dấu các vị trí tương ứng với sự đóng lại của dây thanh (tức là xung pitch) trong dạng sóng tín hiệu tiếng nói. Các vị trí đánh dấu này được sử dụng để tạo ra các phân đoạn cửa sổ của dạng sóng tín hiệu cho mỗi chu kỳ. Với mỗi chu kỳ, hàm cửa sổ phải được chỉnh trùng với trung tâm của vùng có biên độ tín hiệu cực đại và hình dạng của hàm cửa sổ phải được chọn thích hợp. Ngoài ra, độ dài hàm cửa sổ phải dài hơn một chu kỳ nhằm tạo ra một sự chồng lấn nhỏ giữa các cửa sổ tín hiệu cạnh nhau.

Hình 4.2 minh họa nguyên lý làm việc của phương pháp TD-PSOLA trong đó sử dụng hàm cửa sổ Hanning.



Hình 4.2 Nguyên lý phương pháp TD-PSOLA

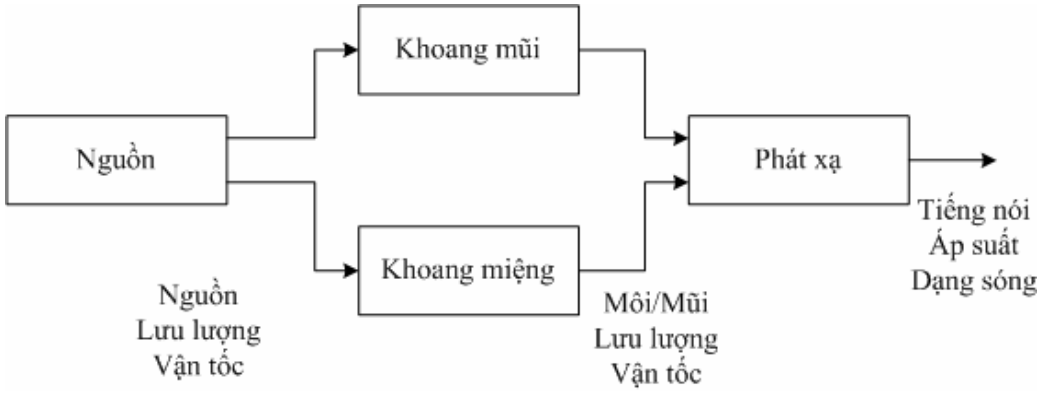
Từ minh họa, chúng ta thấy rằng, bằng cách chắp nối dãy các phân đoạn cửa sổ tín hiệu sóng theo các vị trí tương đối cho trước theo các điểm dấu pitch đã phân tích, chúng ta có thể tái tạo một cách khá chính xác bản tin theo ý mong muốn. Ngoài ra, bằng cách thay đổi các vị trí tương đối và số lượng các điểm dấu pitch, chúng ta có thể làm thay đổi pitch và thời gian của bản tin được tổng hợp.

4.2.2 Tổng hợp tiếng nói theo Formant

Phương pháp tổng hợp theo Formant là phương pháp tổng hợp đích thực đầu tiên được phát triển và là phương pháp tổng hợp phổ biến cho đến tận những năm đầu của thập kỷ 80. Phương pháp tổng hợp theo Formant còn được gọi là phương pháp tổng hợp theo luật. Nó sử dụng các phương pháp mô-đun (modular), dựa trên mô hình (model-based), mối quan hệ âm thanh-âm tiết để giải các bài toán tổng hợp tiếng nói. Trong phương pháp này, mô hình ống âm thanh được sử dụng một cách đặt biết sao cho các thành phần điều khiển của ống dễ dàng

được liên hệ với các tính chất của mối quan hệ âm thanh-âm tiết (acoustic-phonetic) và có thể quan sát được một cách dễ dàng.

Hình 4.3 mô tả sơ đồ tổng quát một hệ thống tổng hợp theo formant. Nguyên lý tổng quát của hệ thống được mô tả như sau. Âm thanh được phát ra từ một nguồn. Đối với các nguyên âm và các phụ âm hữu thanh thì nguồn âm này có thể được tạo ra hoặc đầy đủ bằng một hàm tuần hoàn trong miền thời gian hoặc bằng một dãy đáp ứng xung đưa qua mạch lọc tuyến tính mô phỏng khe thanh (glottal LTI filter). Đối với các âm vô thanh thì nguồn âm này được tạo ra từ một bộ phát nhiễu ngẫu nhiên. Đối với các âm tắc thì nguồn cơ bản này được tạo ra bằng cách kết hợp nguồn cho âm hữu thanh và nguồn cho âm vô thanh. Tín hiệu âm thanh từ nguồn âm cơ bản được đưa vào mô hình tuyến âm (vocal tract). Để tái tạo tất cả các formant, mô phỏng khoang miệng và khoang mũi được xây dựng song song riêng biệt. Do đó, khi tín hiệu đi qua hệ thống sẽ đi qua mô hình khoang miệng, nếu có yêu cầu về các âm mũi thì cũng đi qua hệ thống mô hình khoang mũi. Cuối cùng kết quả các thành phần âm thanh tạo ra từ các mô hình khoang miệng và mũi được kết hợp lại và được đưa qua hệ thống phát xạ, hệ thống này mô phỏng các đặc tính lan truyền và đặc tính tải của môi và mũi.



Hình 4.3 Sơ đồ phương pháp tổng hợp theo formant

Theo lý thuyết mạch lọc, một formant có thể được tạo ra bằng các sử dụng một mạch lọc IIR bậc hai với hàm truyền:

$$H(z) = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2}} \tag{4.1}$$

Trong đó hàm truyền đạt có thể phân tích thành:

$$H(z) = \frac{1}{\left(1 - p_1 z^{-1}\right)\left(1 - p_2 z^{-1}\right)} \tag{4.2}$$

Chúng ta biết rằng, để xây dựng mạch lọc với các hệ số a_1 và a_2 là thực thì các điểm cực phải có dạng là cặp liên hợp phức. Cần chú ý rằng một bộ lọc bậc hai như trên sẽ có đồ thị phổ với hai formant, tuy nhiên chỉ có một trong hai nằm ở phần tần số dương. Do đó, chúng ta có thể coi bộ lọc trên tạo ra một formant đơn lẻ có ích. Các điểm cực có thể quan sát được trên đồ thị, trong đó độ lớn biên độ của các điểm cực quyết định băng tần và biên độ của cộng hưởng. Độ lớn biên độ càng nhỏ thì cộng hưởng càng phẳng, ngược lại, độ lớn biên độ càng lớn thì cộng hưởng càng nhọn.

Nếu biểu diễn các điểm cực trong tọa độ cực với pha θ và bán kính r và chú ý đến nhận xét cặp điểm cực là liên hợp phức chúng ta có thể viết hàm truyền đạt trong công thức (4.1) như sau:

$$H(z) = \frac{1}{1 - 2r\cos(\theta) + r^2 z^{-2}} \tag{4.3}$$

Từ đây chúng ta thấy cúng ta có thể tạo ra một formant với bất cứ tần số mong muốn nào bằng việc sử dụng trực tiếp giá trị thích hợp của θ. Tuy vậy việc điều khiển băng tần một cách trực tiếp khó khăn hơn. Vị trí của formant sẽ thay đổi hình dạng của phổ do đó một mối quan hệ chính xác cho mọi trường hợp là không thể đạt được. Cũng cần chú ý rằng, nếu hai điểm cực gần nhau, chúng sẽ có ảnh hưởng đến việc kết hợp thành một đỉnh cộng hưởng duy nhất và điều này lại gây khó khăn cho việc tính toán băng tần. Thực nghiệm cho thấy mối liên hệ giữa băng tần chuẩn hóa của formant và bán kính của điểm cực có thể xấp xỉ hợp lý bởi:

$$\hat{B} = -2\ln(r) \tag{4.4}$$

Khi đó ta có thể biểu diễn hàm truyền đạt theo hàm của tần số chuẩn hóa $\hat{F}$ và băng tần chuẩn hóa $\hat{B}$ của formant như sau:

$$H(z) = \frac{1}{1 - 2e^{-2\hat{B}}\cos\left(2\pi\hat{F}\right)z^{-1} + e^{-2\hat{B}}z^{-2}} \tag{4.5}$$

Ở đây, các tần số chuẩn hóa $\hat{F}$ và băng tần chuẩn hóa $\hat{B}$ có thể xác định tương ứng bằng cách chia F và B cho tần số lấy mẫu F_s.

Để có thể tạo ra nhiều formant chúng ta có thể thực hiện bằng một bộ lọc mà hàm truyền đạt là tích của một số hàm truyền đạt bậc hai. Nói một cách khác, hàm truyền cho tuyến âm (vocal tract) có dạng:

$$H(z) = H_1(z)H_2(z)H_3(z)H_4(z) \tag{4.6}$$

Trong đó $H_i(z)$ là hàm của tần số F_i và băng tần B_i của formant thứ i.

Tương ứng biểu thức quan hệ đầu vào đầu ra trong miền thời gian có dạng:

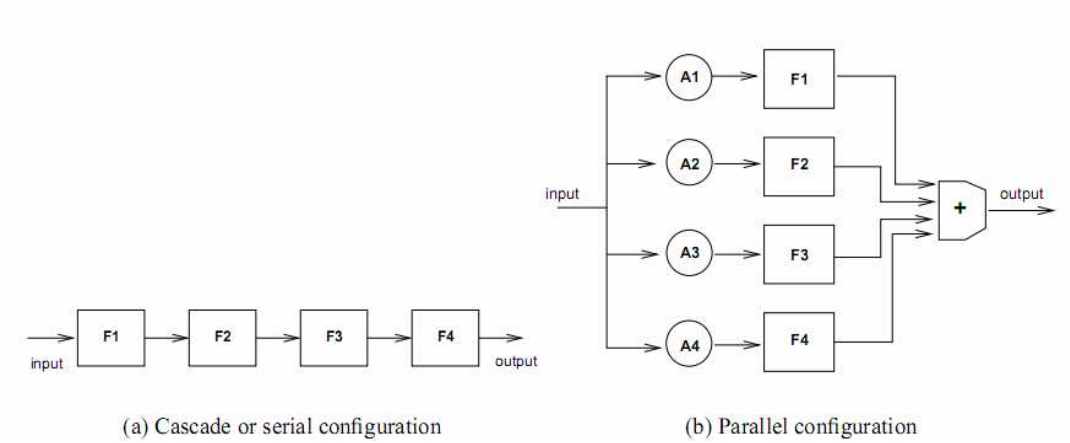
$$y(n) = x(n) + a_1 y(n-1) + a_2 y(n-2) + ... + a_8 y(n-8) \tag{4.7}$$

Một cách tương tự, chúng ta có thể xây dựng hệ thống mô phỏng khoang mũi. Các biểu thức (4.6) và (4.7) biểu diễn kỹ thuật tổng hợp formant theo sơ đồ nối tiếp hay còn gọi là sơ đồ cascade.

Một kỹ thuật khác là tổng hợp formant song song. Phương pháp tổng hợp formant song song mô phỏng mỗi formant riêng rẽ. Nói cách khác, mỗi mô hình có một hàm truyền $H_i(z)$ riêng rẽ. Trong quá trình tạo tín hiệu tiếng nói các nguồn tín hiệu được đưa vào các mô hình một cách riêng rẽ. Sau đó, các tín hiệu từ các mô hình $y_i(n)$ được tổng hợp lại.

$$y(n) = y_1(n) + y_2(n) + ... \tag{4.8}$$

Hình 4.4 minh họa cấu hình tổng quát của phương pháp tổng hợp nối tiếp và song song.



Hình 4.4 Các cấu hình của phương pháp tổng hợp nhiều formant

Phương pháp tổng hợp theo sơ đồ nối tiếp có lợi điểm là với một tập các giá trị formant cho trước, chúng ta có thể dễ dàng xây dựng các hàm truyền đạt và biểu thức quan hệ đầu vào đầu ra (công thức vi sai - difference equation). Việc tổng hợp riêng rẽ các formant trong phương pháp tổng hợp song song cho phép chúng ta xác định một cách chính xác tần số của các formant.

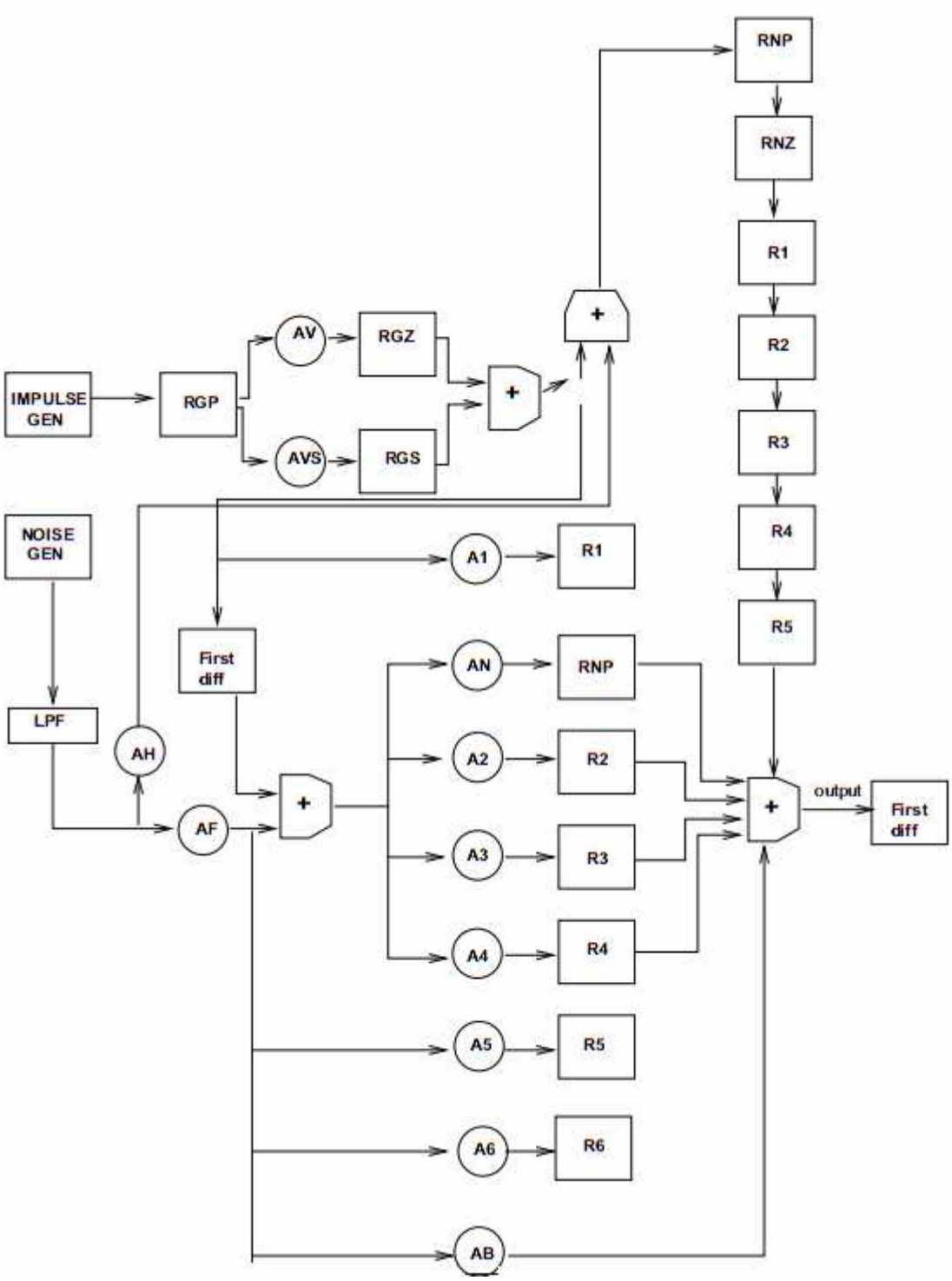
Mặc dù là một phương pháp tổng hợp đơn giản và thường mang lại tín hiệu âm thanh rõ, phương pháp tổng hợp theo formant khó đạt được tính tự nhiên của tín hiệu tiếng nói. Điều này là do mô hình nguồn và mô hình chuyển đổi đã bị đơn giản hóa quá mức và đã bỏ qua nhiều yếu tố phụ trợ góp phần tạo ra đặc tính động của tín hiệu.

Bộ tổng hợp Klatt

Bộ tổng hợp Klatt là một trong các bộ tổng hợp tiến nói dựa trên formant phức tạp nhất đã được phát triển. Sơ đồ của bộ tổng hợp này được trình bày trong hình 4.5 trong đó có sử dụng cả các hệ thống cộng hưởng song song và nối tiếp.

Trong sơ đồ các khối R_i tương ứng với các bộ tạo tần số cộng hưởng formant thứ i; các hộp A_i điều khiển biên độ tín hiệu tương ứng. Bộ cộng hưởng được thiết lập để làm việc ở tần số 10kHz với 6 formant chính được sử dụng.

Cần chú ý rằng, trong thực tế các bộ tổng hợp formant thường sử sụng tần số lấy mẫu khoảng 8kHz hoặc 10kHZ. Điều này không hẳn bởi một lý do nào đặc biệt liên quan đến nguyên tắc về chất lượng tổng hợp mà bởi vì sự hạn chế về không gian lưu trữ, tốc độ xử lý và các yêu cầu đầu ra không cho phép thực hiện với tốc độ lấy mẫu cao hơn. Một điểm khác cũng cần chú ý là, các nghiên cứu đã chúng minh rằng chỉ có ba formant đầu tiên là đủ để phân biệt tín hiệu âm thanh, do đó việc sử dụng 6 formant thì các formant bậc cao đơn giản được sử dụng để tăng thêm tính tự nhiên cho tín hiệu tổng hợp được.



Hình 4.5 Sơ đồ khối bộ tổng hợp Klatt

4.2.3 Tổng hợp tiếng nói theo phương pháp mô phỏng bộ máy phát âm

Một cách hiển nhiên, để tổng hợp tiếng nói thì chúng ta cần tìm một cách nào đó mô phỏng bộ máy phát âm của chúng ta. Đây cũng là nguyên lý của các "máy nói" cổ điển mà nổi tiếng trong số đó là máy do Von Kempelen chế tạo. Các bộ tổng hợp tiếng nói cổ điển theo nguyên lý này thường là các thiết bị cơ học với các ống, ống thổi, ... hoạt động tựa hồ các dụng cụ âm nhạc, tuy nhiên với một chút huấn luyện có thể dùng để tạo ra tín hiệu tiếng nói nhận biết được. Việc điều khiển hoạt động của máy là nhờ con người theo thời gian thực, điều này

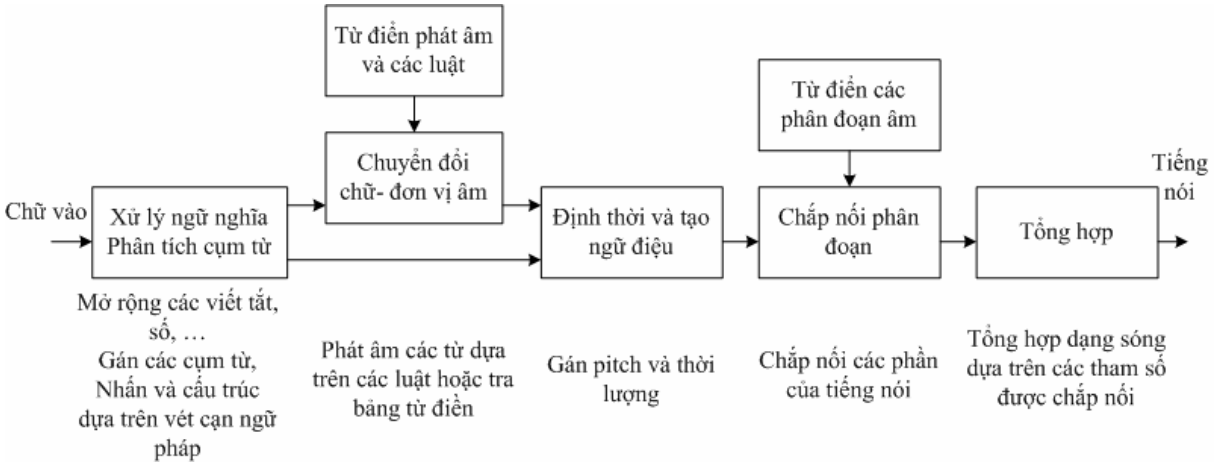
mang lại nhiều thuận lợi cho hệ thống ở khía cạnh con người có thể sử dụng các cơ chế chẳng bạn như thông qua phản hồi để điều khiển và bắt chước quá trình tạo tiếng nói tự nhiên. Tuy nhiên, ngày nay với nhu cầu của các bộ tổng hợp phức tạp hơn, các cỗ máy cổ điển rõ ràng là lỗi thời không thể đáp ứng được.

Cùng với sự hiểu biết của con người về bộ máy phát âm được nâng cao, các bộ tổng hợp sử dụng nguyên lý mô phỏng bộ máy phát âm ngày càng phức tạp và hoàn thiện hơn. Các hình dạng ống phức tạp được xấp xỉ bằng một loạt các ống đơn giản nhỏ hơn. Với mô hình các ống đơn giản, vì chúng ta biết được các đặc tính truyền âm của nó, chúng ta có thể sử dụng để xây dựng các mô hình bộ máy phát âm tổng quát phức tạp.

Một ưu điển của phương pháp tổng hợp mô phỏng bộ máy phát âm là cho phép tạo ra một cách tự nhiên hơn để tạo ra tiếng nói. Tuy nhiên, phương pháp này cũng gặp phải một số khó khăn. Thứ nhất đó là việc quyết định làm thế nào để có được các tham số điều khiển từ các yêu cầu tín hiệu cần tổng hợp. Rõ ràng, khó khăn này cũng gặp phải trong các phương pháp tổng hợp khác. Trong hầu hết các phương pháp tổng hợp khác, chẳng hạn các tham số formant có thể tìm được một cách trực tiếp từ tín hiệu tiếng nói thực, chúng ta chỉ đơn giản ghi âm lại tiếng nói và tính toán rồi xác định chúng. Còn trong phương phương pháp mô phỏng bộ máy phát âm chúng ta sẽ gặp khó khăn hơn vì các tham số về bộ máy phát âm đúng đắn không thể xác định từ việc ghi lại tín hiệu thực mà phải thông qua các đo lường thông qua chẳng hạn ảnh X-ray, MRI... Khó khăn thứ hai là việc cân bằng giữa việc xây dựng một mô hình mô phỏng chính xác cao nhất giống với bộ máy phát âm sinh học của con người và một mô hình thực tiễn dễ thiết kế và thực hiện. Cả hai khó khăn này cho đến nay vẫn được coi là thách thức với các nhà nghiên cứu. Và đây cũng chính là lý do mà cho đến nay có rất ít các hệ thống tổng hợp theo nguyên lý mô phỏng bộ máy phát âm có chất lượng so với các bộ tổng hợp theo nguyên lý khác.

4.3. Hệ thống tổng hợp chữ viết sang tiếng nói

Việc chuyển đổi từ chữ viết sang tiếng nói (TTS) là mục tiêu đầy tham vọng và vẫn đang tiếp tục là tâm điểm chú ý của các nhà nghiên cứu phát triển. TTS có mặt ở nhiều ứng dụng phục vụ cuộc sống. Chẳng hạn như việc các ứng dụng truy cập email qua thoại, các ứng dụng cơ sở dữ liệu cho các dịch vụ hỗ trợ người mù... Một hệ thống TTS điển hình có sơ đồ khối với các thành phần được minh họa trong hình 4.6.

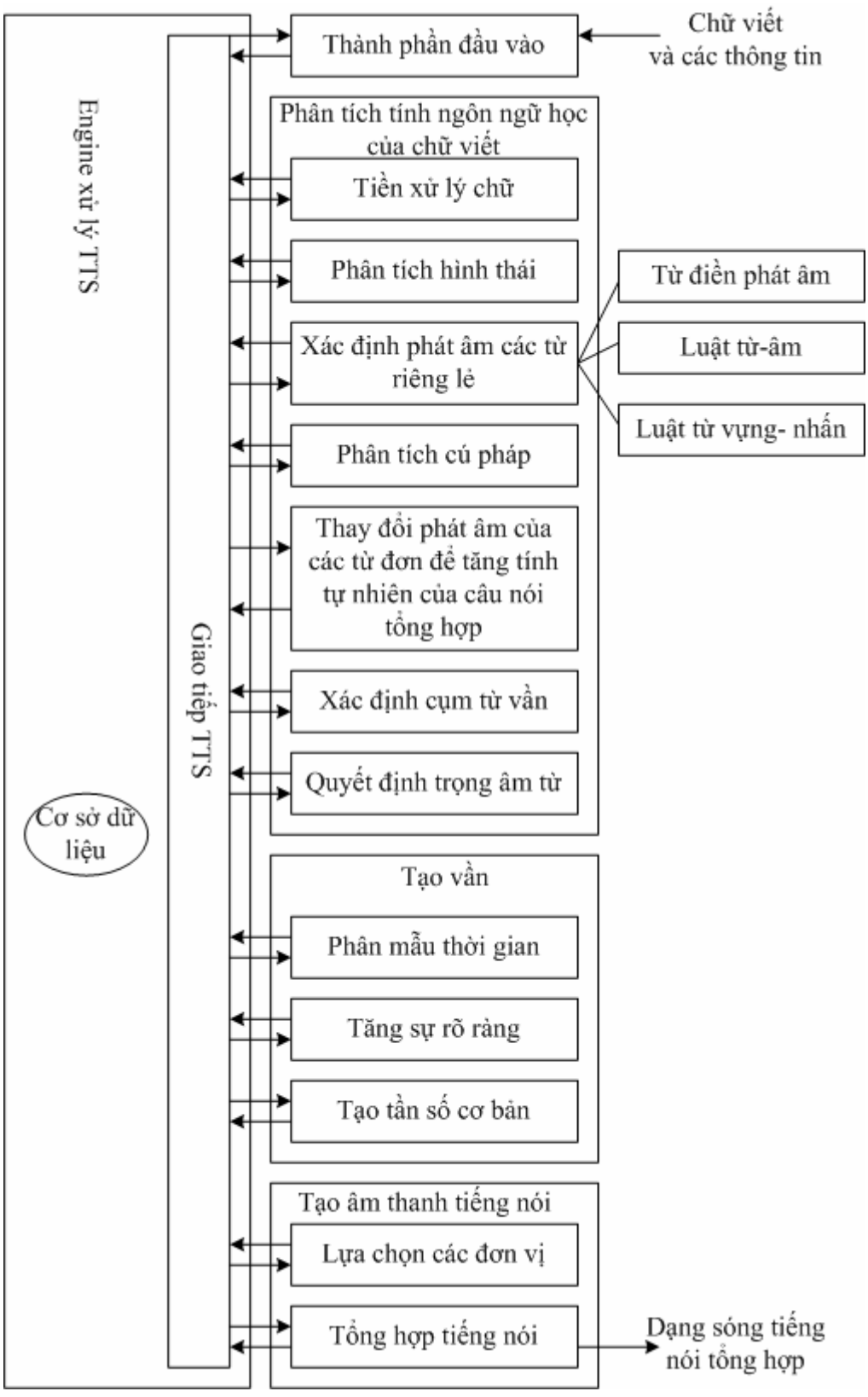


Hình 4.6 Sơ đồ khối một hệ thống TTS

Từ minh họa, chúng ta thấy rằng, hệ thống TTS có thể đặc trưng như một quá trình phân tích-tổng hợp 2-giai đoạn. Giai đoạn một của quá trình thực hiện việc phân tích chữ viết để xác định cấu trúc ngôn ngữ ẩn trong đó. Chữ viết đầu vào thường bao gồm các cụm từ viết tắt, các số La Mã, ngày tháng, công thức, các dấu câu...Giai đoạn phân tích chữ viết phải có khả năng chuyển đổi dạng chữ viết đầu vào thành một dạng chuẩn chấp nhận được để sử dụng cho giai đoạn sau. Các mô tả ngôn ngữ dạng trừu tượng của dữ liệu thu được ở giai đoạn này có thể bao gồm một dãy phoneme và các thông tin khác, chẳng hạn như cấu trúc nhấn, cấu trúc cú pháp...Các mô tả này được chuyển đổi thành một bảng ghi âm tiết nhờ sự giúp đỡ của một từ điển phát âm và các luật phát âm kèm theo. Giai đoạn thứ hai thực hiện việc tổng hợp xây dựng dạng sóng tín hiệu dựa trên các tham số thu được từ giai đoạn trước đó.

Cả quá trình phân tích và tổng hợp của một hệ thống TTS liên quan đến một loạt các hoạt động xử lý. Hầu hết các hệ thống TTS hiện đại thực hiện các hoạt động xử lý được minh họa theo kiến trúc mô-đun như trong hình 4.7.

Hoạt động của sơ đồ khối có thể sơ lược mô tả như sau. Khi dạng dữ liệu chữ viết được đưa vào, mỗi mô-đun trích các thông tin đầu vào hoặc thông tin từ các mô-đun khác liên quan đến chữ viết, và tạo ra các các thông tin đầu ra mong muốn cho việc xử lý ở các mô-đun tiếp theo. Việc trích chuyển được thực hiện cho đến khi dạng tín hiệu tổng hợp cuối cùng được tạo ra. Quá trình xử lý và truyền thông tin từ mô-đun này đến mô-đun khác thông qua một "động cơ" (engine) xử lý riêng biệt. Engine xử lý điều khiển dẫy các hoạt động được thực thi, và lưu trữ mọi thông tin ở dạng cấu trúc dữ liệu thích hợp.



Hình 4.7 Sơ đồ khối kiến trúc mô-đun của một hệ thống TTS hiện đại

a) Phân tích chữ viết

Chúng ta biết rằng, chữ viết bao gồm các ký tự chữ và số, các khoảng trắng, và có thể một loạt các ký tự đặc biệt khác. Như vậy bước đầu tiên trong việc phân tích chữ viết là việc tiền xử lý chữ viết đầu vào (bao gồm thay thế chữ số, các chữ viết tắt bằng dạng viết đầy đủ của

chúng) để chuyển chúng thành một dãy các từ. Quá trình tiền xử lý thông thường còn phát hiện và đánh dấu các vị trí ngắt quãng của câu và các thông tin về định dạng văn bản thích hợp khác chẳng hạn như ngắt đoạn...Các mô-đun xử lý chữ viết tiếp theo sẽ thực hiện việc chuyển dãy từ thành các mô tả ngôn ngữ. Một trong các chức năng quan trọng của các khối này là xác định phát âm tương ứng của các từ riêng lẻ. Trong các ngôn ngữ như ngôn ngữ tiếng Anh, các quan hệ giữa các đánh vần của các từ và dạng ghi âm vị (phonemic transcription) tương ứng là một quan hệ cực kỳ phức tạp. Ngoài ra, mối quan hệ này còn có thể khác nhau với các từ khác nhau có cùng cấu trúc, ví dụ như phát âm của cụm "ough" trong các từ "through", "though", "bough", "rough" và "cough".

Như đã đề cập khái quát trong phần trên, phát âm của từ thường được xác định nhờ việc sử dụng tổng hợp của một từ điển phát âm và các luật phát âm kèm theo. Trong các hệ thống TTS trước khia, nhấn mạnh trong các phát âm xác định được tuân theo luật và bằng cách sử dụng một từ điển các ngoại lệ nhỏ cho các từ chung với cách phát âm bất quy tắc (chẳng hạn như "one", "two", "said", ...). Tuy nhiên ngày nay với sự sẵn có của bộ nhớ máy tính với giá thành rẻ, thường việc xác định phát âm được hoàn thành bằng cách sử dụng một từ điền phát âm rất lớn (có thể gồm hàng vài chục ngàn từ) để đảm bảo rằng từ đã biết được phát âm một cách chính xác. Mặc dù vậy, các luật phát âm vẫn cần thiết để giải quyết vấn đề nảy sinh với các từ không biết vì các từ vựng mới được liên tục thêm vào ngôn ngữ, và cũng như không thể dựa hoàn toàn vào việc thêm vào tất cả các từ vựng là các danh từ riêng trong bộ từ điển. Việc xác định phát âm của từ có thể được thực hiện một cách dễ dàng nếu cấu trúc, hay còn gọi là hình thái học ngôn ngữ (morphology), của từ được biết trước. Hầu hết các hệ thống TTS bao gồm cả các phân tích hình thái ngôn ngữ. Phân tích này xác định dạng gốc (root form của mỗi từ), ví dụ dạng gốc của "gives" là "give", và tránh sự cần thiết phải thêm cả dạng suy ra từ dạng gốc vào trong từ điển. Một số phân tích cú pháp của chữ viết cũng có thể cần được thực hiện nhằm xác định chính xác phát âm của các từ nhất định nào đó. Chẳng hạn, trong tiếng Anh từ "live" được phát âm khác nhau phụ thuộc vào nó đóng vai trò là một động từ hay một tính từ. Các phát âm của từ chúng ta xác định là các phát âm của các từ khi chúng được nói riêng rẽ. Do đó, một số điều chỉnh cần được thực hiện để kết hợp các hiệu ứng âm tiết (phonetic) xảy ra trên vùng biên giữa các từ, nhằm cải thiện tính tự nhiên của tiếng nói tổng hợp được.

Ngoài việc xác định phát âm của dãy từ, giai đoạn phân tích chữ viết cũng phải thực hiện việc xác định các thông tin liên quan đến cách mà chữ viết sẽ được nói. Thông tin này, bao gồm việc phân tiết tấu, dấu nhấn từ (mức từ), và mẫu các ngữ điệu của các từ khác nhau. Các thông tin này sẽ được sử dụng để tạo âm điệu cho tiếng nói được tổng hợp. Các đánh dấu cho dấu nhất từ có thể được thêm vào cho mỗi từ trong từ điển, nhưng các luật cũng sẽ cần để gán dấu nhất từ cho các từ bất kỳ không tìm thấy trong từ điển. Với một số từ, chẳng hạn như từ "permit", về cơ bản có dấu nhất trên các âm tiết khác nhau phụ thuộc vào việc chúng được sử dụng như một danh từ hay một động từ. Và do đó, các thông tin về ngữ pháp cũng cần thiết nhằm gán cấu trúc nhấn một cách chính xác. Kết quả của một phân tích cú pháp cũng có thể được sử dụng để nhóm các từ thành các cụm từ âm điệu, và từ đó quyết định các từ nào sẽ nhấn giọng sao cho mẫu nhấn giọng có thể được gán cho dãy từ. Trong khi cấu trúc cú pháp cung cấp các đầu mối hữu ích cho việc nhấn giọng và phân tiết tấu (và từ đó tạo âm điệu), trong nhiều trường hợp, âm điệu biểu hiện thực có thể không đạt được nếu không thực sự hiểu

nghĩa của chữ viết. Mặc dù một số ảnh hưởng ngữ nghĩa được sử dụng, các phân tích ngữ nghĩa và thực dụng đầy đủ là vượt quá các khả năng của các hệ thống TTS hiện tại.

b) Tổng hợp tiếng nói

Các thông tin được trích từ các phân tích chữ viết được sử dụng để tạo ra âm điều của các đơn vị tiếng nói, bao gồm cả cấu trúc thời gian, mức độ nhấn mạnh toàn bộ và tần số cơ bản. Mô-đun cuối cùng của hệ thống TTS sẽ thực hiện việc tạo âm thanh của tín hiệu tiếng nói bằng cách đầu tiên chọn các đơn vị tổng hợp thích hợp để sử dụng, và sau đó thực hiện việc tổng hợp các đơn vị này với nhau theo thông tin về âm điệu đã biết. Việc tổng hợp có thể được thực hiện bằng một trong các phương pháp đã đề cập ở phần trên.

4.4. Bài thực hành tổng hợp tiếng nói

Sử dụng phương pháp tổng hợp trực tiếp đơn giản

- Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) xây dựng một hệ thống thông báo điểm đỗ xe buýt công cộng.

- Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) xây dựng một hệ thống thông báo số thứ tự khách hàng đến lượt được phục vụ tại một điểm giao dịch ngân hàng.

Chương 5: Nhận dạng tiếng nói

5.1. Mở đầu

Nhu cầu về những thiết bị (máy) có thể nhận biết và hiểu được tiếng nói được nói bởi bất kỳ ai, trong bất kỳ môi trường nào đã trở thành một ước muốn tuột bậc của con người cũng như các nhà nghiên cứu và các dự án nghiên cứu về nhận dạng tiếng nói trong suốt gần một thê kỷ qua. Cho đến nay, mặc dù chúng ta đã đạt được những bước tiến dài trong việc hiểu được quá trình tạo tín hiệu tiếng nói và đưa ra nhiều kỹ thuật phân tích tiếng nói, và thậm chí chúng ta đã đạt được nhiều tiến bộ trong việc xây dựng và phát triển nhiều hệ thống nhận dạng tín hiệu tiếng nói quan trọng, chúng ta vẫn còn đang ở quá xa mục tiêu đặt ra là có thể xây dựng được những cỗ máy có thể giao tiếp một cách tự nhiên với con người. Trong chương này, trước hết chúng ta sẽ xem xét lại lịch sử phát triển của lĩnh vực nghiên cứu nhận dạng tiếng nói, sau đó tìm hiểu sơ bộ một hệ thống nhận dạng tín hiệu tiếng nói tổng quát và một số phương pháp hiện đã đang được sử dụng trong các hệ thống nhận dạng tín hiệu tiếng nói cùng với ưu nhược điểm của nó.

5.2. Lịch sử phát triển các hệ thống nhận dạng tiếng nói

Nghiên cứu về nhận dạng tiếng nói là một lĩnh vực nghiên cứu đã và đang diễn ra được gần một thế kỷ. Trong suốt quá trình đó, chúng ta có thể phân loại các công nghệ nhận dạng thành các thế hệ như sau:

Thế hệ 1: Thế hệ này được đánh mốc bắt đầu từ những năm 30 cho đến những năm 50. Công nghệ của thế hệ này là các phương thức ad hoc để nhận dạng các âm, hoặc các bộ từ vựng với số lượng nhỏ của các từ tách biệt.

Thế hệ 2: Thế hệ thứ hai bắt đầu từ những năm 50 và kết thúc ở những năm 60. Công nghệ của thế hệ này sử dụng các các phương pháp acoustic-phonetic để nhận dạng các phonemes, các âm tiết hoặc các từ vựng của các số.

Thế hệ 3: Thế hệ này sử dụng các biện pháp nhận dạng mẫu để nhận dạng tín hiệu tiếng nói với các bộ từ vựng vừa và nhỏ của các từ tách biệt hoặc dãy từ có liên kết với nhau, bao gồm cả việc sử dụng bộ LPC như là một phương pháp phân tích cơ bản; sử dụng các đo lượng khoảng cách LPC để cho điểm sự tương đồng của các mẫu; sử dụng các giải pháp lập trình động cho việc chỉnh thời gian; sử dụng nhận dạng mẫu cho việc phân hoạch các mẫu thành các mẫu tham chiếu nhất quán, sử dụng phương pháp mã hóa lượng tử hóa véc-tơ để giảm nhỏ dữ liệu và tính toán. Thế hệ thứ ba bắt đầu từ những năm 60 đến những năm 80.

Thế hệ 4: Thế hệ thứ tư bắt đầu từ những năm 80 đến những năm 00. Công nghệ của thế hệ này sử dụng các phương pháp thống kê với mô hình Markov ẩn (HMM) cho việc mô phỏng tính chất động và thống kê của tín hiệu tiếng nói trong một hệ thống nhận dạng liên tục; sử dụng các phương pháp huấn luyện lan truyền xuôi-ngược và phân đoạn K-trung bình (segmental K-mean); sử dụng phương pháp chỉnh thời gian Viterbi; sử dụng thuật toán độ tương đồng tối da (ML) và nhiều tiêu chuẩn chất lượng cùng các giải pháp để tối ưu hóa các mô hình thống kê; sử dụng mạng nơ-ron để ước lượng các hàm mật độ xác suất có điều kiện; sử dụng các thuật toán thích nghi để thay đổi các tham số gắn với hoặc tín hiệu tiếng nói hoặc với mô hình thống kê để nâng cao tính tương thích giữa mô hình và dữ liệu nhằm tăng tính chính xác của phép nhận dạng.

Thế hệ 5: Chúng ta đang chứng kiến sự phát triển của lớp công nghệ nhận dạng tiếng nói thế hệ thứ năm. Công nghệ thế hệ này sử dụng các giải pháp xử lý song song để tăng tính tín cậy trong các quyết định nhận dạng; kết hợp giữa HMM và các phương pháp acoustic-phonetic để phát hiện và sửa chữa những ngoại lệ ngôn ngữ; tăng tính chắc chắn (chín chắn - robustness) của hệ thống nhận dạng trong môi trường có nhiễu; sử dụng phương pháp học máy để xây dựng các kết hợp tối ưu của các mô hình.

Cũng cần chú ý rằng, việc phân chia các giai đoạn chỉ mang tính tương đối về mốc thời gian. Điều này dễ hiểu bởi vì các thế hệ công nghệ không phân tách rạch ròi nhau mà hầu như các ý tưởng cốt lõi của mỗi giai đoạn lại được thai nghén từ giai đoạn trước đó. Các giai đoạn được phân chia chỉ nhằm chỉ ra rằng trong giai đoạn đó nhiều kết quả nghiên cứu liên quan đến công nghệ của giai đoạn đó đựoc đưa ra và trở thành tiêu chuẩn cho hầu hết các hệ thống nhận dạng của thời kỳ đó.

5.3. Phân loại các hệ thống nhận dạng tiếng nói

Tùy theo các cách nhìn mà chúng ta các cách phân loại các hệ thống nhận dạng tiếng nói khác nhau. Xét theo khía cạnh đơn vị tiếng nói được sử dụng trong các hệ thống, thì các hệ thống nhận dạng tiếng nói có thể được phân thành hai loại chính. Loại thứ nhất là các hệ thống nhận dạng từ riêng lẻ, trong đó các biểu diễn từ phân tách đơn lẻ được nhận dạng. Loại thứ hai là các hệ thống nhận dạng liên tục trong đó các câu liên tục được nhận dạng. Hệ thống nhận dạng tiếng nói liên tục còn có thể chia thành lớp nhận dạng với mục đích ghi chép (transcription) và lớp với mục đích hiểu tín hiệu tiếng nói. Lớp với mục đính ghi chép có mục tiêu nhận dạng mỗi từ một cách chính xác. Lớp với mục đích hiểu, cũng còn được gọi là lớp nhận dạng tiếng nói hội thoại, tập trung vào việc hiểu nghĩa của các câu thay vì việc nhận dạng các từ riêng biệt. Trong các hệ thống nhận dạng tiếng nói liên tục, điều quan trọng là phải sử dụng các kiến thức ngôn ngữ phức tạp. Chẳng hạn như việc ứng dụng các luật về ngữ pháp, các luật quy định về việc tổ chức dãy các từ trong câu, là một ví dụ.

Theo cách nhìn khác, các hệ thống nhận dạng tiếng nói có thể được phân chia thành các hệ thống nhận dạng không phụ thuộc vào người nói (speaker-independent) và hệ thống nhận dạng phụ thuộc vào người nói (speaker-dependent). Hệ thống nhận dạng độc lập với người nói có khả năng nhận dạng tiếng nói của bất cứ ai. Trong khi đó, đối với hệ thống nhận dạng phụ thuộc người nói, các mẫu/mô hình tham khảo cần phải thay đổi cập nhật mỗi lần người nói thay đổi. Mặc dù việc nhận dạng độc lập với người nói khó hơn rất nhiều so với việc nhận dạng phụ thuộc người nói, nhưng việc phát triển các phương nhận dạng độc lập là đặc biệt quan trọng nhằm mở rộng phạm vi sử dụng của các hệ thống nhận dạng.

Ngoài ra, các hệ thống tiếng nói cũng có thể phân chia làm các nhóm sau: các hệ thống nhận dạng tiếng nói tự động, các hệ thống nhận dạng tiếng nói liên tục, và các hệ thống xử lý ngôn ngữ tự nhiên (NLP - Natural Language Processing). Các hệ thống nhận dạng tiếng nói tự động, như tên mô tả, là các hệ thống nhận dạng mà không cần thông tin đầu vào của người sử dụng bổ sung vào. Các hệ thống nhận dạng tiếng nói liên tục, như đã đề cập ở phần trên, là các hệ thống có khả năng nhận dạng các câu liên tục. Nói cách khác, về mặt lý thuyết, các hệ thống loại này không yêu cầu người sử dụng (người nói) phải ngừng trong khi nói. Các hệ thống xử lý ngôn ngữ tự nhiên có ứng dụng không chỉ trong các hệ thống nhận dạng tiếng nói. Các hệ thống sử dụng các phương pháp tính toán cần thiết cho các máy có thể hiểu được nghĩa của tiếng nói đang được nói thay vì chỉ đơn giản biết được từ nào đã được nói.

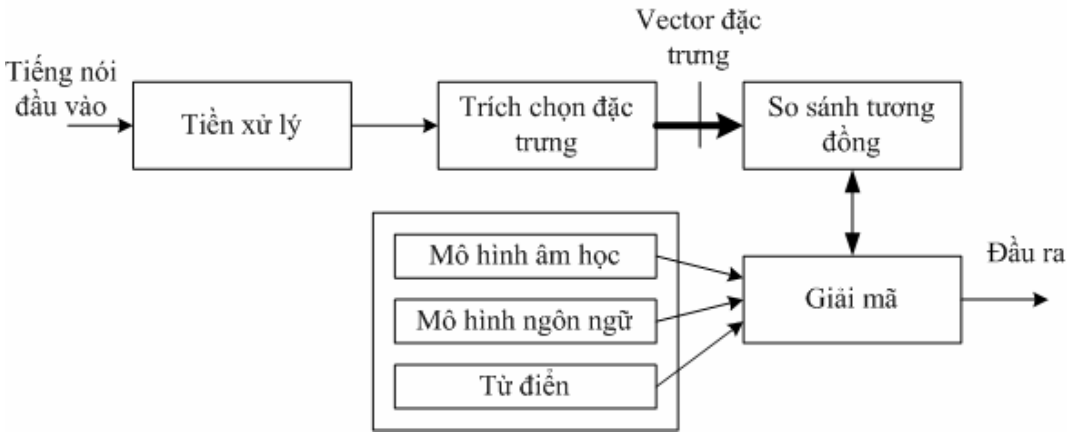
Một cách tổng quát, Victo Zue và đồng nghiệp đã định nghĩa một số tham số và dùng nó để phân chia các hệ thống nhận dạng theo các tham số đó như trình bày trong bảng 5.1.

Tham số	Phân loại điển hình
Đơn vị tiếng nói	Rời rạc (các từ đơn lẻ) – Liên tục (các câu liên tục)
Huấn luyện	Huấn luyện trước khi sử dụng - Huấn luyện liên tục
Người sử dụng	Phụ thuộc - Độc lập
Từ vựng	Số lượng nhỏ - Số lượng lớn
SNR	Thấp – Cao
Bộ chuyển đổi	Hạn chế - Không hạn chế

Bảng 5.1: Các tham số và phân loại hệ thống nhận dạng tương ứng

5.4. Cấu trúc hệ nhận dạng tiếng nói

Hình 5.1 là cấu trúc nguyên lý của một hệ thống nhận dạng tiếng nói. Tín hiệu tiếng nói trước hết được xử lý bằng cách áp dụng một trong các phương pháp phân tích phổ ngắn hạn hay còn được gọi là quá trình trích chọn đặc trưng hoặc quá trình tiền xử lý (front-end processing). Kết quả thu được sau quá trình trích chọn đặc trưng là tập các đặc trưng âm học (acoustic features) được tạo dựng thành một véc-tơ. Thông thường khoảng 100 véc-tơ đặc trưng âm học được tạo ra tại đầu ra của quá trình phân tích trong một đơn vị thời gian một giây.



Hình 5.1 Cấu trúc tổng quát của một hệ thống nhận dạng tiếng nói

Việc so sánh (matching) trước hết thực hiện bằng việc huấn luyện xây dựng các đặc trưng, sau đó sử dụng để so sánh với các tham số đầu vào để thực hiện việc nhận dạng. Trong quá trình huấn luyện hệ thống dòng véc-tơ các đặc trưng được đưa vào hệ thống để ước lượng các tham số của các mẫu tham khảo (reference patterns). Một mẫu tham khảo có thể mô phỏng (model) một từ, một âm đơn (a single phoneme) hoặc một đơn vị tiếng nói nào đó (some other speech unit). Tùy thuộc vào nhiệm vụ của hệ thống nhận dạng, quá trình huấn luyện hệ thống sẽ bao gồm một quá trình xử lý nhiều ít phức tạp. Chẳng hạn với hệ thống nhận dạng phụ thuộc người nói (speaker dependent recognition), có thể chỉ bao gồm một vài hoặc duy nhất

một biểu diễn (utterances) cho mỗi từ cần được huấn luyện. Tuy nhiên, đối với hệ thống nhận dạng độc lập với người nói, có thể bao gồm hàng ngàn biểu diễn tương ứng với tín hiệu của mẫu tham khảo mong muốn. Những biểu diễn này thường là bộ phận (part) của một cơ sở dữ liệu tiếng nói đã được thu thập trước đây. Cần chú ý rằng việc trích chọn các đặc trưng tiêu biểu (representative features) và xây dựng một mô hình tham khảo (a reference model) là một quá trình tốn thời gian và là một công việc phức tạp.

Trong quá trình nhận dạng, dãy các véc-tơ đặc trưng được đem so sánh với các mẫu tham khảo. Sau đó, hệ thống tính toán độ tương đồng (likelihood - độ giống nhau) của dãy véc-tơ đặc trưng và mẫu tham khảo hoặc chuỗi mẫu tham khảo. Việc tính toán độ giống nhau thường được tính toán bằng cách áp dụng các thuật toán hiệu quả chẳng hạn như thuật toán Viterbi. Mẫu hoặc dãy mẫu có độ tương đồng (likelihood) cao nhất được cho là kết quả của quá trình nhận dạng.

Hiện nay, các phương pháp trích chọn đặc trưng phổ biến thường là các mạch lọc Mel (Mel filterbank) kết hợp với các biến đổi phổ Mel sang miền cepstral. Chúng ta sẽ tìm hiểu sơ đồ tiền xử lý được tiêu chuẩn hóa như một phương pháp tiền xử lý bởi ETSI. Mô hình mẫu tham chiếu thường là các mô hình Markov ẩn (HMMs).

5.5. Các phương pháp phân tích cho nhận dạng tiếng nói

5.5.1 Lượng tử hóa véc-tơ

Chúng ta thấy rằng, kết quả của các phép phân tích trích chọn tham số là dãy các véc-tơ đặc trưng của đặc tính phổ thay đổi theo thời gian của tín hiệu tiếng nói. Để thuận tiện, chúng ta kí hiệu các véc-tơ phổ là v_l, l=1,2,…, L, trong đó mỗi véc-tơ thường là một véc-tơ có chiều dài p. Nếu chúng ta so sánh tốc độ thông tin của các biểu diễn véc-tơ và các biểu diễn trực tiếp dạng sóng tín hiệu (uncoded speech waveform), chúng ta thấy rằng các phân tích phổ cho phép chúng ta giảm nhỏ đi rất nhiều tốc độ thông tin yêu cầu. Lấy ví dụ, với tín hiệu tiếng nói được lấy mẫu với tần số lấy mẫu 10kHz, và sử dụng 16bít để biểu diễn biên độ của mỗi mẫu. Khi đó biểu diễn raw cần 160000bps để lưu trữ các mẫu tín hiệu. Trong khi đó, đối với phân tích phổ, giả sử chúng ta sử dụng các véc-tơ có độ dài p=10 và sử dụng 100 véc-tơ phổ trong một đơn vị thời gian một giây. Và chúng ta cũng sử dụng độ chính xác 16 bít để biểu diễn mỗi thành phần phổ, khi đó chúng ta cần 100x10x16bps hay 16000bps để lưu trữ. Như vậy phương pháp phân tích phổ cho phép giảm đi 10 lần. Tỷ lệ giảm này là cực kỳ quan trọng trong việc lưu trữ. Dựa trên khái niệm cần tối thiểu chỉ một biểu diễn phổ đơn lẻ cho mỗi đơn vị tiếng nói, chúng ta có thể làm giảm nhỏ thêm nữa các biểu diễn phổ raw của tín hiệu thành các thành phần từ một tập nhỏ hữu hạn các véc-tơ phổ duy nhất mà mỗi thành phần tương ứng với một đơn vị cơ bản của tín hiệu tiếng nói (tức là các phoneme). Lẽ tất nhiên, một biểu diễn lý tưởng là khó có thể đạt được trong thực tế bởi vì có quá nhiều các biến số trong các tính chất phổ của mỗi một đơn vị tín hiệu tiếng nói cơ bản. Tuy nhiên, khái niệm về việc xây dựng một bộ mã (codebook) gồm các véc-tơ phân tích phân biệt, mặc dù có số từ mã nhiều hơn tập cơ bản các phoneme, vẫn là một ý tưởng hấp dẫn và là ý tưởng cơ bản nằm trong một loạt các kỹ thuật phân tích được gọi chung là các phương pháp lượng tử hóa véc-tơ. Dựa trên các suy luận trên, giả sử chúng ta cần một bộ mã với khoảng 1024 véc-tơ phổ độc nhất (tức là khoảng 25 dạng khác nhau của mỗi tập 40 đơn vị tín hiệu tiếng nói cơ bản). Như thế, để biểu diễn một véc-tơ phổ bất kỳ, tất cả chúng ta cần là một số 10 bít - khi đó chỉ số của véc-tơ bộ mã phù hợp nhất với véc-tơ vào. Giả sử rằng ở tốc độ 100 véc-tơ phổ trong một đơn vị thời gian một

giây, chúng ta cần tổng tốc độ bít vào khoảng 1000bps để biểu diễn các véc-tơ phổ của tín hiệu. Ta thấy rằng, tốc độ này chỉ bằng khoảng 1/16 tốc độ cần thiết của các véc-tơ phổ liên tục. Do đó, phương pháp biểu diễn lượng tử hóa véc-tơ là một phương pháp có khả năng biểu diễn cực kỳ hiệu quả các thông tin phổ của tín hiệu tiếng nói.

Trước khi thảo luận các khái niệm liên quan đến việc thiết kế và thực hiện một hệ lượng tử véc-tơ thực tế, chúng ta điểm lại các ưu điểm và nhược điểm của phương pháp này. Trước hết, các ưu điểm chính của phương pháp biểu diễn lượng tử véc-tơ bao gồm:

Cho phép giảm nhỏ việc lưu trữ thông tin phân tích phổ tín hiệu. Điều này cho phép tạo thuận lợi cho việc áp dụng trong các hệ thống nhận dạng tín hiệu tiếng nói thực tế.

Cho phép giảm nhỏ việc tính toán để xác định sự giống nhau (tương đồng - similarity) của các véc-tơ phân tích phổ. Chúng ta biết rằng, trong phép nhận dạng tín hiệu tiếng nói, một bước quan trọng trong việc tính toán là quyết định tương đồng phổ của một cặp véc-tơ. Dựa trên biểu diễn lượng tử hóa véc-tơ, việc tính toán tính tương đồng phổ tín hiệu thường được giảm xuống thành một phép tra bảng của sự giống nhau giữa các cặp véc-tơ mã.

Cho phép biểu diễn rời rạc tín hiệu âm thanh tiếng nói. Bằng việc gắn một nhãn phonetic (hoặc có thể là một tập các nhãn phonetic hoặc một lớp phonetic) với một véc-tơ mã, quá trình chọn ra một véc-tơ mã biểu diễn một véc-tơ phổ cho trước phù hợp nhất trở thành việc gán một nhãn phonetic cho mỗi khung phổ của tín hiêu. Một loạt các hệ thống nhân dạng tiếng nói tồn tại đã sử dụng những nhãn này để cho phép nhận dạng một cách hiệu quả.

Tuy vậy cũng phải kể đến một số hạn chế của việc sử dụng bộ mã lượng tử hóa véc-tơ để biểu diễn các véc-tơ phổ tín hiệu tiếng nói. Chúng bao gồm:

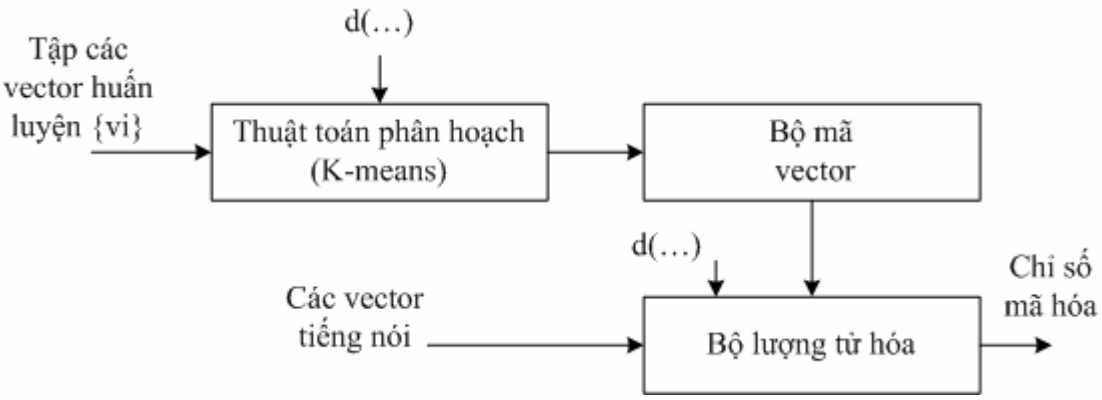
Tồn tại sự méo phổ kế thừa (inherent) trong việc biểu diễn véc-tơ phân tích thực tế. Do chỉ có số lượng hữu hạn véc-tơ mã, quá trình chọn véc-tơ thích hợp nhất biểu diễn một véc-tơ phổ cho trước tương tự như quá trình lượng tử một véc-tơ và kết quả là dẫn đến một sai số lượng tử nào đó. Sai số lượng tử giảm khi số lượng các véc-tơ mã tăng. Tuy nhiên, với mỗi bộ mã có số véc-tơ mã hữu hạn thì luôn tồn tại một mức sai số lượng tử.

Dung lượng lưu trữ cho các véc-tơ mã thường là không bất thường (nontrivial). Nếu bộ mã càng lớn, nghĩa là để càng giảm nhỏ sai số lượng tử, thì dung lượng lưu trữ các thành phần bộ véc-tơ mã yêu cầu càng cao. Với các bộ mã có kích thước lớn hơn hoặc bằng 1000, thì dung lượng lưu trữ thường là không bất thường. Như vậy có một sự mâu thuẫn giữa sai số lượng tử, quá trình lựa chọn véc-tơ mã, và dung lượng lưu trữ các véc-tơ mã. Trong các thiết kế ứng dụng thực tế cần phải cân bằng ba yếu tố này.

a) Sơ đồ thực hiện lượng tử hóa véc-tơ

Sơ đồ khối của cấu trúc phân loại (classification) và huấn luyện sử dụng lượng tử hóa véc-tơ cơ bản được trình bày trong hình 5.2. Một tập lớn các véc-tơ phân tích phổ v_1, v_2, …, v_L tạo thành tập các véc-tơ dùng để huấn luyện. Tập các véc-tơ này dùng để tạo ta một tập tối ưu các véc-tơ mã để biểu diễn các biến phổ quan sát được trong tập huấn luyện. Nếu chúng ta ký hiệu kích cỡ của bộ mã lượng tử hóa véc-tơ là $M=2^B$ (chúng ta gọi đây là một bộ mã B-bít), khi đó chúng ta cần có L>> M để có thể tìm được một tập gồm M véc-tơ phù hợp nhất. Trong thực tế, người ta thấy rằng, để quá trình huấn luyện bộ mã lượng tử véc-tơ hoạt động tốt, L thường phải tối thiểu bằng 10M. Tiếp đến là quá trình đo lường độ giống nhau hay còn gọi là khoảng cách giữa các cặp véc-tơ phân tích phổ nhằm để có thể phân hoạch (cluster) tập các

véc-tơ huấn luyện cũng như gắn hoặc phân loại các véc-tơ phổ thành các thành phần của bộ mã duy nhất. Khoảng cách phổ giữa hai véc-tơ phổ $\mathbf{v}_i$ và $\mathbf{v}_j$ được ký hiệu là $d_{ij}=d(\mathbf{v}_i, \mathbf{v}_j)$. Quá trình tiếp tục phân loại tập L véc-tơ huấn luyện thành M phân hoạch và chúng ta chọn M véc-tơ mã như là tập trung tâm (centroid) của mỗi một phân hoạch đó. Thủ tục phân loại các véc-tơ phân tích phổ tín hiệu tiếng nói xác định thực hiện việc chọn véc-tơ mã gần nhất với véc-tơ nhập vào và sử dụng chỉ số mã như là kết quả biểu diễn phổ. Quá trình này thường được gọi là việc tìm kiếm lân cận gần nhất hoặc thủ tục mã hóa tối ưu. Thủ tục phân loại về cơ bản là một bộ lượng tử hóa với đầu vào là một véc-tơ phổ tín hiệu tiếng nói và đầu ra là chỉ số mã hóa của một véc-tơ mã mà gần giống với đầu vào nhất (best match)



Hình 5.2 Mô hình sử dụng véc-tơ lượng tử huấn luyện và phân loại

b) Tập huấn luyện bộ lượng tử hóa véc-tơ

Để có thể huấn luyện bộ mã lượng tử hóa véc-tơ một cách chính xác, các véc-tơ thuộc tập huấn luyện phải bao phủ (span) các khía cạnh mong muốn như sau:

Người nói, bao gồm các nhóm (ranges) về tuổi tác, trọng âm (accent), giới tính, tốc độ nói, các mức độ và các biến số khác.

Các điều môi trường chẳng hạn như phòng yên lặng hay trên ô-tô (automobile), hoặc khu làm việc ồn ào (noisy workstation).

Các bộ chuyển đổi (transducers) và các hệ thống truyền dẫn, bao gồm cả các mi-cờ-rô băng thông rộng, các ống nghe (handset) điện thoại (với các mi-cờ-rô các-bon và điện than), các truyền dẫn trực tiếp, kênh tín hiệu điện thoại, kênh băng thông rộng, và các thiết bị khác.

Các đơn vị tiếng nói bao gồm các từ vựng sử dụng nhận dạng đặc biệt (chẳng hạn các chữ số) và tiếng nói liên tục (conversational speech)

Mục tiêu huấn luyện càng hẹp càng rõ ràng (chẳng hạn với số lượng người nói hạn chế, tiếng nói trong phòng yên lặng, ...) thì sai số lượng tử khi sử dụng việc biểu diễn phổ tín hiệu với bộ mã kích thước cố định càng nhỏ. Tuy nhiên để có thể ứng dụng giải quyết nhiều loại bài toán thực tế, tập huấn luyện phải càng lớn càng tốt.

c) Đo lường sự tương đồng hay khoảng cách

Khoảng cách phổ giữa các véc-tơ phổ $\mathbf{v}_i$ và v_j được định nghĩa như sau:

$$d\left(v_i, v_j\right) = d_{ij} = \begin{cases} 0 & v_i = v_j \\ > 0 & v_i \neq v_j \end{cases} \tag{5.1}$$

d) Phân hoạch các véc-tơ huấn luyện

Thủ tục phân hoạch tập L véc-tơ huấn luyện thành một tập gồm M bộ véc-tơ mã có thể được mô tả như sau:

Bắt đầu: Chọn M véc-tơ bất kỳ từ tập L véc-tơ huấn luyện tạo thành một tập khởi đầu các từ mã của bộ mã.

Tìm kiếm lân cận gần nhất: Với mỗi véc-tơ huấn luyện, tìm một véc-tơ mã trong bộ đang xét gần nhất (theo nghĩa khoảng cách phổ) và gán véc-tơ đó vào ô tương ứng.

Cập nhật centroid: Cập nhật từ mã trong mỗi ô bằng cách sử dụng centroid của các véc-tơ huấn luyện trong các ô đó.

Lặp: Lặp lại các bước 2 và 3 cho đến khi khoảng cách trung bình nhỏ hơn một khoảng ngưỡng định sẵn.

e) Thủ tục phân loại véc-tơ

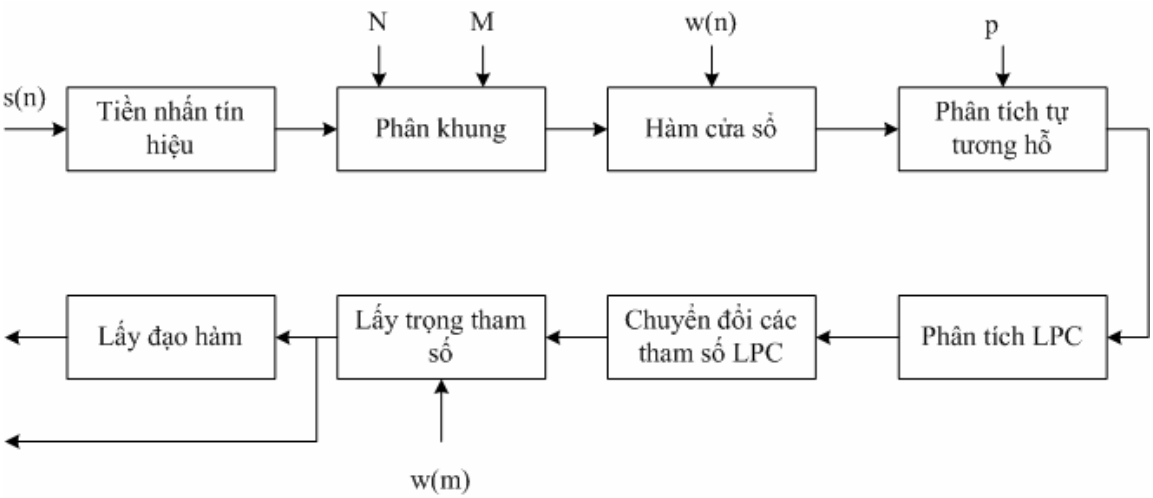
Việc phân loại các véc-tơ đối với các véc-tơ phổ bất kỳ về cơ bản là việc tìm hết trong bộ mã để tìm ra được một véc-tơ tương đồng nhất. Chúng ta ký hiệu bộ véc-tơ mã của một bộ mã M véc-tơ là $\mathbf{y}_m$, ($1 \leq m \leq M$) và véc-tơ phổ cần phân loại (và lượng tự hóa) là **v**, khi đó chỉ số m^* của từ mã phù hợp nhất được xác định như sau:

$$m^* = \arg \min_{1 \leq m \leq M} d(v, y_m) \tag{5.2}$$

Với các bộ mã có giá trị M lớn (chẳng hạn $M \geq 1024$), việc tính toán theo công thức (5.2) sẽ trở lên phức tạp (be excessive), và phụ thuộc vào tính toán chi tiết của quá trình đo lường khoảng cách phổ. Trong thực tế, người ta thường sử dụng các thuật giải cận tối ưu (sub-optimal) để tìm kiếm.

5.5.2 Bộ xử lý LPC trong nhận dạng tiếng nói

Trong phần trước chúng ta thảo luận về các tính chất chung nhất của phương pháp phân tích LPC. Trong phần này chúng ta sẽ mô tả chi tiết việc sử dụng bộ xử lý LPC cho các hệ thống nhận dạng tín hiệu tiếng nói. Sơ đồ khối của khối xử lý LPC được trình bày trong hình 5.3. Các bước cơ bản trong quá trình xử lý của bộ xử lý như sau:



Hình 5.3 Sơ đồ khối bộ xử lý LPC trong nhận dạng tiếng nói

a) Tiền nhấn tín hiệu

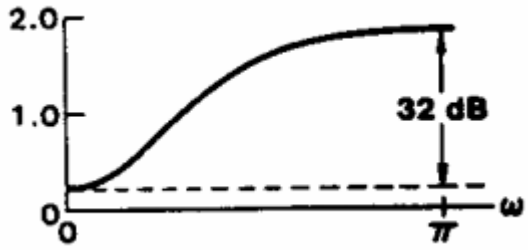
Đầu tiên tín hiệu tiếng nói dạng số hóa s(n) được đưa qua một hệ thống lọc số bậc thấp, thường là bộ lọc đáp ứng xung hữu hạn (FIR) bậc nhất, nhằm làm phẳng phổ tín hiệu. Điều này sẽ giúp cho tín hiệu ít bị ảnh hưởng của các phép biến đổi xử lý tín hiệu có độ chính xác hữu hạn trong suốt quá trình sau đó. Bộ lọc số sử dụng cho việc tiền nhấn tín hiệu có thể là một bộ lọc với các tham số cố định hoặc có thể là một bộ lọc thích nghi có các tham số thay đổi chậm. Trong xử lý tín hiệu tiếng nói, người ta thường dùng một hệ thống mạch lọc bậc nhất có các tham số cố định có dạng:

$$H(z) = 1 - \tilde{a}z^{-1} \quad (0,9 \le \tilde{a} \le 1,0) \tag{5.3}$$

Khi đó, tín hiệu đầu ra của bộ tiền nhấn $\tilde{s}(n)$ có thể tính như sau:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \tag{5.4}$$

Giá trị phổ biến của hệ số cố định $\tilde{a}$ là khoảng 0,95 (trong các ứng dụng thực thi với dấu phẩy tĩnh giá trị của $\tilde{a}$ thường được chọn là 15/16=0.9375). Hình 5.4 biểu diễn biên độ đặc tính hàm truyền đạt $H(e^{j\omega})$ với giá trị $\tilde{a} = 0,95$. Từ hình vẽ, chúng ta có thể quan sát thấy rằng tại $\omega = \pi$, tức là bằng một nửa tốc độ lấy mẫu, có sự gia tăng (boost) biên độ khoảng 32dB so với biên độ ở tần số $\omega = 0$.



Hình 5.4 Phổ biên độ của mạch tiền nhấn tín hiệu

Trong trường hợp mạch lọc thích nghi được sử dụng, hàm truyền đạt của nó thường có dạng:

$$H(z) = 1 - \tilde{a}_n z^{-1} \tag{5.5}$$

Trong đó $\tilde{a}_n$ thay đổi theo thời gian n theo một tiêu chí thích nghi được thiết kế trước. Một giá trị điển hình thường được sử dụng là $\tilde{a} = r_n(1) / r_n(0)$.

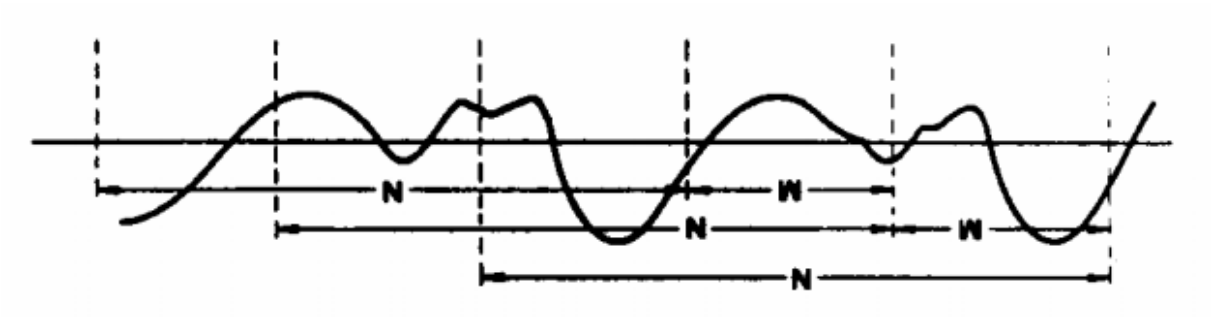
b) Phân khung tín hiệu

Kết quả tín hiệu sau khối tiền nhấn tín hiệu là một khung tín hiệu $\tilde{s}(n)$ gồm các khung có N mẫu, trong đó các khung cạnh nhau cách biệt nhau M mẫu. Hình 5.5 mô tả các khung tín hiệu trong trường hợp M=N/3. Ta thấy, khung thứ nhất gồm N mẫu, khung thứ hai bắt đầu sau khung thứ nhất M mẫu và có chung N-M mẫu với khung thứ nhất. Tương tự như vậy, khung thứ 3 bắt đầu sau khung thứ nhất 2M mẫu hay bắt đầu sau khung thứ hai M mẫu và có chung với khung thứ nhất và thứ hai tương ứng là N-2M và N-M mẫu. Quá trình này được tiếp tục cho đến khi toàn bộ tín hiệu của một hoặc một số khung được phân khung xong. Dễ dàng thấy rằng, nếu M≤N thì các khung cạnh nhau sẽ có sự bao trùm lẫn nhau, và kết quả là

các ước lượng phổ của LPC sẽ có sự tương quan giữa các khung; nếu M< <N thì các ước lượng phổ LPC giữa các khung sẽ tương đối trơn tru (smooth). Mặt khác, nếu M>N, khi đó sẽ không có sự bao trùm lẫn nhau giữa các khung; trong thực tế khi đó một phận tín hiệu sẽ bị mất hoàn toàn (tức là không xuất hiện trong bất cứ một khung phân tích nào), và khi đó tính tương hỗ giữa các ước lượng phổ LPC thu được của các khung cạnh nhau sẽ chứa một thành phần nhiễu mà biên độ của nó tăng khi M tăng (tức là khi số lượng mẫu tín hiệu bị bỏ qua càng nhiều). Đây là trường hợp không thể chấp nhận được (intolerable) trong bất cứ phép phân tích LPC nào sử dụng cho hệ thống nhận dạng tín hiệu tiếng nói. Gọi khung tín hiệu thứ l là $x_l(n)$ và giả sử có toàn bộ L khung tín hiệu, khi đó:

$$x_l(n) = \tilde{s}(Ml+n) \quad n = 0,1,...,N-1;\ l = 0,1,...,L-1 \tag{5.6}$$

Điều này có nghĩa là khung tín hiệu đầu tiên $x_0(n)$ bao gồm các mẫu $\tilde{s}(0)$, $\tilde{s}(1)$, …, $\tilde{s}(L-1)$; khung tín hiệu thứ hai $x_l(n)$ bao gồm các mẫu $\tilde{s}(M)$, $\tilde{s}(M+1)$, …, $\tilde{s}(M+N-1)$; và khung tín hiệu thứ L bao gồm các mẫu $\tilde{s}(M(L-1))$, $\tilde{s}(M(L-1)+1)$, …, $\tilde{s}(M(L-1)+N-1)$. Đối với tín hiệu tiếng nói có tốc độ lấy mẫu 6.67kHz thì giá trị của N và M thường được chọn tương ứng là 300 và 100, nghĩa là tương ứng với các khung 45 mili-giây và khoảng cách giữa các khung là 15mili-giây.



Hình 5.5 Phân khung tín hiệu trong phân tích LPC cho nhận dạng tiếng nói

c) Lấy cửa sổ tín hiệu

Bước tiếp theo trong quá trình xử lý phân tích LPC là việc lấy cửa sổ của các khung tín hiệu riêng rẽ nhằm mục đích giảm nhỏ sự không liên tục của tín hiệu ở phần đầu và cuối mỗi khung. Điều nãy cũng tương tự như đã đề cập trong phần giới thiệu chung khi xem xét trong miền tần số: việc lấy cửa sổ tín hiệu nhằm mục đích cắt bỏ tín hiệu về 0 ở phần bắt đầu và kết thúc của mỗi khung. Giả sử hàm cửa sổ được định nghĩa là w(n) (0≤n≤N-1), khi đó kết quả tín hiệu thu được sau khi lấy cửa sổ là:

$$\tilde{x}_l(n) = x_l(n)\,\mathrm{w}(n) \quad 0 \le n \le N-1 \tag{5.7}$$

Hàm cửa sổ phổ biến dùng cho phương pháp tự tương quan trong LPC sử dụng trong các hệ thống nhân dạng tiếng nói là hàm cửa sổ Hamming, trong đó biểu thức hàm được cho bởi:

$$\mathrm{w}(n) = 0,54 - 0,46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \tag{5.8}$$

d) Phân tích tính tự tương quan

Kết quả tự tương quan của mỗi khung tín hiệu sau phép lấy cửa sổ là:

$$\Phi_l(n) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\,\tilde{x}_l(n+m) \qquad m = 0,1,\ldots,p \tag{5.9}$$

Trong đó, giá trị tự tương quan cao nhất p là bậc của phân tích LPC. Thông thường, p được chọn từ 8 đến 16. Cần chú ý đến một lợi ích phụ của việc sử dụng phương pháp tự tương quan là thành phần tự tương quan bậc 0, tức là $\Phi_l(0)$, chính là năng lượng của khung thứ l. Năng lượng của khung tín hiệu là một tham số quan trọng trong các hệ thống phát hiện tín hiệu tiếng nói.

e) Phân tích LPC

Bước tiếp theo trong quá trình phân tích là phép phân tích LPC, trong đó mỗi khu của p+1 tham số tự tương quan được chuyển đổi thành một tập các tham số LPC. Tập các tham số LPC có thể là tập các hệ số LPC, hoặc tập các hệ số phản ánh, hoặc các hệ số tỉ lệ log, hoặc các hệ số cepstral, hoặc bất cứ biến đổi mong muốn nào đó từ các tập nêu trên. Việc thực hiện biến đổi này thường được thực hiện bằng cách áp dụng phương pháp Durbin được diễn giải như sau. Để thuận tiện, chúng ta tạm bỏ chỉ số l trong biểu thức $r_l(m)$.

$$E^{(0)} = \Phi_l(0) \tag{5.10}$$

$$k_i = \frac{\{\Phi_l(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} \Phi_l(|i-j|)\}}{E^{(i-1)}} \qquad (1 \le i \le p) \tag{5.11}$$

$$\alpha_i^{(i)} = k_i \tag{5.12}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \tag{5.13}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{5.14}$$

Trong công thức tính tổng của công thức thư hai ở trên, (5.11), chúng ta bỏ qua trường hợp i=1. Hệ các phương trình trên dược giải theo phương pháp truy hồi với i=1,2,…, p và kế quả cuối cùng thu được là:

$$a_m = \alpha_m^{(p)} \qquad (1 \le m \le p) \tag{5.15}$$

$$k_m = R_{\text{coef}} \tag{5.16}$$

$$g_m = \log\left(\frac{1-k_m}{1+k_m}\right) \tag{5.17}$$

(5.15) là các hệ số LPC, (5.16) là các hệ số phản xạ, và (5.17) là lô-ga-rít các hệ số tỷ lệ diện tích.

f) Chuyển đổi các tham số LPC sang các hệ số Cepstral

Một tập tham số quan trọng có thể xây dựng trực tiếp từ tập các tham số LPC là tập các hệ số cepstral LPC. Công thức xác định sử dụng phép đệ quy được cho như sau:

$$c_0 = \ln(\sigma^2) \tag{5.18}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \qquad (1 \le m \le p) \tag{5.19}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \qquad (m > p) \tag{5.20}$$

Ở đây, σ^2 là độ lợi của việc sử dụng mô hình LPC. Các hệ số cepstral chính là các hệ số tương ứng của biến đổi Fourier của các giá trị lô-ga-rít của biên độ phổ. Tập các hệ số cepstral được chứng minh rằng là một tập các đặc trưng đáng tin cậy và robust hơn tập các hệ số LPC, hay tập các hệ số phản xạ cũng như tập các hệ số tỉ lệ log diện tích trong việc nhận dạng tín hiệu tiếng nói. Thường một biểu diễn gồm Q>p hệ số cepstral được sử dụng, trong đó phổ biển Q≈3p/2.

g) Lấy trọng các tham số - Parameter Weighting

Trong các hệ số cepstral, các hệ số bậc thấp rất nhạy cảm với độ dốc (slope) của toàn dải phổ, trong khi đó các hệ số bậc cao thì lại rất nhạy cảm với nhiễu. Chính vì lý do này, nó dường như trở thành một tiêu chuẩn của các phép xử lý là sử dụng lấy trọng số các hệ số cepstral bằng một hàm cửa sổ nhằm giảm nhỏ các nhạy cảm nói trên. Một cách thông thường cho việc thay đổi việc sử dụng một cửa sổ cepstral là xem xét biểu diễn Fourier của lô-ga-rít phổ biên độ và các đạo hàm lô-ga-rít của phổ biên độ. Nghĩa là:

$$\log \left| S\left(e^{j\omega} \right) \right| = \sum_{m=-\infty}^{\infty} c_m e^{j\omega m} \tag{5.21}$$

$$\frac{\partial}{\partial \omega} \left[\log \left| S\left(e^{j\omega} \right) \right| \right] = \sum_{m=-\infty}^{\infty} (-jm) c_m e^{-j\omega m} \tag{5.22}$$

Thành phần vi phân của lô-ga-rit phổ biên độ có một tính chất đặc biệt là bất cứ độ dốc phổ cố định nào trong lô-ga-rít biên độ phổ sẽ trở thành một hằng số. Hơn nữa, bất cứ thành phần đỉnh phổ nào trong lô-ga-rít biên độ phổ, tức là các formant, đều được bảo đảm giữ nguyên trong vi phân của lô-ga-rít biên độ phổ. Do đó, bằng việc nhân biểu diễn vi phân của lô-ga-rít biên độ phổ với -jm, chúng ta đã thực hiện việc thay đổi trọng các tham số. Kết quả chúng ta có:

$$\frac{\partial}{\partial \omega} \left[\log \left| S\left(e^{j\omega} \right) \right| \right] = \sum_{m=-\infty}^{\infty} \hat{c}_m e^{-j\omega m} \tag{5.23}$$

Trong đó:

$$\hat{c}_m = c_m (-jm) \tag{5.24}$$

Để có thể đạt được tính robustness cho các giá trị m lớn, tức là các trọng số nhỏ ở gần m=Q, và có thể cắt bỏ được phần tính toán vô định trong công thức (5.23), chúng ta cần phải đưa ra một dạng tổng quát hơn đối với các hệ số trọng số:

$$\hat{c}_m = \mathrm{w}_m c_m \tag{5.25}$$

Một phép lấy trọng số thích hợp chính là một bộ lọc thông dải (bộ lọc trong miền cepstral) có dạng:

$$\mathrm{w}_m = \left[1 + \frac{Q}{2} \sin \left(\frac{\pi m}{Q} \right) \right] \qquad (1 \le m \le Q) \tag{5.26}$$

Hàm tính toán trọng số cho ở công thức (5.26) có khả năng cắt bỏ phần tính toán vô hạn và giải nhấn (de-emphasizes) các hệ số c_m xung quan m=1 và m=Q.

h) Các đạo hàm Cepstral

Các biểu diễn cepstral của phổ tín hiệu tiếng nói là một biểu diễn thích hợp cho phép đặc tả được các tính chất phổ cục bộ của tín hiệu trong một khung tín hiệu phân tích xác định. Tuy nhiên có thể tăng chất lượng của các biểu diễn này bằng các mở rộng các phân tích bao gồm các thông tin về đạo hàm của cepstral theo thời gian (the temporal cepstral derivative). Thực tế cho thấy rằng cả các đạo hàm cấp một và cấp hai đều mang lại khả năng làm gia tăng chất lượng hoạt động của hệ thống nhận dạng tín hiệu tiếng nói. Để đưa khái niệm thời gian vào các biểu diễn cepstral, chúng ta kí hiệu hệ số cepstral thứ m ở thời điểm t là $c_m(t)$. Trong thực tế, thời điểm lấy mẫu t gắn với khung tín hiệu phân tích chứ không phải là một thời điểm bất kỳ. Việc tính đạo hàm các hệ số cepstral theo thời gian được thực hiện một các xấp xỉ như sau: Đạo hàm theo thời gian của lô-ga-rít biên độ phổ có biểu diễn chuỗi Fourier tương ứng:

$$\frac{\partial}{\partial t}\left[\log\left|S\left(e^{j\omega},t\right)\right|\right]=\sum_{m=-\infty}^{\infty}\frac{\partial c_m(t)}{\partial t}e^{-j\omega m} \tag{5.27}$$

Do đó, đạo hàm cepstral theo thời gian cũng sẽ được xác định một cách tương tự. Vì $c_m(t)$ là một biểu diễn thời gian rời rạc (trong đó t là chỉ số khung tín hiệu), chúng ta không thể áp dụng trực tiếp các vi phân cấp một và cấp hai để xấp xỉ với các đạo hàm (vì điều này dẫn đến kết quả nhiễu rất lớn it is very noisy). Do đó, một các tính toán hợp lý là xấp xỉ $\partial c_m(t)/\partial t$ bởi một đa thức nội suy trực giao gần đúng (an orthogonal polynomial fit), một ước lượng bình phương tối thiểu của các đạo hàm (a least-squared estimate of the derivative), trên toàn khoảng cửa sổ hữu hạn. Nghĩa là:

$$\frac{\partial c_m(t)}{\partial t}=\Delta c_m(t)\approx\mu\sum_{k=-K}^{K}kc_m(t+k) \tag{5.28}$$

Trong đó, μ là một hằng số chuẩn hóa thích hợp và (2K+1) là số khung tín hiệu mà trên đó chúng ta thực hiện việc tính toán. Thông thường, giá trị của K thường được lấy bằng 3 và thấy rằng giá trị này thích hợp cho việc tính toán các đạo hàm cấp một. Từ thủ tục tính toán ở trên, với mỗi khung tín hiệu t, kết quả của phép phân tích LPC là một véc-tơ gồm Q hệ số cepstral đã được kể đến trọng và một véc-tơ mở rộng của Q thành phần đạo hàm theo thời gian được kí hiệu là:

$$o'_t=\left(\hat{c}_1(t),\hat{c}_2(t),...,\hat{c}_Q(t),\Delta\hat{c}_1(t),\Delta\hat{c}_2(t),...,\Delta\hat{c}_Q(t)\right) \tag{5.29}$$

Trong công thức (5.29), o'_t là một véc-tơ gồm 2Q thành phần và (.)' biểu diễn phép chuyển vị ma trận.

Một cách tương tự, nếu chúng ta thực hiện việc tính toán các đạo hàm cấp hai $\Delta^2c_m(t)$ và thêm các giá trị này vào véc-tơ o_t ta sẽ thu được một véc-tơ mới gồm 3Q thành phần.

i) Bảng các giá trị phổ biến của các tham số trong phân tích LPC

Trong các phân tích tính toán theo phương pháp phân tích LPC, chúng ta thấy rằng các tính toán phụ thuộc vào số lượng các tham số biến số bao gồm: số mẫu trong khung tín hiệu phân tích N, số mẫu phân cách điểm bắt đầu của các khung liền kề M, bậc của phân tích LPC p, kích cỡ của véc-tơ cepstral được xây dựng Q, số lượng khung K mà trên đó các đạo hàm theo thời gian của các hệ số cepstral được tính toán. Mặc dù mỗi một giá trị của các tham số

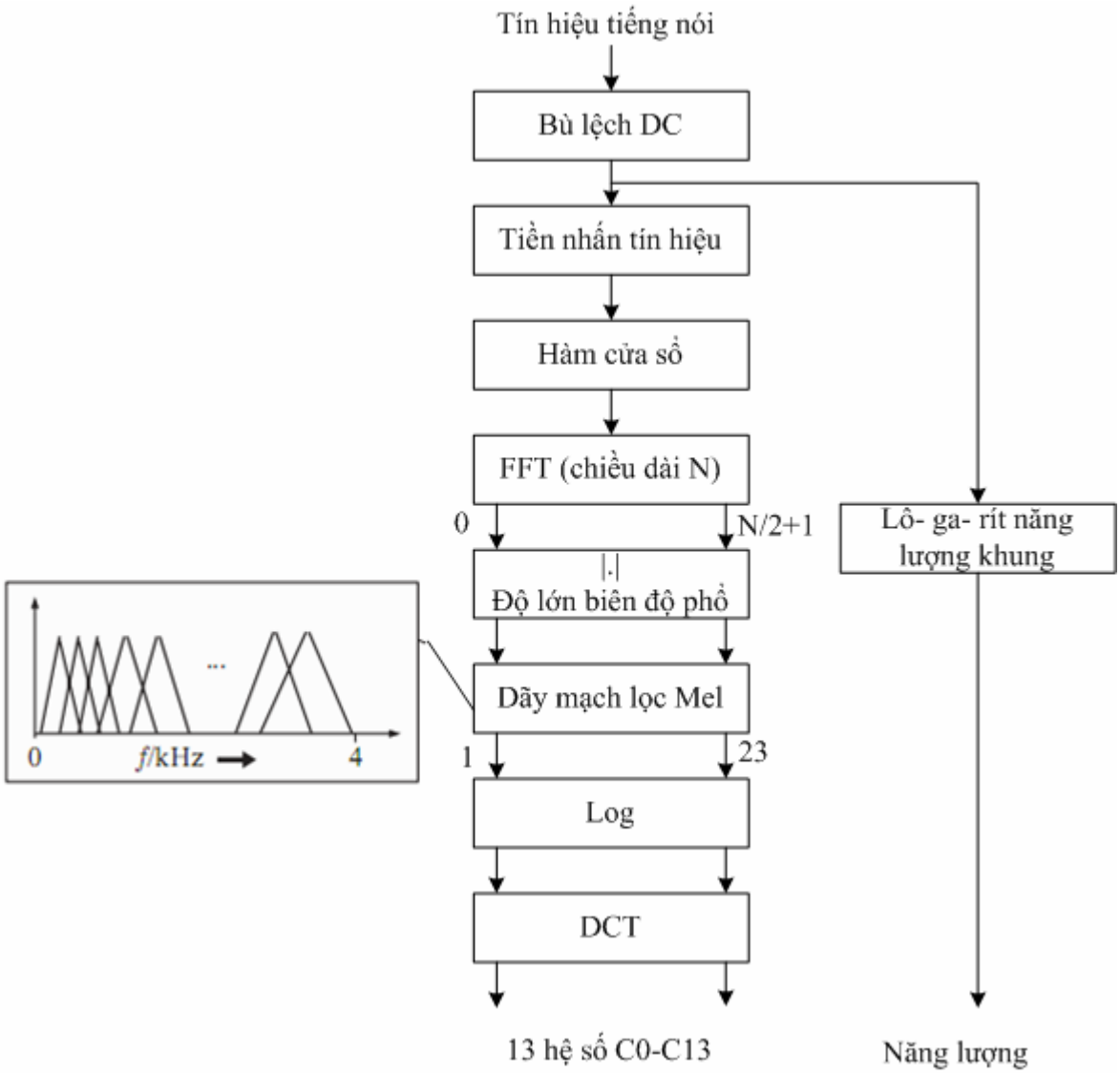
vừa kể thay đổi trên một dải rất lớn phụ thuộc vào các hệ thống cụ thể, một số giá trị phổ biến đối với ba tần số lấy mẫu tương ứng là 6,67kHz, 8kHz và 10kHz được cho trong bảng sau.

Giá trị tham số	F_s=6,67kHz	F_s=8kHz	F_s=10kHz
N	300 (45ms)	240 (30ms)	300 (30ms)
M	100 (15ms)	80 (10ms)	100 (10ms)
p	8	10	10
Q	12	12	12
K	3	3	3

Bảng 5.2: Một số giá trị tham số phổ biến của phép phân tích LPC

5.5.3 Phân tích MFCC trong nhận dạng tiếng nói

Sơ đồ khối phương pháp phân tích cepstral tần số Mel (Mel frequency Cepstral analysis) dùng để trích chọn đặc trưng tín hiệu tiếng nói được trình bày trong hình 5.6. Đây là một kỹ thuật phổ biến đại diện cho lớp phương pháp trích chọn đặc trưng có tên gọi là MFCCs (Mel frequency cepstral coefficients). Đầu tiên, tín hiệu tiếng nói được lọc bởi một mạch lọc thông cao (high-pass filter) với tần số cắt (cut-off frequency) rất thấp nhằm loại bỏ thành phần tín hiệu một chiều mà có thể do bộ chuyển đổi ADC tạo ra. Đặc biệt việc lọc này là cần thiết để tăng tính chính xác khi thực hiện tính toán năng lượng tín hiệu theo khung trong các phân tích ngắn hạn. Năng lượng tín hiệu cũng như các tham số cepstral được tính đối với mọi khung cửa sổ dịch với khoảng dịch d_{shift}=10ms. Do việc cảm nhận âm thanh của con người theo thang không tuyến tính nên việc tính năng lượng tín hiệu thường là dùng thang lô-ga-rít. Năng lượng khung theo lô-ga-rít (logarithmic frame energy) được sử dụng như một thành phần của véc-tơ đặc trưng tín hiệu. Sau đó một mạch lọc thông cao khác được sử dụng để tiền nhấn tín hiệu nhằm mục đích tăng cường các thành phần tín hiệu ở vùng tần cao vùng mà tín hiệu có xu thế có năng lượng thấp. Phổ tín hiệu ngắn hạn được tính sau đó bằng cách nhân các mẫu của khung tín hiệu với một cửa sổ Hamming và sử dụng phép biến đổi Fourier nhanh (FFT). Đến đây chỉ có biên độ phổ được lấy ra bởi vì phổ pha ngắn hạn không chứa các thông tin có ích của tín hiệu tiếng nói. Chúng ta biết rằng, hệ thống âm thanh (auditory) của con người tích lũy (accumulate) các năng lượng theo những dải chính (critical bands). Dựa vào đặc điểm này, hệ mạch lọc thang Mel (Mel-scale filterbank) được sử dụng. Hệ mạch lọc này gồm 23 băng con (subbands). Các thành phần FFT phổ được nhân với một hàm tam giác và được accumulated vào một vùng tần số xác định tào thành một thành phần phổ Mel. Bề rộng của các dải tần tăng dần khi tần số tăng theo quan hệ tuyến tính và tần số Mel. Với năng lượng tín hiệu người ta tính toán lô-ga-rít của các phổ Mel. Các thành phần tần Mel cạnh nhau có tính tương quan cao (fairly correlated). Để trích chọn các thành phần đặc trưng tương đối độc tập thống kê với nhau, người ta áp dụng phép biến đổi Cosine rời rạc (DCT) cho các lô-ga-rít phổ Mel. Các đặc trưng độc lập thống kê này sẽ tạo thuận lợi cho việc mô hình các đặc tính của tín hiệu tiếng nói trong các mô hình tham chiếu (reference models) và việc tính toán các độ tương đồng trong quá trình so sánh đối chiếu mẫu.



Hình 5.6 Sơ đồ khối quá trình phân tích MFCC

Với phương pháp tiền xử lý theo tiêu chuẩn đưa ra bởi ETSI thì có 13 hệ số cepstral được tính toán bao gồm cả hệ số cepstral thứ 0. Chú ý rằng hệ số cepstral thứ 0 biểu diễn giá trị trung bình (mean) của lô-ga-rít phổ Mel. Do đó, giá trị này có quan hệ mật thiết với năng lượng khung. Thường thì hoặc là lô-ga-rít năng lượng khung được tính từ tín hiệu thời gian hoặc là hệ số cepstral thứ 0 được sử dụng như một tham số trong quá trình nhận dạng tín hiệu tiếng nói. Các véc-tơ đặc trưng cho việc nhận dạng tiếng nói thường bao gồm lô-ga-rít năng lượng khung và 12 hệ số cepstral C_1 đến C_{12}. Để áp dụng các kỹ thuật thích ghi nhằm nâng cao chất lượng hệ thống nhận dạng, chúng ta cần thiết biết tham số C_0. Và do đó C_0 thường được trích ra một cách đặc biệt để sử dụng cho quá trình huấn luyện, và C_0 trở thành một tham số của HMM. Nghĩa là một tập các hệ số cepstral trong các mẫu tham chiếu có thể được biến đổi ngược lại thành phổ Mel. Tuy nhiên cần chú ý rằng thành phần C_0 không được sử dụng cho quá trình nhận dạng mẫu.

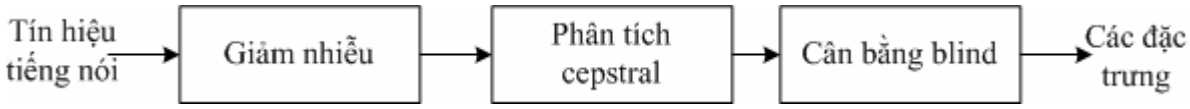
Các tham số âm học giới thiệu phần trên được gọi là các tham số tĩnh vì chúng được tính từ tín hiệu tiếng nói cho một khung ngắn khoảng 25ms. Do đó, để tăng chất lượng hệ thống nhận dạng, một loạt các tham số động cần được quan tâm. Điều này có thể được hiện thực bằng việc quan sát đường biến đổi (contour) của mỗi tham số tĩnh theo thời gian và tính toán vi phân (derivative) của các đường dịch chuyển này. Các tham số được tính toán theo

cách này được gọi là các hệ số đen-ta. Ta có vi phân bậc nhất $\Delta C_i(k)$ của hệ số cepstral C_i được tính theo công thức:

$$\Delta C_i(k) = \frac{\sum_{j=1}^{N_\Delta} j\left[C_i(k+j) - C_i(k-j)\right]}{\sum_{j=1}^{N_\Delta} j^2} \tag{5.30}$$

Hệ số N_Δ trong công thức (5.30) thường được chọn bằng 3. Khi đó các hệ số đen-ta có thể được tính từ 7 khung. Nghĩa là chúng chứa đựng thông tin về các biểu hiện động của tín hiệu trong khoảng thời gian khoảng 85ms. Một cách tương tự, các vi phân cấp hai cũng có thể được tính bằng cách áp dụng (5.30) cho các đường biến đổi của các vi phân cấp một. Các hệ số thu được từ các vi phân cấp hai này được gọi là các hệ số đen-ta-đen-ta. Thời gian cho việc tính toán các vi phân cấp hai thường là thấp hơn cho việc tính toán vi phân cấp một, do đó tổng khoảng thời gian cho việc xác định các hệ số đen-ta-đen-ta của một đoạn tín hiệu khoảng 150ms. Các hệ số đen-ta và đen-ta-đen-ta được thêm vào cùng với các tham số tĩnh để tạo thành các véc-tơ đặc trưng. Thông thường, véc-tơ đặc trưng phổ biến gồm khoảng 39 thành phần bao gồm cả lô-ga-rít năng lượng khung và 12 hệ số cepstral từ C_1 đến C_{12}.

Để có thể tăng tính nhất quán (robust) của việc trích chọn đặc trưng tín hiệu khi có nhiễu nền (background noise) và các hàm truyền đạt không biết trước người ta sử dụng sơ đồ trích chọn được trình bày trong hình 5.7. Đây cũng là sơ đồ tiền xử lý tín hiệu được tiêu chuẩn hóa bởi ETSI. Trong sơ đồ này, ngoài khối trích trọng chúng ta đã đề cập đến ở phần trên, hai khối xử lý được thêm vào. Thứ nhất đó là khối giảm nhiễu, nó bao gồm một mạch lọc Wiener hai tầng (2-stage). Tín hiệu sau khi được giảm nhiễu được đưa vào khối phân tích cepstral như đã mô tả. Để giảm nhỏ ảnh hưởng của các hàm truyền đạt không biết (unknown) đối với các tham số trích chọn ra, một khối cân bằng mờ (blind equalization) được sử dụng. Khối này làm việc trên nguyên lý so sánh phổ tiếng nói với một phổ phẳng và sử dụng thuật toán sai số bình phương nhỏ nhất (LMS - Least mean square) để điều chỉnh bộ lọc cân bằng.



Hình 5.7 Sơ đồ khối cải thiện phương pháp phân tích Cepstral

5.6. Giới thiệu một số phương pháp nhận dạng tiếng nói

Trong phần này, chúng ta sẽ tìm hiểu sơ lược một số phương pháp sử dụng trong các hệ thống nhận dạng tín hiệu tiếng nói. Ngoài phần sơ lược về nguyên lý chúng ta cũng sẽ xem xét đến các điểm mạnh và điểm yếu của mỗi phương pháp.

Một cách khái quát, có ba hướng chính được sử dụng trong các hệ thống nhận dạng tiếng nói. Đó là: phương pháp âm thanh - âm vị (acoustic-phonetic); phương pháp nhận dạng mẫu (pattern recognition) và phương pháp sử dụng trí tuệ nhân tạo.

Phương pháp acoustic-phonetic là phương pháp dựa trên cơ sở lý thuyết âm vị trong đó giả thiết rằng ngôn ngữ tiếng nói tồn tại một số đơn vị âm vị phân biệt và hữu hạn, và rằng

các đơn vị âm tiết (phonetic) được đặc tả một cách đầy đủ bởi một tập các tính chất phù hợp với tín hiệu tiếng nói, hoặc phổ của chúng. Mặc dù các đặc tính âm học của các đơn vị âm tiết thay đổi rất lớn đối với cả người nói (speaker) và với các đơn vị âm tiết lân cận (còn gọi là co-articulation of sound), chúng ta giả thiết rằng những quy luật quản lý sự thay đổi trên có thể suy ra một cách dễ dàng và có thể học và áp dụng vào các tính huống thực tế. Và do đó, bước đầu tiên trong việc sử dụng phương pháp acoustic-phonetic vào việc nhận dạng tín hiệu tiếng nói là việc phân đoạn (segmentation) và gán nhãn. Quá trình này nhằm phân đoạn tín hiệu tiếng nói thành các vùng rời rạc (theo thời gian) trong đó các đặc tính âm học của tín hiệu là đại diện của một (hoặc vài) đơn vị âm tiết (hoặc các lớp). Sau đó gắn một hoặc nhiều nhãn âm tiết với mỗi đoạn tùy theo các tính chất âm học của đoạn đó. Bước tiếp theo trong quá trình nhận dạng là việc cố gắng quyết định một từ hợp lệ (hoặc một chuỗi từ) từ một dãy các nhãn âm tiến được tạo ra từ bước đầu tiên.

Phương pháp nhận dạng mẫu trong nhận dạng tiếng nói là phương pháp trong đó các mẫu tiếng nói được sử dụng trực tiếp mà không cần phải xác định rõ ràng đặc trưng (theo nghĩa đặc trưng âm học) và không cần quá trình phân đoạn. Cũng giống như mọi phương pháp nhận dạng mẫu khác, phương pháp này gồm hai bước: huấn luyện các mẫu tín hiệu tiếng nói; nhận dạng các mẫu thông qua việc sô sánh các mẫu. Thông tin (hiểu biết - knowledge) về tín hiệu tiếng nói được đưa vào hệ thống trong quá trình huấn luyện hệ thống. Nguyên lý của việc này là nếu có đủ các phiên bản của một mẫu cần nhận dạng (mẫu của âm, của từ, hoặc của một cụm từ ...) trong tập dùng để huấn luyện, thì quá trình huấn luyện sẽ có thể đặc tả một cách chính xác các đặc tính âm học của mẫu (mà không cần quan sát hoặc thông tin của bất cứ mẫu nào khác trong quá trình huấn luyện). Quá trình so sánh mẫu thực hiện việc so sánh trực tiếp tín hiệu tiếng nói chưa biết (tín hiệu tiếng nói cần nhận dạng) với mỗi một mẫu học được trong quá trình huấn luyện và phân loại tín hiệu tiếng nói chưa biết theo độ tương hợp với mẫu. Phương pháp nhận dạng mẫu có các ưu điểm:

- Sử dụng đơn giản.

- Nhất quán và không thay đổi với các bộ từ vựng, người sử dụng, tập các đặc trưng khác nhau. Điều này cho phép thuật toán có thể áp dụng một cách rộng rãi với các loại đơn vị tín hiệu tiếng nói (từ các đơn vị phonemelike, từ, cụm từ hoặc câu), các bộ từ vựng, số đông người nói, các môi trường nền khác nhau...

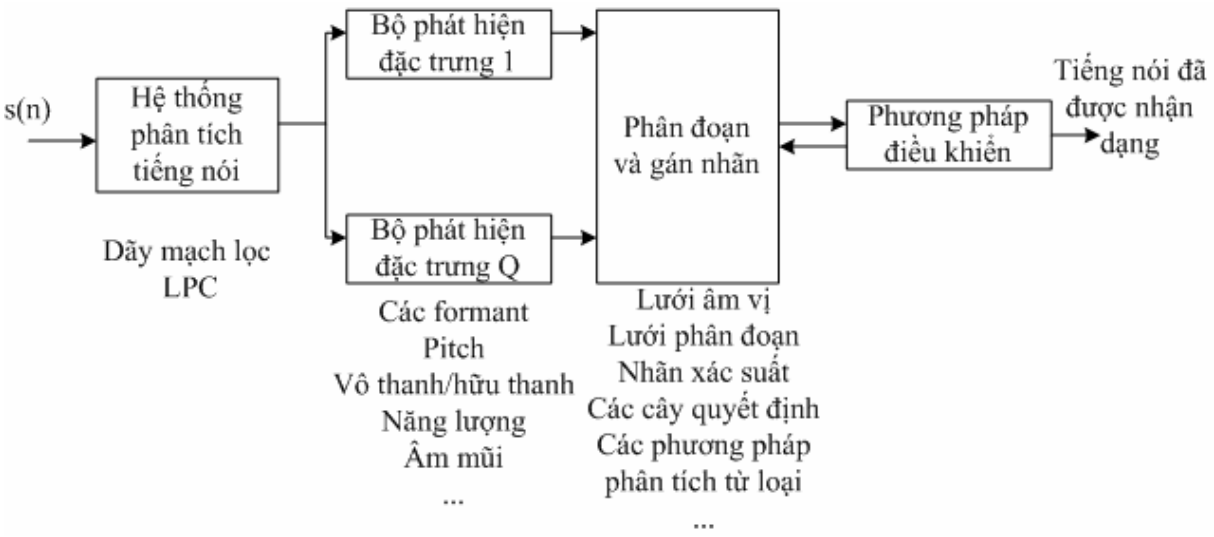
- Có chất lượng tốt. Người ta đã chỉ ra rằng việc sử dụng phương pháp nhận dạng mẫu trong nhận dạng tiếng nói luôn cho phép hệ thống hoạt động tốt đối với bất kỳ nhiệm vụ nào với yêu cầu công nghệ vừa phải.

Phương pháp sử dụng trí tuệ nhân tạo trong nhận dạng tín hiệu tiếng nói là phương pháp lai ghép giữa hai phương pháp kể trên. Phương pháp này cố gắng cơ chế hóa thủ tục nhận dạng tương tự như cách thức con người áp dụng trí tuệ vào việc quan sát (visualizing), phân tích và cuối cùng là ra quyết định trên các đặc tính âm học đo lường được. Đặc biệt một trong các kỹ thuật được sử dụng cho các phương pháp thuộc lớp phương pháp này là việc sử dụng hệ chuyên gia để phân đoạn và gán nhãn. Bằng cách này, bước khó khăn nhất và quan trọng nhất trong quá trình nhận dạng có thể được thực hiện không chỉ với các thông tin âm học như trong các phương acoustic-phonetic thuần túy; học và thích ứng theo thời gian; sử dụng mạng nơ-ron cho việc học các mối quan hệ giữa các âm tiết và tất cả các đầu vào đã biết cũng như cho việc phân biệt sự giống nhau giữa các lớp âm.

Việc sử dụng mạng nơ-ron có thể tạo ra một phương pháp cấu trúc riêng rẽ cho việc nhận dạng tín hiệu tiếng nói hoặc có thể được coi như một cấu trúc có thể thực thi được, cấu trúc mà có thể tích hợp vào một trong ba phương pháp vừa kể.

5.6.1 Phương pháp acoustic-phonetic

Hình 5.8 miêu tả sơ đồ khối của một hệ thống nhận dạng tín hiệu tiếng nói sử dụng phương pháp acoustic-phonetic. Bước đầu tiên trong quá trình xử lý, cũng giống như trong tất cả các phương pháp nhận dạng tín hiệu tiếng nói khác, đó là việc phân tích tín hiệu tiếng nói. Việc phân tích tín hiệu tiếng nói (còn được gọi là phương pháp đo lường các đặc trưng của tín hiệu) đưa ra một biểu diễn phổ phù hợp nhất đối với các đặc trưng của tín hiệu tiếng nói thay đổi theo thời gian. Như đã đề cập, các phương pháp phổ biến nhất trong việc phân tích phổ tín hiệu tiếng nói trong một hệ thống nhận dạng tín hiệu tiếng nói là phương pháp phân tích LPC. Nói một cách tổng quát, việc phân tích phổ tín hiệu tiếng nói có nhiệm vụ đưa ra được các biểu diễn phổ thích hợp của tín hiệu tiếng nói theo thời gian.



Hình 5.8 Sơ đồ khối một hệ thống nhận dạng tiếng nói theo phương pháp acoustic-phonetic

Bước tiếp theo trong quá trình xử lý là giai đoạn phát hiện các đặc trưng. Ý tưởng ở đây là chuyển đổi các đo lường phổ thành một tập các đặc trưng sao cho có thể mô tả một cách bao trùm các tính chất âm học của các đơn vị âm tiết khác nhau. Trong các đặc trưng sử dụng cho việc nhận dạng tín hiệu tiếng nói phải kể đến âm mũi (nasality) tức là sự có mặt hoặc không của cộng hưởng khoang mũi, âm căng (frication) tức là sự có mặt hoặc không của nguồn kích thích ngẫu nhiên trong tín hiệu, vị trí các tần số cộng hưởng bộ máy phát thanh (formant) tức là các tần số của ba đỉnh cộng hưởng đầu tiên, tín hiệu hữu thanh hay vô thanh tức là nguồn kích thích là tuần hoàn hay không tuần hoàn, và tỉ lệ giữa năng lượng của tần cao và tần thấp. Một số đặc trưng bản chất là nhị phân (binary) chẳng hạn như âm mũi, âm căng, âm hữu thanh-âm vô thanh, tuy nhiên một số khác là liên tục chẳng hạn như vị trí các formant, tỷ số năng lượng. Tầng phát hiện các đặc trưng thường bao gồm một tập các bộ phát hiện (detector) hoạt động song song và xử dụng phép xử lý thích hợp và lô-gic để đưa ra quyết định về sự có mặt hoặc không, hoặc giá trị, của một đặc trưng. Các thuật toán dùng cho việc phát biện các đặc trưng riêng biệt thường là rất phức tạp và chúng thường thực hiện rất nhiều

phép biến đổi tín hiệu, trong một số trường hợp chúng có thể là các thủ tục ước lượng tầm thường (thông thường - trivial).

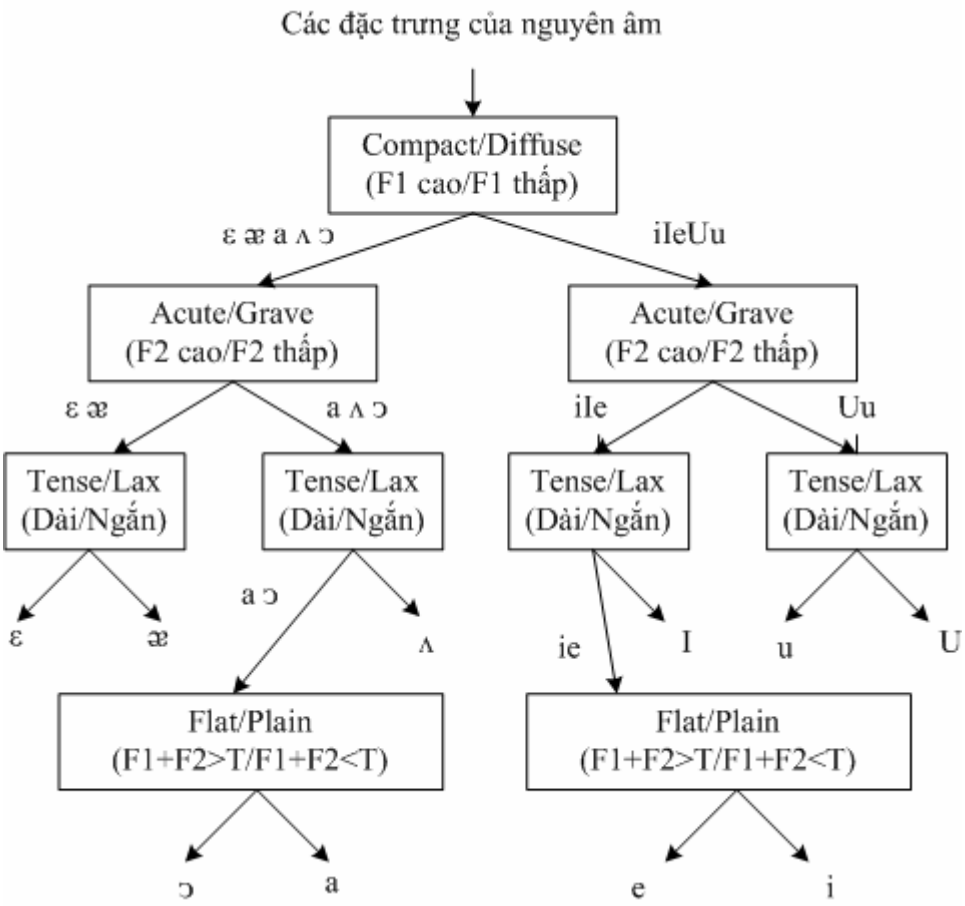
Bước thứ ba trong quá trình là việc phân đoạn và gán nhãn. Hệ thống cố gắng tìm ra vùng ổn định, vùng mà các đặc trưng thay đổi rất nhỏ, và sau đó gán nhãn cho các vùng vừa được phân ra tương ứng sao cho các đặc trưng trong vùng này tương đồng tốt với các đặc trưng tương ứng của các đơn vị âm tiết riêng rẽ. Giai đoạn này là giai đoạn trung tâm của quá trình nhận dạng tín hiệu tiếng nói theo phương pháp acoustic-phonetic và nó cũng là một giai đoạn khó khăn nhất để có thể triển khai một cách tin cậy. Vì lý do đó, nhiều chiến thuật (strategy) điều khiển đã được sử dụng để hạn chế khoảng của các điểm phân đoạn cũng như các khả năng gán nhãn. Chẳng hạn, đối với việc nhận dạng các từ riêng rẽ, các giới hạn chẳng hạn như một từ có chứa ít nhất hai đơn vị âm tiết và không thể nhiều hơn sáu đơn vị âm tiết cho phép chiến lược điều khiển chỉ cần quan tâm đến các kết quả với khoảng giữa một và năm khoảng điểm phân đoạn. Hơn nữa, chiến thuật gán nhãn có thể tận dụng các giới hạn về từ vựng (lexical) của các từ để chỉ cần xem xét các từ với n đơn vị âm tiết, trong đó việc phân đoạn cho ta n-1 điểm phân đoạn. Những điều kiện hạn chế vừa nêu có vai trò quan trọng cho phép chúng ta giảm nhỏ không gian tìm kiếm và tăng đáng kể chất lượng hoạt động của hệ thống.

Kết quả của giai đoạn phân đoạn và gán nhãn thương là một lưới phoneme (phoneme lattice). Lưới này được sử dụng để thực hiện thủ tục truy xuất từ vựng (a lexical access procedure) nhằm xác định được một từ hoặc một dãy từ tương đồng nhất. Ngoài các kiểu lưới phoneme, người ta còn có thể xây dựng lưới từ hoặc syllable bằng cách kết hợp các điều kiện giới hạn từ vựng và cú pháp vào chiến thuật điều khiển vừa được đề cập ở trên. Chất lượng của việc so sánh tương đồng của các đặc trưng với các đơn vị âm tiết trong một phân đoạn có thể được sử dụng để gán xác suất cho các nhãn và các nhãn này sau đó có thể được sử dụng trong thủ tục truy xuất từ vựng thống kê (a probabilistic lexical access procedure). Đầu ra của hệ thống nhận dạng là một từ hoặc một dãy từ mà tương đồng nhất theo một khía cạnh định trước với dãy các đơn vị âm tiết trong lưới phoneme.

a) Bộ phân loại các âm vị nguyên âm

Chúng ta cùng xem xét thủ tục gán nhãn trên một phân đoạn được phân loại như một nguyên âm. Sơ đồ hình 5.9 mô tả lưu đồ phân loại nguyên âm theo phương pháp acoustic-phonetic. Chúng ta giả sử rằng có ba đặc trưng đã được phát hiện trong phân đoạn là formant thứ nhất F_1, formant thứ hai F_2 và chiều dài của phân đoạn D. Thêm nữa chúng ta chỉ xem xét tập các nguyên âm ổn định (steady), tức là loại bỏ các nguyên âm kép (diphthongs). Để phân loại một phân đoạn nguyên âm trong 10 nguyên âm ổn định, một số phép thử cần phải thực hiện để phân tách các nhóm nguyên âm. Như trình bày trong hình 5.9, phép thử đầu tiên tách các nguyên âm có tần số F_1 thấp (còn gọi là các nguyên âm khuếch tán (diffuse) chẳng hạn như /i/, /i/, /u/, ...) với các nguyên âm có tần số cao (còn gọi là các nguyên âm gọn (compact) bao gồm /a/, ...). Mỗi tập con này lại được phân tách thêm dựa vào tần số F_2, trong đó các nguyên âm acute (âm sắc) có tần số F_2 cao và các nguyên âm grave (âm huyền) có tần số F_2 thấp. Phép kiểm tra thứ ba dựa trên khoảng thời gian của phân đoạn sẽ phân tách các nguyên âm căng (tense vowel), tức là các nguyên âm có giá trị D lớn với các nguyên âm lax (thả lỏng), tức là các nguyên âm có giá trị D nhỏ. Cuối cùng, một phép kiểm tra mịn hơn (finer) đối với các giá trị formant để phân tách các nguyên âm chưa phân tách còn lại tạo ra lớp các nguyên

âm bằng (flat) tức là các nguyên âm có F_1+F_2 lớn hơn một ngưỡng T nào đó và các nguyên âm đơn giản (plain) (các nguyên âm có F_1+F_2 nằm dưới một ngưỡng T nào đó)



Hình 5.9 Một phương pháp đơn giản phân loại nguyên âm tiếng Anh

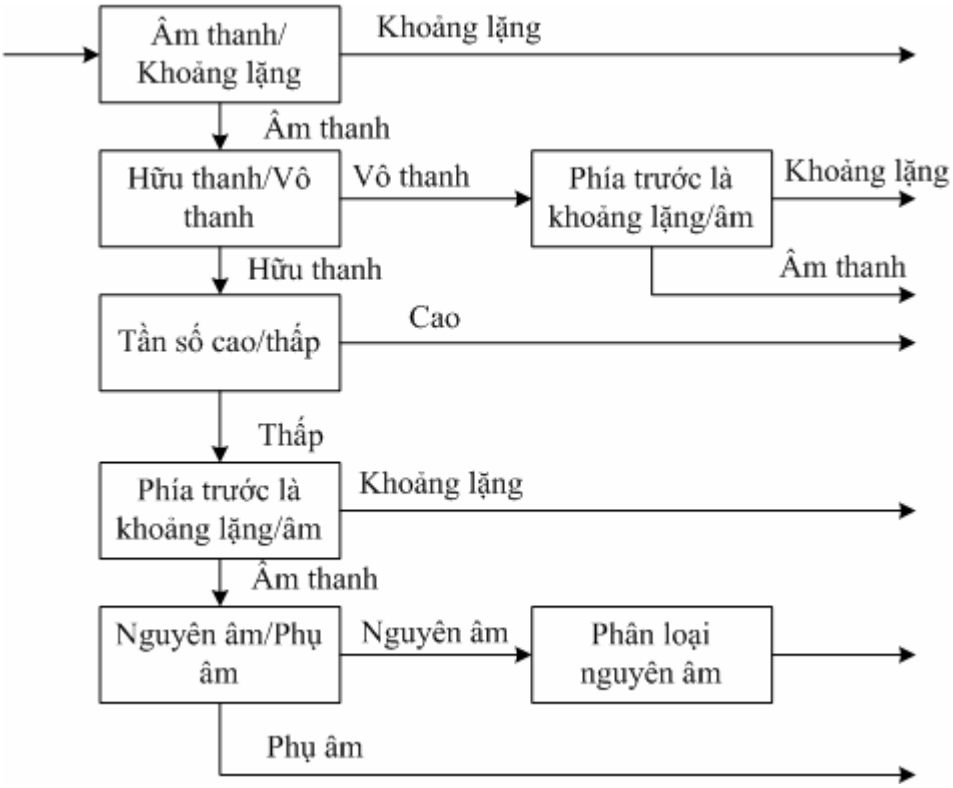
Cần chú ý rằng, có một số mức ngưỡng được sử dụng trong bộ phân loại nguyên âm. Các mức ngưỡng này thường được xác định bằng thực nghiệm sao cho có thể tăng tối đa tính chính xác của phép phân loại trên một tập tín hiệu tiếng nói cho trước.

b) Phân loại âm thanh tiếng nói

Việc phân loại nguyên âm chỉ là một phần nhỏ trong quá trình gán nhãn âm tiết của phương pháp nhận dạng tín hiệu tiếng nói acoustic-phonetic. Về mặt lý thuyết, chúng ta cần phải có một phương pháp phân loại một phân đoạn bất kỳ nào đó thành một hoặc nhiều hơn một trong số hơn 40 đơn vị âm tiết được thảo luận trước đây. Trong phần này chúng ta xem xét một bài toán phân loại đơn giản hơn nhằm phân loại một phân đoạn tiếng nói thành một hoặc một số lớp tín hiệu tiếng nói, chẳng hạn như các âm vô thanh ngắt (unvoiced stop), âm hữu thanh ngắt (voiced stop), âm vô thanh xát (unvoiced fricative). Chúng ta biết rằng không tồn tại một thủ tục đơn giản hoặc tổng quát được chấp nhận rộng rãi để thực hiện tác vụ này, tuy vậy, hình 5.10 mô tả một phương pháp đơn giản trực giác để hoàn thành việc phân loại như vậy.

Phương pháp này sử dụng một cây nhị phân để ra quyết định các lớp tín hiệu khác nhau. Quyết định đầu tiên là phân chia lớp âm thanh/yên lặng (sound/silence). Ở quyết định này các đặc trưng tín hiệu tiếng nói (về cơ bản là năng lượng trong trường hợp này) được so sánh với

một ngưỡng đã được lựa chọn, các tín hiệu yên lặng được tách ra nếu như phép thử là âm đối với âm thanh tiếng nói. Quyết định thứ hai là việc phân lớp các âm hữu thanh và vô thanh (cơ sở dựa trên việc xuất hiện tính tuần hoàn của tín hiệu trong phân đoạn đang xét). Kết quả của quyết định này là các âm vô thanh được tách khỏi các âm hữu thanh. Bước tiếp theo là thực hiện một phép thử để phân tách các phụ âm vô thanh ngắt (unvoiced stop consonants) khỏi các phụ âm vô thanh xát (unvoiced fricatives). Bằng phép thử tần số cao thấp/tần số thấp (năng lượng), chúng ta có thể phân tách các âm hữu thanh xát (voiced fricatives) khỏi các âm hữu thanh. Các âm hữu thanh ngắt (voiced stop) có thể được phân tách bằng cách kiểm tra xem âm vị trước đó có phải là yên lặng (hoặc gần giống yên lặng). Cuối cùng một phép kiểm tra phổ nguyên âm/phụ âm được tiến hành (tìm kiếm khe phổ) nhằm tách các nguyên âm khỏi các phụ âm.



Hình 5.10 Phương pháp phân loại âm thanh tiếng nói dựa vào cây nhị phân

Thủ tục phân tách nguyên âm được trình bày trong hình 5.9 có thể được sử dụng thêm như một phép phân loại mịnh các nguyên âm.

Chú ý là thủ tục phân loại đề cập trên và minh hoạ trong hình 5.10 chỉ mang tính minh họa sơ lược và có nhiều lỗi. Chẳng hạn, một số âm hữu thanh ngắt không phải bắt đầu bằng khoảng lặng hoặc âm giống khoảng lặng. Một vấn đề nữa là không đưa ra được một cách nào có thể phân biệt các âm kép (diphthongs) từ các nguyên âm.

c) Một số tồn tại trong phương pháp nhận dạng acoustic-phonetic

Có rất nhiều vấn đề tồn tại trong phương pháp nhận dạng tín hiệu tiếng nói acoustic-phonetic. Những vấn đề này làm cho phương pháp thiếu sự thành công trong các hệ thống nhận dạng tín hiệu tiếng nói thực tế. Trong các tồn tại phải kể đến là:

1. Phương pháp này yêu cầu một khối lượng thông tin lớn (extensive) về các tính chất âm học của các đơn vị âm tiết. Những thông tin này thường là không đầy đủ và không sẵn sàng ngoại trừ những trường hợp đơn giản.

2. Việc chọn các đặc trưng được thực hiện chủ yếu dựa trên các xem xét ad hoc. Với hầu hết các hệ thống, việc chọn các đặc trưng dựa trên các nhận thức chứ không phải tối ưu theo một tiêu chí định sẵn và có nghĩa (a well-defined and meaningful sense)

3. Thiết kế các bộ phân loại âm thanh cũng không phải là các thiết kế tối ưu. Phương pháp ad hoc thường được sử dụng để xây dựng các cây nhị phân quyết định. Gần đây, các phương pháp cây hồi quy (regression) và phân loại (CART) được sử dụng thay thế để cho phép các cây quyết định nhất quán hơn. Tuy vậy, vì việc lựa chọn các đặc trưng chủ yếu là cận tối ưu, các thực thi tối ưu của CART thường ít khi đạt được.

4. Không tồn tại một thủ tục định sẵn và tự động nào cho việc điều chỉnh phương pháp (chẳng hạn như chỉnh các ngưỡng quyết định, ...) trên các tín hiệu được gán nhãn thực. Thực tế, thậm chí còn không có một phương pháp lý tưởng của việc gán nhãn tín hiệu tiếng nói huấn luyện một cách nhất quán và được sự đồng ý rộng rãi của các chuyên gia ngôn ngữ học.

Do các tồn tại nêu trên, mặc dù phương pháp nhận dạng acoustic-phonetic là một ý tưởng khá thú vị nhưng cần có nhiều nghiên cứu hiểu biết hơn nữa để có thể thực hiện thành công các hệ thống nhận dạng thực tế dựa trên phương pháp này.

5.6.2 Phương pháp nhận dạng mẫu thống kê

Hình 5.11 mô tả sơ đồ khối một hệ thống nhận dạng sử dụng phương pháp nhận dạng mẫu. Phương pháp nhận dạng mẫu bao gồm bốn bước:

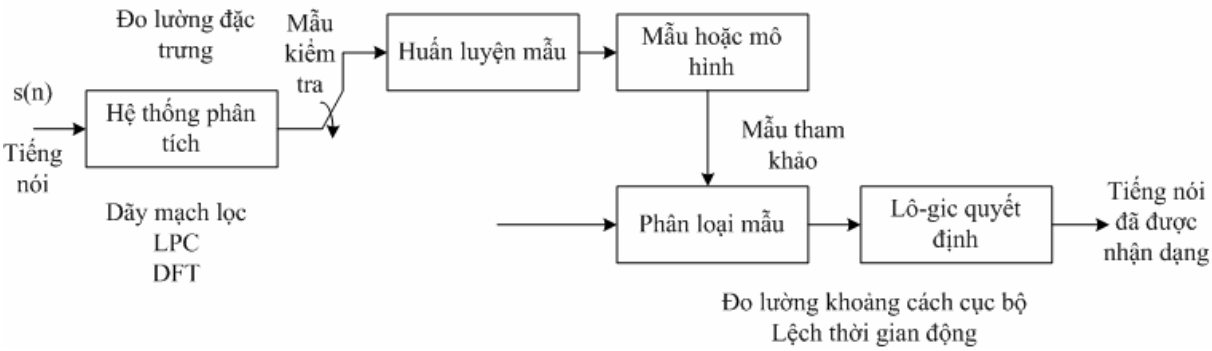
1. Đo lường các đặc trưng, trong đó một dãy các phép đo lường được thực hiện trên tín hiệu vào để định ra các mẫu cần thử. Đối với tín hiệu tiếng nói, các đo lường đặc trưng thường là các đầu ra của một số phương pháp phân tích phổ nào đó, chẳng hạn bộ phân tích mạng mạch lọc, một bộ phân tích LPC, hoặc là một phân tích DFT.

2. Huấn luyện mẫu, trong đó một hoặc nhiều mẫu kiểm tra tương ứng với các âm thanh tín hiệu tiếng nói của cùng một lớp được sử dụng để tạo ra một mẫu đại diện của các đặc trưng của lớp đó. Mẫu kết quả thu được, thường được gọi là mẫu tham khảo (hoặc tham chiếu), có thể trở thành một ví dụ (examplar) hoặc một mẫu (template) được suy ra (derived) từ một số phương pháp tính trung bình hoặc có thể trở thành một mô hình đặc tả tính thống kê của các đặc trưng của mẫu tham khảo.

3. Phân loại mẫu, trong đó mẫu cần kiểm tra chưa biết được so sánh với mỗi lớp (âm) mẫu tham khảo và một đo lường độ tương đồng (khoảng cách) giữa mẫu kiểm tra và mỗi mẫu tham khảo được tính toán. Để so sánh các mẫu tín hiệu tiếng nói (các mẫu bao gồm một dãy các véc-tơ phổ), chúng ta cần cả đo lường khoảng cách cục bộ, với khoảng cách cục bộ được định nghĩa là khoảng cách phổ giữa hai véc-tơ phổ được xác định rõ, và một thủ tục sắp xếp thời gian toàn cục (a global time alignment procedure) (thường được gọi là một thuật toán lệch (warping) thời gian động) nhằm bù lại sự khác biệt tốc độ tiếng nói (tỷ lệ thời gian) của hai mẫu.

4. Quyết định lô-gic, trong đó điểm số về tính tương đồng của mẫu tham chiếu được sử dụng để quyết định xem mẫu tham chiếu nào (hoặc có thể một dãy mẫu tham chiếu) tương đồng nhất với mẫu kiểm tra chưa biết.

Các yếu tố phân biệt các phương pháp nhận dạng mẫu khác nhau là các kiểu đo lường đặc trưng, sự lựa chọn các mẫu (template) hoặc các mô hình cho các mẫu tham chiếu, và phương thức được sử dụng để tạo các mẫu tham chiếu và phân loại các mẫu kiểm tra chưa biết.



Hình 5.11 Sơ đồ khối của một hệ thống nhận dạng sử dụng phương pháp nhận dạng mẫu

Các điểm mạnh và điểm yếu của phương pháp nhận dạng mẫu có thể kể đến:

1. Chất lượng của hệ thống nhận dạng theo phương pháp nhận dạng mẫu nhạy cảm (sensitive) với số lượng dữ liệu huấn luyện để tạo ra lớp các mẫu tham chiếu; thông thường, càng huấn luyện, chất lượng của hệ thống càng cao với mọi tác vụ.

2. Các mẫu tham chiếu nhạy cảm với môi trường tiếng nói và các tính chất truyền dẫn của phương tiện truyền dẫn để tạo tiếng nói; điều này là bởi vì các đặc tính phổ tín hiệu tiếng nói thường dễ bị ảnh hưởng bởi quá trình truyền dẫn và nhiễu nền.

3. Vì không có thông tin tiếng nói cụ thể được sử dụng một cách rõ ràng (explicitly) trong hệ thống, phương pháp này tương đối trơ (insensitive) đối với việc chọn các từ vựng, các tác vụ, cú pháp, và các tác vụ ngữ nghĩa.

4. Khối lượng tính toán cho cả quá trình huấn luyện mẫu và phân loại mẫu thường tỷ lệ thuận với số mẫu cần được huấn luyện hoặc được nhận dạng; do đó việc tính toán cho một số lượng lớn lớp tín hiệu âm có thể và thường trở lên không thể thực hiện được (prohibitive)

5. Bởi vì hệ thống trơ với lớp âm thanh, các kỹ thuật cơ bản có thể áp dụng cho nhiều lớp tín hiệu tiếng nói, bao gồm các cụm từ, từ hoàn chỉnh, hoặc các đơn vị con của từ (sub-word). Do đó, chúng ta sẽ thấy cách một tập cơ bản các kỹ thuật được phát triển cho một lớp âm (chẳng hạn cho các từ) có thể được áp dụng trực tiếp cho các lớp âm khác (chẳng hạn cho các đơn bị sub-word) mà không cần thay đổi hoặc thay đổi rất ít đối với thuật toán.

6. Có thể dễ dàng (straightforward) kết hợp các điều kiện hạn chế cú pháp (và thậm chí cả ngữ nghĩa) một cách trực tiếp vào cấu trúc nhận dạng mẫu. Bằng cách đó có thể tăng tính chính xác của việc nhận dạng và giảm nhỏ khối lượng tính toán.

5.6.3 Phương pháp sử dụng trí tuệ nhân tạo

Ý tưởng cơ bản của phương pháp nhận dạng tín hiệu tiếng nói sử dụng trí tuệ nhân tạo là biên dịch và kết hợp thông tin (hiểu biết) từ nhiều nguồn thông tin và dùng nó để giải bài toán. Do đó, chẳng hạn, phương pháp sử dụng trí tuệ nhân tạo việc phân đoạn và gán nhãn có thể được gia tăng (augment) việc sử dụng thông tin âm học tổng quát với thông tin về

phonemic, thông tin về từ vựng, thông tin về cú pháp, thông tin về ngữ nghĩa, và thậm chí cả các thông tin thực dụng (pragmatic knowledge). Để hiểu rõ, chúng ta định nghĩa các nguồn thông tin khác nhau như sau:

- Thông tin âm học là các dữ kiện (evidence) các âm thanh (các đơn vị âm tiết định nghĩa sẵn) được nói trên cơ sở các đo lường phổ và sự có mặt hoặc không của đặc trưng.

- Thông tin từ vựng (lexical) là các thông tin về sự kết hợp giữa các dữ kiện âm học để tạo thành các cấu trúc từ và được cụ thể hóa bởi một bộ từ vựng ánh xạ các âm thanh vào các từ (hoặc tương ứng tách các từ thành các âm tương ứng).

- Thông tin cú pháp là các thông tin về sự kết hợp của các từ để tạo thành một dãy đúng ngữ pháp (theo một mô hình ngôn ngữ nào đó) chẳng hạn như các câu hoặc các cụm từ.

- Thông tin ngữ nghĩa (semantic) là sự hiểu thông tin nhằm có thể đánh giá được các câu hoặc các cụm từ mà nhất quán với tác vụ đang được thực hiện hoặc nhất quán với các câu đã được giải mã trước đó.

- Thông tin thực dụng là các thông tin cho phép có khả năng suy diễn (inference) cần thiết nhằm giải quyết trường hợp có sự mập mờ về nghĩa dựa trên hiểu biết rằng các từ hoặc cụm từ nào thường được dùng nhiều hơn.

Để hiểu đúng và các hạn chế của các khái niệm nguồn thông tin vừa đề cập, chúng ta xem xét các câu tiếng Anh sau:

1) Go to the refrigerator and get me a book.

2) The bears killed the rams.

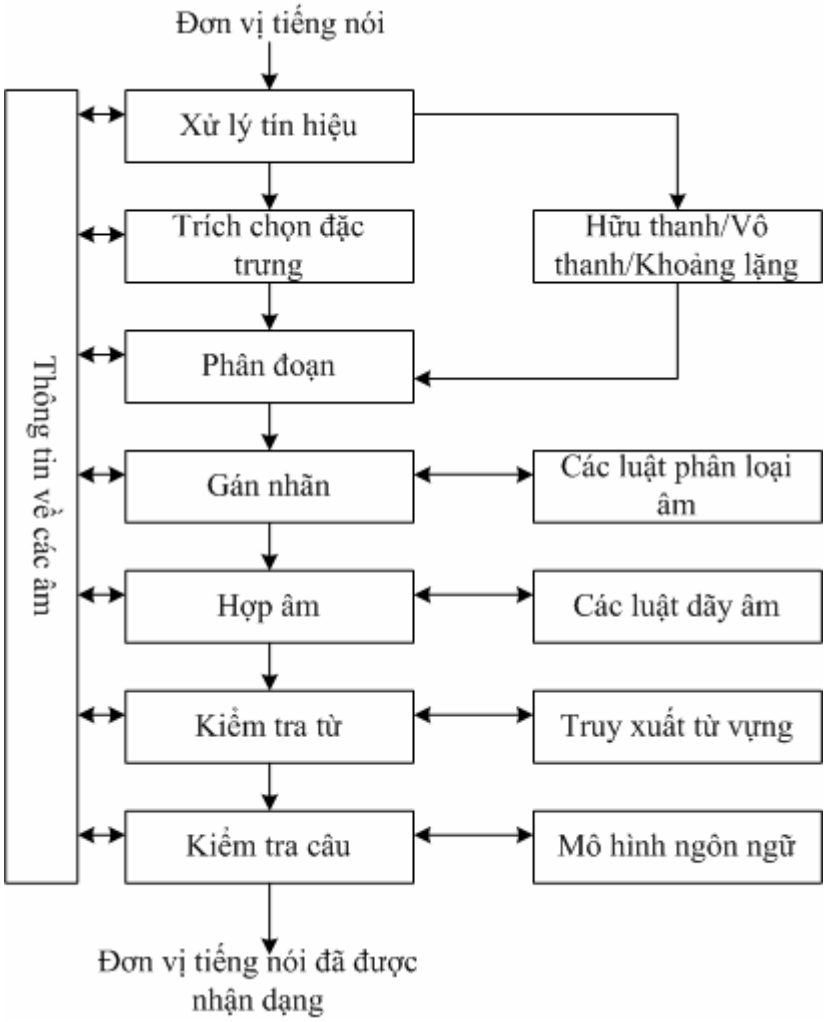
3) Power plants colorless happily old.

4) Good ideas often run when least expected.

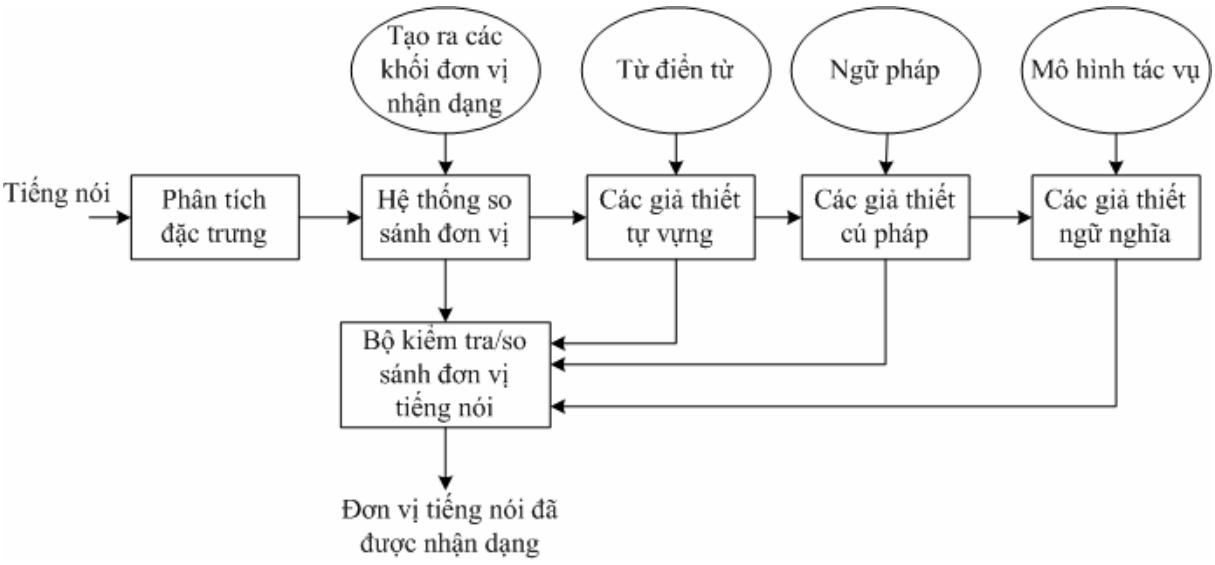
Chúng ta thấy rằng, câu đầu tiên là một câu đúng về mặt cú pháp nhưng không nhất quán về mặt ngữ nghĩa, sách không được mong chờ để ở tủ lạnh. Câu thứ hai tùy thuộc vào ngữ cảnh mà có nghĩa khác nhau. Ví dụ nếu ngữ cảnh là ở rừng thì nó miêu tả sự kiện gấu giết cừu, tuy nhiên nếu chúng ta đang nói đến bóng đá có thể hiểu là đội có tên là những con gấu đã chiến thằng đội có tên là những con cừu. Câu thứ ba thì hoàn toàn không đúng cú pháp cũng như không có nghĩa. Câu thứ tư không nhất quán về mặt ngữ nghĩa, tuy nhiên theo hiểu biết thực dụng có thể đơn giản thay đổi "run" thành "come" thì sẽ có nghĩa mặc dù có chú khác biệt về mặt âm tiết.

Việc kết hợp các điều kiện hạn chế của các nguồn thông tin vừa kể sẽ cho phép hệ thống nhận dạng tín hiệu tiếng nói hoạt động với chất lượng cao hơn. Có nhiều cách kết hợp các nguồn thông tin vừa kể vào một hệ thống nhận dạng. Phương pháp đầu tiên phổ biến nhất có thể kể đến là bộ xử lý "bottom-up" được trình bày trong hình 5.12. Trong phương pháp "bottom-up", các xử lý cấp thấp nhất (chẳng hạn như trích chọn đặc trưng, giải mã âm tiết, ...) được thực hiện trước các phép xử lý cấp cao (giải mã từ vựng, mô hình ngôn ngữ, ...) theo một thứ tự nối tiếp sao cho điều kiện hạn chế của mỗi bước xử lý là nhỏ nhất có thể. Một phương pháp khác là phương pháp xử lý "top-down". Trong phương pháp này mô hình ngôn ngữ tạo ra các giả thuyết từ (word hypotheses) phù hợp với tín hiệu tiếng nói, và tiếp theo là các câu với cú pháp và ngữ nghĩa có nghĩa được xây dựng dựa trên số điểm đánh giá sự tương đồng các từ. Sơ đồ phương pháp xử lý "top-down" được trình bày trong hình 5.13. Một

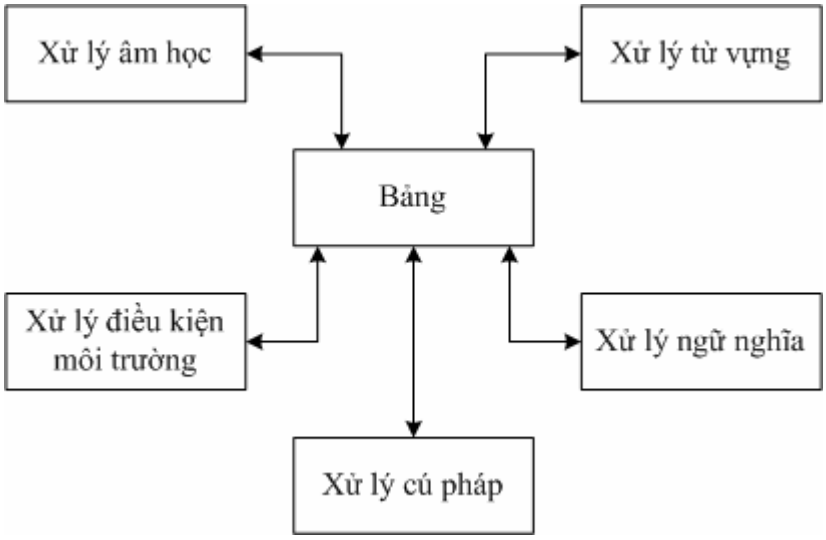
phương pháp thứ ba phải kể đến là phương pháp "blackboard", được mô tả trong hình 5.14. Ở phương pháp này, tất các các nguồn kiến thức được xem xét một các độc lập, một lược đồ giả thiêt-và-kiểm tra có nhiệm vụ thực hiện việc thông tin giữa các nguồn thông tin. Mỗi nguồn thông tin là một nguồn điều khiển dữ liệu dựa trên sự xuất hiện của các mẫu trên "blackboard" mà tương đồng với các mẫu (template) được quy định bởi nguồn thông tin đó. Hệ thống hoạt động theo chế độ cận đồng bộ, các hàm định giá, các xem xét sử dụng và một chính sách đánh giá toàn cục kết hợp và lan truyền việc đánh giá ở mọi mức độ.



Hình 5.12 Phương pháp tích hợp “bottom-up” của hệ thống nhận dạng tiếng nói



Hình 5.13 Phương pháp tích hợp "top-down" của hệ thống nhận dạng tiếng nói



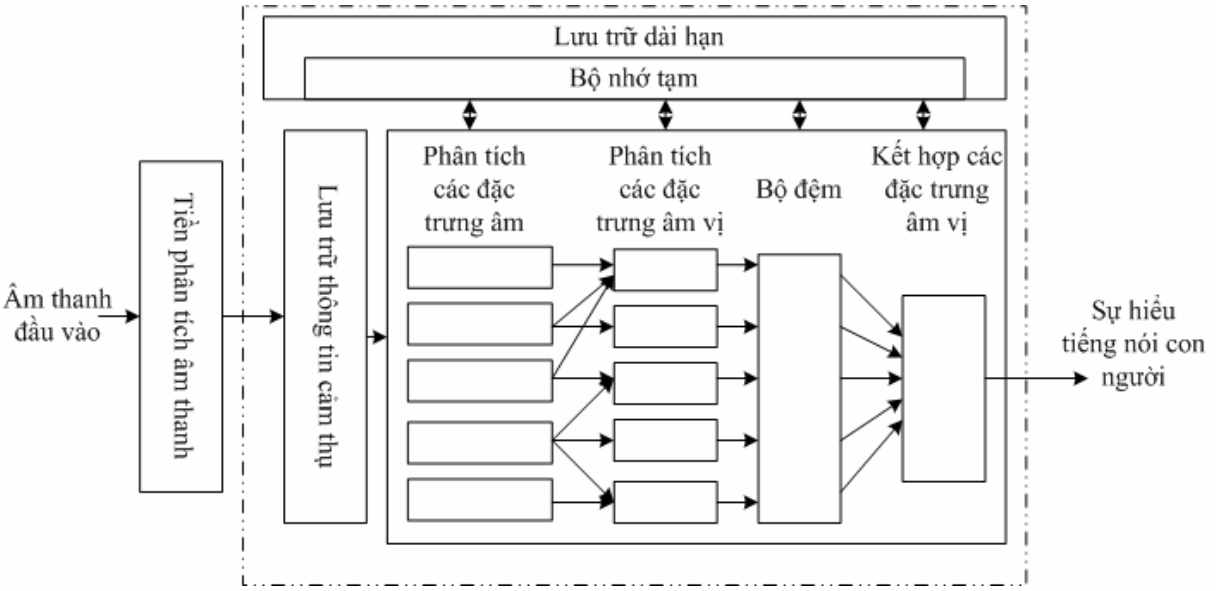
Hình 5.14 Phương pháp tích hợp "blackboard" của hệ thống nhận dạng tiếng nói

5.6.4 Ứng dụng mạng nơ-ron trong hệ thống nhận dạng tiếng nói

Chúng ta biết rằng, có rất nhiều nguồn thông tin (kiến thức) khác nhau cần được thiết lập trong hệ thống nhận dạng tín hiệu tiếng nói sử dụng giải pháp trí tuệ nhận tạo. Do vậy, phương pháp sử dụng trí tuệ nhân tạo có hai khái niệm chính yếu là tự động thu nhận nguồn thông tin (khả năng học) và khả năng thích ứng (adaption). Một giải pháp để thực hiện các yêu cầu này là sử dụng mạng nơ-ron. Trong phần này chúng ta sẽ thảo luận về động lực tại sao người ta nghiên cứu về các mạng nơ-ron và cách mà con người đã áp dụng mạng nơ-ron vào hệ thống nhận dạng tín hiệu tiếng nói.

Hình 5.15 là một mô hình một hệ thống hiểu được tiếng nói con người. Trong hệ thống này, các phân tích âm thanh được dựa một cách không chặt chẽ vào hiểu biết của chúng ta vào quá trình xử lý âm trong tai. Các phân tích đặc trưng khác nhau biểu diễn cho các quá trình xử lý ở nhiều mức độ trong các đường dây thần kinh tới não. Các bộ nhớ ngắn hạn và dài hạn sẽ cho phép điều khiển từ bên ngoài của các quá trình thần kinh được tiến hành theo

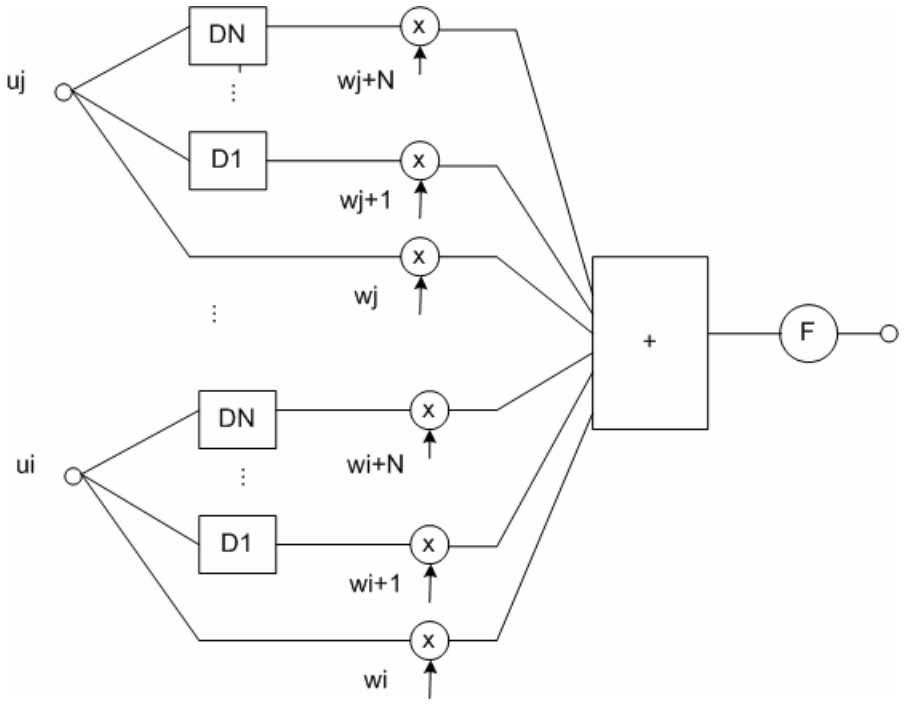
một cách mà chúng ta chưa hiểu biết rõ ràng. Cấu trúc tổng quát của mô hình là một mạng kết nối lan truyền thuận hay còn gọi là mạng nơ-ron.



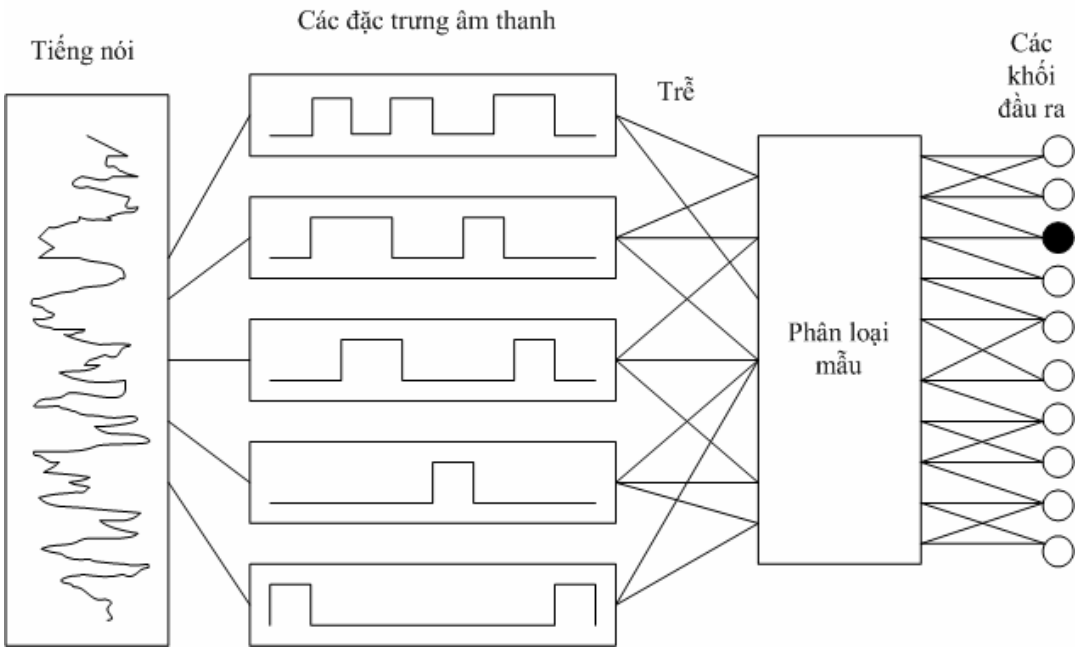
Hình 5.15 Sơ đồ khối ý tưởng của một hệ thống hiểu tiếng nói con người

Các mạng nơ-ron nhân tạo truyền thống (conventional) là các cấu trúc dùng để giải quyết các bài toán liên quan đến các mẫu tĩnh. Do đó, để có thể áp dụng cho tín hiệu tiếng nói, một tín hiệu có bản chất động, chúng ta cần có một số thay đổi trong các cấu trúc mạng truyền thống. Mặc dù cho đến nay chưa có một cách đúng đắn hoặc chính xác để giải quyết vấn đề tính chất động của tín hiệu tiếng nói được biết đến, các nhà nghiên cứu đã đưa ra một số cấu trúc chấp nhận được, trong đó phải kể đến là cấu trúc mạng nơ-ron với thời gian trễ (TDNN - Time delay neural network) được mô tả trong hình 5.16. Cấu trúc này mở rộng đầu vào của mỗi phần tử tính toán để thêm vào N khung tín hiệu tiếng nói (tức là các véc-tơ phổ sẽ bao trùm khoảng thời gian $N\Delta$ giây, trong đó Δ là khoảng thời gian phân tách giữa các thành phần phổ cạnh nhau). Bằng việc mở rộng đầu vào tới N khung (trong đó N thường cỡ 15), các loại bộ phát hiện acoustic-phonetic khác nhau trở thành hiện thực thông qua mạng TDNN.

Một cấu trúc mạng nơ-ron khác cho ứng dụng nhận dạng tiếng nói được trình bày trong hình 5.17. Cấu trúc này kết hợp khái niệm mạch lọc tương hợp (matched filter) với một mạng nơ-ron truyền thống để giải quyết vấn đề tính chất động của tín hiệu tiếng nói. Các đặc trưng âm học của tín hiệu tiếng nói được ước lượng thông qua kiến trúc mạng nơ-ron truyền thống; bộ phân loại mẫu sử dụng các véc-tơ đặc trưng âm học đã được phát hiện (với độ trễ thích hợp) và chập chúng với các mạch lọc tương hợp với các đặc trưng âm học và cộng dồn kết quả theo thời gian. Ở thời điểm thích hợp (tương ứng với thời điểm cuối của một số đơn vị tiếng nói được phát hiện hoặc được nhận dạng), các đơn vị đầu ra diễn tả tín hiệu tiếng nói.



Hình 5.16 Sơ đồ khối một mạng TDNN



Hình 5.17 Sơ đồ khối một hệ thống kết hợp mạng nơ-ron và mạch lọc tương hợp cho việc nhận dạng tiếng nói

Các mạng nơ-ron đã được xem xét và ứng dụng rộng rãi trong nhiều lĩnh vực bởi một số lý do sau:

- Các mạng nơ-ron có thể dễ dàng thực thi với cấp độ rất lớn các tính toán song song. Điều này là bởi vì cấu trúc mạng nơ-ron là một cấu trúc có tính song song cao của các thành phần tính toán tương tự nhau và đơn giản.

- Các mạng nơ-ron kế thừa bản chất là một cấu trúc chịu lỗi tốt (fault tolerance). Vì các thông tin nhúng trong mạng được trải (lan) đến mọi phần tử tính toán trong mạng, điều này

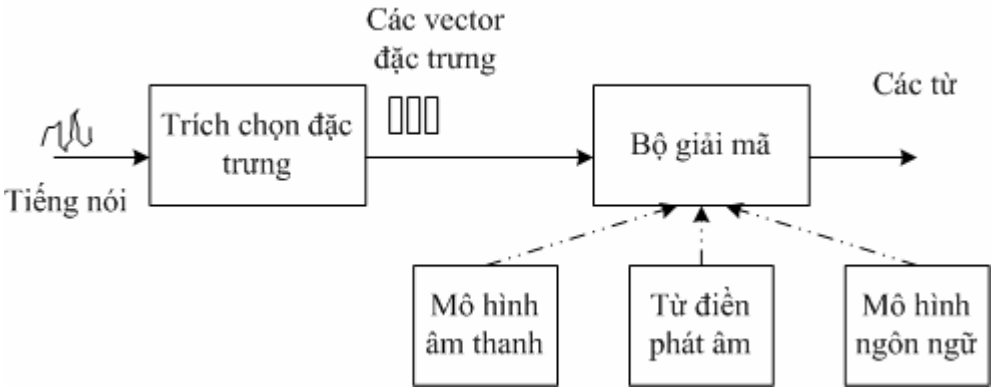
khiến cho cấu trúc khá trơ (least sensitive) với nhiễu hoặc các lỗi không hoàn hảo bên trong cấu trúc.

- Các trọng số kết nối trong mạng không bị hạn chế là phải cố định, chúng có thể thay đổi theo thời gian thực để nâng cao chất lượng của hệ thống. Đây chính là khái niệm cơ bản của việc học thích nghi có tính kế thừa từ cấu trúc của mạng nơ-ron.

- Bởi vì sự không tuyến tính bên trong mỗi phần tử tính toán, một mạng có cấu trúc đủ lớn có thể xấp xỉ (với sự khác biệt nhỏ bất kỳ) mọi cấu trúc không tuyến tính hoặc hệ thống động không tuyến tính. Nói một cách khác, các mạng nơ-ron cho phép thực hiện các phép biến đổi không tuyến tính giữa các tập đầu ra và đầu vào bất kỳ và thường trở lên hiệu quả hơn các phương pháp thực hiện vật lý các biến đổi không tuyên tính khác.

5.6.5 Hệ thống nhận dạng dựa trên mô hình Markov ẩn (HMM)

Hầu hết các hệ thống nhận dạng liên tục hiện nay dựa trên các mô hình Markov ẩn (HMM). Mặc dù nền tảng của các hệ thống nhận dạng liên tục (CSR) dựa trên HMM có trước hàng thập kỷ, đến gần đây mới có được một số tiến bộ trong việc cải thiện công nghệ để giảm nhỏ sự phụ thuộc của các giả thiết cố hữu và thích ứng các mô hình cho các ứng dụng và các môi trường nhất định.



Hình 5.18 Sơ đồ cấu trúc một hệ thống nhận dạng tiếng nói dựa trên mô hình HMM

Các thành phần chính của một hệ thống CSR làm việc với bộ từ vựng lớn được mô tả trong hình 5.18. Dạng sóng âm thanh đầu vào từ một mi-cờ-rô được chuyển đổi thành một dãy có độ dài cố định các véc-tơ âm $\mathbf{y} = y_1, ..., y_T$ nhờ một quá trình trích chọn mẫu. Bộ giải mã sau đó cố gắng tìm kiếm một dãy từ $\mathbf{w} = \mathrm{w}_1, ..., \mathrm{w}_K$ có khả năng cao nhất đã tạo ra $\mathbf{y}$. Nói cách khác, bộ giải mã cố gằng giải bài toán:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left[p(\mathbf{w} \mid \mathbf{y}) \right] \tag{5.31}$$

Tuy nhiên, vì $p(\mathbf{w} \mid \mathbf{y})$ rất khó xác định trong thực tế, do đó bằng cách áp dụng công thức Bayes chúng ta có:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left[p(\mathbf{y} \mid \mathbf{w})\, p(\mathbf{w}) \right] \tag{5.32}$$

Độ tương đồng $p(\mathbf{y} \mid \mathbf{w})$ được xác định bằng một mô hình âm và xác suất tiên nghiệm $p(\mathbf{w})$ được xác định bằng mô hình ngôn ngữ. Trong thực tế, mô hình âm (acoustic model)

không được chuẩn hóa và mô hình ngôn ngữ thường được tỷ lệ bằng một hằng số được xác định một cách thực nghiệm và một tham số bất lợi của việc chèn từ được thêm vào. Nói cách khác, lô-ga-rít của độ tương đồng tổng được tính bằng $\log\left(p(\mathbf{y}\,|\,\mathbf{w})\right)+\alpha p(\mathbf{w})+\beta|\,\mathbf{w}\,|$, trong đó α là giá trị phổ biến trong khoảng 8-20 và β phổ biến trong khoảng từ 0 đến -20. Đơn vị cơ bản của âm được biểu diễn bởi mô hình âm là âm vị (phone). Ví dụ từ *bat* trong tiếng Anh gồm ba âm vị là /b/, /ae/ và /t/. Đối với tiếng Anh cần có khoảng 40 âm vị như vậy.

Với mỗi **w** cho trước, mô hình âm tương ứng được tổng hợp bằng cách chắp nối các mô hình âm vị để tạo ra các từ như đã được quy định bằng một từ điển phát âm. Các thám số của các mô hình âm vị này được ước lượng từ các dữ liệu huấn luyện bao gồm các dạng sóng tín hiệu và các bản ghi hệ thống chính tả của chúng. Mô hình ngôn ngữ thường là một mô hình N-gram trong đó xác suất của mỗi từ chỉ phụ thuộc điều kiện vào N-1 thành phần trước nó. Các tham số của mô hình N-gram được ước lượng bằng cách đếm các tuýp N trong một tập (corpora: corpus - a collection of recorded utterances used as a basis for the descriptive analysis of a language) chữ thích hợp. Bộ giải mã hoạt động bằng cách tìm kiếm qua tất cả các dãy từ có thể, nó sử dụng phương pháp chặt (prune) để loại bỏ các giả thiết gần như không xảy ra và bằng cách đó giữ cho việc tìm kiếm có thể kiểm soát được. Khi việc tìm kiếm đến tiến đến phần cuối cùng, dãy từ có sự tương đồng nhất chính là kết quả. Trong các bộ giải mã hiện đại, thay vì sử dụng các phương pháp vừa nêu, bộ giải mã sinh ra các lưới chứa các biểu diễn gọn của hầu hết các giả thiết có khả năng nhất.

a) Trích chọn đặc trưng

Như đã đề cập, việc trích chọn đặc trưng tìm các tạo ra một biểu diễn (thường là dạng mã hóa) tối ưu tín hiệu tiếng nói. Quá trình này cũng phải đảm bảo giảm thiểu sự mất mát thông tin và tạo ra một sự phù hợp tốt nhất với các giả thiết phân tán tạo ra bởi các mô hình âm. Các véc-tơ đặc trưng thường được tính toán trong mỗi khung có độ dài khoảng 10ms và sử dụng các hàm cửa sổ phân tích chồng lấn nhau. Phương pháp trích trọn phổ biến nhất trong các ứng dụng nhận dạng sử dụng mô hình HMM là phương pháp MFCC như đã trình bày trong phần trên.

b) Các mô hình âm học HMM

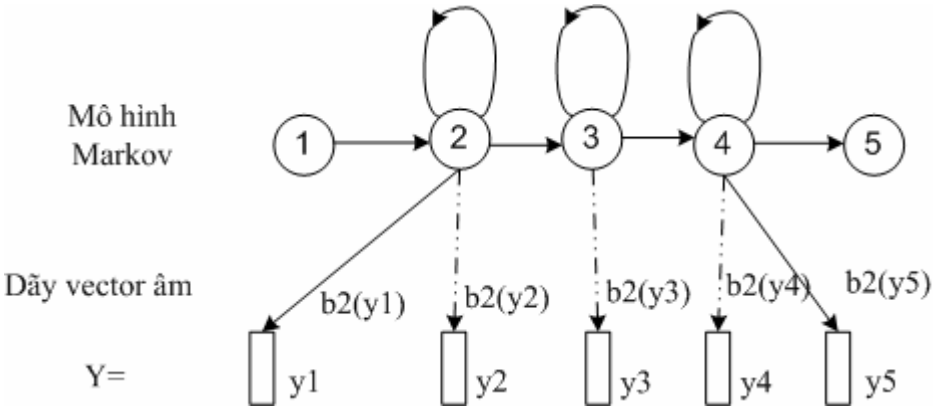
Như đã đề cập, các từ được phát ra trong **w** được phân tách thành một dãy các âm cơ bản được gọi là các âm vị cơ sở. Để cho phép các thay đổi phát âm có thể, độ tương đồng $p(\mathbf{y}\,|\,\mathbf{w})$ có thể được tính trên các phương án phát âm:

$$p(\mathbf{y}\,|\,\mathbf{w})=\sum_{Q} p(\mathbf{y}\,|\,Q)\,p(Q\,|\,\mathbf{w}) \tag{5.33}$$

Các bộ nhận dạng thường xấp xỉ công thức này bằng phép tính cực đại do đó các phương pháp phát âm khác nhau có thể được giải mã như thể chúng là các giả thiết từ thay thế. Mỗi Q là một dãy các phát âm của từ Q_1, …, Q_K trong đó mỗi phương án phát âm là một dãy các âm vị cơ sở $Q_k=q_1^{(k)},q_2^{(k)},\ldots$. Khi đó chúng ta có:

$$p(Q\,|\,\mathbf{w})=\prod_{k=1}^{K} p(Q_k\,|\,w_k) \tag{5.34}$$

Ở đây $p(Q_k \mid w_k)$ là xác suất từ w_k được phát âm dựa trên dãy các âm vị cơ sở Q. Trong thực tế, chỉ có rất ít số khả năng có thể các phương án phát âm Q_k cho mỗi từ w_k, điều này cho phép tổng (5.33) dễ dàng kiểm soát được.



Hình 5.19 Mô hình âm vị cơ sở dựa trên mô hình HMM

Mỗi âm cơ sở q được biểu diễn bởi một mô hình Markov ẩn mật độ liên tục (HMM) được minh họa trong hình 5.19. Trong minh họa này, các tham số dịch chuyển là $\{a_{ij}\}$ và các phân bố quan sát đầu ra $\{b_j(\ )\}$. Các phân bố quan sát đầu ra thường là sự pha trộn của các phân bố chuẩn Gausse:

$$b_j(y) = \sum_{m=1}^{M} c_{jm} \aleph\left(y; \mu_{jm}, \sum_{jm}\right) \tag{5.35}$$

$\aleph$ biểu diễn phân bố chuẩn với giá trị trung bình μ_{jm} và covariance $\sum_{jm}$. Số lượng các thành phần trong công thức (5.35) thường lấy trong khoảng 10 đến 20. Vì kích thước của các véc-tơ âm **y** thường tương đối lớn, các covariance thường được giới hạn là các ma trận đường chéo. Các trạng thái đầu và kết thúc là các trạng thái không phát xạ (nonemitting) và chúng được thêm vào nhằm đơn giản hóa quá trình chắp nối các mô hình âm vị để tạo ra các từ.

Cho trước một HMM tổng hợp với Q được tạo ra bằng các chắp nối tất cả các âm vị cơ sở cấu thành, độ tương đồng âm được tính bởi:

$$p(y \mid Q) = \sum_{X} p(x, y \mid Q) \tag{5.36}$$

Trong đó $X = x(0), ..., x(T)$ là một dãy các trạng thái trong toàn bộ mô hình tổng hợp và

$$p(x, y \mid Q) = a_{x(0),x(1)} \prod_{t=1}^{T} b_{x(t)} a_{x(t),x(t+1)} \tag{5.37}$$

Các tham số mô hình âm $\{a_{ij}\}$ và $\{b_j(\ )\}$ có thể được ước lượng một cách hiệu quả từ tập các bộ huấn luyện bằng phương pháp cực đại kỳ vọng.

5.7. Bài thực hành nhận dạng tiếng nói

Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) thực hiện các công việc sau:

- Xây dựng hệ thống nhận dạng tiếng nói đơn giản (từ vựng hạn chế) dựa vào:
    - Mạng nơ-ron
    - Mô hình HMM

Phụ lục 1: Mạng nơ-ron

Mở đầu

Hoạt động nghiên cứu về cơ chế hoạt động, cấu trúc bộ não con người được chú ý khá sớm. Cùng với sự phát triển của khoa học, chúng ta đã đạt được một số bước tiến quan trọng trong lĩnh vực nghiên cứu này. Tuy nhiên, bộ não con người là một tổ hợp rất phức tạp và cho đến nay hiểu biết của con người về kiến trúc và hoạt động của não vẫn còn chưa đầy đủ. Mặc dù vậy con người ta tạo ra được các máy có một số tính năng tương tự não nhờ mô phỏng các đặc điểm:

- Tri thức thu nhận được nhờ quá trình học
- Tính năng có được nhờ kiến trúc mạng và tính chất kết nối

Các máy mô phỏng này có tên chung là mạng nơ-ron nhân tạo hay đơn giản là mạng nơron. Đặc điểm chính của các mạng nơ-ron:

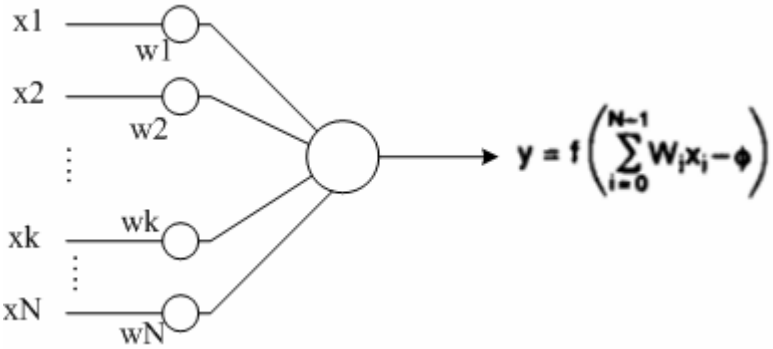
- Phi tuyến. Cho phép xử lý phi tuyến.
- Cơ chế ánh xạ đầu vào - đầu ra cho phép học có giám sát.
- Cơ chế thích nghi. Thay đổi tham số phù hợp với môi trường.
- Đáp ứng theo mẫu huấn luyện.
- Thông tin theo ngữ cảnh.Tri thức được biểu diễn tuỳ theo trạng thái và kiến trúc của mạng.
- Cho phép có lỗi (fault tolerance).
- Phỏng sinh học

Cơ sở mạng về Nơ-ron

Sơ đồ một mạng nơ-ron đơn giản được minh họa trong hình A.1. Giả sử có N đầu vào được đánh nhãn x_1, x_2, …, x_N với các trọng số tương ứng là w_1, w_2, …, w_N. Khi đó quan hệ phi tuyến đầu vào đầu ra được xác định như sau:

$$y = f\left(\sum_{i=1}^{N} w_i x_i - \eta\right)$$

Trong đó η là mức ngưỡng nội tại hay còn gọi là offset, $f(.)$ là một hàm phi tuyến.



Hình A.1: Cấu trúc đơn giản của một mạng nơ-ron N đầu vào

Một số dạng phổ biến của f có thể có dạng như sau:

1. Hàm ngưỡng cứng:

$$f(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

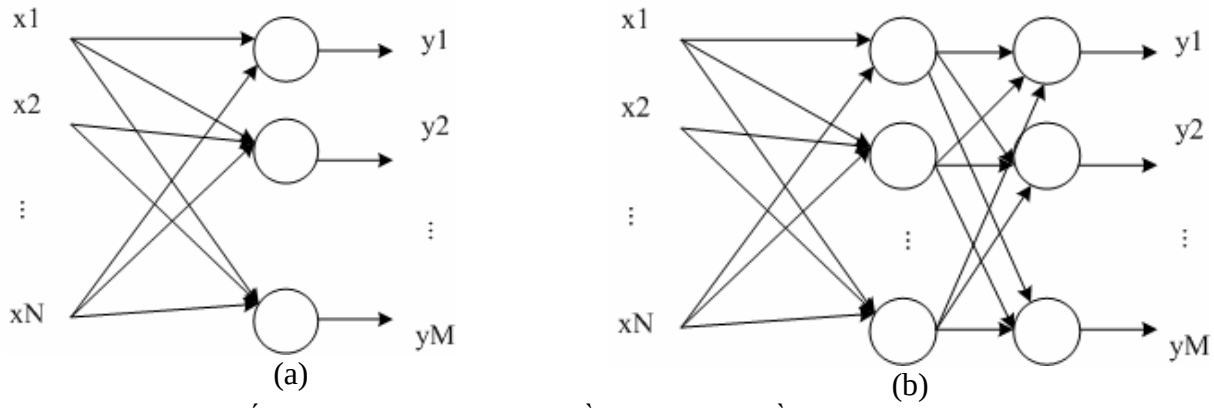
2. Hàm log-sin

$$f(x)=\frac{1}{1+e^{-\beta x}} \quad (\beta>0)$$

Cấu hình mạng Nơ-ron

Một yếu tố quan trọng cho việc thiết lập và ứng dụng của mạng nơ-ron là cấu trúc tô-pô của mạng (network topology). Có ba kiểu cấu trúc cơ bản là:

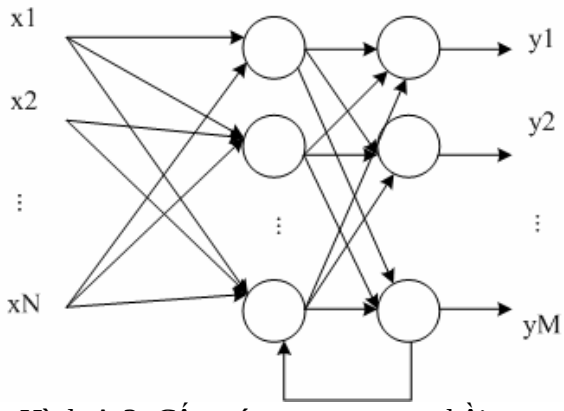
1) Mạng một tầng hoặc nhiều tầng:



(a) (b)

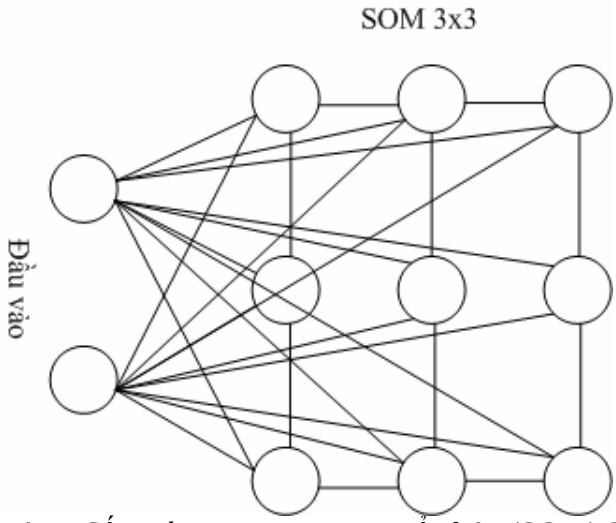
Hình A.2: Cấu trúc mạng nơ-ron một tầng (a) và hai tầng (b)

2) Mạng hồi quy:



Hình A.3: Cấu trúc mạng nơ-ron hồi quy

3) Mạng tự tổ chức:



Hình A.4: Cấu trúc mạng nơ-ron tự tổ chức (SOM) 3x3

Phụ lục 2: Mô hình Markov ẩn

Quá trình Markov

Một quá trình ngẫu nhiên X(t) được gọi là một quá trình Markov nếu tương lai của một quá trình với trạng thái hiện tại đã cho không phụ thuộc vào quá khứ của quá trình. Nói một cách khác, với các thời gian xác định $t_1 < t_2 < ... < t_k < t_{k+1}$ thì:

$$\begin{aligned}\Pr\left[X\left(t_{k+1}\right) = x_{k+1} \mid X\left(t_k\right) = x_k, ..., X\left(t_1\right) = x_1\right] \\ = \Pr\left[X\left(t_{k+1}\right) = x_{k+1} \mid X\left(t_k\right) = x_{k1}\right]\end{aligned}$$

Các giá trị của $X\left(t\right)$ tại thời điểm t thường được gọi là trạng thái của quá trình tại thời điểm t.

Chuỗi Markov với thời gian rời rạc

Giả sử X_n là một chuỗi Markov với giá trị nguyên và thời gian rời rạc với trạng thái bắt đầu tại n=0 có hàm phân bố xác suất rời rạc (pmf):

$$p_j\left(0\right) \triangleq \Pr\left[X_0 = j\right] \quad \left(j = 0,1,...\right)$$

Khi đó, hàm mật độ phân bố xác suất rời rạc hợp của n+1 giá trị đầu tiên của quá trình được tính bằng:

$$\begin{aligned}&\Pr\left[X_n = i_n, ..., X_0 = i_0\right] \\ &\quad = \Pr\left[X_n = i_n \mid X_{n-1} = i_{n-1}\right] ... \Pr\left[X_1 = i_1 \mid X_0 = i_0\right] \Pr\left[X_0 = i_0\right]\end{aligned}$$

Từ công thức trên chúng ta thấy, hàm mất độ phân bố xác suất rời rạc hợp của một dãy xác định là tích của xác suất của trạng thái khởi đầu và các xác suất của các dãy con chuyển đổi trạng thái một bước.

Giả sử các xác suất chuyển đổi trạng thái một bước là cố định và không thay đổi theo thời gian, nghĩa là:

$$\Pr\left[X_{n+1} = j \mid X_n = i\right] = a_{\mathrm{ij}} \qquad \forall n$$

Khi đó X_n được nói là có các xác suất chuyển đổi đồng nhất. Khi đó xác suất phân bố rời rạc hợp cho $X_n, ..., X_0$ trở thành:

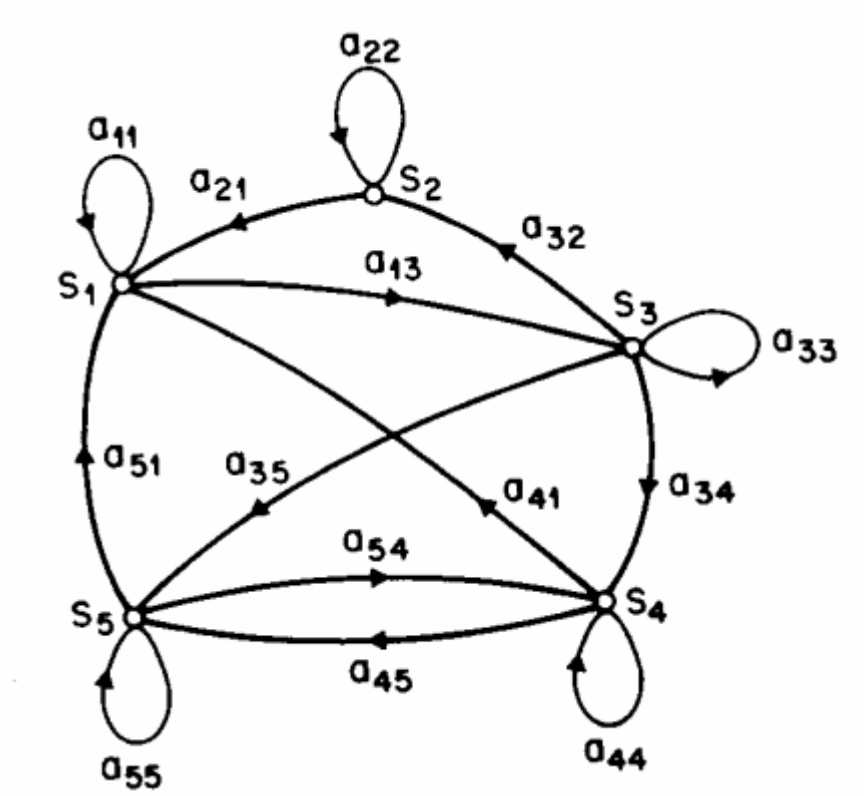
$$\Pr\left[X_n = i_n, ..., X_0 = i_0\right] = a_{i_{n-1} i_n} ... a_{i_0 i_1} p_{i_0}\left(0\right)$$

Như vậy, X_n hoàn toàn được xác định bởi hàm mật độ phân bố xác suất rời rạc khởi đầu $p_i\left(0\right)$ và ma trận các xác suất chuyển một bước $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & ... \\ a_{10} & a_{11} & a_{12} & ... \\ \vdots & \vdots & \vdots & \ddots \\ a_{i0} & a_{i1} & a_{i2} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

P được gọi là ma trận xác suất chuyển. Chú ý rằng, tổng của mỗi hàng của **P** phải bằng 1.

Hình B.1 minh họa sơ đồ một chuỗi Markov rời rạc với 5 trạng thái được gán nhãn S_1 – S_5 và các xác suất chuyển tương ứng là nhãn các nhánh a_{ij}.



Hình B.1: Minh họa một chuỗi Markov rời rạc với 5 trạng thái

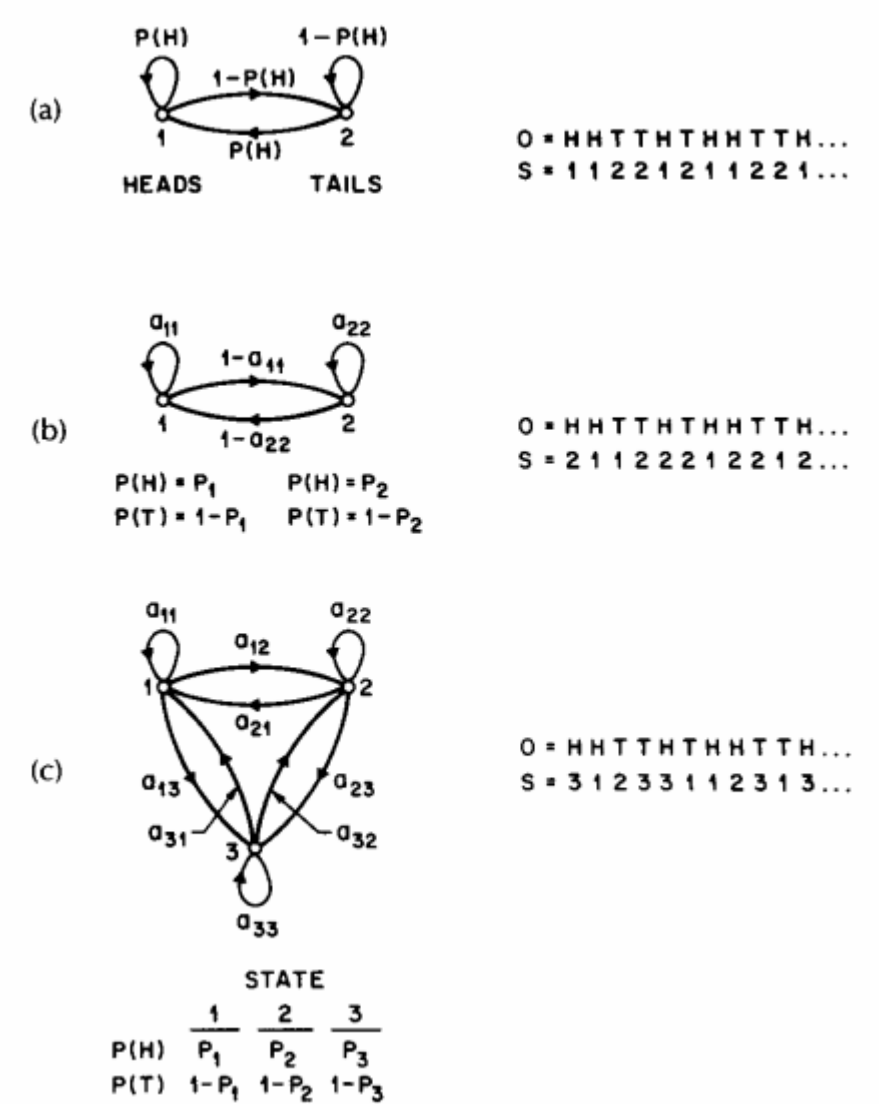
Mô hình Markov ẩn

Trong phần trên chúng ta ví dụ về mô hình Markov mà mỗi trạng thái tương ứng với một sự kiện (vật lý) quan sát được. Tuy nhiên các mô hình như vậy có ứng dụng hạn chế trong các bài toán thực tế. Do đó, mô hình được mở rộng bao gồm cả những trường hợp việc quan sát là một hàm xác suất của trạng thái - tức là mô hình là một quá trình thống kê chồng kép với một quá trình thống kê bên trong mà không quan sát được (ẩn sâu bên trong), nhưng có thể chỉ quan sát được thông qua một tập các quá trình thống kê khác, các quá trình mà tạo ra dãy các quan sát được. Mô hình như vậy được gọi là mô hình Markov ẩn (HMM).

Để minh họa, chúng ta xét ví dụ các mô hình tung đồng xu như sau. Một người thực hiện việc tung đồng xu nhưng không nói cho chúng ta biết anh ta đã làm chính xác những gì. Anh ta chỉ thông báo cho chúng ta kết quả của mỗi đồng xu lật. Như vậy, đối với chúng ta, một loạt các thí nghiệm tung đồng xu được ẩn dấu, mà chỉ có dãy quan sát được về nó là dãy các kết quả chẵn và lẻ. Vấn đề đặt ra làm sao xây dựng một mô hình HMM thích hợp để mô hình dãy chẵn và lẻ quan sát được. Vấn đề đầu tiên là việc quyết định các trạng thái nào trong mô hình tương ứng với và sau đó là quyết định bao nhiêu trạng thái cần thiết trong mô hình.

Hình B.2 minh họa 3 trường hợp ví dụ. Trường hợp thứ nhất tương ứng với giả thiết chỉ một động xu không cân được tung. Mô hình trong trường hợp này là mô hình hai trạng thái trong đó mỗi trạng thái tương ứng với một mặt của đồng xu. Dễ thấy rằng, mô hình Markov trong trường hợp này là quan sát được Cũng cần chú ý rằng, chúng ta có thể sử dụng

mô hình Markov một trạng thái trong đó trạng thái tương ứng với một đồng xu không cân đơn lẻ, và tham số chưa biết là sự không cân của đồng xu.



Hình B.2: Minh họa ba mô hình Markov có thể đối với thí nghiệm tung đồng xu ẩn

Trường hợp thứ hai tương ứng với mô hình hai trạng thái trong đó mỗi trạng thái tương ứng với một đồng xu không cân khác nhau được tung. Mỗi trạng thái được đặc trưng bởi một phân bố xác suất của mặt chẵn và mặt lẻ, và các chuyển đổi giữa các trạng thái được đặc trưng bởi một ma trận chuyển trạng thái.

Trường hợp thứ ba tương ứng với thí nghiệm sử dụng ba đồng xu không cân khác nhau, và việc chọn một trong ba đồng xu này được dựa trên một sự kiện xác suất.

Với một lựa chọn một trong ba trường hợp trên để giải thích dãy mặt chẵn và mặt lẻ quan sát được, câu hỏi đặt ra là mô hình nào mô phỏng tương đồng nhất với các quan sát thực tế. Chúng ta thấy rằng, mô hình trong trường hợp một chỉ có một tham số chưa biết, hay nói cách khác, bậc tự do chỉ bằng một. Trong khi đó các mô hình trường hợp hai và ba có bậc tự do tương ứng là 4 và 9. Do đó, với bậc tự do lớn hơn, mô hình HMM lớn hơn sẽ dường như có khả năng hơn trong việc mô tả một dãy các thí nghiệm tung xu so với các mô hình nhỏ hơn. Tuy nhiên cũng cần chú ý, điều nhận xét trên là đúng về mặt lý thuyết, trong thực tế có một số hạn chế mạnh với kích thước của mô hình.

Một HMM được đặc trưng bởi:

1) Số các trạng thái trong mô hình N. Mặc dù các trạng thái là ẩn, nhưng với một số ứng dụng thực tế thường có một số ý nghĩa vật lý gắn với các trạng thái hoặc một tập các trạng thái của mô hình.

2) Số các ký hiệu quan sát phân biệt với mỗi trạng thái, tức là kích thước bộ chữ rời rạc.
3) Phân bố xác suất chuyển trạng thái **P** trong đó $a_{ij} = \Pr\left[X_{n+1} = S_j \mid X_n = S_i \right]$, $\left(1 \le i, j \le N\right)$. Trong trường hợp đặc biệt trong đó một trạng thái bất kỳ có thể đạt đến bất kỳ trạng thái nào khác trong một bước duy nhất, chúng ta có $a_{ij} > 0$ với mọi i, j. Với các loại HMM khác, chúng ta có $a_{ij} = 0$ cho một hoặc nhiều hơn một cặp (i,j).
4) Phân bố xác suất ký hiệu quan sát ở trạng thái j, $B = \{b_j(k)\}$, trong đó $b_j(k) = \Pr\left[v_k(t) \mid X_t = S_j \right]$, $\left(1 \le j \le N, 1 \le k \le M\right)$.
5) Phân bố trạng thái khởi đầu $\pi = \{\pi_i\}$ trong đó $\pi_i = \Pr\left[X_1 = S_i \right]$, $\left(1 \le i \le N\right)$.

Với các giá trị của N, M, P, B và π cho trước, HMM có thể được sử dụng như một bộ tạo cho một dãy quan sát $O = O_1 O_2 ... O_T$ (với mỗi quan sát O_t là một ký hiệu từ tập v và T là số các quan sát trong dãy) như sau:

1) Chọn một trạng thái khởi đầu $X_1 = S_i$ theo phân bố trạng thái khởi đầu π.
2) Đặt t=1.
3) Chọn $O_t = v_k$ theo phân bố xác suất ký hiệu ở trạng thái S_i, tức là $b_i(k)$.
4) Chuyển sang trạng thái mới $X_{t+1} = S_j$ theo phân bố xác suất chuyển trạng thái cho trạng thái S_j, tức là a_{ij}.
5) Đặt t=t+1; trở lại bước 3 nếu t<T; nếu không kết thúc quá trình.

Tài liệu tham khảo

[1]. John R. Deller, John H. L. Hassen, and John G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE Press, 2000.

[2]. *Editors*: Rainer Martin, Ulrich Heuter and Christiane Antweiler, *Advances in Digital Speech Transmission*, Wiley, 2008.

[3]. Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[4]. *Editors* Jacob Benesty, M. Mohan Sondhi and Yiteng Huang, *Handbook of Speech Processing*, Springer-Verlag Berlin, 2008.

[5]. Antonio M. Peinado and Jose C. Segura, *Speech Recognition over Digital Channels: Robustness and Standards*, John Wiley \& Sons, 2006.

[6]. John Holmes and Wendy Holmes, *Speech Synthesis and Recognition*, second edition, Taylor and Francis, 2001.

[7]. Paul Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[8]. Lawrence R. Rabiner and Ronald W. Schafer, *Introduction to Digital Speech Processing*, Now Publishers Inc., 2007.

[9]. Lawrence R. Rabiner and Ronald Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[10]. Sadaoki Furui, *Digital Speech Processing, Synthesis, and Recognition*, second edition, Marcel Dekker Inc., 2001.

[11]. Lawrence R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of the IEEE, Vol.77, No.2, Feb. 1989, pp.257-286.