

Trường Đại Học Bách Khoa - Đại Học Đà Nẵng

BÁO CÁO BÀI TẬP NHÓM XỬ LÝ TÍN HIỆU SỐ

GVHD: TS. Ninh Khánh Duy

Nhóm 10 - 21Nh10

- Nguyễn An Hưng
- Nguyễn Cửu Nhật Quang
- Nguyễn Thúc Hoàng
- Nguyễn Văn Trường Sơn

BẢNG PHÂN CÔNG CÔNG VIỆC

Thành viên	Nhiệm vụ
Nguyễn An Hưng (Trưởng Nhóm)	Phân chia công việc cho từng thành viên, quản lý tiến độ công việc, set-up source code chung cho toàn bộ dự án, làm slide. Làm bài 3 phần cải thiện độ chính xác
Nguyễn Thúc Hoàng	Làm slide và làm bài 1
Nguyễn Cửu Nhật Quang	Làm slide và làm bài 2
Nguyễn Văn Trường Sơn	Làm slide và làm bài 3 phần MFCC

NỘI DUNG

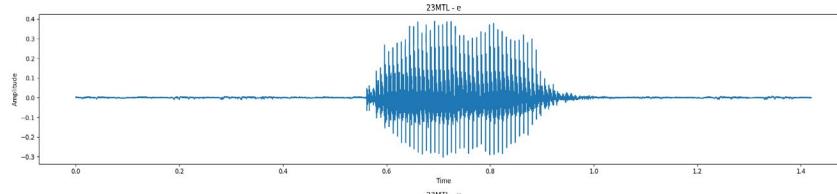
- **PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHỔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI**
 - Xuất ảnh phổ băng rộng (wide - spectrogram)
 - Xuất bảng dữ liệu bộ 3 tần số Formant (thủ công & thuật toán ước tính tự động)
- **PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT**
 - Thuật toán và các bước tiến hành
 - Kết quả thực nghiệm
 - Nhận xét và kết luận
- **PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC**
 - Xây dựng phương pháp trích xuất đặc trưng sử dụng đặc trưng phổ MFCC
 - Mô hình nhận dạng sử dụng đặc trưng phổ MFCC
 - Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means (clustering)
 - Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp KNN (K-Nearest Neighbors)
 - Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp CatBoost
- **PHẦN 4: SO SÁNH VÀ TỔNG KẾT**

PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHÔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI

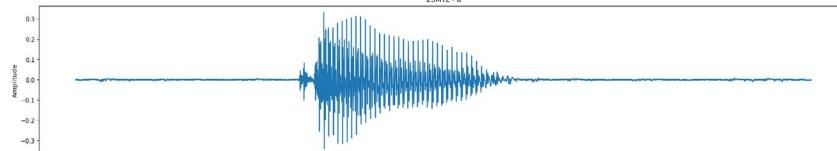
Xuất ảnh phô bắng rộng (wide - spectrogram)

Người nói: 23MTL

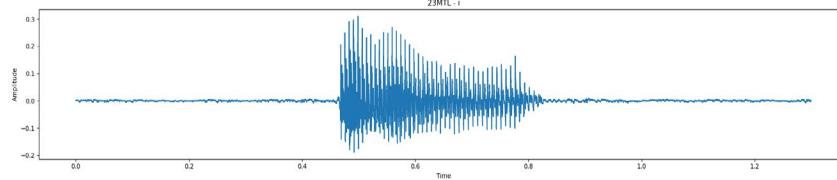
/e/



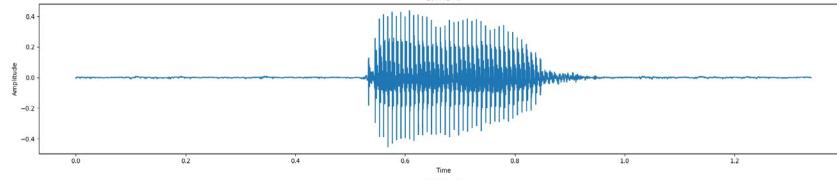
/u/



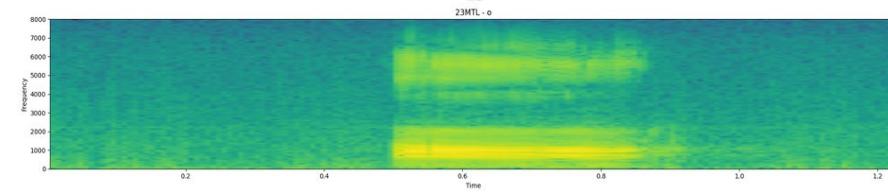
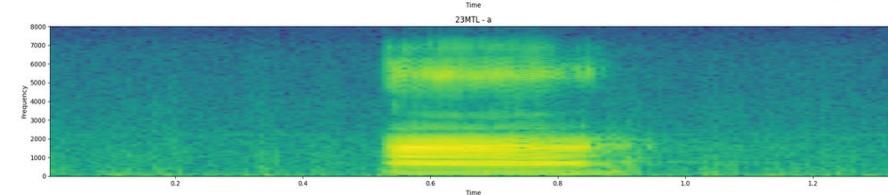
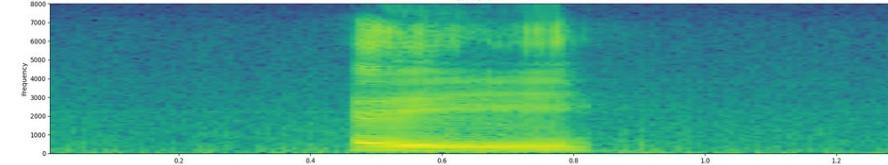
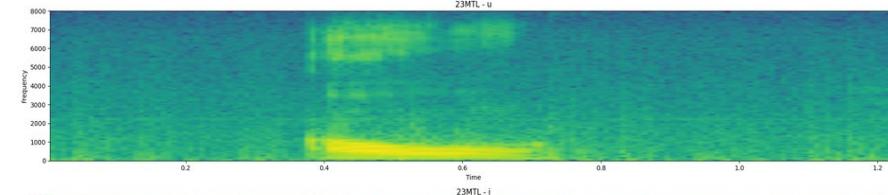
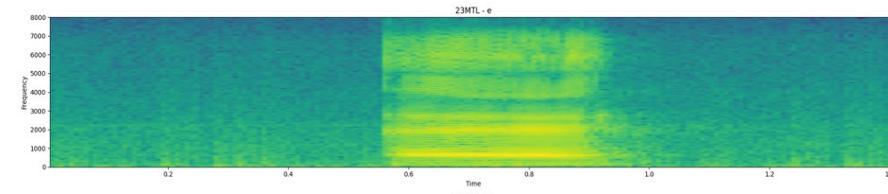
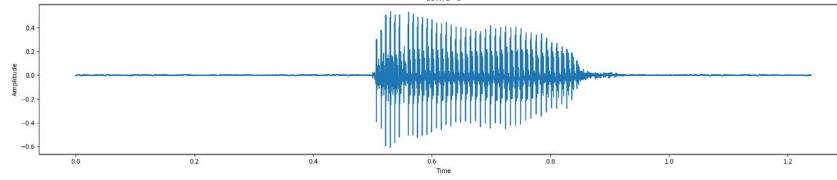
/i/



/a/



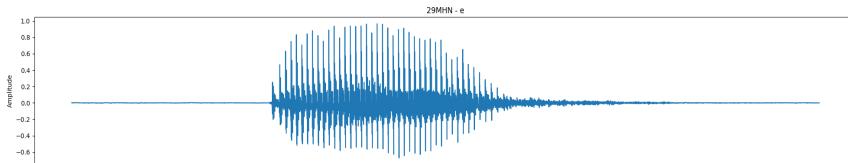
/o/



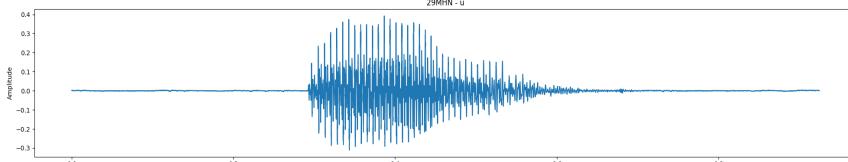
PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHÔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI
Xuất ảnh phô bắng rộng (wide - spectrogram)

Người nói: 29MHN

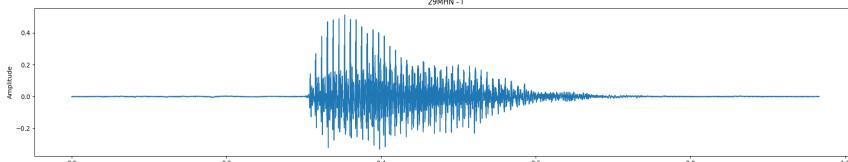
/e/



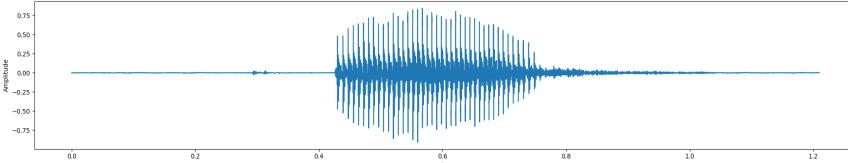
/u/



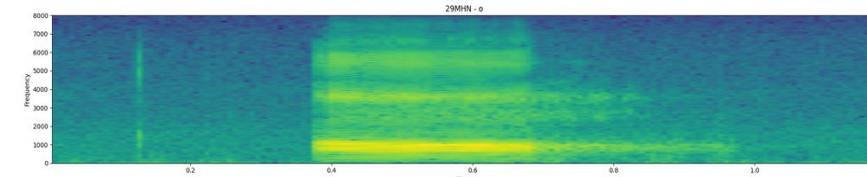
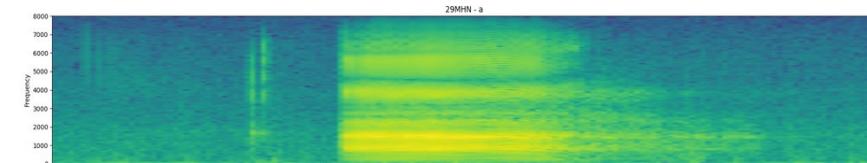
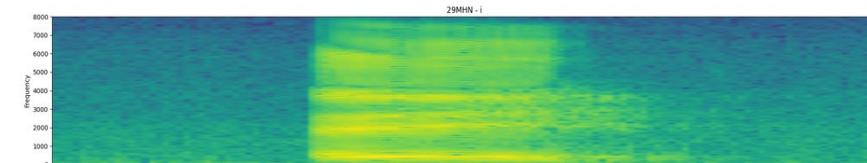
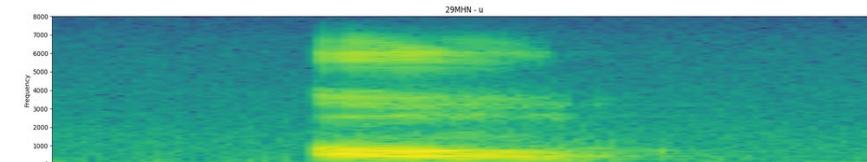
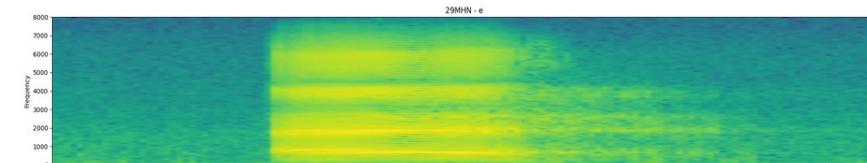
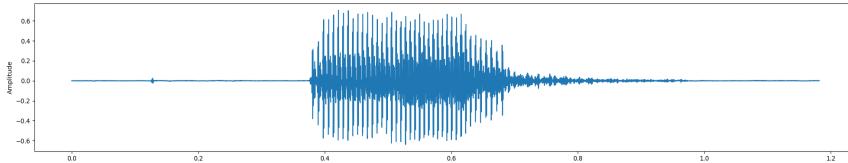
/i/



/a/

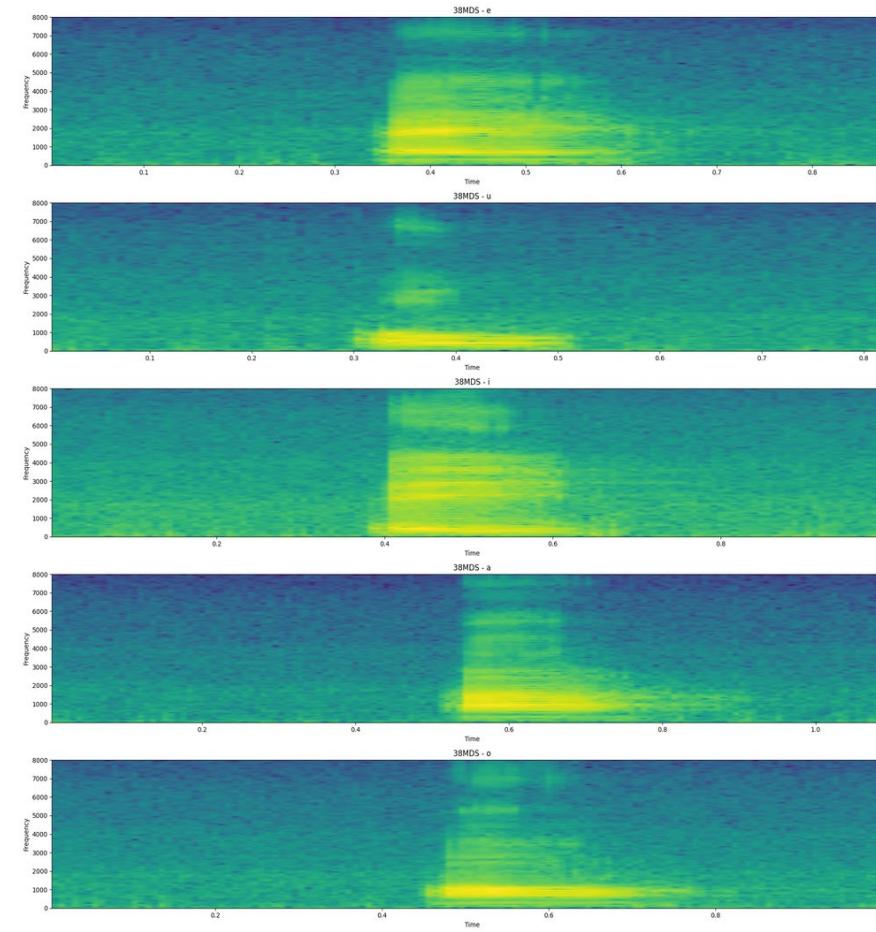
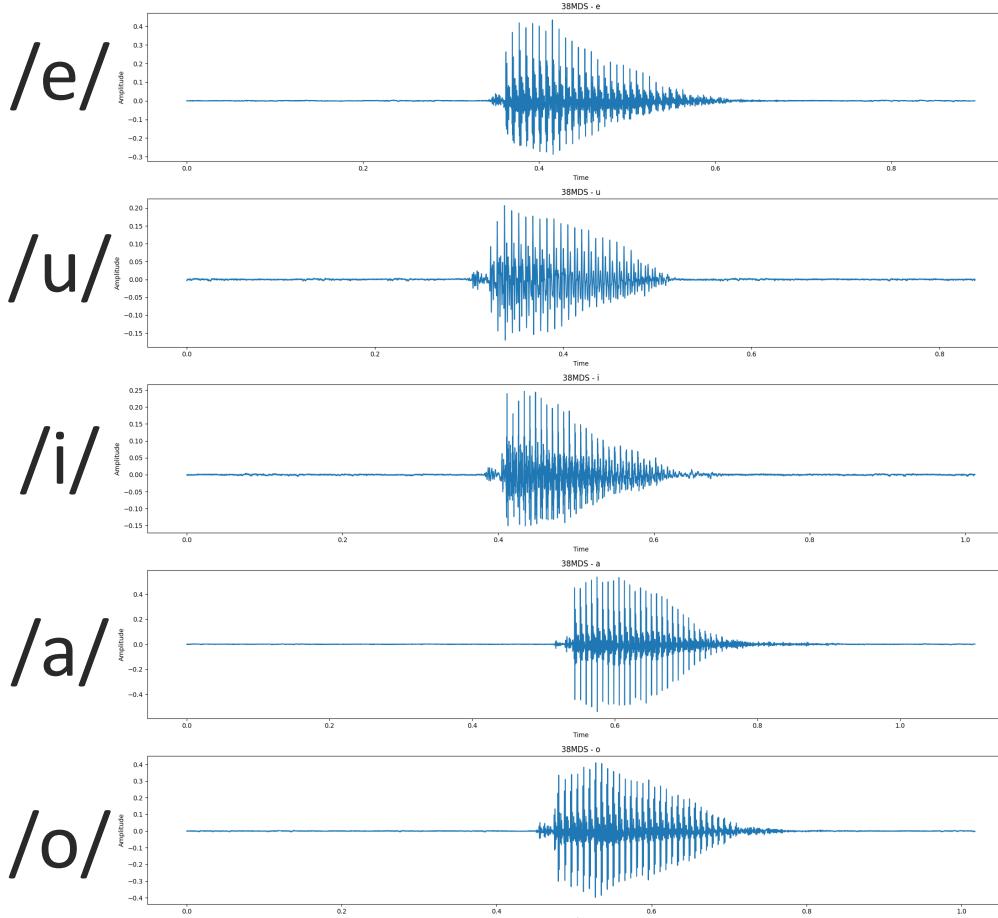


/o/



PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHÔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI
Xuất ảnh phô bắng rộng (wide - spectrogram)

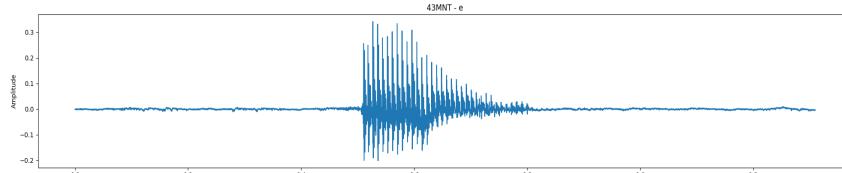
Người nói: 38MDS



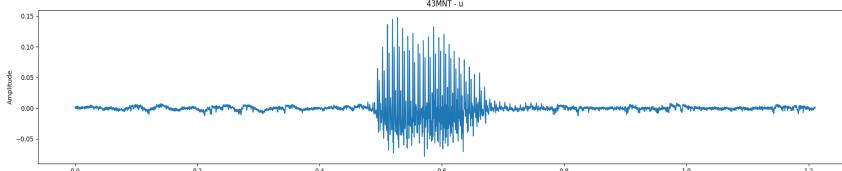
PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHÔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI
Xuất ảnh phô bắng rộng (wide - spectrogram)

Người nói: 43MNT

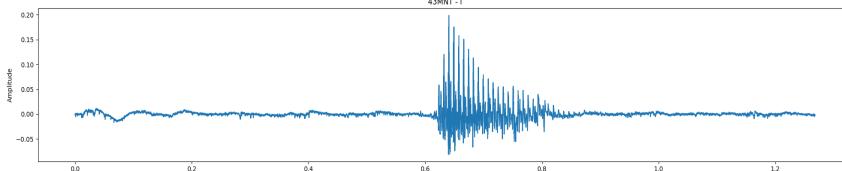
/e/



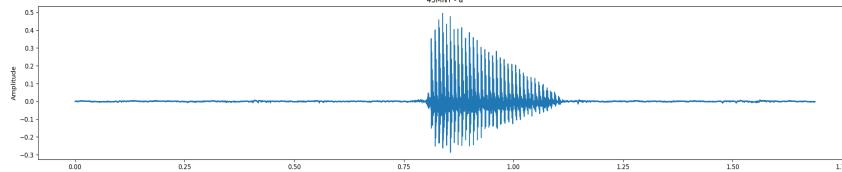
/u/



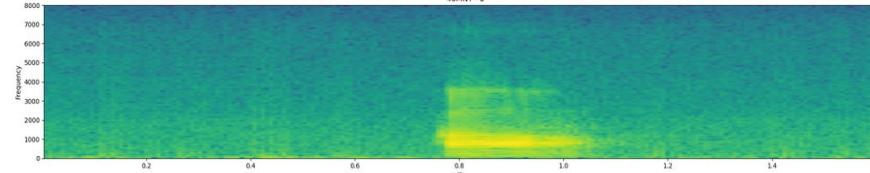
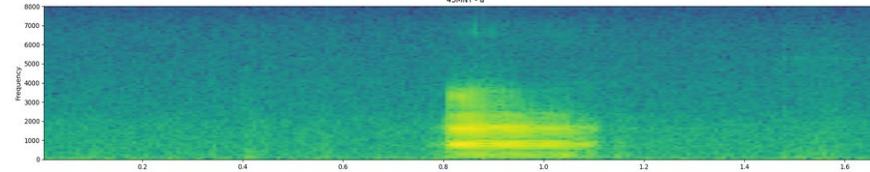
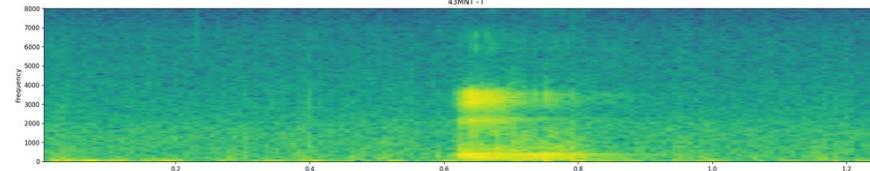
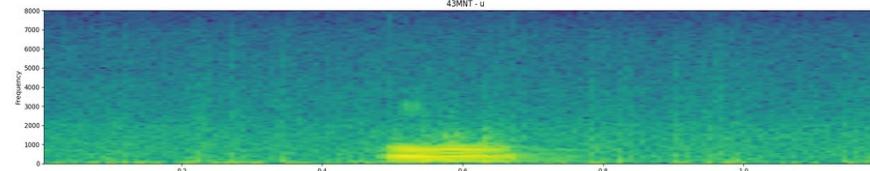
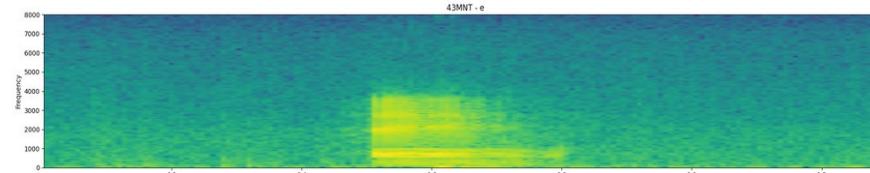
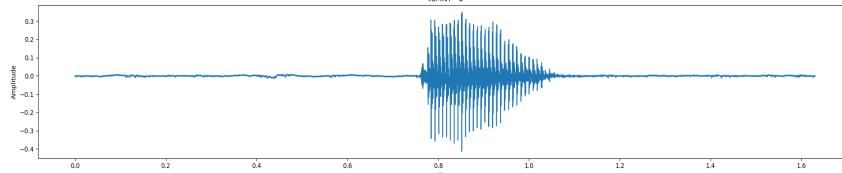
/i/



/a/



/o/

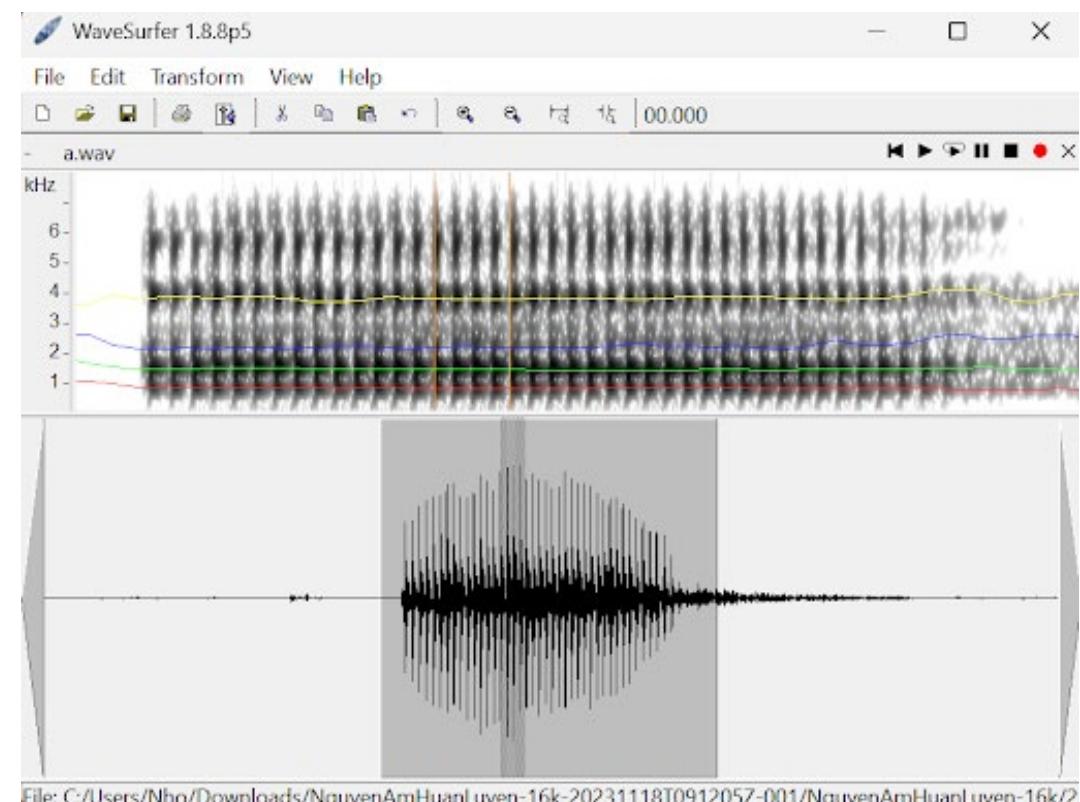


PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHỔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI

Xuất bảng dữ liệu bộ 3 tần số Formant (thủ công & thuật toán ước tính tự động)

Dữ liệu bộ 3 tần số formant 5 nguyên âm của 1 người nói

```
Statistics for data in pane 1
Column 0 mean: 854.678772 sd: 3.758072
Column 1 mean: 1484.319661 sd: 2.846339
Column 2 mean: 2210.901204 sd: 23.627358
Column 3 mean: 3806.316895 sd: 16.242733
-----
Statistics computed between 0.5432784576534576 and 0.573356157731
1577
```



Minh họa khảo sát tín hiệu người nói: 29MHZ
(Đo thủ công bằng phần mềm WaveSurfer)

PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHỔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI

Xuất bảng dữ liệu bộ 3 tần số Formant (thủ công & thuật toán ước tính tự động)

Đo thủ công bằng WaveSurfer

		/a/	/e/	/i/	/o/	/u/
29MHN	F1mean(Hz)	856	727	433	740	488
	F2mean(Hz)	1478	1772	1941	979	770
	F3mean(Hz)	2202	2438	2610	2384	2559

Sử dụng LPC

		/a/	/e/	/i/	/o/	/u/
29MHN	F1mean(Hz)	834	663	408	728	481
	F2mean(Hz)	1464	1828	2047	960	633
	F3mean(Hz)	2366	2548	2799	2745	2840

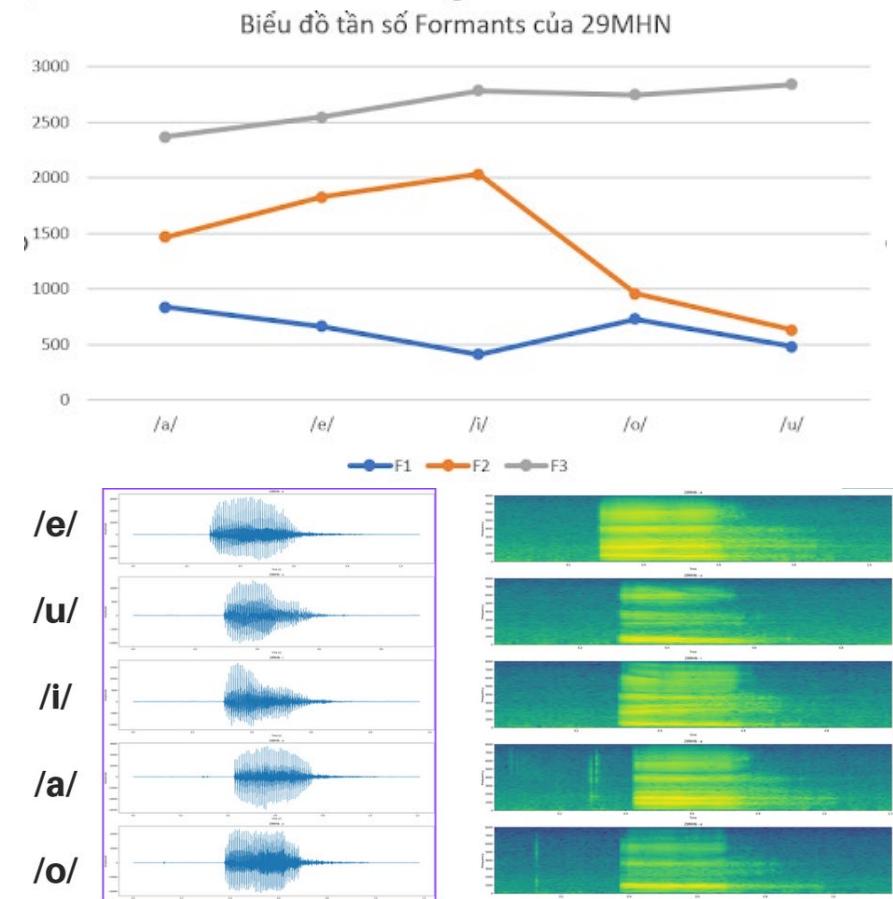
Kết quả khảo sát tín hiệu người nói: 29MHN

PHẦN 1: PHÂN TÍCH ĐẶC TRƯNG PHỔ CÁC NGUYÊN ÂM NHIỀU NGƯỜI NÓI

Xuất bảng dữ liệu bộ 3 tần số Formant (thủ công & thuật toán ước tính tự động)

NHẬN XÉT:

- Tần số F1 và F3 của 5 nguyên âm có sự khác biệt không đáng kể ($\Delta \sim 500\text{Hz}$).
- Tần số F2 của 5 nguyên âm có sự khác biệt đáng kể ($\sim 1400\text{Hz}$).
- Nguyên âm /i/ có tần số F1 thấp nhất
- Nguyên âm /o/ và /u/ có sự khác biệt đáng kể giữa F2 và F3

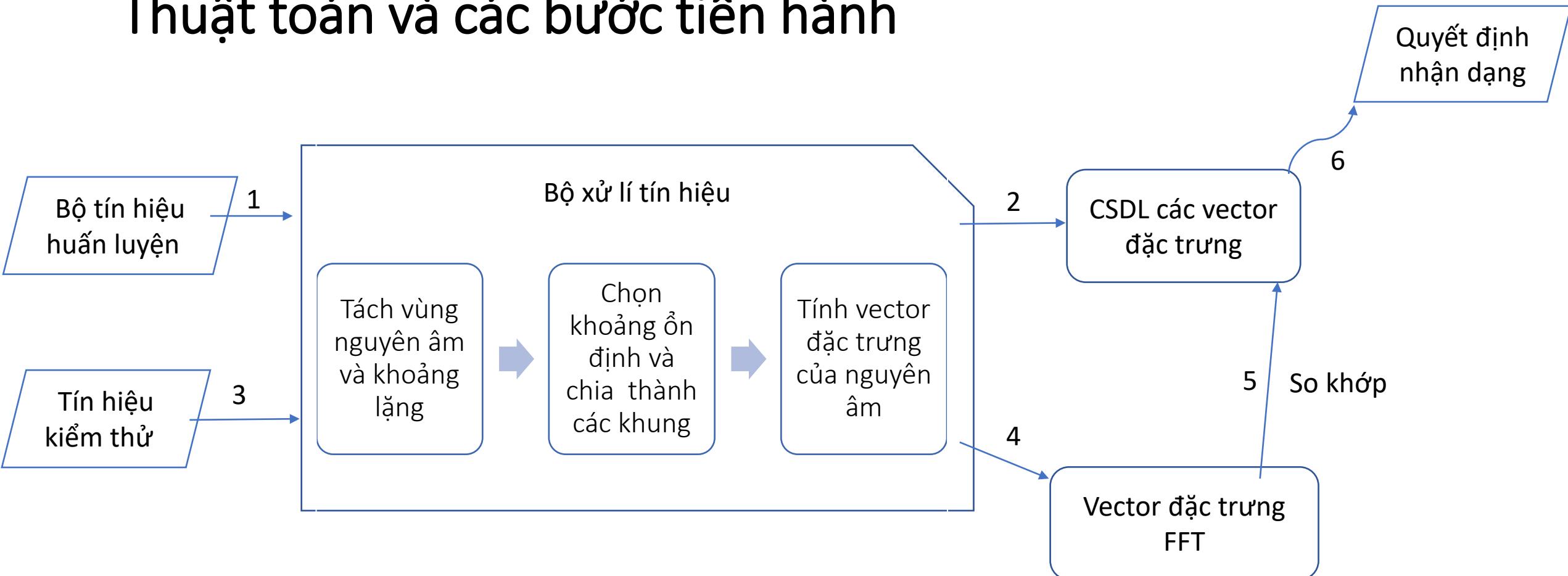


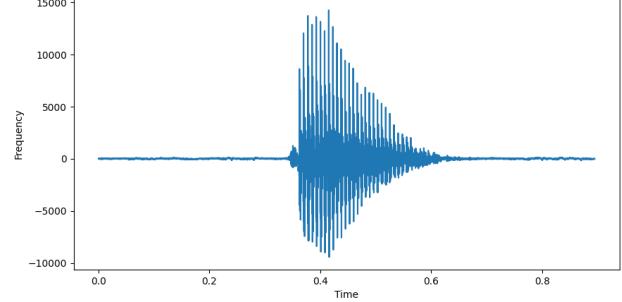
PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

- Thuật toán và các bước tiến hành
- Kết quả thực nghiệm
- Nhận xét và kết luận

PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Thuật toán và các bước tiến hành

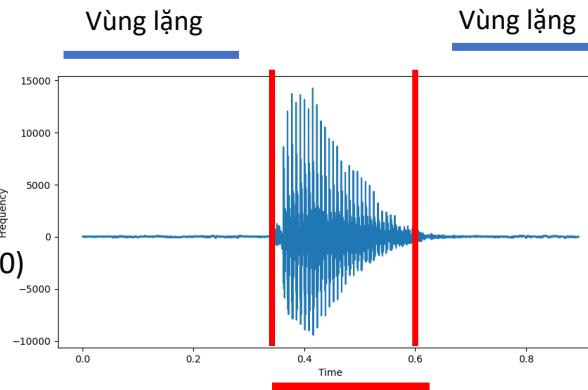




INPUT

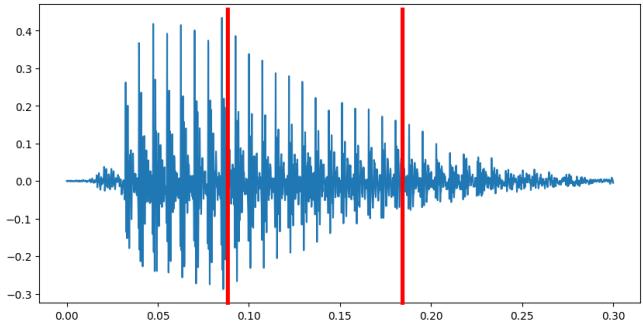
$$\varepsilon_i = STE(X_i)$$

(thr:0.01, w_len:0.03, ovrlap: 0.0)



Vùng âm thanh

$$\varepsilon_i > STE_{threshold}$$

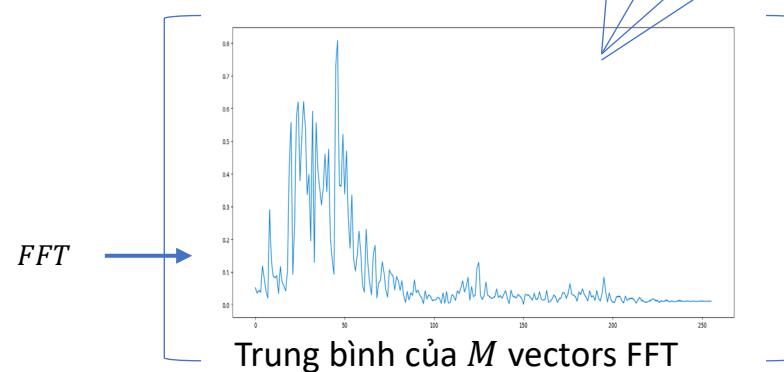


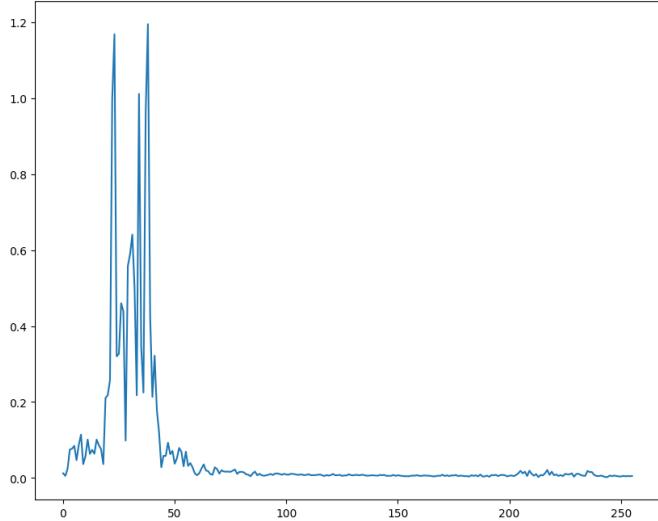
Vùng âm thanh

Chọn phần ổn định

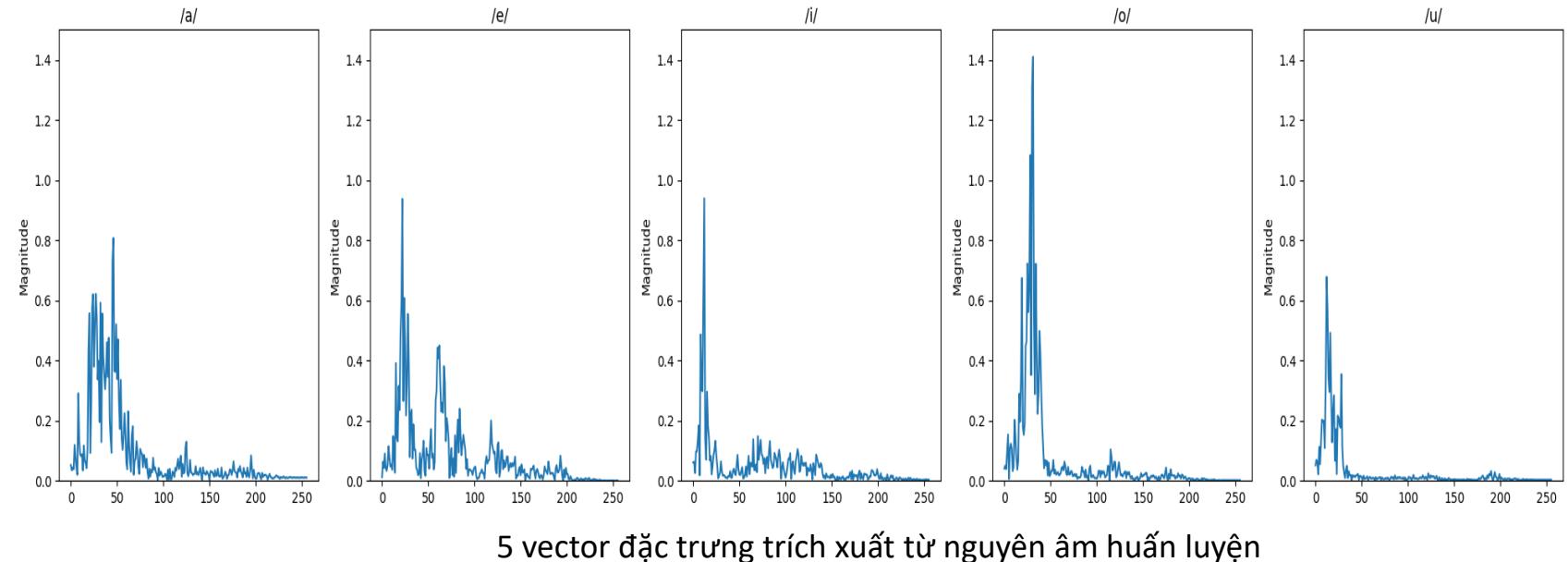


Chia thành M frames





Vector đặc trưng của nguyên âm kiểm thử



5 vector đặc trưng trích xuất từ nguyên âm huấn luyện

Tính khoảng cách Euclidean

- $d_k = \sum_{i=1}^{NFFT} |v_i - Vt_i|$

Dự đoán nhãn với khoảng cách nhỏ nhất

- $\min(d) \Rightarrow label_{predict}$

Kiểm tra kết quả

- So sánh $label_{input}$ và $label_{predict}$

PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Kết quả thực nghiệm

Mức độ ảnh hưởng của tham số:

1. Phân biệt vùng nguyên âm và vùng khoảng lặng bằng Short-Time Energy
 - Kích thước khung quá lớn làm giảm độ chính xác khi tách vùng nguyên âm.
 - Threshold phụ thuộc mật thiết vào kích thước khung.
 - Threshold quá nhỏ hoặc quá lớn làm giảm độ chính xác phân biệt
2. Tính vector đặc trưng bằng FFT
 - Kích thước khung quá lớn ($>80\text{ms}$) xảy ra hiện tượng nội suy vector đặc trưng, giảm độ chính xác
 - Kích thước khung quá nhỏ ($<10\text{ms}$) không lấy tối ưu được đặc trưng của nguyên âm
 - Độ dịch khung: tối ưu khi mức độ overlap xấp xỉ 30-50% kích thước khung
 - Hàm cửa sổ: các hàm cửa sổ có làm thay đổi kết quả nhận dạng nhưng không quá lớn (trong ngưỡng 1-3%)
 - Số điểm lấy mẫu: số điểm lấy mẫu càng nhỏ thì độ chính xác càng giảm

PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

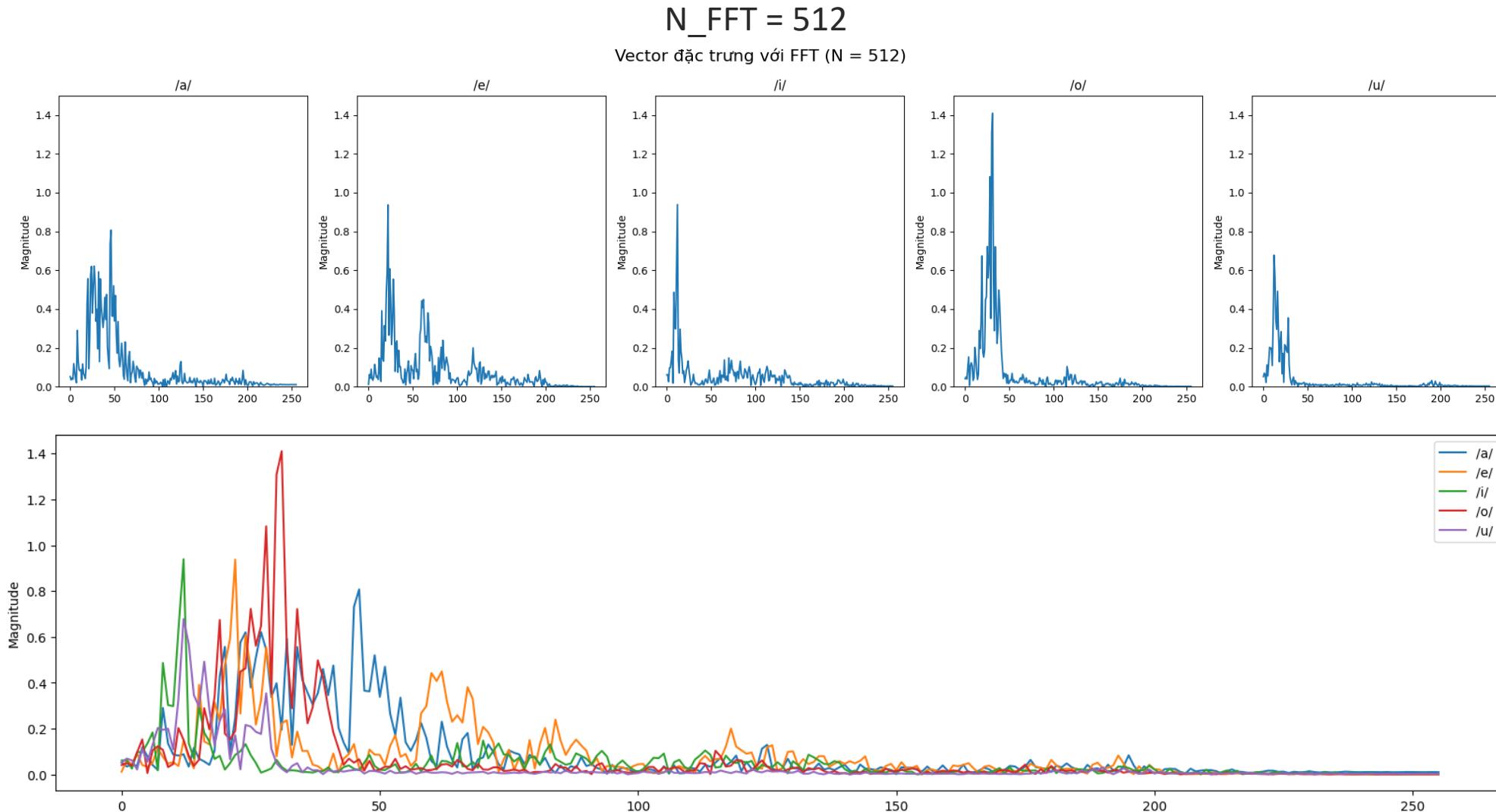
Kết quả thực nghiệm

Bộ tham số sử dụng:

1. Phân biệt vùng nguyên âm và vùng khoảng lặng bằng Short-Time Energy
 - Kích thước khung: 30ms
 - Độ dịch khung: 30ms
 - STE threshold: 0.01
2. Tính vector đặc trưng bằng FFT
 - Khoảng ổn định: 1/3 chính giữa vùng nguyên âm
 - Kích thước khung: 60ms
 - Số khung: $M = 3$
 - Hàm cửa sổ: Triangular window (Bartlett window)

PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Kết quả thực nghiệm

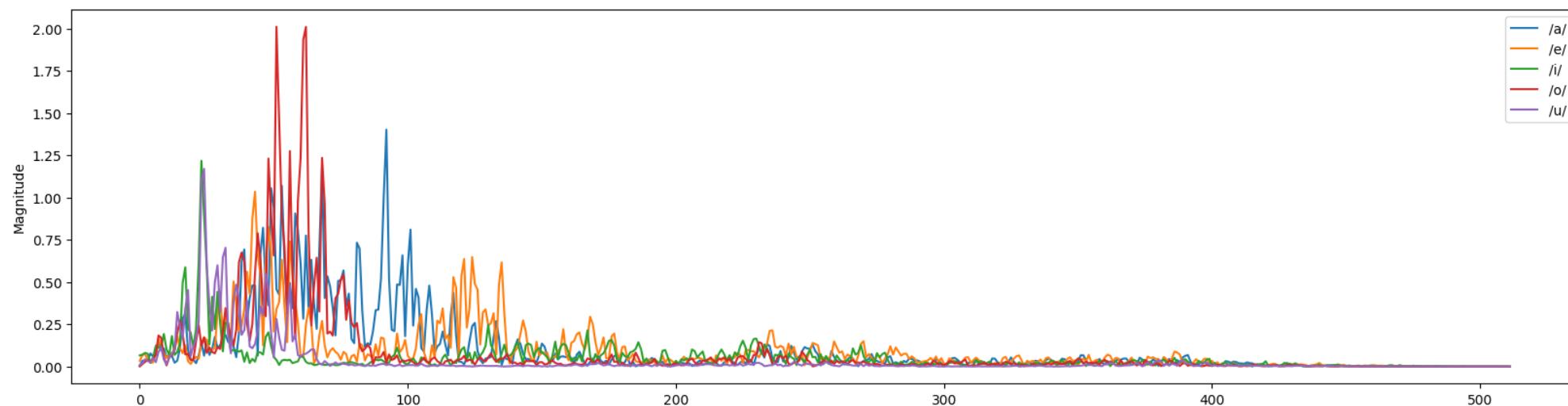
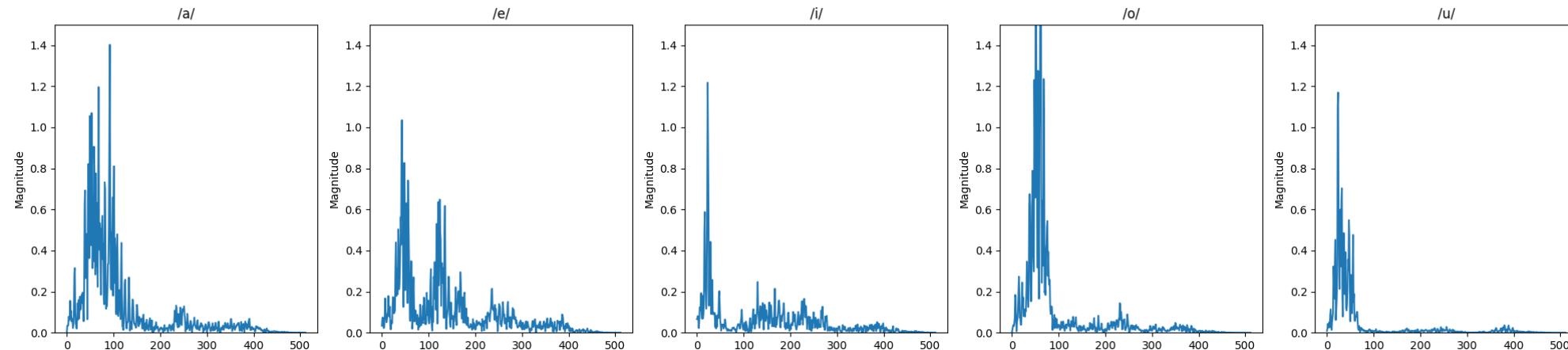


PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Kết quả thực nghiệm

N_FFT = 1024

Vector đặc trưng với FFT (N = 1024)

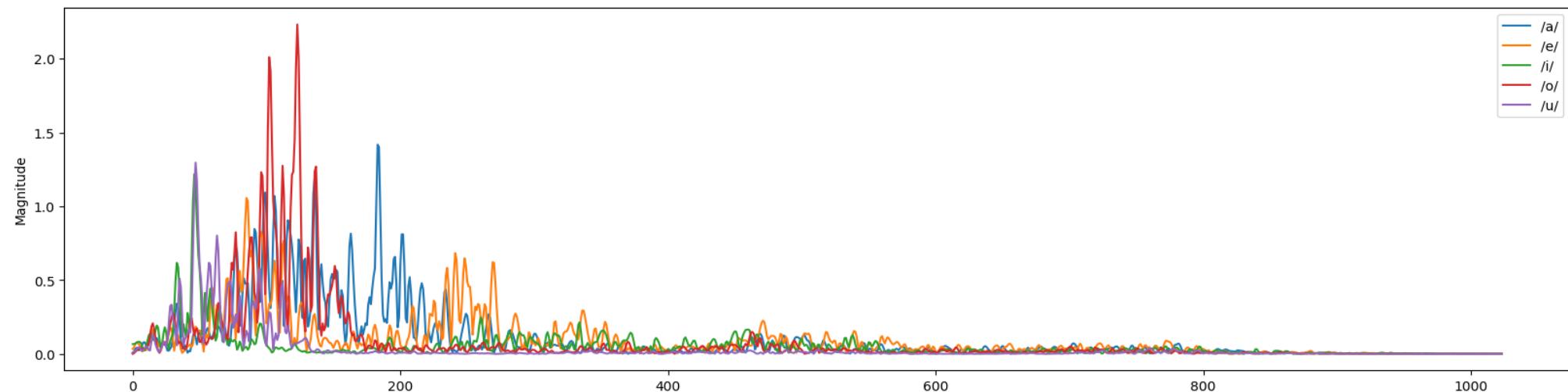
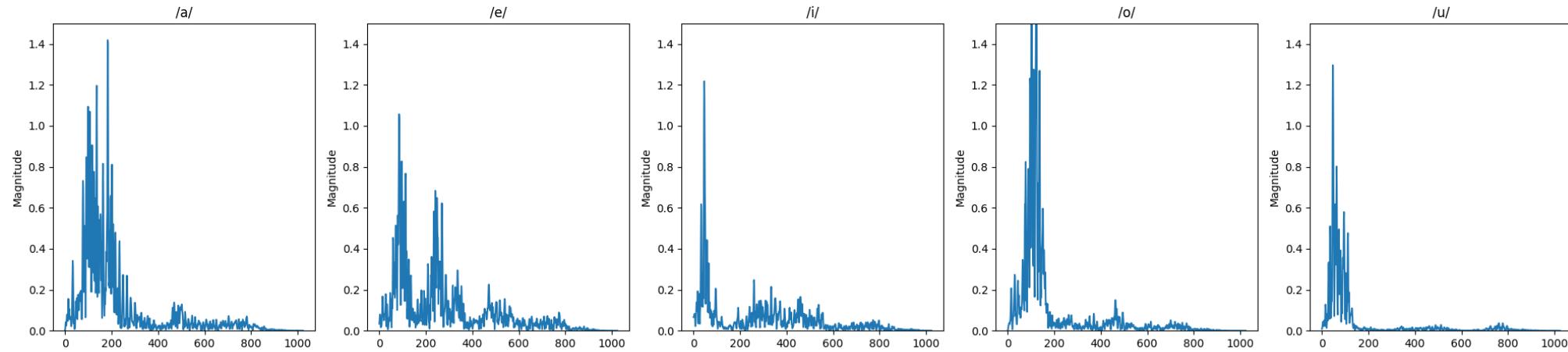


PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Kết quả thực nghiệm

N_FFT = 2048

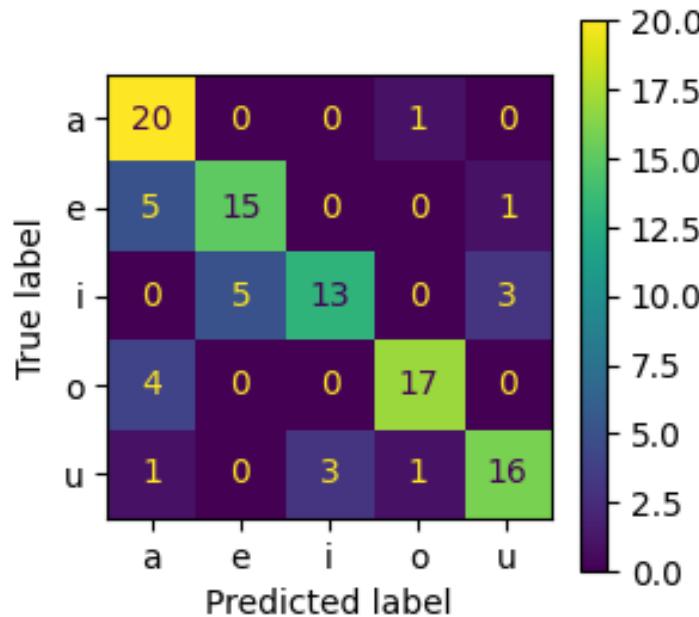
Vector đặc trưng với FFT (N = 2048)



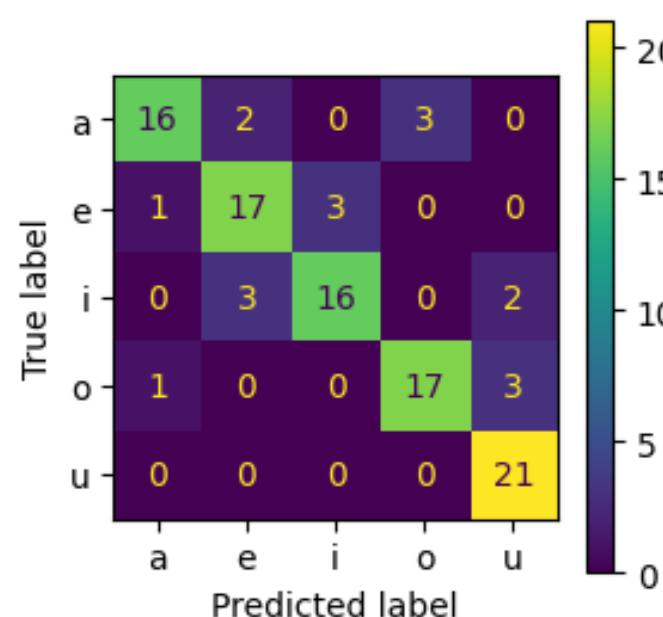
PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Kết quả thực nghiệm

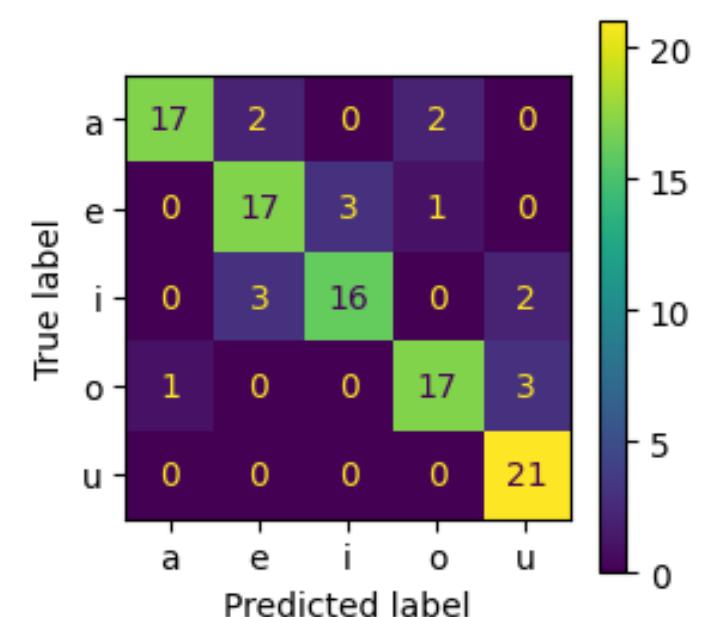
(N_FFT = 512)



(N_FFT = 1024)



(N_FFT = 2048)



Ma trận nhầm lẫn

PHẦN 2: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI DÙNG ĐẶC TRƯNG PHỔ FFT

Nhận xét và kết luận

	/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
N_FFT = 512	95,24%	71,43%	61,90%	80,95%	76,19%	77,14%
N_FFT = 1024	76,19%	80,95%	76,19%	80,95%	100%	82,86%
N_FFT = 2048	80,95%	80,95%	76,19%	80,95%	100%	83,81%

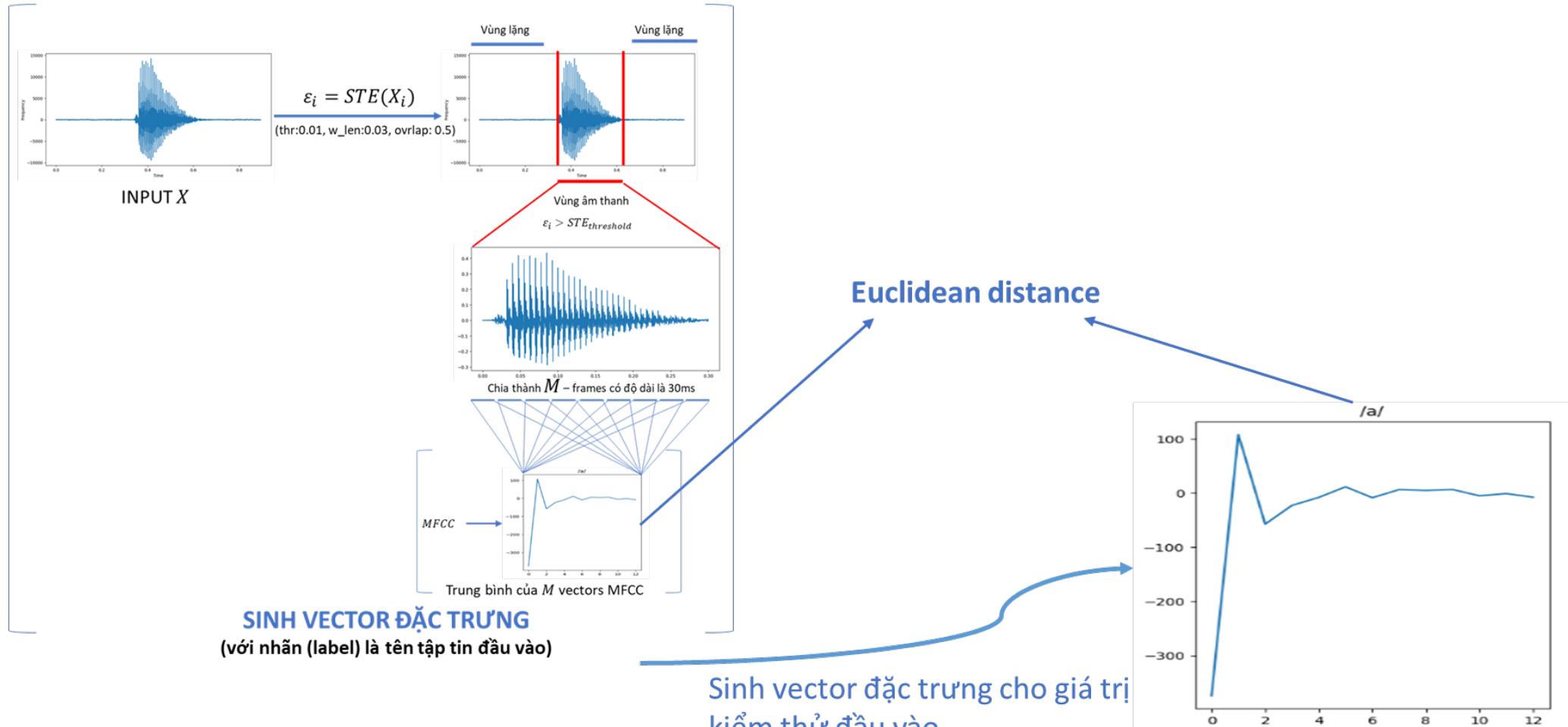
- Nguyên âm /i/ có độ dự đoán chính xác trung bình thấp nhất (71.42%).
- Nguyên âm /u/ có độ dự đoán chính xác trung bình cao nhất (92.06%).
- Độ dự đoán chính xác trung bình khi N_FFT = 512 nhỏ nhất, N_FFT = 1024 và 2048 có độ chính xác cao hơn.
- Với N_FFT = 1024 và 2048, kết quả trung bình tương tự nhau (chênh lệch không đáng kể).
- Nhìn chung độ dự đoán chính xác trung bình khá cao (81.27%).

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

- Xây dựng phương pháp trích xuất đặc trưng sử dụng đặc trưng phổ MFCC
- Mô hình nhận dạng sử dụng đặc trưng phổ MFCC
- Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means clustering
- Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp KNN (K-Nearest Neighbors)
- Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp CatBoost

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

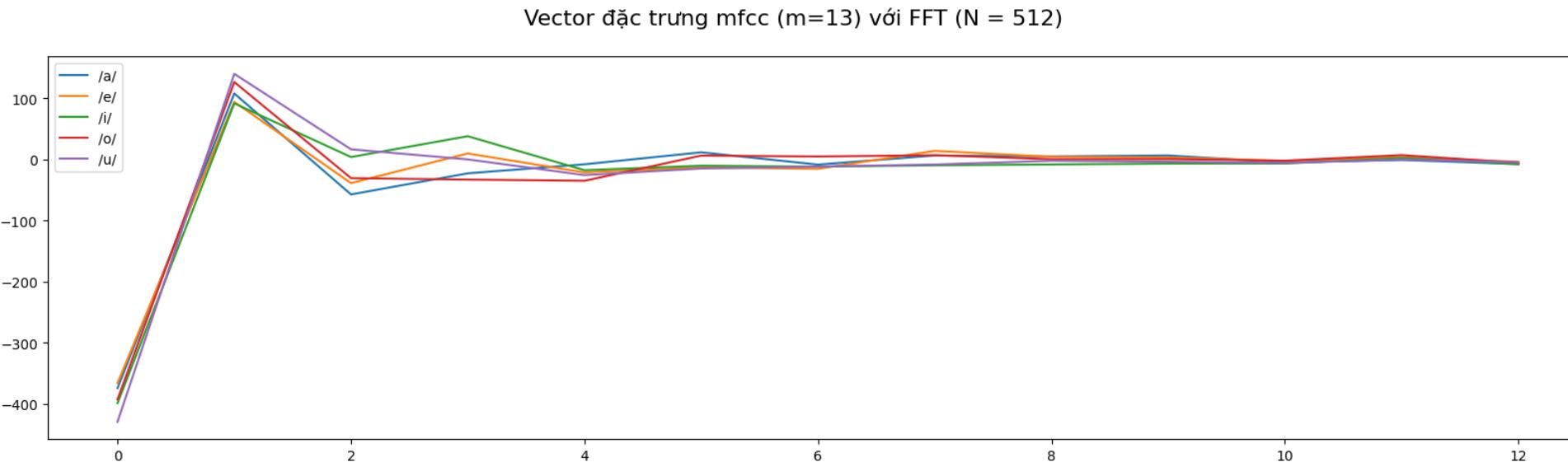
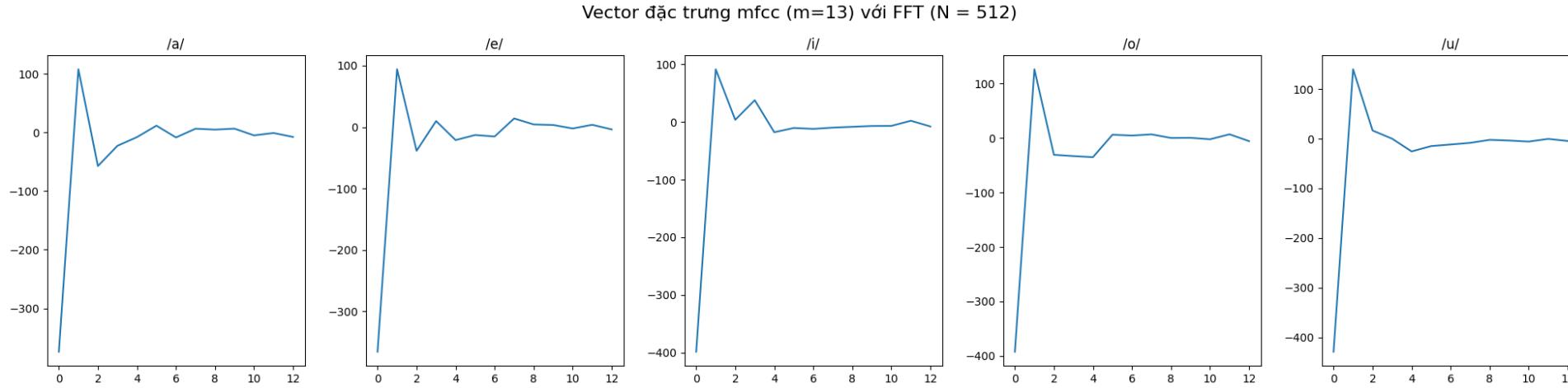
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC



Thực hiện với những nhãn khác (còn lại)
Sau đó xem xét giá trị Euclidean distance thấp nhất là nhãn của đầu vào

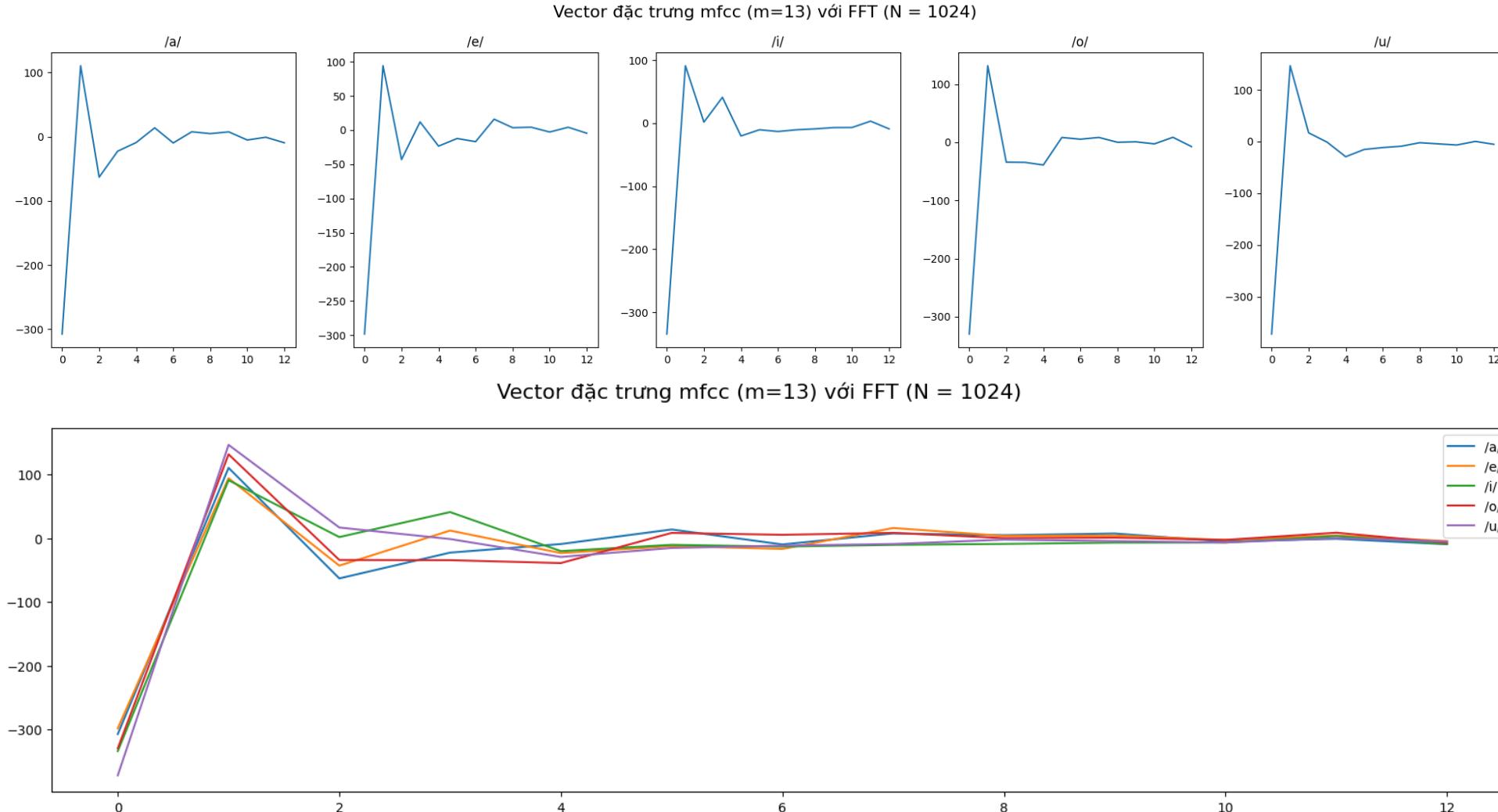
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC



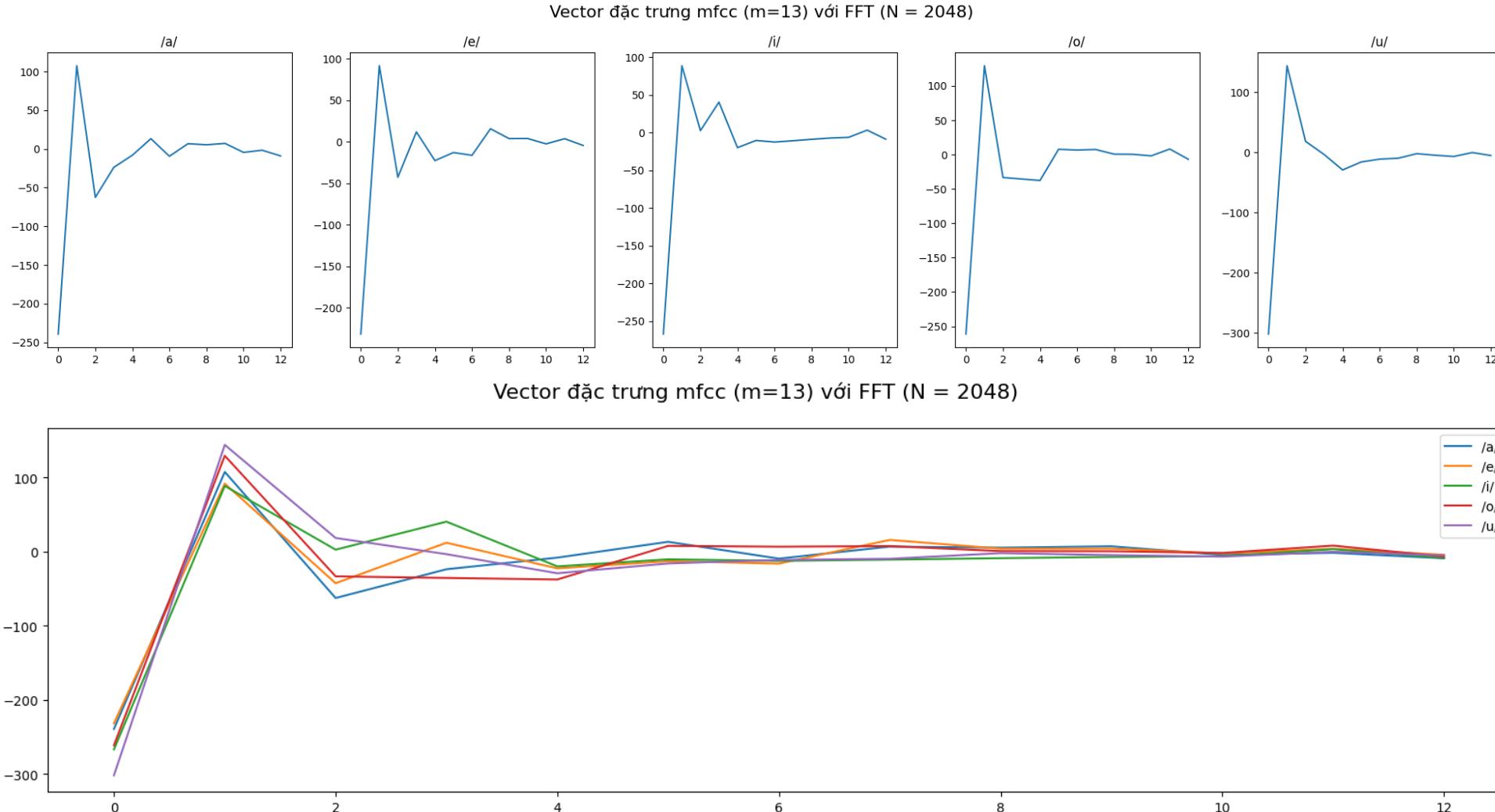
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC



PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

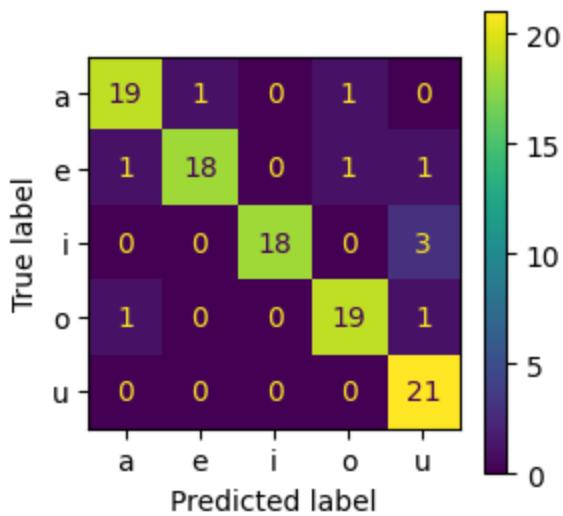
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC



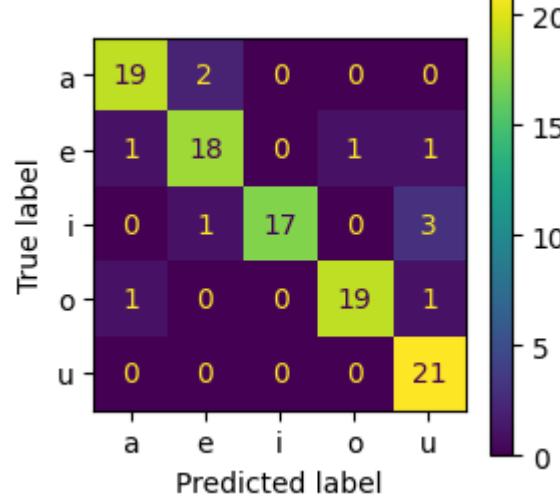
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC

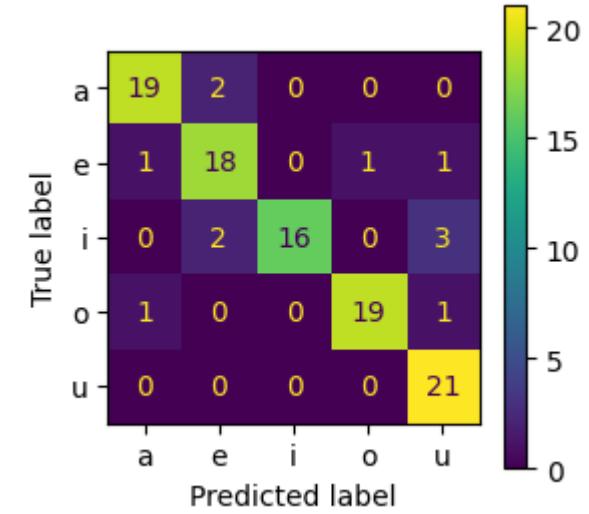
Kết quả thực nghiệm (Confusion Matrix)



(N_FFT = 512)
(acc: 0.904761904762)



(N_FFT = 1024)
(acc: 0.895238095238)



(N_FFT = 2048)
(acc: 0.885714285714)

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC

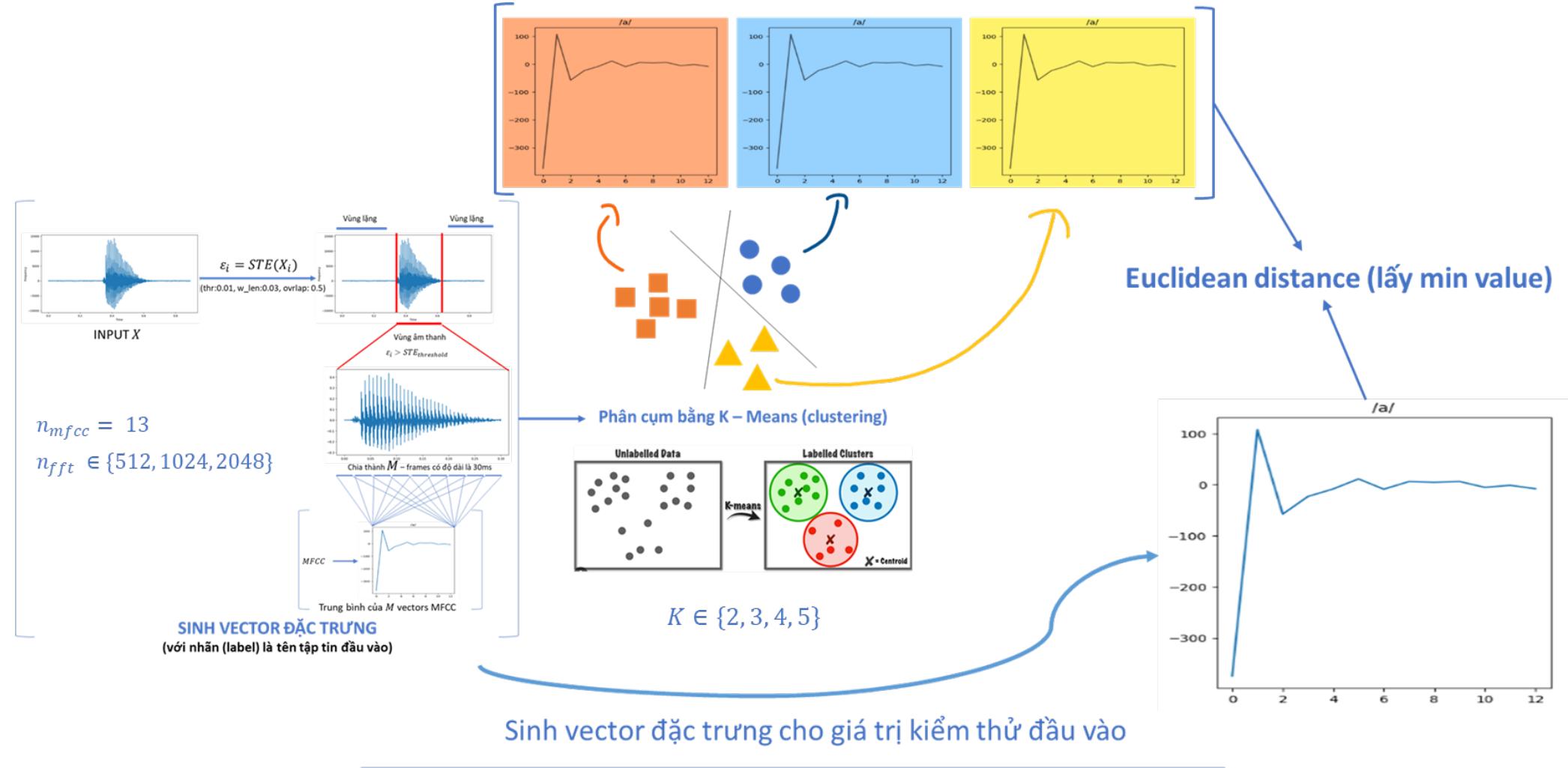
Tổng hợp & nhận xét (trên tập kiểm thử)

	/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
N_FFT = 512	90,48%	85,71%	85,71%	90,48%	100%	90,48%
N_FFT = 1024	90,48%	85,71%	80,95%	90,48%	100%	89,52%
N_FFT = 2048	90,48%	85,71%	76,19%	90,48%	100%	88,57%

- Độ chính xác dự đoán trung bình của **N_FFT = 512** là cao nhất với acc: **90.48%**, thấp nhất là N_FFT = 2048 với acc: 88.57% .
- Độ chính xác dự đoán nguyên âm **/u/** là 100%, **/a/** và **/o/** là 90.48%, **/e/** và **/i/** là 85.71% (Tại N_FFT = 512)

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means



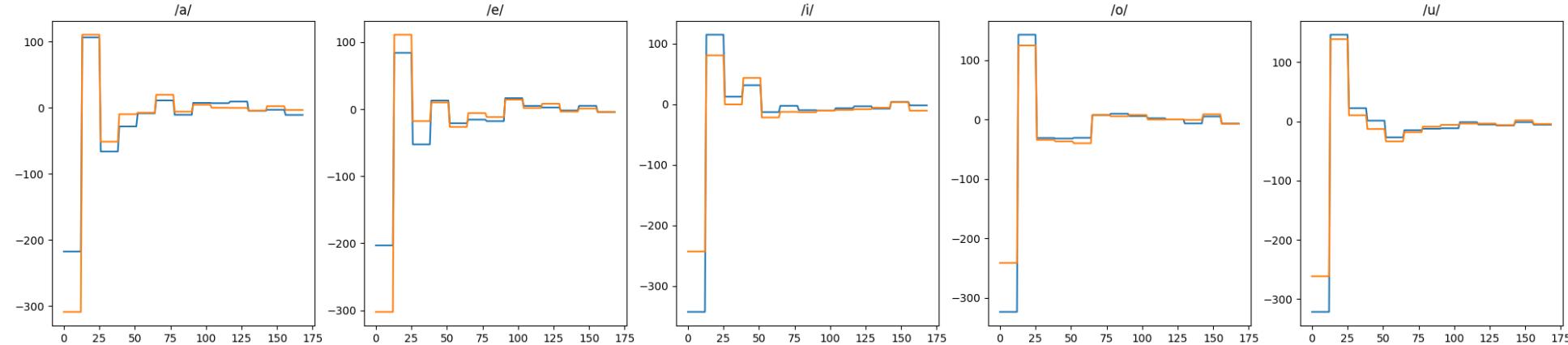
Thực hiện với những nhãn khác (còn lại)

Sau đó xem xét giá trị Euclidean distance thấp nhất là nhãn của đầu vào

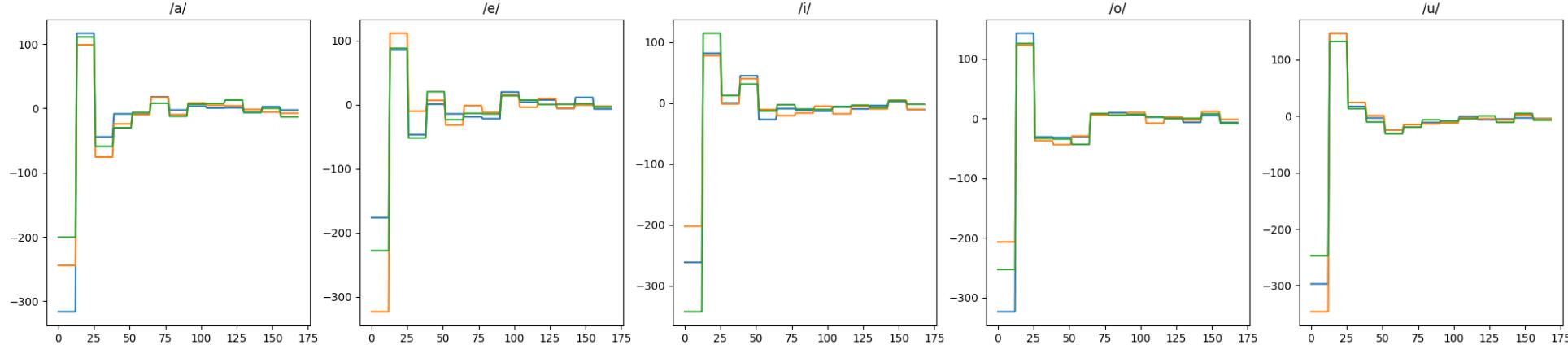
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

Vector đặc trưng mfcc ($m=13$) với FFT ($N = 2048$) - Phân cụm $K = 2$



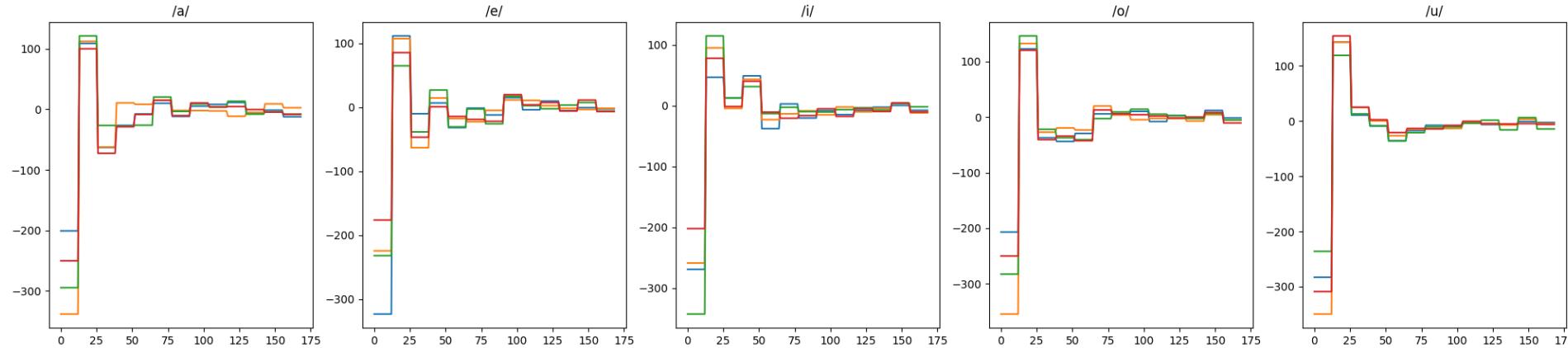
Vector đặc trưng mfcc ($m=13$) với FFT ($N = 2048$) - Phân cụm $K = 3$



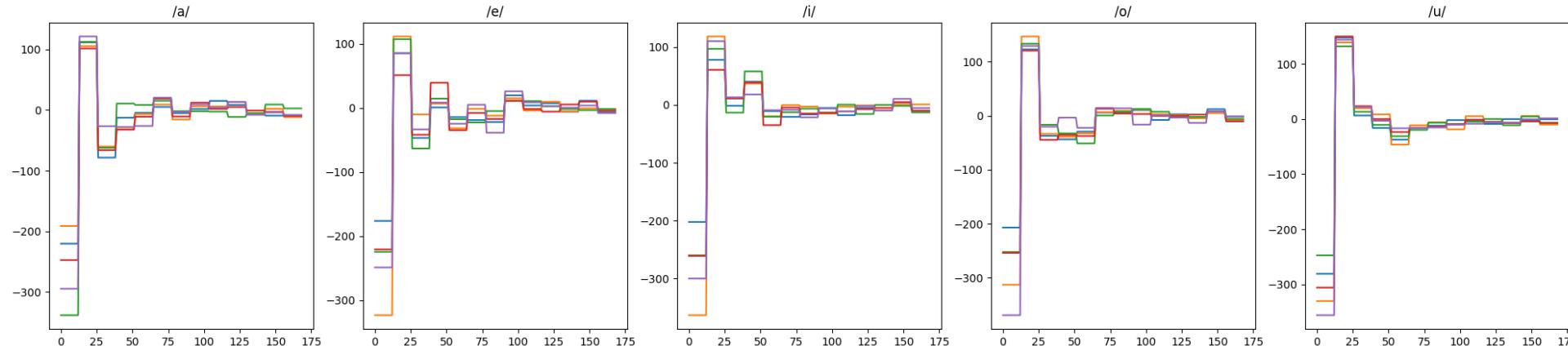
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

Vector đặc trưng mfcc ($m=13$) với FFT ($N = 2048$) - Phân cụm $K = 4$



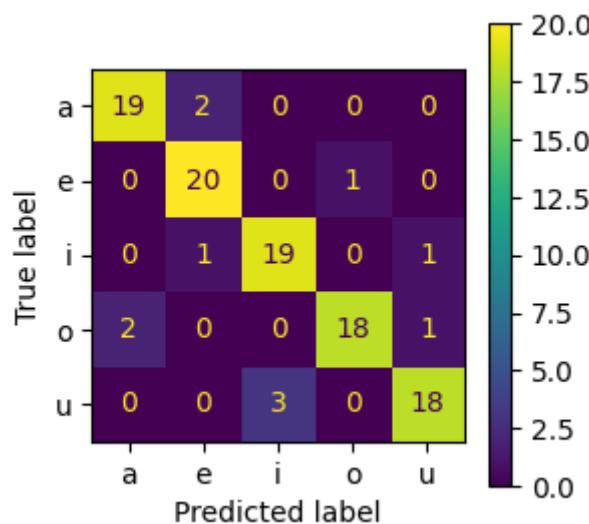
Vector đặc trưng mfcc ($m=13$) với FFT ($N = 2048$) - Phân cụm $K = 5$



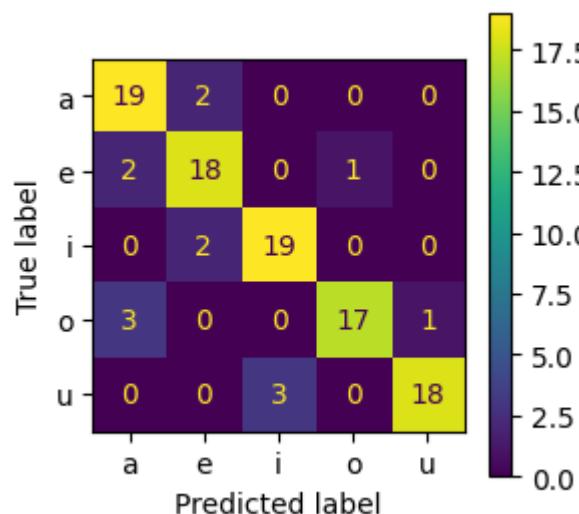
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

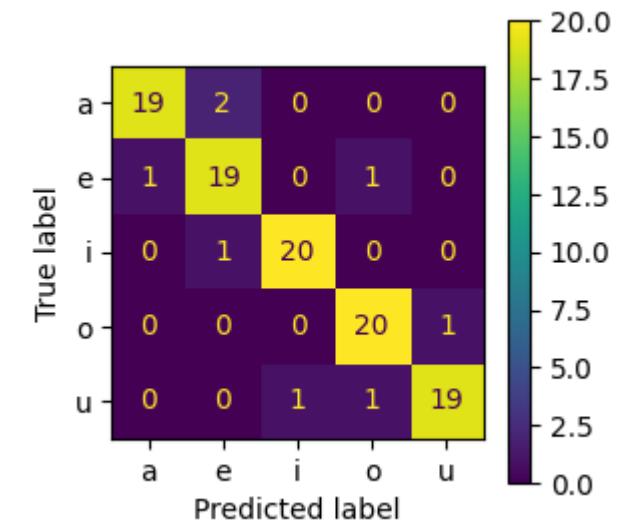
Kết quả thực nghiệm tốt nhất (Confusion Matrix)



(N_FFT = 512, K = 5)
(acc: 0.895238095238)



(N_FFT = 1024, K = 5)
(acc: 0.8666666666667)



(N_FFT = 2048, K = 5)
(acc: 0.923809523809)

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

Bảng tổng hợp kết quả độ chính xác sau lần 1
(PP. K – means có sự khác nhau sau mỗi lần chạy)

		/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
K = 2	N_FFT = 512	95,24%	85,71%	90,48%	85,71%	90,48%	89,52%
	N_FFT = 1024	90,48%	85,71%	90,48%	85,71%	90,48%	88,57%
	N_FFT = 2048	90,48%	85,71%	76,19%	90,48%	90,48%	86,67%
K = 3	N_FFT = 512	95,24%	85,71%	90,48%	95,24%	90,48%	91,43%
	N_FFT = 1024	90,48%	80,95%	85,71%	95,24%	90,48%	88,57%
	N_FFT = 2048	90,48%	80,95%	76,19%	95,24%	90,48%	86,67%
K = 4	N_FFT = 512	85,71%	90,48%	90,48%	90,48%	90,48%	89,53%
	N_FFT = 1024	85,71%	90,48%	85,71%	85,71%	90,48%	87,62%
	N_FFT = 2048	85,71%	90,48%	85,71%	85,71%	90,48%	87,62%
K = 5	N_FFT = 512	85,71%	95,24%	80,95%	90,48%	90,48%	88,57%
	N_FFT = 1024	90,48%	90,48%	95,24%	80,95%	90,48%	89,53%
	N_FFT = 2048	85,71%	90,48%	90,48%	90,48%	100%	91,43%

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

Bảng tổng hợp kết quả độ chính xác sau lần 2
(PP. K – means có sự khác nhau sau mỗi lần chạy)

K-Means		/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
K = 2	N_FFT = 512	85,71%	85,71%	95,24%	80,95%	95,24%	88,57%
	N_FFT = 1024	90,48%	85,71%	90,48%	85,71%	90,48%	88,57%
	N_FFT = 2048	90,48%	85,71%	76,19%	85,71%	90,48%	85,71%
K = 3	N_FFT = 512	95,24%	85,71%	90,48%	90,48%	90,48%	90,48%
	N_FFT = 1024	90,48%	80,95%	90,48%	90,48%	90,48%	88,57%
	N_FFT = 2048	90,48%	80,95%	80,95%	95,24%	90,48%	87,62%
K = 4	N_FFT = 512	85,71%	90,48%	95,24%	85,71%	85,71%	88,57%
	N_FFT = 1024	85,71%	90,48%	80,95%	80,95%	90,48%	85,71%
	N_FFT = 2048	85,71%	90,48%	80,95%	85,71%	90,48%	86,67%
K = 5	N_FFT = 512	85,71%	95,24%	90,48%	85,71%	85,71%	88,57%
	N_FFT = 1024	90,48%	85,71%	80,95%	80,95%	90,48%	85,71%
	N_FFT = 2048	90,48%	90,48%	95,24%	90,48%	85,71%	90,48%

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

**Bảng tổng hợp kết quả độ chính xác tốt nhất sau nhiều lần chạy thử nghiệm
(PP. K – means có sự khác nhau sau mỗi lần chạy)**

K-Means		/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
K = 2	N_FFT = 512	95,24%	85,71%	95,24%	85,71%	95,24%	91,43%
	N_FFT = 1024	90,48%	85,71%	90,48%	85,71%	90,48%	88,57%
	N_FFT = 2048	90,48%	85,71%	76,19%	85,71%	90,48%	85,71%
K = 3	N_FFT = 512	90,48%	80,95%	85,71%	95,24%	90,48%	88,57%
	N_FFT = 1024	85,71%	80,95%	85,71%	95,24%	90,48%	87,62%
	N_FFT = 2048	85,71%	80,95%	80,95%	95,24%	90,48%	86,67%
K = 4	N_FFT = 512	90,48%	95,24%	80,95%	85,71%	90,48%	88,57%
	N_FFT = 1024	85,71%	90,48%	90,48%	85,71%	90,48%	88,57%
	N_FFT = 2048	85,71%	90,48%	85,71%	85,71%	90,48%	87,62%
K = 5	N_FFT = 512	90,48%	95,24%	90,48%	85,71%	85,71%	89,52%
	N_FFT = 1024	90,48%	85,71%	90,48%	80,95%	85,71%	86,67%
	N_FFT = 2048	90,48%	90,48%	95,24%	95,24%	90,48%	92,38%

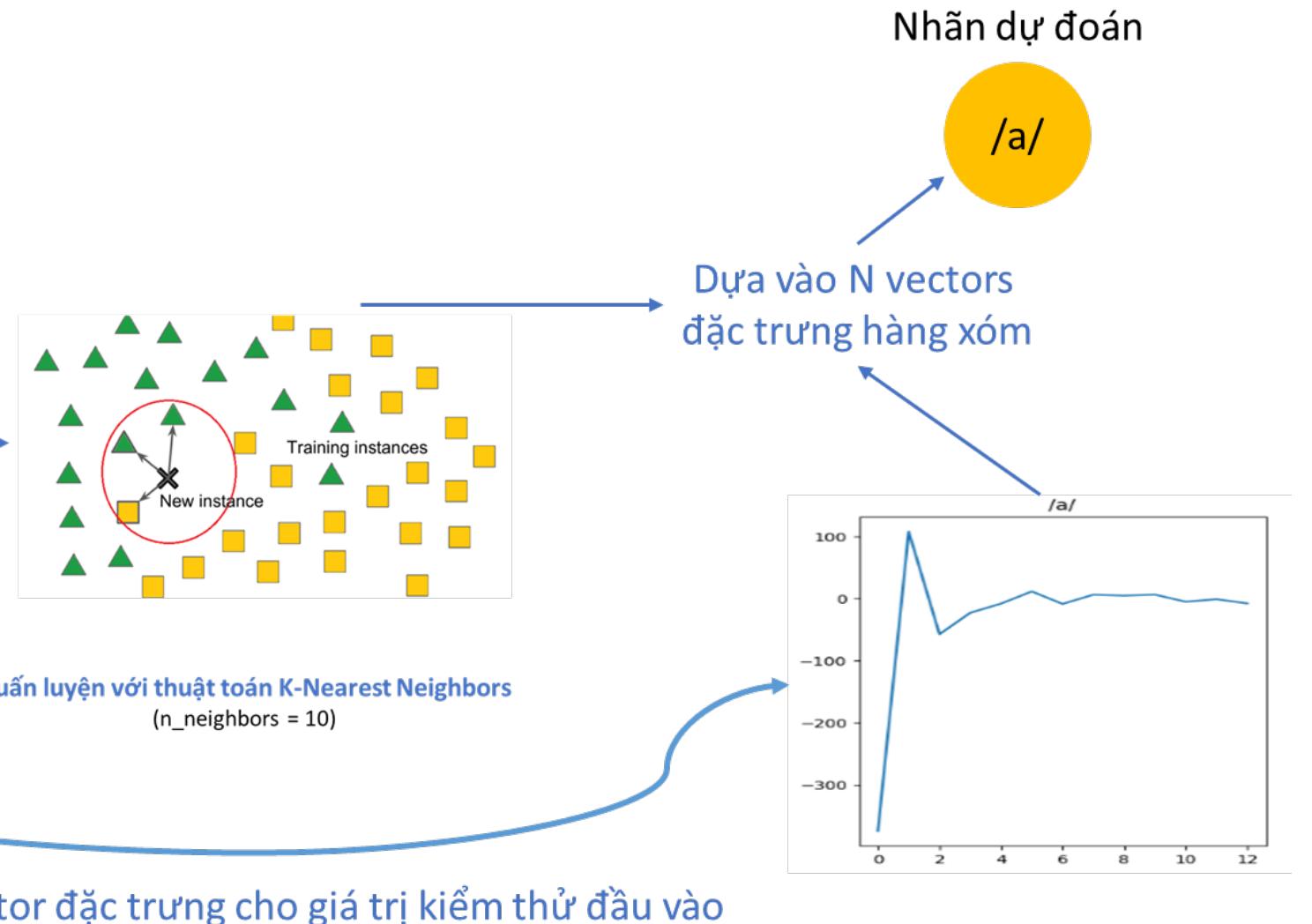
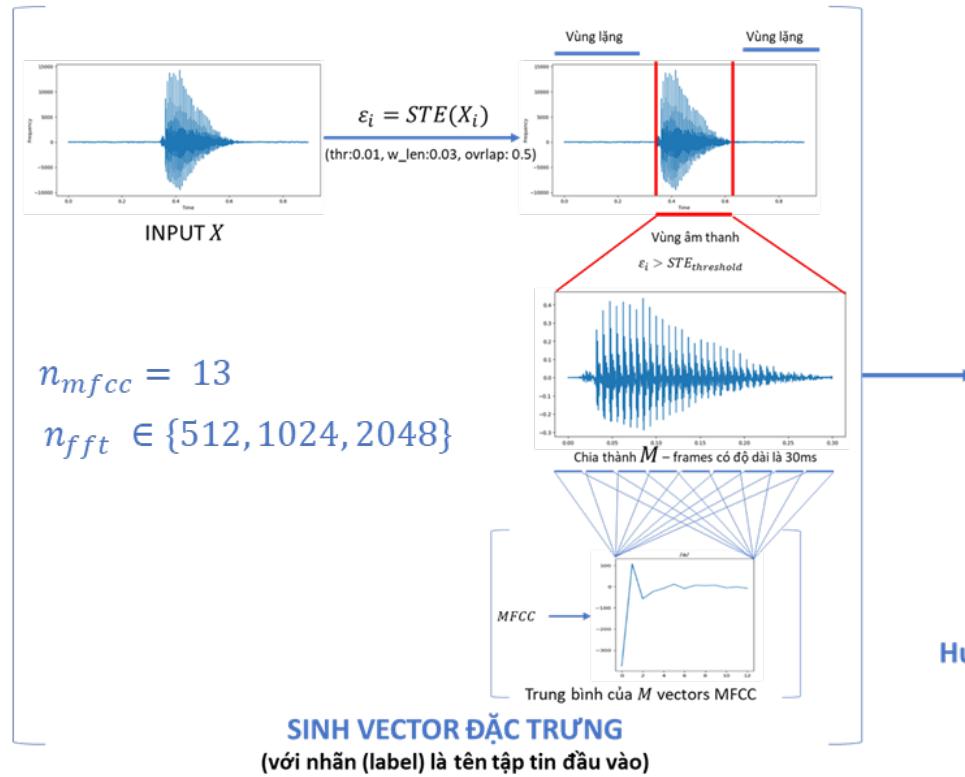
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp K-means

NHẬN XÉT

- Phương pháp kết hợp với K-means (trong đó K-means dùng để phân cụm) **có sự thay đổi kết quả sau những lần chạy khác nhau.**
- Các nhóm được phân cụm có các vectors đặc trưng khác nhau.
- Độ chính xác trung bình các nhãn là **92.38% (N_FFT = 2048, K = 5)** – (Cao nhất trong N lần chạy)
- Ở kết quả tốt nhất cho thấy **cả 5 âm đều cho độ chính xác trên 90%**, trong đó **/i/ và /o/ cho kết quả trên 95%**

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

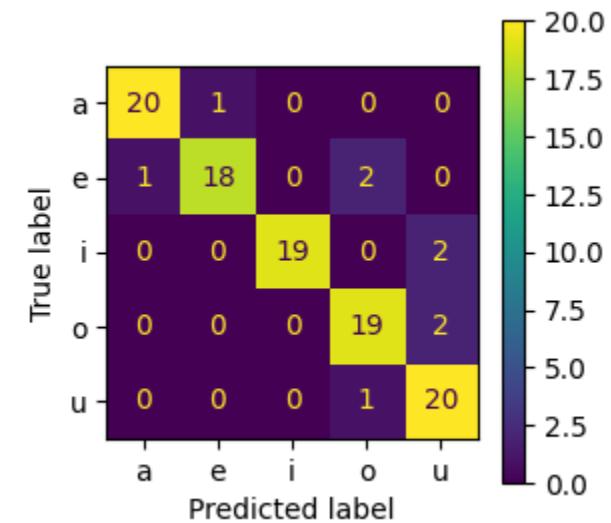
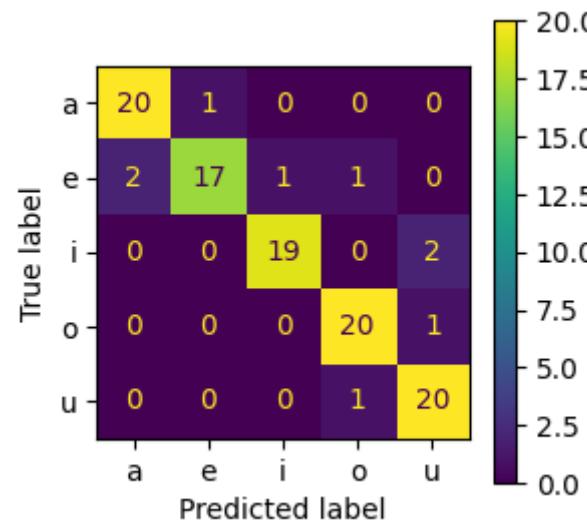
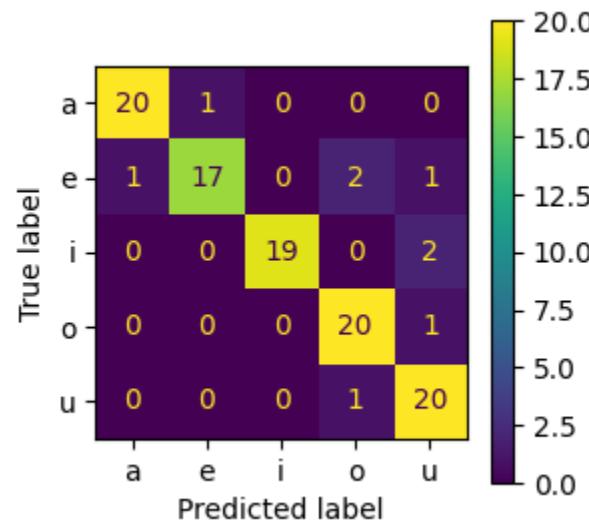
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp KNN



PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp KNN

Kết quả thực nghiệm (Confusion Matrix)



PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp KNN

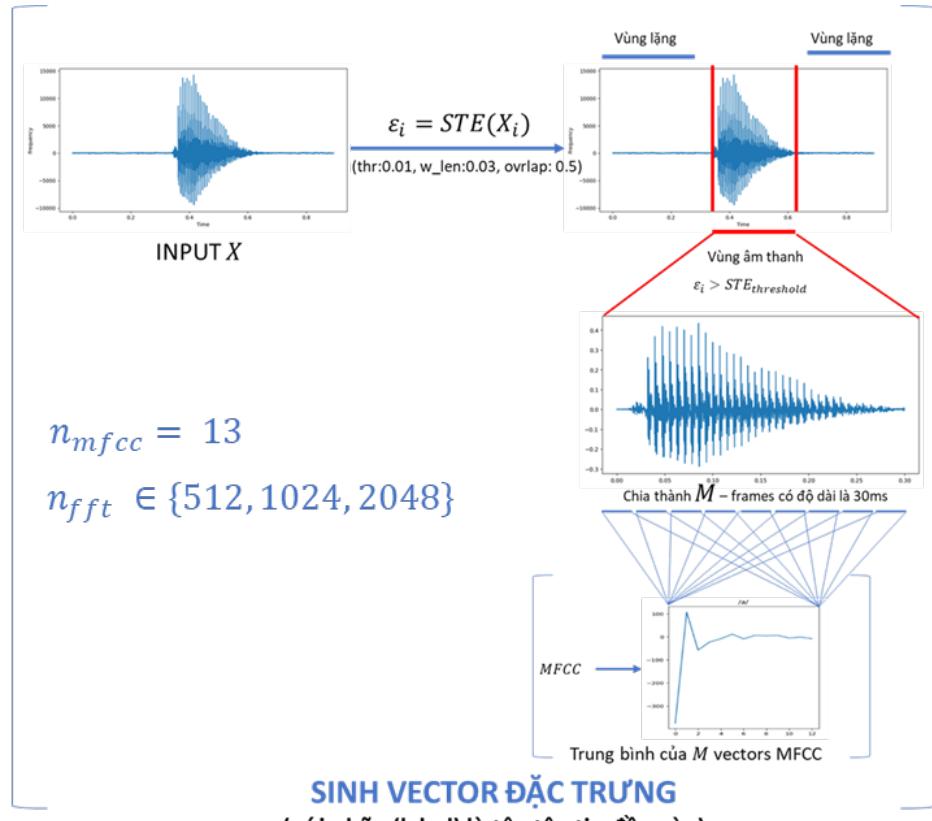
Tổng hợp & nhận xét (trên tập kiểm thử)

	/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
N_FFT = 512	95,24%	80,95%	90,48%	95,24%	95,24%	91,43%
N_FFT = 1024	95,24%	80,95%	90,48%	95,24%	95,24%	91,43%
N_FFT = 2048	95,24%	85,71%	90,48%	90,48%	95,24%	91,43%

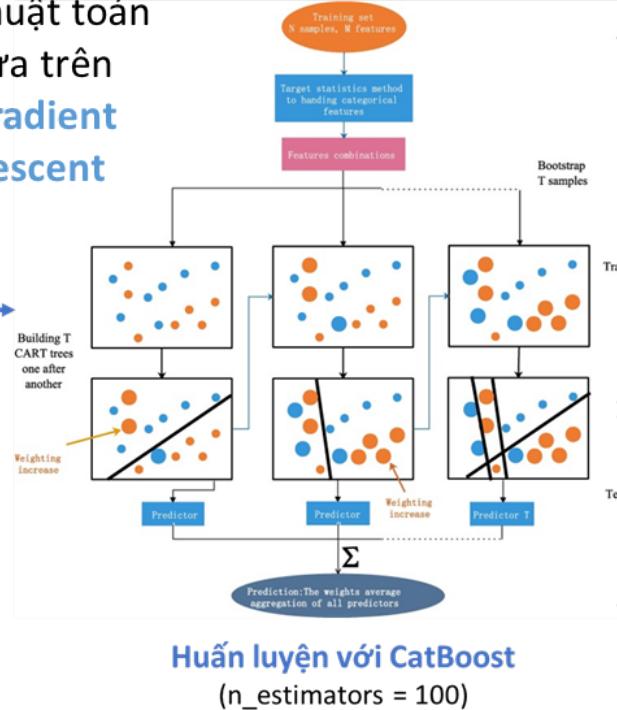
- Độ chính xác trung bình của cả 3 tham số NFFT đều không có thay đổi, tuy nhiên **các nhãn (nguyên âm) lại có thay đổi về kết quả dự đoán.**
- Kết quả dự đoán không thay đổi với mỗi lần chạy thuật toán KNN (Tổng quan).
- Nguyên âm /a/ và /u/ có độ dự đoán chính xác cao nhất 95,24% và /e/ có độ chính xác thấp nhất 85.71%. (Xét N_FFT = 2048)

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

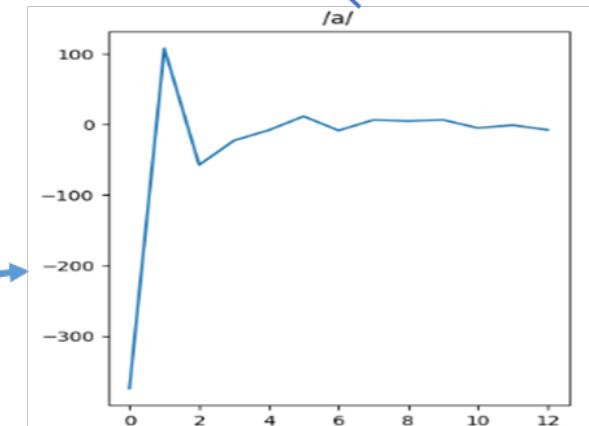
Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp CatBoost



Thuật toán
dựa trên
**Gradient
Descent**



Nhận dự đoán

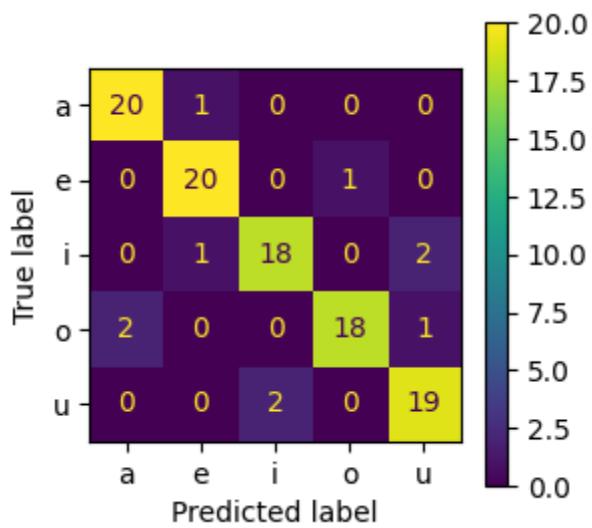


Sinh vector đặc trưng cho giá trị kiểm thử đầu vào

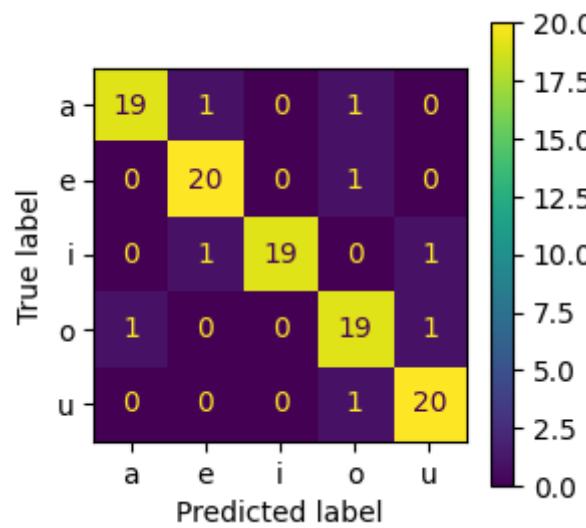
PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp CatBoost

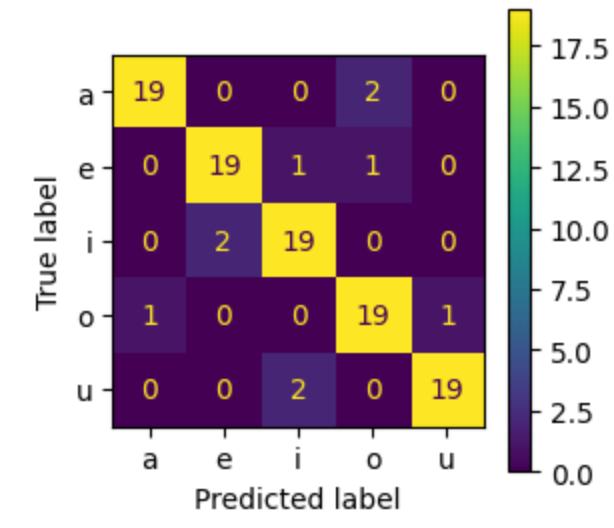
Kết quả thực nghiệm (Confusion Matrix)



(N_FFT = 512)
(acc: 0.904761904762)



(N_FFT = 1024)
(acc: 0.92380952381)



(N_FFT = 2048)
(acc: 0.904761904762)

PHẦN 3: NHẬN DẠNG NGUYÊN ÂM KHÔNG PHỤ THUỘC NGƯỜI NÓI ĐẶC TRƯNG PHỔ MFCC

Mô hình nhận dạng sử dụng đặc trưng phổ MFCC kết hợp CatBoost

Tổng hợp & nhận xét (trên tập kiểm thử)

	/a/	/e/	/i/	/o/	/u/	Độ chính xác trung bình (%)
N_FFT = 512	95,24%	95,24%	85,71%	85,71%	90,48%	90,48%
N_FFT = 1024	90,48%	95,24%	90,48%	90,48%	95,24%	92,38%
N_FFT = 2048	90,48%	90,48%	90,48%	90,48%	90,48%	90,48%

- Độ chính xác dự đoán trung bình khi kết hợp **CatBoost cho kết quả trên 90%**.
- Độ chính xác dự đoán trung bình khi sử dụng **CatBoost cao nhất là 92,38%** với N_FFT = 1024
- Độ chính xác dự đoán của các nguyên âm (nhãn) đều trên 90%**
- Độ chính xác dự đoán cao nhất là hai nguyên âm /e/ và /u/ là 95.24%.** (Xét N_FFT = 1024)

PHẦN 4: SO SÁNH VÀ TỔNG KẾT

Phương pháp nhận dạng	Thời gian chạy	Độ chính xác	So với FFT
FFT	1.33s	83.81%	
MFCC	1p8s	90.48%	↑ 6.67%
MFCC + K-means (Clustering)	1p9s	92.38%	↑ 8.57%
MFCC + KNN	59s	91.43%	↑ 7.62%
MFCC + CatBoost	54s	92.38%	↑ 8.57%

- Hai phương pháp kết hợp MFCC với **K-means** và **CatBoost** đều có kết quả cao là **92.38%**, tuy nhiên **K-means** bị ảnh hưởng bởi các **lần chạy khác nhau**, còn CatBoost thì không.
- Phương pháp kết hợp cho kết quả tốt hơn so với việc không sử dụng.
- Thời gian chạy các thuật toán cho thấy:
 - FFT chạy nhanh nhất trong các phương pháp => phù hợp sử dụng trên các thiết bị nhỏ gọn, cấu hình yếu.
 - MFCC và MFCC + K-means phải duyệt hết toàn bộ mẫu feature vectors huấn luyện nên thời gian dự đoán lâu nhất.
 - MFCC + KNN & MFCC + CatBoost cho tốc độ cao hơn MFCC và MFCC + K-means vì chỉ dự đoán dựa vào tham số huấn luyện.

* Thời gian chạy được tính toán trên máy tính: Windows 11 ProW (WSL2 - Ubuntu)- CORE I7 6820HQ (single core) – 32GB RAM

THANK YOU FOR LISTENING...

QUESTION & ANSWER

GIT: <https://github.com/DUT-21TDT/XLTHS>