BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

****

BÀI GIẢNG

XỬ LÝ TIẾNG NÓI

BIÊN SOẠN:

PHẠM VĂN SỰ

LÊ XUÂN THÀNH

HÀ NỘI - 2014

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 khoa học, 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, hệ thống xử lý tiếng nói. 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 hơn nữa việc hoàn thiện 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 trở nên 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 cũng như Khoa học Máy tính 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 một cách tổng quan về các công nghệ tiên tiến, xu thế nghiên cứu và phát triển của lĩnh vực xử lý tiếng nói. Trong lần tái bản này, cuốn sách được phân chia lại thành 5 chương:

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

Cuốn bài giảng này là những kinh nghiệm đúc rút của các tác giả trong quá trình giảng dạy và nghiên cứu tại Học viện Công nghệ Bưu chính Viễn thông. Cuốn bài giảng còn là kết quả của những nỗ lực đóng góp đầy nhiệt huyết của các thầy cô giáo, những đồng nghiệp tại Khoa Kỹ thuật Điện tử, của các em sinh viên. 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 Xử lý Tín hiệu và Truyền thông, 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 supv@ptit.edu.vn.

Hà Nội, tháng 12 năm 2014

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

DANH MỤC CÁC TỪ VIẾT TẮT

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

MỤC LỤC

CHƯƠNG 1. MỘT SỐ KHÁI NIỆM CƠ BẢN

1.1. MỞ ĐẦU

Tiếng nói là phương tiện trao đổi thông tin chính yếu giữa con người và con người. Phương thức thông tin bằng tiếng nói được sử dụng một cách rộng rãi. Việc trao đổi thông tin thông qua tín hiệu tiếng nói cho phép truyền tải thông tin một cách nhanh chóng hơn. Một người bình thường có thể nói trung bình hơn 100 từ trong một phút, trong khi đó chỉ có thể viết được trung bình khoảng 50 từ trong vòng một phút.

Thông tin tiếng nói đơn giản mà hiệu quả. Tiếng nói là phương tiện trao đổi đầy ma lực: Bản thân ngôn từ (cách hành văn) đã vốn chứa đựng một sắc thái biểu cảm, nhưng thông qua ngôn ngữ nói nó còn có khả năng truyền tải cả sắc thái, thái độ (vui, buồn,...)

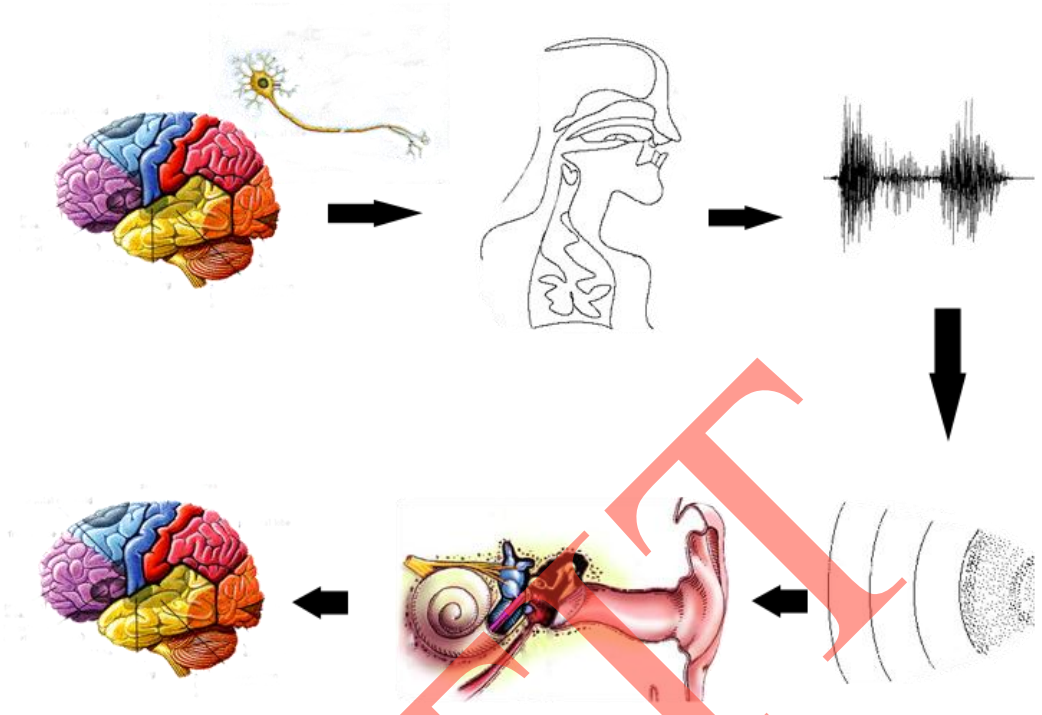
Mặt khác, con người có vẻ ngày càng lười hơn. Nhu cầu sử dụng tiếng nói thay vì các thao tác bằng tay để thực hiện công việc, chẳng hạn như điều khiển, đang tăng một cách mạnh mẽ hơn bao giờ hết. Điều này đặc biệt càng đúng với sự phát triển nhanh chóng của công nghệ khoa học hiện nay. Chúng ta không còn lạ lẫm với các ứng dụng điều khiển các thiết bị trong nhà thông minh bằng cử chỉ và giọng nói. Thậm chí, Google còn cho phép chúng ta có khả năng lái xe bằng cách chỉ cần ra lệnh bằng giọng nói.

Để có thể phát huy được thế mạnh, sự tiện dụng của phương tiện giao tiếp này, đặc biệt là có thể hiểu, nắm bắt và từng bước có khả năng xây dựng và triển khai các hệ thống giao tiếp bằng giọng nói thì rất cần thiết phải có được những kiến thức cơ bản về xử lý tiếng nói. Trong chương này, trước hết chúng ta sẽ làm quen với một số khái niệm cơ bản của hệ thống xử lý tiếng nói. Những khái niệm cơ bản này sẽ là nền tảng để nghiên cứu và tìm hiểu sâu hơn trong các chương tiếp theo.

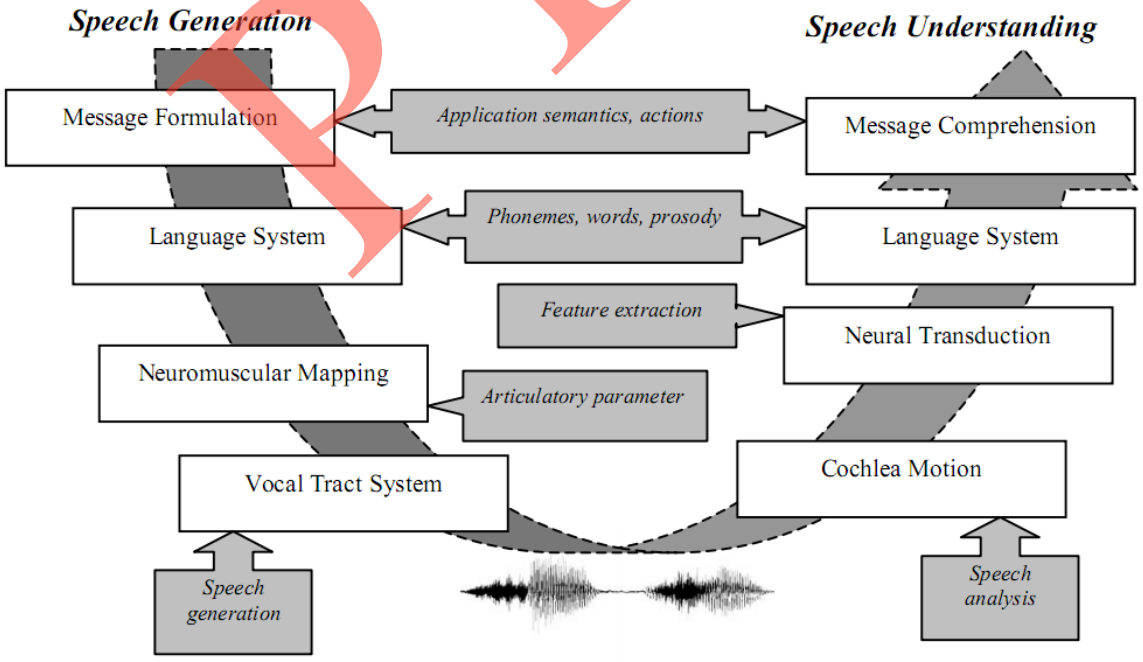
1.2. TỔNG QUAN VỀ XỬ LÝ TIẾNG NÓI

Để đơn giản có cái nhìn tổng quát về hệ thống xử lý tiếng nói và trả lời được câu hỏi "Xử lý tiếng nói là gì?", hãy quan sát quá trình chúng ta thực hiện giao tiếp bằng giọng nói. Nếu chúng ta đóng vai trò người nói, những thông điệp mong muốn truyền tải được định hình tại bộ não. Não sẽ thực hiện việc phân tích thông điệp này và đưa các tín hiệu để điều khiển các bộ phận phát âm tương ứng hoạt động nhằm "tổng hợp" ra âm thanh mong muốn để truyền tải thông điệp. Ở phía người nghe, âm thanh mang thông tin được thu nhận bởi cơ quan cảm thụ sẽ cảm thụ, thông qua các tín hiệu thần kinh truyền đến não để "nhận dạng" và "suy diễn" nhằm hiểu thông tin. Một cách tổng quát, hệ thống thông tin bằng tiếng nói của con người có thể mô tả như hình 1.1. Mặc dù cho đến nay, con người vẫn chưa hoàn toàn hiểu một cách toàn diện về quá trình tạo, cảm nhận

tiếng nói của con người nhưng một số quá trình và cách thức thực hiện cơ bản có thể được tóm lược như hình 1.2.



Hình 1.1 Sơ lược hệ thống thông tin tiếng nói của con người



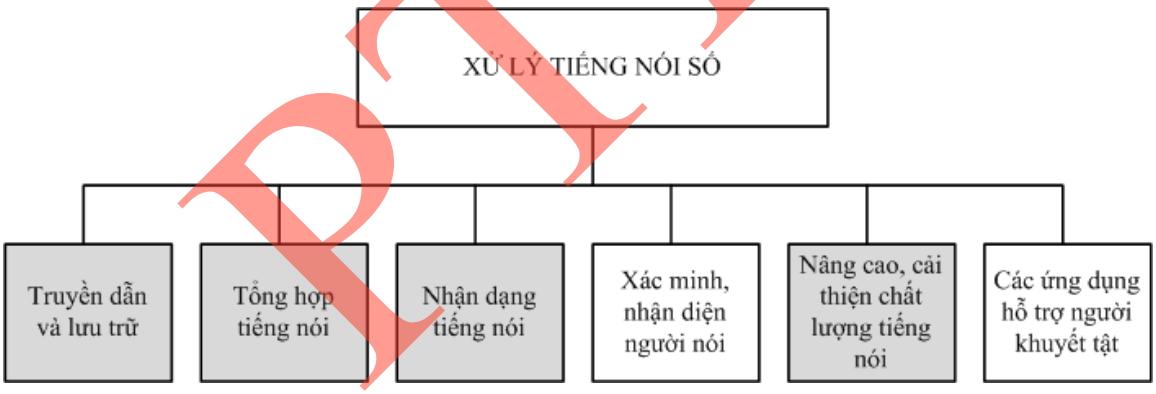
Hình 1.2 Tóm lược một số quá trình xử lý trong hệ thống thông tin bằng tiếng nói

Như vậy, bản chất của "xử lý tiếng nói" là việc thực hiện các phép thao tác nào đó nhằm tạo ra tiếng nói để truyền tải tin tức, và/hoặc bóc tách thông tin từ tín hiệu tiếng nói.

Từ bản chất nói trên, chúng ta có thể dễ dàng xây dựng các hệ thống xử lý tiếng nói trong đó có thể tái tạo một phần hoặc toàn bộ các thao tác xử lý của hệ thống thông tin tiếng nói tự nghiên.

Nói tóm lại, xử lý tiếng nói là lĩnh vực khoa học nghiên cứu về tiếng nói (cả khía cạnh ngôn ngữ và khía cạnh tín hiệu), và các phương pháp xử lý các khía cạnh của tiếng nói.

Cũng như vốn dĩ sự phức tạp của hệ thống thông tin tiếng nói (ngôn ngữ) của con người, xử lý tiếng nói là một lĩnh vực phức tạp và bao trùm tương đối rộng. Đầu tiên có thể kể đến là xử lý tín hiệu tiếng nói về mặt vật lý như giảm/loại bỏ nhiễu, giảm méo, … trong lĩnh vực tăng cường nâng cao chất lượng tiếng nói nhằm cải thiện tín dễ nghe dễ hiểu của tín hiệu tiếng nói. Hoặc có thể kể đến là việc tìm cách biểu diễn tín hiệu tiếng nói ở dạng tín hiệu số sao cho dung lượng nhỏ nhất trong lĩnh vực mã hóa lưu trữ và truyền tải tín hiệu thoại. Không chỉ dừng lại ở đó, khi công nghệ phát triển, xử lý tiếng nói cho phép các hệ thống có thể tái tạo tiếng nói (tổng hợp tiếng nói), hiểu được tiếng nói (nhận dạng tiếng nói). Hình 1.3 mô tả tóm lược các lĩnh vực chủ yếu của xử lý tiếng nói số.



Hình 1.3 Một số lĩnh vực cơ bản của Xử lý tiếng nói số

1.3. QUÁ TRÌNH TẠO VÀ CẢM NHẬN TIẾNG NÓI

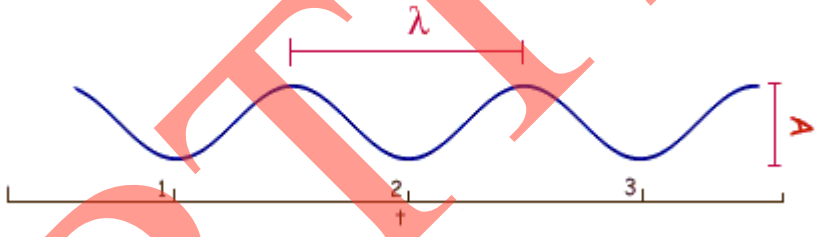
Như đã đề cập ở phần đầu của chương, tiếng nói là một phương tiện thông tin hiệu quả, nhưng quá trình xử lý cũng rất phức tạp. Để có thể hiểu và có thể áp dụng tốt những kỹ thuật, phương pháp xử lý cho tín hiệu tiếng nói, chúng ta không thể không hiểu về quá trình tạo và cảm nhận tiếng nói của con người. Những hiểu biết về cách thức xử lý tuyệt vời của hệ thống cảm nhận của hệ thống phát âm, hệ thống thính giác của con người sẽ là một tham khảo đáng giá. Hơn nữa, một số đặc tính cảm nhận và xử lý có thể sẽ tạo những cơ hội xử lý thuận tiện và hiệu quả nếu được khai thác một cách hợp lý.

1.3.1 Bản chất của tiếng nói

Âm thanh tiếng nói cũng như âm thanh nói chung trong thế giới tự nhiên xung quanh ta, về bản chất đều là những sóng âm được lan truyền trong một môi trường vật lý nhất định (thường là không khí).

Tuy nhiên đó là những hiểu biết phía bên ngoài, phần kết quả, về hệ thống tạo tín hiệu tiếng nói. Để đơn giản, chúng ta bỏ qua khía cạnh tâm thần (neurology) của quá trình tạo tiếng nói. Do đó, có thể coi nguồn gốc của quá trình tạo tín hiệu tiếng nói là quá trình hoạt động của hệ thống phát âm. Khi ta nói dây thanh trong hầu dao động. Những dao động này được truyền qua hệ thống tuyến âm, một hệ thống đóng vai trò như một bộ lọc cơ học, tạo nên những sóng âm truyền tải thông tin tiếng nói. Sóng âm này, về bản chất là những dao động cơ học, lan truyền trong không khí đến phía người nghe.

Như chúng ta đã được học trong chương trình vật lý phổ thông, sóng âm là sóng cơ học và thuộc loại sóng dọc. Sóng âm chỉ có thể lan truyền trong môi trường có vật chất (không khí, nước, …). Về cơ bản nó cũng có các tham số như một sóng cơ học thông thường như tần số, chu kỳ, bước sóng. Một số tham số cơ bản của sóng được minh họa trong hình 1.4.



Hình 1.4 Một số tham số cơ bản của sóng cơ học

Cũng cần lưu ý rằng, sóng âm thanh tiếng nói phức tạp hơn rất nhiều. Bản chất của sự thay đổi liên tục để truyền tải thông điệp khiến cho các tham số cơ bản đề cập ở trên luôn thay đổi thậm chí ngay trong khoảng thời gian rất ngắn.

Sóng âm thanh mà con người có thể cảm nhận được nằm trong một dải tần số rất rộng, khoảng từ 16Hz đến 20000Hz. Những sóng âm dao động có tần số nhỏ hơn 16Hz được gọi là sóng hạ âm. Những sóng âm có tần số lớn hơn 20000Hz được gọi là sóng siêu âm. Mặc dù hầu hết con người không cảm nhận được sóng hạ âm và không sử dụng trong thông tin, một số người có khả năng cảm nhận sóng hạ âm sẽ có những cảm giác bồn chồn lo lắng áp lực. Cũng tương tự, con người không cảm nhận được sóng siêu âm, nhưng sóng siêu âm có khá nhiều ứng dụng thực tế như phát hiện chẩn đoán trong ảnh y tế, định vị phát hiện kẻ thù trong hệ thống sonar trên các tàu ngầm, …

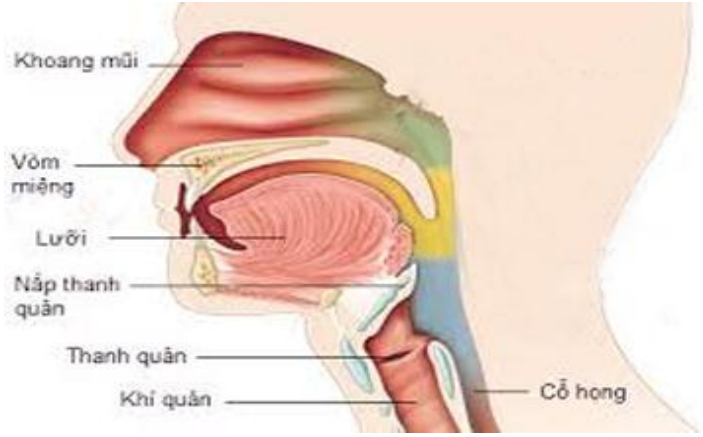
1.3.2 Cấu tạo của hệ thống phát âm

Tiếng nói là kết quả của sự phối hợp hoạt động giữa não, hệ dây thần kinh và các bộ phận trong hệ thống phát âm. Hệ thống phát âm gồm hai phần chính là phổi và hệ thống tuyến âm.

Phổi có nhiệm vụ giãn/ép hơi nhằm tạo lực cần thiết cho dây thanh thực hiện dao động. Nó được coi là nguồn kích thích dao động của dây thanh. Khi nói, lồng ngực mở rộng và thu hẹp, không khí được đẩy từ phổi vào khí quản, luồng khí này bị ép và đi qua cặp dây thanh tạo ra dao động. Dao động này tạo ra sự xáo trộn của luồng hơi, sau khi truyền qua hệ thống tuyến âm thì phát xạ ra ở môi.

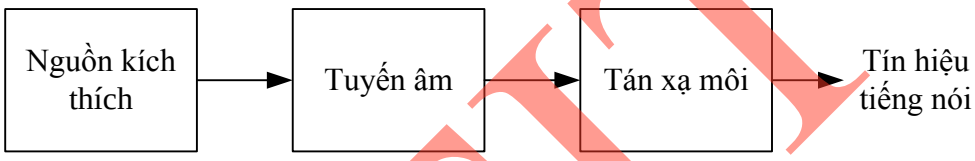
Tuyến âm có thể được coi như một ống âm học (gồm các đoạn ống với độ dài bằng nhau và thiết diện các mặt cắt khác nhau mắc nối tiếp, còn gọi là bộ lọc cơ học) với đầu vào là các dây thanh (còn gọi là thanh môn) và đầu ra là môi. Hình 1.5 minh họa cấu trúc và các bộ phận của hệ thống tuyến âm. Tuyến âm có hình dạng thay đổi và được điều khiển co thắt để 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 tiết 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 một tuyến âm phụ trợ cho việc 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.

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. Cặp môi này có thể khép kín hoàn toàn thanh quản hoặc mở ra tạo ra độ mở hình tam giác gọi là thanh môn. Bình thường không khí qua thanh quản một cách tự do trong quá trình thở hoặc trong quá trình phát âm những âm câm hoặc vô thanh. Khi phát âm những âm hữu thanh, cặp môi này đóng mở liên tục một cách không tuần hoàn (còn gọi là dao động) để tạo ra âm thanh. Những rung động dây thanh liên tiếp được truyền qua tuyến âm. Dao động dây thanh sẽ được điều biến thông qua sự thay đổi hình dạng và tiết diện của tuyến âm để tạo ra những âm khác nhau.



Hình 1.5 Hệ thống phát âm của con người

Tóm lại, tín hiệu tiếng nói được tạo ra từ hệ thống phát âm của con người có thể mô tả đơn giản là một quá trình gồm ba khối như hình 1.6.



Hình 1.6 Quá trình cơ bản tạo tín hiệu tiếng nói

1.3.3 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 bằng tiếng nói, thông tin tiếng nói gồm có nhiều câu nói, mỗi câu gồm nhiều từ, mỗi từ lại có thể gồm một hay nhiều đơn vị âm. Để thuận tiện trong quá trình nghiên cứu, người ta thực hiện việc phân chia tiếng nói theo một số đặc trưng. Tùy theo các đặc trưng được sử dụng để phân loại mà chúng ta có các loại âm thanh tiếng nói khác nhau. Một cách đơn giản nhất là dựa vào đặc trưng phát âm, người ta chia tiếng nói thành 3 loại cơ bản như sau:

- **Âm hữu thanh**: Là âm khi phát ra có thanh, ví dụ như ta phát âm những nguyên âm như "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 với tần số cơ bản.

Âm vô thanh: Là âm khi phát ra không có thanh, dây thanh không rung hoặc rung đôi chút hoặc dao động không có tần số cơ bản. Khi phát âm các âm vô thanh, chúng ta tạo ra giọng như giọng thở, ví dụ "h", "p" hay "th".

Âm bật: Để phát ra âm bật (còn gọi âm nổ), đầu tiên dây thanh đóng kín, tạo nên một áp suất không khí lớn, sau đó có sự mở khiến không khí được giải phóng một cách đột ngột tạo ra các âm thanh bật.

Cũng cần chú ý, có một số âm khác không đơn giản phân loại được vào một trong ba nhóm âm trên bởi vì chúng là âm tổ hợp của các yếu tố của các âm đó. Chẳng hạn âm thanh khi phát âm chữ "kh", âm được tạo ra do sự mở hẹp của thanh môn và sự co thắt và mở hẹp của vòm miệng.

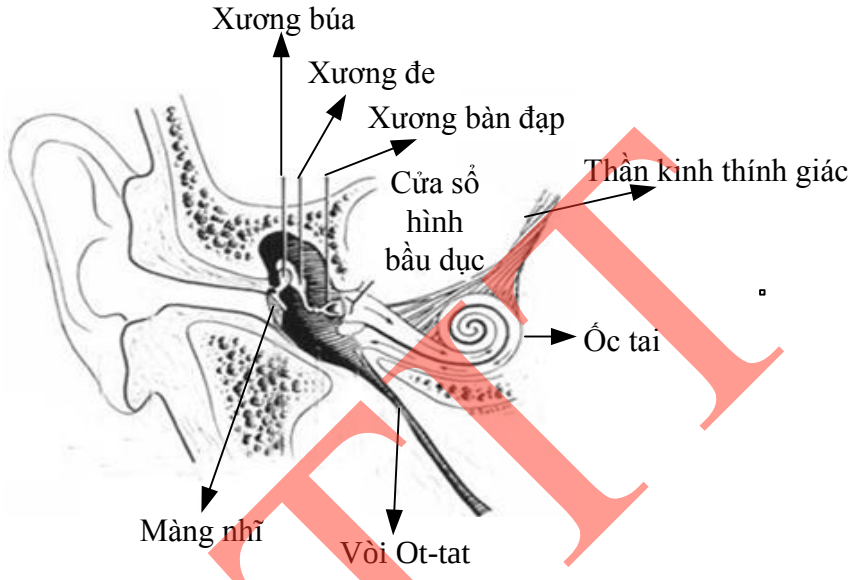
1.3.4 Cấu tạo của hệ thống cảm nhận tiếng nói

Trong hệ thống cảm nhận tiếng nói, tai là một bộ phận quan trọng và là khối đầu tiên trong hệ thống. Không giống như các cơ quan tham gia vào quá trình tạo ra tiếng nói như miệng, mũi, phổi, các cơ quan mà ngoài chức năng tham gia tạo tín hiệu tiếng nói còn thực hiện các chức năng khác như ăn, ngửi, thở. Tai, một cơ quan trong hệ thống thính giác của con người, chỉ sử dụng cho chức năng nghe. Tai người đặc biệt nhạy cảm với những tần số tín hiệu tiếng nói nằm trong vùng nghe (trong khoảng xấp xỉ từ 200 – 5600Hz). Tai người là một máy thu tự nhiên tuyệt hảo, nó có thể phân biệt được những sự khác biệt rất nhỏ về 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 làm nhiệm vụ dẫn hướng những thay đổi áp xuất tiếng nói vào trong màng nhĩ. Nói cách khác, tai ngoài giống như một bộ ăn-ten làm nhiệm vụ thu nhận những dao động âm của tiếng nói truyền đến. Dao động âm, thể hiện ở áp suất hay dao động các phần tử không khí sẽ được biến đổi thành chuyển động cơ học ở tai giữa. Những chuyển động cơ học ở tai giữa được chuyển đổi thành những luồng điện trong nơron thính giác dẫn đến não để thực hiện quá trình phân tích và bóc tách thông tin.

Tai ngoài: là phần phía bên ngoài của tai, bao gồm loa tai (pinna – vành tai) và lỗ tai (meatus - ống tai ngoài). Loa tai hầu như không hoặc rất ít có vai trò đối với 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. Với cấu trúc vành rộng cùng các rãnh xoáy, nó có nhiệm vụ như một ăn-ten thực hiện thu tập năng lượng âm và dẫn hướng vào tai giữa thông qua ống tai ngoài. Ống tai ngoài được nối ở phần cuối hõm của vành tai, nó là một ống ngắn có hình dáng thay đổi có chiều dài khoảng 2.5cm làm đường dẫn cho các tín hiệu âm thu nhận được đến tai giữa. Ống tai ngoài có hai chức năng chính. Chức năng thứ nhất là bảo vệ các cấu trúc phức tạp và dễ bị tổn thương cơ học của tai giữa. Chức năng thứ hai là đóng vai trò như một bộ lọc cơ học cộng hưởng hình ống vốn

ưu tiên cho việc truyền các âm có tần số cao giữa 3000 Hz và 12000Hz. Chức năng này là quan trọng đối với việc tiếp nhận tiếng nói và đặc biệt trợ giúp cho việc tiếp nhận các âm xát, vì đặc điểm của các âm này được tạo ra bởi nguồn kích thích không có chu kỳ và phổ năng lượng của chúng nằm trong trong khu phổ này. Sự cộng hưởng, nói cách khác là khuếch đại, ở ống tai ngoài góp phần vào độ thính chung của tai ở vùng tần số giữa 500Hz và 4000Hz, vốn là một dải tần có chứa nhiều dấu hiệu chính đối với cấu trúc âm vị học.



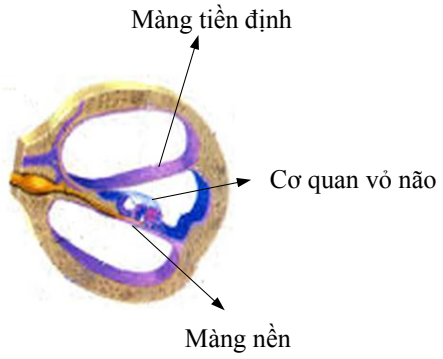
Hình 1.7 Cấu trúc hệ thính giác ngoài

Tai giữa bao gồm một khoang nằm trong cấu trúc hộp sọ có chứa màng nhĩ (eardrum) - màng ở đầu phía trong của ống tai ngoài, một bộ ba khúc xương liên kết với nhau, còn 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 biến đổi những thay đổi áp suất âm (những dao động âm) được thu nhận từ tai ngoài dẫn vào thành những dịch chuyển cơ khí tương ứng. Quá trình biến đổi này bắt đầu ở màng nhĩ, dao động âm làm dịch chuyển màng nhĩ. Sự dịch chuyển này được truyền đến các xương tai, vốn đóng vai trò như một hệ thống đòn bẩy cơ học khéo léo truyền những dịch chuyển này đến cửa hình bầu dục, ô cửa ở giao tiếp giữa tai trong và chất dịch trong lỗ tai.

Với cơ chế hoạt động đòn bẩy của các xương tai, và đặc biệt là vùng diện tích bề mặt của màng nhĩ lớn hơn nhiều so với cửa hình bầu dục, việc truyền hiệu ứng của năng lượng âm học giữa 500Hz và 4000Hz được đảm bảo. Kết quả 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 dao động âm lớn nhờ hoạt động của cơ chế phản xạ âm học. 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), kết quả là những dao động âm quá mạnh bị giảm khi đến cửa hình bầu dục. Tai giữa được nối với họng bằng một ống hẹp gọi là vòi ốc tai (eustachian tube). Việc kết nối này hình thành một đường khí và 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 được 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 tại 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 cụt với một cửa sổ có một màng linh hoạt ở mỗi đầu. Ở bên trong, ốc tai chia thành hai màng, một trong số đó là màng nền (basilar membrane). Đây là màng 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 chất dịch 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. Chính sự dịch chuyển và sự kích thích của các tế bào tóc này biến sự dịch chuyển của màng nền thành các tín hiệu thần kinh. Vì màng nền được dịch chuyển mạnh yếu ở các 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 có khả năng phân biệt. Nhưng cần phải nhấn mạnh rằng sự tái hiện thông tin cuối cùng về tần số cảm nhận từ tín hiệu thần kinh không chỉ đơn thuần phụ thuộc vào vị trí cũng như không chỉ phụ thuộc riêng vào sự dịch chuyển màng nền, mà đây là một quá trình diễn giải phức tạp. Hơn nữa, cho đến nay, hiểu biết của chúng ta về cách thức tần số được lập, mã và giải mã thông qua hệ thống thính giác vẫn chưa hoàn thiện.



Hình 1.8 Mặt cắt ngang của ốc tai

Những nghiên cứu đầu tiên về cảm nhận tiếng nói quan tâm rất ít đến các thuộc tính cảm nhận cơ bản của tai. Những nghiên cứu này đã cố gắng gắn kết các thuộc tính cảm nhận của tín hiệu tiếng 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 tiếng 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.

1.3.5 Đặc điểm cảm nhận tiếng nói của người

Tín hiệu tiếng nói được truyền tải đến tai người nghe thông qua các dao động tạm thời của các phần tử vật chất dọc theo đường truyền tạo ra một áp suất âm đến tai. Tai con người có thể cảm nhận được một dải áp suất âm rộng hơn 7 đơn vị đề-các, bắt đầu từ ngưỡng nghe (còn gọi là TOH – Threshold of hearing) với áp suất âm 10^-5Pa đến ngưỡng nghe gây đau với áp suất âm 10^2Pa. Ngưỡng nghe là ngưỡng áp suất âm thấp nhất mà tai con người có thể cảm nhận được. Ngược lại, ngưỡng nghe gây đau (hay đơn giản gọi là ngưỡng gây đau) là mức ngưỡng áp suất âm mà con người bắt đầu có cảm giác đau ở tai.

Để đơn giản trong đánh giá độ lớn của âm, thay vì sử dụng áp suất âm người ta sử dụng một đại lượng mức áp suất âm (ký hiệu là SPL, Lp – Sound Pressure Level). Mức áp suất âm là một đo lường theo tỷ lệ lô-ga-rít của áp suất âm tương đối so với một quá trị tham chiếu. Nói một cách cụ thể, SPL là một đại lượng đo lường tương đối có đơn vị là dB. Giá trị tham chiếu thường là ngưỡng nghe. SPL được xác định bởi công thức:

$$\mathrm{SPL[dB]} = 10\log\frac{P_{rms}^2}{P_0^2} = 20\log\frac{P_{rms}}{P_0}$$

trong đó, P_{rms} là áp suất âm trung bình quân phương, P_0 là áp suất âm tham chiếu.

Một đại lượng đo lường khác là mức cường độ âm (ký hiệu là SIL, Li – Sound Intensity Level) được xác định bởi công thức:

$$\mathrm{SIL[dB]} = L_I = 10\log_{10}\frac{I}{I_0}$$

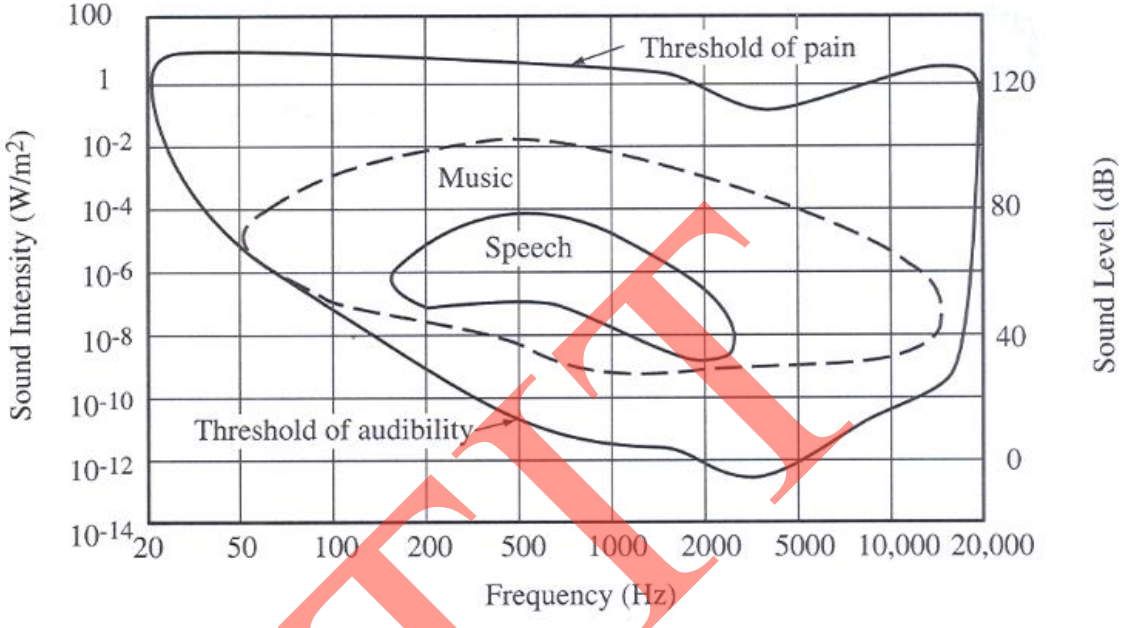
trong đó, I là mức cường độ âm, I0 là mức cường độ âm tham chiếu.

Mức cường độ âm tham chiếu thường là mức cường độ âm ứng với ngưỡng nghe. Giá trị này vào khoảng 10^-12W/m2.

Khi sóng âm lan truyền trong môi trường không khí tự do, giá trị của SPL và SIL bằng nhau. Tuy nhiên, trong không hạn chế điều này không còn đúng do có sự phản xạ âm.

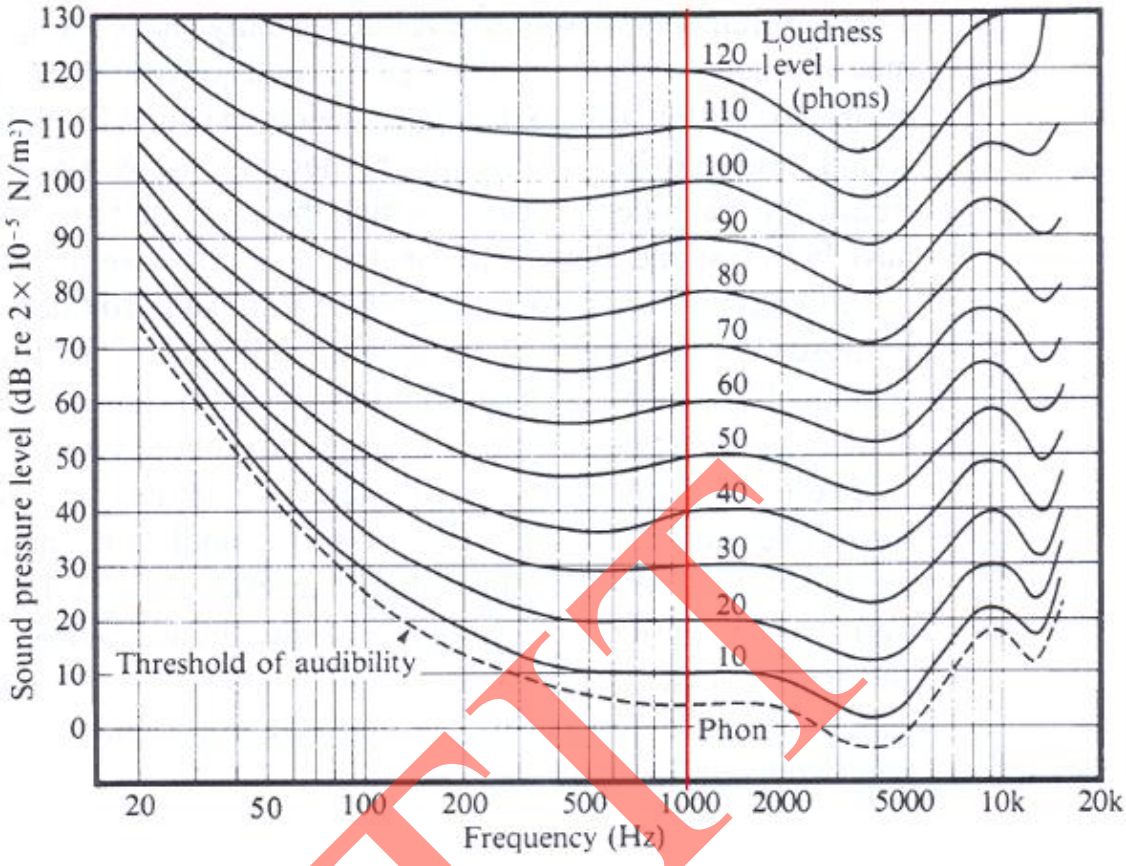
Hầu hết các microphone, một trong nhiều loại thiết bị biến đổi áp suất âm thành tín hiệu điện, làm việc theo nguyên lý nhạy cảm/đáp ứng với kích thích là áp suất âm. Nghĩa là những thiết bị này sẽ đo lường/xác định SPL chứ không phải SIL.

Trong nhiều tài liệu kỹ thuật, người ta thường đồng nhất độ to của âm chính là mức cường độ âm. Mối quan hệ có thể được minh họa trong hình vẽ 1.9.



Hình 1.9 Mối quan hệ giữa cường độ âm, mức cường độ âm và tần số trong vùng nghe

Sự cảm nhận âm thanh của một người bình thường với một mức độ to âm thanh xác định (chính là mức cường độ âm, hay SIL) không độc lập với tần số. Tai người rất kém nhạy với các âm có tần số rất nhỏ (<20Hz) hoặc rất lớn (>20kHz). Nói cách khác, sự cảm nhận âm thanh của con người không phải như trong trong toàn dải tần của vùng nghe. Do đó, rõ ràng mức độ to của âm thanh phụ thuộc vào tần số của âm. Bằng các thí nghiệm, ở cùng một mức cảm nhận về cùng độ to của âm thanh của tai người, sự thay đổi SPL theo tần số được minh họa trong hình 1.10.



Hình 1.10 Mức áp suất âm cần thiết ở các vùng tần số khác nhau để tai người cảm nhận cùng độ to của âm

Sự cảm nhận về độ to của âm phụ thuộc vào tần số có thể xấp xỉ bằng công thức hàm ngưỡng nghe tuyệt đối như sau:

$$T_q(f) \approx 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$

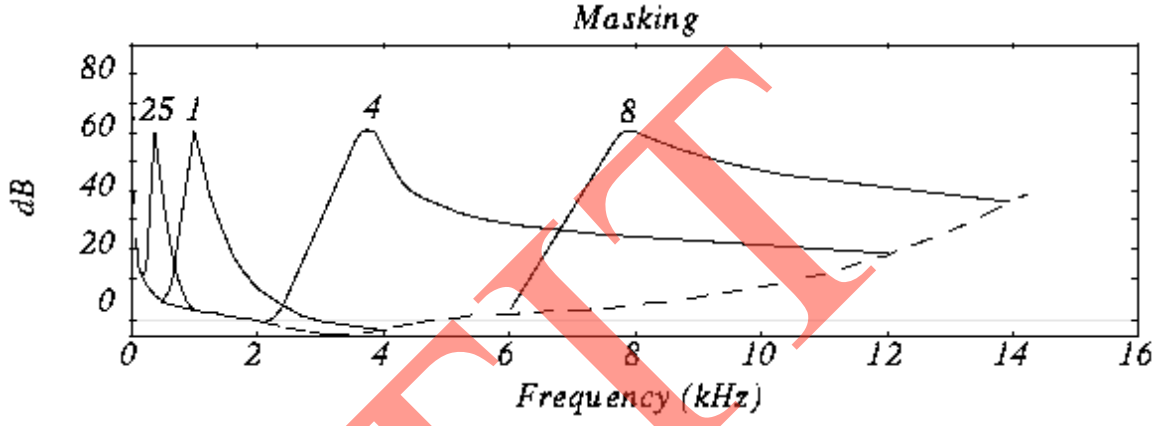
Người ta định nghĩa ngưỡng nghe tuyệt đối là mức năng lượng tối đa của một tín hiệu đơn âm cơ bản (pure tone) mà người nghe không thể cảm nhận được trong môi trường tự do.

Trong quá trình cảm nhận âm thanh của tai người, có một hiện tượng rất quan trọng khác được phát hiện đó là hiện tượng che lấp âm thanh (gọi tắt là hiện tượng che lấp). Hiện tượng che lấp có thể quan sát trong miền tần số, còn gọi là che lấp tần số, hoặc quan sát trong miền thời gian, còn gọi là hiện tượng che lấp thời gian.

Hiện tượng che lấp thời gian xảy ra khi chúng ta nghe một âm rất lớn, sau đó âm đó tắt đột ngột nhưng tai chúng ta vẫn cảm nhận về âm này trong một khoảng thời gian sau đó. Giả sử ngay sau khi âm thanh lớn tắt đột ngột, chúng ta phát một âm thanh khác

nhưng với mức thấp hơn. Khi đó tai chúng ta sẽ không thể cảm nhận được âm thanh khác đó. Người ta nói âm thanh tiếp sau đó đã bị che lấp.

Hiện tượng che lấp tần số là hiện tượng một âm thanh bị làm mờ hoặc mất hẳn không thể cảm nhận được khi xuất hiện một âm thanh có tần số khác. Hay nói một cách khác, sự xuất hiện một âm thanh sẽ làm tăng mức ngưỡng nghe của một âm thanh ở tần số khác. Các âm tần số thấp thường che lấp các âm tần số cao hơn, trong đó hiệu ứng che lấp lớn nhất tại vùng gần các thành phần hài của âm che lấp. Các dải tín hiệu âm băng tần rộng che lấp các dải tín hiệu âm băng tần hẹp hơn. Hình 1.11 minh họa hiện tượng che lấp ở một số tần số xác định.



Hình 1.11 Hiện tượng che lấp ở các tần số khác nhau

Một điểm thú vị từ quan sát của hình 1.11 ở trên là độ rộng vùng tần số che lấp ở các tần số che lấp khác nhau không đồng nhất. Độ rộng vùng tần số che lấp gần như không đổi cỡ khoảng 100Hz với các tần số che lấp <500Hz, và độ rộng vùng này càng tăng rất nhanh theo hàm lô-ga-rít khi tần số che lấp tăng. Độ rộng vùng tần số che lấp được gọi là băng tần cơ bản (critical band).

Với sự cảm nhận không tuyến tính vừa đề cập ở trên, Zwicker sử dụng một đơn vị đo lường mới cho tần số âm: thang tần số Bark. Đơn vị này được đặt tên theo Barkhausen, một nhà vật lý người Đức. Một cách đơn giản, 1 Bark chính là độ rộng của một băng tần cơ bản. Với định nghĩa này, toàn dải nghe của người được chia thành 24 thang tương ứng với 24 băng tần cơ bản. Mối quan hệ giữa thang tần Hz và Bark được cho bởi công thức:

$$\text{Bark} = 13a\tan(0.00076f) + 3.5a\tan((f/7500)^2)$$

$$W[\text{Hz}] = 52548/(b^2 - 52.56b + 690.39)$$

Ngoài thang tần Bark, trong phân tích âm thanh tiếng nói người ta còn hay sử dụng thang tần số Mel. Khác với thang tần Bark, thang tần Mel tuyến tính trong một khoảng nhỏ hơn 1kHz, và thay đổi theo quy luật lô-ga-rít ở vùng lớn hơn 1kHz. Thang Mel được xây dựng từ thí nghiệm với các tân đơn (pure sine tone) trong đó người cảm nhận được

yêu cầu chia vùng tần số thành 4 vùng cảm nhận tương đồng nhau. Thang tần Mel được cho là mô phỏng gần với đặc tính độ nhạy của tai hơn so với thang tần Bark. Thang tần Mel có mối liên hệ với thang tần Hz theo các công thức:

$$m[Mel] = 2595 \log_{10}(1 + \frac{f}{700})$$

$$f[Hz] = 700(10^{m/2595} - 1)$$

Trong một số kỹ thuật xử lý tiếng nói hiện đại, chẳng hạn như phân tích cepstral, phân tích đặc trưng động (dynamic feature), …, thường sử dụng thang tần này.

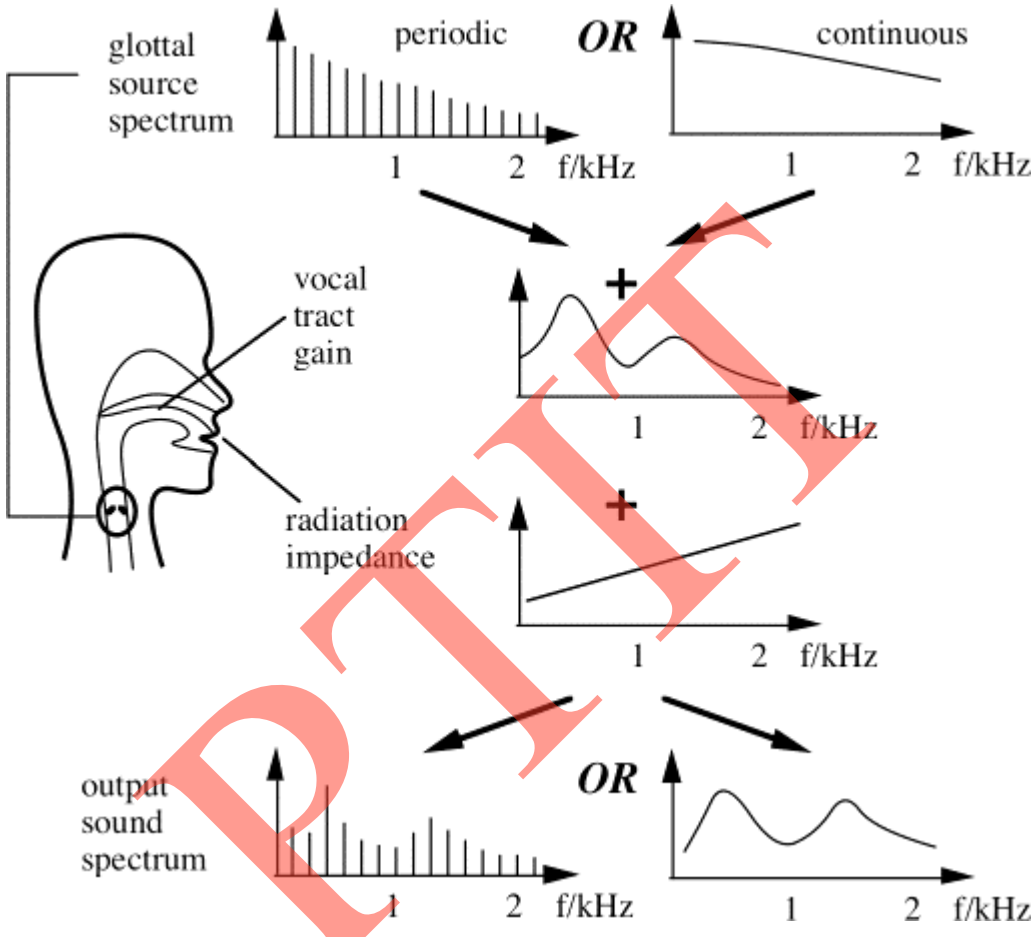
Cũng cần nhấn mạnh, có một sự khác biệt cơ bản giữa các thuộc tính cảm nhận một tín hiệu âm thanh, đặc biệt là tín hiệu tiếng nói, và các thuộc tính vật lý có thể đo lường của âm. Sự tương ứng giữa các thuộc tính và các đại lượng vật lý được cho trong bảng 1.1. Mỗi thuộc tính dường như có mối liên hệ mật thiết với một tính chất vật lý, tuy nhiên mối quan hệ này thường rất phức tạp. Điều này dễ hiểu vì các tính chất vật lý của âm thành có thể ảnh hưởng đến việc cảm nhận âm thanh theo một cách thức rất phức tạp. Lấy ví dụ, chúng ta thưởng cho rằng cường độ âm càng lớn thì âm thanh cảm nhận càng to. Tuy nhiên như minh họa trong hình 1.10 ở trên, điều này không đơn giản như vậy. Rõ ràng là có một sự khác biệt rõ ràng giữa cảm nhận âm to và đại lượng vật lý mức áp suất âm/mức cường độ âm. Hoặc lấy một ví dụ khác, đó là cảm nhận về cao độ của âm thanh. Rõ ràng cao độ âm thanh mà ta có thể cảm nhận được có một mối quan hệ mật thiết với tần số cơ bản. Dường như tần số cơ bản càng cao thì âm mà chúng ta cảm nhận được càng cao. Tuy nhiên, sự phân biệt giữa hai cao độ sẽ phụ thuộc vào tần số của cao độ có tần số thấp hơn. Cao độ mà chúng ta cảm nhận được sẽ thay đổi khi cường độ âm tăng lên trong khi tần số giữ cố định. Hoặc một ví dụ khác nữa là hiện tượng che lấp đã đề cập ở trên.

Bảng 1.1: Sự liên quan giữa các đại lượng vật lý và thuộc tính cảm nhận

Đại lượng vật lý	Chất lượng cảm nhận
Mức cường độ âm	Độ to (loudness)
Tần số cơ bản	Cao độ (pitch)
Hình dạng phổ	Âm sắc (timbre)
Độ lệch thời gian	Cảm giác về thời gian (timing)
Sự lệch pha	Vị trí âm (location)

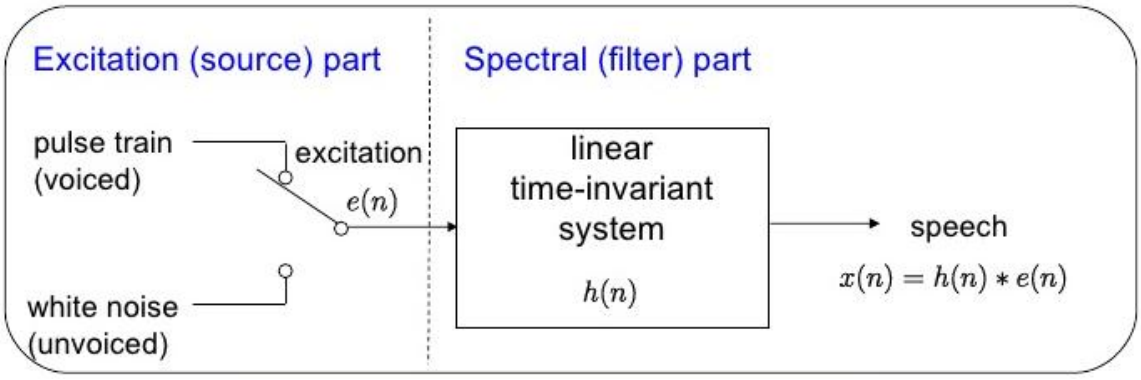
1.4. MÔ HÌNH HÓA HỆ THỐNG CƠ QUAN PHÁT ÂM

Trong phần trên chúng ta đã tìm hiểu về cơ chế hoạt động của bộ máy phát âm. Hoạt động này gồm hai quá trình: nguồn tạo dao động âm và cấu trúc phổ định hình hay còn gọi là bộ lọc. Cơ chế hoạt động có thể tóm lược như minh họa hình 1.12.



Hình 1.12 Minh họa tóm lược cơ chế phát âm

Để đơn giản trong quá trình phân tích, người ra thực hiện mô hình hóa quá trình làm việc của bộ máy phát âm như sơ đồ hình 1.13.



Hình 1.13 Mô hình nguồn-bộ lọc mô phỏng bộ máy phát âm

Trong mô hình này, nguồn tương ứng với dao động dây thanh được mô tả tương ứng với hai trường hợp: (1) với các âm hữu thanh, dao động dây thanh có tần số cơ bản xác định, khi đó nó được mô tả bởi một dãy xung tuần hoàn; (2) với các âm vô thanh, dao động dây thanh không xác lập tần số, nó được mô tả tương ứng như là nhiễu trắng.

Tín hiệu dao động dây thanh sẽ đưcc lọc bởi bộ lọc tuyến âm để tạo ra tín hiệu tiếng nói mong muốn. Bản chất bộ lọc tuyến âm là một bộ lọc cơ học (bộ lọc âm), ta có thể mô tả bởi một bộ lọc có đáp ứng xung tương ứng h(n).

Việc xác định hàm đáp ứng xung của bộ lọc tuyến âm tương đối phức tạp. Mặc dù đã có rất nhiều nghiên cứu, cùng với đó là có khá nhiều phương pháp để xấp xỉ bộ lọc này, nhưng cho đến nay vẫn chưa có một mô hình hoàn toàn đúng nào được đề ra. Bởi đặc tuyến của bộ lọc phụ thuộc không những sự co thắt của tuyến âm mà còn phụ thuộc rất lớn vào hiệu quả phát xạ âm tại môi hoặc/và mũi và những tương tác giữa các bộ phận này.

Thông thường, để có thể nhấn được các đỉnh cộng hưởng của bộ lọc tuyến âm, người ta thường xấp xỉ nó bằng bộ lọc toàn điểm cực (all-pole). Bằng cách tổng hợp mạch lọc IIR bậc hai, chúng ta có thể mô tả một cách đầy đủ một tần số formant.

Khi có kể đến khoang mũi, hoạt động của khoang miệng trở nên phức tạp cũng như sự tương tác giữa khoang miệng và khoang mũi rất khó quan sát. Để đơn giản trong nghiên cứu, người ta coi khoang mũi là khoang tĩnh, và bỏ qua sự tương tác. Khi đó, khoang mũi được xem như một bộ lọc mắc song song với khoang miệng. Quá trình thực nghiệm xác định hàm truyền đạt tổng hợp thường được tiến hành bằng cách xấp xỉ hàm truyền đạt của từng bộ lọc.

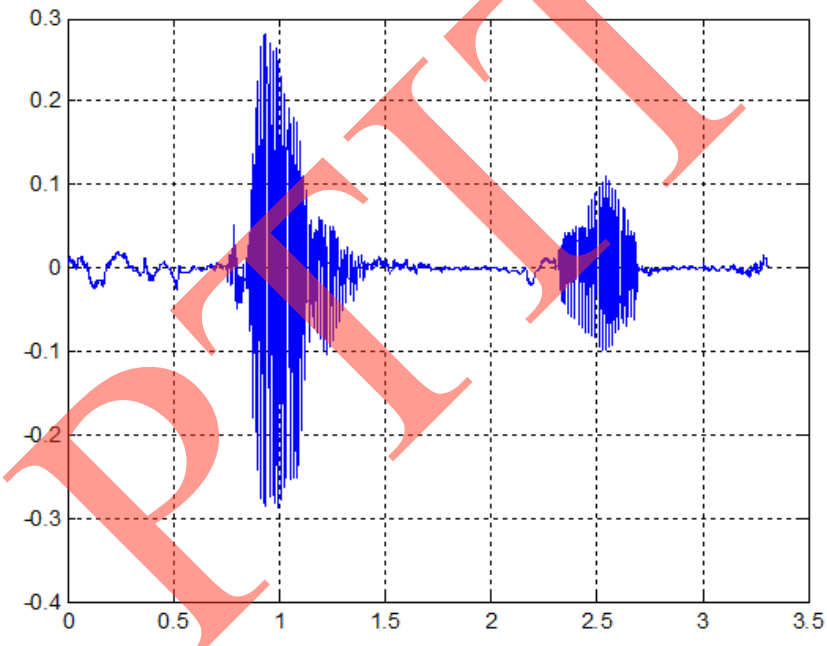
1.5. BIỂU DIỄN TÍN HIỆU TIẾNG NÓI

Có 3 phương pháp cơ bản thường được dùng để biểu diễn tín hiệu tiếng nói: Biểu diễn dạng sóng tín hiệu trong miền thời gian; Biểu diễn phổ trong miền tần số; Biểu diễn spectrogram.

1.5.1 Biểu diễn dạng sóng tín hiệu trong miền thời gian

Tín hiệu tiếng nói cũng giống như các tín hiệu thông thường, có thể coi là là một hàm của thời gian s(t) (nếu xem xét tín hiệu tiếng nói liên tục, tiếng nói tự nhiên) hoặc s(n) (nếu xem xét tin hiệu tiếng nói số, tiếng nói trong các hệ thống xử lý tín hiệu số). Trong khuôn khổ bài giảng này, chúng ta sẽ chỉ xem xét tín hiệu tiếng nói số s(n). s(n) là kết quả lấy mẫu và lượng tử hóa của s(t).

Khi thực hiện biểu diễn tín hiệu tiếng nói s(n) theo thời gian hoặc chỉ số thời gian, người ta gọi đó là biểu diễn dạng sóng tín hiệu trong miền thời gian, hay đơn giản là biểu diễn dạng sóng. Đây là phương thức biểu diễn trực quan và đơn giản nhất. Biểu diễn này có thể cho biết được sự thay đổi về biên độ tín hiệu, sự dao động nhanh hay chậm của tín hiệu theo thời gian. Hình 1.14 minh họa một biểu diễn theo thời gian của cụm từ "không một".

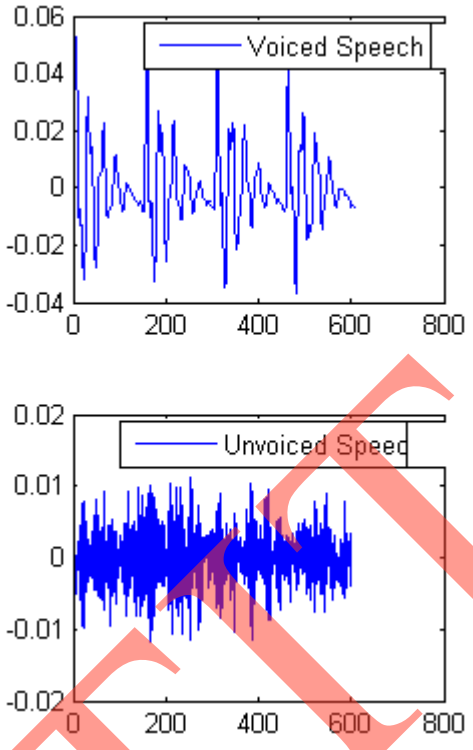


Hình 1.14 Biểu đồ dạng sóng của cụm từ "không một"

Từ biểu diễn trên, chúng ta có thể thấy có sự phân biệt tương đối giữa các từ. Ở trước, sau và giữa các từ có một khoảng tín hiệu ở đó biên độ rất nhỏ gần như bằng không, chúng ta gọi đó là các khoảng lặng (silent).

Khi quan sát đơn lẻ dạng sóng tín hiệu tiếng nói là phát âm của một từ, chẳng hạn cụm từ "không một" như minh họa trong hình 1.14, chúng ta thấy có một đoạn tín hiệu ngay sau khoảng lặng, phần bắt đầu của âm có biên độ khác không tuy nhiên rất nhỏ (chỉ cỡ 1/3 lần) so với phần chính của âm. Phần này tương ứng với sự phát âm của âm vô thanh. Nói một cách khác, từ biểu đồ dạng sóng chúng ta có thể phân biệt được âm vô thanh và hữu thanh. Phần âm vô thanh tương ứng với dạng tín hiệu có biên độ thấp,

không có dạng tuần hoàn mà có dạng ngẫu nhiên. Hình 1.15 minh họa sự khác biệt dạng sóng của âm vô thanh và hữu thanh.



Hình 1.15 Sự khác biệt dạng sóng tín hiệu âm hữu thanh và vô thanh

Cũng cần lưu ý là việc phân biệt giữa khoảng lặng và âm vô thanh chỉ mang tính tương đối và chỉ có thể cho kết quả chấp nhận được khi nhiễu đủ nhỏ. Điều này là bởi vì bản chất của nhiễu cũng có tính ngẫu nhiên, khi nhiễu có biên độ lớn (nhiễu lớn) có thể khiến ta quan sát nhầm giống như phần phát âm của âm vô thanh.

Chúng ta thường cho rằng, giọng điệu tiếng nói của một người gần như không thay đổi: một người nói hay hai người cùng nói từ "một" thì nó luôn có nghĩa là "một" và dạng sóng tín hiệu của phát âm tương ứng phải giống hệt nhau. Tuy nhiên, khi quan sát dạng sóng của những lần thu âm khác nhau thì điều này không đúng. Ta có thể thấy, ngay cùng với một từ và một người phát âm, nhưng dạng sóng ở hai thời điểm khác nhau có sự khác nhau nhất định. Quan sát tương tự cũng thấy khi hai người phát âm cùng một từ, dạng sóng cũng có sự khác nhau tương đối.

Ngoài ra, dạng sóng tín hiệu tiếng nói cũng có sự khác biệt đáng kể khi sử dụng các thiết bị thu âm, mã hóa có chất lượng khác nhau.

Chính từ những khác nhau nhất định của dạng sóng này cho ta thấy ở chương 5 việc nhận dạng bằng cách sử dụng trực tiếp dạng sóng, còn gọi là sử dụng dữ liệu thô, là không khả thi.

Dữ liệu dạng sóng tín hiệu tiếng nói số thường được lưu trữ trong máy tính dưới nhiều định dạng, phổ biến nhất là *.wav. Tín hiệu này là kết quả của việc lấy mẫu tín hiệu tiếng nói với tần số lấy mẫu phổ biến là 8000Hz, 10000Hz, 11025Hz, 16000Hz, 22050Hz, 32000Hz, 44100Hz,…, với độ phân giải bít phổ biến là 8bit, 16bit, 24bit, … và có thể là một kênh (mono) hoặc hai kênh (stereo)

1.5.2 Biểu diễn phổ tín hiệu tiếng nói

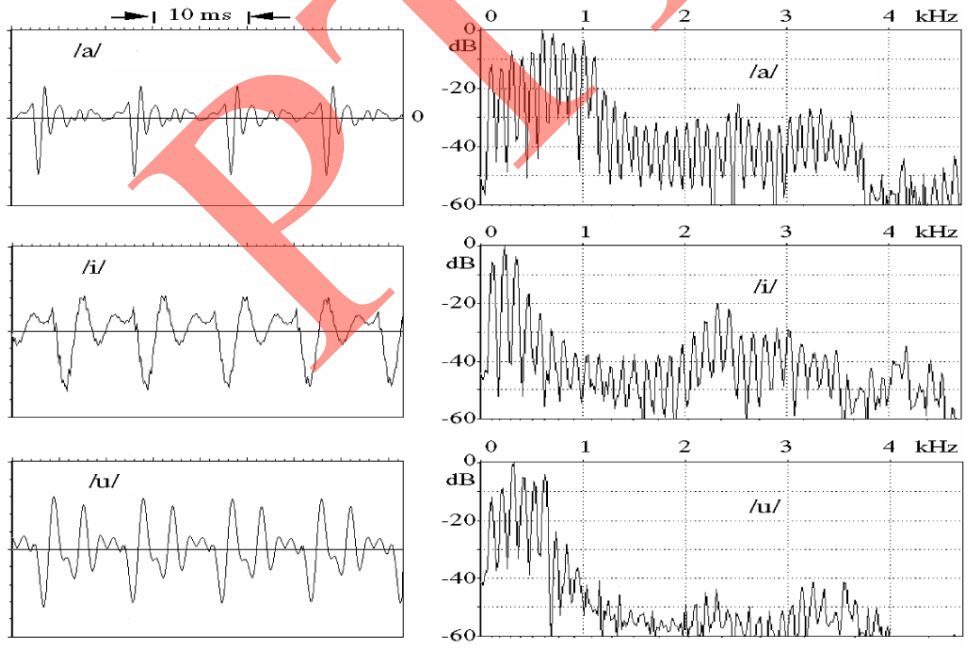
Như chúng ta đã biết trong môn học Xử lý tín hiệu số, việc biểu diễn phổ, hay nói cách khác là biểu diễn tín hiệu tiếng nói trong miền tần số có thể cho phép việc phân tích và tìm hiểu tín hiệu tiếng nói được thuận tiện và dễ dàng hơn.

Với tín hiệu tiếng nói số s(n), thực hiện biến đổi Fourier, ta được:

$$S(j\omega) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n}$$

Khi đó phổ biên độ và phổ pha của tín hiệu tiếng nói tương ứng là biểu diễn $|S(j\omega)|$, và $\arg\{S(j\omega)\}$. Trong phân tích tín hiệu tiếng nói, thông tin tiếng nói được chứa chủ yếu trong phổ biên độ, do đó người ta rất ít quan tâm đến phổ pha.

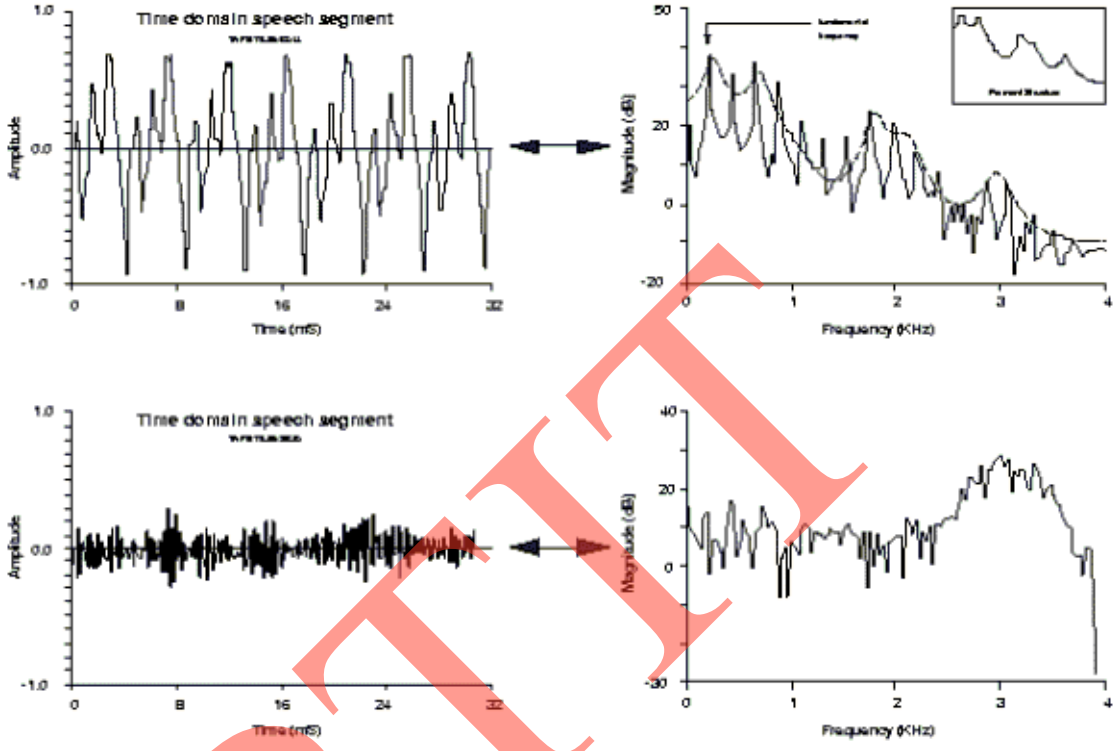
Biểu diễn phổ biên độ của một phân đoạn tiếng nói ứng với phát âm của âm hữu thanh được minh họa trong hình 1.16.



Hình 1.16 Minh họa phổ tín hiệu tiếng nói

Từ quan sát biểu diễn phổ biên độ, ta có thể thấy phổ biên độ có thể tách thành hai thành phần: đường bao phổ và những dao động phổ nhỏ hay còn gọi là phổ nhỏ. Đường

bao phổ tương ứng là dạng phổ của một tín hiệu biến đổi chậm (tần số thấp). Nó tương ứng là hàm truyền đạt của bộ lọc tuyến âm. Phần phổ nhỏ tương ứng là dạng phổ của một tín hiệu biến đổi nhanh (tần số cao). Nó tương ứng là phổ của tín hiệu tạo bởi dao động của dây thanh.



Hình 1.17 Minh họa sự khác biệt phổ giữa âm vô thanh và hữu thanh

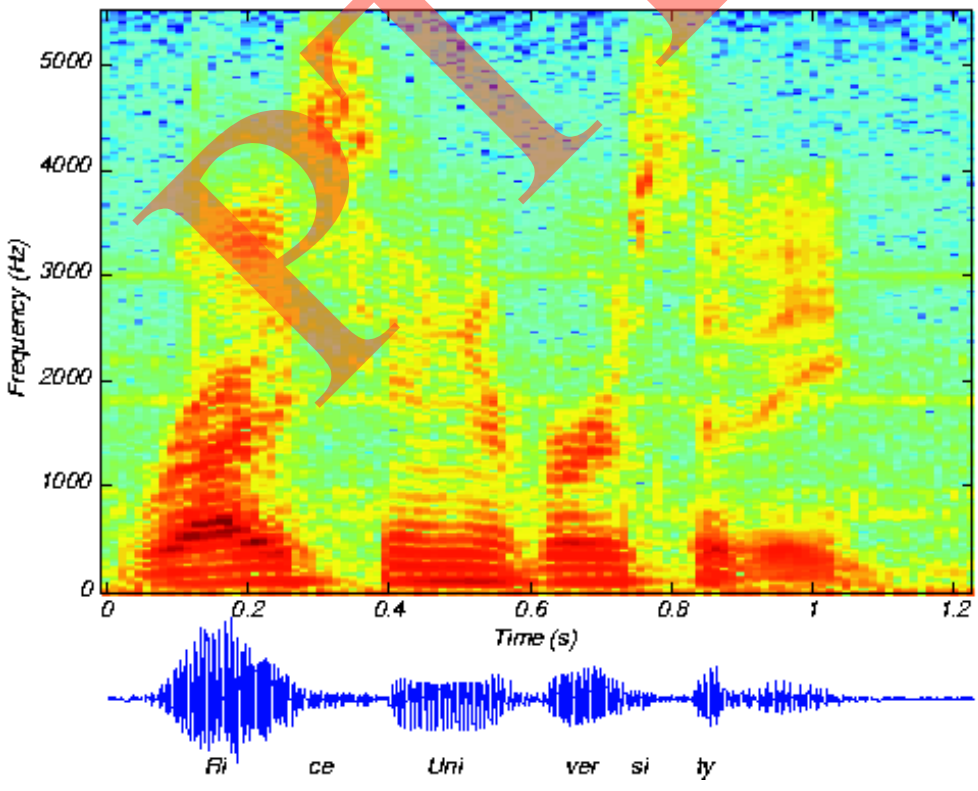
Phổ biên độ của phân đoạn âm hữu thanh và vô thanh được minh họa trong hình 1.17. Từ biểu diễn phổ biên độ, chúng ta thấy có thể dựa trên phổ biên độ để phân biệt một cách tương đối giữa âm vô thanh và hữu thanh. Phổ biên độ của phân đoạn tín hiệu ứng với âm hữu thanh có xuất hiện các cực trị của đường bao phổ. Trong các cực trị này những đỉnh cực đại được gọi là các đỉnh formant, tương ứng là các đỉnh cộng hưởng của bộ lọc tuyến âm, những tần số được tăng cường; các rãnh cực tiểu xen kẽ giữa các đỉnh cực đại được gọi là các phản formant (anti-formant), những tần số bị suy giảm. Ngoài ra năng lượng phổ của phân đoạn tín hiệu này cũng có sự tập trung chủ yếu ở phần tần thấp. Ngược lại, phổ biên độ của phân đoạn tín hiệu âm vô thanh không xuất hiện các cực trị phân biệt trong đường bao phổ. Nói cách khác không tồn tại các formant trong biểu diễn phổ của âm vô thanh. Ngoài ra, năng lượng phổ của âm vô thanh phân bố đều trên toàn dải tần số và có xu thế tập trung ở vùng tần số cao.

Cũng dễ dàng quan sát thấy rằng, mặc dù dải tần số tín hiệu tiếng nói rất rộng (20-20000Hz), nhưng năng lượng phổ của tín hiệu tiếng nói chỉ tập trung trong một khoảng từ 300-3400Hz.

1.5.3 Biểu diễn spectrogram

Như đã đề cập ở trên, bản chất của tiếng nói là bán tĩnh (quasi-static), nghĩa là các tham số thay đổi theo thời gian, chỉ có thể coi là các tham số không thay đổi nếu xem xét tín hiệu trong một khoảng thời gian đủ nhỏ. Do đó, việc chỉ phân tích trong miền thời gian, hoặc chỉ trong miền tần số là không đủ để tìm hiểu về các đặc trưng của tín hiệu. Spectrogram hay còn gọi là sonogram là một phân tích thời gian-tần số của tín hiệu, hay phân tích hai chiều. Với phân tích này, các đặc trưng phổ thay đổi theo thời gian có thể dễ dàng quan sát được.

Để thực hiện biểu diễn này, người ta thực hiện chia tín hiệu thành các phân đoạn ngắn hạn bằng các hàm cửa sổ. Độ rộng của cửa sổ phân tích thường được chọn tương ứng với độ rộng của 10-30ms tín hiệu. Các phân đoạn không tách rời nhau mà thường có sự bao trùm nhau tương ứng với khoảng 10ms tín hiệu. Sau đó, mỗi phân đoạn tín hiệu được thực hiện biến đổi Fourier để tìm phổ biên độ. Tại mỗi phân đoạn tương ứng với trục thời gian theo phương ngang, thực hiện biểu diễn phổ theo phương thẳng đứng với biểu diễn độ đậm nhạt màu tỷ lệ thuận với năng lượng phổ biên độ.



Hình 1.18 Minh họa spectrogram của phân đoạn âm thanh

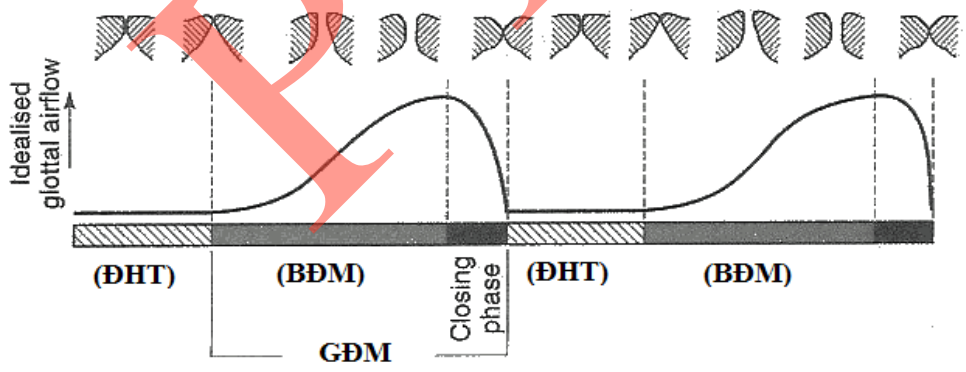
Từ biểu diễn spectrogram, chúng ta có thể thấy đây là một công cụ rất thuận tiện để quan sát và phân tích tín hiệu. Chẳng hạn, chúng ta có thể phân biệt một cách tương đối âm vô thanh với âm hữu thanh dựa trên biểu diễn spectrogram. Ở những phân đoạn tín hiệu ứng với âm hữu thanh thì spectrogram tương ứng là những dải đậm màu có những vằn (còn gọi là những cực trị) tương ứng với tính tuần hoàn của tín hiệu. Những vạch này cho thấy có sự phân bố không đồng đều của tần số tín hiệu như đã quan sát trong biểu diễn phổ biên độ. Còn ở những phân đoạn tín hiệu tương ứng với âm vô thanh thì spectrogram tương ứng là những dải đặc nhạt màu. Dải đặc này tương ứng với sự phân bố tần số không có các cực trị và trải đều trên toàn trục trùng với quan sát trong biểu diễn phổ biên độ.

1.6. CÁC THAM SỐ CƠ BẢN CỦA TÍN HIỆU TIẾNG NÓI

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), ...

1.6.1 Tần số cơ bản

Với phần tín hiệu tiếng nói bán tuần hoàn, giá trị trung bình chu kỳ của tín hiệu được gọi là chu kỳ cơ bản hay chu kỳ pitch (T0). Chu kỳ cho bản tương ứng với chu kỳ đóng mở của dây thanh.



Hình 1.19 Minh họa đóng mở thanh môn và chu kỳ cơ bản

Tần số cơ bản F0 được định nghĩa là nghịch đảo của chu kỳ cơ bản: F0=1/T0. Tần số cơ bản có sự khác nhau giữa các giới và độ tuổi và người nói. Các số liệu thống kê cho thấy tần số cơ bản của nam giới vào khoảng 85-180Hz, trong khi giá trị này là khoảng 165-255Hz. Tần số cơ bản của tín hiệu tiếng nói trẻ em lớn cỡ gấp hai lần tần số cơ bản tiếng nói của người lớn, cỡ 350-850Hz. Giá trị trung bình tần số cơ bản thay đổi theo độ

tuổi. Với nam giới, tần số cơ bản có sự giảm mạnh trong thời từ tuổi kỳ dậy thì đến khoảng tầm 35 tuổi. Tuy nhiên, sau tuổi 55, tần số cơ bản của tiếng nói của nam giới lại bắt đầu có sự tăng trở lại. Với nữ giới, tần số cơ bản giữ ổn định cho đến tuổi trung niên, và sau đó bắt đầu có sự suy giảm.

Tần số cơ bản (chu kỳ cơ bản) là một trong các đặc trưng cơ bản và được sử dụng nhiều trong các phân tích cũng như xây dựng các ứng dụng tiếng nói.

1.6.2 Tần số formant

Như đã đề cập trong phần biểu diễn tín hiệu tiếng nói trong miền tần số, đường bao phổ tần số có những đỉnh cực đại gọi là các tần số formant. Tại các tần số này tín hiệu dao động dây thanh được tăng cường.

Các tần số formant được biết đến như những đặc trưng quan trọng trong việc xác định nội dung về khía cạnh âm học của các âm. Và do đó tần số formant thường được sử dụng vào nhận dạng tiếng nói.

Việc xác định tần số formant thường được dựa vào phân tích phổ của tín hiệu tiếng nói. Đỉnh cộng hưởng đầu tiên, ứng với đỉnh cộng hưởng có tần số thấp nhất được ký hiệu là F1, tiếp đến là tần số formant F2, F3, … Trong các phát âm của nguyên âm, người ta thấy rằng luôn có bốn hoặc nhiều hơn bốn tần số formant phân biệt. Nhiều nghiên cứu chỉ ra rằng, chỉ cần hai tần số formant đầu tiên là đủ để phân biệt các nguyên âm. Hai formant đầu tiên này cũng quyết định chất lượng của các nguyên âm theo khía cạnh tính đóng/mở và vị trí phát âm trước/sau trong vòng miệng. Tuy nhiên, những phân biệt này chỉ mang tính tương đối.

1.7. MỘT SỐ ĐẶC ĐIỂM NGỮ ÂM

Trong phần này, chúng ta sẽ tìm hiểu một số khái niệm về mặt ngữ âm của ngôn ngữ. Những khái niệm cơ bản này sẽ được sử dụng trong các chương 4 và 5.

1.7.1 Một số định nghĩa cơ bản về đơn vị ngữ âm

Âm vị (phoneme): chỉ một đơn vị trừu tượng phân biệt về mặt cảm nhận nhỏ nhất của âm thanh tiếng nói trong một ngôn ngữ cho phép phân biệt một từ này với một từ khác. Nói cách khác, nó là một đơn vị nhỏ nhất của tiếng nói được sử dụng để tạo ra sự khác biệt của một từ với một từ khác. Âm vị không phải là các phân đoạn âm về mặt vật lý thông thường mà chúng được phân loại dựa trên nhận thức. Chẳng hạn như phần đơn vị âm thanh ứng với phát âm các âm b, p, t, đ trong phát âm của các từ bố, phố, tố, đố

Âm tố (phone): ám chỉ một thực hiện vật lý về mặt âm học của một âm vị, tức là là một phân đoạn vật lý cụ thể biểu diễn âm vị. Ví dụ, trong tiếng Anh, âm vị /t/ có hai thực hiện về mặt âm học (âm tố) rất khác nhau trong các phát âm của các từ sat và meter.

Cần chú ý rằng tập các âm vị sẽ có các thực hiện về mặt âm học (âm tố) khác nhau tùy theo người nói, nhưng chúng luôn có một chức năng mang tính hệ thống cho phép phân biệt nghĩa của các từ.

Bán âm tố kép (diphone): là cụm kết hợp của một nửa cuối của âm tố phía trước và một nửa đầu của âm tố phía sau. Bán âm tố kép cho phép giữ được sự thay đổi về mặt phát âm giữa các âm tố, do đó có khả năng làm tăng độ chính xác trong việc tổng hợp tiếng nói

Âm tiết (syllable): là một đơn vị phát âm gồm có một âm của nguyên âm đứng một mình hoặc kết hợp với các phát âm của các phụ âm để tạo thành một từ hoặc một phần của một từ có nghĩa. Nói cách khác, âm tiết là một phần phát âm của một từ mà có thể phân tách một cách tự nhiên. Ví dụ, từ *doctor* trong tiếng Anh gồm hai âm tiết.

Từ (word): là một đơn vị ngôn ngữ nói hoặc viết mang ý nghĩa xác định. Ví dụ *work* trong tiếng Anh là một từ.

Câu (sentence): là một tập hợp các từ với một tổ chức hoàn chỉnh được cấu thành bởi một cấu trúc chủ ngữ - vị ngữ và mang một ý hoàn chỉnh mang tính trần thuật, hoặc mệnh lệnh, hoặc câu hỏi, …

1.7.2 Đặc điểm ngữ âm của tiếng Việt

Tiếng Việt là một ngôn ngữ thuộc nhóm ngôn ngữ Nam Á (còn gọi là Mon-Khmer). Tiếng Việt được xem là một ngôn ngữ đơn lập (mono-syllabic language) tiêu biểu mà đặc điểm cơ bản của nó là mỗi đơn vị từ được phát âm bởi một âm tiết. Nói cách khác, mỗi âm tiết trong tiếng Việt đều có khả năng trở thành một từ. Do đó, âm tiết giữ một vai trò cơ bản trong hệ thống các đơn vị ngôn ngữ. Theo thống kê, tiếng Việt gồm có 2500 âm tiết. So với số lượng âm tiết, số lượng từ thì lớn hơn rất nhiều bởi trong tiếng Việt cũng tồn tại nhiều từ ghép. Một đặc điểm nữa là các từ tiếng Việt không có sự biến hình, một âm tiết cũng đồng thời là một hình vị và ý nghĩa ngữ pháp được thể hiện chủ yếu bằng trật tự của từ.

Âm tiết tiếng Việt có cấu trúc đơn giản, luôn gắn liền với thanh điệu. Tiếng Việt gồm có sáu thanh điệu: Thanh ngang, thanh bằng, thanh sắc, thanh hỏi, thanh ngã, thanh nặng. Ngữ nghĩa của một từ thay đổi khi thanh điệu thay đổi.

Tiếng Việt là một ngôn ngữ đánh vần được, các từ được cấu thành từ các cụm phụ âm – nguyên âm – (phụ âm). Nguyên âm trong tiếng Việt thường được chia thành hai nhóm: nguyên âm đơn, nguyên âm kép. Phụ âm thường được phân loại theo cấu hình của các bộ phân trong hệ thống phát âm và phương thức phát âm: phụ âm bật (còn gọi là phụ âm nổ), phụ âm mũi, phụ âm xát, phụ âm bật rung, phụ âm xát tắc.

Phương thức cấu âm				**Các phụ âm**
Ồn	Tắc	Vô thanh	Bật hơi	th
			Không bật hơi	p, t, h, c, k
		Hữu thanh		b,d
	Xát	Vô thanh		s, x, tr, ch
		Hữu thanh		v, r, g, ph
Vang	Mũi			m, n, nh, ng
	Không mũi			l

1.8. CÂU HỎI VÀ BÀI TẬP CUỐI CHƯƠNG

1. Các bộ phận chính và vai trò của chúng trong bộ máy phát âm?
2. Môi, khoang mũi có vai trò gì trong quá trình phát âm?
3. Các bộ phận chính và vai trò của chúng trong cơ quan cảm nhận tiếng nói?
4. Đặc điểm nghe của tai người? Mối quan hệ giữa các đặc tính cảm nhận âm và các đại lượng vật lý của âm?
5. Mô hình nguồn-bộ lọc mô phỏng bộ máy phát âm?
6. Hiện tượng che lấp là gì? Hiện tượng này có vai trò gì?
7. Các phương pháp biểu diễn cơ bản tín hiệu tiếng nói?
8. Một số khái niệm ngữ âm cơ bản? Đặc điểm ngữ âm tiếng Việt?
9. Các tham số cơ bản của tín hiệu tiếng nói?
10. Phân biệt âm vô thanh và hữu thanh?
11. (Matlab) Sử dụng Matlab (hoặc bộ công cụ thích hợp khác, chẳng hạn Octave), thực hiện các công việc sau:
    a. Ghi âm một đoạn tiếng nói sao cho có cả âm vô thanh và hữu thanh và lưu dưới dạng file *.wav
    b. Đọc file vừa ghi và thực hiện biểu diễn dạng sóng tín hiệu trong miền thời gian
    c. Đọc file vừa ghi, tách các phân đoạn tương ứng với âm vô thanh, hữu thanh và biểu diễn phổ tương ứng

d. Đọc file vừa ghi, thực hiện biểu diễn spectrogram và quan sát đặc điểm của nó. Đối chiếu với những nhận xét có được trong phần học lý thuyết ở trên.

CHƯƠNG 2. PHÂN TÍCH TÍN HIỆU TIẾNG NÓI

2.1. MỞ ĐẦU

Trong chương này ta sẽ xem xét các phương pháp phân tích tín hiệu tiếng nói. Phân tích tiếng nói thực hiện việc giải quyết các vấn đề để tìm ra một dạng thức tối ưu biểu diễn được tín hiệu tiếng nói một các hiệu quả. Mục tiêu của việc thực hiện phân tích tín hiệu tiếng nói là nhằm trích chọn các đặc trưng của tín hiệu tiếng nói. 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 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 trò 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 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.

2.2. KHÁI NIỆM CHUNG VỀ PHÂN TÍCH TIẾNG NÓI

2.2.1 Mô hình phân tích tín hiệu tiếng nói

Mô hình tổng quát cho việc phân tích tiếng nói được trình bày trong hình 2.1. Các dạng tín hiệu tại các bước cũng được trình bày kèm theo trong minh họa.

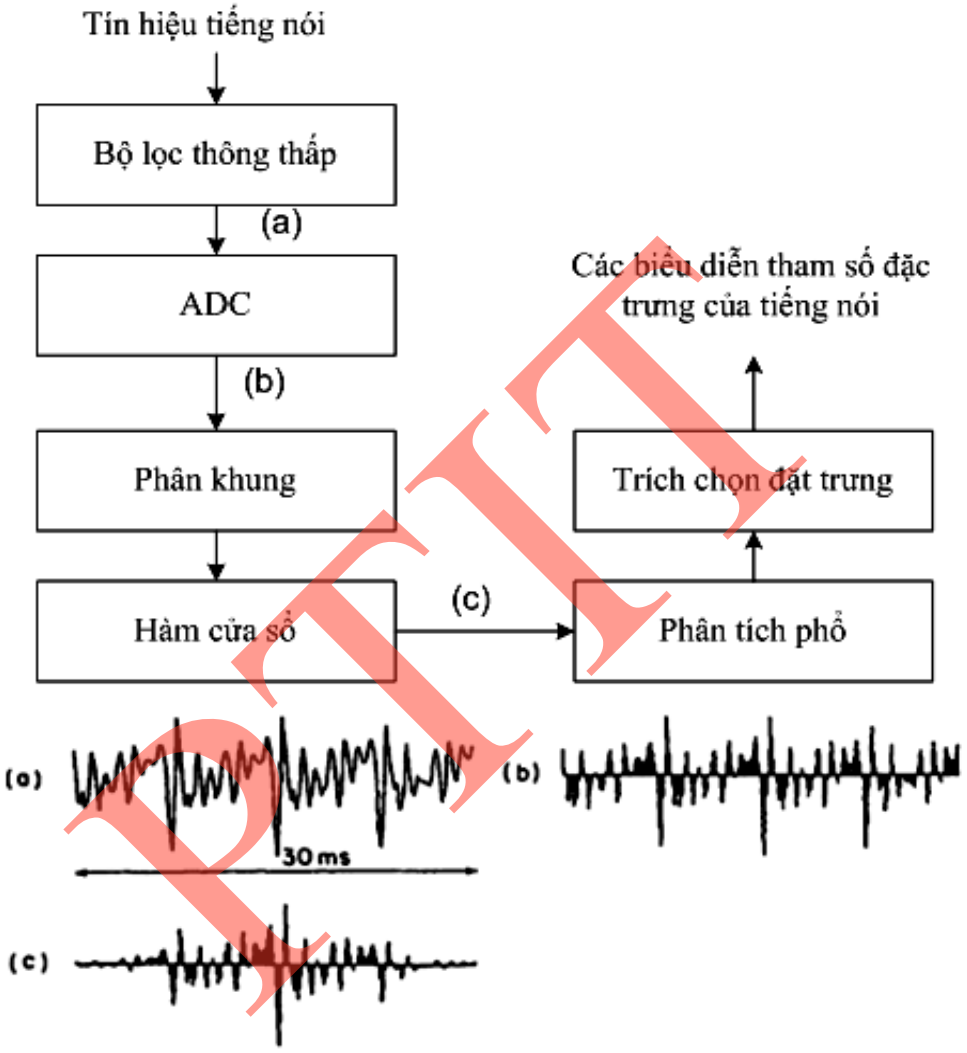
Tín hiệu tiếng nói tương tự (tự nhiên) được tiền xử lý bằng cách cho qua một bộ lọc thông thấp với tần số cắt thích hợp (thường khoảng 8kHz). Tín hiệu thu được sau đó được thực hiện quá trình biến đổi sang dạng tín hiệu tiếng nói số nhờ bộ biến đổi ADC. Thông thường, tần số lấy mẫu bằng 16kHz với tốc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

Tín hiệu tiếng nói dạng số được phân khung với chiều dài khung thường tương ứng với khoảng 30ms tín hiệu và khoảng lệch giữa các khung thường bằng ½-1/2 khung phân tích (khoảng 10ms tín hiệu). 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 thích hợp, chẳng hạn phân tích phổ như STFT, LPC,... Hoặc sau khi thực hiện các phép phân tích cơ bản, tín hiệu tiếp tục được đưa đến các khối để trích chọn các đặc trưng.

2.2.2 Phân tích ngắn hạn

Tín hiệu tiếng nói được tạo ra từ một hệ thống tuyến âm thay đổi theo thời gian cùng với tín hiệu kích thích cũng thay đổi theo thời gian. Trong khi đó, hầu hết các công cụ phân tích tín hiệu đã học khi nghiên cứu về hệ thống và xử lý tín hiệu đều giả thiết rằng

chúng không đổi theo thời gian, tức là giả thiết chúng là các thể hiện của quá trình dừng. Điều này có nghĩa là những công cụ đã học không thể đưa váo áp dụng một cách trực tiếp cho xử lý phân tích tín hiệu tiếng nói. Trong trường hợp vẫn áp dụng một cách vô thức thì kết quả tính toán được cũng không có hoặc có rất ít ý nghĩa cho việc phân tích tín hiệu.



Hình 2.1 Sơ đồ khối quá trình phân tích tín hiệu tiếng nói

Khi nói đến các phân tích tín hiệu tiếng nói, người ta thường mặc định các phân tích này được tiến hành trong một phân đoạn tín hiệu tương ứng với thời gian rất nhỏ, cỡ khoảng 10-30ms. Và do đó, các phân tích này được gọi là phân tích ngắn hạn. Sở dĩ như vậy là vì bản chất của tín hiệu tiếng nói, như đã đề cập trong chương trước, nó là tín hiệu bán tĩnh: các tham số chỉ có thể coi là không thay đổi nếu thời gian quan sát đủ ngắn.

Việc thực hiện phân tích ngắn hạn có thể được thực hiện trong miền thời gian hoặc miền tần số. Việc được thực hiện phân tích trong miền nào phụ thuộc vào những thông

tin/đặc trưng của tín hiệu tiếng nói mà ta mong muôn trích xuất. Chẳng hạn, các tham số như năng lượng ngắn hạn, tốc độ trở về không ngắn hạn, giá trị hàm tự tương quan ngắn hạn được tính toán và xác định trong miền thời gian. Trong khi đó, phổ ngắn hạn được tính toán xác định bằng phân tích ngắn hạn trong miền tần số.

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

$$X(n) = \sum_{m=-\infty}^{\infty} T\{s_n(m)\}$$

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 trên được tính với giới hạn vô cùng được hiểu là phép lấy tổng của tất cả các thành phần khác không của khung tín hiệu thu được sau phép lấy cửa sổ. Nói cách khác, tổng được thực hiện với mọi giá trị của m trong vùng xác định (support) của hàm cửa sổ.

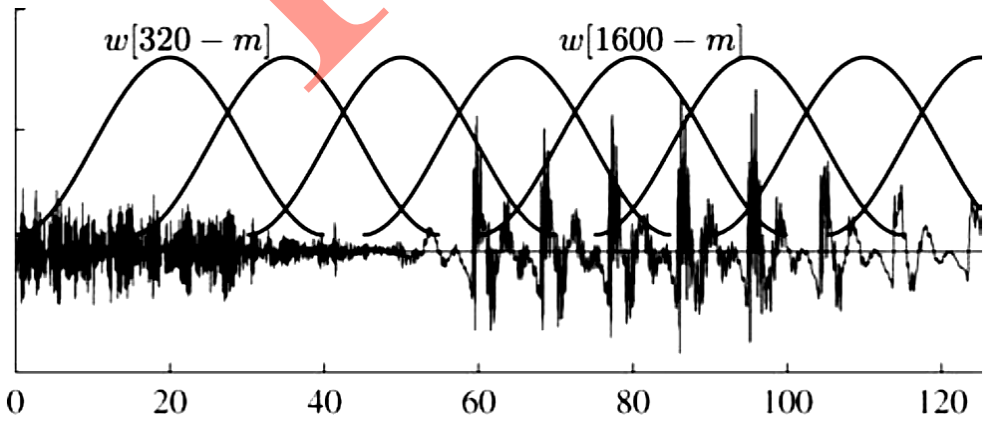
2.2.3 Hàm cửa sổ phân tích

Để thực hiện các phân tích trên các phân đoạn tín hiệu ngắn hạn, chúng ta phải thực hiện việc "cắt" ra các đoạn tín hiệu này. Việc "cắt" này có thể thực hiện được thông qua một phép nhân với hàm cửa sổ. Giả sử tín hiệu tiếng nói số s(n), khi đó phân đoạn tín hiệu có độ dài N mẫu có thể xác định bởi công thức:

$$s_N(n) = s(m) \times w(n-m)$$

trong đó, w(n) là hàm cửa sổ, hay còn gọi là cửa sổ phân tích có độ dài N mẫu. Để đơn giản chúng ta ký hiệu $s_N(n) = s_n(m)$ để vừa có thông số về vị trí của các mẫu s(m) trong của sổ phân tích ở vị trí n.

Hình 2.2 minh họa việc phân chia khung với hàm cửa sổ.



Hình 2.2 Minh họa của sổ phân tích tín hiệu với các đoạn bao trùm nhau

Tùy theo mục đích nghiên cứu mà hàm cửa sổ phân tích có các hình dạng khác nhau. Hình dạng đơn giản nhất là cửa sổ hình chữ nhật. Tuy nhiên, để đạt được hiệu quả mong muốn, người ta thường hay sử dụng cửa sổ Hamming, hoặc Hanning.

Độ rộng của cửa sổ được quyết định bởi việc lựa chọn phân tích ngắn hạn.

2.3. CÁC PHÂN TÍCH CƠ BẢN TRONG MIỀN THỜI GIAN

Phân tích tiếng nói trong miền thời gian là phân tích trực tiếp trên dạng sóng tín hiệu sau khi thực hiện việc lấy cửa sổ tín hiệu trong miền thời gian. Như đã đề cập trong phần trước, ta chỉ xem xét các phân tích ngắn hạn của tín hiệu. Do đó, để đơn giản trong trình bày 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.

2.3.1 Năng lượng ngắn hạn

Tham số đầu tiên cần quan tâm trong phân tích tín hiệu tiếng nói trong miền thời gian đó là *năng lượng ngắn hạn.*

Năng lượng gắn với tín hiệu tiếng nói cũng là một đại lượng thay đổi theo thời gian. Năng lượng của một phân đoạn tín hiệu tiếng nói gồm N mẫu được xác định bởi công thức:

$$E_T = \sum_{n=0}^{N-1} s_N^2(n)$$

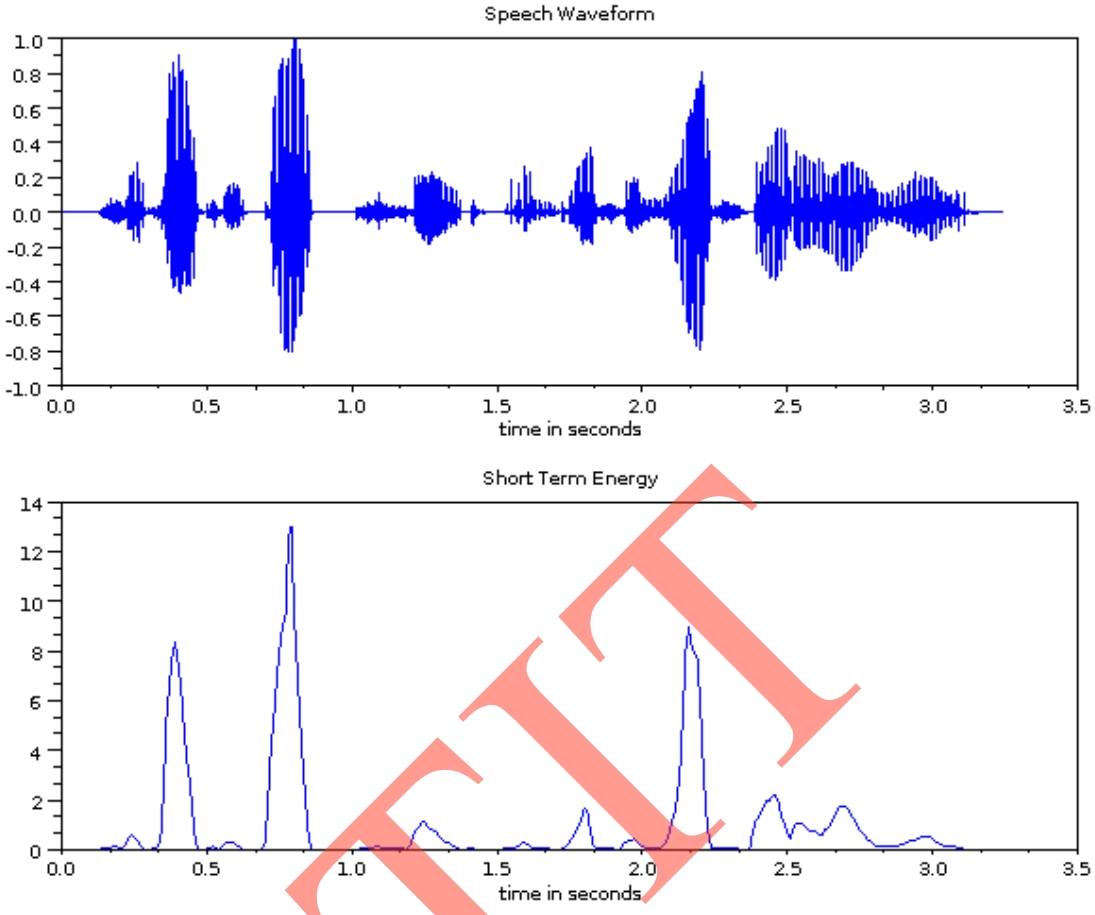
Giá trị này còn được gọi là năng lượng tổng của một phân đoạn tín hiệu

Mở rộng biểu thức trên, chúng ta có công thức tính năng lượng ngắn hạn như sau:

$$E_T(n) = E_n = \sum_{m=-\infty}^{\infty} s_n^2(m) = \sum_{m=-\infty}^{\infty} (s(m)w(n-m))^2$$

Trong công thức này, chỉ số n chạy/dịch trên trục các mẫu tại những vị trí mà chúng ta quan tâm đến giá trị năng lượng ngắn hạn. n có thể bằng 1, ứng với mỗi lần dịch một mẫu, hoặc có thể bằng N (bằng kích thước cửa sổ phân tích), hoặc lớn hơn. Giá trị n rất nhỏ thường là không cần thiết vì các mức năng lượng trong khoảng thời gian nhỏ gần như không thay đổi. Ngược lại, nếu rất lớn (>=N), tức là các khung phân tích không có sự bao trùm nhau, có thể dẫn đến sự mất thông tin. Điều này là bởi vì sự thay đổi quan sát được có thể bắt đầu từ phần cuối của đoạn trước, nhưng bị ngắt quãng sang đến đầu khung sau. Thường giá trị n được thiết lập sao cho sự bao trùm giữa các khung phân tích tín hiệu khoảng bằng ½-1/3 của khung.

Hình 2.3 minh họa năng lượng ngắn hạn của một đoạn âm thanh.



Hình 2.3: Minh họa năng lượng ngắn hạn của tín hiệu tiếng nói

Từ minh họa chúng ta thấy, những phân đoạn tương ứng với âm hữu thành (nguyên âm), mức năng lượng ngắn hạn rất lớn. Ở những phân đoạn tương ứng với âm vô thanh, mức năng lượng ngắn hạn rất nhỏ. Ở những phân đoạn tương ứng với khoảng lặng, mức năng lượng ngắn hạn bằng không (xấp xỉ bằng không).

Như vậy, việc xác định năng lượng ngắn hạn 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 là một công cụ hữu ích để phát hiện một tín hiệu âm là của âm hữu thanh, âm vô thanh hay một khoảng lặng.

Cần chú ý rằng độ dài cửa sổ phân tích phải được chọn thích hợp theo nguyên tắc của phân tích ngắn hạn đã đề cập ở trên. 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 đáng kể giá trị ước lượng năng lượng toàn khung.

2.3.2 Độ lớn biên độ ngắn hạn

Từ phần trên thấy rằng năng lượng ngắn hạn của 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 ngắn hạn*, được xác định bởi:

$$M_n = \sum_{m=-\infty}^{\infty} |s_n(n)| = \sum_{m=-\infty}^{\infty} |s(m)|w(n-m)$$

2.3.3 Vi sai độ lớn biên độ ngắn hạn

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

$$\Delta M_n(\eta) = \sum_{m=-\infty}^{\infty} |s_n(m) - s_n(m-\eta)| = \sum_{m=-\infty}^{\infty} |s(m) - s(m-\eta)|w(n-m)$$

Công thức trên cho thấy giá trị hàm vi sai 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 sai 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.

2.3.4 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 - ZCR). Sự kiện trở về không xảy ra khi 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:

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

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. Z_n là tổng trọng số của tất cả các thay đổi dấu của các mẫu trong vùng xác định bởi cửa sổ phân tích. 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 giữa 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 cho phép phát hiện xem một tín hiệu là tín hiệu của âm vô thanh, hữu thanh hay khoảng lặng.

2.3.5 Giá trị 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:

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

Công thức trên 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 đã biết trong môn học Xử lý tín hiệu số, có một mối quan hệ giữa hàm tự tương quan và năng lượng tín hiệu:

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

2.4. PHÂN TÍCH PHỔ TÍN HIỆU TIẾNG NÓI

2.4.1 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:

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

Biểu thức trên có thể viết lại thành:

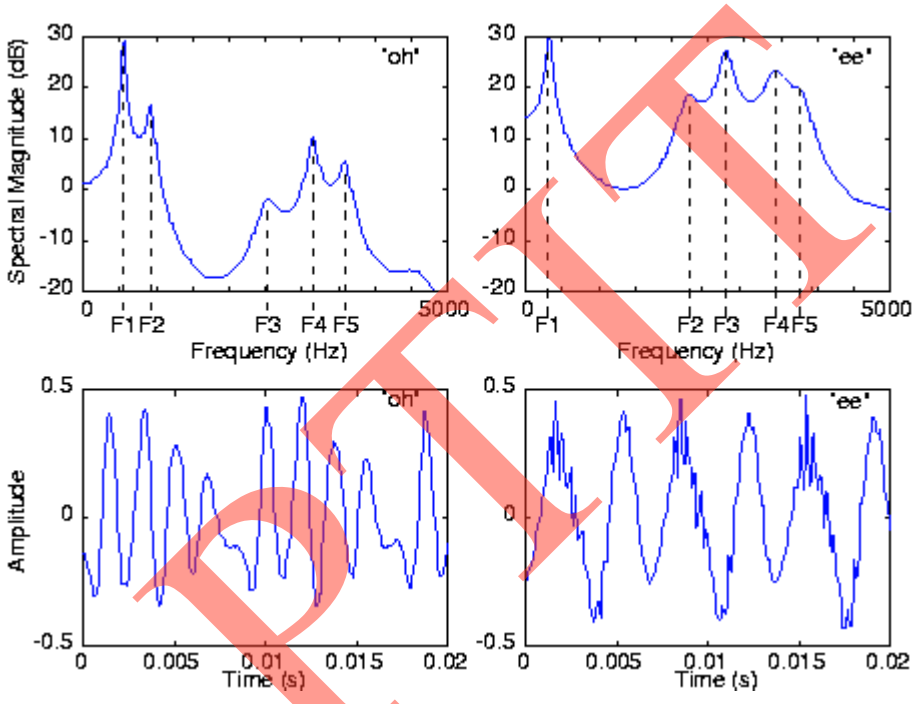
$$S_n\left(e^{j\omega}\right) = \left(s(\tilde{n})e^{-j\omega\tilde{n}}\right) * \mathrm{w}(\tilde{n})|_{\tilde{n}=n}$$

Biểu thức này 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 biến đổi phổ cũng có thể viết là:

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

Công thức trên có thể diễn giải như sau: Tín hiệu $s(\tilde{n})$ được đưa qua bộ lọc thông dải có tần số trung tâm ω và đáp ứng xung $w(\tilde{n})e^{j\omega\tilde{n}}$. 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.



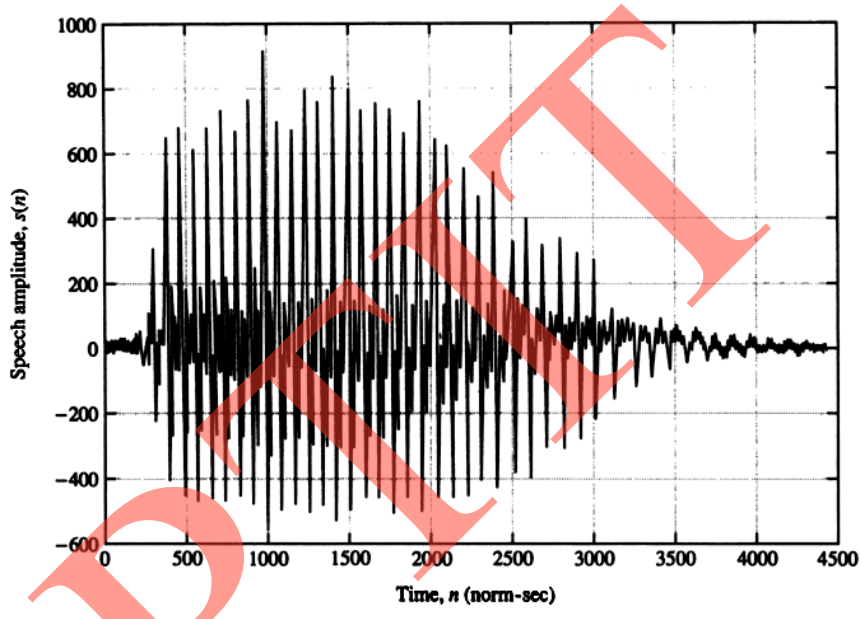
Hình 2.3 Minh họa một khung tín hiệu và phổ tương ứng.

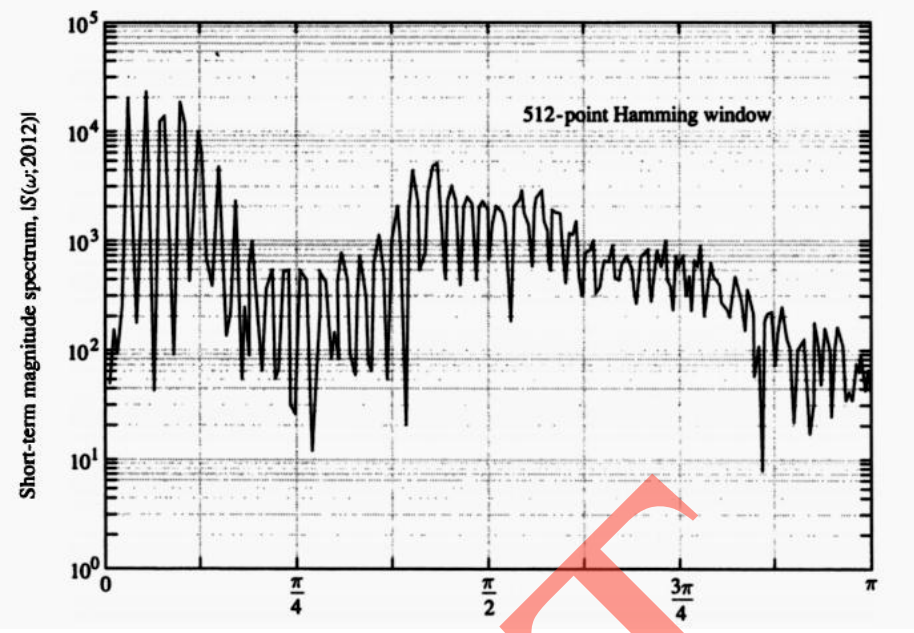
Mật độ phổ công suất trong một khoảng thời gian ngắn, tức là phổ ngắn hạn của tín hiệu tiếng nói, có thể được xem như là tích của hai thành phần: thành phần thứ nhất là đường biên phổ thay đổi chậm theo tần số; thành phần thứ hai là cấu trúc phổ mịn (spectral fine structure) thay đổi rất nhanh theo tần 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 thanh môn (glottal) ở môi và khoang mũi. Trong khi đó, cấu trúc phổ mịn mô tả tính tuần hoàn của nguồn âm.

Công thức đầu tiên 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 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 ta có:

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

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 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 2.4 Khung tín hiệu và phổ tương ứng

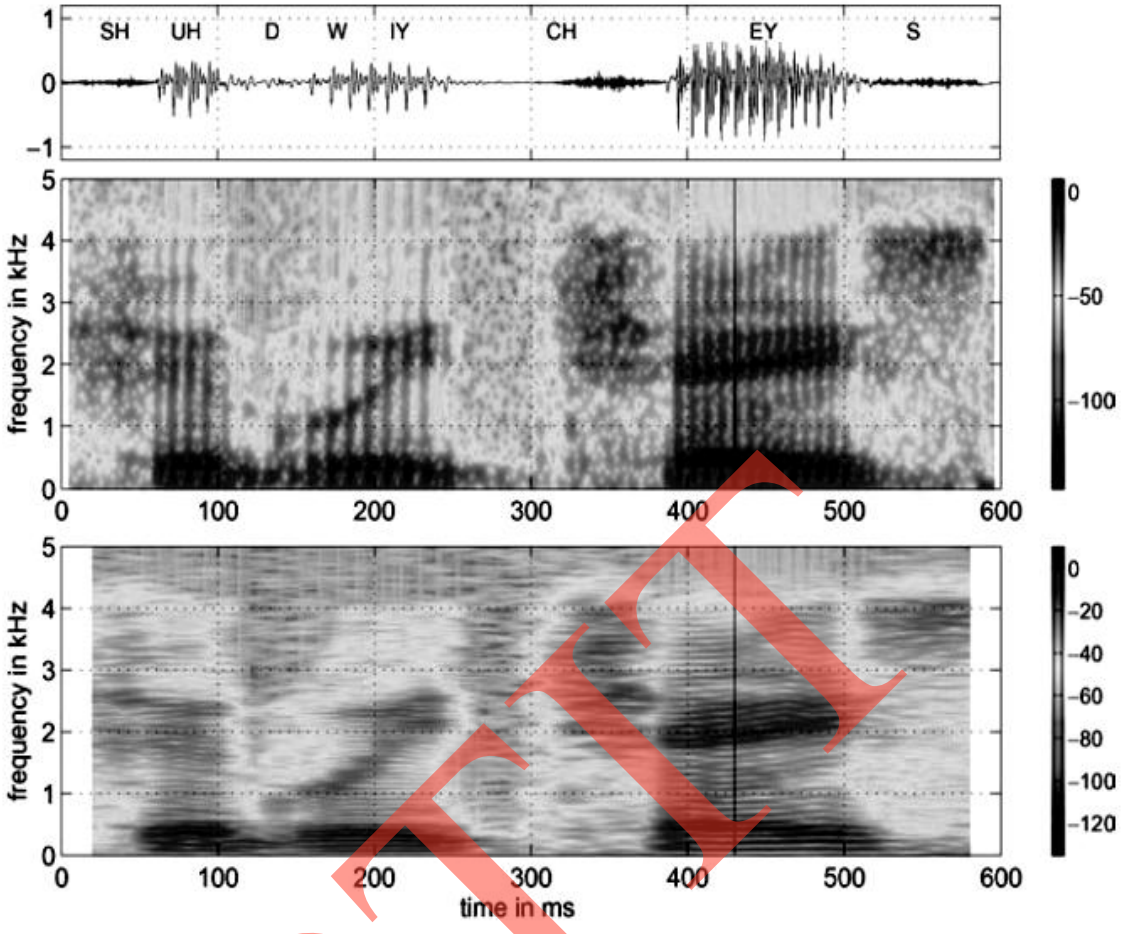
2.4.2 Phân tích spectrogram

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 hình 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 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 vì các thông tin này 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 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)

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

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

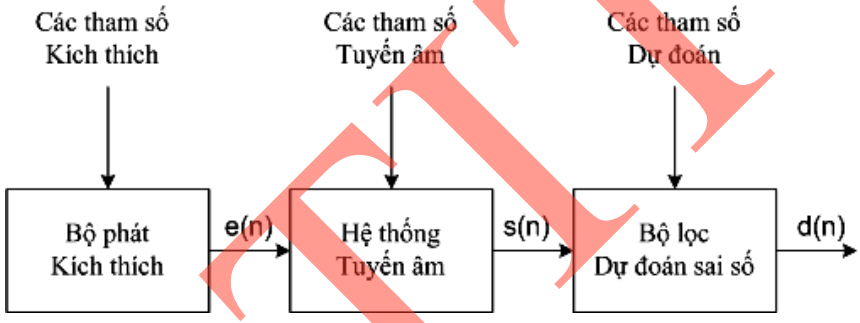
Hai lược đồ spectrogram được xây dựng với các hàm cửa sổ có độ dài khác nhau. Lược đồ spectrogram phía trên là kết quả khi sử dụng cửa sổ có chiều dài 101 mẫu tương ứng với 10ms. Chiều dài của cửa sổ phân tích này xấp xỉ bằng chu kỳ của dạng sóng trong các khoảng tín hiệu âm hữu thanh. Kết quả là trong 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 gồm hầu hết các mẫu có biên độ lớn, lúc gồm 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 hình 2.5 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 ta thấy, lược đồ spectrogram tương ứng không còn nhạy với sự thay đổi về thời gian nữa.

2.5. PHÂN TÍCH DỰ ĐOÁN TUYẾN TÍNH

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 là 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 2.6. 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:

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



Hình 2.6 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 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:

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

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

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

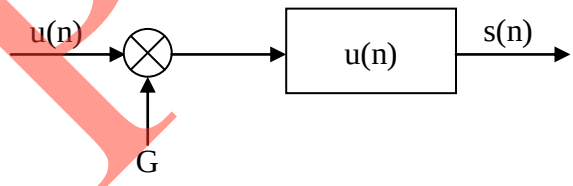
Trong công thức trên, u(n) là thành phần kích thích chuẩn và G là hệ số khuếch đại của thành phần kích thích. Nếu xem xét biểu thức trên trong miền z ta có biểu thức:

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

Hay hàm truyền đạt tương ứng là:

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

Hàm truyền đạt thu được biểu diễn trong sơ đồ khối trong hình 3.6. Nguyên lý hoạt động của sơ đồ khối như sau: Nguồn kích thích chuẩn 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). 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, ta xây dựng được mạch tổng hợp tín hiệu tiếng nói dựa vào mô hình phân tích LPC như trong hình 2.7. 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 2.7 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 $\tilde{s}(n)$:

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

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

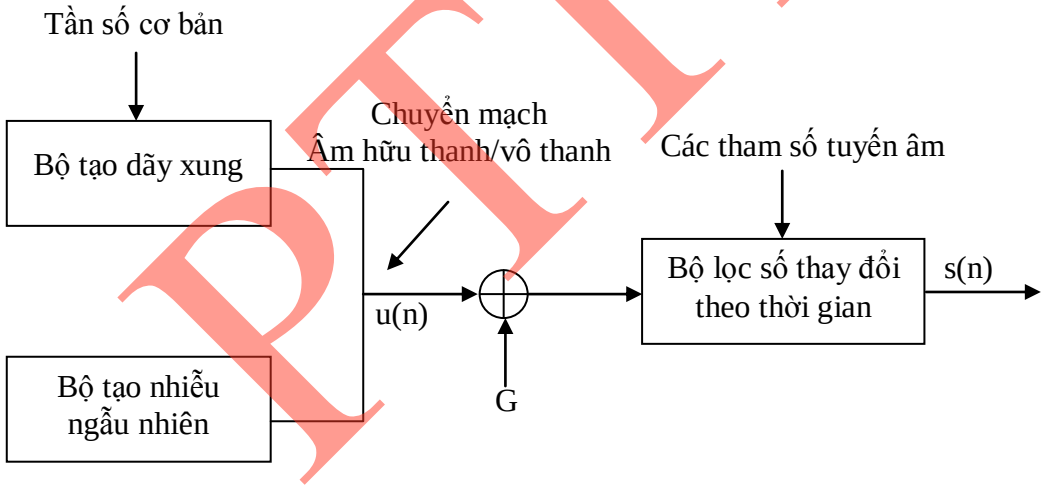
$$e(n)=s(n)-\tilde{s}(n)=s(n)-\sum_{k=1}^{p}a_k s(n-k)$$

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

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

Từ đó 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 2.8 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 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 2.8 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, ta định nghĩa các khung tín hiệu ngắn hạn và tương ứng là các sai số ngắn hạn:

$$s_n(m)=s(n+m)$$

$$e_n(n)=e(n+m)$$

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

$$\varepsilon_n=\sum_m e_n^2(m)$$

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

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

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

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

Khi đó ta có:

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

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:

$$\Psi_n(i,k) = \sum_m s_n(m-i) s_n(m-k)$$

Ta có thể thu gọn biểu thức như sau:

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

Biểu thức tính trên 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:

$$\hat{\varepsilon}_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m) s_n(m-k)$$

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

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.

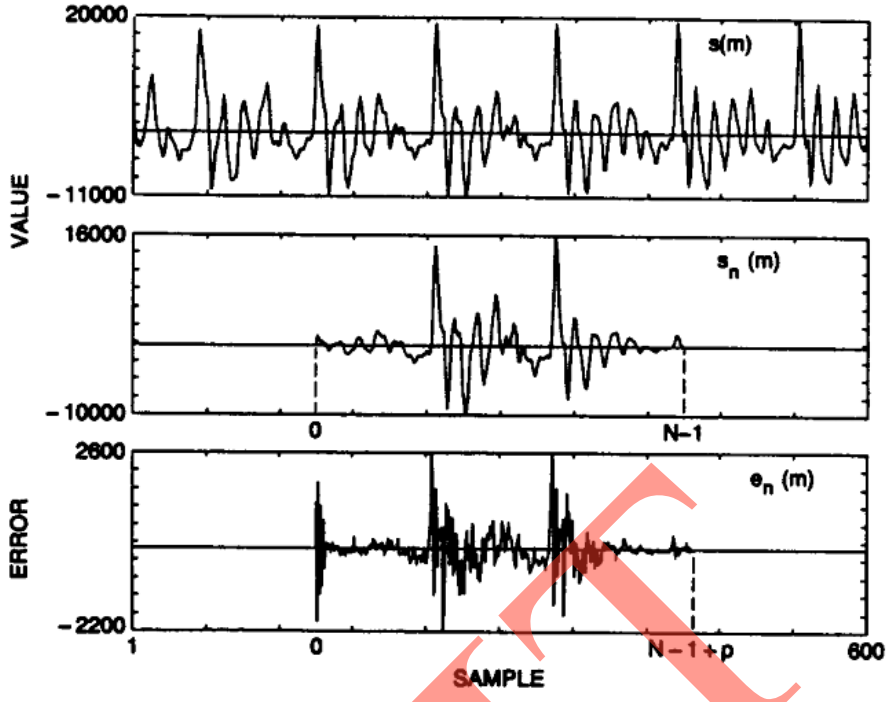
Để tìm các hệ số dự đoán tối ưu $\hat{a}_k$ trước hết ta tính $\Psi_n(i,k)$ ($1 \le i \le p$ và $0 \le k \le p$) và sau đó giải hệ đồng thời của p biểu thức. Trong thực tế, việc giải hệ và tính toán các thành phần Ψ phụ thuộc rất nhiều vào khoảng thời gian m được sử dụng để định ra khung tín hiệu phân 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.

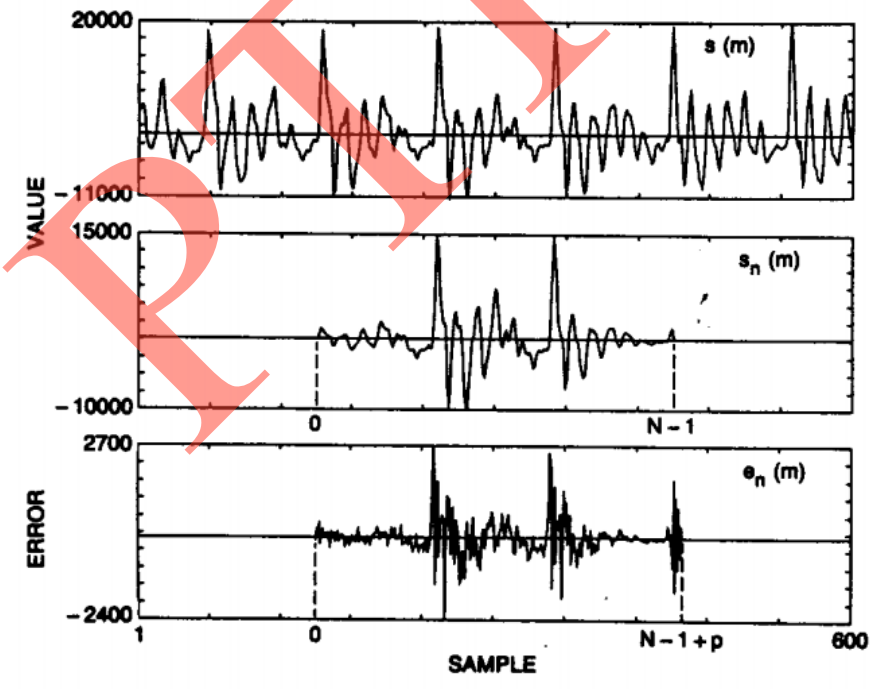
Phương pháp sử dụng hàm tự tương quan xuất phát trực tiếp từ việc định ra khoảng giới hạn m trong tổ hợp tuyến tính sao cho đoạn tín hiệu tiếng nói $s_n(m)$ bằng 0 ở ngoài khoảng $0 \le m \le N-1$. Điều này tương đương với việc giả thiết tín hiệu tiếng nói s(n+m) được nhân với hàm của sổ w(m) hữu hạn 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:

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

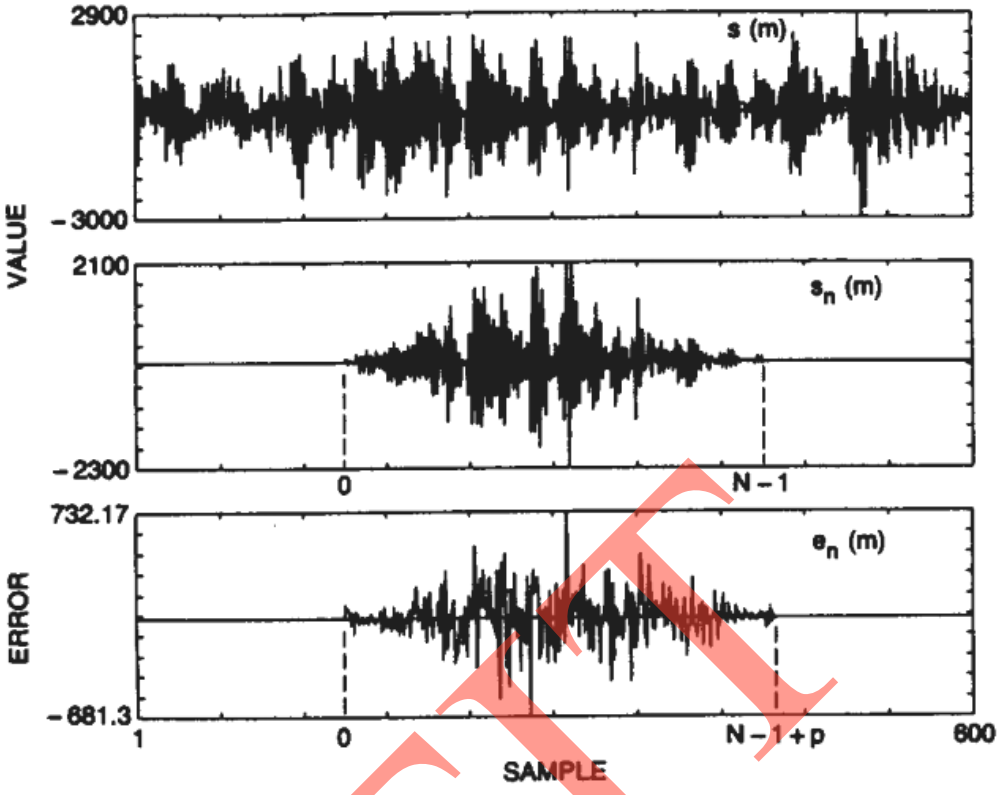
Từ công thức (3.31), khi m<0 tín hiệu sai số $e_n(m)$ bằng 0 vì khi đó $s_n(m)=0$. Mặt khác, cũng tương tự khi m>N-1+p sẽ không có sai số dự đoán bởi vì khi đó ta cũng có $s_n(m)=0$. Tuy nhiên trong vùng m=0 (tức là từ m=0 đến m=p-1) tín hiệu thu được sau khi thực hiện việc lấy cửa sổ có thể được dự 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 2.9 - 2.11.



Hình 2.9 Minh họa trường hợp sai số dự đoán lớn ở đầu khung với tín hiệu âm hữu thanh



Hình 2.10 Minh họa trường hợp sai số dự đoán lớn ở cuối khung với tín hiệu âm hữu thanh



Hình 2.11 Minh họa trường hợp sai số dự đoán lớn với tín hiệu âm vô thanh

Mục đích của việc lấy của sổ 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.

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

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

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

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

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

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

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

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

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:

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

Nếu biểu diễn dưới dạng ma trận ta có:

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

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:

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

Và covariance được tính bởi:

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

Hoặc bằng cách đổi chỉ số:

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

Để ý thấy rằng việc tính toán theo biểu thức trên 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:

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

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.

2.6. XỬ LÝ ĐỒNG HÌNH

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:

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

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

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

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

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

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 trên là các định nghĩa dựa trên các công thức toán học. Để công thức có ý nghĩa trong các phân tích thực tế, 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.

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

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

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

2.7. ÁP DỤNG MỘT SỐ PHÉP PHÂN TÍCH ĐỂ XÁC ĐỊNH CÁC THAM SỐ CƠ BẢN CỦA TÍN HIỆU TIẾNG NÓI

2.7.1 Một số phương pháp xác định các 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, 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).

2.7.2 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.

2.7.3 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 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 ưu 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.

2.7.4 Một số phương pháp xác định tần số cơ bản

Tần số cơ bản F_0 là tần số dao 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).

2.7.5 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.

2.7.6 Sử dụng Vi sai độ lớn biên độ ngắn hạn

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, 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.

2.7.7 Sử dụng tốc độ trở về không

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. 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.

2.7.8 Sử dụng phân tích 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.

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

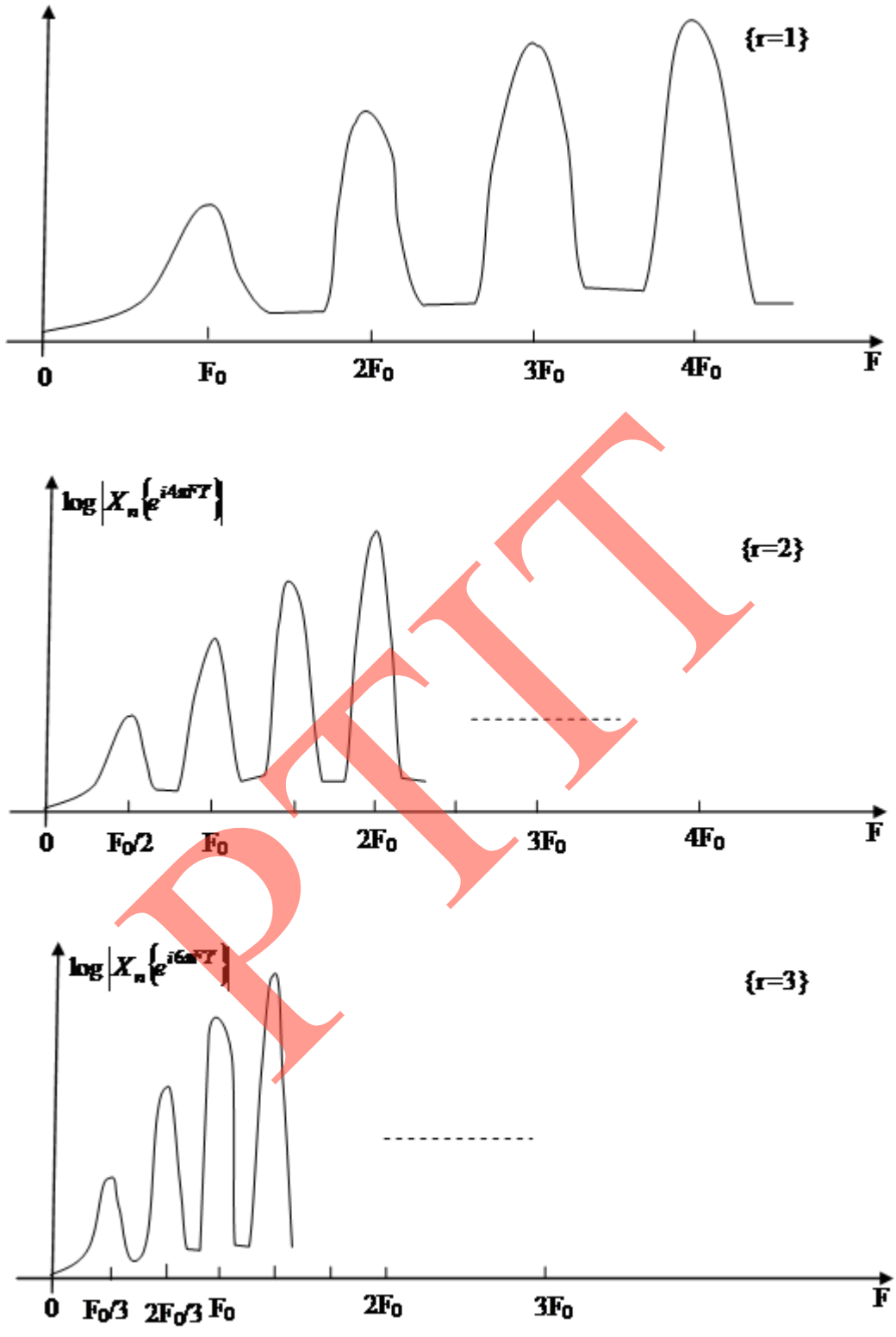
$$P_n\left(e^{j\omega}\right)=\prod_{r=1}^{K}\left|S_n\left(e^{j\omega r}\right)\right|$$

Lấy lô-ga-rít của phổ tích các hài, thu được phổ tích các hài trong thang lô-ga-rít:

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

Hàm $\hat{P}_n\left(e^{j\omega}\right)$ trong công thức trên là một tổng của K phổ nén tần số của $|S_n(e^{j\omega})|$.

Việc sử dụng hàm trong công thức trên 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 2.12 minh họa nhận xét vừa nêu.

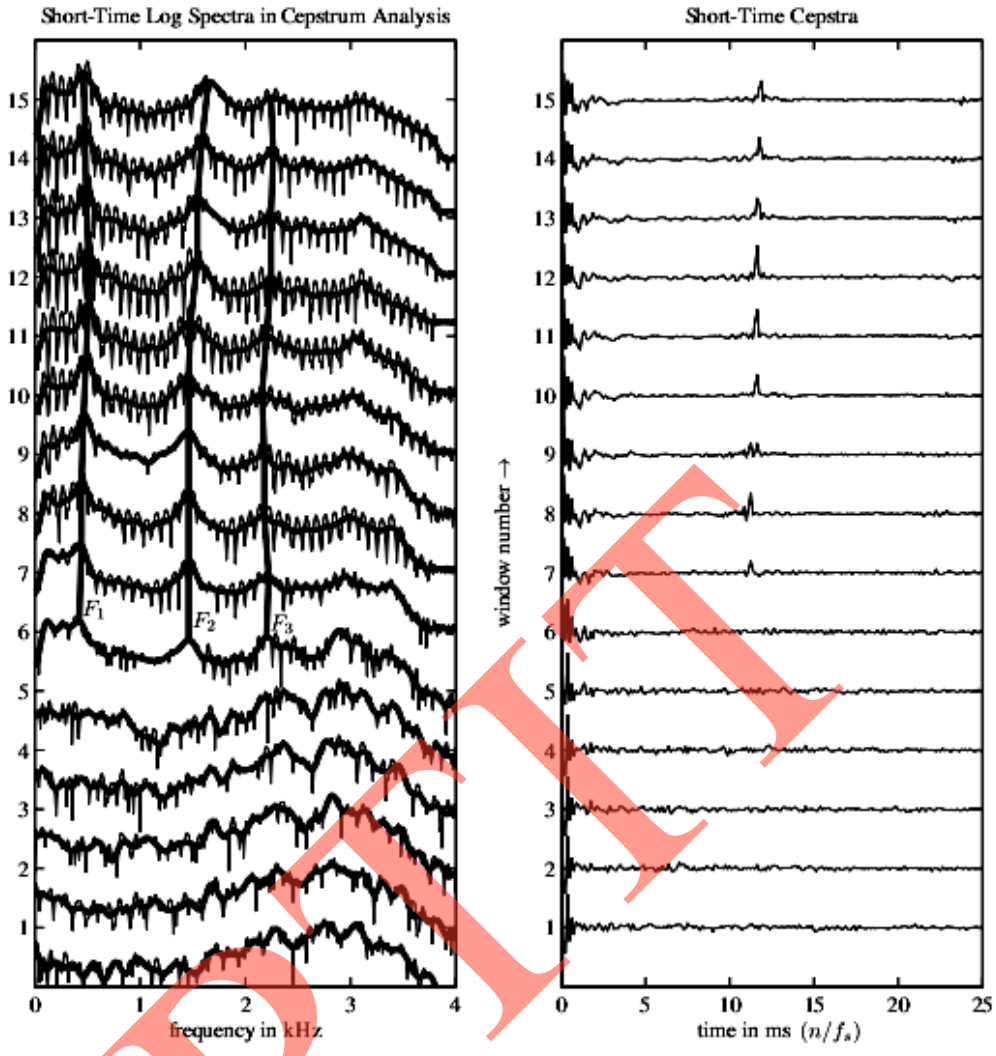


Hình 2.12 Minh họa sự nén tần số

2.7.9 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 2.13 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, 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 2.13 Lô-ga-rít các thành phần hài trong phổ tín hiệu

2.8. CÂU HỎI VÀ BÀI TẬP CUỐI CHƯƠNG

1. Mục đích của việc Xử lý tiếng nói? Liệt kê một số phép xử lý phân tích tiếng nói cơ bản
2. Các phương pháp phân tích tiếng nói trong miền thời gian? Ứng dụng của các phương pháp này?
3. Phương pháp phân tích phổ tín hiệu tiếng nói?
4. Tại sao với tiếng nói phải thực hiện phân tích ngắn hạn?
5. Có thể dùng những tham số nào để xác định điểm đầu cuối trong một đoạn âm thanh?
6. Phân tích LPC: nguyên lý, hệ phương trình, áp dụng?

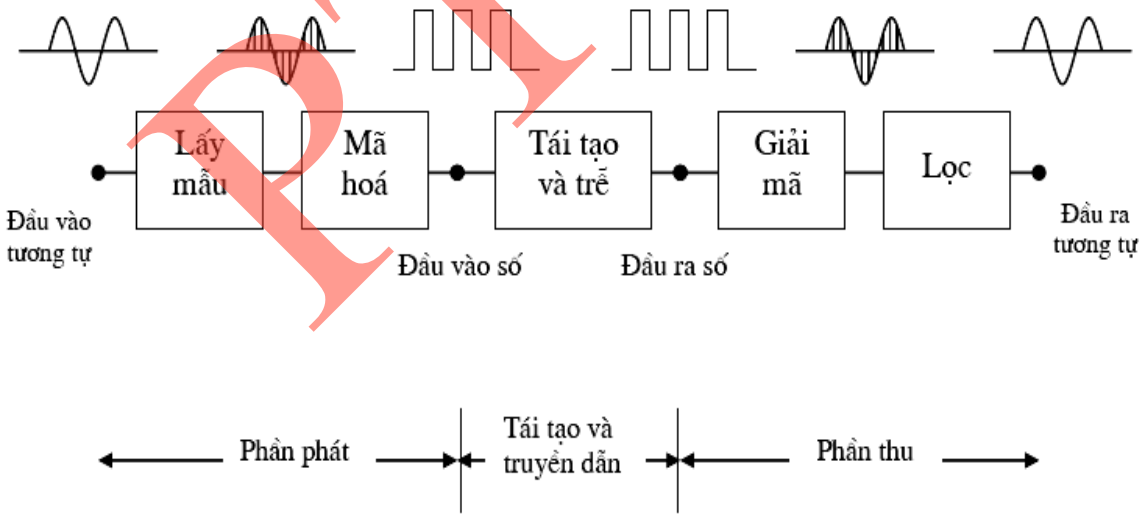
7. Phân tích cepstral: nguyên lý, công thức tính, áp dụng?
8. Xét một phân đoạn tín hiệu tiếng nói sau {0 0.6442 0.9854 0.8632 0.3350 -0.3508 -0.8716 -0.9825 -0.6313}. Biết đây là mẫu của một phân đoạn tín hiệu tiếng nói được lấy mẫu với tần số lấy mẫu là 8000Hz. Hãy xác định các thông số cơ bản cho phân đoạn tín hiệu bằng phân tích trong miền thời gian. Giả sử cửa sổ phân tích là cửa sổ chữ nhật có chiều rộng N=4 điểm mẫu.
9. (Matlab) Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) cùng công cụ chỉnh sửa âm thanh Audicity (hoặc công cụ khác) thực hiện các công việc sau:
    i. Với cùng một nội dung thông tin, các thành viên trong nhóm lần lượt phát âm (đọc/nói) và ghi âm phát âm của các nguyên âm tiếng Việt. Lưu tệp ở định dạng *.wav
    ii. Sử dụng phần mềm Matlab (hoặc các bộ công cụ, ngôn ngữ lập trình khác) và kiến thức đã học trong chương này:
        1. Xác định tần số cơ bản của phát âm tương ứng của mỗi thành viên
        2. Xác định formant đầu tiên (F1) trong phát âm của mỗi thành viên. Từ kết quả đó, lập bản đồ phân bố tần số formant của các nguyên âm tiếng Việt của các thành viên trong nhóm
10. (Matlab) Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc công cụ thích hợp):
    i. Ghi một file tín hiệu tiếng nói của cụm từ “Xin chào các bạn”, ghi file dưới dạng *.wav
    ii. Sử dụng thư viện của Matlab (hoặc các công cụ thích hợp) thực hiện phân tích LPC của đoạn tín hiệu tiếng nói trên
    iii. Sử dụng thư viện của Matlab (hoặc các công cụ thích hợp) thực hiện phân tích LPC của đoạn tín hiệu tiếng nói trên

CHƯƠNG 3: MÃ HÓA TIẾNG NÓI

3.1. KHÁI NIỆM CHUNG VỀ MÃ HÓA TIẾNG NÓI

Mã hoá là quá trình biến đổi các giá trị rời rạc thành các mã tương ứng. Mã hóa tín hiệu tiếng nói (gọi tắt là mã hóa tiếng nói), còn được biết đến là mã hóa tín hiệu thoại, được biết đến từ rất sớm. Ngay từ những năm 1930, mã hóa tín hiệu tiếng nói đã được nhiều nhà nghiên cứu và vận hành hệ thống liên lạc điện thoại quan tâm. Sự bùng nổ về các thuật toán mã hóa tín hiệu thoại phải kể đến khi có sự phát triển mạnh của hệ thống thông tin di động và sau đó là sự tích hợp dịch vụ đa phương tiện. Không chỉ có một vai trò quan trọng trong các mạng thông tin dân dụng, mã hóa tiếng nói cũng được ứng dụng và có mặt ở trong hầu hết các hệ thống thông tin số cả dân sự và quân sự.

Mục tiêu của việc mã hóa tiếng nói là nhằm giảm nhỏ lượng dữ liệu biểu diễn thông tin tiếng nói cần lưu trữ hoặc truyền tải mà không làm giảm chất lượng cảm thụ của tiếng nói khôi phục được sau mã hóa. Nói một cách khác, mã hóa tiếng nói là quá trình tìm kiếm biểu diễn số nhỏ gọn nhất có thể của tín hiệu tiếng nói mà vẫn không làm mất hoặc làm mất đi thông tin (méo) ít nhất có thể. Về cơ bản thì mã hóa tín hiệu tiếng nói cũng giống với mã hóa dữ liệu thông thường. Tuy nhiên, với đặc trưng của tín hiệu tiếng nói, bao gồm cả đặc trưng của quá trình tạo và cảm nhận tiếng nói của con người, mã hóa tiếng nói sẽ có nhiều điểm khác biệt và cũng cần những cách tiếp cận riêng biệt để có thể khai thác tốt các đặc trưng.



Hình 3.1 Sơ đồ tổng quan hệ thống mã hóa tiếng nói

Nhìn chung, mã hóa tín hiệu tiếng nói (hay gọi tắt là mã hóa tiếng nói) liên quan đến quá trình xử lý số tín hiệu tiếng nói trong đó có việc lấy mẫu và lượng tử hóa. Nói một cách khác, quá trình mã hóa tiếng nói liên quan trước hết tới quá trình biến đổi các tín hiệu tiếng nói liên tục thành các tín hiệu tiếng nói rời rạc cả về thời gian (lấy mẫu) và

chuẩn hóa về biên độ (lượng tử hóa). Với tín hiệu tiếng nói, từ đặc trưng nghe của tai con người trong đó nhạy với vùng tín hiệu tiếng nói ở tần số 0.3-3.4kHz, do đó trong các hệ thống thông tin thoại người ta thường chỉ quan tâm đến khoảng tín hiệu này. Từ đó, theo định lý lấy mẫu Shannon/Nyquist, tần số lấy mẫu với tín hiệu tiếng nói tối thiểu là 8kHz. Sơ đồ khối tổng quan của hệ thống mã hóa tiếng nói được minh họa trong hình 3.1.

Tín hiệu tiếng nói tương tự được thực hiện tiền xử lý: lọc hạn biên (Anti-aliasing filter), tiền nhấn, khuếch đại, … Sau đó được thực hiện việc số hóa (lấy mẫu và lượng tử hóa). Ở một dạng thức đơn giản nhất, việc thực hiện số hóa này có thể coi là một quá trình mã hóa. Tuy nhiên, để đạt được các hiệu quả mã hóa tốt hơn, một loạt các quá trình phân tích khác sẽ được áp dụng trên tín hiệu tiếng nói số thu được. Quá trình giải mã nhằm tái tạo tín hiệu tiếng nói thực hiện các thao tác ngược lại với quá trình mã hóa. Cũng cần chú ý rằng trong quá trình mã hóa, có một khâu mà không thể thực hiện chính xác quá trình ngược lại, đó chính là quá trình lượng tử hóa.

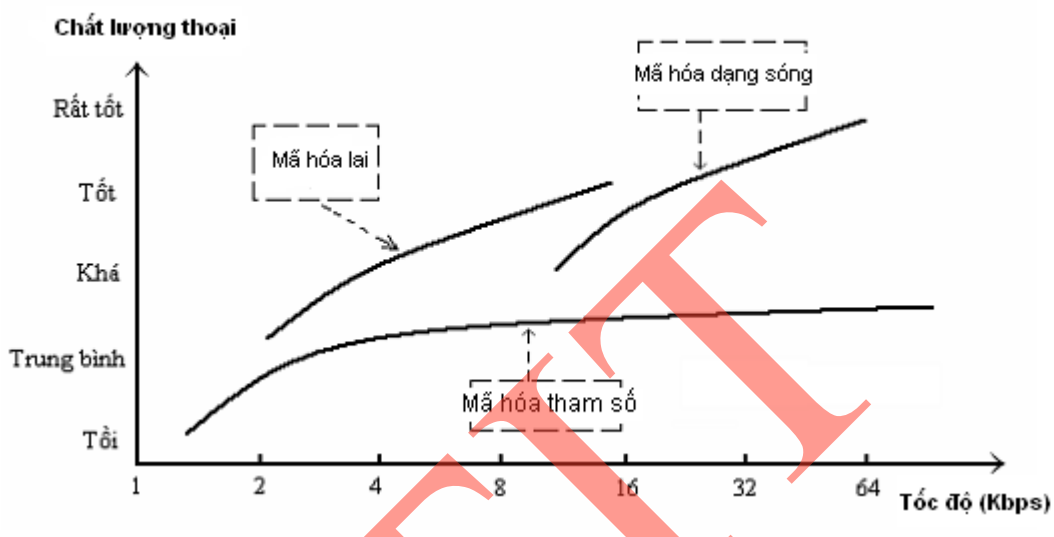
Như đã đề cập, có rất nhiều cách tiếp cận bài toán mã hóa tiếng nói. Kết quả là có rất nhiều phương pháp mã hóa. Việc phân loại các phương pháp mã hóa do đó không hề đơn giản. Tùy vào cách nhìn nhận về vấn đề, hay tùy vào sự quan tâm trong quá trình mã hóa, người ta có nhiều cách phân loại.

Nếu dựa trên cách tiếp cận và miền tiếp cận tín hiệu chúng ta có: (1) mã hóa trực tiếp dạng sóng miền thời gian, (2) mã hóa dạng sóng miền tần số. Mã hóa trực tiếp dạng sóng (waveform coding) là kỹ thuật mã hóa khai thác đặc trưng về hình dạng sóng tín hiệu trực tiếp trong miền thời gian. Đây là một cách tiếp cận phổ dụng và có thể áp dụng cho bất cứ loại tín hiệu nào chứ không riêng gì cho tín hiệu tiếng nói. Các phương thức mã hóa phổ biến thuộc lớp này như PCM, DPCM, …Đây là phương pháp mã hóa không hiệu quả nếu xét về mặt nén dữ liệu. Tuy nhiên, chất lượng theo nghĩa độ trung thực khi khôi phục tín hiệu mã hóa của các phương pháp này khá cao. Khác với lớp mã hóa trực tiếp dạng sóng, mã hóa dạng sóng trong miền tần số thực hiện việc mã hóa tín hiệu dựa trên các đặc trưng phổ của tín hiệu. Phương thức mã hóa này còn được gọi là mã hóa chuyển đổi (transform coding).

Nếu phân loại theo tốc độ mã hóa, chúng ta có: (1) mã hóa tốc độ cao, (2) mã hóa tốc độ trung bình, (3) mã hóa tốc độ thấp, (4) mã hóa tốc độ rất thấp. Xu hướng phát triển hiện nay của các phương pháp mã hóa tiếng nói là các phương pháp tiếp cận mã hóa tốc độ rất thấp, khoảng 2.4kbps hoặc thấp hơn.

Ngoài ra, người ta cũng thường phân loại các phương pháp mã hóa dựa trên phương thức tiếp cận. Với cách phân loại này, chúng ta có: (1) mã hóa trực tiếp dạng sóng, (2) mã hóa dựa trên tham số tín hiệu tiếng nói, (3) phương pháp mã hóa lai ghép. Khác với mã hóa trực tiếp dạng sóng, phương thức mã hóa mà chúng ta đã biết trong phần trên, phương pháp mã hóa dựa trên tham số tín hiệu (gọi tắt là mã hóa tham số) sử dụng

nguyên lý của mô hình nguồn-bộ lọc mô tả bộ máy phát âm. Phương thức mã hóa lai ghép thực hiện việc kết hợp giữa phương pháp mã hóa trực tiếp dạng sóng và phương thức mã hóa tham số để có thể kết hợp được ưu điểm của các phương thức mã hóa thành phần nhằm đạt được hiệu quả mã hóa tốt nhất. So sánh chất lượng về khía cạnh chất lượng tiếng nói tái tạo sau mã hóa của ba phương pháp mã hóa trên được minh họa trong hình 3.2.



Hình 3.2 So sánh chất lượng thoại và tốc độ mã hóa của ba phương pháp mã hóa

Trong các phần tiếp theo, chúng ta sẽ tìm hiểu về các phương pháp mã hóa theo cách phân loại này.

3.2. MỘT SỐ PHƯƠNG PHÁP MÃ HÓA DẠNG SÓNG

Như đã đề cập ở trên, mã hóa dạng sóng thực hiện việc khai thác trực tiếp dạng sóng tín hiệu (độ lớn biên độ, sự thay đổi độ lớn biên đồ, đường bao phổ, …) để thực hiện phương pháp mã hóa. Lấy một ví dụ phương pháp mã hóa dự đoán tuyến tính: bộ mã hóa sẽ sử dụng tổ hợp tuyến tính các mẫu tín hiệu quan sát được ở thời điểm trước đó, cố gắng dự đoán giá trị tín hiệu (độ lớn biên độ) ở thời điểm tiếp theo. Các phương pháp mã hóa trực tiếp dạng sóng tương đối đơn giản, dễ triển khai thực hiện. Tuy nhiên các phương pháp mã hóa thuộc nhóm này không hiệu quả trong việc loại bỏ độ dư thừa dữ liệu. Kết quả là, các phương pháp mã hóa này không hiệu quả khi xét về khía cạnh nén dữ liệu.

Các phương pháp mã hóa trực tiếp dạng sóng thường được thực hiện dựa trên tiêu chí tối thiểu hóa sai số giữa tín hiệu mã hóa và dạng sóng tín hiệu gốc. Nói cách khác, lớp phương pháp mã hóa này cố gắng bảo toàn dạng sóng của tín hiệu gốc. Đây cũng chính là lý do mà lớp phương pháp mã hóa này cho tín hiệu tiếng nói có chất lượng cảm nhận cao. Do đó, một số phương pháp mã hóa thuộc lớp mã hóa này thường được sử dụng cho

mã hóa âm thanh, âm nhạc chất lượng cao. Một số phương pháp mã hóa dạng sóng còn có khả năng chịu được nhiễu lớn. Hơn nữa, các phương pháp mã hóa thuộc lớp mã hóa này hoạt động độc lập với cách mà tín hiệu được tạo ra. Chúng không những được sử dụng để mã hóa tiếng nói, âm thanh mà còn được sử dụng để mã hóa các tín hiệu khác nữa.

Một số phương pháp mã hóa thuộc lớp mã hóa này có thể kể đến như: PCM tuyến tính (ITU G.711, 64kbps), ADPCM (CCITT/ITU G.721, 32kbps; CCITT/ITU G.726/727, 16/24/32/40kbps).

3.2.1 PCM

Phương pháp mã hóa PCM (Pulse Code Modulation), còn gọi là phương pháp điều chế xung mã (hay đơn giản là đều xung mã) là phương pháp mã hóa dạng sóng đơn giản nhất. Phương pháp này còn được biết đến với chuẩn G.711 của ITU. Phương pháp này chỉ đơn thuần bao gồm việc lấy mẫu và lượng tử hóa để chuyển thành mã tương ứng.

Một tín hiệu tiếng nói băng hẹp (0.3-3.4kHz) được lấy mẫu với tần số thỏa mãn tiêu chuẩn Nyquist (~8kHz). Sau đó mỗi mẫu được thực hiện việc lượng tử hóa.

Quá trình lượng tử hóa là quá trình không khả nghịch, nghĩa là không tồn tại phép toán ngược để khôi phục một cách chính xác. Như vậy, có thể nói khâu lượng tử hóa là khâu gây tổn thất thông tin trong quá trình mã hóa.

Cách đơn giản nhất là thực hiện việc lượng tử hóa tuyến tính, còn gọi là lượng tử hóa đều. Khi đó khoảng tín hiệu quan tâm (min-max) được chia đều thành 2^b mức, với b là số bít sử dụng để biểu diễn một mẫu. Khi đó, độ phân giải, hay còn gọi là bước lượng tử hóa được xác định bởi:

$$\Delta = \frac{s_{max} - s_{min}}{2^b}$$

Mối quan hệ đầu vào-ra của hàm lượng tử có thể mô tả bởi hàm $y_i = Q(s)$ nếu $s \in [d_i, d_{i+1}]$. Hàm này thường có dạng hình bậc thang như minh họa trong hình 3.x.

Từ hình này, dễ dàng thấy, ngoại trừ có thể hai khoảng ngoài cùng bên trái và bên phải, tất cả các khoảng khác dọc trục tín hiệu vào có độ dài bằng nhau. Quan sát tương tự với trục tín hiệu ra.

Có hai loại đặc tuyến lượng tử hóa tuyến tính: (1) lượng tử hóa bước cân (midtread quantizer), (2) lượng tử hóa bước lệch (midrise quantizer). Lượng tử hóa bước cân thường được sử dụng cho trường hợp số mức lượng tử lẻ và trong các mức lượng tử có mức giá trị bằng 0. Ngược lại, lượng tử hóa bước lệnh sử dụng trong trường hợp số mức lượng tử là số chẵn và trong các mức lượng tử có mức lượng tử có giá trị bằng 0. Sơ đồ minh họa đặc tuyến hàm lượng tử bước cân và bước lệch cho trong hình 3.3.

Sai số của quá trình lượng tử là sự khác biệt giữa mẫu thu được so với giá trị tín hiệu thực ở cùng thời điểm. Gọi $\hat{s}(n)$ là giá trị tín hiệu lượng tử thu được ứng với giá trị tín hiệu vào s(n), khi đó sai số lượng tử:

$$e(n) = s(n) - \hat{s}(n)$$

Dễ dàng có $-\frac{\Delta}{2} \le e(n) \le \frac{\Delta}{2}$. Để đơn giản, giả thiết sai số lượng tử là một quá trình dừng với giá trị trung bình bằng 0, không tương quan với tín hiệu, có phân bố đều. Nghĩa là

$$p_e(e) = \begin{cases} \frac{1}{\Delta} & \text{khi} -\frac{\Delta}{2} \le e \le \frac{\Delta}{2} \\ 0 & \text{truong hop khac} \end{cases}$$

$$\bar{e} = 0,\ \sigma_e^2 = \frac{\Delta^2}{12}$$

σ_e^2 còn gọi là công suất nhiễu lượng tử.

Khi đó, để đánh giá chất lượng mã hóa người ta sử dụng một hệ số tỷ lệ công suất trung bình của tín hiệu trên công suất nhiễu lượng tử chuẩn hóa SNR

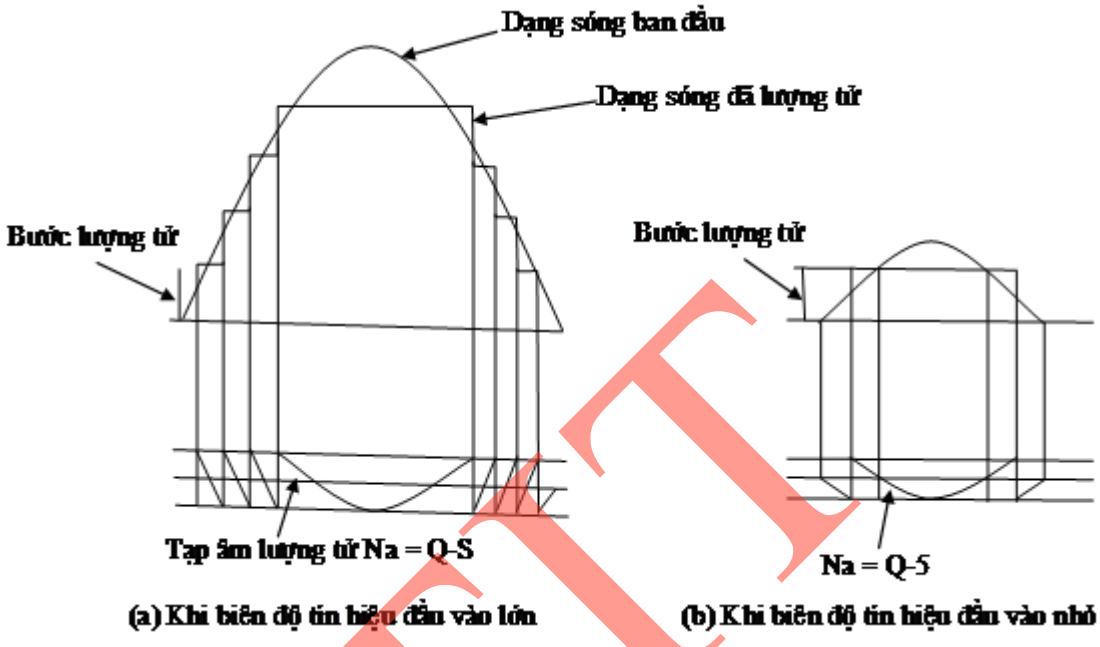
$$SNR = \frac{S}{N_q} = \frac{\sigma_s^2}{\sigma_e^2} = \frac{\sigma_s^2}{\frac{\Delta^2}{12}}$$

Dễ dàng có, $SNR = 6b + 4.77 - 20\log_{10}\left(\frac{s_{max}}{\sigma_s}\right)$

Như vậy, nếu cứ tăng thêm một bít cho biểu diễn mẫu thì SNR sẽ tăng 6dB.

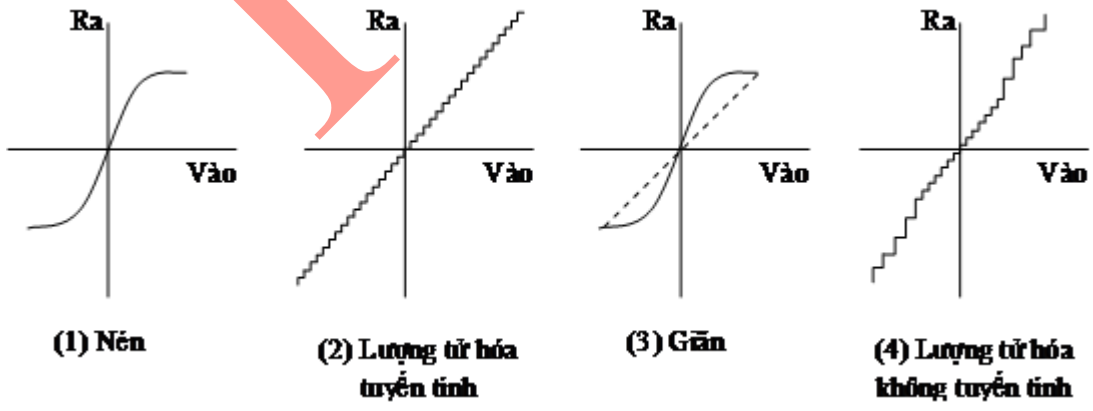
Chúng ta đã đề cập ở trên, phần tín hiệu tiếng nói có biên độ nhỏ (phần các phụ âm vô thanh,..) thường xảy ra thường xuyên hơn so với phần tín hiệu có biên độ lớn. Hơn nữa, đặc điểm cảm nhận của hệ thống thính giác người có đặc tuyến lô-ga-rít trong đó các tín hiệu có bên độ lớn được xử lý với độ phân giải khác với các tín hiệu có biên độ nhỏ. Nói cách khác, cùng múc nhiễu lượng tử, tai người nhạy cảm với nhiễu lượng tử của tín hiệu nhỏ hơn là tín hiệu lớn. Khi bước lượng tử là một hằng số, SNR 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 cần được giảm và đối với các tín hiệu mức cao nó được tăng để ít hoặc nhiều cân bằng SNR với mức tín hiệu đầu vào. Hình 3.3 minh họa sự thay đổi SNR theo mức tín hiệu mã hóa.

Như vậy, cần phải có một phương pháp lượng tử sao có có thể phản ánh được đặc tính cảm nhận này. Phương pháp lượng tử thỏa mãn điều này cần có bước lượng tử thay đổi theo mức tín hiệu. Do đó, phương pháp này được gọi là phương pháp lượng tử hóa phi tuyến.



Hình 3.3 Minh họa sự phụ thay đổi của sai số lượng tử và mức tín hiệu

Về nguyên tắc, phương pháp lượng tử phi tuyến 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 như minh họa trong hình 3.4.



Hình 3.4 Minh họa sự nén và giãn tín hiệu trong lượng tử hóa phi tuyến

Với cách tiếp cận lượng tử hóa phi tuyến, tốc độ mã hóa cũng được giảm xuống một cách đáng kể. Người ta thấy rằng, chỉ cần sử dụng 8 bít mã hóa cho một mẫu là đủ đảm bảo chất lượng thoại và gần như rất khó phân biệt giữa tín hiệu mã hóa và tín hiệu gốc.

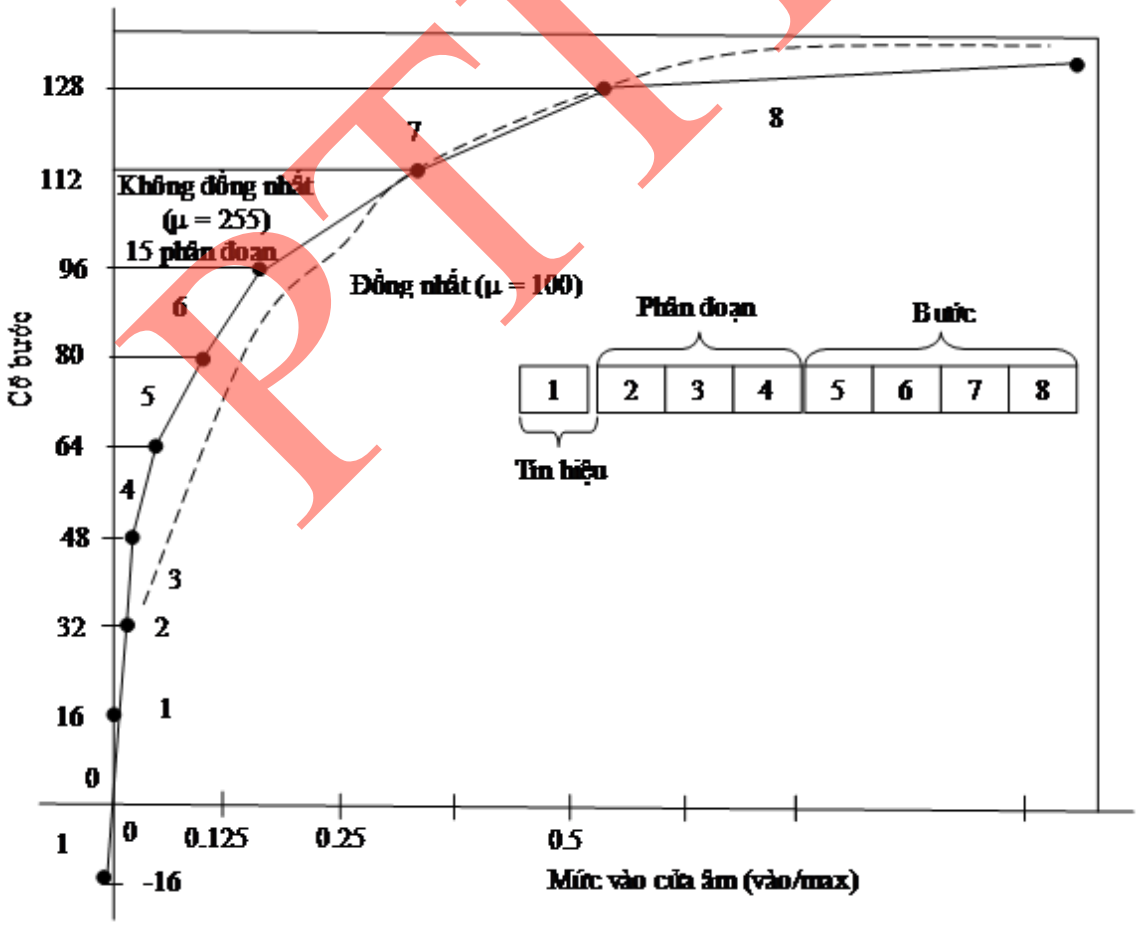
Có hai luật lượng tử hóa phi tuyến phổ biến là luật μ và luật A. Luật μ được dùng phổ biến tài Bắc Mỹ, trong khi luật A được áp dụng ở Châu Âu. Cả hai luật lượng tử này đều có đặc điểm là thực hiện đơn giản, đảm bảo được chất lượng thoại, có độ trễ thấp.

Luật μ, với μ=255, thực hiện nén tín hiệu vào theo công thức:

$$y(n) = s_{max} \frac{\log(1+\mu \frac{s(n)}{s_{max}})}{\log(1+\mu)} sgn(s(n))$$

Trong đó, mỗi mẫu tuyến tính 14 bit gồm cả bít dấu sẽ được ánh xạ thành một từ mã gồm 8 bít bao gồm cả một bit dấu có dạng SABCXYZW. S là bit dấu, ABC là các bít xác định phân đoạn (gồm 15 phân đoạn), XYZW là các bít xác định mức trong phân đoạn.

Hình 3.5 minh họa việc mã hóa tín hiệu theo luật μ



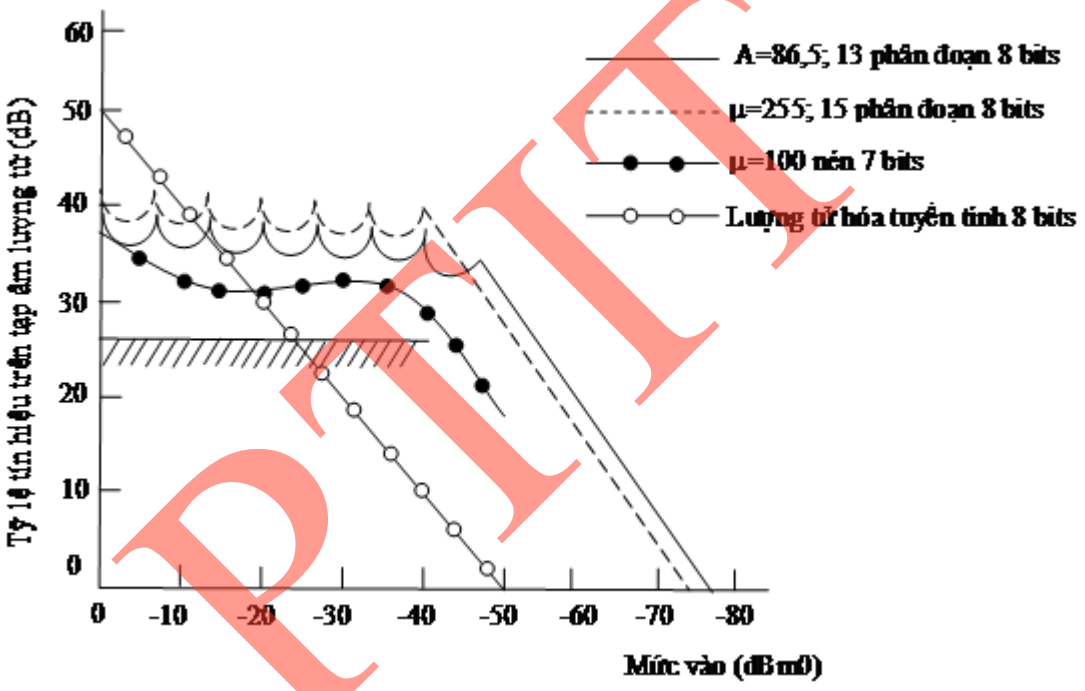
Hình 3.5 Minh họa việc mã hóa PCM với lượng tử phi tuyến theo luật μ

Luật A, với A=87.56 gồm 13 phân đoạn, thực hiện việc nén tín hiệu theo công thức:

$$y(n)=\begin{cases}\dfrac{A\,|s(n)|}{1+\log(A)} & 0\le\dfrac{|s(n)|}{s_{max}}\le\dfrac{1}{A}\\[2ex] s_{max}\dfrac{1+\log(A\dfrac{|s(n)|}{s_{max}})}{1+\log(A)} & \dfrac{1}{A}\le\dfrac{|s(n)|}{s_{max}}\le 1\end{cases}$$

Trong đó, mỗi mẫu tuyến tính 13 bít bao gồm cả bít dấu được ánh xạ thành một từ mã 8 bít có dạng SABCXYZW.

Sự thay đổi SNR của các phương pháp lượng tử được so sánh và minh họa trong hình 3.6.



Hình 3.6 So sánh SNR của các phương pháp lượng tử hóa khác nhau

Việc giải mã cho mã thu được bằng cách tiếp cận lượng tử hóa phi tuyến khá đơn giản. Bằng cách tách ra ba cụm: cụm dấu (bít S), cụm phân đoạn (cụm bít ABC), và cụm mức trong phân đoạn (cụm XYZW), sau đó thực hiện việc ánh xạ ngược lại.

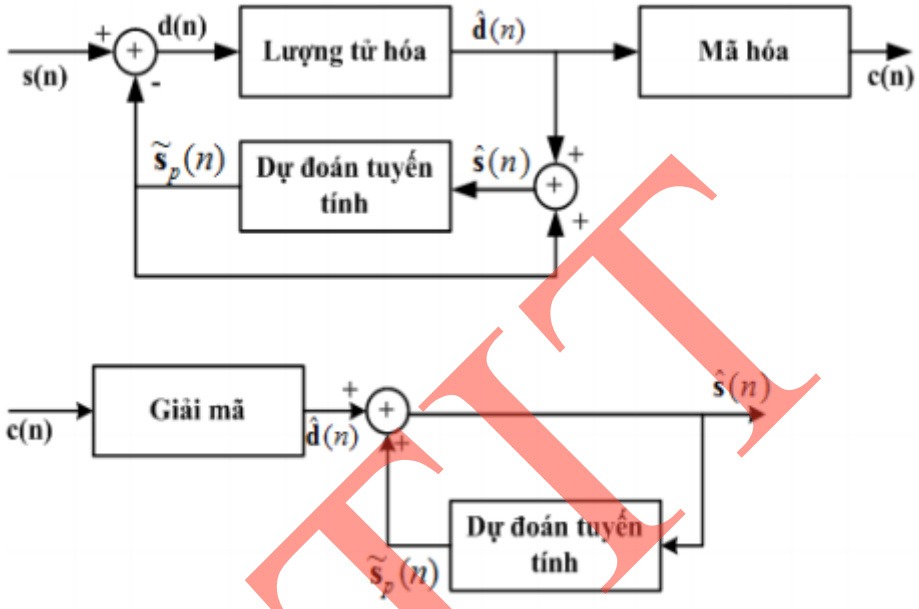
3.2.2 DPCM

Đây là một kỹ thuật cũng được sử dụng phổ biến trong mã hóa thoại nhằm mục tiêu giảm nhỏ tốc độ dữ liệu sau mã hóa.

Ý tưởng của phương pháp mã hóa điều chế xung mã vi sai là tận dụng tự tương quan giữa các mẫu tín hiệu lân cận nhau. Bằng cách sử dụng dự đoán tuyến tính đơn giản về giá trị mẫu tiếp theo từ những mẫu đã biết trước đó, sau đó chỉ thực hiện mã hóa và

truyền đi độ chênh lệch giữa các mẫu cạnh nhau của tín hiệu. Rõ ràng, sự khác biệt giữa các mẫu lân cận nhau phần lớn sẽ nhỏ hơn so với chính giá trị các mẫu. Như vậy, số bít cần thiết để mã hóa sự khác biệt này chắc chắn sẽ thường cần ít hơn mã hóa trực tiếp thông thường.

Sơ đồ của bộ mã hóa và giải mã DPCM cho tín hiệu tiếng nói được cho trong hình 3.7.



Hình 3.7 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), sau đó được lấy mẫu để tạo các giá trị mẫu s(n). Đồng thời, bộ mã hóa thực hiện dự đoán giá trị mẫu theo công thức:

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

Trong đó, a_k là hệ số của các bộ dự đoán.

Độ chênh lệch giữa xung lấy mẫu đầu vào và tín hiệu ra lấy mẫu là:

$$d(n) = s(n) - \tilde{s}_p(n)$$

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. Sai số lượng tử trong trường hợp này được xác định bởi:

$$e(n) = d(n) - \hat{d}(n)$$

Việc sử dụng vòng phản hồi giúp cho bộ lượng tử thỏa mãn biểu thức lỗi lượng tử:

$$e(n) = d(n) - \hat{d}(n) = \hat{s}(n) - s(n)$$

Nói cách khác, vòng hồi tiếp cho phép hạn chế sự khác biệt giữa sai số e(n) và sai số về độ chênh lệch giữa các mẫu. Như vậy, nhiễu lượng tử không phụ thuộc vào việc sử dụng bộ dự đoán, ngoài ra, nhiễu lượng tử không bị tích lũy. Rõ ràng, 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 giảm đi một nửa.

SNR của phương pháp này được xác định theo công thức:

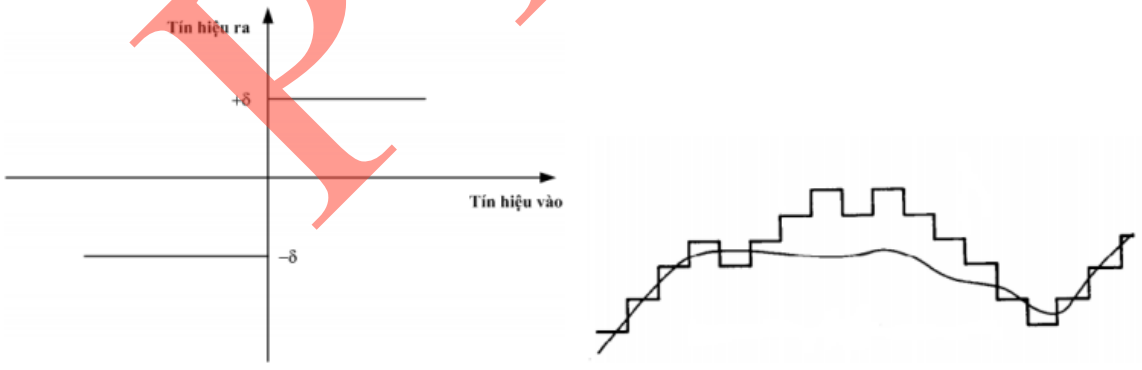
$$SNR_{DPCM} = 6n + 4.77 - 10\log_{10}\left(\frac{\sigma_s^2}{d_{max}^2}\right)$$

Hay có thể viết: $SNR_{DPCM} = SNR_{PCM} + 10\log_{10} G_p$

Trong đó, Gp là độ lợi thu được từ việc sử dụng bộ dự đoán tuyến tính.

3.2.3 DM

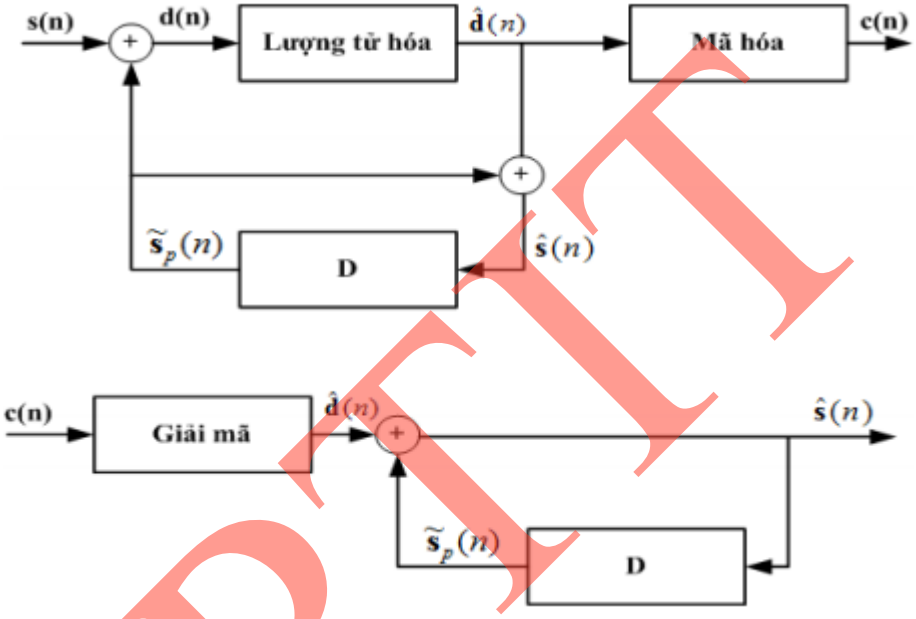
Điều chế DM là một loại điều chế DPCM đơn giản trong đó mỗi từ mã chỉ có một bít nhị phân. Phương pháp này có ưu điểm là việc thực hiện mạch điện rất dễ dàng chỉ cần một bộ so sánh phân ngưỡng như minh họa trong hình 3.8.



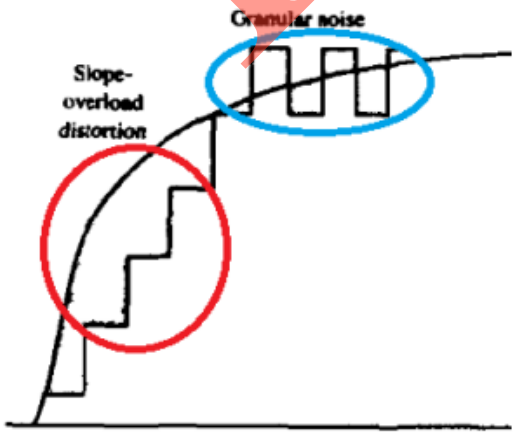
Hình 3.8 Mã hóa DM và sự tương đương phân ngưỡng

Ý tưởng cơ bản của phương pháp là dựa trên nhận xét: các mẫu liền kề nhau có một sự tương quan rất lớn. Khi đó việc dự đoán các mẫu sẽ đơn giản hơn chỉ cần. Sơ đồ tổng quát của bộ mã hóa DM được cho trong hình 3.9. Ở đây, sai số dự đoán là sự khác biệt giữa mẫu hiện tại và giá trị dự đoán xấp xỉ sau cùng nhất từ các mẫu trước đó. Dễ thấy, khi đó sai số lượng tử tỷ lệ với biên độ bước lượng tử.

Mặc dù khá đơn giản, nhưng phương pháp mã hóa DM mắc phải hai lọa méo nghiêm trọng. Thứ nhất là méo quá độ dốc (slope-overload distortion). Nếu bước lượng tử quá nhỏ thì đường xấp xỉ bậc thang, chính là đường kết quả mã hóa, sẽ không bắt kịp sự thay đổi (tăng/giảm) của tín hiệu. Điều này dẫn đến đường mã hóa thu được không phản ánh trung thực tín hiệu gốc. Dạng thứ hai là méo dạng nhiễu (granular noise). Đây là trường hợp xảy ra khi tín hiệu gốc có độ bằng phẳng lớn, nếu bước lượng tử lớn thì tại vùng này đường mã hóa xuất hiện các đỉnh nhấp nhô. Nghĩa là tín hiệu mã hóa bị nhiễu thay vì bằng phẳng như tín hiệu gốc. Hình 3.10 minh họa những sai số vừa đề cập của phương pháp mã hóa DM.



Hình 3.9 Sơ đồ tổng quát mã hóa và giải mã DM



Hình 3.10 Minh họa nhược điểm của mã hóa DM

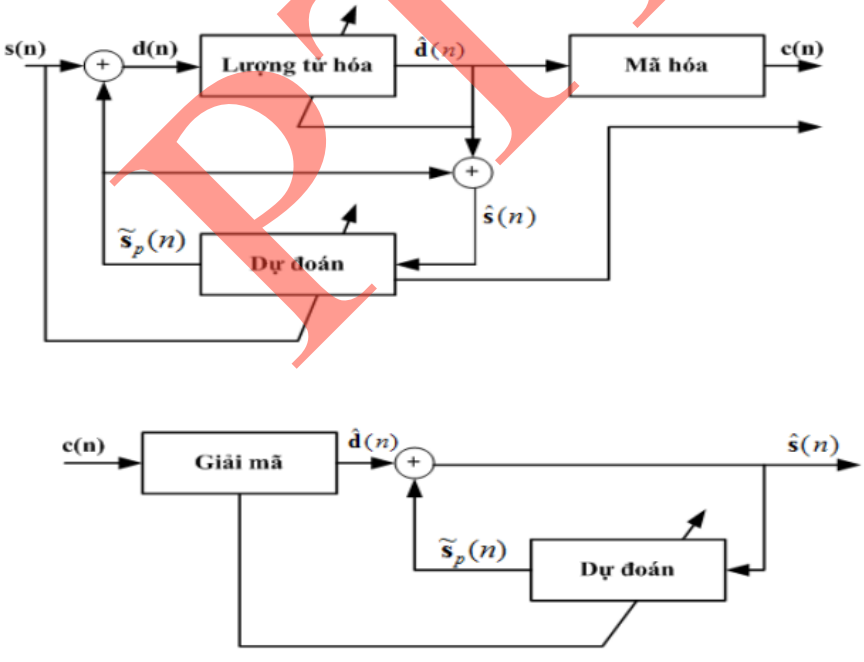
Mặc dù vậy, tốc độ bit của phương pháp mã hóa DM có thể đạt được rất thấp, cỡ bằng tốc độ của tần số lấy mẫu, tức là 8 kbps. Đây là phương pháp duy nhất của phương pháp mã hoá dạng sóng có thể so sánh về tốc độ mã hóa với phương pháp tham số nguồn sẽ tìm hiểu trong phần sau của chương.

3.2.4 APCM

Trong các cách tiếp cận của phương pháp mã hóa PCM, DPCM mặc định với giả thiết tín hiệu mã hóa là một thể hiện của một quá trình dừng. Tuy nhiên, điều này không đúng với tín hiệu tiếng nói. Như vậy, nếu kể đến yếu tố này thì chúng ta có thể thực hiện việc tăng hiệu quả và chất lượng tín hiệu mã hóa bằng cách thay đổi thích nghi theo đặc trưng thống kê của tín hiệu. Vì tín hiệu tiếng nói là một tín hiệu bán dừng (quasi-stationary) nên các thông số thống kê thay đổi chậm theo thời gian.

Nếu thực hiện phép lượng tử hóa đều thì sai số lượng tử sẽ có phương sai thay đổi theo thời gian, cũng tức là công suất nhiễu lượng tử thay đổi theo thời gian. Điều này dẫn đến tỷ số SNR thay đổi theo thời gian. Để giảm nhỏ điều này, tức là làm giảm nhỏ khoảng động của nhiễu lượng tử, chúng ta có thể thực hiện bằng phép lượng tử thích nghi. Ở đây, trong phương pháp APCM, bước lượng tử được thay đổi theo phương sai các mẫu tín hiệu.

Sơ đồ tổng quát của bộ mã hóa APCM như hình 3.11.



Hình 3.11 Sơ đồ tổng quát của phương pháp mã hóa và giải mã APCM

Có hai phương pháp lượng tử thích nghi được sử dụng trong mã hóa APCM: thích nghi forward, và thích nghi backward.

Ở phương pháp thích nghi forward, một bước lượng tử mới được xác định theo công thức:

$$\Delta = \Delta_{ref}\sqrt{\sum_{k=1}^{N} s_n^2(k)}$$

Nói cách khác, bước lượng tử được xác định dựa trên các mẫu s(n) ở thời điểm sau đó. Phương pháp này sẽ cho phép thích ứng nhanh với sự thay đổi hình dạng phổ và cho phép cải thiện SNR khoảng 5dB so với phương pháp PCM luật μ thông thường. Tuy nhiên, phương pháp này cần phải truyền tải thông tin về bước lượng tử. Điều này sẽ làm tăng đáng kể tốc độ bít sau mã hóa trong một số trường hợp.

Ngược lại với phương pháp thích nghi forward, phương pháp lượng tử thích nghi backward ước lượng bước lượng tử từ các mẫu ở thời điểm trước đó theo công thức:

$$\Delta = \Delta_{ref}\sqrt{\sum_{k=n-N}^{n-1} \hat{s}_n^2(k)}$$

Như vậy phương pháp này không cần truyền tải thông tin về bước lượng tử. Tuy nhiên, do bước lượng tử được ước lượng từ các mẫu ở thời điểm trước đó nên phương pháp này thích nghi chậm hơn với sự thay đổi của hình dạng phổ.

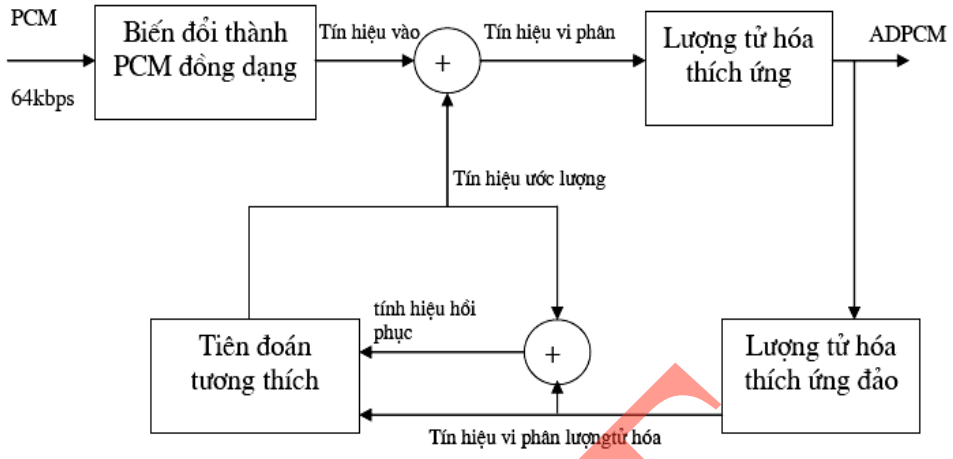
3.2.5 ADPCM

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. Các tốc độ chuẩn của chuẩn mã hóa này là 40, 32, 24, và 16kbps.

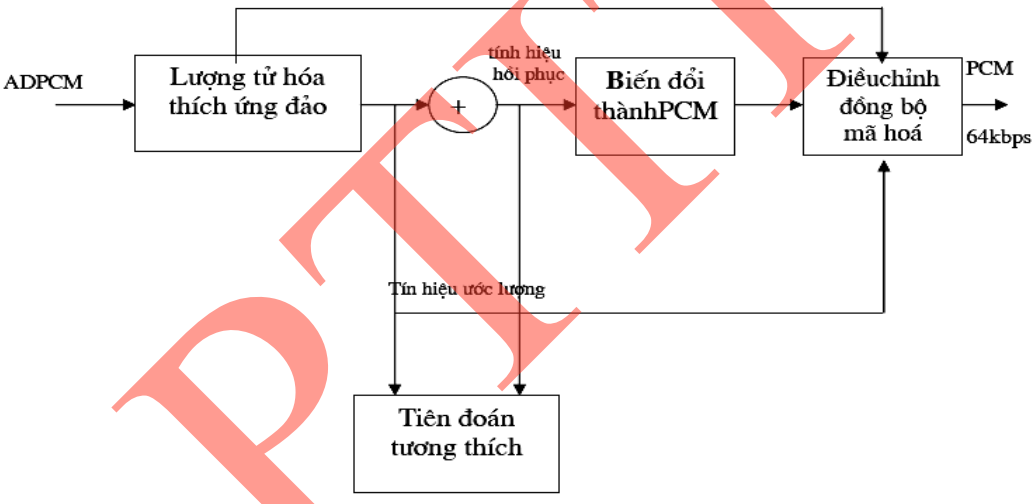
Về cơ bản, cũng như phương pháp mã hóa DPCM, phương pháp mã hóa này thực hiện việc mã hóa sự sai khác giữa tín hiệu và tín hiệu dự đoán. Như vậy, chất lượng mã hóa phụ thuộc khá lớn vào tính chính xác của bộ dự đoán. Mặc khác, nếu sự dự đoán có độ chính xác cao thì sự khác biệt này càng nhỏ, nghĩa là số bít cần thiết để biểu diễn mẫu càng ít. Như vậy, tùy thuộc vào các chỉ tiêu kỹ thuật yêu cầu, cũng như tùy thuộc vào yêu cầu chất lượng tín hiệu ra chúng ta có thể thực hiện việc tùy biến (thay đổi thích nghi) dự đoán hoặc/và bước lượng tử. Khi đó, chúng ta có phương pháp mã hóa điều chế xung mã vi sai thích nghi (ADPCM – Adaptive Differential PCM).

Cách tiếp cận thực hiện phổ biến của 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 3.12 Sơ đồ mã hoá ADPCM



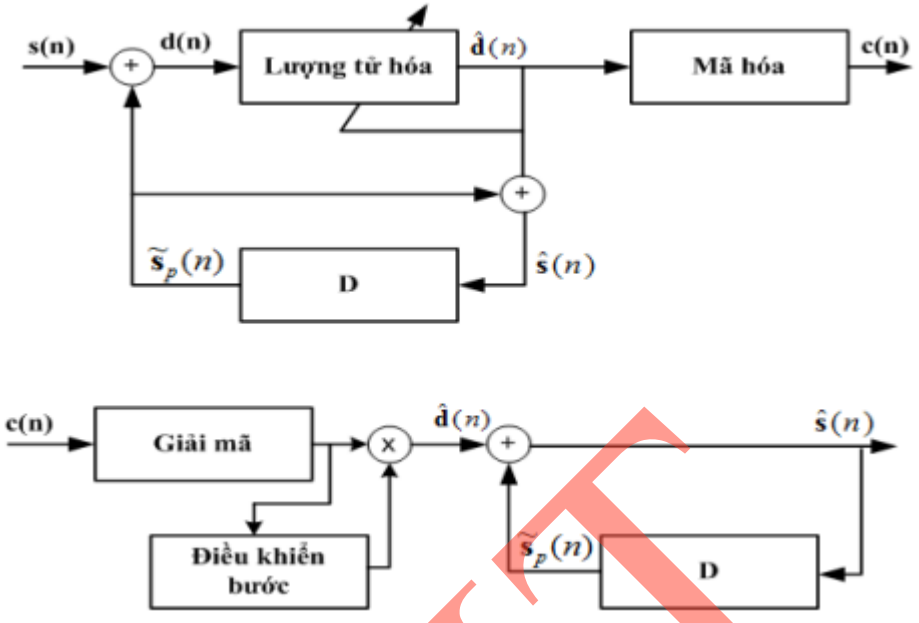
Hình 3.13 Sơ đồ giải mã ADPCM

Ngoài ra, để cải thiện và thích nghi khả năng dự đoán, người ta cũng thường hay sử dụng các sơ đồ dự đoán khác nhau. Chẳng hạn như dự đoán thích nghi Forward, Backward, ….

3.2.6 ADM

Để cải tiến và khắc phục nhược điểm của phương pháp DM, người ta áp dụng phương pháp ADM (điều chế Delta thích nghi). Phương pháp này còn gọi là phương pháp điều chế delta có độ dốc thay đổi liên tục. 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.

Sơ đồ tổng quát của bộ mã hóa ADM cho trong hình 3.14.



Hình 3.14 Sơ đồ mã hoá và giải mã Delta thích nghi

Luật thay đổi bước lượng tử đơn giản nhất được Jayant đề xuất vào năm 1970, trong đó bước lượng tử ở thời điểm n được xác định theo công thức:

$$\Delta_n = \Delta_{n-1} K^{d(n)d(n-1)}$$

Trong đó, K là một hằng số được chọn để giảm méo thỏa mãn >=1

Ngoài ra, Greefkes đưa ra luật thay đổi bước liên tục:

$$\Delta_n = \begin{cases} \alpha\Delta_{n-1} + k_1 & \text{sgn}(d(n)) = \text{sgn}(d(n-1)) = \text{sgn}(d(n-2)) \\ \alpha\Delta_{n-1} + k_2 & \text{con lai} \end{cases}$$

Trong đó, α, k1, k2 là các hằng số 0<α<1, 0<k2<<k1

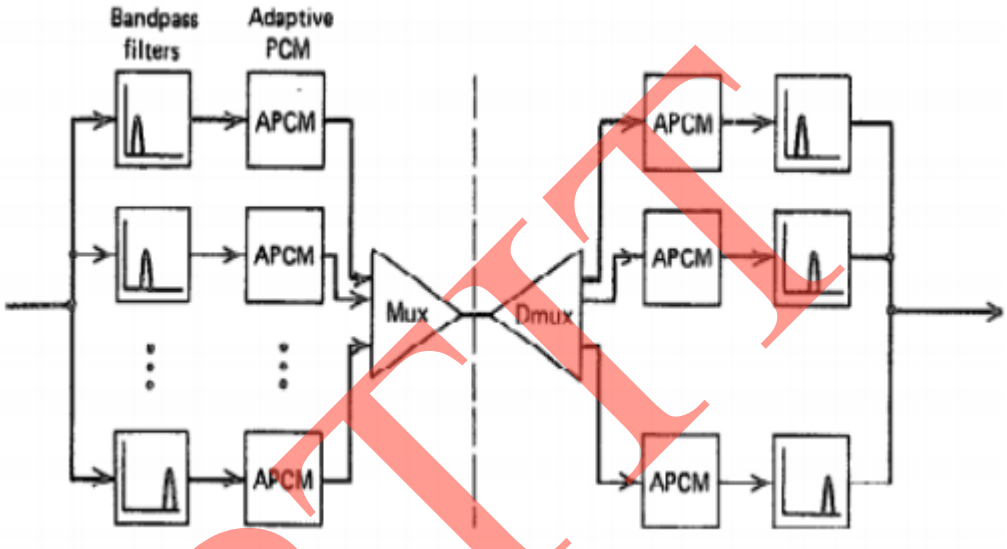
3.2.7 Mã hóa dạng sóng trong miền tần số

Việc mã hóa trực tiếp dạng sóng có thể tiếp cận trong miền tần số. Khi đó, thay vì dựa trên dạng sóng tín hiệu, các phương pháp mã hóa thuộc lớp tiếp cận này dựa vào đặc trưng phổ của tín hiệu. Lợi điểm của phương pháp mã hóa trong miền tần số là có thể khai thác một cách triệt để đặc điểm của tín hiệu trong miền tần số. Thứ nhất, các thành phần tín hiệu trong miền tần số được giải tương quan, tức là gần như không có sự tương hỗ. Hơn nữa, với hiện tượng che lấp tần số đã xem xét trong chương 1, chúng ta có thể thực hiện mã hóa với lượng thông tin ít nhất mà vẫn đảm bảo được chất lượng cảm nhận.

Có rất nhiều cách thực hiện việc mã hóa dạng sóng trong miền tần số, chẳng hạn như phương pháp mã hóa băng con (Subband coding) sử dụng dãy mạch lọc, phương pháp mã hóa chuyển đổi, …

Phương pháp mã hóa băng con tận dụng đặc điểm cảm nhận tiếng nói của tai người: tai người có độ nhạy âm ở các tần số khác nhau là khác nhau, tai người cảm nhận âm chịu tác động bởi hiện trượng che lấ tần số. Từ đó cho phép chỉ mã hóa ở những vùng tần số mà tai người nhạy hơn, hoặc không cần mã hóa các âm bị che lấp.

Sơ đồ tổng quát của một hệ thống mã hóa băng con cho trong hình 3.15.



Hình 3.15 Sơ đồ tổng quát của phương pháp mã hóa băng con

Tín hiệu thoại đầu vào được phân chia thành một số dải băng tần nhỏ hơn gọi là các băng con thông qua các bộ lọc số. Sau đó mỗi một băng con được mã hóa độc lập bằng việc sử dụng các bộ mã hóa dạng sóng như ADPCM.

Phương pháp mã hóa này thực hiện việc kết hợp loại bỏ dư thừa dữ liệu về mặt tần số và thời gian. Do đó, nó có thể đạt được tốc độ mã hóa cỡ 16kbps nhưng chất lượng tín hiệu có thể so sánh với phương pháp mã hóa PCM 64kbps thông thường.

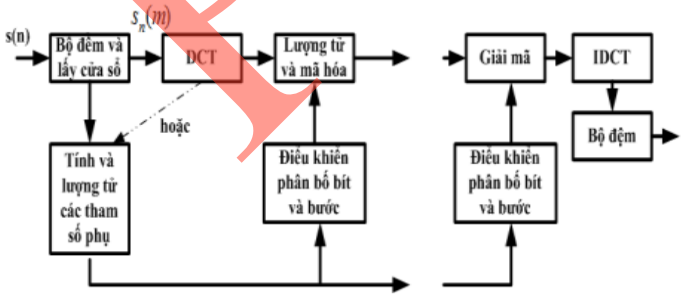
Ngoài phương pháp mã hóa băng con ở trên, người ta có thể thực hiện cải tiến để có được phương thức mã hóa tốt hơn. Cách đơn giản nhất là mã hóa băng con với sự phân bố bít thay đổi thích nghi theo băng tần số tín hiệu (gọi là ASBC – Adaptive Subband coding). Ở đây, các băng con tương ứng với phổ tần số thấp chứa hầu hết năng lượng của tín hiệu thoại sẽ được cấp phát với số bit mã hóa lớn, còn các băng con tương ứng với các phổ tần số cao, chứa ít năng lượng tín hiệu sẽ

được mã hóa với số bit nhỏ hơn. Kết quả là tổng số bit dùng cho mã hóa băng con sẽ ít hơn so với trường hợp mã hóa trên toàn dải phổ của tín hiệu. Tại phía thu, các tín hiệu băng con được giải mã và kết hợp lại để khôi phục lại tín hiệu thoại ban đầu (G. 722 1988).

Một ưu điểm khác của mã hóa băng con là nhiễu trong mỗi băng con chỉ phụ thuộc vào mã hóa sử dụng trong băng con đó. Bởi vậy chúng ta có thể cấp phát nhiều bit hơn cho các băng con quan trọng sao cho nhiễu trong những vùng tần số này là nhỏ, trong khi đó ở các băng con khác, chúng ta có thể cho phép có nhiễu mã hóa cao vì nhiễu ở những tần số này có tầm quan trọng thấp hơn. Các mô hình cấp phát bit thích ứng có thể được sử dụng để khai thác thêm ý tưởng này. Các bộ mã hóa băng con cho chất lượng thoại tốt trong phạm vi tốc độ từ 16 – 32 kbps.

Tuy nhiên, do phải cần đến bộ lọc, một khâu mà việc thực thi không hề đơn giản, để tách tín hiệu thoại trong các băng con nên mã hóa băng con phức tạp hơn bộ mã hóa DPCM thông thường và có thêm độ trễ mã hóa. Tuy nhiên, độ phức tạp và độ trễ là tương đối thấp so với các bộ mã hóa lai ghép mà chúng ta sẽ tìm hiểu trong phần sau của bài giảng.

Trong thực tế, sơ đồ mã hóa băng con được biết đến khá nhiều đó là sơ đồ MUSICAM được phát triển bởi hảng Philips. Trong sơ đồ này bộ mã hóa sử dụng một dãy gồm 32 bộ lọc. Sơ đồ này đã trở thành tiêu chuẩn mã hóa âm thanh ISO/IEC, một cơ sở của mã hóa MPEG-1,2 Layer I,II với độ trễ thấp, cỡ khoảng 10.66ms.



Hình 3.16 Sơ đồ mã hóa MUSICAM

Khác với phương pháp mã hóa băng con, phương pháp mã hóa chuyển đổi và chuyển đổi thích nghi xử lý và mã hóa trực tiếp mẫu ở miền tần số. Các mẫu tín hiệu được phân chia thành các nhóm gồm N mẫu. Các nhóm mẫu này được chuyển đổi sang miền tần số bằng các phép biến đổi thông thường như DFT, FFT, ..Kết quả biến đổi là các hệ số sẽ được lựa chọn, mã hóa để truyền đi. Dễ dàng thực hiện mã hóa thích nghi với phương

pháp mã hóa này. Chúng ta chỉ cần thay đổi số bít cho mã hóa: những thành phần phổ quan trọng sẽ dùng nhiều bít, những thành phần phổ ít quan trọng sẽ dùng ít bít.

Sơ đồ tổng quát của bộ mã hóa chuyển đổi thích nghi được minh họa trong hình 3.x.

Phương pháp mã hóa chuyển đổi thích nghi (ATC) cho phép kết quả mã hóa với tốc độ rất thấp, cỡ 9.6kbps với chất lượng khá tốt.

3.3. MỘT SỐ PHƯƠNG PHÁP MÃ HÓA THAM SỐ

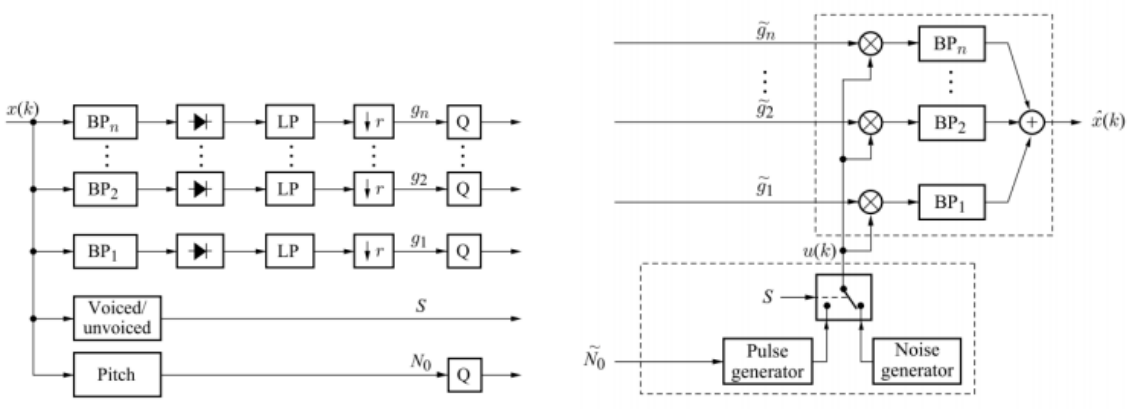
Mã hóa tham số còn gọi là mã hóa phân tích-tổng hợp. Ý tưởng của phương pháp mã hóa này bắt nguồn từ mô hình của bộ máy phát âm.

Chúng ta đã biết, việc tạo ra tín hiệu tiếng nói có thể mô hình bằng sơ đồ nguồn-bộ lọc. Nguồn đóng vai trò tín hiệu kích thích là dao động của dây thanh (dao động bán tuần hoàn với âm hữu thanh, không xác định – giống nhiễu – với âm vô thanh). Âm của tín hiệu được quyết định bởi sự co thắt, hay một cách cụ thể là đặc điểm cộng hưởng của bộ lọc tuyến âm. Như vậy, nếu chúng ta biết được một âm là vô thanh hay hữu thanh và bộ tham số điều khiển sự cộng hưởng của tuyến âm (phân tích), chúng ta hoàn toàn có thể tái tạo lại âm đó (tổng hợp). Và như vậy, thay vì phải truyền đi toàn bộ tín hiệu hoặc đặc trưng dạng sóng của tín hiệu, chúng ta chỉ cần truyền đi thông tin về các tham số của âm. Các bộ mã hóa tham số còn được gọi là các bộ mã hóa Vocoder.

Ưu điểm của loại mã hóa này là nó rất có hiệu quả đối với âm tiếng nói, dễ hiểu, trong khi nó lại có nhược điểm là phức tạp hơn nhiều so với phương pháp mã hóa dạng sóng. Mã hóa tham số có thể đạt được tốc độ bit rất thấp (xuống đến 2.4 Kbps) trong khi vẫn đảm bảo là tiếng nói được tái tạo lại là hoàn toàn dễ hiểu. Tuy nhiên, tính tự nhiên của tiếng nói được tái tạo thì khác xa với tín hiệu tiếng nói con người.

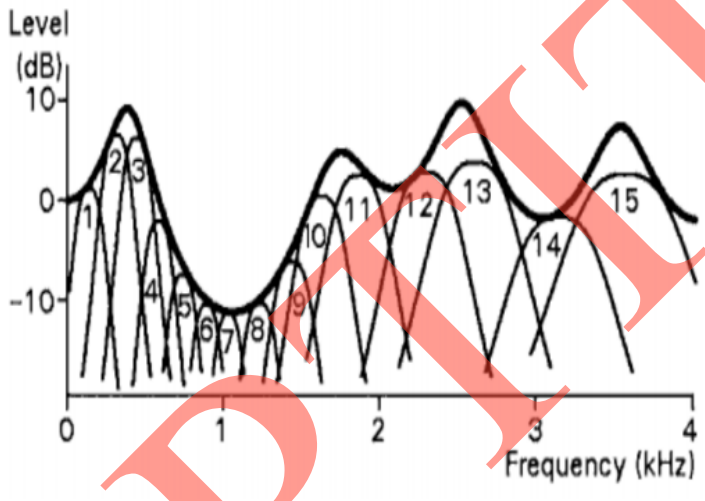
Có rất nhiều cách tiếp cận thực hiện phương pháp mã hóa tham số.

Sơ đồ tổng quát của một hệ thống mã hóa tham số có sử dụng dãy mạch lọc được minh họa trong hình 3.17. Tín hiệu vào được đưa vào đồng thời 3 phân tích để trích chọn đặc trưng. Thứ nhất là phát hiện xem phân đoạn tín hiệu cần mã hóa là của âm vô thanh hay hữu thanh (S), với âm hữu thanh thì tiếp tục xác định tần số cơ bản (pitch) (N0). Đồng thời tín hiệu được phân tách thành những băng tần nhỏ. Mỗi băng tần tín hiệu ứng với một vùng tần số quan tâm. Và mỗi tần số quan tâm chúng ta được bộ đặc trưng g. Toàn bộ các tham số trích chọn được sẽ được mã hóa và gửi đến phía thu để thực hiện tái tạo tín hiệu tiếng nói.



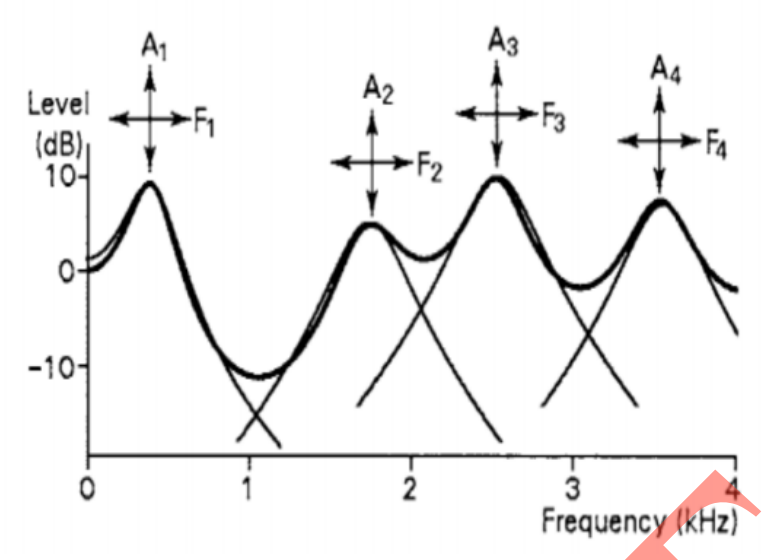
Hình 3.17 Sơ đồ tổng quát một phương pháp mã hóa tham số phân kênh

Hình 3.18 minh họa các tham số g là những đặc tuyến phổ mong muốn



Hình 3.18 Các đặc trưng phổ trong mã hóa tham số phân kênh

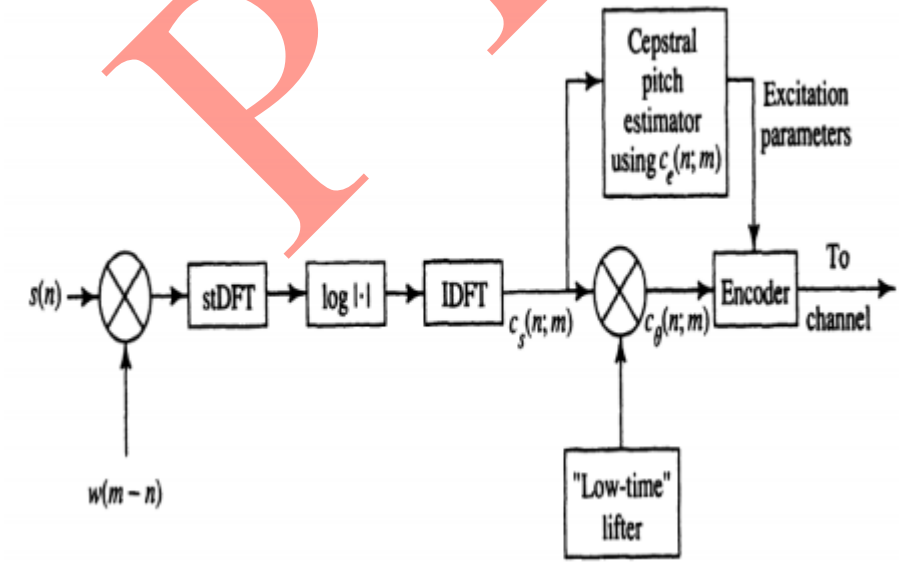
Hoặc các đặc trưng là các tần số formant như minh họa trong hình 3.9.



Hình 3.19 Các đặc trưng formant trong mã hóa tham số phân kênh

Một phương pháp tiếp cận khác cũng khá phổ biến trong các chuẩn mã hóa tiếng nói được sử dụng gần đây là phương pháp mã hóa dựa trên phân tích cepstral. Sơ đồ tổng quát của hệ thống mã hóa được minh họa trong hình 3.20.

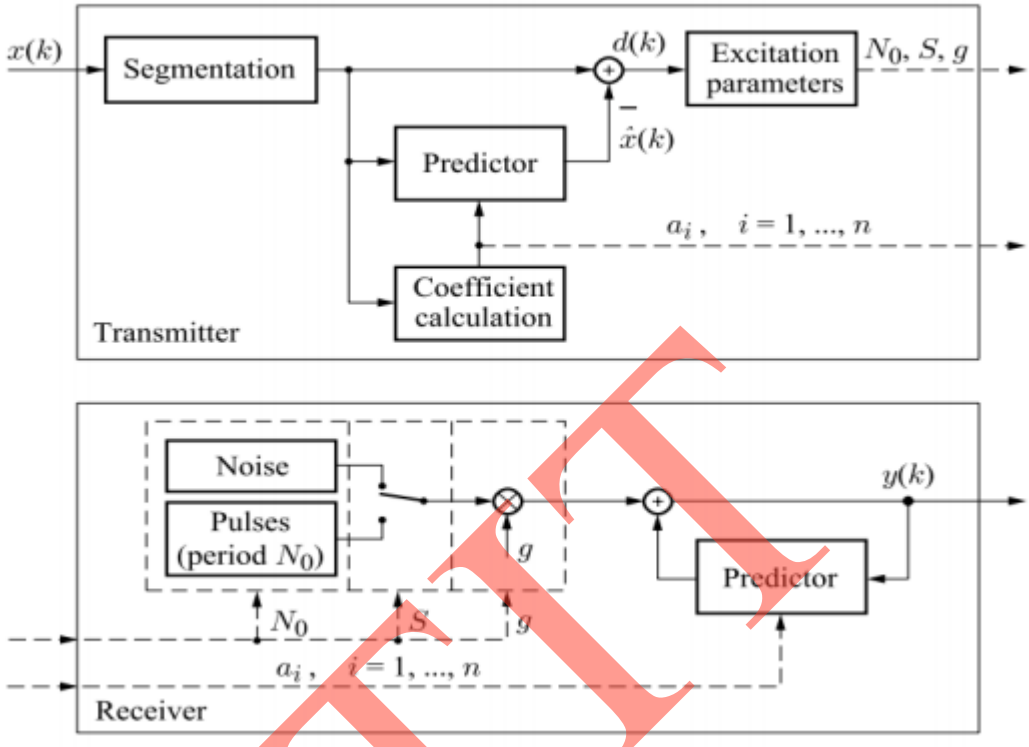
Trong phương pháp này, dựa trên sự khác nhau cơ bản giữa sự thay đổi của biên phổ (đường bao phổ) và xung kích thích (thành phần phổ nhỏ) các đặc tính đường bao phổ và thành phần kích thích được phân tích (phân tách) trích chọn bằng phép tích cepstral mà chúng ta đã xem xét trong chương 2.



Hình 3.20 Sơ đồ mã hóa phân tích cepstral

Một phương pháp tiếp cận khác cũng khá phổ biến đó là mã hóa tham số dựa trên phân tích LPC. Cũng tương tự với đa số các phương pháp mã hóa tham số, phương pháp

này cũng cố gắng mô phỏng quá trình tạo tiếng nói của hệ thống phát âm. Sơ đồ tổng quát của phương pháp mã hóa này được minh họa trong hình 3.x.



Hình 3.21 Minh họa mã hóa tham số LPC

Các thông tin mã hóa của bộ mã hóa tham số LPC là: thông tin về loại âm (hữu thanh/vô thanh) của phân đoạn tín hiệu; độ lớn của tín hiệu; tập các hệ số bộ lọc LPC; chu kỳ pitch (tần số cơ bản) của tín hiệu.

Có rất nhiều phiên bản mã hóa tham số dựa trên LPC, chẳng hạn như LPC-10, CELP, MELP, …

Với phương pháp mã hóa tham số LPC, chúng ta có thể đạt được tốc độ mã hóa thoại bằng 2.4kbps.

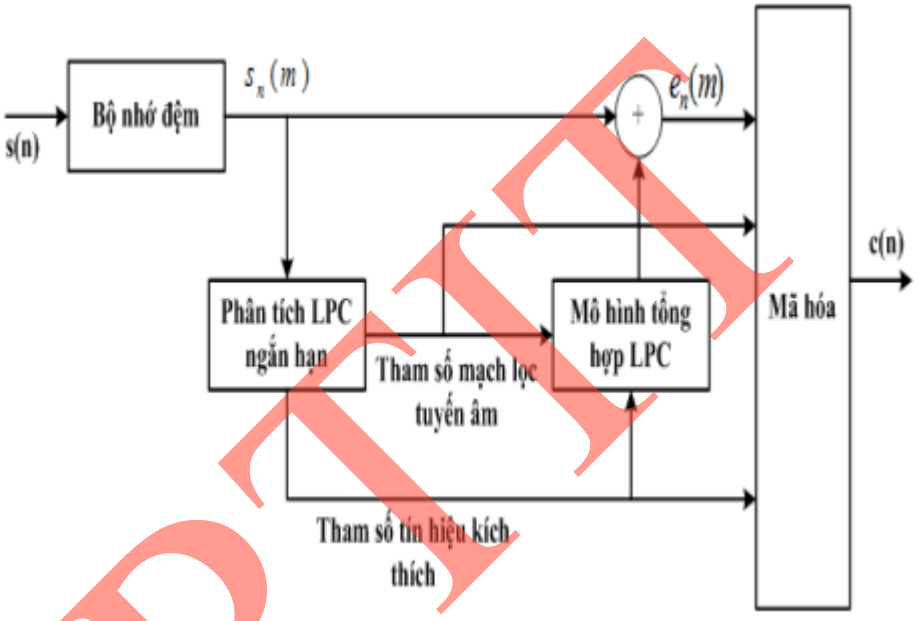
3.4. PHƯƠNG PHÁP MÃ HÓA LAI GHÉP

Mã hóa lai cố gắng lấp khoảng cách ranh giới giữa mã hóa dạng sóng và mã hóa nguồn: đạt được tốc độ mã hóa thấp; tăng được chất lượng tín hiệu tiếng nói mã hóa. Các phương pháp mã hóa thuộc nhóm này thường được áp dụng trong các hệ thống thông tin di động.

Sự kết hợp lai ghép có thể được thực hiện trong miền tần số, hoặc miền thời gian.

Mặc dù có nhiều cách tiếp cận thực hiện mã hóa lai, nhưng thành công và thường được sử dụng nhiều nhất là các bộ mã hóa kết hợp trong miền thời gian "thực hiện các phép phân tích thông qua việc tổng hợp" - AbS (Analysic - by - Synthesis). Những bộ mã hóa này sử dụng mô hình bộ lọc dự đoán tuyến tính cho cơ quan phát âm như được trong các bộ mã thoại LPC. Tuy nhiên, để thay thế cho việc ứng dụng mô hình 2 trạng thái đơn giản - hữu thanh/vô thanh, mô hình này cố gắng giảm tối đa sai lệch giữa dạng sóng tín hiệu đầu vào và dạng sóng tín hiệu được xây dựng lại bằng việc tìm kiếm tín hiệu kích thích lý tưởng. Nói cách khác, phương pháp mã hóa này không sử dụng ước lượng đơn giản là âm hữu thanh hay vô thanh.

Sơ đồ tổng quát bộ mã hóa lai ghép RELP được minh họa trong hình 3.22.



Hình 3.22 Minh họa phương pháp mã hóa lai ghép RELP

Trước tiên, bộ mã hóa thực hiện phân tích tín hiệu thoại đầu vào thành các khung ngắn có độ dài khoảng 10-30 ms. Các tham số của một khung sẽ xác định một bộ lọc tổng hợp tương ứng với khung đó và tín hiệu kích thích tương ứng cho mỗi bộ lọc này sẽ được xác định thông qua một vòng lặp. Tín hiệu kích thích phải đảm bảo rằng sai lệch giữa tín hiệu đầu vào và tín hiệu được tái tạo lại là nhỏ nhất. Cuối cùng bộ mã hóa sẽ truyền đi những thông tin liên quan đến các bộ lọc bao gồm các tham số và tín hiệu kích thích tương ứng với mỗi bộ lọc gửi cho bộ giải mã. Ở bộ giải mã, tín hiệu kích thích sẽ được đưa qua bộ lọc tổng hợp để xây dựng lại tín hiệu thoại ban đầu. Bộ lọc tổng hợp thường là một bộ lọc tuyến tính, ngắn hạn nhưng nó cũng có thể bao gồm một bộ lọc độ cao âm thanh (pitch filter) liên quan đến mô hình tuần hoàn dài hạn của tín hiệu thoại. Phương pháp này cung cấp tín hiệu thoại có chất lượng cao tại tốc độ bit thấp. Tuy nhiên độ phức tạp của phương pháp này là khá lớn bởi vì tất cả các tín hiệu kích thích có thể có đều phải được đưa qua bộ lọc tổng hợp để tìm ra tín hiệu kích thích thích hợp nhất.

3.5. MỘT SỐ PHƯƠNG PHÁP MÃ HÓA TIẾNG NÓI TỐC ĐỘ THẤP

Để thực hiện phương pháp mã hóa tiếng nói tốc độ thấp, xu hướng tiếp cận của các phương pháp là sự kết hợp giữa các phương pháp mã hóa tham số cùng với một số phương pháp khác.

Nhóm đầu tiên có thể kể đến là một số phương pháp mã hóa lai sử dụng: bộ mã hóa kích thích đa xung - MPE (Multi – Pulse – Excited); bộ mã hóa kích thích xung đều – RPE (Regular – Pulse – Excited); bộ mã hóa dự đoán tuyến tính kích thích mã - CELP (Code - Excited – Linear – Predictive).

Trong phương pháp MPE tín hiệu kích thích u(n) được xác định bằng một số lượng cố định các xung tương ứng đối với mỗi khung tín hiệu. Do vậy thông tin cần truyền đi sẽ bao gồm thông tin về độ lớn và về vị trí của các xung này. Phương pháp này cung cấp chất lượng thoại khá tốt tại tốc độ bit khoảng 10 Kbits/s.

Phương pháp RPE tương tự như MPE tuy nhiên các xung kích thích sử dụng trong phương pháp này được sắp xếp cách đều nhau một khoảng cố định do đó phía phát chỉ cần truyền đi thông tin về độ lớn của các xung và vị trí của xung đầu tiên. Như vậy ở cùng một tốc độ bit cho trước thì RPE sẽ có thể sử dụng nhiều xung kích thích hơn so với MPE. Điều này cho phép mã hóa RPE cung cấp chất lượng thoại tốt hơn so với phương pháp MPE song nó lại có độ phức tạp lớn hơn. Mặc dù hai phương pháp MPE và RPE có thể cung cấp chất lượng thoại tốt tại tốc độ bit vào khoảng 10 Kbits/s hoặc cao hơn tuy nhiên chúng lại không thích hợp cho việc sử dụng ở tốc độ bit giảm thấp hơn nữa.

Phương pháp CELP khác với hai phương pháp MPE và RPE ở chỗ tín hiệu kích thích được lượng tử hóa vector một cách hiệu quả. Các tín hiệu này được xác định bởi một mã nằm trong bộ mã lượng tử vector và một hệ số khuếch đại để điều khiển công suất của tín hiệu. Bộ mã lượng tử vector thường được mã hóa bằng 10 bit và hệ số khuếch đại được mã hóa bởi 5 bit tín hiệu do đó sẽ làm giảm đáng kể tốc độ bit dùng để truyền thông tin đi. Tuy nhiên việc phải đưa tất cả các chuỗi tín hiệu kích thích (tương ứng với số lượng tất cả các mã trong bộ mã lượng tử) qua bộ lọc tổng hợp sẽ khiến cho mã hóa CELP có độ phức tạp rất cao. Những nghiên cứu gần đây nhằm cải tiến cấu trúc của bộ mã hóa lượng tử và những tiến bộ trong việc chế tạo các chip vi xử lý đã giúp cho việc thực hiện mã hóa CELP trong thời gian thực. Phương pháp này cung cấp tín hiệu thoại chất lượng tốt ở tốc độ 4,8 Kbps và 16 Kbps. Các nghiên cứu trong thời gian gần đây nhằm cải tiến phương pháp mã hóa CELP đã cho phép cung cấp tín hiệu thoại tại tốc độ 2,4 Kbps.

Ngoài ra, dựa trên đặc trưng của tín hiệu tiếng nói là tổng hòa của hai thành phần với sự thay đổi chậm theo thời gian, người ta còn sử dụng phương pháp mã hóa dựa trên phân tích các sóng nhỏ (wavelets)

3.6. ĐÁNH GIÁ CHẤT LƯỢNG MÃ HÓA TIẾNG NÓI

Một đánh giá đơn giản và hay sử dụng là cách đánh giá định lượng thông qua tỷ số SNR: tỷ số công suất trung bình tín hiệu trên nhiễu. Như đã đề cập trong phần mã hóa PCM, SNR được xác định theo công thức tổng quát:

$$SNR = \frac{E\{s^2(n)\}}{E\{e^2(n)\}}$$

Trong đó $E\{\}$ là giá trị trung bình thống kê.

SNR là một thông số mang tính chất kỹ thuật mang tính chất khách quan mà gần như không có một mối quan hệ chặt chẽ đến sự cảm nhận của tai người. Do đó, ngoài đánh giá khách quan bằng tỷ số SNR, người ta còn đánh giá chất lượng mã hóa thông qua một thông số mang tính chất chủ quan là thang đo điểm ý kiến (còn được biết đến là thang đo độ hài lòng – Mean Opinion Score). Đây là thang đo đánh giá tính chủ quan cảm nhận của người nghe sau khi được hỏi ý kiến về chất lượng tiếng nói thu được của bộ mã hóa và giải mã. Thông thường thang này gồm có 5 cấp độ: 1- Tồi; 2-Kém; 3-Chấp nhận được; 4-Tốt; 5-Rất tốt. Mặc dù nó phản ánh được đặc điểm nghe của con người, nhưng đây là một tham số mang tính định tính, khó có thể có được công thức tính trực tiếp. Như vậy, nó không thể được dùng như là một điều kiện trong bài toán thiết kế xây dựng bộ mã tối ưu.

Một đánh giá nữa là tốc độ mã hóa: là số bít trung bình cần phải truyền trong một đơn vị thời gian.

Trong các ứng dụng mã hóa tiếng nói của các hệ thống thông tin, một yêu cầu quan trọng không kém đó là khả năng đáp ứng thời gian thực, hay độ trễ của phép mã hóa. Trong mã hóa tiếng nói của hệ thống thoại tương tác thời gian thực, độ trễ >150ms là không thể chấp nhận được.

3.7. CÂU HỎI VÀ BÀI TẬP CUỐI CHƯƠNG

1. Mục đích của việc mã hóa tín hiệu tiếng nói?
2. Có những lớp mã hóa tiếng nói nào?

3. Các phương pháp mã hóa dạng sóng tín hiệu tiếng nói: ý tưởng, nguyên lý thực hiện, ưu/nhược điểm?
4. Các phương pháp mã hóa tham số: ý tưởng, nguyên lý thực hiện, ưu/nhược điểm?
5. Các phương pháp mã hóa lai ghép: ý tưởng, nguyên lý thực hiện, ưu/nhược điểm?
6. (Matlab) Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các ngôn ngữ lập trình khác) thực hiện các công việc sau:
    i. Ghi âm một đoạn tín hiệu tiếng nói bất kỳ, lưu ở định dạng *.wav
    ii. Sử dụng hàm thư viện của Matlab hoặc công cụ thích hợp:
        1. Kiểm nghiệm một số phương pháp mã hóa dạng sóng cơ bản (PCM, DPCM, …), đánh giá SNR, chất lượng âm thanh cảm thụ, dung lượng file dữ liệu sau mã hóa
        2. Kiểm nghiệm một số phương pháp mã hóa tham số cơ bản (LPC, CELP, …), đánh giá SNR, chất lượng âm thanh cảm thụ, dung lượng file dữ liệu sau mã hóa

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, đến 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 kết 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.

4.2.1.1 Phương pháp tổng hợp trực tiếp đơn giản

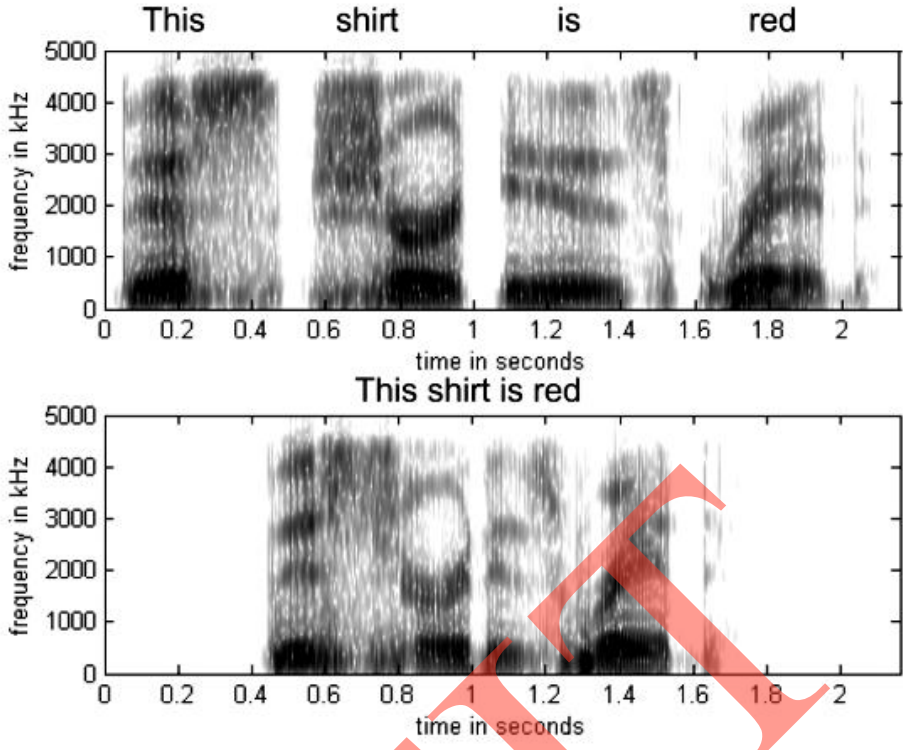
Phương pháp đơn giản nhất để tạo các bản tin tiếng nói là ghi và lưu trữ tiếng nói của con người theo các đơn vị từ riêng lẻ khác nhau và sau đó chọn phát lại các từ theo thứ tự mong muốn nào đó. Phương pháp này được đưa vào sử dụng trong hệ thống điện thoại của nước Anh từ những năm 36 của thế kỷ trước, 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 thiết 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

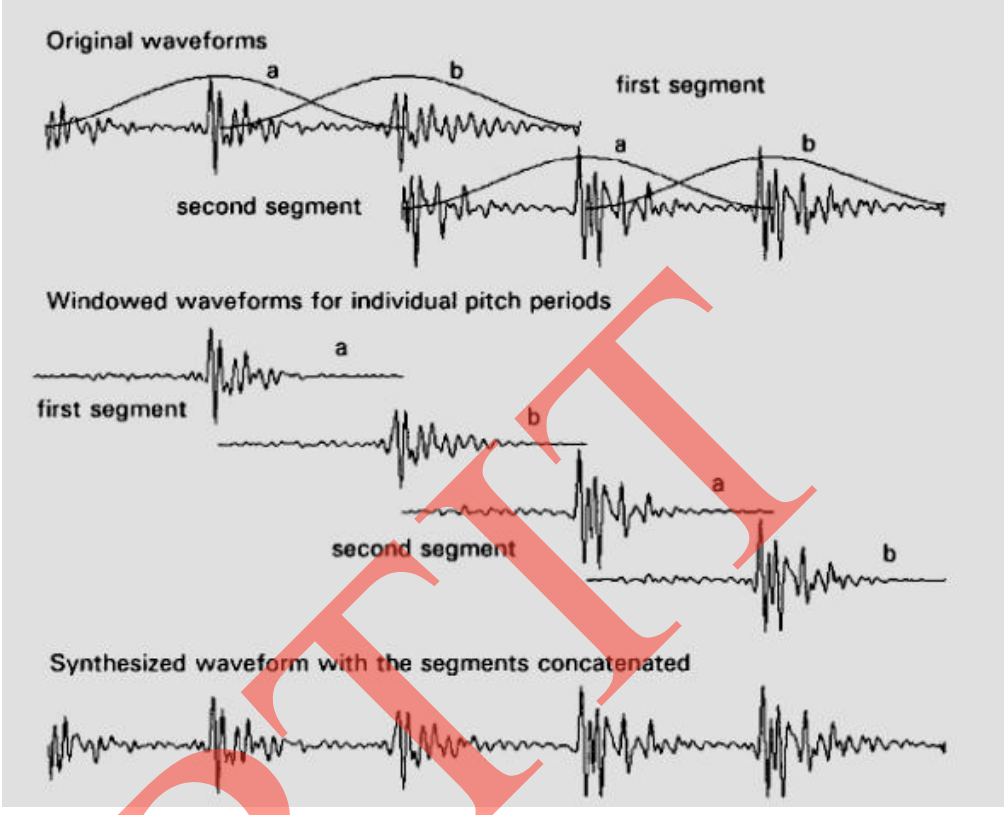
4.2.1.2 Phương pháp tổng hợp trực tiếp từ các phân đoạn dạng sóng

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. Xét bài toán nối hai phân đoạn của dạng sóng tín hiệu của nguyên âm, 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 (dao động thanh môn) 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 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

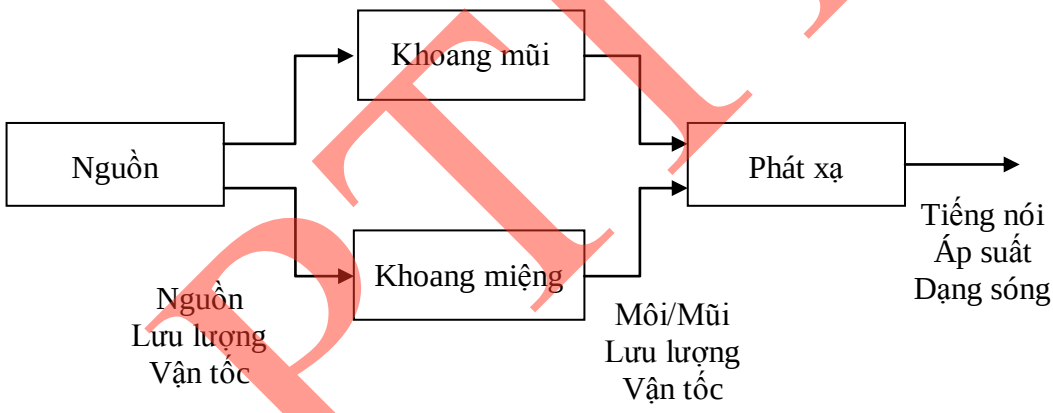
Từ minh họa, ta thấy rằng, bằng cách nối dãy các phân đoạn cửa sổ tín hiệu sóng theo các vị trí tương đối cho trước theo các điểm dấu pitch đã phân tích, ta có thể tái tạo một cách khá chính xác bản tin theo ý mong muốn. Ngoài ra, bằng cách thay đổi các vị trí tương đối và số lượng các điểm dấu pitch, 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 tuyến âm thanh được sử dụng một cách đặt biệt 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 môn (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:

$$H(z)=\frac{1}{1-a_1z^{-1}-a_2z^{-2}}$$

Trong đó hàm truyền đạt có thể phân tích thành:

$$H(z)=\frac{1}{\left(1-p_1z^{-1}\right)\left(1-p_2z^{-1}\right)}$$

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, 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 góc pha θ và bán kính r và chú ý đến nhận xét cặp điểm cực là liên hợp phức ta có thể viết hàm truyền đạt trong công thức (4.1) như sau:

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

Từ đây 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:

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

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:

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

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 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:

$$H(z) = H_1(z)H_2(z)H_3(z)H_4(z)$$

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:

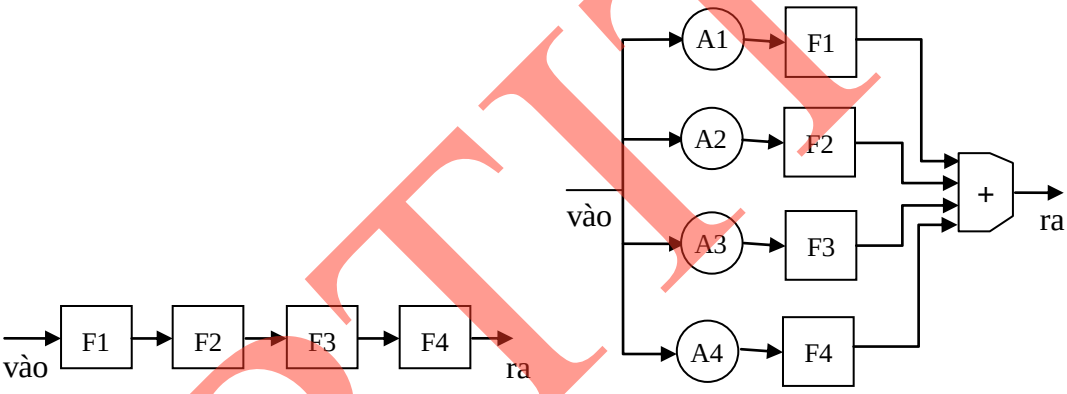
$$y(n) = x(n) + a_1 y(n-1) + a_2 y(n-2) + ... + a_8 y(n-8)$$

Một cách tương tự, ta có thể xây dựng hệ thống mô phỏng khoang mũi. Các biểu thức **Error! Reference source not found.** và **Error! Reference source not found.** 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.

$$y(n) = y_1(n) + y_2(n) + ...$$

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



(a) Cấu hình tổng quát của phương pháp tổng hợp nối tiếp

(b) Cấu hình tổng quát của phương pháp tổng hợp song song

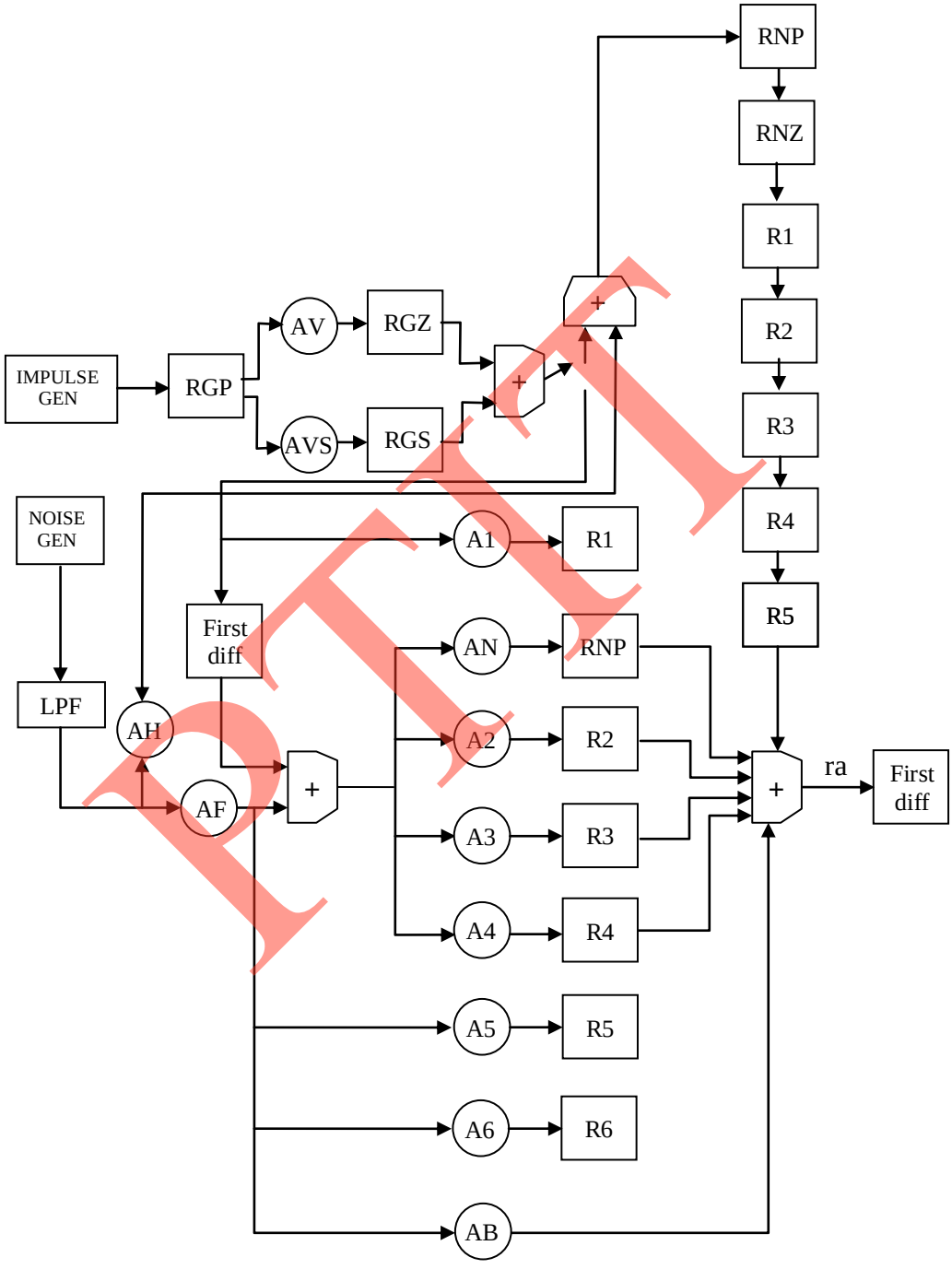
Hình 4.4 Các cấu hình của phương pháp tổng hợp nhiều formant

Phương pháp tổng hợp theo sơ đồ nối tiếp có ưu điểm là với một tập các giá trị formant cho trước, 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 ta xác định một cách chính xác tần số của các formant.

Mặc dù là phương pháp tổng hợp đơn giản và mang lại tín hiệu âm thanh rõ nhưng phương pháp tổng hợp theo formant khó đạt được tính tự nhiên của tín hiệu tiếng nói. Nguyên nhân là do mô hình nguồn và mô hình chuyển đổi 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.



Hình 4.5 Sơ đồ khối bộ tổng hợp Klatt

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 cần 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.

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 ta cần tìm một cách nào đó mô phỏng bộ máy phát âm của ta. Đây cũng là nguyên lý của các "máy nói" cổ điển mà nổi tiếng trong số có 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 như 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 hạn 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.

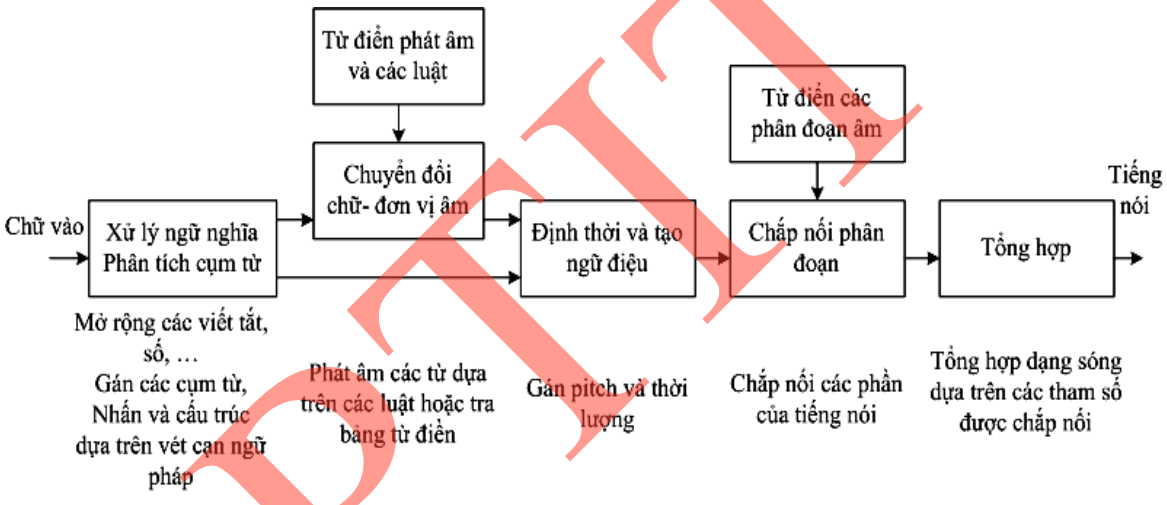
Cùng với sự hiểu biết của con người về bộ máy phát âm được nâng cao, các bộ tổng hợp sử dụng nguyên lý mô phỏng bộ máy phát âm ngày càng phức tạp và hoàn thiện hơn. Các hình dạng ống phức tạp được xấp xỉ bằng một loạt các ống đơn giản nhỏ hơn. Với mô hình các ống đơn giản, vì ta biết được các đặc tính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

Một ưu điểm 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, 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 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 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

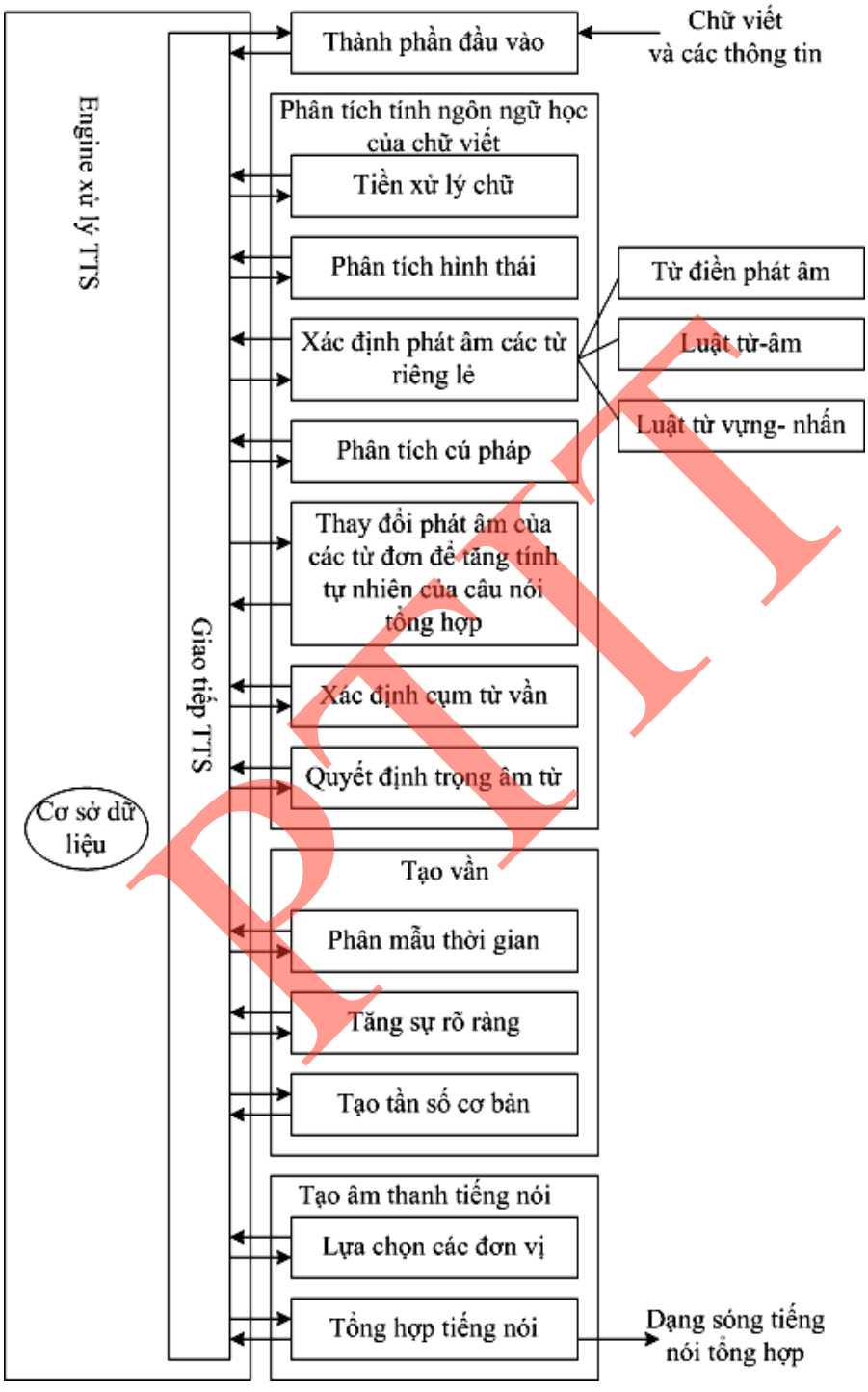
Việc chuyển đổi từ chữ viết sang tiếng nói (TTS) là mục tiêu đầy tham vọng và vẫn đang tiếp tục là tâm điểm chú ý của các nhà nghiên cứu phát triển. TTS có mặt ở nhiều ứng dụng phục vụ cuộc sống. Chẳng hạn như việc các ứng dụng truy cập email qua thoại, các ứng dụng cơ sở dữ liệu cho các dịch vụ hỗ trợ người khiếm thị... 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, 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.



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

Hoạt động của sơ đồ khối có thể mô tả sơ lược 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 "cơ chế" (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.

4.3.1. Phân tích chữ viết

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 điển 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 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 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 nhấn từ có thể được thêm vào cho mỗi từ trong từ điển, nhưng các luật cũng sẽ cần để gán dấu nhấn 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 nhấn 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 đã được sử dụng, các phân tích ngữ nghĩa và tính thực dụng một cách đầy đủ là vượt quá các khả năng của các hệ thống TTS hiện tại.

4.3.2. Tổng hợp tiếng nói

Các thông tin được trích từ các phân tích chữ viết được sử dụng để tạo ra âm điệu 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 được cung cấp từ các mô-đun trước đó. Việc tổng hợp có thể được thực hiện bằng một trong các phương pháp đã đề cập ở phần trên.

4.4. MỘT SỐ ĐẶC ĐIỂM CỦA VIỆC TỔNG HỢP TIẾNG VIỆT

Một điểm đầu tiên cần chú ý trong việc thực hiện tổng hợp tiếng Việt là sự khác biệt trong ngôn ngữ văn bản, văn phạm câu, khái niệm từ so với các ngôn ngữ tiếng Anh hoặc một số ngôn ngữ phổ biến khác. Ngoài ra, cấu trúc âm của tiếng Việt cũng có cách cấu âm, với các âm vị khác biệt rõ rệt. Đặc biệt là phải kể đến hiện tượng thanh điệu trong tiếng Việt.

Theo một số nghiên cứu thì thanh điệu trong tiếng Việt được quyết định bởi sự phân bố năng lượng tín hiệu và tần số cơ bản. Tuy nhiên, cho đến thời điểm này vẫn chưa có một phương pháp tổng hợp chính xác nào có thể tạo được thanh điệu với các âm sắc tự nghiên.

4.5. CÂU HỎI VÀ BÀI TẬP CUỐI CHƯƠNG

1. Mục đích của tổng hợp tiếng nói? Nêu một số ứng dụng của tổng hợp tiếng nói?
2. Có những phương pháp tổng hợp tiếng nói nào? Ý tưởng của từng phương pháp?
3. (Matlab) Sử dụng phương pháp tổng hợp trực tiếp đơn giản:
    i. Sử dụng máy tính cá nhân và phần mềm Matlab (hoặc các công cụ khác) xây dựng một hệ thống dừng đỗ xe buýt công cộng:
        1. Lưu file âm thanh các cụm từ thông báo (ví dụ: Điểm dừng tiếp theo", …), các địa danh
        2. Viết chương trình: chuẩn hóa dữ liệu tiếng Việt, phân tích văn bản, và ghép nối âm thanh để khi người nhập một cụm từ, chương trình sẽ thông báo về điểm dừng xe buýt.
4. (Matlab) Tương tự như bài 3, nhưng với hệ thống thông báo về số thứ tự khách hàng, thông tin về bàn phục vụ tại một điểm giao dịch ngân hàng
5. (Matlab) Tương tự như bài 3, nhưng với hệ thống thông báo số điện thoại của khách hang

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 thế kỷ qua. Cho đến nay, mặc dù 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, 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, tuy nhiên, 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 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, ta có thể phân loại các công nghệ nhận dạng thành các thế hệ như sau:

Thế hệ 1: Thế hệ này được đánh dấu 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 lường 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 đa (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: 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 trên đây 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 đuợc 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 ta có 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 này 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.

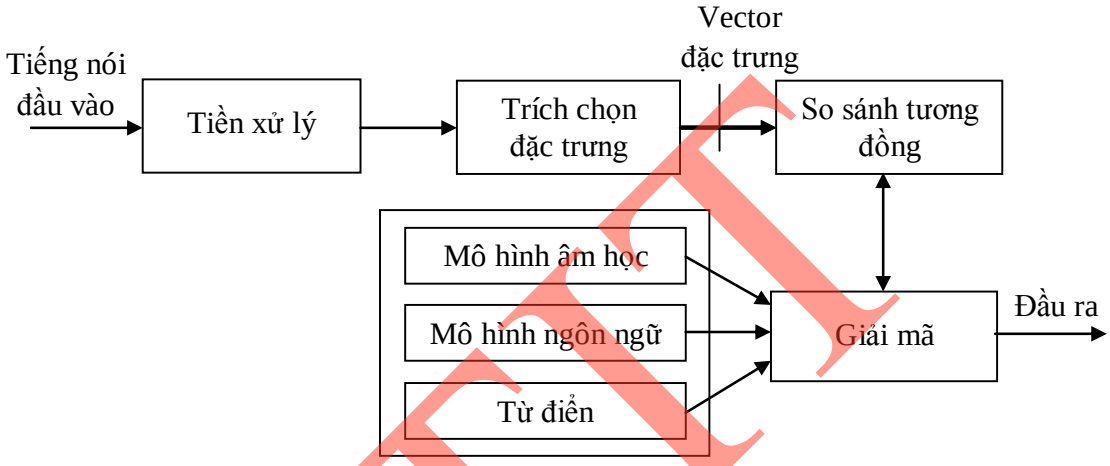
Một cách tổng quát, Victo Zue và đồng nghiệp đã định nghĩa một số tham số và dùng nó để phân chia các hệ thống nhận dạng theo các tham số đó như trình bày trong bảng 5.1.

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 trình bày 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 chuỗi 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 phức tạp hoặc không. 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. 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

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, 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 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), ta thấy rằng các phân tích phổ cho phép 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 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 ta cũng sử dụng độ chính xác 16 bít để biểu diễn mỗi thành phần phổ, khi đó 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, ta có thể làm giảm nhỏ thêm nữa các biểu diễn phổ thô 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 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 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, 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, 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. 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.

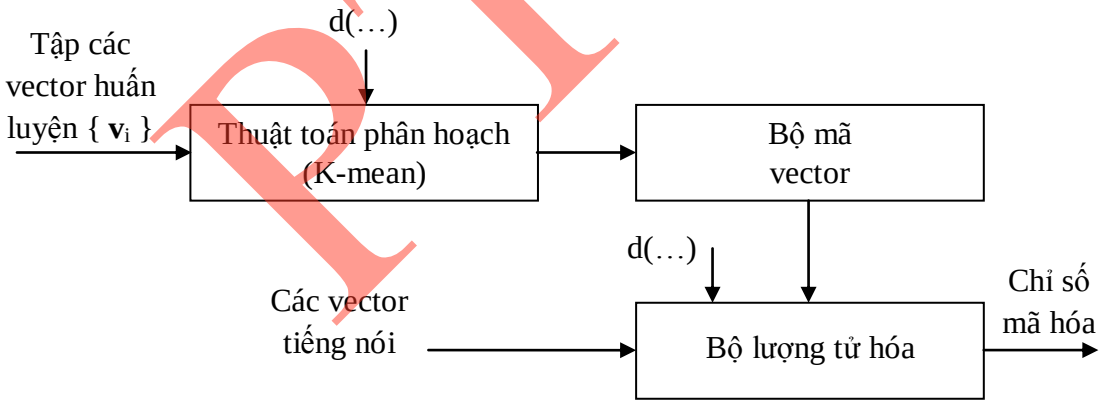
Tuy vậy cũng phải kể đến một số hạn chế của việc sử dụng bộ mã lượng tử hóa véc-tơ để biểu diễn các véc-tơ phổ tín hiệu tiếng nói. Chúng bao gồm:

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.

5.5.1.1. 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 ra một tập tối ưu các véc-tơ mã để biểu diễn các biến phổ quan sát được trong tập huấn luyện. Nếu ta ký hiệu kích cỡ của bộ mã lượng tử hóa véc-tơ là $M=2^B$ (ta gọi đây là một bộ mã B-bít), khi đó 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 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

5.5.1.2. 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 kiện 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.

5.5.1.3. 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:

$$d\left(v_i, v_j\right) = d_{\text{ij}} = \begin{cases} 0 & v_i = v_j \\ > 0 & v_i \neq v_j \end{cases} \tag{3.1}$$

5.5.1.4. 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.

5.5.1.5. *Thủ tục phân loại véc-tơ*

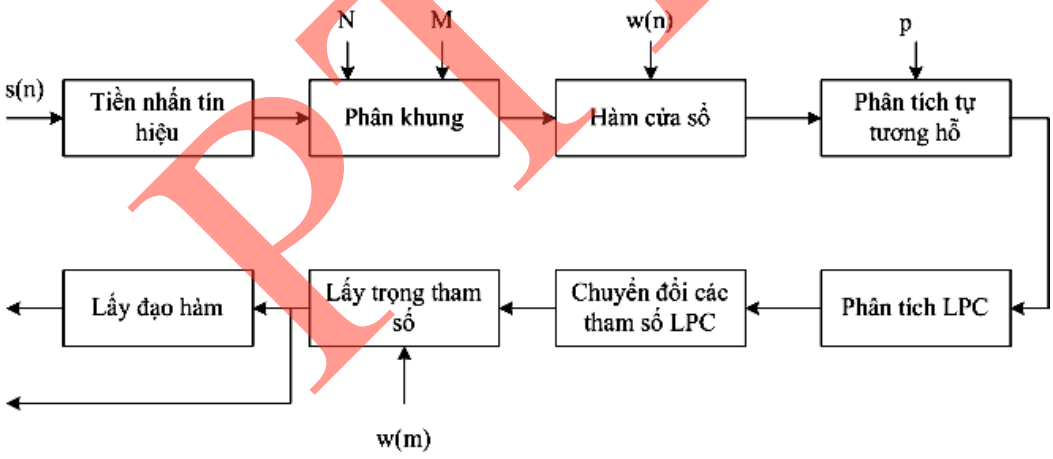
Việc phân loại các véc-tơ đối với các véc-tơ phổ bất kỳ về cơ bản là việc tìm hết trong bộ mã để tìm ra được một véc-tơ tương đồng nhất. Ta ký hiệu bộ véc-tơ mã của một bộ mã M véc-tơ là $\mathbf{y}_m$, ($1 \leq m \leq M$) và véc-tơ phổ cần phân loại (và lượng tự hóa) là **v**, khi đó chỉ số m^* của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

$$m^* = \arg \min_{1 \leq m \leq M} d(v, y_m) \tag{3.2}$$

Với các bộ mã có giá trị M lớn (chẳng hạn $M \geq 1024$), việc tính toán theo công thức (3.2) sẽ trở lên quá 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 ta thảo luận về các tính chất chung nhất của phương pháp phân tích LPC. Trong phần này 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

5.5.2.1. *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

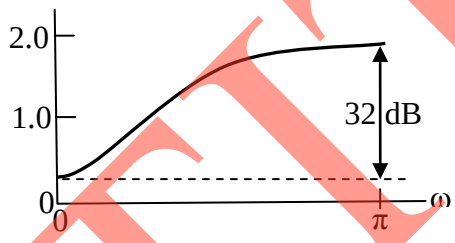
nhấn tín hiệu có thể là một bộ lọc với các tham số cố định hoặc có thể là một bộ lọc thích nghi có các tham số thay đổi chậm. Trong xử lý tín hiệu tiếng nói, người ta thường dùng một hệ thống mạch lọc bậc nhất có các tham số cố định có dạng:

$$H(z) = 1 - \tilde{a}z^{-1} \quad (0{,}9 \le \tilde{a} \le 1{,}0) \tag{3.3}$$

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

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \tag{3.4}$$

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 $H\left(e^{j\omega}\right)$ với giá trị $\tilde{a} = 0{,}95$. Từ hình vẽ, 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:

$$H(z) = 1 - \tilde{a}_n z^{-1} \tag{3.5}$$

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 $\tilde{a} = r_n(1)/r_n(0)$.

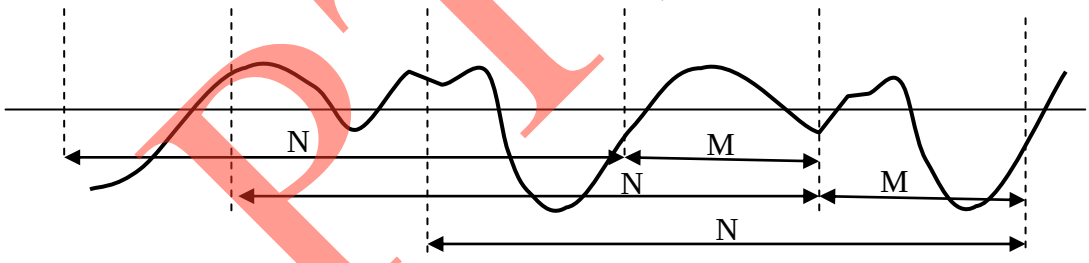
5.5.2.2. Phân khung tín hiệu

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 phần 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:

$$x_l(n) = \tilde{s}(Ml + n) \quad n = 0,1,...,N-1;\ l = 0,1,...,L-1 \tag{3.6}$$

Điều này có nghĩa là khung tín hiệu đầu tiên $x_0(n)$ bao gồm các mẫu $\tilde{s}(0), \tilde{s}(1), \ldots, \tilde{s}(L-1)$; khung tín hiệu thứ hai $x_l(n)$ bao gồm các mẫu $\tilde{s}(M), \tilde{s}(M+1), \ldots, \tilde{s}(M+N-1)$; và khung tín hiệu thứ L bao gồm các mẫu $\tilde{s}(M(L-1)), \tilde{s}(M(L-1)+1), \ldots, \tilde{s}(M(L-1)+N-1)$; 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

5.5.2.3. 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:

$$\tilde{x}_l(n) = x_l(n)\mathrm{w}(n) \quad 0 \le n \le N-1 \tag{3.7}$$

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:

$$w(n) = 0,54 - 0,46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \tag{3.8}$$

5.5.2.4. Phân tích tính tự tương quan

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

$$\Phi_l(n) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \quad m = 0,1,...,p \tag{3.9}$$

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.

5.5.2.5. Phân tích LPC

Bước tiếp theo trong quá trình phân tích là phép phân tích LPC, trong đó mỗi khung 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 thuật toán Durbin được diễn giải như sau. Để thuận tiện, ta tạm bỏ chỉ số l trong biểu thức $r_l(m)$.

$$E^{(0)} = \Phi_l(0) \tag{3.10}$$

$$k_i = \frac{\{\Phi_l(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} \Phi_l(|i-j|)\}}{E^{(i-1)}} \quad (1 \le i \le p) \tag{3.11}$$

$$\alpha_i^{(i)} = k_i \tag{3.12}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \tag{3.13}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{3.14}$$

Trong công thức tính tổng của công thức thứ hai ở trên, (3.11), ta bỏ qua trường hợp i=1. Hệ các phương trình trên được giải theo phương pháp truy hồi với i=1,2,…, p và kết quả cuối cùng thu được là:

$$a_m = \alpha_m^{(p)} \quad (1 \le m \le p) \tag{3.15}$$

$$k_m = R_{\text{coef}} \tag{3.16}$$

$$g_m = \log\left(\frac{1-k_m}{1+k_m}\right) \tag{3.17}$$

(3.15) là các hệ số LPC, (3.16) là các hệ số phản xạ, và (3.17) là lô-ga-rít các hệ số tỷ lệ diện tích.

5.5.2.6. 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:

$$c_0 = \ln\left(\sigma^2\right) \tag{3.18}$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \quad (1 \le m \le p) \tag{3.19}$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \quad (m > p) \tag{3.20}$$

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 với biến đổi Fourier của các giá trị lô-ga-rít của biên độ phổ. Tập các hệ số cepstral được chứng minh là một tập các đặc trưng đáng tin cậy và chắc chắn (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 biến Q≈3p/2.

5.5.2.7. 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:

$$\log\left|S\left(e^{j\omega}\right)\right| = \sum_{m=-\infty}^{\infty} c_m e^{j\omega m} \tag{3.21}$$

$$\frac{\partial}{\partial\omega}\left[\log\left|S\left(e^{j\omega}\right)\right|\right] = \sum_{m=-\infty}^{\infty} (-jm) c_m e^{-j\omega m} \tag{3.22}$$

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, ta đã thực hiện việc thay đổi trọng các tham số. Kết quả ta có:

$$\frac{\partial}{\partial\omega}\left[\log\left|S\left(e^{j\omega}\right)\right|\right] = \sum_{m=-\infty}^{\infty} \hat{c}_m e^{-j\omega m} \tag{3.23}$$

trong đó:

$$\hat{c}_m = c_m(-jm) \tag{3.24}$$

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 (3.23), ta cần phải đưa ra một dạng tổng quát hơn đối với các hệ số trọng số:

$$\hat{c}_m = \mathrm{w}_m c_m \tag{3.25}$$

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

$$\mathrm{w}_m = \left[1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right)\right] \quad (1 \le m \le Q) \tag{3.26}$$

Hàm tính toán trọng số cho ở công thức (3.26) 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.

5.5.2.8. 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, 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:

$$\frac{\partial}{\partial t}\left[\log\left|S\left(e^{j\omega},t\right)\right|\right]=\sum_{m=-\infty}^{\infty}\frac{\partial c_m(t)}{\partial t}e^{-j\omega m} \tag{3.27}$$

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), 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). 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:

$$\frac{\partial c_m(t)}{\partial t}=\Delta c_m(t)\approx\mu\sum_{k=-K}^{K}kc_m(t+k) \tag{3.28}$$

Trong đó, μ là một hằng số chuẩn hóa thích hợp và (2K+1) là số khung tín hiệu mà trên đó 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:

$$o'_t=\left(\hat{c}_1(t),\hat{c}_2(t),...,\hat{c}_Q(t),\Delta\hat{c}_1(t),\Delta\hat{c}_2(t),...,\Delta\hat{c}_Q(t)\right) \tag{3.29}$$

Trong công thức (3.29), o'_t là một véc-tơ gồm 2Q thành phần và (.)' biểu diễn phép chuyển vị ma trận.

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

5.5.2.9. 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, 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

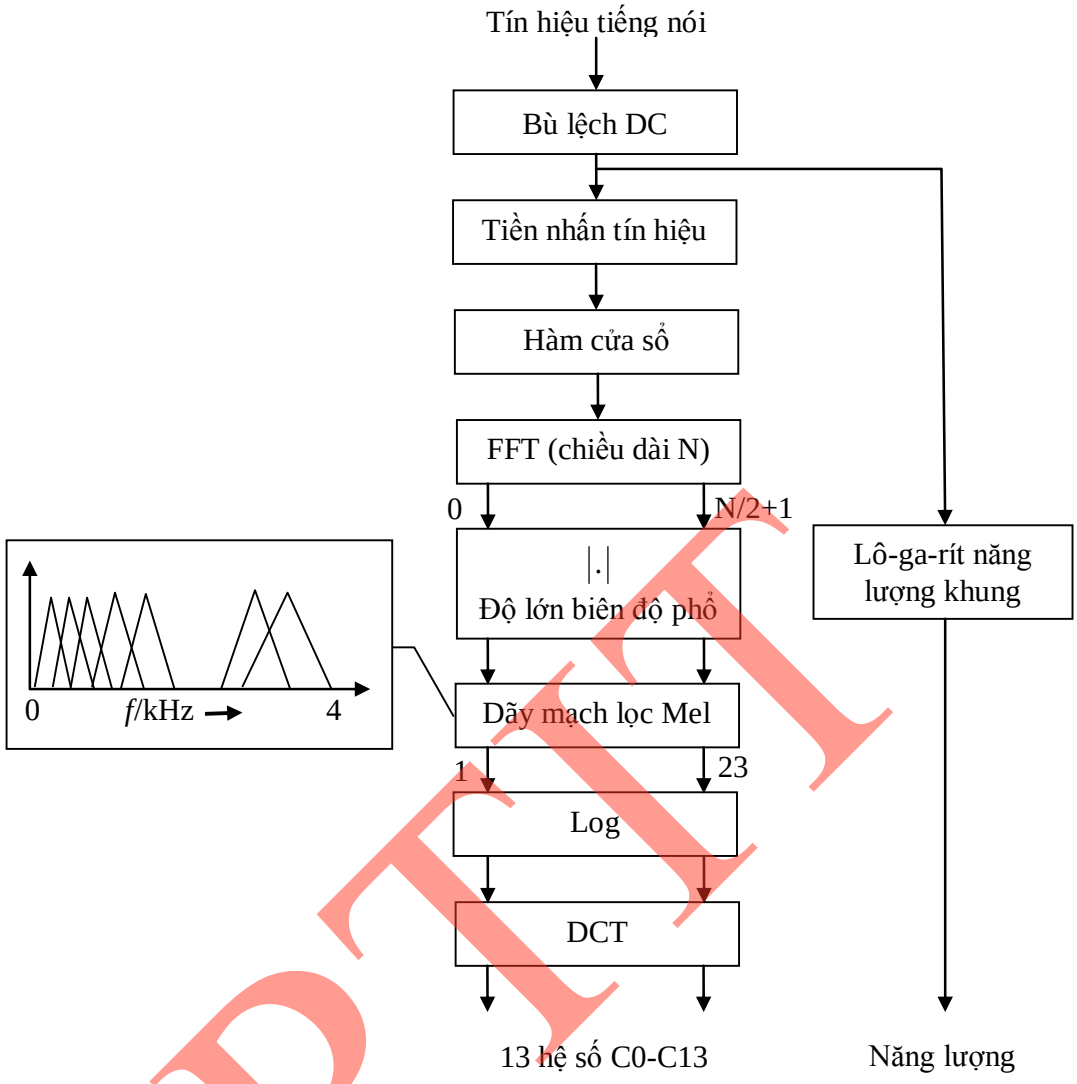
Bảng 5.2: Một số giá trị tham số phổ biến của phép phân tích LPC

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. Ta biết rằng, hệ thống cảm nhận 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 tích lũy vào một vùng tần số xác định tạo 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.

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 trong miền 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, 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.



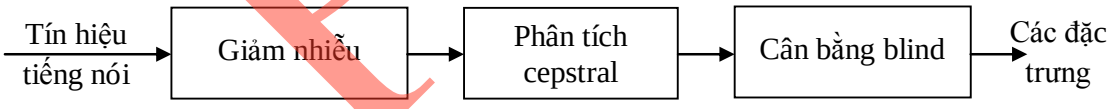
Hình 5.6 Sơ đồ khối quá trình phân tích MFCC

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:

$$\Delta C_i(k) = \frac{\sum_{j=1}^{N_\Delta} j\left[C_i(k+j) - C_i(k-j)\right]}{\sum_{j=1}^{N_\Delta} j^2} \tag{3.30}$$

Hệ số N_Δ trong công thức (3.30) 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 (3.30) 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 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 trung bình 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, 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 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), ta giả thiết rằng những quy luật quản lý sự thay đổi trên có thể suy ra một cách dễ dàng, 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 tiết 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 so 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...

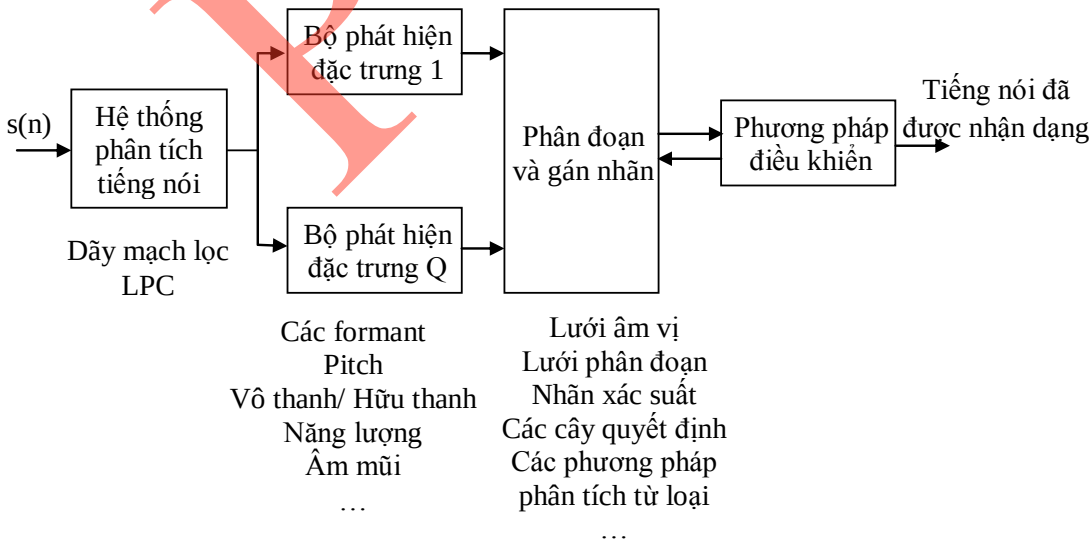
- Có chất lượng tốt. Người ta đã chỉ ra rằng việc sử dụng phương pháp nhận dạng mẫu trong nhận dạng tiếng nói luôn cho phép hệ thống hoạt động tốt đối với bất kỳ nhiệm vụ nào với yêu cầu công nghệ vừa phải.

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 các 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.



Hình 5.8 Sơ đồ khối một hệ thống nhận dạng tiếng nói theo 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.

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 xát (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 tắc, 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 sử 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 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, 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

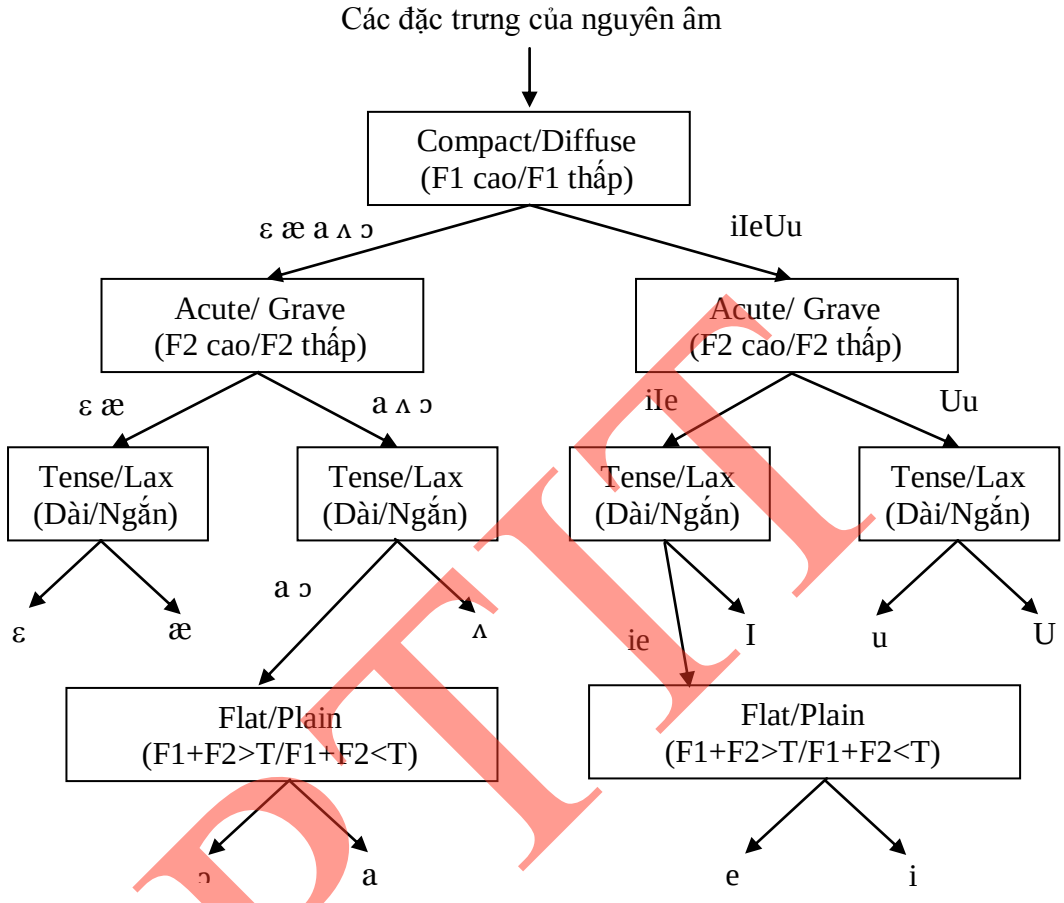
gán nhãn có thể tận dụng các giới hạn về từ vựng (lexical) của các từ để chỉ cần xem xét các từ với n đơn vị âm tiết, trong đó việc phân đoạn cho ta n-1 điểm phân đoạn. Những điều kiện hạn chế vừa nêu có vai trò quan trọng cho phép ta giảm nhỏ không gian tìm kiếm và tăng đáng kể chất lượng hoạt động của hệ thống.

Kết quả của giai đoạn phân đoạn và gán nhãn thường 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.

5.6.1.1. Bộ phân loại các âm vị nguyên âm

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. 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 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)

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.

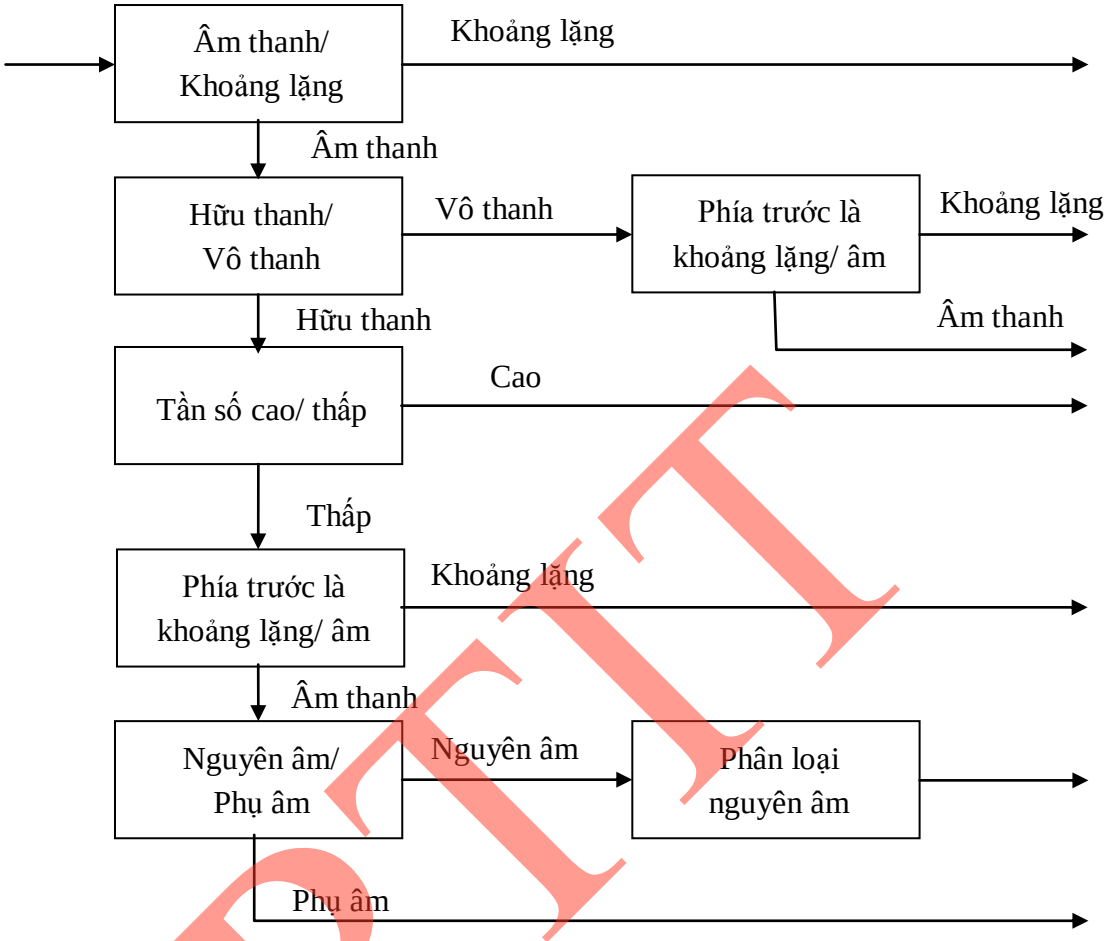


Hình 5.9 Một phương pháp đơn giản phân loại nguyên âm tiếng Anh

5.6.1.2. 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, 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 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). 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.



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

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/khoảng 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 khoảng 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), 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.

Thủ tục phân tách nguyên âm được trình bày trong hình 5.9 có thể được sử dụng thêm như một phép phân loại mịn 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 thủ tục minh họa không đưa ra được một cách nào có thể phân biệt các âm kép (diphthongs) từ các nguyên âm.

5.6.1.3. 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)

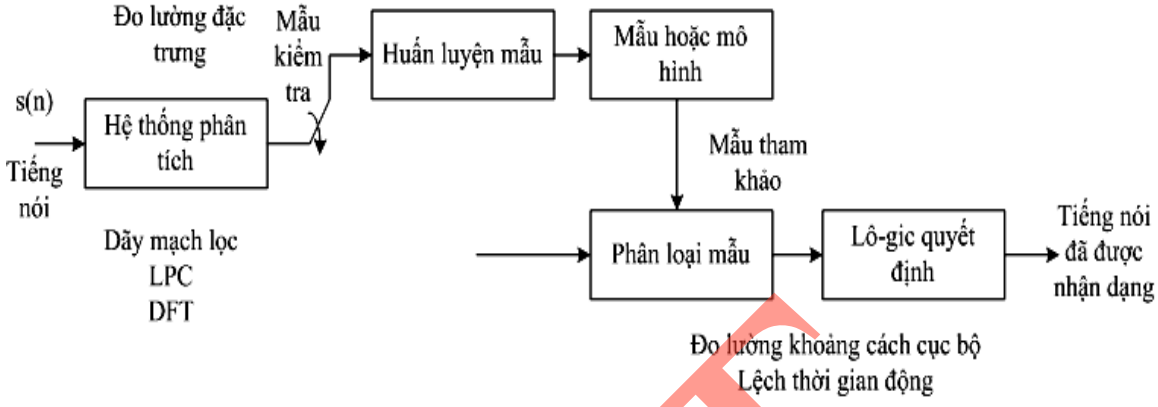
3. Thiết kế các bộ phân loại âm thanh cũng không phải là các thiết kế tối ưu. Phương pháp ad hoc thường được sử dụng để xây dựng các cây nhị phân quyết định. Gần đây, các phương pháp cây hồi quy (regression) và phân loại (CART) được sử dụng thay thế để cho phép các cây quyết định nhất quán hơn. Tuy vậy, 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.



Hình 5.11 Sơ đồ khối của 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 (dãy) 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), 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 chỉnh (chỉnh 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.

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, 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, 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 về các khái niệm nguồn thông tin vừa đề cập cũng như hạn chế của chúng, 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.

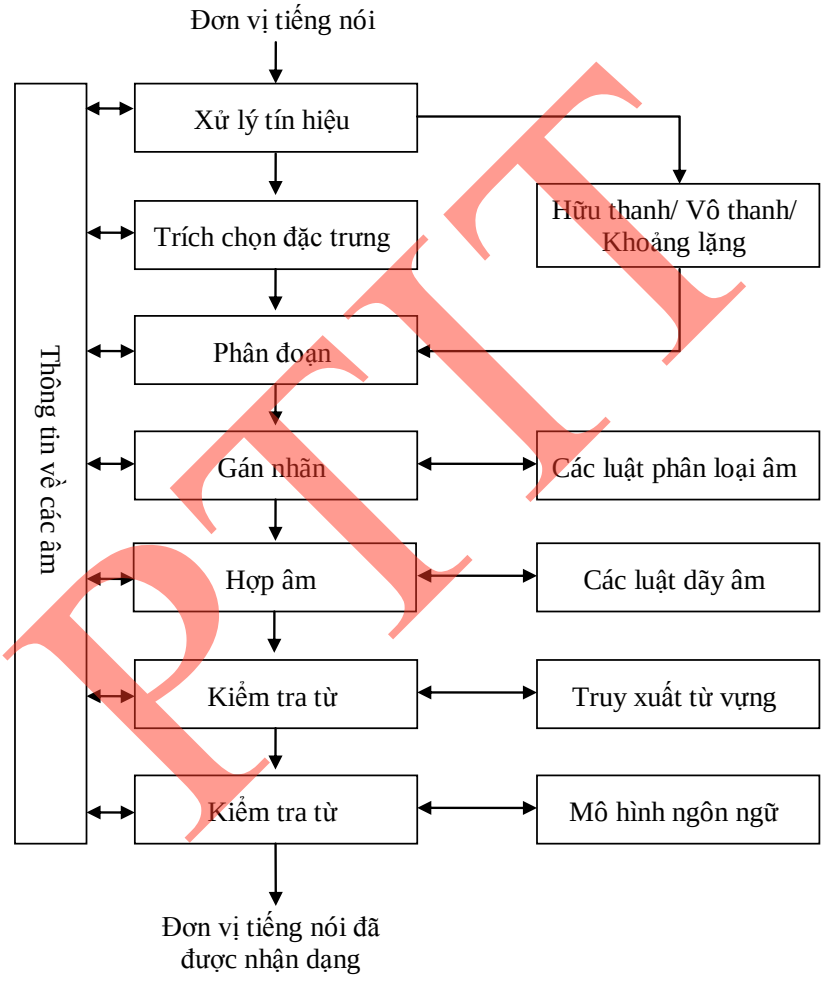
3. Power plants colorless happily old.

4. Good ideas often run when least expected.

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 ta đang nói đến bóng đá có thể hiểu là đội có tên là những

con gấu đã chiến thắng 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.

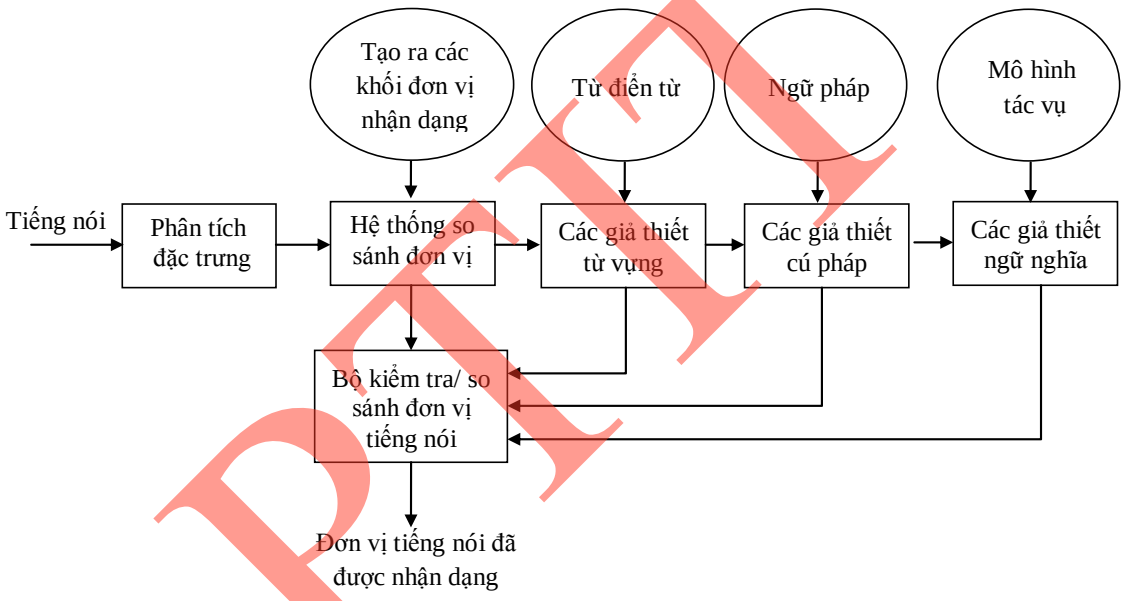
Việc kết hợp các điều kiện hạn chế của các nguồn thông tin vừa kể sẽ cho phép hệ thống nhận dạng tín hiệu tiếng nói hoạt động với chất lượng cao hơn. Có nhiều cách kết hợp các nguồn thông tin vừa kể vào một hệ thống nhận dạng. Phương pháp đầu tiên phổ biến nhất có thể kể đến là bộ xử lý "bottom-up" được trình bày trong hình 5.12.



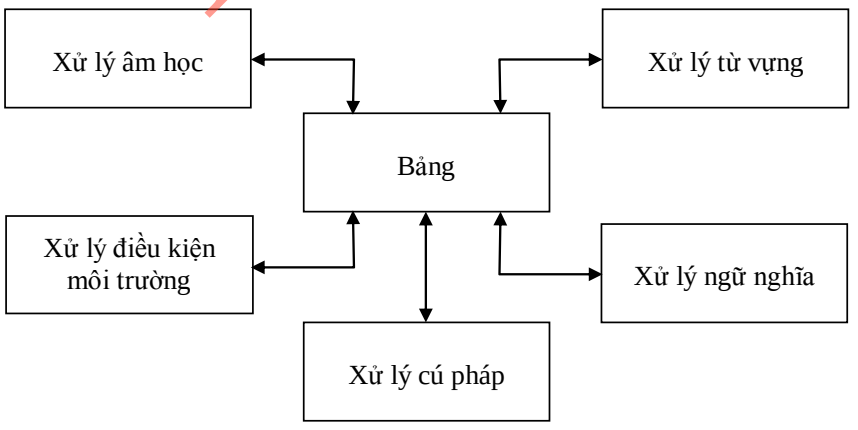
Hình 5.12 Phương pháp tích hợp "bottom-up" của hệ thống nhận dạng tiếng nói

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

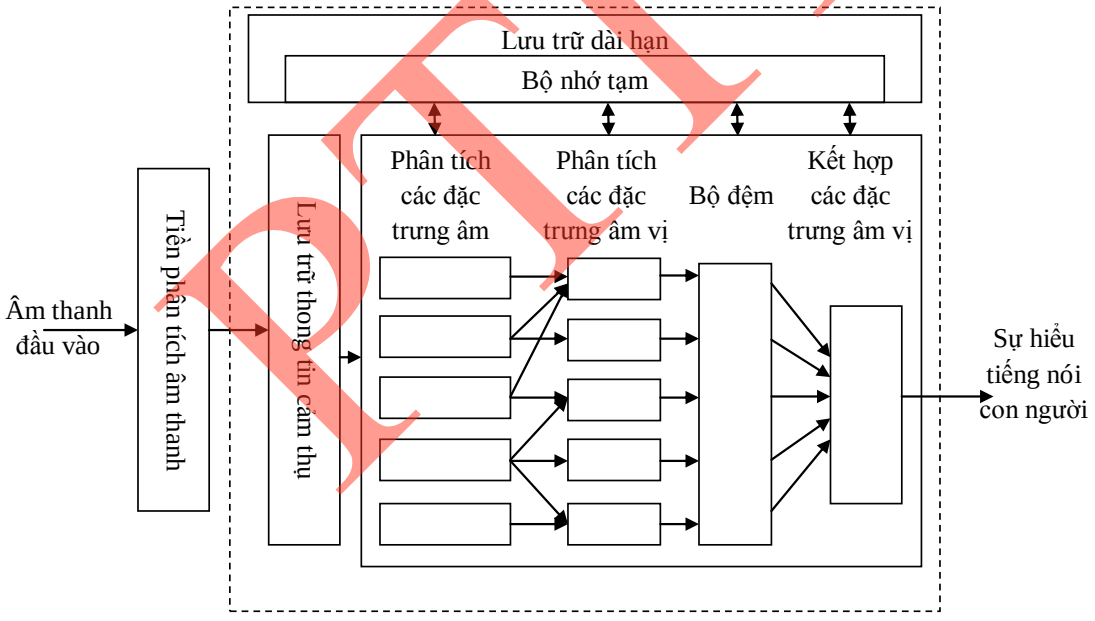


Hình 5.14 Phương pháp tích hợp "blackboard" của hệ thống nhận dạng tiếng nói

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

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 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 con người 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 cho đến nay con người 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.

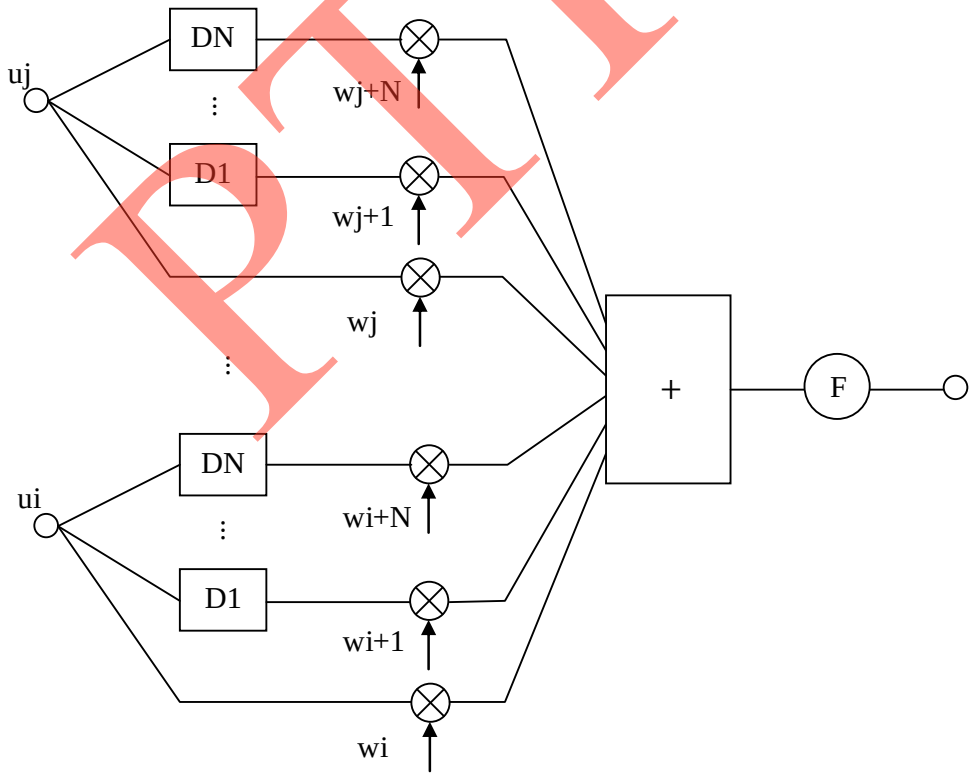


Hình 5.15 Sơ đồ khối ý tưởng của một hệ thống hiểu tiếng nói con người

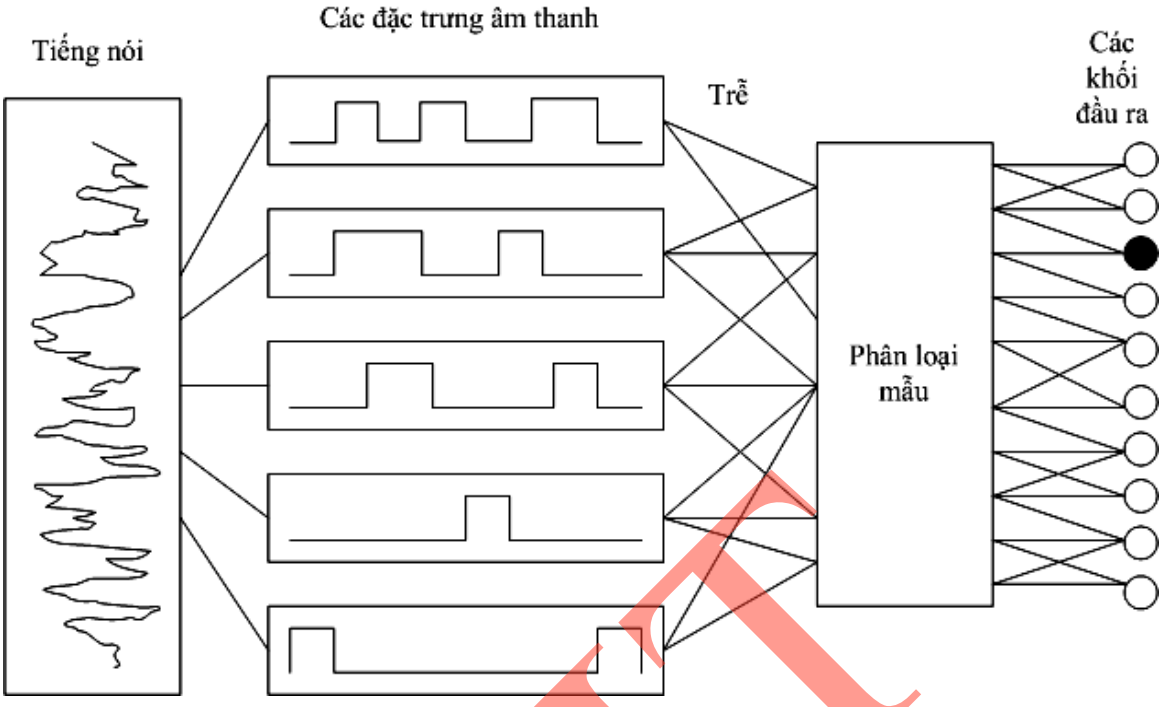
Các mạng nơ-ron nhân tạo truyền thống (conventional) là các cấu trúc dùng để giải quyết các bài toán liên quan đến các mẫu tĩnh. Do đó, để có thể áp dụng cho tín hiệu tiếng nói, một tín hiệu có bản chất động, ta cần có một số thay đổi trong các cấu trúc mạng truyền thống. Mặc dù cho đến nay chưa có một 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.

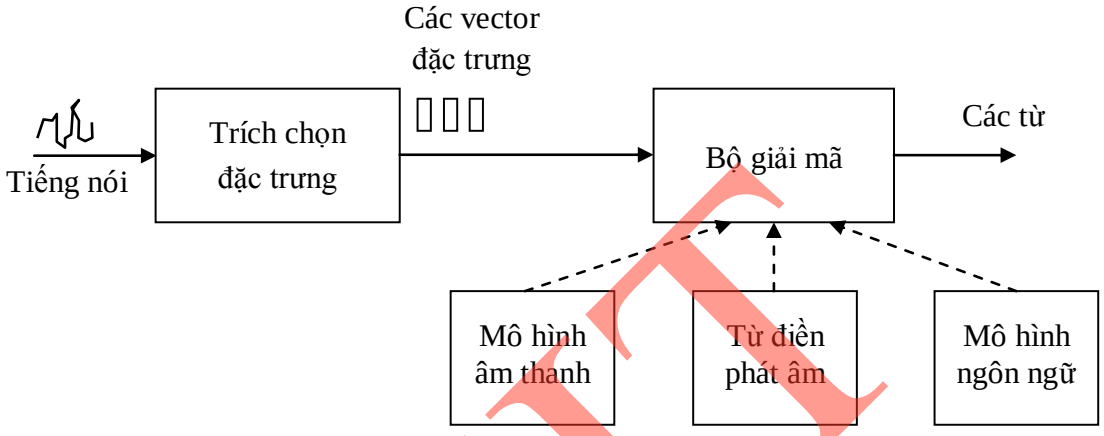
- Các mạng nơ-ron kế thừa bản chất là một cấu trúc chịu lỗi tốt (fault tolerance). Vì các thông tin nhúng trong mạng được trải (lan) đến mọi phần tử tính toán trong mạng, điều này khiến cho cấu trúc khá trơ (least sensitive) với nhiễu hoặc các lỗi không hoàn hảo bên trong cấu trúc.

- 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 tuyến 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 tính 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 $\mathbf{w} = w_1,...,w_K$ có khả năng cao nhất đã tạo ra $\mathbf{y}$. Nói cách khác, bộ giải mã cố gắng giải bài toán:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left[p(\mathbf{w}|\mathbf{y}) \right] \tag{3.31}$$

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

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left[p(\mathbf{y}|\mathbf{w})\, p(\mathbf{w}) \right] \tag{3.32}$$

Độ tương đồng $p(\mathbf{y}|\mathbf{w})$ 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 $\log(p(\mathbf{y}|\mathbf{w})) + \alpha p(\mathbf{w}) + \beta p(\mathbf{w})$, 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 tham 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.

5.6.5.1. 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.

5.6.5.2. 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:

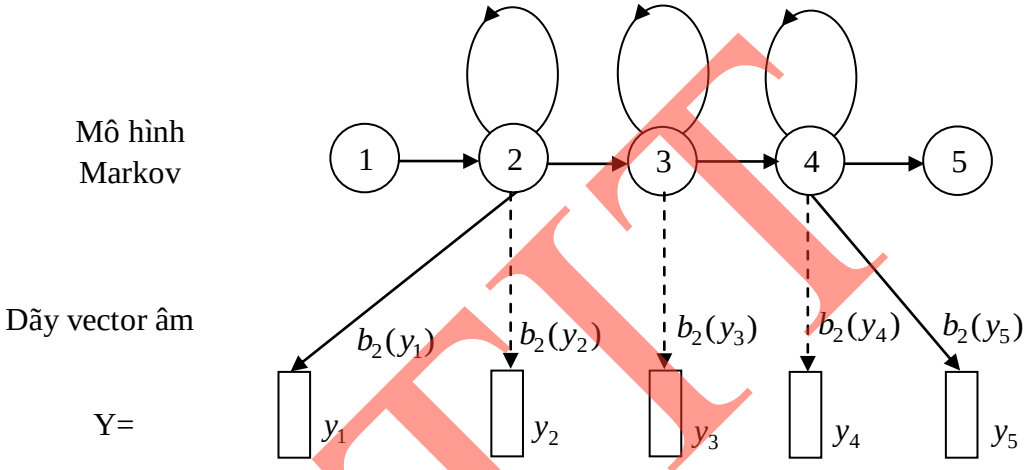
$$p(\mathbf{y}\,|\,\mathbf{w}) = \sum_{Q} p(\mathbf{y}\,|\,Q)\,p(Q\,|\,\mathbf{w}) \tag{3.33}$$

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 $Q_K = q_1^{(k)}, q_1^{(k)},....$ Khi đó ta có:

$$p(Q|\mathbf{w}) = \prod_{k=1}^{K} p(Q_k | w_k) \tag{3.34}$$

Ở đây $p(Q_K|w_k)$ 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 (3.33) 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:

$$b_j(y) = \sum_{m=1}^{M} c_{jm} \aleph\left(y; \mu_{jm}, \sum_{jm}\right) \tag{3.35}$$

$\aleph$ biểu diễn phân bố chuẩn với giá trị trung bình μ_{jm} và covariance $\sum_{jm}$. Số lượng các thành phần trong công thức (3.35) 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:

$$p(y|Q) = \sum_X p(x, y|Q) \tag{3.36}$$

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

$$p(x, y|Q) = a_{x(0),x(1)} \prod_{t=1}^{T} b_{x(t)} a_{x(t),x(t+1)} \tag{3.37}$$

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. MỘT SỐ ĐẶC ĐIỂM CỦA VIỆC NHẬN DẠNG TIẾNG VIỆT

Việc xây dựng một hệ thống nhận dạng tiếng Việt một cách chính xác với lượng từ vựng lớn và có đáp ứng thời gian thực là rất khó khăn vì tính phức tạp của ngôn ngữ. Cùng một âm vị phát ra bởi nhiều người sẽ có những đặc điểm về mặt âm học khác nhau. So với ngôn ngữ của nhiều nước, thì tiếng Việt có sự phân hóa về mặt thổ ngữ tương đối lớn. Có một sự thay đổi lớn giữa cách phát âm giữa ba miền Bắc, Trung, Nam. Ngay trong một miền, ở các vùng địa phương khác nhau cũng có sự phát âm dẫn khác nhau.

Thêm nữa, cũng giống như ngôn ngữ của một số nước khu vực Châu Á, tiếng Việt có thanh điệu. Sự khác biệt giữa các thanh điệu có khi rất nhỏ khi được phát âm bởi một số vùng miền. Chẳng hạn, phía Bắc có sự phát âm s và x tương đương nhau; hoặc dấu "?" và "~" được phát âm giống nhau ở vùng Bắc Trung bộ.

Sự phức tạp này khiến cho những phương pháp nhận dạng của các ngôn ngữ khác không hiệu quả khi áp dụng với tiếng Việt

5.8. CÂU HỎI VÀ BÀI TẬP CUỐI CHƯƠNG

1. Ý tưởng cơ bản của phương pháp đối sánh mẫu trong nhận dạng tiêng nói?

2. Ý tưởng cơ bản của phương pháp sử dụng mạng nơ-ron trong nhận dạng tiếng nói?

3. Ý tưởng cơ bản của việc sử dụng HMM trong nhận dạng tiếng nói?

4. Sự khác biệt của giác hệ thống nhân dạng tiếng nói: rời rạc và liên tục; nhận dạng tiếng nói và nhận dạng người nói?

5. (Matlab) 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:

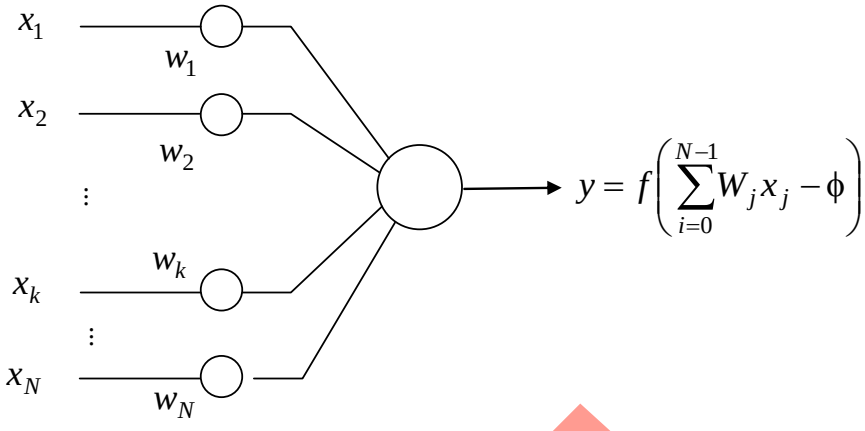
- Phi tuyến. Cho phép xử lý phi tuyến.
- Cơ chế ánh xạ đầu vào - đầu ra cho phép học có giám sát.
- Cơ chế thích nghi. Thay đổi tham số phù hợp với môi trường.
- Đáp ứng theo mẫu huấn luyện.
- Thông tin theo ngữ cảnh.Tri thức được biểu diễn tuỳ theo trạng thái và kiến trúc của mạng.
- Cho phép có lỗi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

CƠ SỞ VỀ MẠNG 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:

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

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:

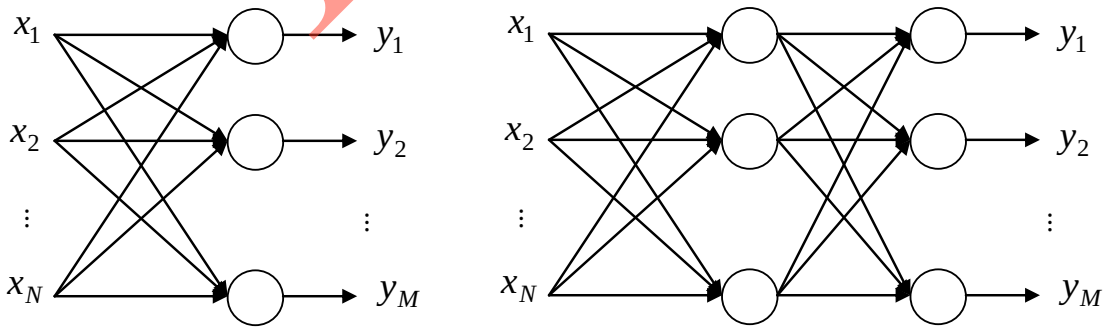
6. Hàm ngưỡng cứng: $f(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$

7. Hàm log-sin: $f(x) = \dfrac{1}{1+e^{-\beta x}} \quad (\beta > 0)$

CẤU HÌNH MẠNG NƠ-RON

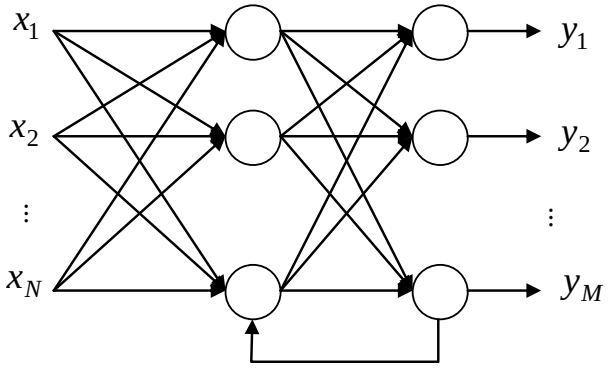
Một yếu tố quan trọng cho việc thiết lập và ứng dụng của mạng nơ-ron là cấu trúc tô-pô của mạng (network topology). Có ba kiểu cấu trúc cơ bản là:

7. Mạng một tầng hoặc nhiều tầng:



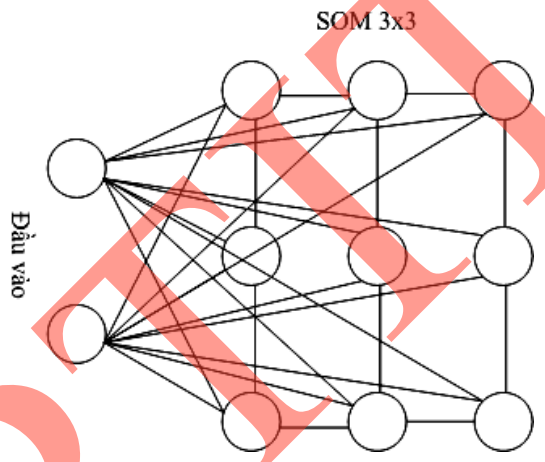
Hình A.2: Cấu trúc mạng nơ-ron một tầng (a) và hai tầng (b)

8. Mạng hồi quy:



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

9. 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:

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

Các giá trị của $X(t)$ 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):

$$p_j(0) = \Pr[X_0 = j] \quad (j=0,1,\ldots)$$

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:

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

Từ công thức trên ta thấy, hàm mật độ phân bố xác suất hợp rời rạc 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:

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

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

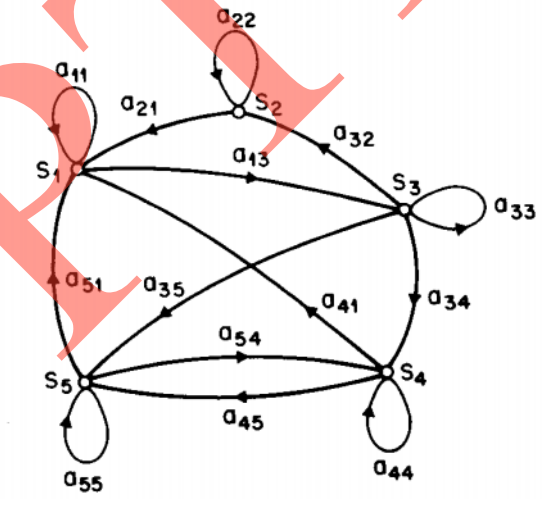
$$\Pr\left[X_n = i_n,...,X_0 = i_0\right] = a_{i_{n-1}i_n}...a_{i_0i_1}p_{i_0}(0)$$

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

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

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}.



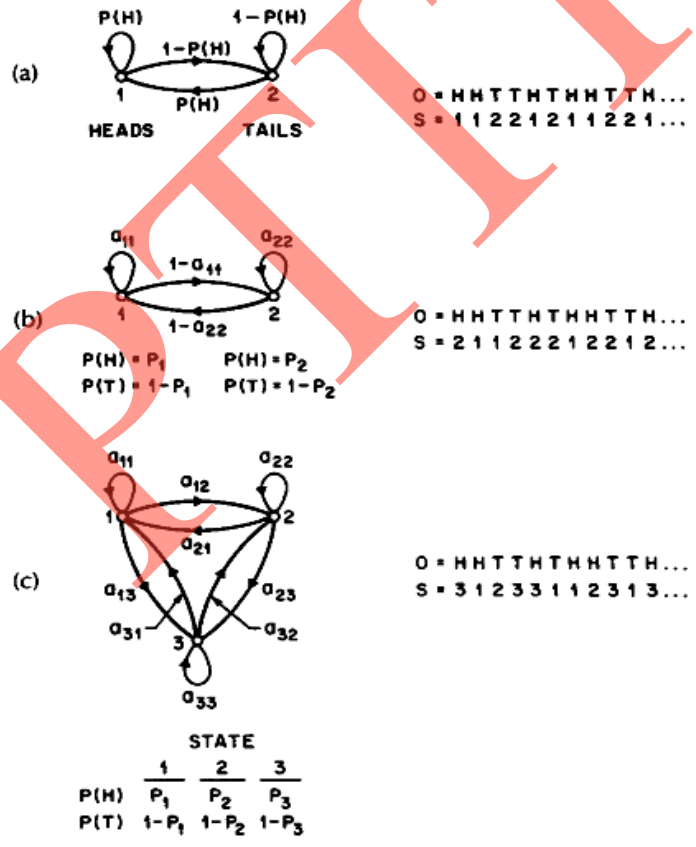
Hình B.1: Minh họa một chuỗi Markov rời rạc với 5 trạng thái

MÔ HÌNH MARKOV ẨN

Trong phần trên 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, ta xét ví dụ các mô hình tung đồng xu như sau. Một người thực hiện việc tung đồng xu nhưng không nói cho ta biết anh ta đã làm chính xác những gì. Anh ta chỉ thông báo cho ta kết quả của mỗi đồng xu lật. Như vậy, đối với 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 ba mô hình Markov có thể đối với thí nghiệm tung đồng xu ẩn

Hình B.2 minh họa 3 trường hợp ví dụ. Trường hợp thứ nhất tương ứng với giả thiết chỉ một đồng 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, 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.

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. 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 với kích thước của mô hình.

Một HMM được đặc trưng bởi:

11. 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.
12. Số các ký hiệu quan sát phân biệt với mỗi trạng thái, tức là kích thước bộ chữ rời rạc.
13. Phân bố xác suất chuyển trạng thái **P** trong đó $a_{ij} = \Pr[X_{n+1} = S_j | X_n = S_i]$, $(1 \le i, j \le N)$. 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, ta có $a_{ij} > 0$ với mọi i, j. Với các loại HMM khác, ta có $a_{ij} = 0$ cho một hoặc nhiều hơn một cặp (i,j).
14. Phân bố xác suất ký hiệu quan sát ở trạng thái j, $B = \{b_j(k)\}$, trong đó $b_j(k) = \Pr[v_k(t) | X_t = S_j]$, $(1 \le j \le N, 1 \le k \le M)$.
15. Phân bố trạng thái khởi đầu $\pi = \{\pi_i\}$ trong đó $\pi_i = \Pr[X_1 = S_i]$, $(1 \le j \le N)$.

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.