Kiểm tra và đánh giá:

GV kiểm tra chương trình và báo cáo CHUNG của mỗi nhóm. Các SV trong nhóm tự phân công nhiệm vụ. Khi GV chấm bài, mỗi SV trình bày riêng phần code và slide phần việc mà SV đó làm.

Chú ý:

- Nhóm SV chỉ báo cáo cách làm step-by-step và kết quả (bằng hình ảnh và số liệu cụ thể), KHÔNG báo cáo lý thuyết.
- Nhóm SV nộp chung mã nguồn và slide báo cáo (file PDF). Bảng dữ liệu (hoặc đồ thị 2D/3D) chứa kết quả đo được tổng hợp để đưa ra nhận xét chung.
- Timeslot cho mỗi nhóm để trình bày slide (trên laptop GV): 4 phút cho các nhóm chỉ làm BT1, 8 phút cho các nhóm làm BT1 và BT2, 12 phút cho các nhóm làm cả 3 BT.
- Demo trên laptop của nhóm SV: max. 2 phút.

(Xem nội dung 03 BT ở các trang sau)

Bài 1: Phân tích đặc trưng phổ các nguyên âm của nhiều người nói (trọng số điểm: 40% cột điểm BT)

Input: tín hiệu tiếng nói

Output:

- ảnh phổ (spectrogram) và bộ 3 tần số formant {F1,F2,F3} của mỗi nguyên âm được đo (hoặc đánh dấu) trực tiếp trên ảnh phổ
- nhận xét về sự khác biệt đặc trưng phổ giữa các nguyên âm khác nhau của cùng một người nói.

SV chỉ cần kết luận bộ {F1,F2,F3} có khác nhau đáng kể (hoặc ko đáng kể) giữa các nguyên âm của cùng 1 người nói dựa trên bảng số liệu (hoặc chuyển bảng thành đồ thị 2D/3D thì trực quan hơn). Các giá trị tần số đo thủ công chỉ có tính chất ước lượng nên ko thể chính xác, do đó ko nên để số lẻ nào ở phần thập phân. SV nhận xét lan man như trong luận văn cao học sẽ bị trừ điểm.

Yêu cầu:

- Xuất ảnh phổ băng rộng (wideband spectrogram) của các file DL huấn luyện (của ~~08~~ 04 người nói chọn ngẫu nhiên từ 21 người trong thư mục **NguyenAmHuanLuyen-16k**).
- Xuất bảng dữ liệu (hoặc đánh dấu trên ảnh phổ) bộ 3 tần số formant của mỗi nguyên âm và đưa ra nhận xét về sự khác biệt giữa đặc trưng phổ giữa 05 nguyên âm của một người nói trong ~~08~~ 04 người đã chọn.
- ~~So sánh bộ 3 tần số formant của một nguyên âm của các người nói khác nhau và đưa ra nhận xét về sự khác biệt giữa đặc trưng phổ của 08 người nói đã chọn (bỏ, ko cần làm)~~.
- Dữ liệu 3 tần số formant chỉ cần đo thủ công (cách đo xem phần 3.7.2 trong [3]) hoặc dùng đoạn thẳng nằm ngang để đánh dấu trên ảnh phổ.

TLTK:

[1] Hands-on lab on Speech Processing-Frequency-domain processing_2021 (phần 1 về spectrogram,n guồn: https://www.csd.uoc.gr/~hy578/2021/Project0_Part2.pdf)).
[2] Spectrogram-Cepstrum-and Mel-Frequency Analysis_CMU (slide 11-18 về ý tưởng phân biệt các nguyên âm với nhau bằng bộ 3 tần số formant và cách đo tần số formant trên spectrogram).
[3] Phân tích formant các nguyên âm của nhiều người nói_Luận văn_PĐThiện_2021 (phần 2.2.1 về khái niệm formant, phần 3.5 và 3.6.2 về một số kết quả đo để tham khảo về dải giá trị của F1,F2 và F3 của các nguyên âm, chú ý tránh dùng các bình luận "dài dòng" như trong luận văn).

Bài 2: Nhận dạng nguyên âm không phụ thuộc người nói dùng đặc trưng phổ FFT
(trọng số điểm: 40% cột điểm BT)

Input:
tín hiệu tiếng nói (chứa 01 nguyên âm và khoảng lặng) của tập kiểm thử (thư mục **NguyenAmKiemThu-16k** gồm 21 người, 105 file test).

Output:

- Kết quả nhận dạng (dự đoán) nhãn nguyên âm của mỗi file test (/a/, …,/u/), Đúng/Sai (dựa vào nhãn tên file).
- Xuất 05 vector đặc trưng biểu diễn 05 nguyên âm trên cùng 01 đồ thị.
- Bảng thống kê độ chính xác nhận dạng tổng hợp (%) theo số chiều của vector đặc trưng (là số điểm lấy mẫu trên miền tần số N_FFT với 03 giá trị 512, 1024, 2048).
- Ma trận nhầm lẫn (confusion matrix) của trường hợp có độ chính xác tổng hợp cao nhất trong 3 giá trị N_FFT trên: ma trận này thống kê số lần nhận dạng đúng/sai của mỗi cặp nguyên âm (có highlight nguyên âm đc nhận dạng đúng và bị nhận dạng sai nhiều nhất) theo format sau:

		Nhãn dự đoán				
		/a/	/e/	/i/	/o/	/u/
Nhãn đúng	/a/	Số lần /a/ nhận dạng đúng là /a/	Số lần /a/ nhận dạng sai thành /e/	Số lần /a/ nhận dạng sai thành /i/	Số lần /a/ nhận dạng sai thành /o/	Số lần /a/ nhận dạng sai thành /u/
	/e/	Số lần /e/ nhận dạng sai thành /a/	Số lần /e/ nhận dạng đúng là /e/	…	…	…
	/i/	Số lần /i/ nhận dạng sai thành /a/	…	Số lần /i/ nhận dạng đúng là /i/	…	…
	/o/	…	…	…	Số lần /o/ nhận dạng đúng là /o/	…
	/u/	…	…	…	…	Số lần /u/ nhận dạng đúng là /u/

Dữ liệu sử dụng để thiết lập và hiệu chỉnh các thông số của thuật toán:
tín hiệu tiếng nói (mỗi tín hiệu chứa 01 nguyên âm ở giữa và 2 khoảng lặng ở 2 đầu) của tập huấn luyện (thư mục **NguyenAmHuanLuyen-16k** gồm 21 người, 105 file huấn luyện).

Yêu cầu:
Cài đặt BT nhận dạng theo mô hình tương tự như BT tìm kiếm âm thanh (trong TLTK [4]) gồm 3 thuật toán sau:

1. Phân đoạn tín hiệu thành nguyên âm và khoảng lặng (slide Chapter6_SPEECH SIGNAL PROCESSING).
2. Trích xuất vector đặc trưng phổ của 05 nguyên âm dựa trên tập huấn luyện (gồm 21 người, 105 file huấn luyện):

a. Đánh dấu vùng có đặc trưng phổ ổn định đặc trưng cho nguyên âm: chia vùng chứa nguyên âm tìm được ở bước 1 thành 3 đoạn có độ dài bằng nhau và lấy đoạn nằm giữa (giả sử gồm M khung).
b. Trích xuất vector FFT của 1 khung tín hiệu với số chiều là N_FFT (=512, 1024, 2048) dùng các hàm thư viện.
c. Tính vector đặc trưng cho 1 nguyên âm của 1 người nói = Trung bình cộng của M vector FFT của M khung thuộc vùng ổn định.
d. Tính vector đặc trưng cho 1 nguyên âm của nhiều người nói = Trung bình cộng của các vector đặc trưng cho 1 nguyên âm của 21 người nói (trong tập huấn luyện).

3. So khớp vector FFT của tín hiệu nguyên âm đầu vào (thuộc tập kiểm thử) với 5 vector đặc trưng đã trích xuất của 5 nguyên âm (dựa trên tập huấn luyện) để đưa ra kết quả nhận dạng nguyên âm bằng cách tính 5 khoảng cách Euclidean giữa 2 vector và đưa ra quyết định nhận dạng dựa trên k/c nhỏ nhất (hàm này SV tự cài đặt).

TLTK:

[1] Hands-on lab on Speech Processing-Frequency-domain processing_2021 (phần 1 về spectrogram, nguồn: https://www.csd.uoc.gr/~hy578/2021/Project0_Part2.pdf).
[4]CITA_SoSanhPhuongPhapDuongBaoPhovaPhuongPhapAnhPhoTrongTimKiemAmNhac_2021 (phần 2 về mô hình của BT tìm kiếm/nhận dạng dựa trên so khớp mẫu).

Bài 3: Nhận dạng nguyên âm không phụ thuộc người nói dùng đặc trưng phổ MFCC
(trọng số điểm: 20% cột điểm BT)

BT3 là phần mở rộng của BT2 nhằm cải thiện độ chính xác nhận dạng với mô tả các phần Input, Output và Yêu cầu tương tự như BT2. Điểm khác biệt ở chỗ vector đặc trưng phổ FFT (phản ánh nội dung chi tiết của phổ) được thay bằng vector đặc trưng phổ MFCC (phản ánh đường bao phổ) và thuật toán phân cụm K-mean được sử dụng để gom các người nói có chất giọng giống nhau vào từng cụm. Do đó các mô tả được cập nhật như sau:

Input:
Tín hiệu tiếng nói (chứa 01 nguyên âm và 2 khoảng lặng) trong tập kiểm thử (gồm 21 người, 105 file test).

Output:
- Kết quả nhận dạng (dự đoán) nhãn nguyên âm của mỗi file test (/a/, …,/u/), Đúng/Sai (dựa vào nhãn tên file).
- Xuất 05 vector đặc trưng MFCC biểu diễn 05 nguyên âm trên cùng 01 đồ thị.
- Kết quả độ chính xác nhận dạng tổng hợp (%) theo số chiều của vector đặc trưng N_MFCC (N_MFCC cố định là 13) và K (là số cụm với 04 giá trị K=2,3,4,5).
- Ma trận nhầm lẫn (confusion matrix) của trường hợp có độ chính xác tổng hợp cao nhất trong 4 giá trị K trên: ma trận này thống kê số lần nhận dạng đúng/sai của mỗi cặp nguyên âm (có highlight nguyên âm đc nhận dạng đúng và bị nhận dạng sai nhiều nhất).

Dữ liệu sử dụng để thiết lập và hiệu chỉnh các thông số của thuật toán:
tín hiệu tiếng nói (mỗi tín hiệu chứa 01 nguyên âm ở giữa và 2 khoảng lặng ở 2 đầu) của tập huấn luyện (thư mục **NguyenAmHuanLuyen-16k** gồm 21 người, 105 file huấn luyện).

Yêu cầu:
Cài đặt BT nhận dạng theo mô hình tương tự như BT tìm kiếm âm thanh (trong TLTK [4]) gồm 3 thuật toán sau:

1. Phân đoạn tín hiệu thành nguyên âm và khoảng lặng (slide Chapter6_SPEECH SIGNAL PROCESSING).
2. Trích xuất vector đặc trưng phổ của 05 nguyên âm dựa trên tập huấn luyện (gồm 21 người, 105 file huấn luyện):
    a. Đánh dấu vùng có đặc trưng phổ ổn định đặc trưng cho nguyên âm: chia vùng chứa nguyên âm tìm được ở bước 1 thành 3 đoạn có độ dài bằng nhau và lấy đoạn nằm giữa (giả sử gồm M khung).
    b. Trích xuất vector MFCC (mel-frequency cepstral coefficients) của 1 khung tín hiệu với số chiều (chính là số lượng hệ số MFCC) là N_MFCC =13, dùng các hàm của thư viện Voicebox (Matlab) hoặc libbrosa (python).
    c. Tính vector đặc trưng cho 1 nguyên âm của 1 người nói = Trung bình cộng của M vector MFCC của M khung thuộc vùng ổn định.
    d. Tính vector đặc trưng cho 1 nguyên âm của nhiều người nói = Trung bình cộng của các vector đặc trưng cho 1 nguyên âm của 21 người nói (trong tập huấn luyện).

3. So khớp vector MFCC của tín hiệu nguyên âm đầu vào (thuộc tập kiểm thử) với 5 vector đặc trưng đã trích xuất của 5 nguyên âm (dựa trên tập huấn luyện) để đưa ra kết quả nhận dạng nguyên âm bằng cách tính 5 khoảng cách Euclidean giữa 2 vector và đưa ra quyết định nhận dạng dựa trên k/c nhỏ nhất (hàm này SV tự cài đặt).

Phần nâng cao (dành cho các nhóm SV muốn làm thêm để cải thiện độ chính xác nhận dạng hơn nữa và nhận điểm tối đa):

- Mục 2c và 2d: Nếu chỉ tính 1 vector đặc trưng cho 1 nguyên âm của nhiều người nói thì độ chính xác biểu diễn không cao do các người nói có chất giọng ít/nhiều khác nhau → làm giảm độ chính xác nhận dạng. Do đó, có thể tăng độ chính xác biểu diễn bằng cách tính K vector đặc trưng cho 1 nguyên âm của nhiều người nói dùng thuật toán phân cụm K-trung bình (K-mean clustering) với K=2,3,4,5. Chạy K-mean clustering trên tất cả các vector MFCC của các khung nằm trong phần ổn định của 1 nguyên âm của 21 người trong tập huấn luyện để thu được K vector trung bình làm K vector đặc trưng cho 1 nguyên âm.
- Mục 3: So khớp vector MFCC của tín hiệu nguyên âm đầu vào (thuộc tập kiểm thử) với 5xK vector đặc trưng đã trích xuất của 5 nguyên âm (dựa trên tập huấn luyện) để đưa ra kết quả nhận dạng nguyên âm: tính 5xK khoảng cách Euclidean giữa 2 vector và đưa ra quyết định nhận dạng dựa trên k/c nhỏ nhất (SV tự cài đặt).
- Thuật toán phân cụm K-trung bình: SV có thể tự cài đặt hoặc dùng hàm thư viện có sẵn.
- Lập bảng báo cáo kết quả độ chính xác nhận dạng tổng hợp (%) theo số cụm K.

TLTK:

[2] Spectrogram-Cepstrum-and Mel-Frequency Analysis_CMU (slide 16-49 về đường bao phổ và các hệ số MFCC).

[4]CITA_SoSanhPhuongPhapDuongBaoPhovaPhuongPhapAnhPhoTrongTimKiemAmNhac_2021 (phần 2 và 3.1 về mô hình tìm kiếm/nhận dạng và thuật toán phân cụm K-trung bình).