BÁO CÁO BÀI TẬP LỚN SPEECH PROCESSING
NHẬN DIỆN NGƯỜI NÓI VÀ ỨNG DỤNG

Thành viên nhóm: Phạm Lê Việt Anh - 17021180
Nguyễn Tiến Đạt - 17021185
Phạm Thanh Tùng - 17020042
Nguyễn Huy Hoàng - 17021191
Nguyễn Tuấn Anh - 15020971

1. Mô tả bài toán

Xây dựng mô hình nhận diện người nói. Kết hợp với mô hình nhận diện một số từ để tạo ứng dụng cụ thể với các chức năng: khóa màn hình máy tính, tìm kiếm người nói trên Google, đăng nhập và đăng xuất trang web.

2. Dữ liệu

Bài toán yêu cầu xây dựng 2 mô hình. Một mô hình nhận diện người nói và một mô hình nhận diện từ.

Đối với mô hình nhận diện người nói, dữ liệu đầu vào là các đoạn âm thanh độ dài trong khoảng 8 đến 15 giây chứa câu bất kỳ do đối tượng nói. Cụ thể có phân phối như sau:

Người nói	Tập dữ liệu huấn luyện	Tập dữ liệu thử nghiệm
Việt Anh	35	15
Tiến Đạt	35	15
Thanh Tùng	35	15
Huy Hoàng	35	15

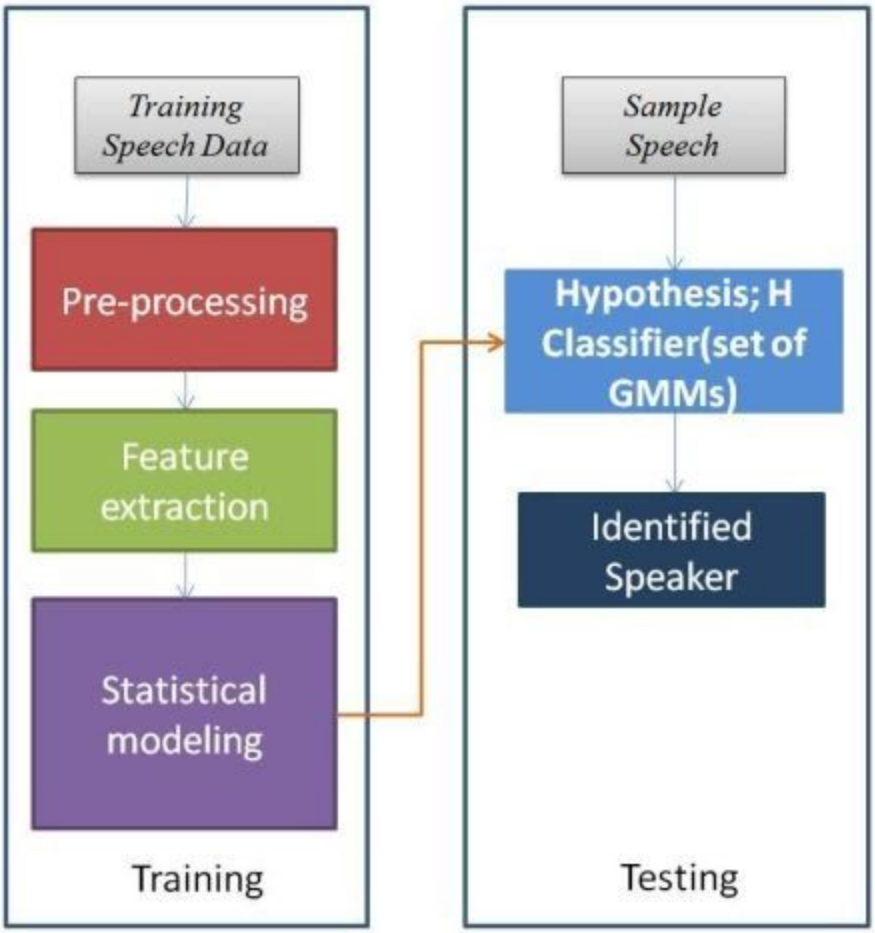
Tuấn Anh	35	15
Vân Dung	35	15
Tự Long	35	15
Công Lý	35	15
MC Thảo Vân	35	15
Trường Giang	35	15

Mô hình thứ hai là mô hình nhận diện từ với dữ liệu đầu vào của mô hình là các file âm thanh chứa các từ ghép mô tả các yêu cầu của người sử dụng với ứng dụng. Cụ thể là các từ: “đăng nhập”, “đăng xuất”, “khóa máy”, “tìm kiếm”. Với mô hình thứ hai này, các thành viên trong nhóm sẽ ghi âm mỗi từ 20 mẫu. Trong đó 80% được sử dụng để huấn luyện mô hình và 20% dùng làm tập thử nghiệm. Số lượng file cụ thể cho tập huấn luyện và tập thử nghiệm của mỗi từ như sau:

Từ	Tập huấn luyện	Tập thử nghiệm
đăng nhập	60	20
đăng xuất	60	20
khóa máy	60	20
tìm kiếm	60	20

3. Xây dựng mô hình nhận dạng người nói

Mô hình tổng quát



Trích xuất đặc trưng MFCC

Mỗi file dữ liệu âm thanh được xử lý để lấy 20 đặc trưng mfcc với win_length = 25ms và hop_length = 10ms. Sau đó các đặc trưng mfcc này được chuẩn hóa bằng cách trừ đi giá trị mean của chúng. Sau đó lấy giá trị mfcc sau trừ đi trước để được delta1. Lấy giá trị delta1 sau trừ đi trước để có được delta2. Nối các đặc trưng trên lại, ta có:

20 mfcc + 20 delta1 + 20 delta2 (delta của delta) = 60 đặc trưng

Mô hình hóa người nói bằng gaussian mixture model (GMM)

Có hai nguyên nhân chính khiến cho gaussian mixture model được sử dụng cho mô hình hóa người nói. Người ta thấy rằng tiếng nói cũng được tạo thành từ nhiều lớp

âm thanh khác nhau, được tạo thành khi đi qua lưỡi, thanh quản, miệng tạo ra nguyên âm, phụ âm, hơi khác nhau. Mặc khác, việc sử dụng GMM cho phép ta biểu diễn số lượng lớn những mô hình phân phối khác nhau tương ứng với những người nói khác nhau. Do đó, GMM có thể được sử dụng để mô hình hóa các người nói khác nhau.

Việc xây dựng mô hình người nói được dựa trên các vectors MFCCs được lấy từ giai đoạn rút trích đặc trưng. Phương pháp thường được sử dụng đó là phương pháp maximum likelihood nhằm tìm những hệ số của mô hình gaussian sao cho xác suất của các vector huấn luyện là cao nhất.

Để xây dựng mô hình GMM để nhận dạng người nói, ta dùng hàm GMMHMM trong thư viện hmmlearn. Với các thông số chung như sau:

- n_mix = 2. Thể hiện 2 miền giọng nói khác nhau trong dữ liệu (nam, nữ)
- random_state = 42.
- n_iter = 1000. Số lần lặp tối đa
- verbose = true.
- n_components = 16
- Các thông số khác để mặc định

4. Xây dựng mô hình nhận dạng từ

Trích xuất đặc trưng MFCC

Mỗi file dữ liệu âm thanh được xử lý để lấy 12 đặc trưng mfcc với win_length = 25ms và hop_length = 10ms. Sau đó các đặc trưng mfcc này được chuẩn hóa bằng cách trừ đi giá trị mean của chúng. Sau đó lấy giá trị mfcc sau trừ đi trước để được delta1. Lấy giá trị delta1 sau trừ đi trước để có được delta2. Nối các đặc trưng trên lại, ta có:

12 mfcc + 12 delta1 + 12 delta2 (delta của delta) = 36 đặc trưng

Từ đây, ta nhận được một ma trận X có cỡ T*36 (với T là số frame) để đưa vào huấn luyện mô hình hmm.

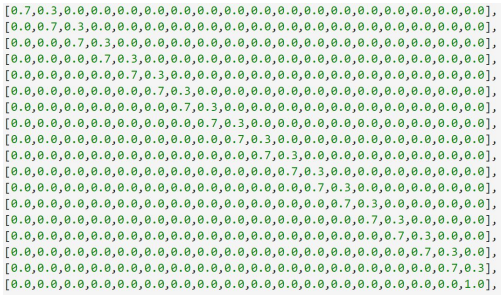
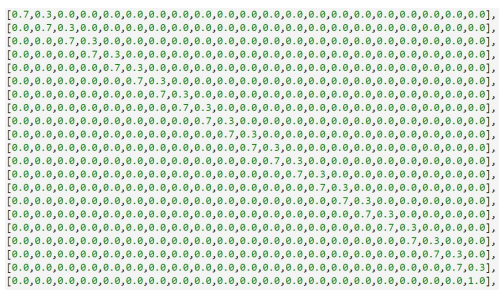
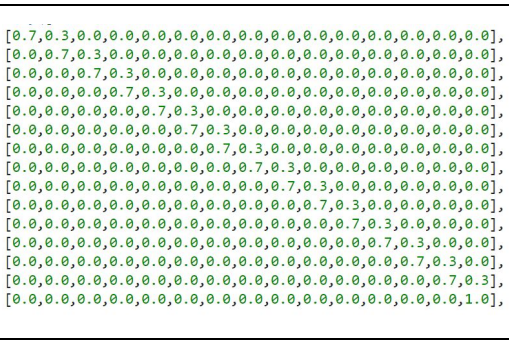
Xây dựng mô hình bằng GMMHMM

Để xây dựng mô hình gmm, nhóm sử dụng GMMHMM trong gói hmmlearn. Với mỗi từ, nhóm xây dựng một mô hình hmm từ trái sang phải với các parameter chung:

- n_mix = 4. Thể hiện 4 miền giọng nói khác nhau trong dữ liệu
- random_state = 42.

- n_iter = 1000. Số lần lặp tối đa
- verbose = true.
- params = ‘mctw’. Cho phép huấn luyện m: means; c: covars; t: transmat; w: GMM mixing weights.
- init_params = ‘mct’. Cho phép mô hình tự khởi tạo m, c và t.

Các parameter riêng:

Từ	Theo âm vị	n_ components	startprob_	transmat_
đăng nhập	|d||aˇ||ŋ| |ɲ||ɤˇ||p|	18	[1.0,0.0,0.0,0.0 ,0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0]	
đăng xuất	|d||aˇ||ŋ| |s||u||ɤˇ||t|	21	[1.0,0.0,0.0,0.0 ,0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0]	
khóa máy	|χ||ɔ||a| |m||a||i|	15	[1.0,0.0,0.0,0.0 ,0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0]	

tìm kiếm	|t||i||m| |k||ie||m|	18	[1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0]	[0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3], [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],

5. Giảm nhiễu và cắt âm thanh thừa

Nhiễu và âm thanh thừa có thể ảnh hưởng lớn tới độ chính xác của quá trình nhận diện. Để tránh những sai sót này, nhóm sử dụng giảm nhiễu và cắt âm thanh thừa khỏi các file âm thanh.

Để tìm ra các đoạn nhiễu, nhóm sử dụng bộ lọc median của thư viện scipy với cỡ kernel là 3. Sau đó cắt các đoạn âm thanh thừa ở đầu và cuối file để được file chỉ chứa đoạn âm thanh cần nhận diện.

6. Kết quả thử nghiệm

Kết quả thử nghiệm của mô hình nhận diện người nói trên bộ dữ liệu đã được chuẩn bị như sau:

Người nói	Correct Predict	Accuracy
Việt Anh	15/15	1.0
Tiến Đạt	15/15	1.0
Thanh Tùng	12/15	0.8
Huy Hoàng	11/15	0.73
Tuấn Anh	13/15	0.86
Vân Dung	15/15	1.0
Tự Long	15/15	1.0
Công Lý	15/15	1.0
MC Thảo Vân	15/15	1.0

Trường Giang	15/15	1.0

Kết quả thử nghiệm của mô hình nhận diện từ trên bộ dữ liệu được đã chuẩn bị như sau:

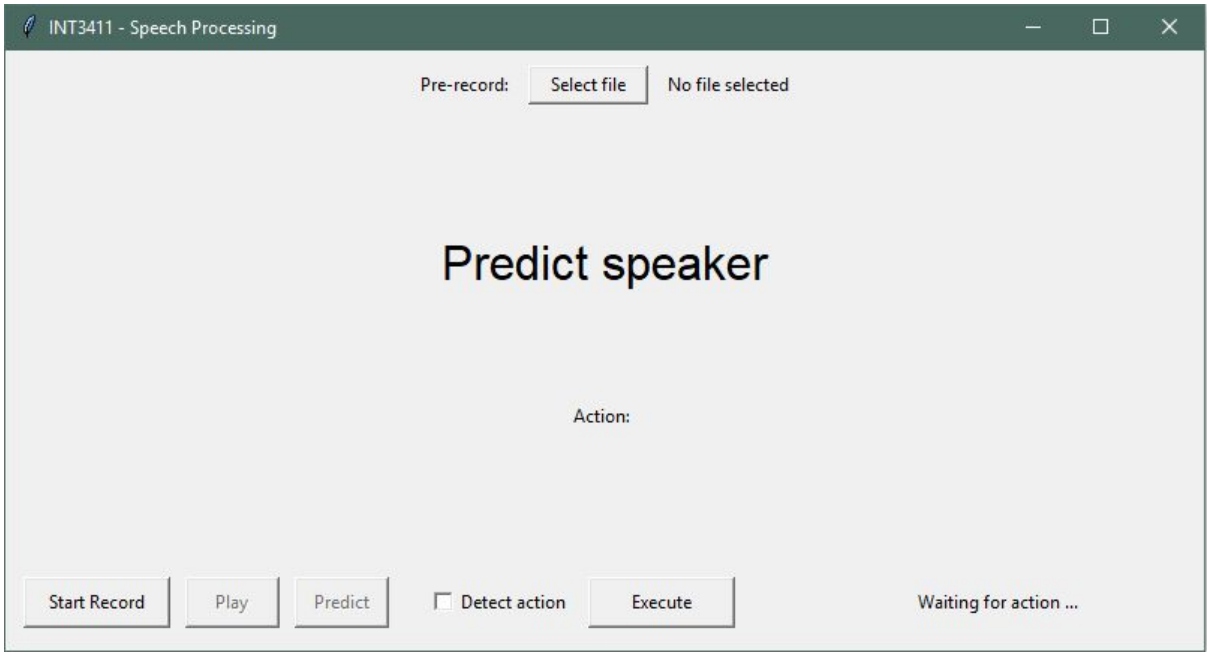
Từ	Kết quả thử nghiệm	
	Correct Predict	Accuracy
đăng nhập	10/10	1.0
đăng xuất	10/10	1.0
khóa máy	10/10	1.0
tìm kiếm	10/10	1.0

7. Sản phẩm

Với hai mô hình đã có, nhóm xây dựng một ứng dụng trên desktop với ba phần chính:

- Nhận diện người nói.
- Nhận diện từ.
- Thực hiện mệnh lệnh tương ứng với từ đã nhận diện được.

Giao diện của ứng dụng:



Để sử dụng, đầu tiên cần cung cấp cho ứng dụng một file ghi âm. Có thể chọn file ghi âm đã có trước hoặc ghi âm trực tiếp.

Với trường hợp ghi âm trực tiếp, ấn Start Record và bắt đầu nói. Khi muốn kết thúc ấn lại nút đó (đã chuyển thành Stop record).

Sau khi đã cung cấp file, ấn Play để nghe lại. Nếu muốn dừng, ấn lại nút đó (đã chuyển thành Stop).

Để dự đoán, ấn Predict. Nếu không chọn Detect action, ứng dụng sẽ chỉ dự đoán người nói trong file ghi âm được cung cấp. Nếu chọn Detect action, ứng dụng sẽ dự đoán cả một mệnh lệnh trong file được cung cấp.

Khi có mệnh lệnh được xác định, Ấn Execute để thực hiện. Danh sách mệnh lệnh gồm có:

- Đăng nhập: Đăng nhập vào trang web uetcodehub.xyz.
- Đăng xuất: Đăng xuất khỏi trang web uetcodehub.xyz.
- Khóa máy: Khóa desktop nếu người ra lệnh là chủ máy.
- Tìm kiếm: Tìm kiếm người nói trên Google.

Đối với người nổi tiếng: Chỉ cho phép tìm kiếm trên Google.

Có thể sử dụng bàn phím để điều khiển các chức năng:

- Record: sử dụng phím "r".
- Stop record: sử dụng phím "s"
- Play audio: sử dụng phím "p"
- Predict: sử dụng phím "f"

8. Đóng góp của các thành viên

Nhìn chung các thành viên trong nhóm đều có ý thức tham gia đóng góp cho bài tập.
Đóng góp cụ thể của từng thành viên như sau:

Công việc	Cắt dữ liệu	Ghi âm dữ liệu test	Làm app	Hiệu chỉnh mô hình	Khử nhiễu và cắt âm thừa	Viết báo cáo
Phạm Lê Việt Anh	✓	✓	✓	✓	✓	✓
Nguyễn Tiến Đạt	✓	✓	✓	✓		✓
Phạm Thanh Tùng	✓	✓	✓			
Nguyễn Huy Hoàng	✓	✓		✓		✓
Nguyễn Tuấn Anh		✓			✓	✓