HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN XUÂN HÀ

NGHIÊN CỨU NHẬN DẠNG NGÔN NGỮ NÓI TỰ ĐỘNG DỰA TRÊN TẦN SỐ CƠ BẢN

LUẬN VĂN THẠC SĨ KỸ THUẬT

HÀ NỘI - 2020

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN XUÂN HÀ

NGHIÊN CỨU NHẬN DẠNG NGÔN NGỮ NÓI TỰ ĐỘNG DỰA TRÊN TẦN SỐ CƠ BẢN

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

LUẬN VĂN THẠC SĨ KỸ THUẬT
(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC:
PGS.TS. HÀ HẢI NAM

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều đã trình bày là của cá nhân tôi hoặc được tôi tổng hợp từ nhiều nguồn tài liệu. Tất cả các nguồn tài liệu tham khảo có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin chịu toàn bộ trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của tôi.

Hà Nội, tháng 04 năm 2020

Tác giả luận văn

Trần Xuân Hà

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc đến **PGS.TS. Hà Hải Nam**, người đã giúp tôi chọn đề tài, định hình hướng nghiên cứu, tận tình hướng dẫn và chỉ bảo tôi trong suốt quá trình thực hiện luận văn tốt nghiệp.

Tôi xin bày tỏ lòng biết ơn trân thành tới các thầy, cô giáo trong trường Học viện Công nghệ và Bưu chính Viễn thông. Các thầy, cô giáo đã dạy bảo và truyền đạt cho tôi rất nhiều kiến thức, giúp tôi có được một nền tảng kiến thức vững chắc sau những ngày tháng học tập tại trường. Và xin gửi lời cảm ơn đến Ban Lãnh đạo và các đồng chí, đồng nghiệp tại Phòng Giám định Kỹ thuật số và điện tử - Viện Khoa học hình sự - Bộ Công đã hết sức tạo điều kiện thuận lợi cho tôi trong suốt quá trình học tập và thực hiện luận văn. Tôi xin gửi lời cảm ơn sâu sắc tới các bạn khóa 2018 đợt 2 đã ủng hộ khuyến khích tôi trong suốt quá trình học tập tại trường.

Cuối cùng, tôi muốn gửi lời cảm ơn sâu sắc nhất đến gia đình và bạn bè, những người thân yêu luôn kịp thời động viên và giúp đỡ tôi vượt qua những khó khăn trong học tập cũng như trong cuộc sống.

Hà Nội, tháng 04 năm 2020

Tác giả luận văn

Trần Xuân Hà

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
ACF	Autocorreclation Function	Hàm tự tương quan
AMDF	Averaged Magnitude Differentiate Function	Hàm vi sai biên độ trung bình
BPN	Back- propagation Network	Mạng lan truyền ngược
C-V	Consonant - Vowel	Nguyên âm - Phụ âm
DCT	Discrete Cosine Transform	Biến đổi cosin rời rạc
DFT	Discrete Fourier Transform	Biến đổi Fourier rời rạc
DFFT	Discrete Fast Fourier Transform	Biến đổi Fourier nhanh rời rạc
FFT	Fast Fourier Transform	Biến đổi Fourier nhanh
F0	Fundamental Frequency of Speech	Tần số cơ bản
Frame	Frame	Khung
ITU	Upper Energy threshold	Ngưỡng năng lượng trên
ITL	Low Energy threshold	Ngưỡng năng lượng dưới
IZCT	Zero Crossing Rate threshold	Ngưỡng năng lượng thấp hơn
Nơ ron	Neural	Tế bào thần kinh
MFCC	Mel frequency cepstrum computation	Tính toán cepstrum tần số Mel
PIP	Preferred Installer Program	Trình quản lý gói thư viện của ngôn ngữ lập trình Python
STFT	Short-Time Fourier Transform	Biến đổi Fourier thời gian ngắn

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Công tác giám định âm thanh ở Việt Nam được Viện Khoa học hình sự - Bộ Công an bắt đầu triển khai từ năm 1998, đến nay đã được 20 năm, số lượng vụ án hàng năm ngày càng tăng, nhu cầu phân loại tự động tiếng nói ban đầu trước khi tiến hành giám định là rất lớn.

Công việc giám định âm thanh nghiên cứu phạm vi ổn định của một số tham số tiếng nói để nhận dạng người nói hoặc một nhóm người nói. Luận văn này nghiên cứu về một trong các tham số tiếng nói nếu trên, đó là tần số cơ bản. Xuất phát từ thực tế trên, tôi chọn đề tài **"Nghiên cứu nhận dạng ngôn ngữ nói tự động dựa trên tần số cơ bản".**

2. Tổng quan về đề tài nghiên cứu

Từ lâu, người ta đã nhận ra rằng thông tin ngôn điệu (nghĩa là thông tin có nguồn gốc từ các đặc điểm của giọng nói như cao độ, biên độ và tốc độ nói) góp phần lớn vào việc nhận dạng giọng nói cũng như nhận dạng ngôn ngữ nói. Thực tế là ngữ điệu lời nói đóng một vai trò quan trọng trong việc hiểu ngôn ngữ nói, cho thấy các đặc trưng ngôn điệu cũng có thể là yếu tố cơ bản của nhận dạng ngôn ngữ nói. Ngoài ra, các tham số có nguồn gốc từ tần số cơ bản (cao độ giọng nói) là ổn định và được cho là mạnh hơn tần số định dạng. Vì thế, người ta chấp nhận rằng các mẫu biến thể của tần số cơ bản là một trong những tham số tốt nhất để thể hiện các đặc trưng ngôn điệu của ngôn ngữ nói. Chúng ta đã cố gắng đạt được một hệ thống nhận dạng ngôn ngữ nói tự động bằng cách sử dụng thông tin ngôn điệu bắt nguồn từ tần số cơ bản hay cao độ giọng nói.

Hiện nay ở Việt Nam có rất ít nghiên cứu về tần số cơ bản nói chung cũng như việc áp dụng tần số cơ bản trong nhận dạng tiếng nói. Luận văn này có phạm vi nghiên cứu phân biệt với 02 ngôn ngữ nói là tiếng Việt và tiếng Pháp. Tiếng Việt là một ngôn ngữ có thanh điệu, do đó tần số cơ bản của nó thay đổi rất nhiều trong một âm tiết cũng như từ âm tiết này sang âm tiết khác. Tiếng Pháp là một ngôn ngữ có trọng âm, do đó tần số cơ bản của nó thay đổi không nhiều từ âm tiết này đến âm

tiết khác. Chúng ta sử dụng các đặc điểm biến đổi tần số cơ bản để phân biệt các ngôn ngữ.

3. Mục đích nghiên cứu

Mục đích của đề tài Nghiên cứu nhận dạng ngôn ngữ nói tự động dựa trên tần số cơ bản trước tiên là để rèn luyện phương pháp và khả năng nghiên cứu, sau đó tìm hiểu về cơ quan cấu âm của con người, nghiên cứu tần số tiếng nói cơ bản, nghiên cứu một số thuật toán phân tích và xử lý tiếng nói, áp dụng vào một bài toán cụ thể. Đây là những nghiên cứu bước đầu về tần số cơ bản để áp dụng vào thực tiễn tại đơn vị công tác .

4. Đối tượng và phạm vi nghiên cứu

- Bài toán nhận dạng ngôn ngữ nói tự động dựa trên tần số cơ bản và các vấn đề liên quan. Cụ thể phân biệt ngôn ngữ tiếng Việt và tiếng Pháp.

- Các thuật toán, phương pháp phân tích và xử lý tiếng nói.

- Dữ liệu tiếng nói tiếng Việt trong tàng thư tiếng nói tại Viện Khoa học hình sự - Bộ Công an và dữ liệu tiếng nói tiếng Pháp trên Internet.

5. Phương pháp nghiên cứu

- Nghiên cứu lý thuyết.

- Thực nghiệm và phân tích kết quả.

6. Cấu trúc của luận văn

Luận văn ngoài phần mở đầu và kết luận gồm 3 chương chính:

- Chương 1: Tổng quan về bài toán nhận dạng ngôn ngữ nói tự động dựa trên tần số cơ bản.

- Chương 2: Thuật toán và mô hình hệ thống nhận dạng ngôn ngữ nói tự động dựa trên tần số cơ bản.

- Chương 3: Ứng dụng.

Trong đó, luận văn tập trung vào chương 2 và chương 3 với mục đích nghiên cứu tần số cơ bản để nhận dạng ngôn ngữ nói tiếng Việt và tiếng Pháp, sau đó thực nghiệm nhằm đánh giá mô hình này. Mặc dù có nhiều cố gắng nhưng do thời gian có hạn. Luận văn chắc chắn còn nhưng hạn chế, khiếm khuyết. Kính mong các thầy cô và đồng nghiệp thông cảm và góp ý. Xin trân trọng cảm ơn!

CHƯƠNG 1 - TỔNG QUAN VỀ BÀI TOÁN NHẬN DẠNG NGÔN NGỮ NÓI TỰ ĐỘNG DỰA TRÊN TẦN SỐ CƠ BẢN

Để có thể nghiên cứu nhận dạng ngôn ngữ tự động dựa trên tần số cơ bản nói chung và ứng dụng tần số cơ bản để phân biệt tiếng Việt và tiếng Pháp nói riêng, trước hết chúng ta cần phải rõ các khái niệm về âm thanh, các đặc trưng của tiếng nói và đặc điểm của ngôn ngữ tiếng Việt và tiếng Pháp.

1.1 Tổng quan về tiếng nói và các đặc trưng của tiếng nói

1.1.1 Nguồn gốc của âm thanh

Âm thanh là do vật thể rung động, phát ra tiếng ra tiếng và lan truyền đi trong không khí. Sở dĩ tai ta nghe được âm thanh là nhờ có màng nhĩ. Màng nhĩ nối liền với hệ thống thần kinh.

Làn sóng âm thanh từ vật thể rung động phát ra, được lan truyền trong không khí, tới tai ta làm rung động màng nhĩ theo đúng nhịp điệu rung động của vật thể đã phát ra tiếng. Nhờ đó, tai ta nghe được âm thanh. Không khí là môi trường truyền dẫn âm thanh, tuy nhiên, không phải tất cả các âm thanh đều được con người thu nhận mà chỉ những âm thanh có tần số trong một phạm vi nhất định. Như vậy bản chất âm thanh là một dao động có tần số, con người có thể cảm nhận được từ dao động này. Nếu dao động có biên độ càng lớn thì âm lượng càng lớn và ngược lại. Tần số dao động của các âm thanh trong tự nhiên có phạm vi rộng, tuy nhiên con người chỉ cảm nhận trong một phạm vi nhất định.

Âm thanh được lan truyền trong các chất khí, lỏng, rắn… nhưng không lan truyền được trong khoảng chân không. Một số chất truyền dẫn âm kém. Các chất dẫn âm kém thường là loại mềm, xốp như bong, dạ, cỏ khô. Các chất này gọi là chất hút âm, được dùng lót tường các rạp hát, phòng cách âm… để giảm tiếng vang.

Vận tốc truyền lan của âm thanh phụ thuộc vào chất truyền âm, ví dụ tốc độ truyền âm trong không khí là 340 m/s, trong nước là 1480 m/s, trong sắt là 5000 m/s. Trong quá trình truyền lan, nếu gặp phải các vật chướng ngại như tường, núi

đá,… thì phần lớn năng lượng của âm thanh sẽ bị phản xạ trở lại, một phần nhỏ tiếp tục truyền lan về phía trước. Còn một phần nhỏ nữa của năng lượng âm thanh bị cọ sát với vật chướng ngại biến thành nhiệt năng tiêu tan đi.

1.1.2 Bộ máy phát âm

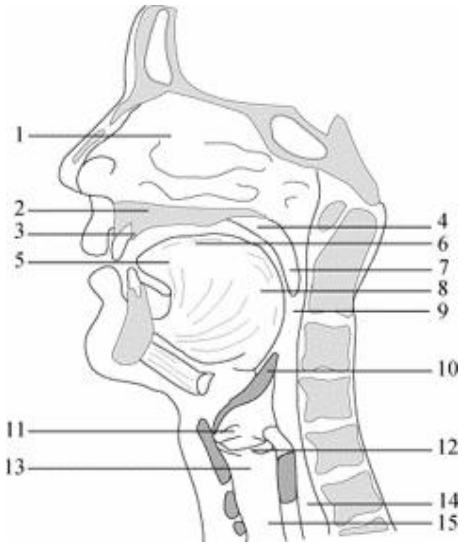
Bộ máy phát âm của con người bao gồm các thành phần riêng rẽ như phổi, khí quản, thanh quản và các đường dẫn miệng, mũi. Trong đó:

- Thanh quản chứa hai dây thanh có thể dao động tạo ra sự cộng hưởng cần thiết để tạo ra âm thanh.

- Tuyến âm là ống không đều bắt đầu từ môi, kết thúc bởi dây thanh hoặc thanh quản.

- Khoang mũi là ống không đều bắt đầu từ môi, kết thúc bởi vòm miệng, có độ dài cố định khoảng 12cm đối với người lớn.

- Vòm miệng là các nếp cơ chuyển động.



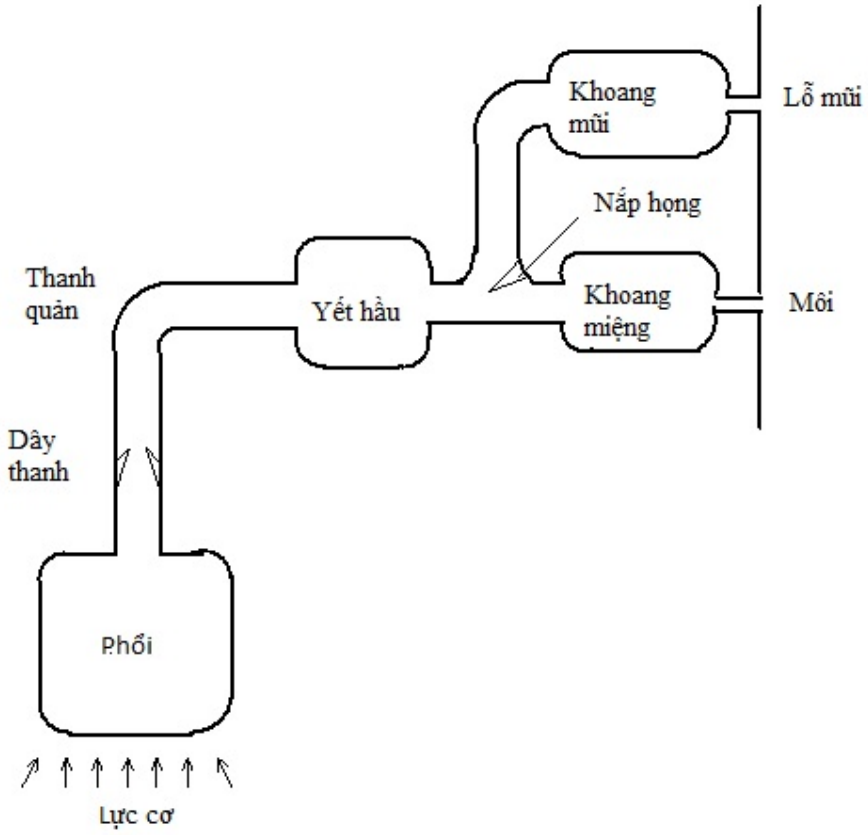
1. Hốc mũi
2. Vòm miệng trên
3. Ổ răng
4. Vòm miệng mềm
5. Đầu lưỡi
6. Thân lưỡi
7. Lưỡi gà
8. Cơ miệng
9. Yết hầu
10. Nắp đóng của thanh quản
11. Dây thanh giả
12. Dây thanh
13. Thanh quản
14. Thực quản

Hình 1.1 Sơ đồ bộ máy phát âm của con người.

1.1.3 Cơ chế phát âm

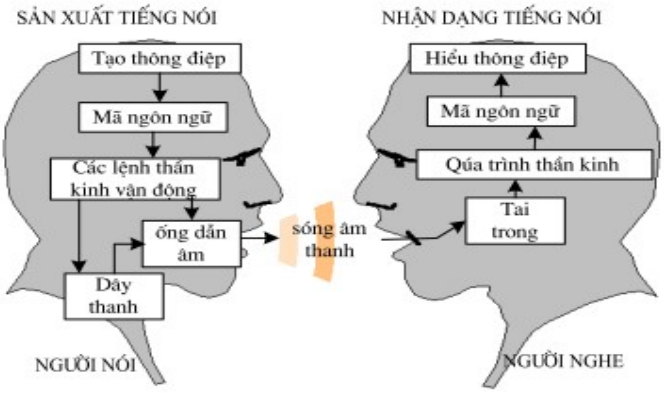
Trong quá trình tạo âm thanh không phải là âm mũi, vòm miệng mở, khoang mũi đóng lại, dòng khí sẽ chỉ đi qua khoang mũi. Khi phát âm mũi, vòm miệng hạ thấp và dòng khí sẽ chỉ đi qua khoang mũi.

Tuyến âm sẽ được kích thích bởi nguồn năng lượng chính tại thanh môn. Tiếng nói được tạo ra do tín hiệu nguồn từ thanh môn phát ra, đẩy không khí có trong phổi lên tạo thành dòng khí, va chạm vào hai dây thanh trong tuyến âm. Hai dây thanh dao động sẽ tạo ra cộng hưởng, dao động âm sẽ được lan truyền theo tuyến âm (tính từ tuyến âm đến khoang miệng) và sau khi đi qua khoang mũi và môi, sẽ tạo ra tiếng nói. Các âm thanh khác nhau được tạo ra khi các cơ hoạt động để thay đổi hình dạng của dây thanh âm, và do đó thay đổi tần số cộng hưởng của nó, hoặc tần số định dạng. Tốc độ của các xung được gọi là tần số cơ bản hoặc cao độ. Cơ chế sản xuất giọng nói được mô tả trong hình 1.2.



Hình 1.2 Sơ đồ cơ chế phát âm

1.1.4 Quá trình sản xuất tiếng nói và thu nhận tiếng nói



Hình 1.3 Sơ đồ biểu diễn quá trình sản xuất thu nhận tiếng nói của con người.

Quá trình sản xuất tiếng nói bắt đầu từ khi người nói tạo ra một thông điệp (trong ý nghĩ của người nói) và muốn chuyển tải nó cho người nghe thông qua tiếng nói. Tổ chức thần kinh tương ứng chịu trách nhiệm tạo ra thông điệp dưới dạng văn bản biểu diễn các từ của thông điệp. Bước tiếp theo của quá trình là chuyển đổi thông điệp sang dạng một mã ngôn ngữ. Điều này gần như tương đương với việc chuyển đổi các biểu diễn văn bản của thông điệp thành một chuỗi các âm vị tương ứng với những âm thanh tạo nên các từ; Đồng thời với việc ghi nhận âm điệu nhằm xác định sự kéo dài, sự nhấn mạnh, và trọng âm cao thấp của âm thanh. Khi một mã ngôn ngữ được lựa chọn, người nói phải thực hiện một loạt các lệnh thần kinh vận động để làm cho các dây thanh dao động, đồng thời cấu trúc hình dạng ống dẫn âm thanh phát ra một chuỗi các âm thanh. Như vậy, đầu ra cuối cùng của quá trình là một tín hiệu âm học. Các lệnh thần kinh vận động phải điều khiển một cách đồng bộ tất cả các khâu vận động như sự hoạt động của môi, hàm, lưỡi,…

Khi tín hiệu tiếng nói đã được sinh ra và được truyền cho người nghe, quá trình thu nhận tiếng nói (hay nhận dạng tiếng nói) bắt đầu. Đầu tiên, người nghe xử lý tín hiệu âm thanh thông qua màng nền của tai trong, nó có khả năng cung cấp một phân tích phổ cho tín hiệu tới. Một quá trình xử lý thần kinh chuyển đổi tín hiệu phổ tại đầu ra của màng nền thành các tín hiệu hoạt động đối với thần kinh

thính giác, có thể coi đây như một quá trình lấy ra các đặc trưng. Bằng một phương pháp đặc biệt các tín hiệu hoạt động đi qua hệ thần kinh thính giác được chuyển đổi thành một mã ngôn ngữ cho những trung tâm xử lý cấp cao hơn bên trong bộ não, và cuối cùng là việc hiểu được nội dung thông điệp.

Từ sự minh họa quá trình nhận dạng tiếng nói thông qua hệ thống thần kinh con người như trên, chúng ta có thể có một chút ý niệm về khả năng ứng dụng mạng neuron nhân tạo trong việc mô phỏng một số tổ chức thần kinh như một phần của hệ thần kinh thính giác.

1.1.5 Đặc tính âm học của tiếng nói

a. Âm hữu thanh

Âm hữu thanh được tạo ra từ các dây thanh bị căng đồng thời và chúng rung động ở chế độ dãn khi không khí tăng lên làm thanh môn mở ra và sau đó thanh môn xẹp xuống do không khí chạy qua.

Do sự cộng hưởng của dây thanh, sóng âm tạo ra có dạng tuần hoàn hoặc gần như tuần hoàn. Phổ của âm hữu thanh có nhiều thành phần hài tại giá trị bội số của tần số cộng hưởng, còn gọi là tần số cơ bản (pitch).

b. Âm vô thanh

Khi tạo ra âm vô thanh dây thanh không cộng hưởng. Âm vô thanh có hai loại cơ bản là âm xát và âm tắc.

Âm xát (ví dụ như âm s) được tạo ra khi có sự co thắt tại vài điểm trong tuyến âm. Không khí khi đi qua điểm co thắt sẽ chuyển thành chuyển động hỗn loạn tạo nên kích thích giống như nhiễu ngẫu nhiên. Thông thường điểm co thắt xảy ra gần miệng nên sự cộng hưởng của tuyến âm ảnh hưởng rất ít đến đặc tính của âm xát được tạo ra.

Âm tắc (ví dụ như âm p) được tạo ra khi tuyến âm đóng tại một số điểm làm cho áp suất không khí tăng lên và sau đó được giải phóng đột ngột. Sự giải phóng đột ngột này tạo ra kích thích nhất thời của tuyến âm. Sự kích thích này có thể xảy ra với sự cộng hưởng hoặc không cộng hưởng của dây thanh tương ứng với âm tắc hữu thanh hoặc vô thanh.

c. Âm vị

Tín hiệu tiếng nói là tín hiệu tương tự biểu diễn cho thông tin về mặt ngôn ngữ và được mô tả bởi các âm vị khác nhau. Như vậy, âm vị là đơn vị nhỏ nhất của ngôn ngữ. Tuỳ theo từng ngôn ngữ cụ thể mà số lượng các âm vị nhiều hay ít (thông thường số lượng các âm vị vào khoảng 20 - 30). Các âm vị được chia thành hai loại: nguyên âm và phụ âm.

- Nguyên âm là âm hữu thanh được tạo ra bằng sự cộng hưởng của dây thanh khi dòng khí được thanh môn đẩy lên. Khoang miệng được tạo lập thành nhiều hình dạng nhất định tạo thành các nguyên âm khác nhau. Số lượng các nguyên âm phụ thuộc vào từng ngôn ngữ nhất định.

- Phụ âm được tạo ra bởi các dòng khí hỗn loạn được phát ra gần những điểm co thắt của đường dẫn âm thanh do cách phát âm tạo thành. Phụ âm có đặc tính hữu thanh hay vô thanh tuỳ thuộc vào việc dây thanh có dao động để tạo nên cộng hưởng không. Dòng không khí tại chỗ đóng của vòm miệng tạo ra phụ âm tắc. Phụ âm xát được phát ra từ chỗ co thắt lớn nhất.

1.1.6 Các đặc tính khác

a. Tỷ suất thời gian

Trong khi nói chuyện, khoảng thời gian nói và khoảng thời gian nghỉ xen kẽ nhau. Tỷ lệ % thời gian nói trên tổng số thời gian nói và nghỉ được gọi là tỷ suất thời gian. Giá trị này biến đổi tuỳ thuộc vào tốc độ nói và từ đó ta có thể phân loại thành nói nhanh, nói chậm hay nói bình thường.

b. Tần số lấy mẫu

Bản chất của âm thanh là các sóng âm. Đây là tín hiệu tương tự. Để có thể biểu diễn âm thanh trong máy tính và áp dụng kĩ thuật xử lý tín hiệu số thì bước đầu tiên là phải chuyển đổi các tín hiệu tương tự thành các dãy số. Quá trình này được thể hiện bằng cách lấy mẫu tín hiệu âm thanh theo chu kỳ (được gọi là chu kỳ lấy mẫu).

Với tín hiệu tương tự x(t), chu kỳ lấy mẫu T (tần số lấy mẫu 1/T) thu được dãy số X(n): $X(n) = x(n*T)$ *với* $-\infty < n < \infty$

Để đảm bảo quá trình số hóa không làm mất mát thông tin của phổ tín hiệu thì tần số lấy mẫu Fs = 1/T phải đủ lớn. Giá trị đủ lớn của Fs phải tuân theo định lý lấy mẫu: Tín hiệu liên tục theo thời gian có bề rộng phổ hữu hạn với tần số cao nhất f Hz có thể được khôi phục một cách duy nhất từ các mẫu nếu quá trình lấy mẫu thực hiện với tốc độ Fs >= 2f mẫu trên một giây. Đối chuẩn của file âm thanh thì tần số lấy mẫu thấp nhất là 800 Hz điều này nghĩa là quá trình số hóa chỉ được áp dụng với tín hiệu tương tự có tần số cao nhất là 4000 Hz phù hợp với tiếng nói con người có tần số từ 40 Hz - 4000 Hz.

c. Formant

Formant hay còn gọi là các họa âm, đóng vai trò tạo nên âm sắc của âm thanh. Formant là giải tần số được tăng cường do hiện tượng cổng hưởng, đặc trưng cho âm sắc của mỗi nguyên âm. Trong mỗi dải tần như thế có một tần số được tăng cường hơn cả và được gọi là đỉnh của formant, một nguyên âm do người phát ra có nhiều formant, trong đó có 2 formant tương ứng với hộp cổng hưởng miệng và hộp cộng hưởng yết hầu, các formant khác đặc trưng cho giọng nói của từng người.

Với phổ của tín hiệu tiếng nói, mỗi đỉnh có biên độ lớn nhất xét trong một khoảng nào đó (cực đại khu vực) tương ứng với một formant. Ngoài tần số, formant còn được xác định bởi biên độ và dải thông. Về mặt vật lý các formant tương ứng với các tần số cộng hưởng của tuyến âm. Trong xử lý tiếng nói và nhất là trong tổng hợp tiếng nói, để mô phỏng lại tuyến âm người ta phải xác định được các tham số formant đối với từng loại âm vị, do đó việc đánh giá, ước lượng các formant có ý nghĩa rất quan trọng.

Tần số formant biến đổi trong một khoảng rộng phụ thuộc vào giới tính của người nói và phụ thuộc vào các dạng âm vị tương ứng với formant đó. Đồng thời, formant còn phụ thuộc các âm vị trước và sau đó. Về cấu trúc tự nhiên, tần số formant có liên hệ chặt chẽ với hình dạng và kích thước tuyến âm. Thông thường phổ của tín hiệu tiếng nói có khoảng 5 formant nhưng chỉ có 3 formant đầu tiên ảnh hưởng quan trọng đến các đặc tính của các âm vị, các formant còn lại cũng có ảnh hưởng song rất ít.

Tần số formant đặc trưng cho các nguyên âm biến đổi tuỳ thuộc vào người nói trong điều kiện phát âm nhất định. Mặc dù phạm vi của các tần số formant tương ứng với mỗi nguyên âm có thể trùm lên nhau nhưng vị trí giữa các formant là không đổi vì sự xê dịch của các formant là song song.

d. Tần số cơ bản

Sóng âm do con người phát ra rất phức tạp. Nó có dạng đường cong phức tạp có chu kỳ. Khi phát ra một âm có tần số F0 thì đồng thời cũng phát ra âm có tần số 2F0, 3F0, 4F0, ... Âm có tần số F0 được gọi là âm cơ bản, tần số F0 được gọi là tấn số cơ bản các âm khác được gọi là các họa âm (Formant) thứ nhất, họa âm thứ 2... Âm cuối cùng (âm nghe được) là âm tổng hợp của âm cơ bản và các họa âm. Do đó đường biểu diễn của nó là một đường cong phức tạp có tần số là tần số cơ bản.

Đối với những người nói khác nhau, tần số cơ bản cũng khác nhau. Dưới đây là một số giá trị tần số cơ bản tương ứng với giới tính và tuổi:

Bảng 1.1 Bảng giá trị tần số cơ bản

Giá trị tần số cơ bản	Người nói
80 - 200 Hz	Nam giới
150 - 450 Hz	Phụ nữ
200 - 600 Hz	Trẻ em

e. Chu kỳ cao độ (Pitch)

- Định nghĩa 1: Chu kỳ cao độ của tín hiệu tiếng nói là thời gian trôi qua giữa hai xung thanh môn liên tiếp. Việc đo bắt đầu ở một thời điểm xác định trong một chu trình thanh môn, tốt nhất ở thời điểm đóng thanh môn hay nếu thanh môn không đóng hoàn toàn thì ở điểm mà diện tích thanh môn nhỏ nhất. Thuật toán phát hiện cao độ của theo định nghĩa này. Nếu chỉ căn cứ vào giá trị tức thời của chu kỳ cao độ để xác định đường vận động chu kỳ cao độ mà không loại bỏ các chu kỳ cao độ bị biến dạng thì đường này sẽ không trơn nên nhận dạng sẽ kém chính xác.

- Định nghĩa 2: Chu kỳ cao độ là độ dài trung bình của một vài chu kỳ, là thời gian trôi qua trung bình của một số ít chu trình kích thích liên tiếp. Xác định giá trị trung bình như thế nào và trên bao nhiêu chu kỳ phụ thuộc từng phương pháp

trích chu kỳ cao độ. Các thuật toán xác định chu kỳ cao độ trung bình theo hàm tự tương quan, hàm hiệu biên độ trung bình theo định nghĩa này.

g. Biên độ

Biên độ là một đặc trương quan trọng của sóng âm. Sóng âm thanh khi thu vào máy tính khi được thu vào máy tính sẽ được số hóa thành một chuỗi các số rời rạc với miền giá trị tùy theo độ phân giải. Độ phân giải được hiểu theo nghĩa là số bit được dùng để lưu trữ một mẫu thu được trong quá trình lấy mẫu. Với độ phân giải 8 bit, được gọi là âm thanh mono, miền giá trị của mẫu là khoảng đóng [0,255]; độ phân giải 16 bit (stereo), miền giá trị này là khoảng đóng [0,65535]. Do đó xác định chính xác biên độ của sóng là một bài toán khó và trên thực tế không phải giải quết tuyệt đối chính xác vì cái mà ta cần thực sự quan tâm là sự biến thiên của biên độ. Do đó bài toán này thường được giải quyết bằng bài toán gần đúng. Trước hết ta xác định ngưỡng gần đúng ngưỡng không, sau đó biên độ sẽ được tính bằng trị tuyệt đối hiệu giá trị số hóa trừ đi giá trị ngưỡng không. Giá trị ngưỡng không tùy thuộc vào từng SoundCard.

h. Nhiễu

Nhiễu là một trong các yếu tố làm cho bài toán nhận dạng trở lên vô cùng phức tạp. Đại lượng nhiễu được xem như một đại lượng ngẫu nhiên, làm biến đổi tín hiệu cần nhận dạng. Do đó lọc nhiễu là một khâu cần thiết phải tiến hành trong quá trình xử lý tín hiệu.

1.2 Đặc điểm của ngôn ngữ tiếng Việt và tiếng Pháp

Nguyên tắc của hệ thống nhận dạng ngôn ngữ tự động dựa trên các tính năng của một ngôn ngữ cụ thể. Đặc trưng của giọng nói cũng như phát âm khác nhau từ ngôn ngữ này sang ngôn ngữ khác. Do đó, để xây dựng một hệ thống nhận dạng ngôn ngữ tự động cho một số ngôn ngữ nhất định, chúng ta phải nắm vững các đặc trưng của các ngôn ngữ đó. Bằng cách nắm vững các đặc trưng của ngôn ngữ, chúng ta có thể rút ra các đặc điểm cụ thể của từng ngôn ngữ hữu ích cho việc giải quyết vấn đề nhận dạng ngôn ngữ.

1.2.1 Đặc điểm của ngôn ngữ tiếng Việt

Tiếng Việt là ngôn ngữ sử dụng các ký tự Latinh và nó có một số đặc điểm như sau:

- Tiếng Việt là một ngôn ngữ đơn âm tiết.

- Tiếng Việt là một ngôn ngữ ngữ điệu.

- Trong tiếng Việt, âm tiết là những đơn vị nhỏ nhất mang thông tin có ý nghĩa.

- Trong tiếng Việt, không có sự thay đổi về âm tiết trong thì, giới tính, số ít hoặc số nhiều. (Ví dụ, trong tiếng Pháp có danh từ chỉ giống đực và danh từ chỉ giống cái: étudiant - étudiante, nouveau - nouvelle, danh từ số ít và danh từ số nhiều: amie - amies).

- Cấu trúc từ trong tiếng Việt không có các phụ tố (tiền tố, hậu tố, trung tố). Ví dụ trong tiếng Anh hay tiếng Pháp, các từ trái nghĩa được tạo ra bằng cách thêm tiền tố "im-", "ir-", "un-": impolite, unreadable, irregular….

- Mỗi âm tiết có âm điệu riêng.

- Hệ thống âm điệu tiếng việt Việt gồm sáu âm.

- Một âm tiết tiếng Việt có cấu trúc C-V.

a. Cấu trúc của tiếng Việt

Một âm tiết tiếng Việt được kết hợp chặt chẽ bởi ba thành phần chính với các mức độ độc lập khác nhau; đó là các phụ âm chính, vần và thanh điệu. Trong đó phần vần lại được chia thành 3 phần nhỏ hơn là nguyên âm chính, nguyên âm đệm và phần cuối cùng. Các thành phần của một âm tiết được trình bày trong bảng 1.2.

Bảng 1.2 Sơ đồ tiếng Việt

<table>
<tr><td colspan="4">Thanh</td></tr>
<tr><td rowspan="2">PHỤ ÂM</td><td colspan="3">VẦN</td></tr>
<tr><td>Âm đệm</td><td>Âm chính</td><td>Âm cuối</td></tr>
</table>

Ví dụ âm tiết "Toán" trong đó phần trước là /t/, vần là /oan/. Trong vần /oan/, âm đệm là /o/, nguyên âm chính là /a/, và phần âm cuối là /n/ và thanh sắc.

b. Hệ thống âm đầu

Tiếng Việt có 22 phụ âm đầu bao gồm: /b, m, f, v, t, t', d, n, z, ʐ, s, ʂ, c, ʈ, ɲ, l, k, χ, ŋ, ɣ, h, ʔ/. Các tiền tố phụ âm được phân biệt như là phụ âm xát, dừng và âm mũi. Do đó, chúng ta có thể phân loại tiền tố phụ âm tiếng Việt thành các loại như mô tả trong bảng 1.3

Bảng 1.3 Bảng hệ thống âm đầu tiếng Việt

Phương thức \ Vị trí				Môi	Đầu lưỡi		Mặt lưỡi	Gốc lưỡi	Thanh hầu
					Bọt	Lưỡi			
Tắc	Ồn	Bật hơi			t'				
		Không bật hơi	Vô thanh		t	ʈ	c	k	ʔ
			Hữu thanh	b	d				
	Vang			m	n		ɲ	ŋ	
Xát	Ồn	Vô thanh		f	s	ʂ		χ	h
		Hữu thanh		v	z	ʐ		ɣ	
	Vang				l				

c. Hệ thống âm đệm

Âm đệm /w/ có chức năng làm trầm hoá âm sắc của âm tiết.

d. Hệ thống âm chính

Tiếng Việt có 13 nguyên âm đơn và 3 nguyên âm đôi làm âm chính: /i, e, ɛ, ɤ, ɤ̆, a, ɯ, ă, u, o, ɔ, ɔ̆, ɛ̆, ie, ɯɤ, uo/

Bảng 1.4 Bảng hệ thống âm nguyên âm tiếng Việt

Âm sắc	Vị trí lưỡi, hình dáng môi / Độ mở của miệng	Trước, không tròn môi	Sau	
			Không tròn môi	Tròn môi
Cố định	Nhỏ	i	ɯ	u
	Lớn vừa	e	ɤ/ɤ̆	o
	Lớn	ɛ/ɛ̆	a/ă	ɔ/ɔ̆
Không cố định		i̯e	ɯ̯ɤ̯	u̯o

e. Hệ thống âm cuối

Hệ thống âm cuối tiếng Việt có 6 phụ âm /m, n, ŋ, p, t, k/ và hai bán nguyên âm /-w, -j/.

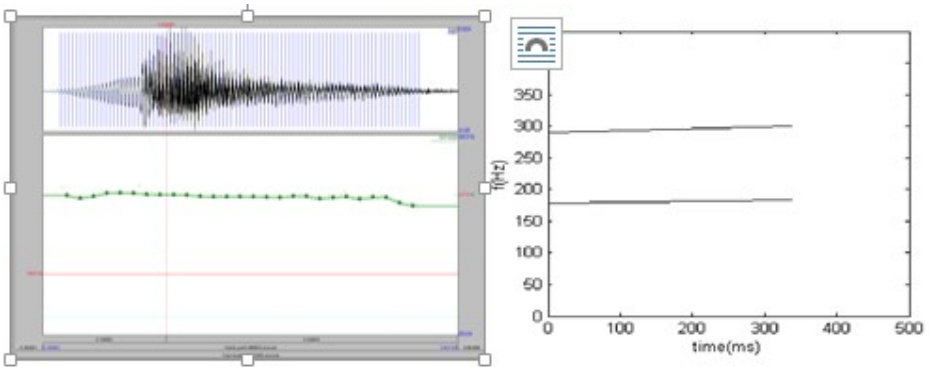
Bảng 1.5 Bảng hệ thống âm cuối tiếng Việt

Phương thức \ Vị trí		Môi	Lưỡi	
			Đầu lưỡi	Gốc lưỡi
Ốn		p	t	k
Vang	Mũi	m	n	ŋ
	Không mũi	-w	-j	

g. Hệ thống thanh điệu

Tiếng Việt là ngôn ngữ có thanh điệu, ngữ nghĩa của một từ phụ thuộc vào thanh điệu, khi thanh điệu thay đổi, nghĩa của từ cũng thay đổi theo. Ở cấp độ vật lý, thanh điệu là đường cong của tần số cơ bản (F0), tương ứng với mỗi thanh điệu, tần số cơ bản thay đổi theo một quy luật riêng. Hệ thống thanh của tiếng Việt tương đối phức tạp. Nó thay đổi theo từng vùng miền. Số lượng các thanh có thể thay đổi từ 6 (giọng Hà Nội) đến 5 (giọng Thành phố Hồ Chí Minh) hoặc đến 4 (giọng miền Trung). Bởi vì giọng Hà Nội được coi là phương ngữ chuẩn của Việt Nam, nên phần sau ta sẽ chỉ quan tâm đến các thuộc tính của giọng Hà Nội. Tiếng Việt có 6 thanh điệu được phân thành hai nhóm: nhóm có quãng âm cao (ngang, ngã, sắc) và nhóm có quãng âm thấp (huyền, hỏi, nặng).

- Thanh "ngang": Đây là một thanh cao. Điểm bắt đầu đường F0 của thanh này cao hơn các thanh khác, dáng điệu đường F0 của thanh này là thẳng và ổn định.

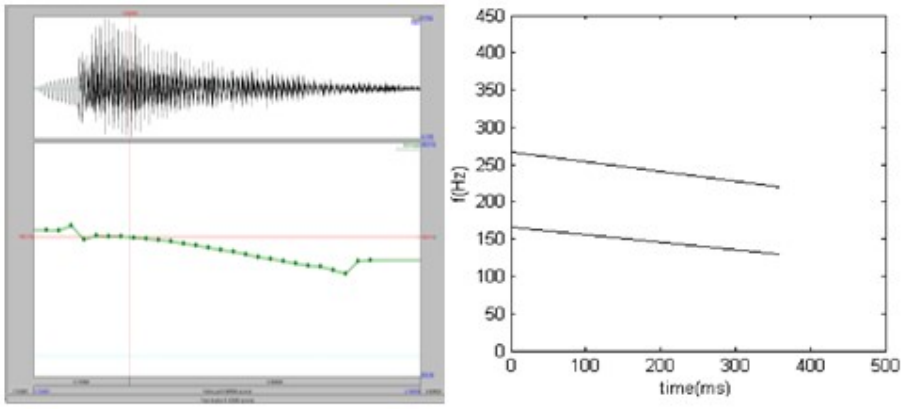


Hình 1.4 Dáng điệu đường F0 của thanh "ngang"

Một ví dụ về dáng điệu của đường F0 của âm tiết /ba/ với thanh ngang được mô tả trong Hình 1.3. Hai đường trong hình bên phải thể hiện đường ngữ điệu của hai giọng nữ cao nhất và thấp nhất. Nếu gọi F0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu huyền có thể được mổ tả như sau:

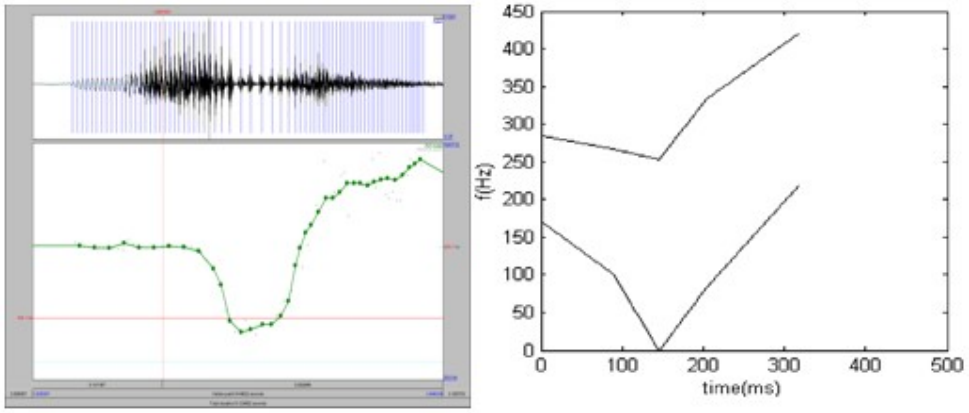
F0, F0-10, F0-20, F0-30, F0-40, F0-50, F0-60

- Thanh "huyền": Điểm bắt đầu của thanh này thấp hơn so với của thanh "ngang". Dáng điệu đường F0 chung của thanh này giảm dần đến cuối âm tiết.



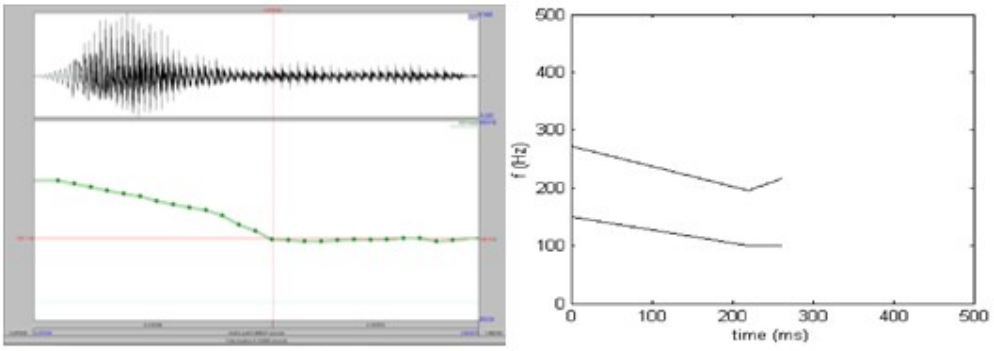
Hình 1.5 Dáng điệu đường F0 của thanh "huyền"

- Thanh "ngã": Giá trị bắt đầu của thanh ngã cao hơn của thanh "huyền". Đoạn giữa của thanh ngã bị gãy là do có sự di chuyển co thắt thanh môn. Dáng điệu đường F0 chung của thanh này thấp hơn ở giữa và sau đó tăng lên ở cuối.



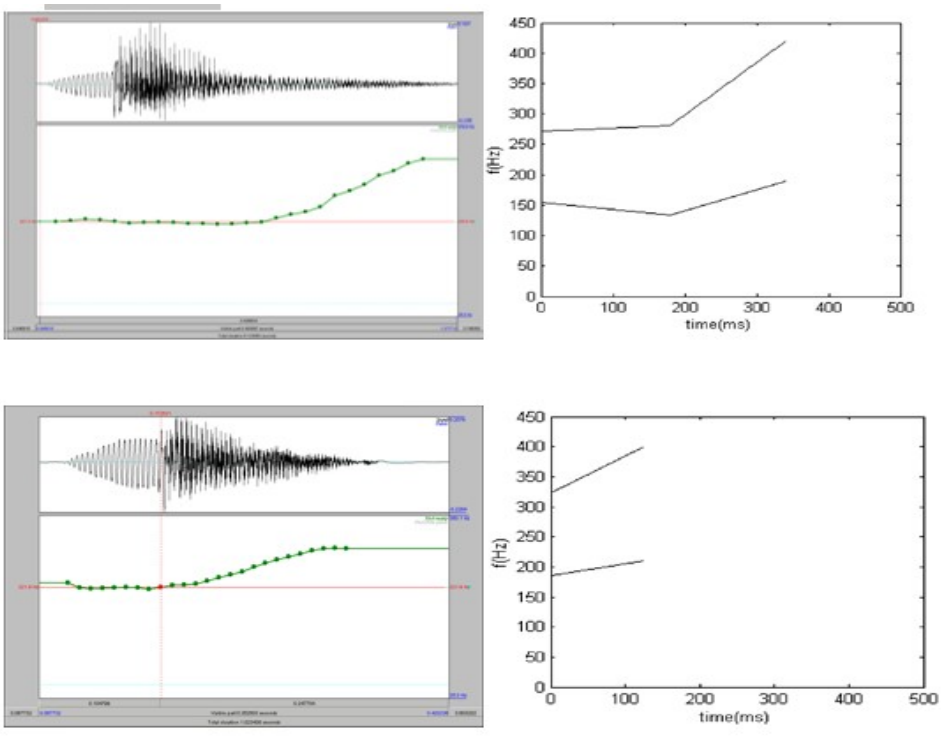
Hình 1.6 Dáng điệu đường F0 của thanh "ngã"

- Thanh “hỏi”: Giá trị khởi đầu F0 của thanh hỏi là nhỏ nhất trong 6 thanh. Giá trị F0 giảm dần dần cho đến hơn 2/3 âm tiết, sau đó bắt đầu tăng trở lại cho đến cuối âm tiết.



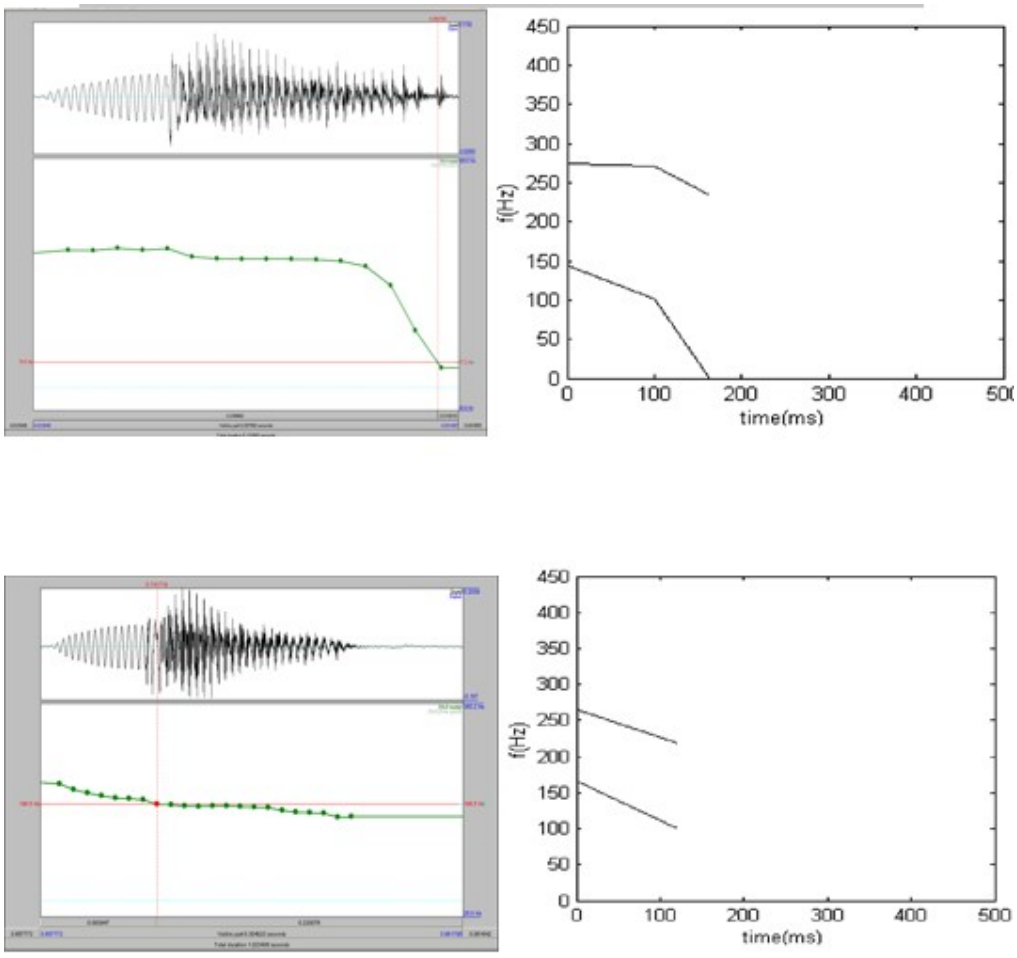
Hình 1.7 Dáng điệu đường F0 của thanh “hỏi”

- Thanh “sắc”: Điểm bắt đầu của F0 là cao, thanh sắc có 2 dạng khác nhau trong các âm tiết mở và trong các âm tiết đóng. Dáng điệu đường F0 chung của thanh này giữ ổn định từ đầu đến giữa, và sau đó tăng lên ở cuối.



Hình 1.8 Dáng điệu đường F0 của thanh “sắc”

- Thanh “nặng”: Dáng điệu đường F0 chung của thanh này giảm mạnh ở cuối đối với các âm tiết mở. Đối với các âm tiết đóng, đường viền cao độ chung của những âm này ổn định ở âm thấp.



Hình 1.9 Dáng điệu đường F0 của thanh “nặng”

1.2.2 Đặc điểm của ngôn ngữ tiếng Pháp

a. Một vài đặc trưng của tiếng Pháp

Các từ tiếng Pháp chỉ người, địa điểm và sự vật (danh từ) được phân loại là giống đực hoặc giống cái. Nói chung, tính từ được sử dụng để mô tả các từ giống cái kết thúc bằng *e*.

Le (hình thức giống cái của *the*) được sử dụng với các từ giống đực. *La* (hình thức giống cái của *the*) được sử dụng với các từ giống cái. Nhưng *l’*, được sử dụng với một trong hai nếu từ bắt đầu bằng một nguyên âm. Ví dụ, từ *enfant* có nghĩa là trẻ

em hoặc trẻ sơ sinh, giống đực hoặc giống cái. Nhưng *l'enfant est né* (đứa trẻ được sinh ra) được sử dụng với một đứa trẻ nam, và *l'enfant est née* với một đứa trẻ nữ.

b. Các cấu trúc đa dạng của từ

Trong tiếng Pháp, hình thức của một số từ sẽ thay đổi tùy theo cách chúng được sử dụng trong một câu. Danh sách từ này cung cấp các hình thức tiêu chuẩn của mỗi từ tiếng Pháp. Khi bạn đọc đoạn ghi âm tiếng Pháp, bạn sẽ cần lưu ý rằng một số từ thay đổi theo cách sử dụng.

Các dạng số nhiều của các từ tiếng Pháp thường được tạo bằng cách thêm *s* hoặc *x* vào các từ số ít. Do đó *frère* trở thành *frères*, và *beau* trở thành *beaux*. Số nhiều của *beau-frère* (anh rể) là *beaux-frères* (anh rể).

Trong tiếng Pháp có năm dấu phụ (dấu). Chúng được đặt trên các nguyên âm hoặc dưới chữ c để chỉ ra sự thay đổi trong cách phát âm. Các dấu phụ sau đây được sử dụng trong tiếng Pháp: *à, â, é, è, ê, ë, î, ï, ô, ö, û, ù và ç*. Các *ç* được phát âm như là một *s*. Những dấu phụ này không ảnh hưởng đến trật tự chữ cái

c. Đặc tính [±clitic]

Một từ (hoặc một âm tiết) là nhấn âm hoặc không nhấn âm tùy thuộc vào các thuộc tính từ vựng hoặc hình thái. Những từ nhấn âm được cho là mang trọng âm của từ, nhưng đây thực sự chỉ là một trọng âm tiềm tàng vì những từ có trọng âm không cần phải luôn luôn được nhấn mạnh. Từ ngữ không có trọng âm được tổ chức lại xung quanh những từ có trọng âm. Đơn vị kết quả, nhóm trọng âm (SG), vẫn quan tâm đến trọng âm thật vì lý do nêu trên. Một nhóm nhịp điệu (IG) có được khi người nói chọn một chuỗi âm cụ thể (từ chuỗi được cho phép bởi ngữ pháp ngữ điệu) và kết hợp nó với một phần của chuỗi phân đoạn tương ứng với một hoặc nhiều nhóm trọng âm liền kề. Trong khi việc sử dụng chuỗi âm tạo ra sự hình thành nhóm nhịp điệu, sự lựa chọn thực tế của âm được nhấn mạnh (trong số các âm khác từ vị trí AF) sẽ xác định khả năng nhóm nhịp điệu này với các nhóm nhịp điệu liền kề trong chuỗi. Chúng ta sử dụng gói thuật ngữ để chỉ một nhóm gồm một hoặc nhiều nhóm nhịp điệu được liên kết bởi cơ chế phân nhóm ngôn điệu. Các phần sau đây mô tả các đơn vị và quy tắc cho các cấp khác nhau, bắt đầu với cấp thấp hơn.

Khi một từ có thể bằng chính nhóm trọng âm của nó hoặc có thể cấu tạo thành trung tâm của một nhóm trọng âm và do đó trở thành một nhóm nhịp điệu, nó không phải là từ ngữ thông thường. Một số tiêu chí đã được đưa ra để xác định các lớp từ nhấn âm và không nhấn âm:

- Nhận dạng từ vựng: Đối với mỗi hình thái từ vựng, hãy nêu đặc trưng [±clitic] của nó (ví dụ: man man là [-clitic]).

- Thể loại ngữ pháp: ví dụ các danh từ là [-clitic].

- Chức năng cú pháp: ví dụ chủ đề là [-clitic]

- Danh mục cú pháp: ví dụ Cụm danh từ là [-clitic].

Trong tiếng Pháp, một ngôn ngữ có quy tắc trọng âm, vị trí trọng âm có thể dự đoán được từ các giới hạn của nhóm trọng âm; vì vậy không cần chỉ định âm tiết nào trong từ mang trọng âm. Điều này sẽ là cần thiết trong ngôn ngữ có trọng âm tự do. Vì lý do này, người ta có thể nói về các âm tiết [±clitic], trong đó một âm tiết có trọng âm thường là một âm tiết mang trọng âm trong một từ nhấn âm.

d. Thành lập nhóm trọng âm

Trong các ví dụ sau đây, các âm tiết không có trọng âm được biểu thị bằng các dấu chấm trên lớp "WS" (đối với trọng âm từ) và các âm tiết có trọng âm bằng kí hiệu o; trên lớp "SG", các ranh giới của SGs được biểu thị bằng dấu ngoặc và các âm tiết có trọng âm bằng dấu trừ.

Ví dụ (1): Tu le vois, cet admirable bateau? Vois-tu le problème?

WS • • o • • • o • o o • • • o
SG (• • -) (• • • -) (• -) (• -)(• • -)

Quy tắc 1: Một nhóm trọng âm được tạo thành từ một từ nhấn âm N1 và tất cả các từ không nhấn âm được chi phối bởi N1 (như trong *tu le vois* trong *vois-tu*) hoặc bởi một từ nhấn âm N2 khác lần lượt chi phối N1 (như trong *cet admirable*).

Như có thể thấy *vois-tu* trong ví dụ (1), vị trí trọng âm có thể chuyển từ một âm tiết nhấn âm sang một âm tiết không nhấn âm, ít nhất là trong các ngôn ngữ trọng âm bị ràng buộc. Chính hiện tượng này giải thích sự cần thiết của hai cấp độ (WS và SG) trong các ngôn ngữ đó.

Tuy nhiên, khi một từ không nhấn âm được tách từ một từ nhấn âm của mình bởi một hoặc nhiều thành phần không có quan hệ cú pháp, thì từ không nhấn âm đó cũng tạo thành một SG (như đối với *qui*).

Ví dụ (2): Rene', qui, tu le sais bien, ne nous aime pas tellement …

WS • o • • • o o • • • • • o
SG (• -) (-) (• • -) (•-) (• • -) (-) (• -)

e. Thành lập nhóm ngữ điệu

Ví dụ (3a-c) cho thấy các hình thức khác nhau của cùng một cách nói, với số lượng các nhóm ngữ điệu khác nhau (IGs). Các nhóm ngữ điệu được bao quanh bởi các dấu ngoặc trên lớp có nhãn là "IG". Dấu cộng biểu thị một âm tiết với âm điệu của mô hình giai điệu AF. Người nói có thể sắp đặt một nhóm ngữ điệu tới mỗi nhóm trọng âm (như trong ví dụ 3a) hoặc anh ta có thể kết hợp nhiều nhóm trọng âm vào một nhóm ngữ điệu (như trong ví dụ 3b), với điều kiện các ràng buộc cú pháp sau được đáp ứng; một SG nên được nhóm với SG, nó phụ thuộc vào cú pháp trước khi có thể được nhóm với bất kỳ SG nào khác. Vì vậy (ví dụ 3c) là sai vì *ainsi* bị chi phối bởi *voir* chứ không phải bởi *attriste*.

Ví dụ (3): de la voir ainsi m'attriste beaucoup

WS • • o • o • o • o
SG (• • -) (• -) (• -) (• -)
(a) IG (• • +) (• +) (• +) (• +)
(b) IG (• • • • +) (• • • +)
(c)*IG (• • +) (• • • +) (• +)

<u>*Quy tắc 2*</u>: Một IG được tạo thành từ một hoặc nhiều SG liền kề được điều chỉnh về mặt cú pháp bởi một yếu tố xuất hiện trong chuỗi tuyến tính tạo nên IG.

Vì một chuỗi các SG có thể được sắp xếp theo nhiều cách, tức là với số lượng Ig khác nhau, câu hỏi đặt ra về tác dụng ngữ nghĩa của các sắp xếp khác nhau. Nhiều SG được hợp nhất vào một IG đơn lẻ cho thấy sự gắn kết ngữ nghĩa lớn hơn; IG tạo thành một đơn vị ngữ nghĩa.

Ngoài ra, khi các thanh điệu có hiệu ứng nghịch lý (như tiêu cự hoặc độ tương phản nghịch lý) được sử dụng, phạm vi của chúng bị giới hạn ở các yếu tố

trong IG, vì vậy các giới hạn của IG rất cần thiết cho việc giải thích ngữ nghĩa của mẫu ngữ điệu.

g. Các gói và nhóm ngôn điệu

Bất kỳ nhóm IGs nào cũng cho thấy một nhóm nội bộ phụ thuộc vào thanh điệu được sử dụng. Trong tiếng Pháp, nhóm này được xác định bởi các thanh điệu của vị trí AF. Do đó, cơ chế nhóm ngôn điệu là một quy tắc của tiếng Pháp.

Quy tắc 3: Quy tắc trên nhóm ngôn điệu: Đối với bất kỳ hai IGs liên tiếp nào: nếu thanh điệu ở vị trí AF của IG cuối cùng chiếm ưu thế so với IG đầu tiên, sau đó sẽ có hiệu ứng ẩn của IG đầu tiên trong giây thứ hai; mặt khác, hai IG là độc lập.

Nhóm ngôn điệu được đệ quy: Nó có thể được áp dụng cho các hình thức đơn vị bởi một bước trước đó. Gói được sử dụng để chỉ kết quả của một hoạt động nhóm; một gói chứa một hoặc nhiều IGs.

h. Nhóm ngôn điệu và cấu trúc cú pháp

Nhiều tác giả lưu ý rằng việc phân nhóm ngôn điệu phải tuân theo cấu trúc cú pháp. Người ta thường cho rằng các ranh giới ngôn điệu (và do đó phân nhóm ngôn điệu) phải tỷ lệ thuận với các ranh giới cú pháp.

Vì các giới hạn của gói được xác định bởi các IGs và cuối cùng là bởi các SGs và vì sau này có thể là một thành phần, một phần của thành phần, hoặc nhiều thành phần hơn, các gói không cần phải có kích thước của các thành phần.

Quan điểm tương ứng ngụ ý sự bất khả thi của việc có một ranh giới ngôn điệu chính tại một ranh giới bên trong của một thành phần cú pháp phức tạp. Tuy nhiên, dữ liệu giọng nói cho thấy các trường hợp trong đó một gói hợp nhất một thành phần đầu tiên chỉ với một phần của thành phần tiếp theo.

Vì việc phân nhóm ngôn điệu chỉ có thể đưa ra một số lượng hạn chế các quan hệ phân cấp, nên cơ chế không thể tái tạo toàn bộ cấu trúc cú pháp, ngay cả đối với các câu có độ phức tạp cú pháp vừa phải. Vì vậy, đến một lúc nào đó sự đồng nhất sẽ thất bại.

Đặc biệt với các thanh điệu tương phản, người ta có thể tìm thấy một hoặc nhiều thành phần là phần không nhấn âm của một IG, ngay cả với các yếu tố chi

phối về mặt cú pháp. Những sự thật này chỉ ra một tiêu chí mới cho sự đồng nhất về cú pháp ngôn từ.

Quy tắc 4: Nếu các phần tử được nhóm được liên kết bởi một mối quan hệ hóa trị, các IGs có thể được nhóm trong một gói và các gói thành các gói lớn hơn. Không có yêu cầu nào cho việc bao gồm các thành phần hoàn chỉnh.

Sự thể hiện rõ ràng của các cấp độ WS, SG; IG và các gói trong cấu trúc ngôn điệu cho phép các mối quan hệ giữa ngữ điệu, cú pháp và hình thái được xác định chính xác và hy vọng sẽ mang lại hiểu biết tốt hơn về tương tác của chúng.

1.3 Kết luận chương 1

Trong chương này luận văn đã giới thiệu tổng quan về tiếng nói, các đặc điểm và sự khác nhau của ngôn ngữ tiếng Việt và tiếng Pháp. Tiếng Việt là một ngôn ngữ có thanh điệu, do đó tần số cơ bản của nó thay đổi rất nhiều trong một âm tiết cũng như từ âm tiết sang âm tiết. Tiếng Pháp là một ngôn ngữ mà ngôn điệu có trọng âm, do đó tần số cơ bản của nó thay đổi không nhiều từ âm tiết đến âm tiết. Trong chương tiếp theo luận văn sẽ trình bày các thuật toán và mô hình hệ thống của bài toán nhận dạng tiếng nói dựa trên tần số cơ bản.

CHƯƠNG 2 - THUẬT TOÁN VÀ MÔ HÌNH HỆ THỐNG NHẬN DẠNG NGÔN NGỮ NÓI TỰ ĐỘNG DỰA TRÊN TẦN SỐ CƠ BẢN

2.1 Phân tích dữ liệu tiếng nói

Trong xử lý tiếng nói bao gồm: phân tích tiếng nói, tổng hợp tiếng nói và nhận dạng tiếng nói. Việc phân tích tiếng nói là vấn đề quan trọng quyết định đến kết quả của xử lý tiếng nói theo hướng nhận dạng hay tổng hợp. Việc phân tích tiếng nói tốt sẽ cho ta trích chọn các đặc trưng cơ bản, quan trọng nhất của tiếng nói để phục vụ cho công việc nhận dạng.

Như vậy mục đích của việc phân tích tín hiệu tiếng nói nhằm tách ra được các tham số đặc trưng cho tín hiệu tiếng nói. Các tham số này sẽ được ứng dụng trong nhận dạng hay tổng hợp tiếng nói. Mục đích của đồ án là trích chọn ra các đặc trưng: chu kỳ cao độ của tiếng nói từ đó xác định được tần số cơ bản, bởi tần số cơ bản đặc trưng cho thanh điệu của tiếng nói. Việc xác định tốt đặc trưng trên sẽ cho phép ta xây dựng được ứng dụng nhận dạng thanh điệu đạt chất lượng tốt.

2.1.1 Trích rút đặc trưng trong miền thời gian

a. Hàm tự tương quan (ACF)

Trong xử lý tín hiệu số, hàm tự tương quan của tín hiệu x(n) được định nghĩa như sau:

$$R(k) = \sum_{m=-\infty}^{\infty} x(m).x(m+k) \qquad (2.1.1)$$

Dễ thấy rằng nếu tín hiệu x(n) tuần hoàn với chu kỳ P thì hàm tự tương quan cũng tuần hoàn với chu kỳ P: R(k) = R(k+P)

Hơn nữa hàm tự tương quan còn có những tính chất quan trọng sau:

- Là hàm chẵn R(k) =R(-k)

- R(k) đạt giá trị cực đại tại 0 : |R(k)| < R(0) với mọi k

- Giá trị R(0) chính bằng năng lượng tín hiệu

$$R\left(k\right)=\sum_{m=-\infty}^{\infty}x^{2}\left(m\right) \tag{2.1.2}$$

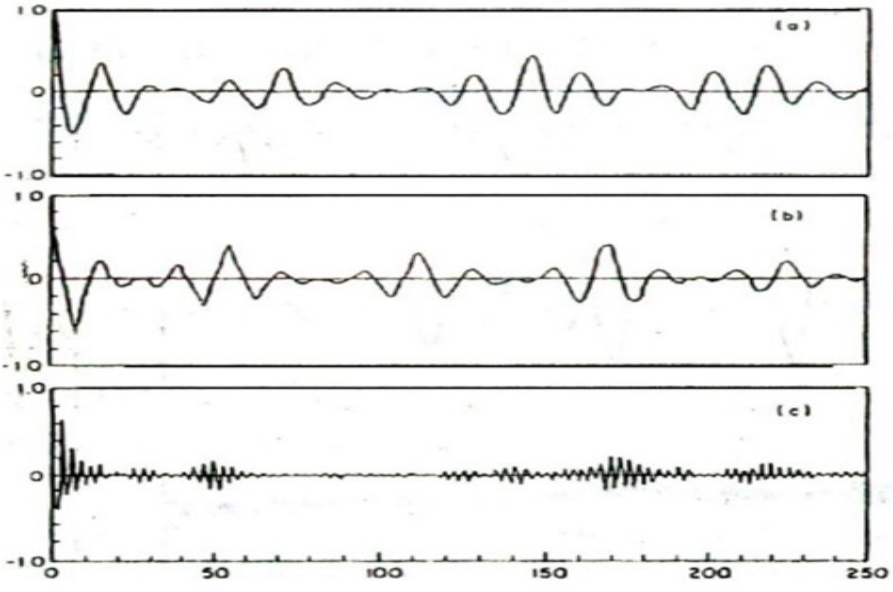
Dựa vào các tính chất trên ta có nhận xét: Hàm tự tương quan sẽ đạt giá trị cực đại tại các mẫu 0, ±P, ±2P,… và bằng giá trị năng lượng của tín hiệu, các điểm cực đại được gọi là các đỉnh (peak). Như vậy việc xác định chu kỳ cơ bản của tín hiệu tiếng nói sẽ đưa về việc xác định chu kỳ của hàm tự tương quan.

Để áp dụng cho một đoạn tín hiệu tiếng nói, ta phải xác định hàm tự tương quan thời gian ngắn.

Trước hết ta nhân tín hiệu với hàm cửa sổ thích hợp w(n), khi đó hàm tự tương quan được biểu diễn bằng công thức:

$$R_{n}\left(k\right)=\sum_{m=-\infty}^{\infty}x(m).\mathrm{w}(n-m).x(m+k).\mathrm{w}(n-m-k) \tag{2.1.3}$$

Biểu thức trên có thể hiểu như sau: đầu tiên một đoạn của tín hiệu tiếng nói được lựa chọn bằng cách nhân với cửa sổ; sau đó việc xác định hàm tự tương quan theo công thức định nghĩa được áp dụng cho đoạn tín hiệu đã qua cửa sổ.



Hình 2.1: Mô tả hàm tự tương quan

Trong hình 2.1: a,b là âm hữu thanh, c là âm vô thanh với của sổ N=40.

Dễ thấy:

$$R_n(-k) = R_n(k)$$

Và $$R_n(k) = R_n(-k) = \sum_{m=-\infty}^{\infty} [x(m).w(m-k)].[x(n-m).w(n-m+k)]$$

Nếu định nghĩa $h_k(n) = w(n).w(n+k)$

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m).x(m-k)].h_k(n-m) \quad (2.1.4)$$

Tức là $R_n(k)$ đạt được bằng cách cho x(m).x(m-k) qua bộ lọc có đáp ứng xung $h_k(n)$.

Việc tính toán hàm tự tương quan thời gian thực được tiến hành bằng việc sử dụng biểu thức định nghĩa được viết lại như sau:

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m+m).w'(m)].[x(n+m+k).w'(m+k)] \quad (2.1.5)$$

với w'(n) = w(-n).

Nếu w' là cửa sổ Hamming hoặc chữ nhật thì biểu thức trên có thể biểu diễn như sau:

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(m+m).w'(m)].[x(n+m+k).w'(m+k)] \quad (2.1.6)$$

Khi tính toán hàm tự tương quan việc lựa chọn N là rất quan trọng. Do sự không ổn định của tín hiệu tiếng nói nên giá trị N càng nhỏ càng tốt. Mặt khác để hàm tự tương quan tuần hoàn thì cửa sổ phải có chiều dài ít nhất 2 nửa chu kỳ của sóng tín hiệu. Mặt khác khi tính toán hàm tự tương quan thường được chuẩn hoá về 1 đơn vị.

b. Hàm vi sai biên độ trung bình (AMDF)

Xét chuỗi vi sai sau:

$$d(n) = x(n) - x(n-k) \quad (2.1.7)$$

Dễ thấy rằng d(n) tuần hoàn cùng chu kỳ P với tín hiệu gốc x(n) và đạt giá trị bằng 0 tại các mẫu 0, ±kP,…

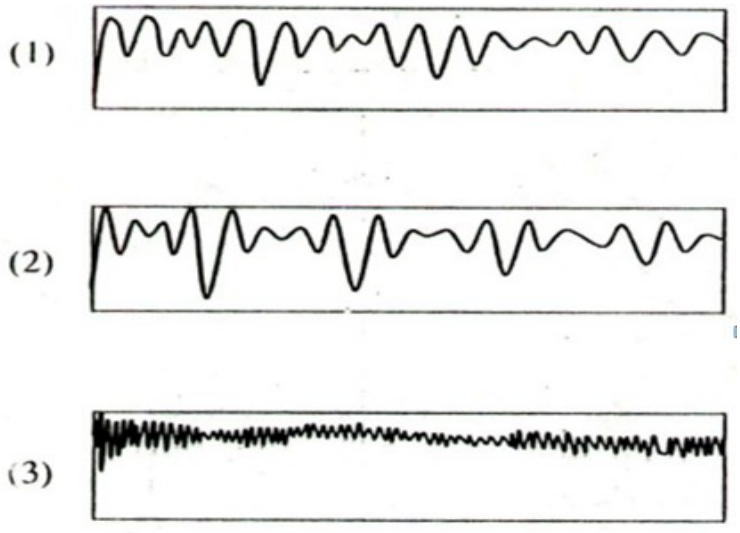
Biên độ trung bình thời gian ngắn của d(n) là một hàm của k có giá trị nhỏ khi k gần chu kỳ. Hàm vi sai biên độ trung bình thời gian ngắn (AMDF) được định nghĩa như sau:

$$d(p) = \sum_{n=0}^{N-1-p} |x(n) - x(n+p)| \qquad (2.1.8)$$

Nếu x(n) là tín hiệu tuần hoàn với chu kỳ T (mẫu) thì AMDF sẽ đạt cực tiểu nếu tín hiệu bị dời đi một đoạn đúng bằng T mẫu. Nhận dạng giọng của người có tần số cơ bản từ 80Hz (tương ứng với số mẫu là n1=Fs/80) đến 200Hz (tương ứng n2=Fs/200, Fs là tần số lấy mẫu).

Sẽ tính AMDF của tín hiệu với độ dời thay đổi từ n2 đến n1. Giả sử AMDF đạt cực tiểu ứng với độ dời P0 (mẫu). Đó chính là chu kỳ của tín hiệu (hoặc gần với chu kỳ của tín hiệu nhất), và tần số cơ bản của tín hiệu là F0=Fs/P0. Giá trị này chính là đặc trưng của tín hiệu về mặt thanh điệu.

Chu kỳ cao độ P0 được chọn sao cho d(P0) đạt giá trị nhỏ nhất.



Hình 2.2 Mô tả hàm vi sai biên độ trung bình

Trong hình 2.2: (1),(2) - âm hữu thanh, (3) - âm vô thanh

Hàm vi sai biên độ trung bình khác với hàm tự tương quan ở chỗ nó dùng phép trừ thay cho phép nhân nên tốc độ tính toán rất nhanh. Khi nhận dạng thanh

điệu tiếng Việt, phương pháp trích chu kỳ cao độ tức thời của không thể tốt bằng phương pháp trích chu kỳ cao độ trung bình theo AMDF vì đường vận động chu kỳ cao độ của AMDF mang tính trung bình nhiều hơn nên trơn hơn.

Do tiếng nói là tín hiệu không dừng nên cứ mỗi 30ms phải tín lại các giá trị mới. Tất cả các giá trị tính được sẽ là đặc trưng của một từ và được dùng để huấn luyện mạng nơ ron.

c. Trích chu kỳ cơ bản bằng AMDF

Tiếng nói được lấy mẫu ở tần số 11.025 kHz, mono 8 bit cho một mẫu, độ dài khung tiếng nói là 200 mẫu.

Bước 1: Lọc nhiễu sử dụng bộ lọc thông cao có tần số cắt là 60Hz do tần số cơ bản của người từ 80Hz đến 200Hz.

Bước 2: Tín hiệu sẽ được xén theo ngưỡng động để giữ lại các đỉnh lớn và làm nổi rõ chu kỳ cơ bản.

$$y(n)=\begin{cases} x(n)-C & \text{nêu } x(n)>C \\ 0 \text{ nêu} & -C \le x(n) \le C \\ C-x(n) & \text{nêu } x(n) < -C \end{cases} \qquad (2.1.9)$$

trong đó giá trị C khoảng 1/3 biên độ cực đại của tín hiệu.

Bước 3: Tín hiệu sau khi xén được đưa đến hàm lấy hiệu biên độ trung bình :

$$d(p)=\sum_{n=0}^{N-1-p} |x(n)-x(n+p)| \qquad (2.1.10)$$

trong đó N là độ dài khung và p được lấy trong khoảng pitch tương ứng với tần số cơ bản 80-200Hz. Chọn P0 có d cực tiểu, đó chính là chu kỳ pitch và tần số cơ bản là Fs/P0. Đối với các khung có d(P0) > 0.7 dmax(p) được phân loại là khung vô thanh và gán F0 = 0.

Bước 4: Sau khi đã xác định F0 của toàn bộ âm tiết, cần phải xử lý các khung có F0 = 0. Nếu các khung là vô thanh ở đầu hay ở cuối âm tiết, thay các khung đó bằng F0 của khung hữu thanh kế cận. Nếu các khung vô thanh ở giữa âm tiết thì thay F0 của khung đó bằng trung bình của hai khung hữu thanh ở hai bên.

Bước 5: Tín hiệu sau khi xử lý các khung có F0 = 0 sẽ được loại bỏ các đỉnh do nhận nhầm Pitch

Bước 6: Đường nét F0 được làm trơn bằng bộ lọc trung bình có trọng số với đáp ứng xung : h=[0.1 0.2 0.4 0.2 0.1]

Bước 7: Do số ngõ vào của mạng nơ ron là cố định nên cần chuẩn hóa kích thước F0.

Bằng các thử nghiệm trên các mẫu tiếng nói tôi nhận thấy với tần số lấy mẫu là 11kHz và tần số cơ bản nằm trong khoảng 80Hz-200Hz thì kích thước mỗi Frame tiếng nói là 200 mẫu sẽ cho kết quả đường tần số cơ bản trơn hơn. Số Frame tiếng nói với các mẫu âm thanh trung bình là 15 frames/mẫu tiếng nói. Do đó kích thước véc tơ F0 được chuẩn hóa với độ lớn là 15 phần tử. Vector này được lấy làm đặc trưng đầu vào để huấn luyện mạng nơ ron trong nhận dạng tiếng nói.

d. Hàm năng lượng thời gian ngắn và cường độ trung bình *(Short Time Energy and Average Magnitude*)

Chúng tôi đã quan sát và thấy rằng biên độ của tín hiệu lời nói thay đổi đáng kể theo thời gian. Cụ thể, biên độ của các phân đoạn vô thanh nói chung thấp hơn nhiều so với biên độ nếu các phân đoạn hữu thanh. Năng lượng thời gian ngắn của tín hiệu lời nói cung cấp một mô tả thuận tiện, phản ánh các biến đổi biên độ này. Nói chung, chúng ta có thể định nghĩa năng lượng thời gian ngắn là:

$$E = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (2.1.11)$$

Sự mô tả này có thể được viết dưới dạng:

$$E = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \qquad (2.1.12)$$

Trong đó:

$$h(n) = w^2(n) \qquad (2.1.13)$$

Sự lựa chọn của đáp ứng xung, h(n) hoặc tương đương với cửa sổ, xác định tính chất nếu biểu diễn năng lượng thời gian ngắn. Để xem sự lựa chọn của cửa sổ ảnh hưởng đến năng lượng thời gian ngắn như thế nào, chúng ta hãy quan sát rằng

nếu h(n) trong phương trình (2.1.12) rất dài và có biên độ không đổi, E_n sẽ thay đổi rất ít theo thời gian. Chẳng hạn như cửa sổ sẽ tương đương với một bộ lọc băng tần thông thấp rất hẹp. Rõ ràng những gì chúng tôi mong muốn là một số bộ lọc thông thấp nhưng không nhiều đến mức đầu ra không đổi; tức là chúng ta muốn năng lượng thời gian ngắn phản ánh sự thay đổi biên độ của tín hiệu giọng nói. Do đó, lần đầu tiên chúng ta bắt gặp một cuộc xung đột sẽ phát sinh liên tục trong nghiên cứu về các biểu diễn thời gian ngắn của các tín hiệu giọng nói. Nghĩa là, chúng ta muốn có một cửa sổ thời lượng ngắn (đáp ứng xung) để đáp ứng với thay đổi biên độ nhanh, nhưng một cửa sổ quá ngắn sẽ không cung cấp đủ trung bình để tạo ra một hàm năng lượng trơn tru.

Hiệu quả của lấy cửa sổ đối với biểu diễn năng lượng phụ thuộc thời gian có thể được minh họa bằng cách bàn luận về các thuộc tính của hai cửa sổ đại diện, tức là, cửa sổ hình chữ nhật

$$h(n) = 1 \qquad 0 \le n \le N-1 \qquad (2.1.14)$$

Và cửa sổ Hamming:

$$h(n) = 0.54 - 0.46 \cos(2\pi n/(N-1)). \qquad 0 \le n \le N-1 \qquad (2.1.15)$$

Cửa sổ hình chữ nhật tương ứng với việc áp dụng trọng số bằng nhau cho tất cả các mẫu trong khoảng (n-N+1) đến n.

e. Hàm tỷ lệ vượt quá điểm không trung bình thời gian ngắn *(Short Time Average Zero-Crossing Rate)*

Trong bối cảnh các tín hiệu rời rạc, một điểm vượt quá điểm không được cho là xảy ra nếu các mẫu liên tiếp có các dấu hiệu đại số khác nhau. Tỉ lệ tại đó vượt quá điểm không xảy ra là một phép đo đơn giản về nội dung tần số của một tín hiệu. Điều này đặc biệt đúng với tín hiệu băng tần hẹp. Ví dụ, tín hiệu hình sin có tần số F_0, được lấy mẫu ở tốc độ Fs, có các mẫu Fs/F_0 trên mỗi chu kỳ của sóng hình sin. Mỗi chu kỳ có hai điểm vượt quá điểm không sao cho tốc độ trung bình của thời gian vượt quá điểm không được tính như trong phương trình (2.1.16).

$$Z = 2F_s/F_0 \qquad (2.1.16)$$

Do đó, tỉ lệ vượt quá điểm không trung bình đưa ra một cách hợp lý để ước tính tần số của sóng hình sin.

Tín hiệu giọng nói là những tín hiệu băng thông rộng và việc giải thích của tỉ lệ vượt quá điểm không trung bình do đó ít chính xác hơn. Tuy nhiên, ước tính sơ bộ của các tính chất phổ có thể thu được bằng cách sử dụng biểu diễn dựa trên tỷ lệ vượt quá điểm không trung bình thời gian ngắn. Trước khi thảo luận về việc giải thích nếu tỷ lệ vượt quá điểm không đối với giọng nói, trước tiên chúng ta hãy xác định và bàn luận về các tính toán cần thiết. Một định nghĩa thích hợp là:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| w(n-m) \qquad (2.1.17)$$

Where

$$\text{sgn}[x(n)] = 1 \qquad x(n) \geq 0 \qquad (2.1.18)$$

$$= -1 \qquad x(n) < 0$$

And

$$w(n) = \frac{1}{2N} \qquad 0 \leq n \leq N-1 \qquad (2.1.19)$$

Otherwise $= 0$

2.1.2 Trích rút đặc trưng trong miền tần số

a. Xử lý giọng nói

Các đặc điểm của thanh quản xác định âm vị hiện tại. Các đặc điểm như vậy được chứng minh trong miền tần số bằng vị trí của các định dạng, tức là các đỉnh được đưa ra bởi sự cộng hưởng của thanh âm. Tần số cao có biên độ nhỏ tương tự các định dạng tần số thấp mặc dù sở hữu thông tin liên quan. Việc xử lý như vậy thường thu được bằng cách lọc tín hiệu giọng nói với bộ lọc FIR thứ nhất, có chức năng truyền trong miền z là:

$$H(z) = 1\text{-}a.z^{-1} \quad , 0 \leq a \leq 1 \qquad (2.1.20)$$

a là thông số nhấn mạnh trước. Về bản chất, trong miền thời gian, tín hiệu được nhấn mạnh trước có liên quan đến tín hiệu đầu vào theo quan hệ:

$$x'(n) = x(n) - ax(n-1) \qquad (2.1.21)$$

Giá trị điển hình cho a là 0,95, làm tăng mức khuếch đại hơn 20 dB của phổ tần số cao.

b. Lấy cửa sổ (Windowing)

Các phương pháp truyền thống để đánh giá phổ là đáng tin cậy trong trường hợp tín hiệu đứng yên. Đối với giọng nói, điều này chỉ giữ độ ổn định khớp nối trong khoảng thời gian ngắn, trong đó phân tích thời gian ngắn có thể được thực hiện bằng cách Lấy cửa sổ của một tín hiệu *x'(n)* để nối tiếp các chuỗi cửa sổ

$x_t(n)$, t = 1,2, tầm, T, được gọi là khung, sau đó được xử lý riêng lẻ:

$$x'_t \equiv x'(n-tQ) \ , \ 0 \le n \le N, \ \ 1 \le t \le T \qquad (2.1.22)$$

$$x_t(n) \equiv w(n)x'_t(n) \qquad (2.1.23)$$

Trong đó w(n) là đáp ứng xung của cửa sổ. Mỗi khung được dịch chuyển theo độ dài thời gian Q. Nếu Q = N, các khung không trùng nhau theo thời gian trong khi nếu các mẫu Q <N, N - Q ở cuối khung $x'_t(n)$ được sao chép ở đầu của khung sau x $x'_{t+1}(n)$. Chúng ta biết rằng phân tích Fourier được thực hiện thông qua biến đổi Fourier mà đối với tín hiệu thời gian riêng biệt $x_t(n)$ là:

$$X_t(e^{j\omega}) = \sum_{n=0}^{N-1} x_t(n)e^{-j\omega n} = \Im\{x_t(n)\} \qquad (2.1.24)$$

Nơi ω là trục tần số liên tục. Giới thiệu biến đổi Fourier của $w(n)$ và $x'_t(n)$: $W(e^{-j\omega}) = \Im\{w(n)\}$ $X'_t(e^{-j\omega}) = \Im\{x'_t(n)\}$, một sản phẩm trong miền thời gian như trong phương trình (2.1.23) trở thành tích chập trong miền tần số:

$$X_t = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_t'\left(e^{j\theta}\right)W\left(e^{j(\omega-\theta)}\right)d\theta = \Im\{x'_t(n)w(n)\} \qquad (2.1.25)$$

Suy ra từ hàm (2.1.22) và (2.23) , hàm (2.1.24) có thể được viết là:

$$X_t\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x'(n-t.Q)w(n)e^{-j\omega n} \qquad (2.1.26)$$

Công thức trên cũng được đề cập như là Biến đổi Fourier thời gian ngắn (STFT) hoặc Biến đổi Fourier có cửa sổ của *x(n)*.

Có hai loại cửa sổ thường được sử dụng. Chúng là cửa sổ hình chữ nhật và cửa sổ Hamming.

Sự hiện diện của một cửa sổ hình chữ nhật gây ra sự biến dạng trên phổ ước tính, vì $X_t(e^{j\omega})$ là tích chập của phổ $x_t'(n)$ và biến đổi Fourier của cửa sổ hình chữ nhật w (n). $W(e^{j\omega})$ bao gồm một thùy chính năng lượng cao hơn tập trung ở tần số 0 và các thùy bên năng lượng thấp hơn tập trung ở tần số cao hơn. Thùy chính trải ra trong một dải tần số rộng hơn công suất dải hẹp của tín hiệu $x_t'(n)$ mà trong trường hợp của chúng ta được biểu thị bằng các định dạng. Hiện tượng này làm giảm độ phân giải tần số cục bộ. Hơn nữa, các thùy bên của $W(e^{j\omega})$ trao đổi năng lượng từ các tần số xa và khác nhau của $x_t'(n)$. Vấn đề này được gọi là rò rỉ (*leakage*.).

- Hình dạng cửa sổ có thể làm giảm độ méo, nhưng nó có thể làm tăng sự thay đổi hình dạng tín hiệu.

- Độ dài N tỷ lệ thuận với độ phân giải tần số và tỷ lệ nghịch với độ phân giải thời gian.

- N-Q lặp tỷ lệ thuận với tỷ lệ khung, nhưng nó cũng tỷ lệ thuận với tương quan của các khung tiếp theo.

Các thùy bên của cửa sổ Hamming thấp hơn nhiều so với cửa sổ hình chữ nhật mặc dù độ phân giải giảm đáng kể. Điều này là do thùy chính Hamming rộng hơn. Việc làm rõ hay làm trơn tín hiệu có ý nghĩa quan trọng trong bài toán nhận dạng tiếng nói, làm tăng hiệu quả của hệ thống nhận dạng. Cửa sổ Hamming là một lựa chọn tốt trong nhận dạng giọng nói, vì không cần độ phân giải cao.

c. Phân tích phổ

Các phương pháp tiêu chuẩn để phân tích phổ dựa vào biến đổi Fourier $x_t'(n)$: $X_t(e^{j\omega})$. Độ phức tạp tính toán giảm đáng kể nếu $X_t(e^{j\omega})$ chỉ được ước tính cho một số giá trị ω rời rạc.

Nếu các giá trị như vậy cách đều nhau, ví dụ, xem xét $\omega=2\pi k/N$, thì biến đổi Fourier rời rạc của tất cả các khung của tín hiệu được lấy:

$$X_t(k) = X_t\left(e^{j2\pi k/N}\right) \qquad ,k = 0,1,...,N-1 \qquad (2.1.27)$$

Ngoài ra, nếu số lượng mẫu N là lũy thừa 2, $N=2^p$ với p là số nguyên, thì độ phức tạp tính toán có thể được giảm thêm thành một đơn $Nlog(N)$ dùng cho FFT.

Lưu ý rằng nếu $x_t(n)$ là có thật, FFT có thể được tính bằng một nửa độ phức tạp tính toán, trong trường hợp này là $N/2log(N/2)$.

Các đặc điểm của thanh âm có thể được ước tính bằng biểu đồ của $x_t'(n,)$, *mà* đơn giản là cường độ bình phương của DFT: $|X_t(k)|^2$. Xét rằng biểu đồ là một công cụ ước lượng không nhất quán không thiên vị của năng lượng phổ, $|X_t(k)|^2$ là một công cụ ước tính của $P_x(\omega)$ được đưa ra trong phương trình: $P_x(\omega) = P_x(\omega)P_h(\omega)$.

Lưu ý rằng thông tin pha của các mẫu DFT của mỗi khung bị loại bỏ. Điều này phù hợp với thực tế là pha không mang thông tin hữu ích. Các thí nghiệm tri giác đã chứng minh rằng quan niệm về tín hiệu được tái tạo với các pha ngẫu nhiên gần như không thể phân biệt được với bản gốc, nếu tính liên tục của pha giữa các khung liên tiếp được giữ nguyên.

Cũng lưu ý rằng việc sử dụng các bước sóng cosin cục bộ hoàn toàn có thể tăng cường SFFT rời rạc.

d. Hệ thống xử lý băng lọc

Phân tích phổ cho thấy các đặc trưng tín hiệu giọng nói, chủ yếu là do hình dạng của thanh quản. Các đặc trưng phổ của lời nói thường thu được là lối ra của các băng lọc, tích hợp đúng phổ ở các dải tần xác định. Một bộ gồm 24 bộ lọc thông dải thường được sử dụng vì nó mô phỏng quá trình xử lý qua tai của con người.

Các bộ lọc thường được bố trí không đồng dạng với trục tần số. Như một quy luật, phần phổ dưới 1 kHz được xử lí bởi nhiều băng lọc vì nó chứa nhiều thông tin trên thanh âm ví dụ như cấu trúc đầu tiên. Phản ứng tần số của các băng lọc mô phỏng quá trình xử lý cảm nhận được thực hiện trong tai và do đó việc lọc như vậy được gọi là trọng số tri giác (*perceptual weighting*).

Phân tích tần số phi tuyến tính cũng được sử dụng để đạt được độ phân giải tần số/thời gian. Sử dụng các bộ lọc thông dải hẹp ở tần số thấp cho phép xóa sóng hài, nhưng nó cung cấp thông tin khởi phát kém. Sử dụng băng thông dài hơn ở tần số cao hơn cho phép độ phân giải theo thời gian cao hơn.

Thang đo nhận thức được sử dụng rộng rãi nhất trong nhận dạng là thang đo Mel. Tần số trung tâm của mỗi giàn bộ lọc Mel được đặt cách đều nhau trước 1 kHz

và nó tuân theo thang logarit sau 1 kHz. Chúng ta biết rằng trong khoảng thời gian lấy mẫu T_c, tần số ω của tín hiệu thời gian rời rạc có liên quan đến tần số f của tín hiệu thời gian liên kết bằng cách:

$$f = \frac{\omega}{2\pi T_c} \tag{2.1.28}$$

Có một loạt các phương pháp để thực hiện các bộ lọc như vậy. Một phương pháp tính toán rẻ bao gồm việc thực hiện lọc trực tiếp trong miền DFT. Các phản ứng DFT của các bộ lọc chỉ đơn giản là các phiên bản bị thay đổi và bị biến dạng tần số của cửa sổ hình tam giác $U_{\Delta m}(k)$:

$$U_{\Delta m}(k) = \begin{cases} |k| < \Delta m \rightarrow 1 - |k|/\Delta m \\ |k| > \Delta m \rightarrow 0 \end{cases} \tag{2.1.29}$$

Trong đó k là chỉ số miền DFT và $2\Delta m$ là kích thước của cửa sổ tam giác giàn bộ lọc thứ m. Tín hiệu ra của bộ lọc thứ m được đưa ra bởi:

$$Y_t(m) = \sum_{k=b_m-\Delta_m}^{b_m+\Delta_m} X_t(k) U_{\Delta_m}(k + b_m) \tag{2.1.30}$$

Trong đó Xt(k) được cho bởi phương trình (2.1.17) và 1 <m <M. Tần số trung tâm có thể được tính theo $b_m = b_m + \Delta_m$, và, $\frac{\omega}{2\pi T_c} = f < 1kHz$, Δ_m được chọn sao cho thu được 10 bộ lọc cách đều nhau. Đối với $f > 1kHz$, có thể sử dụng ước lượng sau: $\Delta_m = 1.2x\Delta_{m-1}$

e. Phép tính Log năng lượng

Quy trình trước có vai trò làm mịn phổ, thực hiện quá trình xử lý tương tự như quy trình được thực hiện bởi tai người. Bước tiếp theo bao gồm tính toán logarit của cường độ bình phương các hệ số $Y_t(m)$ thu được từ phương trình (2.1.20). Sự giảm này đơn giản là tính toán logarit của cường độ các hệ số, bởi vì tính chất đại số logarit mang lại logarit của công suất nhân với hệ số tỷ lệ. Tương tự, lợi ích của quy trình này có thể được nhìn thấy bằng cách sử dụng khuôn khổ của phân tích cepstral được giới thiệu trong phần tiếp theo. Ở đây chúng ta lưu ý rằng việc xử lý cường độ và logarit cũng được thực hiện bằng tai. Hơn nữa, cường

độ loại bỏ pha thông tin vô dụng trong khi logarit thực hiện một sức ép động học, làm cho việc trích xuất đặc trưng ít nhạy cảm hơn với các biến động trong động lực học.

g. Tính toán cepstrum tần số Mel

Quy trình cuối cùng cho việc tính toán cepstrum tần số Mel bao gồm thực hiện nghịch đảo DFT trên logarit của cường độ tín hiệu bộ lọc đầu ra:

$$y_t^{(m)}(k) = \sum_{m=1}^{M} \{\log|Y_t(m)|\}.\cos\left(k\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right), \qquad k = 0,1,...,L \qquad (2.1.31)$$

Quy trình này có lợi thế lớn. Đầu tiên, lưu ý rằng vì log phổ công suất là thực và đối xứng nên DFT nghịch đảo giảm xuống thành một Biến đổi Cosine rời rạc (Discrete Cosine Transform - DCT). DCT có đặc tính để tạo ra các đặc trưng không tương thích $y_t^{(m)}(k)$. Do đó, đặc tính ngẫu nhiên của quá trình các đặc trưng đơn giản hơn và trong hàm mật độ xác suất của các đặc trưng, thường được mô phỏng bằng các tổ hợp tuyến tính của các hàm Gaussian, ma trận hiệp phương sai có thể được sử dụng thay vì ma trận hiệp phương sai hoàn toàn. Điều này làm giảm đáng kể chi phí tính toán và số lượng tham số được ước tính.

Các hệ số MFCC $y_t^{(0)}(k)$ trong phương trình (2.1.21) gần tương đương với năng lượng log của khung. Hệ số này thường bị loại bỏ vì năng lượng được tính trực tiếp trên tín hiệu thời gian. DCT cũng có tác dụng làm mịn phổ nếu chỉ các hệ số đầu tiên được giữ lại.

h. Hệ số dữ liệu và năng lượng

Đầu tiên cần lưu ý rằng các hệ số cepstral thường là một hệ số năng lượng e_t lấy logarit của năng lượng của khung. Thông số này rất hữu ích vì sự khác biệt về năng lượng được nhìn thấy giữa các âm vị khác nhau.

Một sự cải thiện nữa trong việc thực hiện có được bằng cách xem xét rằng các thông số và năng lượng của cepstral không tính đến sự phát triển động của tín hiệu giọng nói. Do đó, với một vectơ chung u_t được lập trong thời gian t, chênh lệch thời gian thứ i có thể được tính là:

$$\Delta^i\{u_t\} = \Delta^{i-1}\{u_{i+1}\} - \Delta^{i-1}\{u_{t-1}\}, \quad \Delta^0\{u_t\} = u_t \qquad (2.1.32)$$

Lưu ý rằng sự khác biệt i-*th* liên quan đến sự khác biệt *(i-l)-th*. Khoảng cách cao hơn hoặc thấp hơn nên được suy xét theo thời gian chồng chéo. Do khoảng cách thấp có thể bao hàm các khung quá tương quan và do đó, động lực không được bắt bởi các khác biệt, các giá trị cao hơn có thể bao hàm các khung mô tả các trạng thái quá khác nhau.

i. Phân tích Cepstrum

Cepstrum phức tạp (tên là đảo chữ của cepstrum) $\hat{x}(n)$ cho tín hiệu rời rạc x (n) là biến đổi Fourier ngược của logarit phức tạp *log* X($e^{-j\omega}$)

$$\hat{x}(n) = \Im^{-1}\{\log X(e^{j\omega})\} \qquad (2.1.33)$$

Logarit của phổ có tác dụng làm giảm biên độ thành phần ở mọi tần số. Thang đo logarit này cũng là một đặc điểm của hệ thống thính giác của con người. Do đó, những tín hiệu được đặc trưng bởi sự kết hợp của sóng hài được phân tích tốt hơn bằng phổ chứ không phải bằng phổ hoặc tự tương quan.

Việc sử dụng cepstrum lần đầu tiên được giới thiệu để phân biệt các phân đoạn lời nói hữu thanh và vô thanh. Trong thực tế, cepstrum nhấn mạnh các thành phần của thanh âm, ngay cả với tiếng ồn. Ngược lại, cepstrum phẳng cho âm thanh thiếu cấu trúc hài hòa rõ ràng. Bằng cách khám phá các tính chất này, các hệ số cepstrum đã được sử dụng để phân loại các phân đoạn giọng nói, xác định sự phát triển của kỹ thuật cepstrum. Thật vậy, phân tích cepstrum, nghĩa là một phân tích đồng hình với một logarit như là chức năng trung gian, cho phép giải mã các tín hiệu lời nói như được giải thích dưới đây.

Một dạng sóng lời nói có thể được coi là một tổ hợp giữa sự kích thích được tạo ra bởi các dây thanh âm *v(n)* và phản ứng xung của một bộ lọc đại diện cho thanh âm *h(n):*

$$x(n) = v(n)*h(n) \qquad (2.1.34)$$

Do thông tin ngữ âm chủ yếu liên quan đến thanh âm, thuật toán giải mã cho tín hiệu giọng nói được quan tâm đáng kể nhằm cô lập phản ứng của thanh âm. Các

thuật toán này thuộc về nhánh lý thuyết hệ thống được gọi là lọc đồng hình. Phải sử dụng đến cepstrum phức tạp, ta có:

$$\begin{aligned}\hat{x}(n) &= \Im^{-1}\{\log(\Im\{v(n)*h(n)\})\} \\ &= \Im^{-1}\left\{\log(V(e^{j\omega}))+\log(H(e^{j\omega}))\right\} = \hat{v}(n)+\hat{h}(n)\end{aligned} \tag{2.1.35}$$

Trong đó, $\hat{v}(n)$, $\hat{h}(n)$ lần lượt là các cepstrum phức tạp của *v(n)* và *h(n)*. Các cepstrum phức tạp biến đổi tích chập (2.1.24) thành tổng của hai thành phần $\hat{v}(n)$, $\hat{h}(n)$ mà có thể phân tách bằng các bộ lọc tuyến tính băng thông, nếu không có sự chồng chéo tần số.

Đối với tín hiệu giọng nói, điều này là khả thi vì phổ thời gian ngắn cho thấy đường bao của bộ lọc thanh âm *h(n)* thay đổi chậm đối với cấu trúc tinh tế của các sóng hài được tạo ra bởi sự kích thích định kỳ của lời nói *v(n)*.

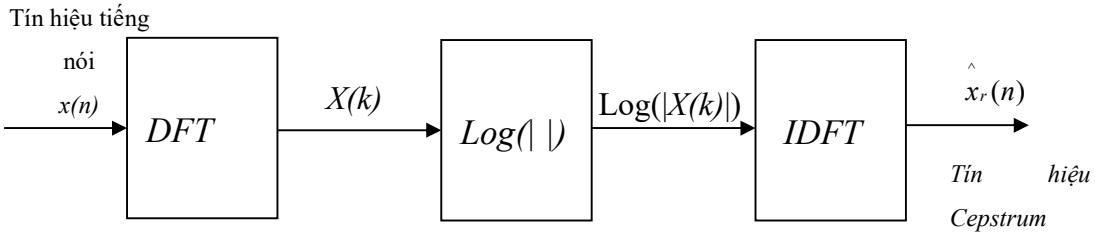
Đối với các tín hiệu pha tối thiểu hoặc khi thông tin pha không được quan tâm, cepstrum thực có thể được sử dụng thay cho cepstrum phức tạp. Cepstrum thực của tín hiệu được xác định bằng biến đổi Fourier ngược của logarit có độ lớn $X(e^{j\omega})$:

$$\hat{x}(n) = \Im^{-1}\left\{\log\left|X(e^{j\omega})\right|\right\} \tag{2.1.36}$$

Như thể hiện trong hình 2.3, cepstrum thực có thể được tính bằng DFT nghịch đảo.

Sự phân giải đồng hình thể hiện trong phương trình (2.1.25) có thể làm nổi bật các thuộc tính có liên quan của MFCC. Đầu tiên lưu ý rằng hằng số nhân được áp dụng cho một tín hiệu giọng nói, logarit của hằng số như vậy được thêm vào tất cả các hệ số của $log|Y_t(m)|^2$. Các hằng số như vậy chỉ ảnh hưởng đến hệ số 0 $y_t^{(m)}(0)$. Cũng lưu ý rằng phản ứng của thanh âm và kích thích tín hiệu được kết hợp một cách cộng gộp trong cepstrum như ở công thức (2.1.25). Phổ log thanh âm có một hành vi trơn tru trong khi kích thích có phổ bán định kỳ biến đổi cao cho âm hữu thanh. Do đó, phản ứng thanh âm có thể thu được bằng cách đơn giản giữ lại các hệ số cepstral đầu tiên $y_t^{(m)}(k)$. Đó là lý do tại sao chỉ các hệ số *k-th*, $k \le L \le 15$, được giữ lại. Cũng lưu ý rằng ảnh hưởng môi trường có thể được mô hình hóa như

một bộ lọc tuyến tính. Sự xuống cấp này trở thành sai lệch trong ước tính phổ log có thể được đánh giá và loại bỏ.



Hình 2.3 Sơ đồ khối của tín hiệu cepstrum thực

2.2 Mạng nơ ron ứng dụng trong nhận dạng tiếng nói

2.2.1 Phương pháp nhận dạng dùng mạng nơ ron

Mạng nơ ron nhân tạo hay thường gọi ngắn gọn là mạng nơ ron là một mô hình toán học hay mô hình tính toán được xây dựng mô phỏng các mạng nơ ron sinh học, là sự liên kết giữa các nơ ron nhân tạo với nhau. Các nơ ron được sắp xếp trong mạng theo các lớp, bao gồm lớp ngoài cùng gọi là lớp ra (output layer), các lớp còn lại gọi là lớp ẩn (hide layer). Các nơ ron trong cùng một lớp thì nhận tín hiệu cùng vào cùng một lúc. Chức năng của mạng được xác định bởi: cấu trúc mạng, quá trình xử lý bên trong của từng nơ ron, và mức độ liên kết giữa các nơ ron.

Các khả năng của mạng nơ ron:

+ Khả năng học: Mạng nơ ron có khả năng tiếp thu sự huấn luyện về mối quan hệ giữa đầu vào và đầu ra của các mẫu học, nếu ta chỉ cho nó đầu vào x tương ứng với đầu ra y thì mạng có thể nhớ lại được điều đó.

+ Khả năng chuẩn hoá: Mạng nơ ron học các mẫu dữ liệu cơ sở nên có khả năng nhận dạng được dữ liệu mới, những dữ liệu mà nó cho rằng gần giống với dữ liệu đã được học. Chính khả năng này của mạng nơ ron rất thuận lợi khi ứng dụng nó nhận dạng tiếng nói vì các mẫu âm học không bao giờ giống nhau một cách tuyệt đối.

+ Khả năng tính toán: Mạng nơ ron có khả năng tính toán song song rất cao, đáp ứng yêu cầu của các giải thuật. Trong nhận dạng tiếng nói khối lượng tính toán

là rất lớn. Và kết quả cần được đưa ra nhanh chóng thì hệ thống mới nâng cao được giá trị sử dụng.

Nhờ những khả năng đó của mạng nơ ron mà con người đã cố gắng mô phỏng để tạo ra mạng nơ ron nhân tạo, kết quả mà chúng ta đạt được là rất khả quan và hiện nay được ứng dụng rông rãi trong nhiều bài toán đặc tả quan trọng, đặc biệt là trong nhận dạng tiếng nói.

2.2.2 Luật học của mạng nơ ron

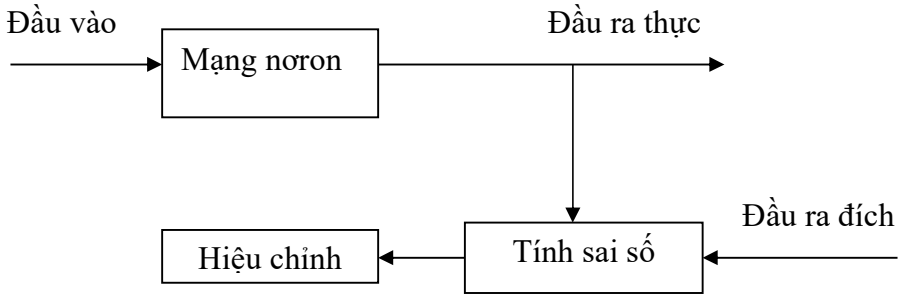
Sau khi lựa chọn kiến trúc của mạng, quá trình huấn luyện mạng (training) là quá trình hiệu chỉnh các trọng số và hệ số bias của mạng bằng một số luật học (thuật toán huấn luyện mạng). Các luật học được chia làm 3 loại: luật học có giám sát luật học không giám sát và luật học tăng cường.

a. Luật học có giám sát

Trong luật học có giám sát, mạng được cung cấp một tập hợp các mẫu chuẩn thể hiện mối quan hệ đầu vào và đầu ra của mạng: $\{p_1,t_1\},\{p_2,t_2\},\ldots,\{p_Q,t_Q\}$.

Trong đó p_q là đầu vào của mạng, t_q là đầu ra đích: là đầu ra mạng cần thích ứng. Khi p_q được đưa vào mạng thì đầu ra thực của mạng là a_q. Sai số e_q giữa t_q và a_q được sử dụng để hiệu chỉnh các trọng số và hệ số bias của mạng.

Sơ đồ khối mô tả luật học có giám sát như sau:



Hình 2.4 Sơ đồ khối luật học có giám sát

Để đánh giá sai số e_q người ta dùng hàm sai số. Hàm này thể hiện chất lượng học của mạng và có nhiều cách để lựa chọn. Một khái niệm liên quan đó là

mặt lỗi (error surface). Mỗi một trọng số và hệ số bias của mạng tương ứng với một chiều không gian. Nếu mạng có tất cả N trọng số và hệ số bias thì chiều thứ N+1 biểu diễn sai số của mạng. Mỗi một bộ trọng số và hệ số bias của mạng ứng với một điểm của mặt lỗi. Mục tiêu của luật học là tìm được điểm thấp nhất (điểm cực tiểu) của mặt lỗi này tức là tìm điểm ứng với bộ trọng số và hệ số bias tối ưu.

b. Luật học không có giám sát

Trong dạng luật học này, mạng không được biết trước về véctơ đầu ra đích đối với từng vectơ đầu vào. Mạng phải tự phân tích các đặc trưng, tính chất của các mẫu đầu vào để phân loại chúng.

Các trọng số và hệ số bias của mạng được thay đổi chỉ để đáp ứng yêu cầu đầu vào của mạng. Hầu hết các luật học này thực hiện thao tác phân loại mẫu: chúng tự động phân loại các mẫu thành một số hữu hạn các lớp. Các mạng với cơ chế học không giám sát gọi là các mạng tự tổ chức.

Trong thực tế người ta thường kết hợp cả hai phương pháp học có giám sát và học không có giám sát. Giả sử số lượng các mẫu là lớn, nếu áp dụng học có giám sát sẽ không thích hợp. Do đó trước hết dùng phương pháp học không giám sát để phân loại các mẫu, sau khi các mẫu được chia thành các nhóm tương đối nhỏ thì có thể áp dụng phương pháp học có giám sát đối với một số nhóm nào đó.

c. Luật học tăng cường

Luật học này tương tự như luật học có giám sát, nhưng thay vì được cung cấp đầu ra đúng, mạng chỉ nhận được đánh giá theo các mức độ (cho điểm) đối với mỗi đầu vào. Đánh giá này cho phép đo hiệu xuất của mạng trên một dãy đầu vào Tín hiệu tăng cường được xử lý bởi bộ xử lý tín hiệu tăng cường tạo ra các tín hiệu đánh giá giúp mạng hiệu chỉnh các trọng số với hy vọng nhận tín hiệu đánh giá tốt hơn trong tương lai.

Luật học này không được sử dụng rộng rãi bằng luật học có giám sát và được dùng chủ yếu trong các hệ thống điều khiển.

2.2.3 Thuật toán lan truyền ngược - Back propagation

Tín hiệu lỗi tại đầu ra của nơ ron j tại vòng lặp thứ n (khi xử lý ví dụ tích luỹ thứ n) được xác định như sau

$$e_j(n) = d_j(n) - y_j(n) \quad \textit{nơ ron } j \textit{ là một nút đầu ra} \qquad (2.2.1)$$

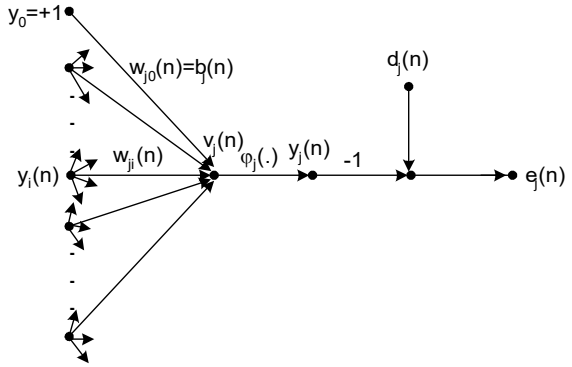
Chúng ta định nghĩa giá trị tức thời của năng lượng lỗi cho nơ ron j là $\frac{1}{2}e_j^2(n)$. Cũng tương tự như vậy, giá trị tức thới τ(n) của năng lượng lỗi tổng cộng nhận được bởi việc tính tổng $\frac{1}{2}e_j^2(n)$ trên tất cả các nơ ron trong mức đầu ra; đây là các nơ ron nhìn thấy duy nhất mà các tín hiệu lỗi có thể được tính toán một cách trực tiếp. Như vậy, chúng ta có thể viết:

$$\tau(n) = \frac{1}{2}\sum_{j \in C} e_j^2(n) \qquad (2.2.2)$$

ở đó tập hợp C bao gồm tất cả các nơ ron trong mức đầu ra của mạng. Đặt N là số các mẫu (các ví dụ) trong tập hợp tích luỹ. Năng lượng lỗi bình phương trung bình nhận được bằng cách tính tổng τ(n) trên tất cả các giá trị của n rồi chia cho kích thước tập hợp N như sau:

$$\tau_{av} = \frac{1}{N}\sum_{n=1}^{N} \tau(n) \qquad (2.2.3)$$

Các giá trị năng lượng lỗi kể trên là hàm của tất cả các các tham số tự do (các trọng số synapse và các hệ số hiệu chỉnh) của mạng. Với một tập hợp tích luỹ cho trước, đại lượng τ_{av} biểu diễn một *hàm giá* như một thước đo tính năng của việc học. Mục đích của quá trình học là để điều chỉnh các tham số tự do của mạng làm cho τ_{av} đạt cực tiểu. Chúng ta xem xét một phương pháp tích luỹ đơn giản mà trong đó các trọng số được cập nhật trên cơ sở từng mẫu một cho tới khi một *Thời kỳ* (toàn bộ tập hợp tích luỹ được sử dụng một lượt) kết thúc. Những sự điều chỉnh trọng số được thực hiện theo các lỗi tương ứng được tính toán cho từng mẫu tích luỹ.



Hình 2.5 Đồ thị luồng tín hiệu chi tiết cho một nơ ron đầu ra

Chúng ta xem xét hình 2.5, trong đó thể hiện nơ ron j với các đầu vào là các tín hiệu chức năng được tạo ra bởi mức nơ ron ở bên trái. Tổ hợp tuyến tính $v_j(n)$ được tạo ra tại đầu vào của hàm kích hoạt của nơ ron j như sau

$$v(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \qquad (2.2.4)$$

ở đó m là số các đầu vào của nơ ron j. Trọng số synapse w_{i0} (tương ứng với đầu vào cố định y_0=+1) là hệ số hiệu chỉnh b_j của nơ ron j. Như vậy tín hiệu chức năng $y_i(n)$ xuất hiện tại đầu ra của nơ ron j tại vòng lặp thứ n là

$$y_j(n) = \varphi(v_j(n)) \qquad (2.2.5)$$

Trong thuật toán back-propagation, hiệu chỉnh $\Delta w_{ji}(n)$ đối với trọng số synapse $w_{ji}(n)$ tỷ lệ với đạo hàm riêng $\partial\tau(n)/\partial w_{ji}(n)$. Theo quy tắc tính toán đạo hàm, chúng ta có thể biểu diễn gradient này như sau

$$\frac{\partial \tau(n)}{\partial w_{ji}(n)} = \frac{\partial \tau(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_j(n)} \qquad (2.2.6)$$

đạo hàm riêng này xác định hướng tìm kiếm trong không gian trọng số đối với trọng số synapse w_{ji}.

Đạo hàm cả hai vế của (2.2.2) theo $e_j(n)$, chúng ta có

$$\frac{\partial \tau(n)}{\partial e_j(n)} = e_j(n) \qquad (2.2.7)$$

Đạo hàm cả hai vế của (2.2.1) theo $y_j(n)$, chúng ta có

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1 \tag{2.2.8}$$

Đạo hàm cả hai vế của (2.2.5) theo $v_j(n)$, chúng ta có

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi'_j(v_j(n)) \tag{2.2.9}$$

Cuối cùng, đạo hàm cả hai vế của (2.2.4) theo $w_{ji}(n)$, chúng ta có

$$\frac{\partial v_j(n)}{\partial w_j(n)} = y_i(n) \tag{2.2.10}$$

áp dụng các công thức từ (2.2.7) đến (2.2.10) cho công thức (2.2.6), ta có

$$\frac{\partial \tau(n)}{\partial w_j(n)} = -e_j(n)\psi'_j(v_j(n))y_i(n) \tag{2.2.11}$$

Hiệu chỉnh $\Delta w_{ji}(n)$ áp dụng cho trọng số $w_{ji}(n)$ được xác định theo *quy tắc delta* như sau

$$\Delta w_{ji}(n) = -\eta \frac{\partial \tau(n)}{\partial w_j(n)} \tag{2.2.12}$$

ở đó η là tham số tốc độ học. Như vậy từ (2.2.11) và (2.2.12), ta có

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_j(n) \tag{2.2.13}$$

ở đó *gradient cục bộ* $\delta_j(n)$ được xác định như sau

$$\begin{aligned} \delta_j(n) &= -\frac{\partial \tau(n)}{\partial v_j(n)} \\ &= -\frac{\partial \tau(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \\ &= e_j(n)\varphi'_j(v_j(n)) \end{aligned} \tag{2.2.14}$$

Gradient cục bộ chỉ ra những thay đổi cần thiết cho các trọng số synapse. Từ các công thức (2.2.13) và (2.2.14), chúng ta thấy rằng một yếu tố chính liên quan tới việc tính toán các điều chỉnh trọng số $\Delta w_{ji}(n)$ là tín hiệu lỗi $e_j(n)$ tại đầu ra của nơ ron j. Ở đây, chúng ta quan tâm đến hai trường hợp riêng biệt. Trường hợp thứ nhất, nơ ron j là một nút đầu ra; và trường hợp thứ hai, nơ ron j là một nút ẩn.

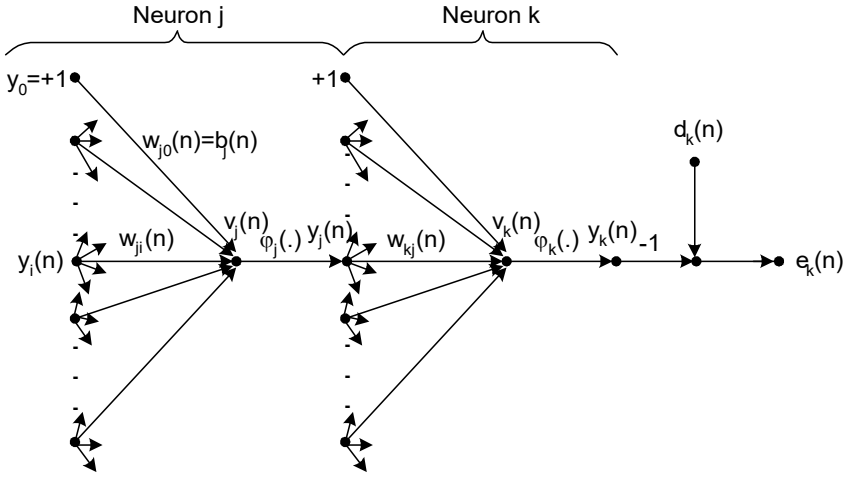
Trường hợp 1. Nơ ron j là một nút đầu ra

Khi nơ ron j nằm ở mức đầu ra của mạng, nó được cung cấp một đáp ứng mong muốn. Chúng ta có thể sử dụng công thức (2.2.1) để tính toán tín hiệu lỗi $e_j(n)$ tương ứng với nơ ron này (xem hình 2.5). Do xác định được $e_j(n)$, chúng ta dễ dàng tính được gradient cục bộ $\delta_j(n)$ theo công thức (2.2.14).

Trường hợp 2. Nơ ron j là một nút ẩn

Khi nơ ron j thuộc một mức ẩn của mạng, không có một đáp ứng mong muốn xác định nào cho nơ ron này. Như vậy, tín hiệu lỗi cho một nơ ron ẩn được xác định một cách đệ quy theo các tín hiệu lỗi của tất cả các nơ ron mà nơ ron đang xét kết nối trực tiếp. Hãy xem xét tình huống được mô tả trong hình 2.6. Theo công thức (2.2.14), chúng ta có thể tính lại gradient cục bộ của nơ ron j như sau

$$\delta_j(n) = -\frac{\partial \tau(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} = -\frac{\partial \tau(n)}{\partial y_j(n)} \varphi'_j(v_j(n)) \quad \text{, nơ ron j là ẩn} \qquad (2.2.15)$$



Hình 2.6 Đồ thị luồng tín hiệu chi tiết cho một nơ ron ẩn j nối với một nơ ron đầu ra k.

Để tính đạo hàm riêng $\partial\tau(n)/\partial w_{ji}(n)$, chúng ta có thể thực hiện như sau. Từ hình 2.6, chúng ta thấy rằng:

$$\tau(n) = \frac{1}{2}\sum_{k\in C} e_k^2(n), \quad \textit{nơ ron k là nút đầu ra} \qquad (2.2.16)$$

Đạo hàm hai vế công thức (2.2.16) theo tín hiệu chức năng $y_j(n)$, chúng ta có

$$\frac{\partial \tau(n)}{\partial y_j(n)} = \sum_k e_k \frac{\partial e_k(n)}{\partial y_j(n)} \qquad (2.2.17)$$

Tiếp theo, chúng ta sử dụng quy tắc tính đạo hàm và viết lại công thức (2.2.17) dưới dạng tương đương như sau

$$\frac{\partial \tau(n)}{\partial y_j(n)} = \sum_k e_k \frac{\partial e_k(n)}{\partial v_k(n)} \frac{\partial v_k(n)}{\partial y_j(n)} \qquad (2.2.18)$$

Tuy nhiên, từ hình 2.7, chúng ta nhận thấy

$$\begin{aligned} e_k(n) &= d_k(n) - y_k(n) \\ &= d_k(n) - \varphi_k(v_k(n)), \textit{ nơ ron k là một nút đầu vào} \end{aligned} \qquad (2.2.19)$$

Như vậy

$$\frac{\partial e_k(n)}{\partial v_k(n)} = -\varphi'_k(v_k(n)) \qquad (2.2.20)$$

Từ hình 2.7, chúng ta cũng thấy rằng đối với nơ ron k

$$v_k(n) = \sum_{j=0}^{m} w_{kj}(n) y_j(n) \qquad (2.2.21)$$

ở đó m là số đầu vào (bao gồm cả hệ số hiệu chỉnh) của nơ ron k. Lấy đạo hàm công thức (2.2.21) theo $y_j(n)$, ta có

$$\frac{\partial v_j(n)}{\partial y_j(n)} = w_{kj}(n) \qquad (2.2.22)$$

Từ các công thức (2.2.20), (2.2.22) và (2.2.18), chúng ta tính được đạo hàm riêng mong muốn

$$\begin{aligned} \frac{\partial \tau(n)}{\partial y_j(n)} &= -\sum_k e_k \varphi'(v_k(n)) w_{kj}(n) \\ &= -\sum_k \delta_k(n) w_{kj}(n) \end{aligned} \qquad (2.2.23)$$

ở đó $\delta_k(n)$ cho bởi công thức (2.2.14) với chỉ số k thay cho chỉ số j.

Cuối cùng, thay công thức (2.2.23) vào công thức (2.2.15), chúng ta được *công thức back-propagation* cho gradiient cục bộ $\delta_j(n)$ như sau

$$\delta_j(n) = \varphi'(v_j(n))\sum_k \delta_k(n) w_{kj}(n), \qquad (2.2.24)$$

Hình 2.6 thể hiện đồ thị luồng dữ liệu của công thức (2.2.24), với giả định rằng mức đầu ra bao gồm m_Lnơ ron.

Tóm tắt lại thuật toán

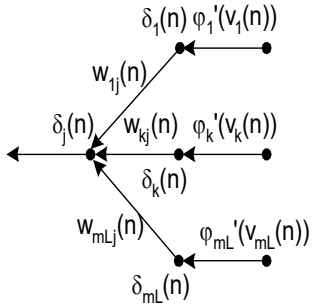
Bây giờ chúng ta tổng kết những gì mà chúng ta vừa tính được cho thuật toán back-propagation. Đầu tiên, hiệu chỉnh $\Delta w_{ji}(n)$ của trọng số $w_{ji}(n)$ mà nối nơ ron i với nơ ron j được xác định bởi *quy tắc delta* như sau:

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_j(n) \qquad (2.2.25)$$

Thứ hai, gradient cục bộ δ_j được xác định tuỳ theo việc nơ ron j là một nút đầu ra hay một nút ẩn:

Nếu nơ ron j là một nút đầu ra, $\delta_j(n)$ được tính theo công thức (2.2.14) bằng tích của đạo hàm $\varphi'(v_j(n))$ với tín hiệu lỗi e(n).

Nếu nơ ron j là một nút ẩn, $\delta_j(n)$ được tính một cách đệ quy theo công thức (2.2.24): bằng tích của đạo hàm riêng $\varphi'(v_j(n))$ với tổng các δ đã được nhân với các trọng số tương ứng ($\sum_k \delta_k(n) w_{kj}(n)$) của tất cả các nơ ron thuộc mức tiếp theo mà được nơ ron j nối tới.



Hình 2.7 Đồ thị luồng tín hiệu của một phần mạng tiến đa mức khi tín hiệu lỗi phản hồi trở lại

a. Hai giai đoạn tính toán của thuật toán

Trong việc áp dụng thuật toán back-propagation, có hai giai đoạn tính toán tách biệt nhau: giai đoạn *tiến (forward)*, và giai đoạn *lùi (backward)*.

Trong giai đoạn tiến, các trọng số synapse giữ nguyên không thay đổi trong toàn bộ giai đoạn, và các tín hiệu chức năng được tính toán dựa trên cơ sở từ nơ ron này tới nơ ron khác theo chiều tiến của mạng. Bắt nguồn từ đầu vào là một mẫu tích luỹ, quá trình tính toán cứ tiếp diễn cho tới khi tính được đầu ra của các nơ ron nằm trong mức đầu ra cuả mạng. Đầu ra này được so sánh với đáp ứng mong muốn $d_i(n)$, và thu được tín hiệu lỗi $e_j(n)$. Như vậy, giai đoạn tính toán tiến bắt đầu từ mức ẩn đầu tiên và kết thúc tại mức đầu ra để tính toán tín hiệu lỗi cho mỗi nơ ron tại mức này.

Trái lại, giai đoạn tính toán lùi bắt đầu tại mức đầu ra bằng cách chuyển tín hiệu lỗi ngược trở lại qua toàn bộ mạng theo từng mức nhằm tính toán một cách đệ quy δ (gradient cục bộ) cho mỗi nơ ron. Quá trình đệ quy này cho phép các trọng số synapse của mạng có thể được điều chỉnh theo quy tắc delta (công thức (2.2.25)). Bằng cách như vậy, sự điều chỉnh các trọng số synapse sẽ được lan truyền trên toàn bộ mạng.

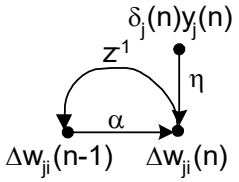
b. Tốc độ học

Thuật toán back-propagation cung cấp một phương pháp tính "xấp xỷ" cho việc lần tìm trong không gian trọng số (nhằm tìm ra các trọng số phù hợp nhất cho mạng). Chúng ta càng lấy giá trị của tham số tốc độ học nhỏ bao nhiêu thì sự thay đổi của trọng số synapse càng nhỏ bấy nhiêu và quỹ đạo trong không gian trọng số sẽ càng trơn. Tuy nhiên điều này lại làm cho tốc độ học chậm đi. Trái lại nếu chúng ta chọn tham số tốc độ học quá lớn, sự thay đổi lớn của các trọng số synapse có thể làm cho mạng trở nên không ổn định. Một phương pháp đơn giản để tăng tốc độ học mà tránh được nguy cơ không ổn định như trên là thay đổi quy tắc delta trong công thức (2.2.13) bằng cách sử dụng thêm một toán hạng moment như sau

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n-1) + \eta \delta(n) y_i(n) \qquad (2.2.26)$$

ở đó α thường được gọi là *hằng số moment*. Nó điều khiển vòng lặp quay lui hoạt động xung quanh Δw_{ji} như được minh hoạ trong hình 2.7 mà ở đó z^{-1} là toán

hạng đơn vị trễ. Đẳng thức (2.2.26) được gọi là *quy tắc delta tổng quát* mà quy tắc delta trong công thức (2.2.13) là một trường hợp đặc biệt (với α=0).



Hình 2.8 Đồ thị luồng tín hiệu minh hoạ tác dụng của hằng số moment α

Để thấy được ảnh hưởng của các hiệu chỉnh trong quá khứ đối với các trọng số synapse, chúng ta viết lại công thức (2.2.26) như một tổng chuỗi theo thời gian với chỉ số thời gian t như sau

$$\Delta w_{ji}(n) = \eta \sum_{t=0}^{n} \alpha^{n-t} \delta_j(t) y_i(t) \qquad (2.2.27)$$

Từ công thức (2.2.11) và (2.2.14), chúng ta thấy rằng toán hạng $\delta_j(n)y_i(n)$ chính bằng $-\partial\tau(n)/\partial w_{ji}(n)$. Chúng ta có thể viết lại công thức (2.2.27) như sau

$$\Delta w_{ji}(n) = -\eta \sum_{t=0}^{n} \alpha^{n-t} \frac{\partial \tau(n)}{\partial w_{ji}(n)} \qquad (2.2.28)$$

Dựa trên quan hệ này, chúng ta có thể đi một số nhận định sau:

Điều chỉnh hiện thời Δw_{ji} biểu diễn một tổng chuỗi theo thời gian với số mũ của hằng số α tăng theo thời gian. Để chuỗi này có thể hội tụ, hằng số moment phải nằm trong giới hạn $0 \le |\alpha| < 1$. Khi α bằng 0, thuật toán back-propagation hoạt động không có moment. α có thể nhận giá trị âm hoặc dương.

Khi đạo hàm $-\partial\tau(n)/\partial w_{ji}(n)$ có cùng dấu đại số trong các vòng lặp kế tiếp nhau, Δw_{ji} tăng theo hàm mũ, và như vậy trọng số w_{ji} được điều chỉnh bởi một giá trị lớn. Trong trường hợp này, việc đưa thêm moment vào thuật toán back-propagation có xu hướng làm cho nó hoạt động một cách nhanh chóng và đều đặn.

Khi đạo hàm $-\partial\tau(n)/\partial w_{ji}(n)$ đổi dấu ở những vòng lặp kế tiếp nhau, Δw_{ji} dao động về giá trị, và như vậy trọng số $w_{ji}(n)$ được điều chỉnh bởi một giá trị nhỏ. Trong trường hợp này, việc đưa thêm moment vào thuật toán back-propagation cho nó khả năng tự ổn định dần theo thời gian.

c. Các thời kì (Epoch)

Trong ứng dụng thực tế của thuật toán back-propagation, việc học thực hiện thông qua rất nhiều biểu hiện của một tập hợp các ví dụ tích luỹ cho trước. Như đã nhắc đến ở trên, một biểu diễn trọn vẹn của toàn bộ tập hợp tích luỹ trong quá trình học được gọi là một *Thời kỳ*. Quá trình học được duy trì trên một cơ sở từ *Thời kỳ* này sang *Thời kỳ* khác cho tới khi các trọng số và các mức hiệu chỉnh của mạng trở nên ổn định và lỗi bình phương trung bình trên toàn bộ tập hợp tích luỹ hội tụ tại một số giá trị cực tiểu. Sẽ tốt hơn nếu chúng ta *chọn một cách ngẫu nhiên thứ tự các thể hiện của các ví dụ tích luỹ* cho từng thời kỳ. Mặc dù điều này không được chứng minh một cách rõ ràng nhưng nó là một trong những kinh nghiệm vô cùng quan trọng đảm bảo cho hiệu quả của thuật toán.

Trong thực tế, đôi khi việc điều chỉnh trọng số của thuật toán back-propagation không được thực hiện ngay trong *giai đoạn lùi* của thuật toán mà sau mỗi Thời kỳ. Khi đó tích $\delta_j(n)y_j(n)$ trong công thức của quy tắc delta (2.2.25) sẽ được tính trung bình trên toàn bộ *Thời kỳ* cho mỗi nơ ron j.

d. Tiêu chuẩn dừng thuật toán

Nói chung, thuật toán back-propagation không thể xác định được là đã hội tụ hay chưa, và như vậy không có một tiêu chuẩn tuyệt đối nào cho việc dừng thuật toán. Tuy nhiên, vẫn có một số tiêu chuẩn có thể coi là chấp nhận được. Để tạo nên một tiêu chuẩn như vậy, một cách logic, chúng ta có thể nghĩ đến những thuộc tính đặc trưng của các *cực tiểu cục bộ* hay *toàn cục* của bề mặt lỗi. Đặt w* là một cực tiểu, có thể là toàn cục hay cục bộ. Một điều kiện cần để w* là một cực tiểu là vector gradient g(w) (đạo hàm riêng bậc nhất) của bề mặt lỗi theo vector trọng số w phải bằng không tại w=w*. Như vậy, chúng ta có thể xây dựng nên một tiêu chuẩn hội tụ nhạy cảm cho thuật toán back-propagation như sau:

Thuật toán back-propagation được xem là hội tụ khi độ lớn Euclide của vector gradient đạt tới một ngưỡng gradient đủ nhỏ.

Một hạn chế của tiêu chuẩn hội tụ này là thời gian học phải dài và cần phải tính toán vector gradient g(w).

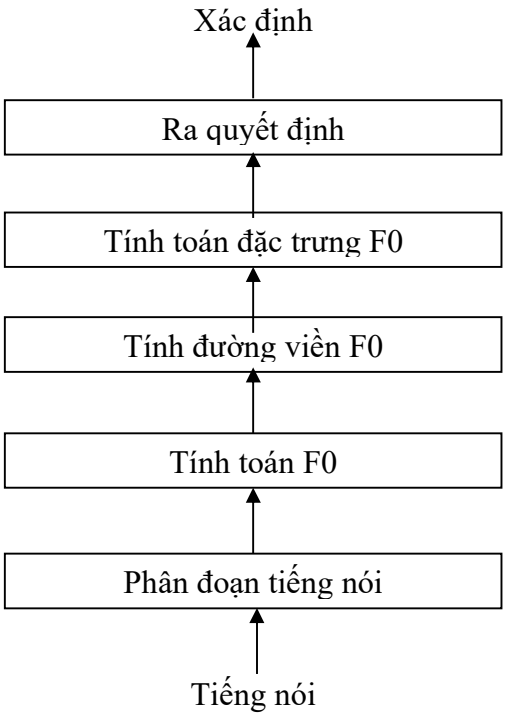
Một thuộc tính duy nhất nữa của cực tiểu là hàm giá $\tau_{av}(w)$ ổn định tại w=w*. Như vậy chúng ta lại có thể đưa ra một tiêu chuẩn khác như sau:

Thuật toán back-propagation được xem là hội tụ khi tỷ lệ thay đổi tuyệt đối của lỗi bình phương trung bình trên một Thời kỳ là đủ nhỏ.

Tỷ lệ thay đổi của lỗi bình phương trung bình được coi là đủ nhỏ nếu nó nằm trong giới hạn từ 0.1 đến 1 phần trăm trên một thời kỳ. Thật không may là tiêu chuẩn này có thể dẫn đến một sự kết thúc vội vàng của quá trình học.

Ngoài ra còn có phương pháp vừa học vừa thực hiện kiểm tra tính năng nhận dạng trên một tập hợp mẫu kiểm tra khác với tập hợp tích luỹ. Trong trường hợp này, thuật toán được xem là hội tụ khi nó đã tích luỹ đủ lâu và tính năng nhận dạng trên tập hợp kiểm tra đạt tới một giá trị cho phép.

2.3 Mô hình hệ thống nhận dạng ngôn ngữ nói tự động



Hình 2.9 Mô hình hệ thống nhận dạng ngôn ngữ nói tự động

Hệ thống nhận dạng ngôn ngữ tự động bao gồm năm giai đoạn chính được mô tả như hình 2.9:

- Phân đoạn tiếng nói: Chức năng của giai đoạn này là phân đoạn tín hiệu đầu vào liên tiếp của tiếng nói thành các phân đoạn tiếng nói rời rạc.

- Tính toán F0: Giai đoạn này chịu trách nhiệm tính toán F0 cho từng phân đoạn tiếng nói rời rạc bằng phương pháp AMDF.

- Tính đường viền F0: Giá trị F0 rút ra từ giai đoạn trước được tính toán, liên kết lại thành đường F0.

- Tính toán đặc trưng F0: Chức năng của giai đoạn này là tính toán hướng đi lên hoặc xuống của đường F0.

- Ra quyết định: Với các đặc trưng xuất phát từ giai đoạn trước, sử dụng mạng nơ ron lan truyền ngược để xác định ngôn ngữ.

2.4 Kết luận chương 2

Qua tìm hiểu cơ sở lý thuyết về phân tích dữ liệu tiếng nói và ứng dụng mạng nơ ron trong nhận dạng tiếng nói, chúng ta đã hiểu được cách máy tính xử lý và nhận dạng tiếng nói, qua đó xây dựng mô hình hệ thống nhận dạng ngôn ngữ nói tự động. Trong chương tiếp theo luận văn sẽ thử nghiệm và đánh giá chương trình nhận dạng tự động tiếng Việt và tiếng Pháp.

CHƯƠNG 3 - ỨNG DỤNG

3.1 Đặt vấn đề

Trong phần trước luận văn đã giới thiệu về kĩ thuật phân tích tiếng nói và trích rút đặc trưng. Tìm hiểu về mạng nơ ron và khả năng học cũng như điểm mạnh của mạng nơ ron trong các bài toán nhận dạng tiếng nói, xây dựng mô hình hệ thống nhận dạng ngôn ngữ nói tự động. Trong chương này luận văn thử nghiệm và đánh giá chương trình tự động nhận dạng tiếng nói tiếng Việt và tiếng Pháp. Mục đích của chương trình là nhằm nhận dạng được 2 ngôn ngữ cho đầu vào là một file hay nhiều file cùng một lúc với tỷ lệ chính xác cao trong thời gian ngắn.

3.2 Chi tiết hệ thống nhận dạng ngôn ngữ tự động phân biệt tiếng Việt và tiếng Pháp

3.2.1 Phân đoạn tiếng nói

Đầu vào liên tục của tín hiệu tiếng nói bao gồm các vùng im lặng và vùng tiếng nói. Sử dụng Thuật toán phát hiện điểm cuối của Rabiner và Sambur để phân đoạn tín hiệu giọng nói đầu vào thành các phân đoạn tiếng nói rời rạc.

Để có được hiệu suất tốt, thuật toán phân đoạn phải có một số thông số đặc biệt để tính toán:

- Những từ bắt đầu bằng hoặc kết thúc bằng âm vị năng lượng thấp.
- Những từ kết thúc bằng một âm bật vô thanh.
- Những từ kết thúc bằng âm mũi.
- Người nói kết thúc các từ với cường độ cao hoặc hơi thở ngắn (tiếng ồn).

Sử dụng phương pháp Zero Crossing Rate và Short-Term Energy để đo tín hiệu giọng nói sau 10ms trên các frames có độ dài 10ms (cho rằng 10 frames đầu tiên là nhiễu nền). Phương pháp này được sử dụng để tìm giá trị trung bình và phương sai của từng đặc trưng, những thống kê này được sử dụng để tính 3 ngưỡng:

- ITU (Upper Energy threshold): Ngưỡng năng lượng trên.

- ITL (Lower Energy threshold): Ngưỡng năng lượng thấp hơn.

- IZCT (Zero Crossing Rate threshold): Ngưỡng tỷ lệ vượt quá điểm không.

Mức năng lượng sau đó được tìm kiếm để tìm điểm giao nhau đầu tiên của ngưỡng trên ITU di chuyển về giữa đoạn từ mỗi đầu. Sau đó, chúng ta quay trở lại xuống điểm giao nhau gần nhất của ITL trong mỗi trường hợp. Quá trình này mang lại điểm cuối dự kiến gọi là N1 và N2. Sau đó di chuyển về phía cuối từ N1 và N2 không quá 25 frames, kiểm tra tỉ lệ vượt quá điểm không để tìm sự xuất hiện của số đếm trên ngưỡng IZTC. Nếu chúng không được tìm thấy, điểm cuối vẫn là ước tính ban đầu. Nếu tìm thấy ba lần xuất hiện, thì ước tính điểm cuối được di chuyển lùi (hoặc chuyển tiếp) đến thời điểm vượt ngưỡng đầu tiên.

3.2.2 Tính toán F0

Tính toán F0 cho từng phân đoạn tiếng nói rời rạc. Để tính F0 cho một phân đoạn tiếng nói rời rạc, chúng ta tính toán F0 cho các khoảng liên tiếp 10ms của từng phân đoạn tiếng nói. Cửa sổ phân tích 50ms và khoảng thời gian 10ms trên các frames được sử dụng để trích xuất F0 bằng phương pháp AMDF. Tần số cơ bản F0 được xác định là số mẫu m đưa ra phương trình tối thiểu (3.2.1).

$$D(m) = \Sigma \mid x(n) - x(n-m) \mid , n = 1,2,\ldots N \qquad (3.2.1)$$

x (n): tín hiệu tiếng nói của frame hiện tại.

N: Chiều dài của frame hiện tại được tính theo mẫu.

Hầu hết các phương pháp trích xuất F0, bao gồm phương pháp ADMF, đều mắc lỗi. Hầu hết trong số đó là lỗi gấp đôi hoặc chia đôi cao độ. Một phương pháp sửa lỗi đơn giản đã được đề xuất. Tần số cơ bản trung bình F_0^{Tb} cho tất cả các frame trong cửa sổ được tính toán. F_0, $2*F_0$, $F_0/2$ của frame hiện tại được so sánh với F_0^T và frame gần nhất với F_0^{Tb} được sử dụng làm giá trị F_0 được sửa. Tính toán F_0^{Tb} trong cửa sổ tín hiệu:

$$F_0^{Tb} = (1/5)\Sigma F_0(i) \quad i = 1,2,\ldots,5 \qquad (3.2.2)$$

$F_0(i)$: Tần số cơ bản của frame thứ i bên trong cửa sổ tín hiệu.

Cuối cùng, F0 được xác định theo phương trình (3.2.3)

$$F_0 = \begin{cases} F_0(1), & \min(F1, F2, F3) = F1 \\ 2*F_0(1), & \min(F1, F2, F3) = F2 \\ F_0(1)/2, & \min(F1, F2, F3) = F3 \end{cases} \quad (3.2.3)$$

Trong đó F1, F2, F3 được tính như sau:

$$F1 = |F_0(1) - F_0^{Tb}|$$
$$F2 = |2*F_0(1) - F_0^{Tb}|$$
$$F3 = |F_0(1)/2 - F_0^{Tb}|$$

Bằng cách dịch chuyển cửa sổ tín hiệu sang toàn bộ phân đoạn tiếng nói với khoảng thời gian 10ms, ta lấy được đường viền F0 của phân đoạn giọng nói rời rạc.

3.2.3 Tính đường viền F0

a. Phân đoạn đường viền cao độ

Bước đầu tiên của giai đoạn tính đường viền F0 là phân đoạn đường viền cao độ vào các phân đoạn định hướng lên hoặc xuống. Trong bước này, tôi sử dụng một quy trình động được mô tả như sau:

Sự thay đổi đường viền cao độ là vị trí của đường viền cao độ mà tại đó tồn tại kết thúc tối đa cục bộ.

Bước 1: Tìm kiếm đường viền cao độ ngay từ đầu để tìm thay đổi đầu tiên về cao độ.

Bước 2: Vị trí bắt đầu của đoạn đầu tiên là vị trí phát hiện thay đổi.

Bước 3: Tiêu chí tìm vị trí kết thúc của một đoạn là vị trí phát hiện thay đổi. Nếu vị trí kết thúc của đoạn hiện tại được phát hiện, đi đến bước 4.

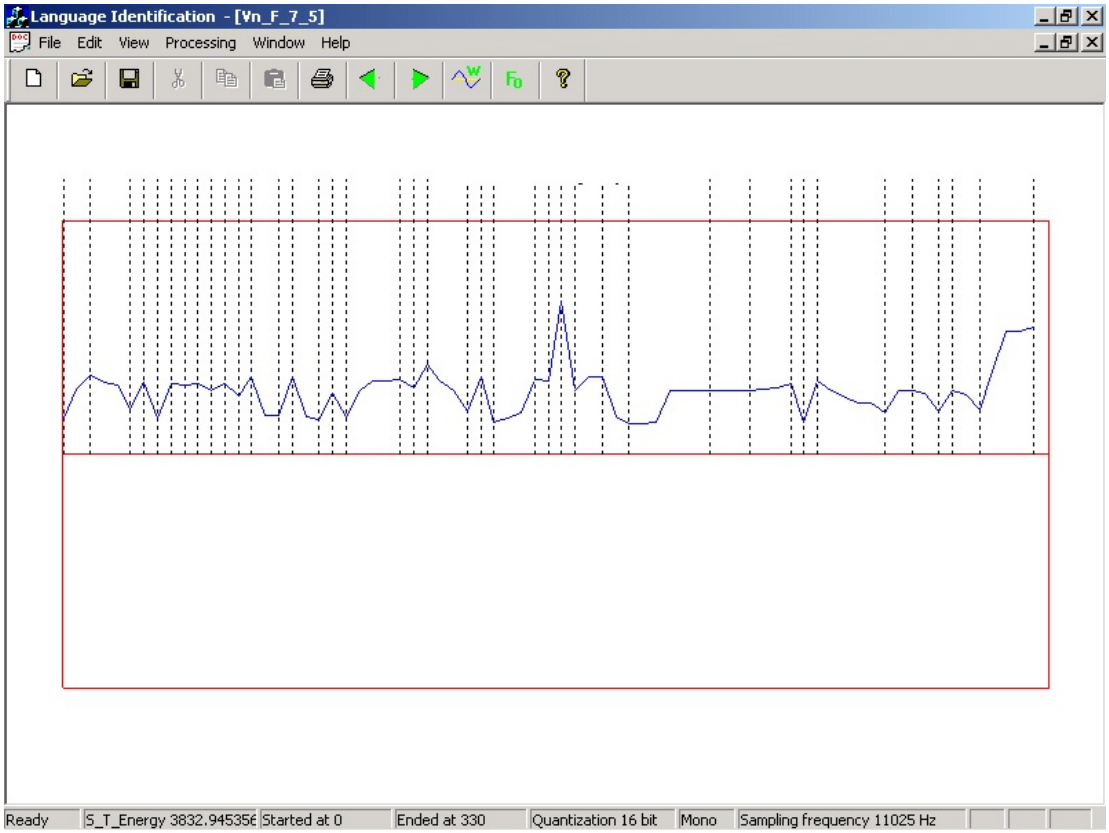
Bước 4: Lưu các tham số phân đoạn hiện tại (bao gồm vị trí bắt đầu, vị trí kết thúc). Thiết lập các tham số ban đầu của phân đoạn mới.

- Đặt vị trí bắt đầu của phân đoạn mới tương đương với vị trí kết thúc của đoạn liền kề.

- Đặt vị trí kết thúc của phân đoạn mới tương đương với vị trí kết thúc của phân đoạn liền kề. Lần lượt, đi đến bước 5.

Bước 5: Kiểm tra xem vị trí hiện tại có phải là kết thúc của đường viền cao độ. Nếu đúng, hãy đến bước 6, nếu không thì chuyển sang bước 3.

Bước 6: Thuật toán kết thúc



Hình 3.1 Ví dụ về kết quả từ quy trình động

b. Ước tính các phân đoạn cao độ theo dòng

Bước thứ hai của giai đoạn ước tính viền F0 là tính các phân đoạn cao độ xuất phát từ bước trước bằng một tập hợp các dòng thứ nhất. Thuật toán ước tính một phân đoạn cao độ theo dòng thứ nhất dựa trên phương pháp lỗi bình phương trung bình.

Giả sử chúng ta có một tập hợp các quan sát M (x_i,y_i), i = 1,2,… M. Bây giờ chúng tôi muốn ước tính tập hợp này bằng một dòng thứ nhất như trong phương trình (3.2.4)

$$f(x) = a_0 + a_1x \qquad (3.2.4)$$

Lỗi bình phương trung bình giữa y_i và giá trị được tính với (3.2.4) như trong phương trình (3.2.5).

$$e_i^2 = [\ y_i - f(x_i)] \qquad (3.2.5)$$

Đối với các quan sát M, sai số tổng là như trong phương trình (3.2.6).

$$E = \Sigma\ e_i^2 = \Sigma\ \{\ y_i - [\ a_0 + a_1\ x_i]\ \}^2 \qquad i = 1,2,\ldots,M \qquad (3.2.6)$$

E là hàm của biến a_0 và a_1. Giá trị chính xác của các biến a_0 và a_1 đưa ra phương trình (3.2.6) tối thiểu. Giá trị của các biến a_0 và a_1 được xác định bằng cách giải phương trình (3.2.7) và (3.2.8).

$$\frac{\partial E}{\partial a_0} = 0 \qquad (3.2.7)$$

$$\frac{\partial E}{\partial a_1} = 0 \qquad (3.2.8)$$

Mỗi phân đoạn cao độ bây giờ có thể được biểu thị bằng một cặp giá trị (a_0^i, a_1^i). Do đó, một đường bao cao độ được xấp xỉ bằng một tập hợp các dòng thứ nhất, được biểu thị bằng một tập hợp S của các cặp giá trị.

$$S = \{(a_0^i, a_1^i)\}\ ,\ i=1,2,\ldots,K \qquad (3.2.9)$$

K- Số dòng thứ tự đầu tiên

3.2.4 Tính toán đặc trưng F0

Một số tính năng hữu ích để nhận dạng ngôn ngữ được trích xuất theo quy trình sau:

- Chia số thực (-∞,+∞) thành 20 vùng:

(-∞,-9], (-9,-8], …, (-1,0), [0,1), …,[8,9), [9,+∞)

Các vùng dương được ký hiệu là P_0, P_1, …,P_9 theo thứ tự của [0,1), …,[8,9), [9,+∞).

Các vùng âm được ký hiệu là N_0, N_1, …, N_9 theo thứ tự (-1,0),…[-9,-∞).

- Đối với một ngôn ngữ nhất định, độ dốc của các đường phân phối trên các vùng trên. Tính số lượng dòng trong mỗi vùng.

- Tính các tỷ lệ để đánh giá sau cùng:

R_i^P = Số dòng trong vùng P_i / Số dòng trong tất cả các vùng.

R_i^N = Số dòng trong vùng N_i / Số dòng trong tất cả các vùng.

Tỷ lệ trên khác nhau từ ngôn ngữ này đến ngôn ngữ khác. Chúng ta có thể sử dụng các tỷ lệ này để xác định ngôn ngữ.

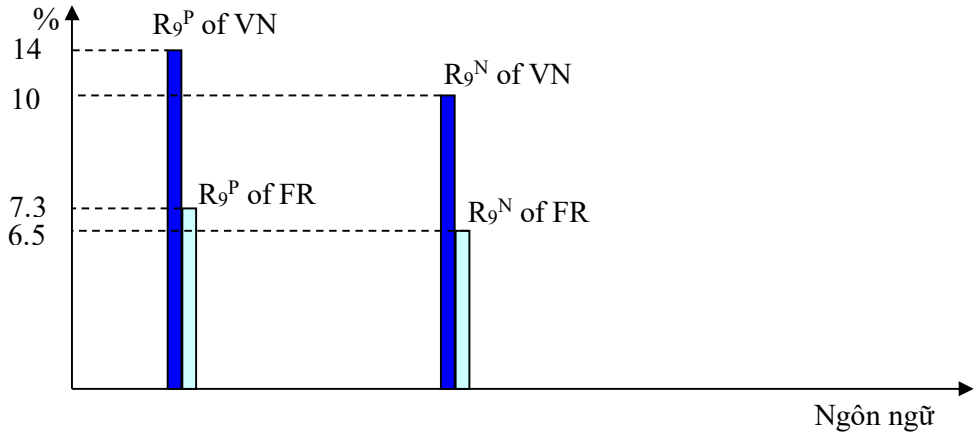
3.2.5 Ra quyết định

Trong giai đoạn này, thiết kế một mạng nơ ron để hoàn thành nhiệm vụ nhận dạng ngôn ngữ.

Tác giả đã thực hiện một số thí nghiệm trong đó tín hiệu giọng nói với độ dài 30 phút đã được quan sát và đưa ra một số kết luận hữu ích cho việc thiết kế giai đoạn này.

- Tỷ lệ R_9^P và R_9^N khác nhau đáng kể giữa tiếng Việt và tiếng Pháp (xem hình 3.2).

- Sử dụng tỷ lệ R_9^P và R_9^N chúng ta có thể phân biệt giữa tiếng Việt và tiếng Pháp. Nói cách khác, tỷ lệ R_9^P và R_9^N chứa hầu hết thông tin để xác định tiếng Việt và tiếng Pháp.



Hình 3.2 So sánh giữa R_9^P và R_9^N của tiếng Việt và tiếng Pháp

Sử dụng Mạng nơ ron truyền ngược để phân loại các điểm của (R_9^P, R_9^N).

- Số lớp: Quan sát kết quả thí nghiệm, tác giả quyết định sử dụng mạng BPN gồm 3 lớp.

- Số lượng đầu vào: Sử dụng tỷ lệ R_9^P và R_9^N, do đó chúng ta có 2 đầu vào.

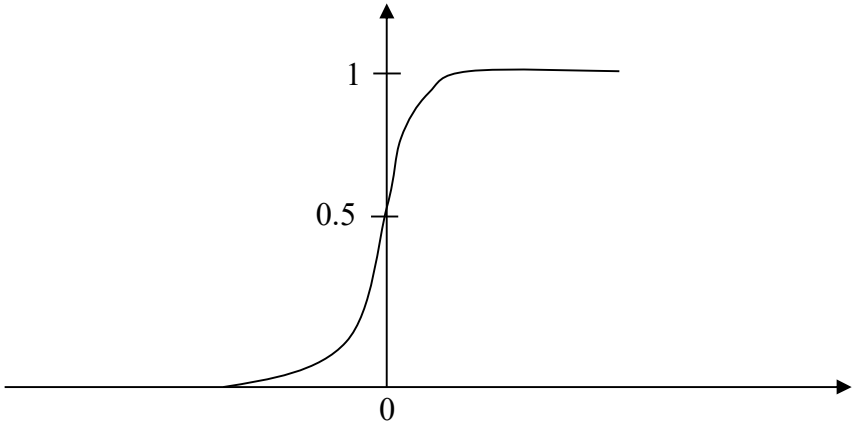
- Số lớp ẩn: 2.

- Số lượng đầu ra: Đầu ra có hai trạng thái, do đó sử dụng một nơron trong lớp đầu ra.

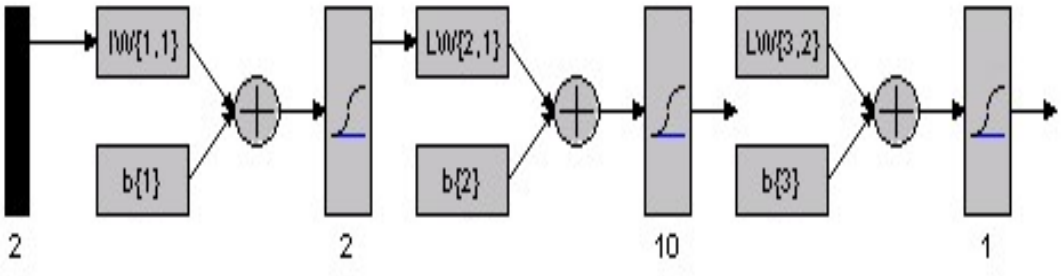
- Chức năng chuyển giao: Sử dụng logarit chuẩn làm chức năng chuyển giao cho tế bào nơ ron ở cả ba lớp. Hình dạng của hàm logarit chuẩn được mô tả trong hình 3.3.

$$f(x) = \frac{1}{1+e^{-x}}$$

Mạng nơ ron truyền bá ngược được mô tả trong hình 3.4.



Hình 3.3 Hình dáng của hàm logarit chuẩn



Hình 3.4 Mạng nơ ron truyền bá ngược sử dụng trong giai đoạn Ra quyết định

3.3 Chương trình nhận dạng ngôn ngữ tự động tiếng Việt và tiếng Pháp

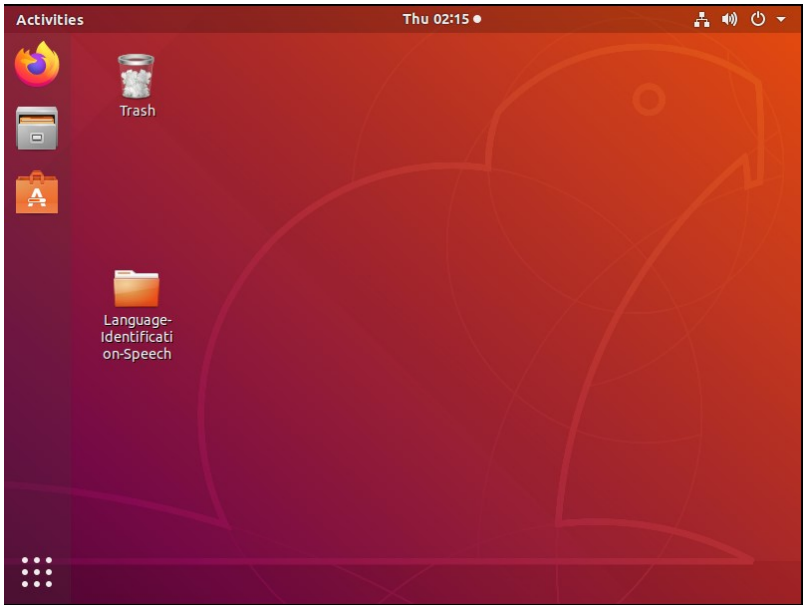
Chương trình nhận dạng ngôn ngữ có tên là "language identifcation speech" được viết trên ngôn ngữ lập trình Python, sử dụng thư viện có sẵn trong PIP và chạy trên hệ điều hành Ubuntu.

Chương trình có sử dụng một số phần mềm hỗ trợ:

- FFmpeg: để ghi, chuyển đổi và truyền phát âm thanh và video.

- SoX: phần mềm chỉnh sửa âm thanh đa nền tảng.

- youtube-dl: là một chương trình dòng lệnh để tải xuống video từ youtube.com và một vài trang web khác.



Hình 3.5 Chương trình nhận dạng

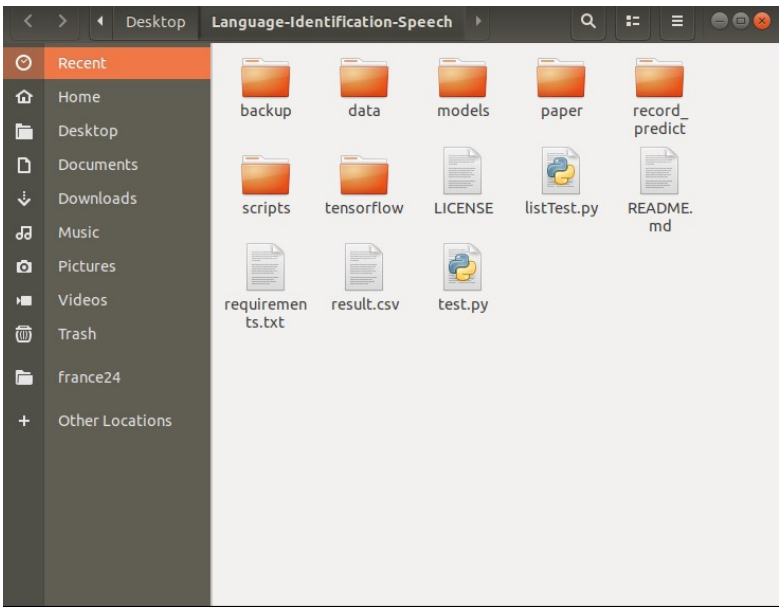
Giao diện chương trình (hình 3.6):

- test.py: để chạy file riêng lẻ cần kiểm tra ngôn ngữ.

- listTest.py: để chạy nhiều file cần kiểm tra ngôn ngữ.

- data: thư mục chứa các file cần chạy.

- tensorflow: chứa code chương trình.

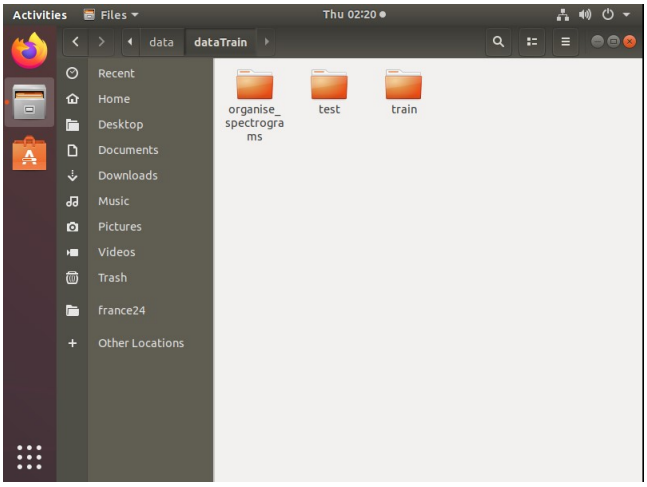


Hình 3.6 Giao diện chương trình

Trong thư mục dataTrain trong data:

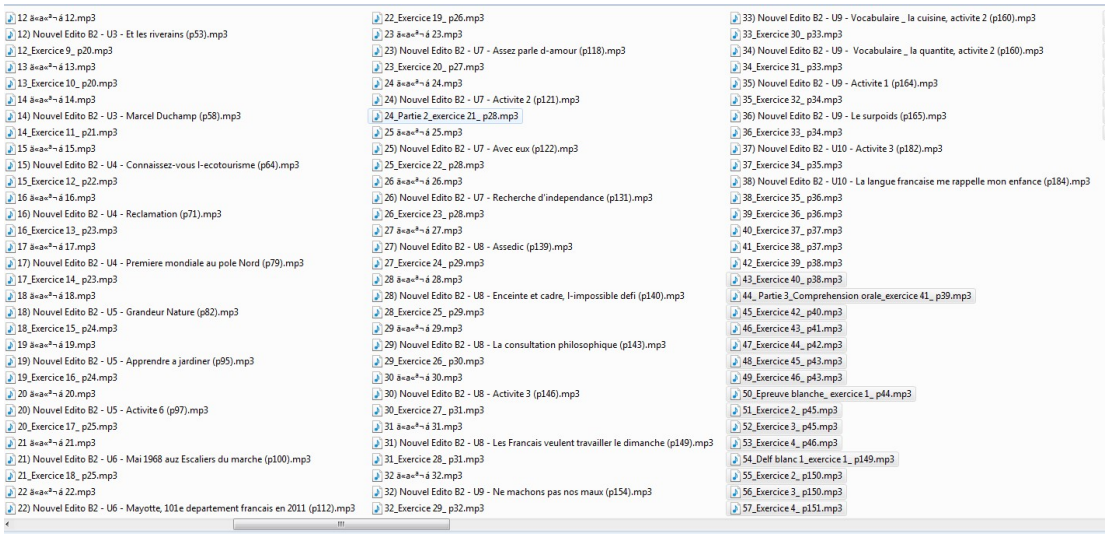
- organise_spectrograms: Nơi lưu các ảnh phổ âm thanh mà chương trình xử lý được.

- test: Nơi lưu các dữ liệu cần phát hiện ngôn ngữ, có thể nạp dữ liệu vào là file video, chương trình sẽ tự động chyển sang định dạng đuôi wav để chạy.



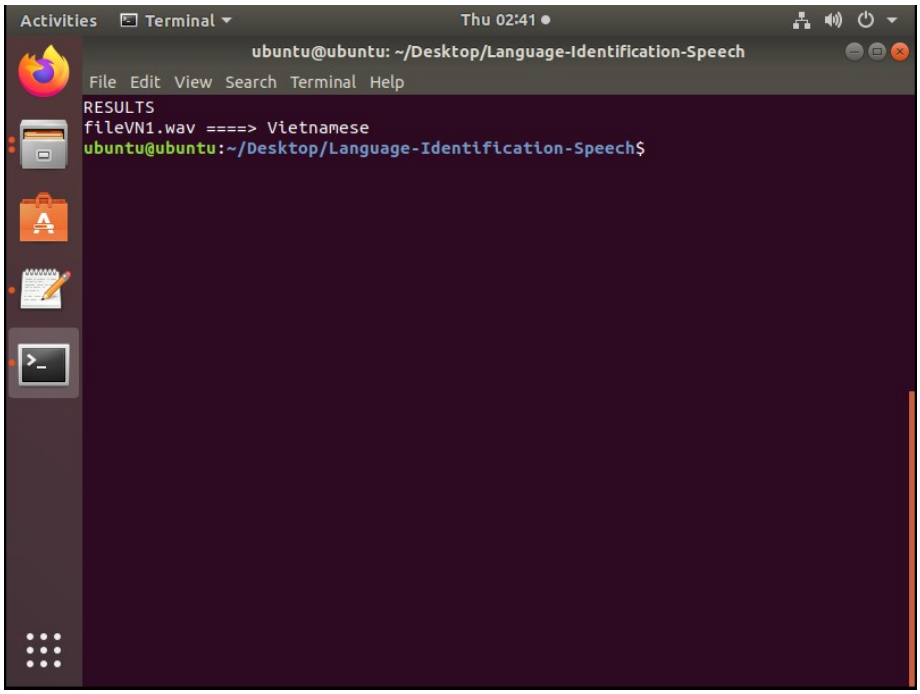
Hình 3.7 Thư mục datatrain của chương trình

- train: Nơi chứa cơ sở dữ liệu đã được training của chương trình, có tiếng Việt và tiếng Pháp, chúng ta có thể thêm dữ liệu training tại đây.

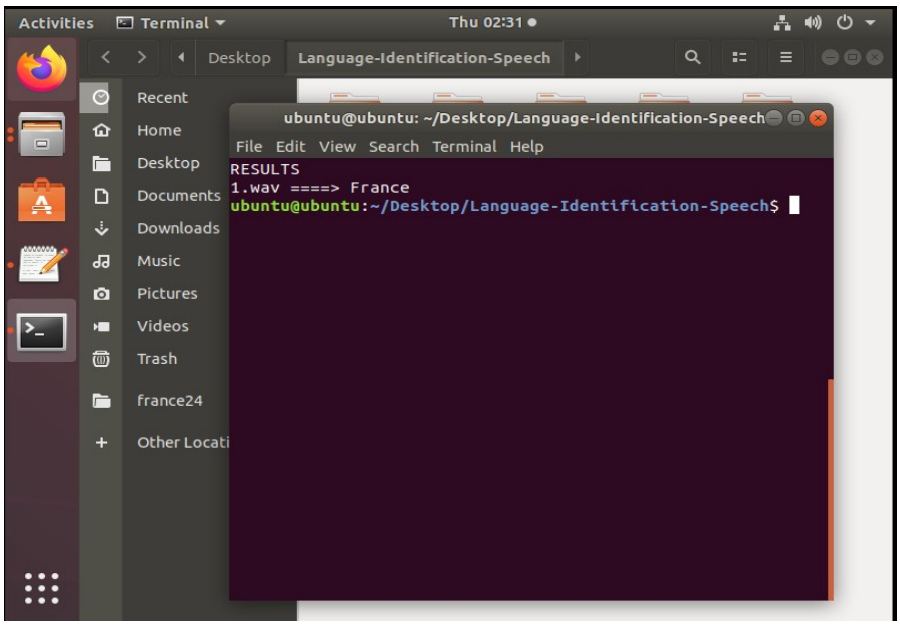


Hình 3.8 Hình ảnh cơ sở dữ liệu tập đào tạo

Chạy chương trình với từng file riêng lẻ cho ra kết quả tiếng Việt và tiếng Pháp trong 2 giây.



Hình 3.9 Hình ảnh kết quả chương trình với file tiếng Việt



Hình 3.10 Hình ảnh kết quả chương trình với file tiếng Pháp

Chạy chương trình với một folder với nhiều file cho kết quả trả ra file exel, thời gian tùy thuộc vào số lượng file chạy nhiều hay ít, trung bình 2 giây 1 file.

Clipboard | Font

A1 | FILE NAME

	A	B	C	D	E
1	FILE NAME	LANGUAGE IDENTIFICATION			
2	file136.wav	Vietnamese			
3	file96.wav	Vietnamese			
4	file175.wav	Vietnamese			
5	file139.wav	Vietnamese			
6	file184.wav	Vietnamese			
7	file6.wav	France			
8	file36.wav	France			
9	file83.wav	France			
10	file8.wav	France			
11	file147.wav	Vietnamese			
12	file138.wav	France			
13	file46.wav	France			
14	file124.wav	France			
15	file171.wav	France			
16	file38.wav	France			
17	file40.wav	France			
18	file47.wav	Vietnamese			
19	file141.wav	Vietnamese			
20	file160.wav	Vietnamese			
21	file30.wav	Vietnamese			
22	file123.wav	France			
23	file162.wav	France			
24	file74.wav	Vietnamese			
25	file65.wav	France			
26	file156.wav	France			
27	file13.wav	Vietnamese			
28	file82.wav	France			
29	file179.wav	France			
30	file172.wav	France			
31	file79.wav	France			

Hình 3.11 Hình ảnh kết chương trình nhiều file đầu vào

3.4 Đánh giá kết quả

Phương pháp thử nghiệm chia tập dữ liệu mẫu làm 2 tập, 1 tập để đào tạo mạng nơ ron (gồm 450 file tiếng Pháp, 200 file tiếng Việt) và một tập thử nghiệm (gồm 164 file tiếng Pháp, 186 file tiếng Việt).

Độ chính xác khi cho đầu vào trộn lẫn cả 2 ngôn ngữ gồm 100 file tiếng Pháp và 100 file tiếng Việt là: 84%.

Độ chính xác khi cho đầu vào tập thử nghiệm ngôn ngữ tiếng Việt là: 80%.

Độ chính xác khi cho đầu vào tập thử nghiệm ngôn ngữ tiếng Pháp là: 87%.

3.5 Kết luận chương 3

Đây là chương cuối và cũng là một trong những chương quan trọng nhất của luận văn để thể hiện được ứng dụng của tần số cơ bản của tiếng nói vào nhận dạng ngôn ngữ. Thử nghiệm với chương trình nhận dạng ngôn ngữ tiếng Việt và tiếng Pháp với kết quả có độ tin cậy cao trong thời gian ngắn.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết quả của luận văn

Luận văn đã giới thiệu những khái niệm và lý thuyết cơ bản về nguồn gốc của âm thanh, bộ máy phát âm, cũng như cơ chế phát âm, các đặc tính âm học của tiếng nói. Luận văn cũng đã giới thiệu các đặc điểm của ngôn ngữ tiếng Việt và tiếng Pháp, giới thiệu phân tích dữ liệu tiếng nói, mạng nơ ron ứng dụng trong nhận dạng tiếng nói, đặc biệt là đặc trưng của tần số cơ bản của tiếng nói với nhận dạng ngôn ngữ.

Song song với nghiên cứu và tìm hiểu lý thuyết luận văn đã thử nghiệm bước đầu là nhận dạng 2 ngôn ngữ tiếng Việt và tiếng Pháp với kết quả nhanh và chính xác cao.

2. Định hướng phát triển

Hướng nghiên cứu tiếp theo của luận văn sẽ tập trung vào việc nghiên cứu các phương pháp nâng cao độ chính xác của chương trình. Qua đặc trưng của tần số cơ bản trong tiếng nói đã trình bày thì còn rất nhiều bài toán có thể tìm hiểu và nghiên cứu thêm trong tương lai như tự động phân biệt giới tính, vùng miền cũng như phân biệt nhiều hơn 2 ngôn ngữ. Với sự hạn chế nhiều mặt về kiến thức cũng như thời gian, đồ án chắc chắn sẽ không thể tránh khỏi những thiếu xót, tôi tin rằng nếu được đầu tư thêm thời gian và được sự hỗ trợ thêm về kiến thức của các thầy cô giáo và các bạn, đồ án sẽ hoàn thành ở mức tốt hơn nữa và trở thành một sản phẩm có tình hoàn thiện cao.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Dương Tử Cường, *"Xử lý tín hiệu số"*, Nhà xuất bản Học viện Kỹ thuật quân sự, 2004.

[2] Ngô Minh Dũng, Đặng Văn Chuyết, *"Khảo sát tính ổn định của một số đặc trưng ngữ âm trong nhận dạng người nói",* 2010.

[3] Nguyễn Viết Sơn, "*Caracteristiques des voyelles et consonnes finales Vietnamiennes: Aspect statiques et dynamiques, Maison d'edition Presses Academiques Francophones"*, 2012.

[4] Bùi Đức Thịnh, *"Văn phạm Việt Nam"*, Culture Publisher, 1996.

[5] Chuya China Bhanja, Dipjyoti Bisharad, Rabul Hussain Laskar, *"Automatic Classification of Indian Language into Tonal and Non-tonal Categories Using Cascade Convolutional Neural Networks (CNN)-Long-Short-Term Memory (LSTM) Recurrent Neural Networks",* International Conference on Signal Processing and Communications (SPCOM), 2018.

[6] Jiangxiong Zhou & Shuichi ITAHASHI, *"Feature extraction for spoken language discrimination using speech fundamental frequency"*, Proc. IWSP.

[7] Key Margarethe Berkling, *"Automatic language identification with sequences of language independent phoneme clusters"*, PhD thesis, 1996.

[8] Liang Wang, Eliatham by Ambikairajah, Eric H.C.Choi, *"Automatic Tonal and Non-Tonal Language Classification and Language Identification Using Prosodic Infomation"*, 15th European Signal Processing Conference, 2007.

[9] Liang Wang, Eliathamby Ambikairajah, Eric H.C.Choi, *"Automatic language recognition with tonal and non-tonal language pre-classification"*, 15th European Signal Processing Conference, 2007.

[10] MICA speech database, Hanoi University of Technology.

[11] Richard E.Crandall, *"Topics in advanced scientific computation"*, Springer-Verlag, 1996.

[12] Rabiner L.R., Shafer R.W., *"Digital Processing of Speech Signal"*, Prentic Hall, 1978.

[13] TimKientzle, *"A programer guide to sound"*, Addison – Wesley, 1996.

[14] Yeshwant K. Muthusamy et all, *"Automatic language identification: A Review/Tutorial"*, OGI.

[15] Y. Vamsi, *"Robust speech recognition system for indian languages"*, Hyderabad institute, 2003.