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

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

HÀ NỘI – NĂM 2020

Luận văn được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. HÀ HẢI NAM

Phản biện 1: PGS.TS. NGUYỄN HẢI CHÂU

Phản biện 2: PGS.TS. BÙI THU LÂM

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 09 giờ ngày 20 tháng 6 năm 2020

Có thể tìm hiểu luận văn tại:
- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

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

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.

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

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, bản chất âm thanh là một dao động có tần số, con người có thể cảm nhận được từ dao động này.

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. Vòm miệng là các nếp cơ chuyển động.

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 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.

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

Trong 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.

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.

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.

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.

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* $-\infty < n < \infty$

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.

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,... 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.

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.

- 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.

g. Biên độ

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].

g. 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

1.2.1. Đặc điểm của ngôn ngữ tiếng Việt

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.

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.

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/

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/.

g. Hệ thống thanh điệu

- 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.

- 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.

- 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.

- 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.

- 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.

- 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.

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 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. 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. Các dạng số nhiều của các từ tiếng Pháp thường được tạo bằng cách thêm *s* hoặc *x* vào các từ số ít. Trong tiếng Pháp có năm dấu phụ (*à, â, é, è, ê, ë, î, ï, ô, ö, û, ù và ç*), chúng được đặt trên các nguyên âm hoặc dưới chữ c để chỉ ra sự thay đổi trong cách phát âm.

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 không có trọng âm được tổ chức lại xung quanh những từ có trọng âm.

d. Thành lập nhóm trọng âm

e. Thành lập nhóm ngữ điệu

g. Các gói và nhóm ngôn điệu

h. Nhóm ngôn điệu và cấu trúc cú pháp

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 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

2.1. Phân tích dữ liệu tiếng nói

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:

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

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)

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

Xét chuỗi vi sai sau:

$$d(n) = x(n) - x(n-k) \tag{2.1.7}$$

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:

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

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.

c. Trích chu kỳ cơ bản bằng AMDF

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.

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

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:

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

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

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

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)

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: $P_x(\omega) = P_x(\omega)P_h(\omega)$.

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.

e. Phép tính log năng lượng

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:

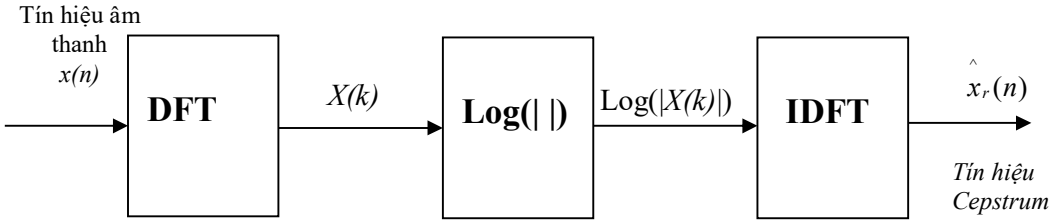
$$y_t^{(m)}(k) = \sum_{m=1}^{M} \{\log|Y_t(m)|\}.\cos\left(k\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right), \qquad k = 0,1,...,L \qquad (2.1.31)$$

h. Hệ số dữ liệu và năng lượng

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:

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

i. Phân tích Cepstrum



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).

2.2.2. Luật học của mạng nơ ron

a. Luật học có giám sát

b. Luật học không có giám sát

c. Luật học tăng cường

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 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 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:

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

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:

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_j(n) \qquad (2.2.12)$$

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 bằng tích của đạo hàm $\varphi'(v_j(n))$ với tín hiệu lỗi e(n) theo công thức:

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

Nếu nơ ron j là một nút ẩn, $\delta_j(n)$ được tính một cách đệ quy theo công thức:
$\delta_j(n) = \varphi'(v_j(n))\sum_k \delta_k(n)w_{kj}(n)$, 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.

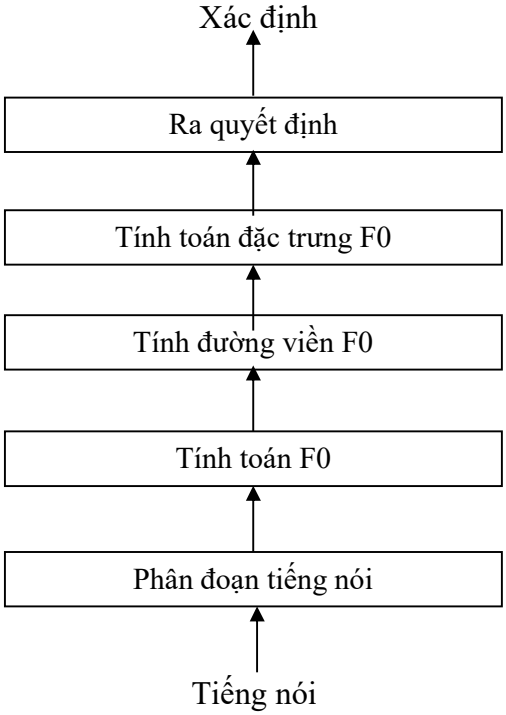
a. Hai giai đoạn tính toán của thuật toán

b. Tốc độ học

c. Các thời kỳ (Epoch)

d. Tiêu chuẩn dừng thuật toán

2.3. Mô hình hệ thống nhận dạng ngôn ngữ 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, 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 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:

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

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)

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

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

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

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 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

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)

$$f(x) = a_0 + a_1 x \qquad (3.2.4)$$

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.

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

K- Số dòng thứ tự đầu tiên

3.2.4. Tính toán đặc trưng F0

Quá 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.

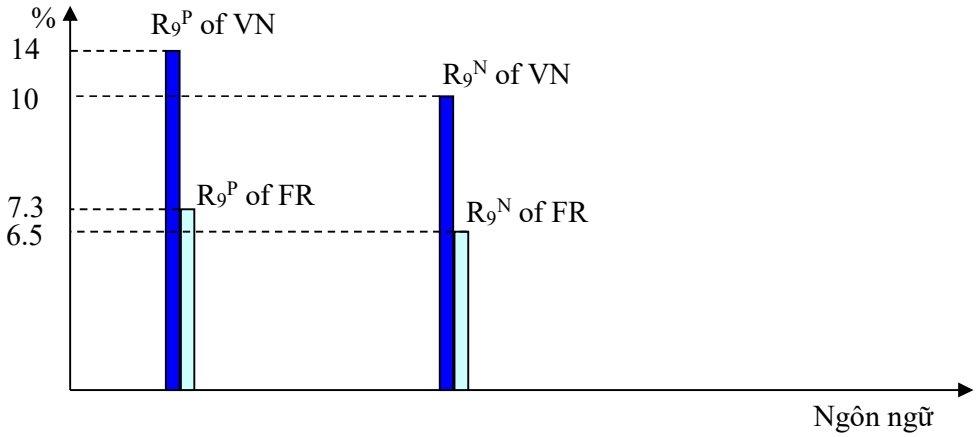
3.2.5. Ra quyết định

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.

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.

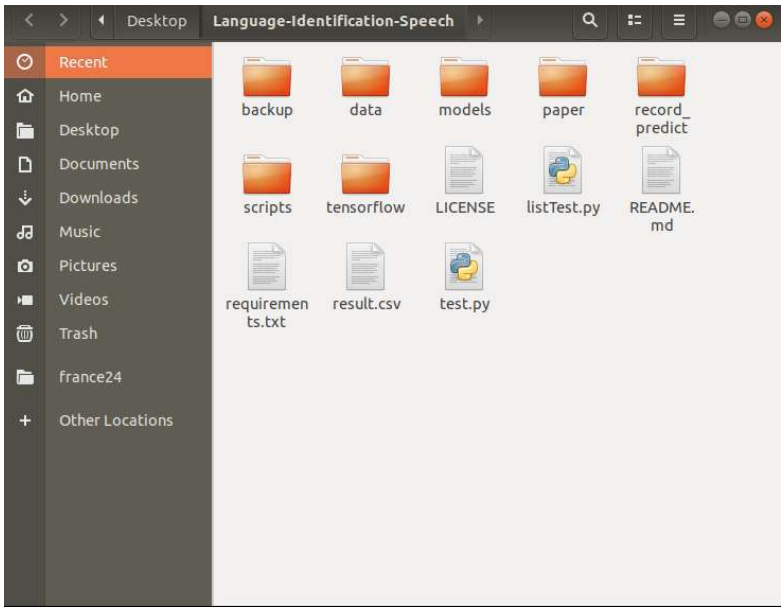
Giao diện chương trình (hình 3.6):

- 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.



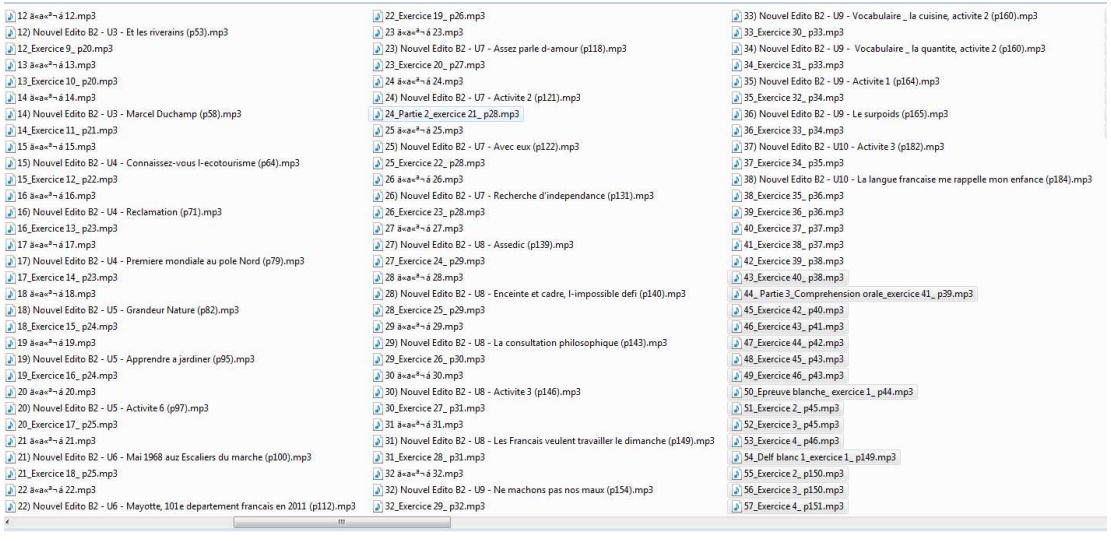
Hình 3.6 Giao diện chương trình

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

- train: Nơi chứa cơ sở dữ liệu tập đào tạo (training) của chương trình, có 450 file tiếng Pháp và 186 file tiếng Việt.



Hình 3.8 Hình ảnh cơ sở dữ liệu tập đào tạo

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 từng ngôn ngữ là: 80% với tiếng Việt và 87% với tiếng Pháp.

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%.

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.