VNU - University of Engineering and Technology

BÁO CÁO CUỐI KHÓA

Môn học: Xử lý tiếng nói

Tên đề tài: Tìm kiếm hình ảnh thông qua nhận diện giọng nói

Sinh viên: Nguyễn Quốc Khánh **MSSV: 18020710**
Nguyễn Tất Đạt **MSSV: 18020009**

MỤC LỤC

1. Giới thiệu bài toán

Trong quá trình học một ngôn ngữ mới, việc học theo phương pháp học theo từng cặp phiên dịch ngoại ngữ - ngôn ngữ dich (ví dụ: căn hộ - apartment) là rất nhàm chán và khá khó để ghi nhớ. Ứng dụng tìm kiếm hình ảnh bằng nhận diện giọng nói ra đời nhằm giúp việc học ngôn ngữ trở nên dễ dàng và trực quan hơn. Trong phạm vi môn học, ứng dụng được xây dựng để có thể nhận diện giọng nói của 4 từ "căn hộ", "cảnh sát", "học sinh", "người". Phía người dùng có thể chọn file audio có sẵn của một trong 4 từ trên ở máy của mình và upload lên ứng dụng hoặc tự mình ghi âm bằng option record. Ứng dụng sẽ nhận diện từ được nói và tìm kiếm hình ảnh dựa trên kết quả nhận diện.

2. Dữ liệu, mô hình và kết quả huấn luyện

2.1.Thu thập dữ liệu

- Dữ liệu huấn luyện bao gồm file wav của các từ "căn hộ", "cảnh sát", "học sinh", "người" được lấy từ các nguồn:
    - Được cắt ra từ file âm thanh record giọng đọc của các thành viên trong lớp.
    - Được thu thập thêm từ nhiều khác, trong đó nguồn dữ liệu tiếng miền Trung khá phong phú.
- Tổng số lượng bản ghi thu thập được:
    - căn hộ: 499
    - cảnh sát: 477
    - học sinh: 488
    - người: 497

2.2.Xử lý dữ liệu

- Chuyển tất cả các file wav sang dạng mfcc để chuẩn bị cho quá trình huấn luyện.
- Bởi vì các file wav có độ dài khác nhau nên khi chuyển sang dạng mfcc, chọn một độ dài cố định cho tất cả các mfcc.
    - Đối với tất cả các mfcc có độ dài nhỏ hơn độ dài mfcc cố định, thực hiện việc padding.
    - Đối với tất cả các mfcc có độ dài lớn hơn độ dài cố định, chỉ lấy phần có độ dài bằng độ dài cố định.

- Việc chọn các tham số về n_mfcc và độ dài cố định là rất quan trọng, chúng có thể được chọn bằng cách thử chọn và quan sát kết quả thí nghiệm.
- Độ dài cố định nên được chọn đủ lớn để có thể bao quát được nhiều thông tin. Tuy nhiên, quan sát trong bộ dữ liệu thu thập được các file thường có khoảng lặng phía sau do thói quen của người ghi âm, cho nên độ dài cố định không nên quá dài để tránh đi việc lấy các khoảng lặng có nhiễu.
- Chúng em cố định n_mfcc là 20.
- Chọn độ dài cố định bằng cách:
    - Với mỗi dãy mfcc của các từ, tính mean, median, tứ phân vị thứ ba của các dãy, lấy ra mean lớn nhất, median lớn nhất và tứ phân vị thứ ba lớn nhất.
    - Thử chọn độ dài cố định bằng các giá trị vừa lấy ra ở trên và quan sát kết quả.

```
import os
import librosa
import numpy as np

means = []
medians = []
third_quatiles = []

for label in ['can ho', 'canh sat', 'hoc sinh', 'nguoi']:
    paths = os.listdir('all/' + label)
    len_list = []
    for path in paths:
        wave, sr = librosa.load('all/'+label+'/'+ path, mono=True, sr=None)
        wave = np.asfortranarray(wave[::3])
        mfcc = librosa.feature.mfcc(wave, sr=16000, n_mfcc=20)
        len_list.append(mfcc.shape[1])
    means.append(sum(len_list)/len(len_list))
    medians.append(np.percentile(len_list, 50))
    third_quatiles.append(np.percentile(len_list, 75))

print(f'Max mean {max(means)}, max median {max(medians)}, max third quatile {max(third_quatiles)}')
```

```
Max mean 27.315573770491802, max median 20.0, max third quatile 37.0
```

Hình 2.1: Code lấy max mean, median và third quartile cho dãy độ dài các mfcc.

- Sau khi convert tất cả các file wav sang mfcc, lưu tất cả các giá trị mfcc trong file "mfccs.npy", tạo một file "label.csv" có chứa label tương ứng với từ trong các file wav.

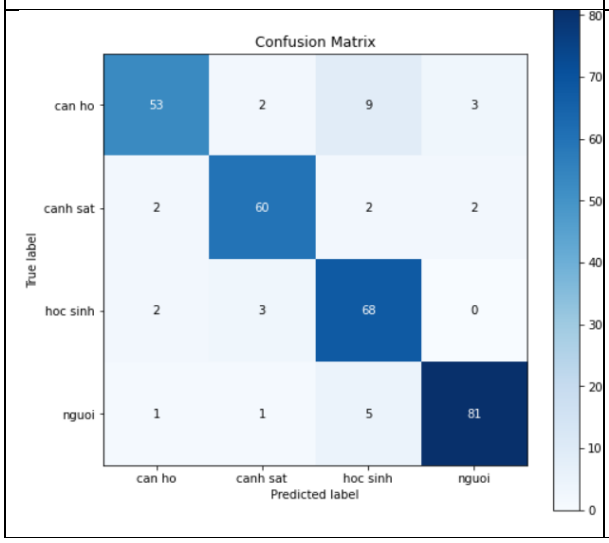
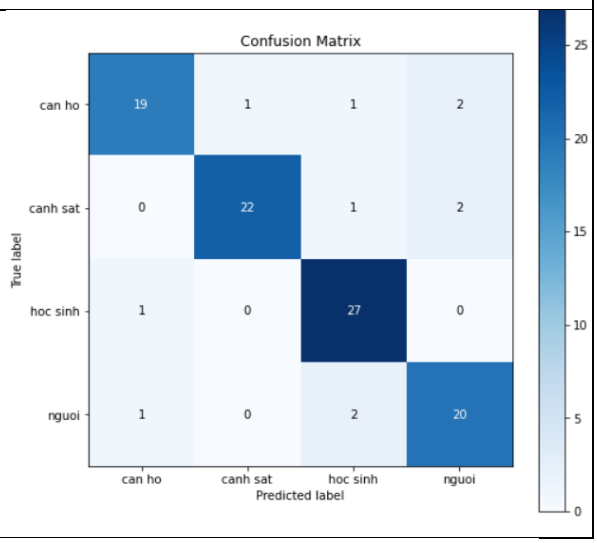
2.3.Mô hình

```
LitVoice(
  (conv1): Conv2d(1, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (conv2): Conv2d(32, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (linear1): Linear(in_features=800, out_features=64, bias=True)
  (linear2): Linear(in_features=64, out_features=5, bias=True)
  (accuracy): Accuracy()
)
```

Hình 2.2: Kiến trúc mô hình huấn luyện nhận diện giọng nói.

2.4.Kết quả huấn luyện

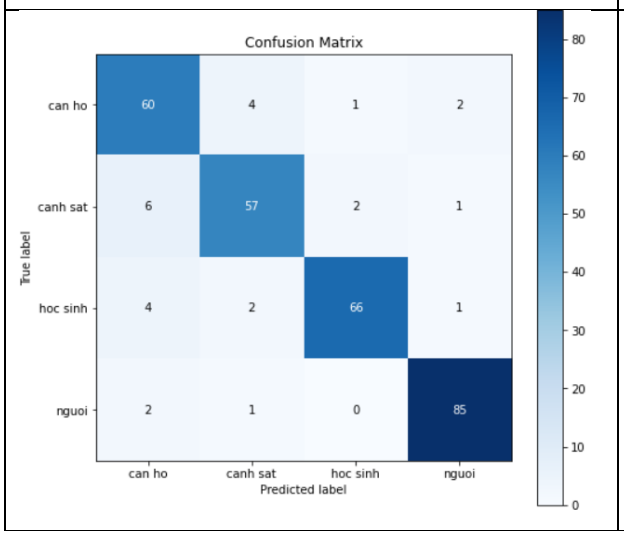
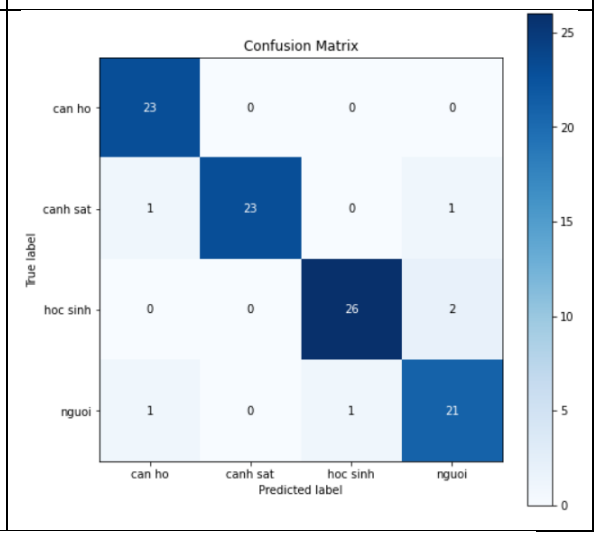
- **Độ dài cố định: 20 (chọn theo max median):**

On validation set	On test set
	


```
[{'Test accuracy': 0.8888888955116272, 'Test loss': 0.449354887008667}]
```

Hình 2.3: Kết quả dự đoán mô hình trên tập validation và test với độ dài cố định là 20.

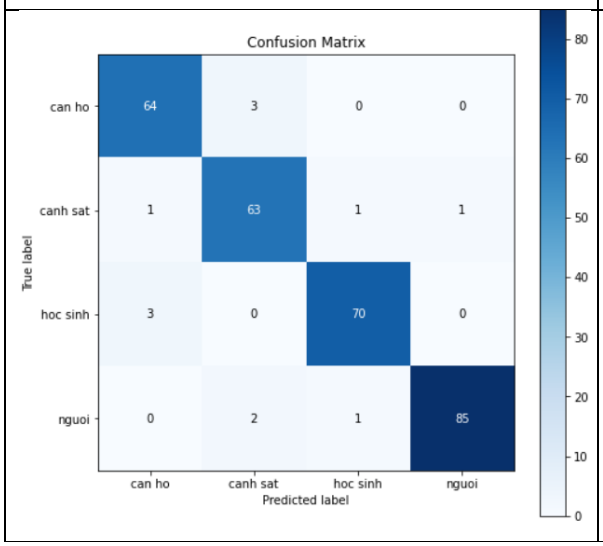
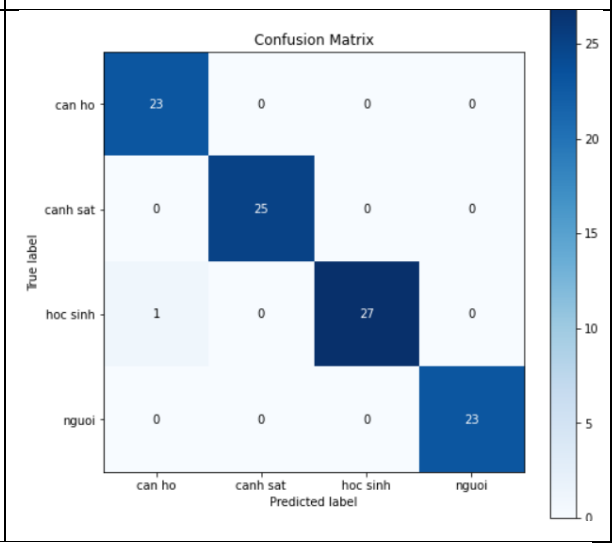
- **Độ dài cố định: 28 (chọn theo max mean):**

On validation set	On test set
	


```
[{'Test accuracy': 0.939393937587738, 'Test loss': 0.31787109375}]
```

Hình 2.4: Kết quả dự đoán mô hình trên tập validation và test với độ dài cố định là 28.

- **Độ dài cố định: 40 (chọn theo max third quantile):**

On validation set	On test set
	

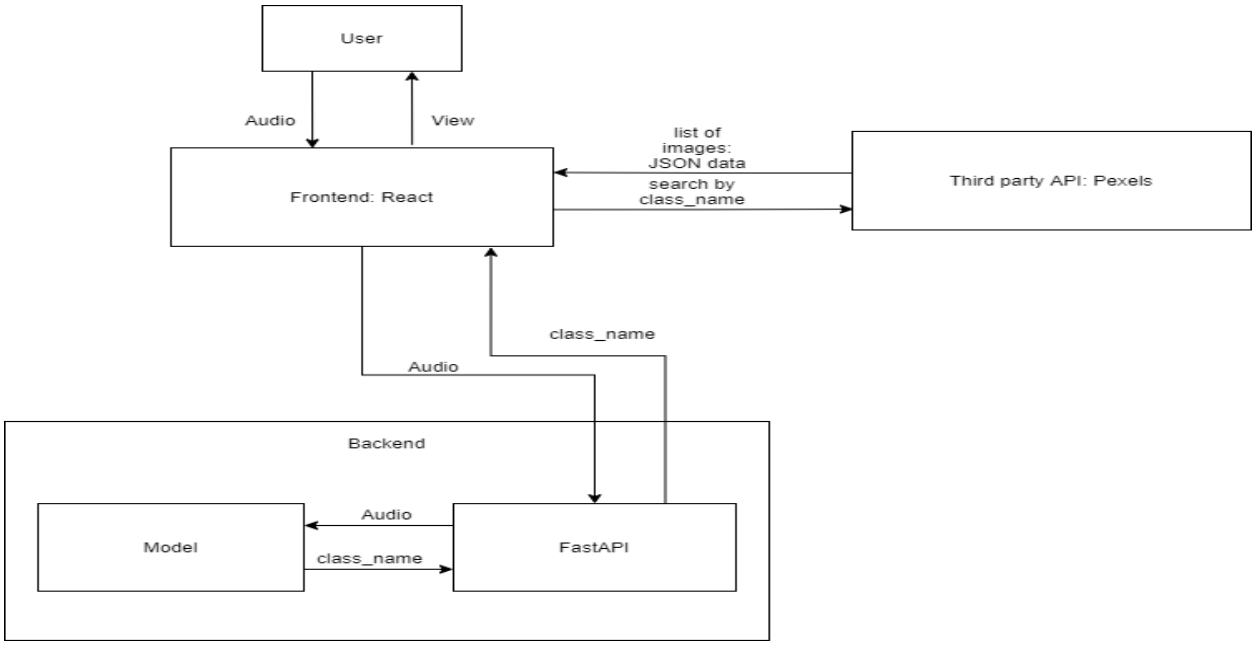

```
[{'Test accuracy': 0.9898989796638489, 'Test loss': 0.03399985656142235}]
```

Hình 2.5: Kết quả dự đoán mô hình trên tập validation và test với độ dài cố định là 28.

Nhận xét: với độ dài cố định là 40 thì mô hình huấn luyện cho kết quả tốt nhất.

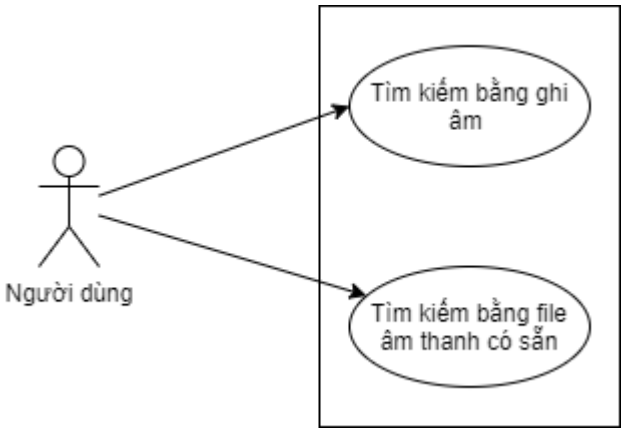
3. Phân tích và thiết kế ứng dụng

3.1.Kiến trúc tổng quan



Hình 3.1: Kiến trúc tổng quát của ứng dụng

3.2.Use case diagram



Hình 3.2: Use case diagram của ứng dụng

3.3.Phân tích ca sử dụng

- Ca sử dụng tìm kiếm bằng ghi âm
  - Trạng thái ban đầu: người dùng ở giao diện của ứng dụng
  - Mô tả: cho phép người dùng ghi âm bằng giọng của mình và dùng file ghi âm đó để tìm kiếm hình ảnh
  - Điều kiện đầu: người dùng ở giao diện của ứng dụng
  - Điều kiện cuối: các hình ảnh tìm kiếm được hiển thị trên ứng dụng
  - Kích bản chính

Người dùng	**Ứng dụng**
1. Bấm vào nút “Record” ở giao diện ứng dụng.	
	2. Chuyển sao giao diện record với các option “Start” và “Stop”.
3. Bấm vào nút “Start” để bắt đầu ghi âm và “Stop” để kết thúc ghi âm.	
	4. Nhận diện từ trong file ghi âm và tìm kiếm hình ảnh theo kết quả nhận diện và hiển thị trên giao diện ứng dụng.

5. Người dùng có thể xem thêm hình ảnh bằng cách nhấn nút “Load more” ở cuối danh sách hình ảnh.	

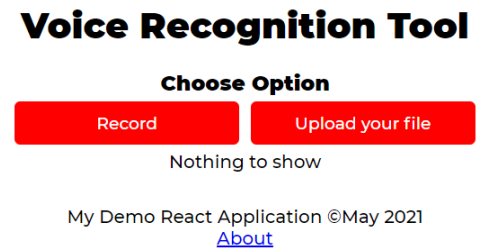
- Luồng thay thế: không
- Yêu cầu: không

- Ca sử dụng tìm kiếm bằng file có sẵn
  - Trạng thái ban đầu: người dùng ở giao diện của ứng dụng
  - Mô tả: cho phép người dùng ghi âm bằng giọng của mình và dùng file ghi âm đó để tìm kiếm hình ảnh
  - Điều kiện đầu: người dùng ở giao diện của ứng dụng
  - Điều kiện cuối: các hình ảnh tìm kiếm được hiển thị trên ứng dụng
  - Kích bản chính

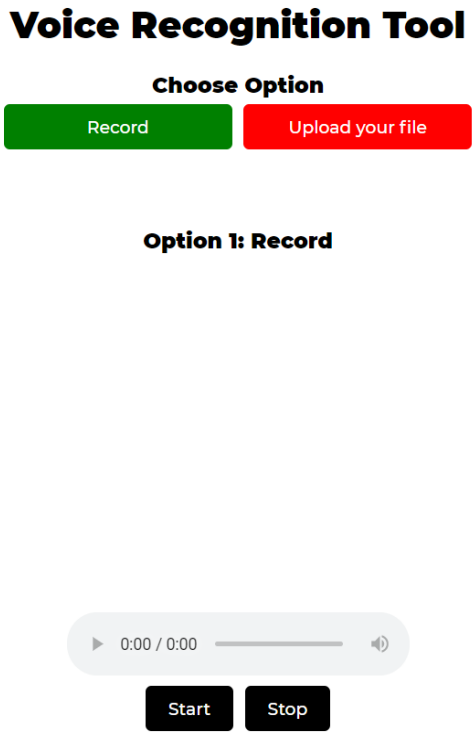
Người dùng	**Ứng dụng**
1. Bấm vào nút “Upload your file” ở giao diện ứng dụng.	
	2. Chuyển sao giao diện upload file với các option “Chọn tệp” và “Upload”.
3. Bấm vào nút “Chọn tệp” để chọn file âm thanh từ phía local của người dùng. 4. Bấm vào nút “Upload” để gửi dữ liệu đến ứng dụng	
	5. Nhận diện từ trong file upload và tìm kiếm hình ảnh theo kết quả nhận diện và hiển thị trên giao diện ứng dụng.
6. Người dùng có thể xem thêm hình ảnh bằng cách nhấn nút “Load more” ở cuối danh sách hình ảnh.	

  - Luồng thay thế: không
  - Yêu cầu: không

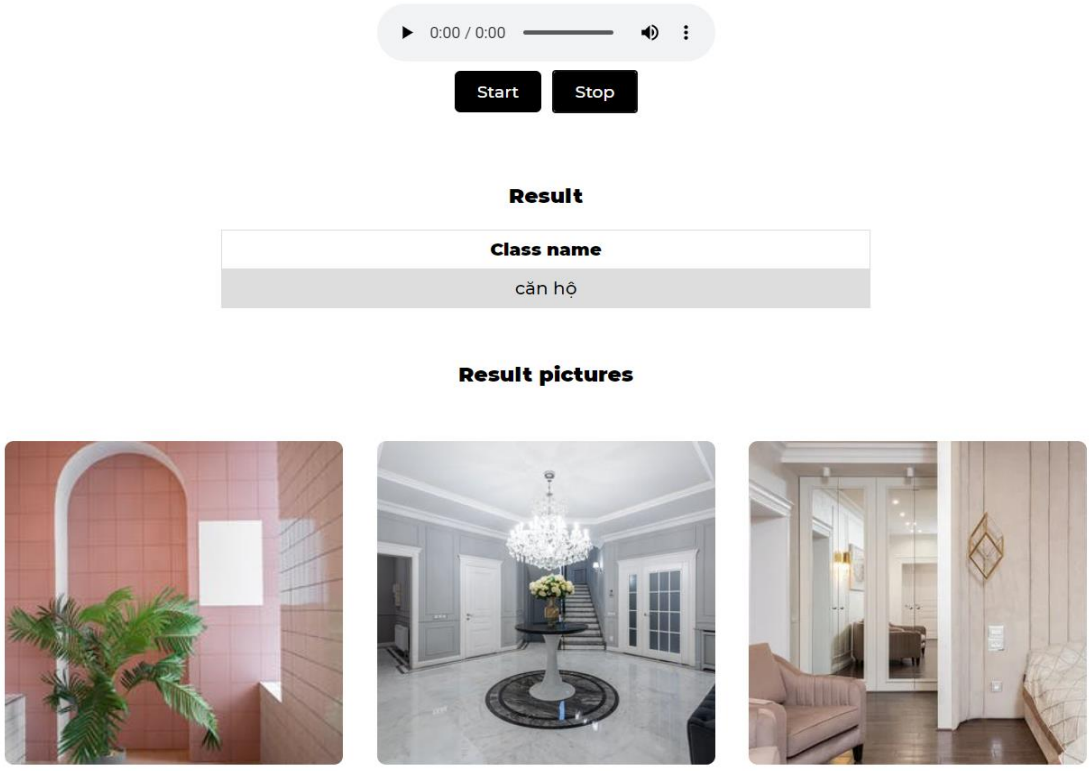
4. Giao diện ứng dụng



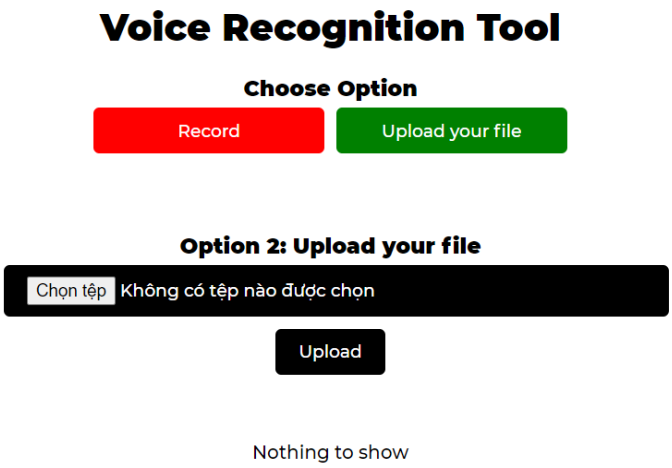
Hình 4.1: Giao diện chính.



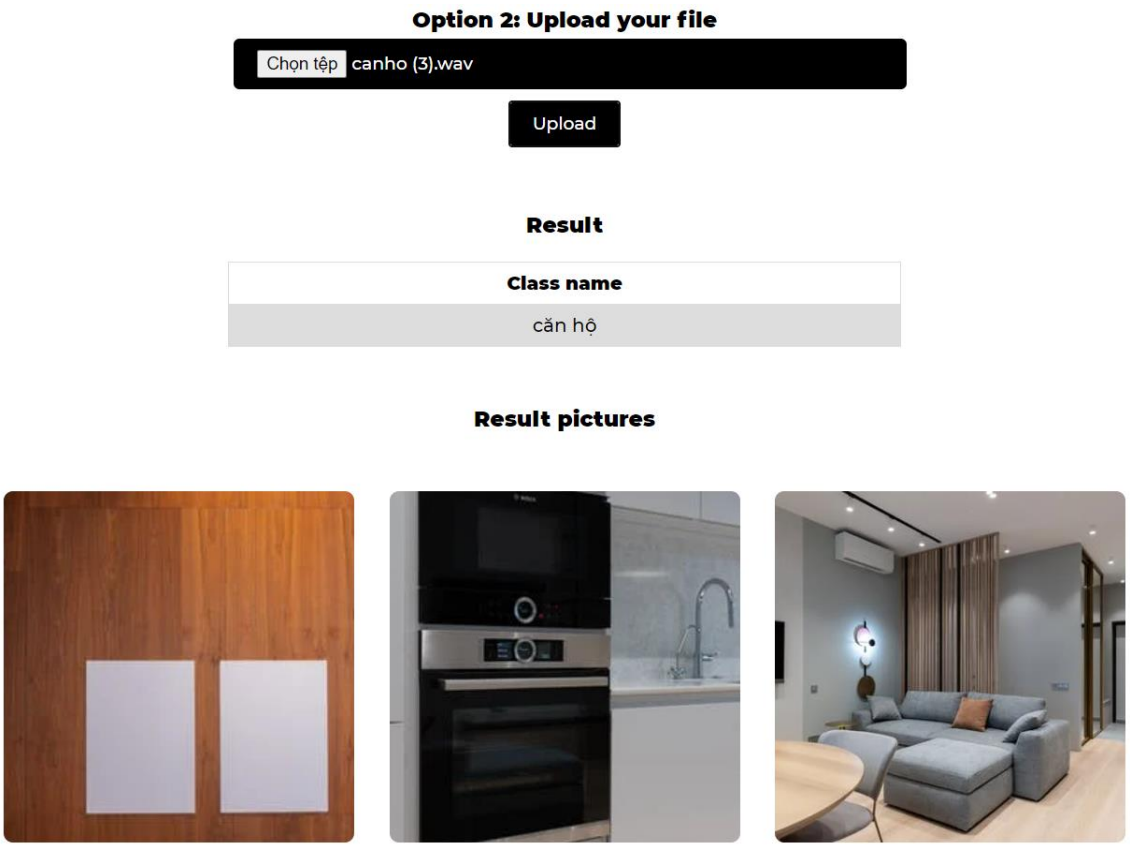
Hình 4.2: Giao diện record.



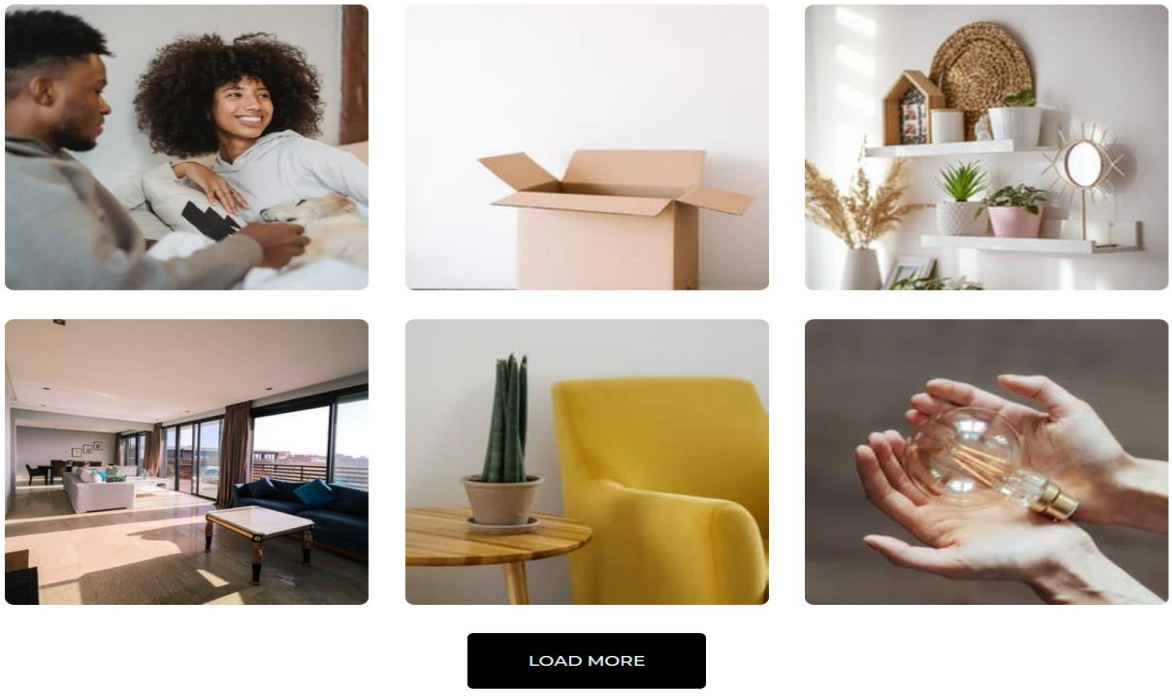
Hình 4.3: Kết quả tìm kiếm bằng ghi âm.



Hình 4.4: Giao diện upload file.



Hình 4.5: Kết quả tìm kiếm bằng file.



Hình 4.6: Lựa chọn load more để xem thêm hình ảnh.