GENDER RECOGNITION BY VOICE

Sinh viên thực hiện:
Nguyễn Duy Vũ 18120264
Nguyễn Chiêu Bản 18120283

NỘI DUNG

1. Thu thập dữ liệu
2. Đưa ra câu hỏi cần trả lời
3. Khám phá dữ liệu
4. Tiền xử lý dữ liệu
5. Mô hình hóa dữ liệu
6. Nhìn lại quá trình làm đồ án

1.Thu thập dữ liệu

- Thu thập dữ liệu bằng cách parse HTML Voxforge để thu thập file chứa dữ liệu về các bản ghi âm giọng nói
- Sau đó giải nén cái các file bằng thư viện tarfile
- Tiếp đó rút trích các đặc trưng của trong giọng nói bằng thư viện scipy

1.Thu thập dữ liệu

Data –Set thu được

- Dữ liệu thu được: 5209 mẫu
- Cột dữ liệu (13 cột): nobs, mean, skewness, kurtosis, median, mode, std, low, peak, q25, q75, iqr, lable

Ý nghĩa của các cột

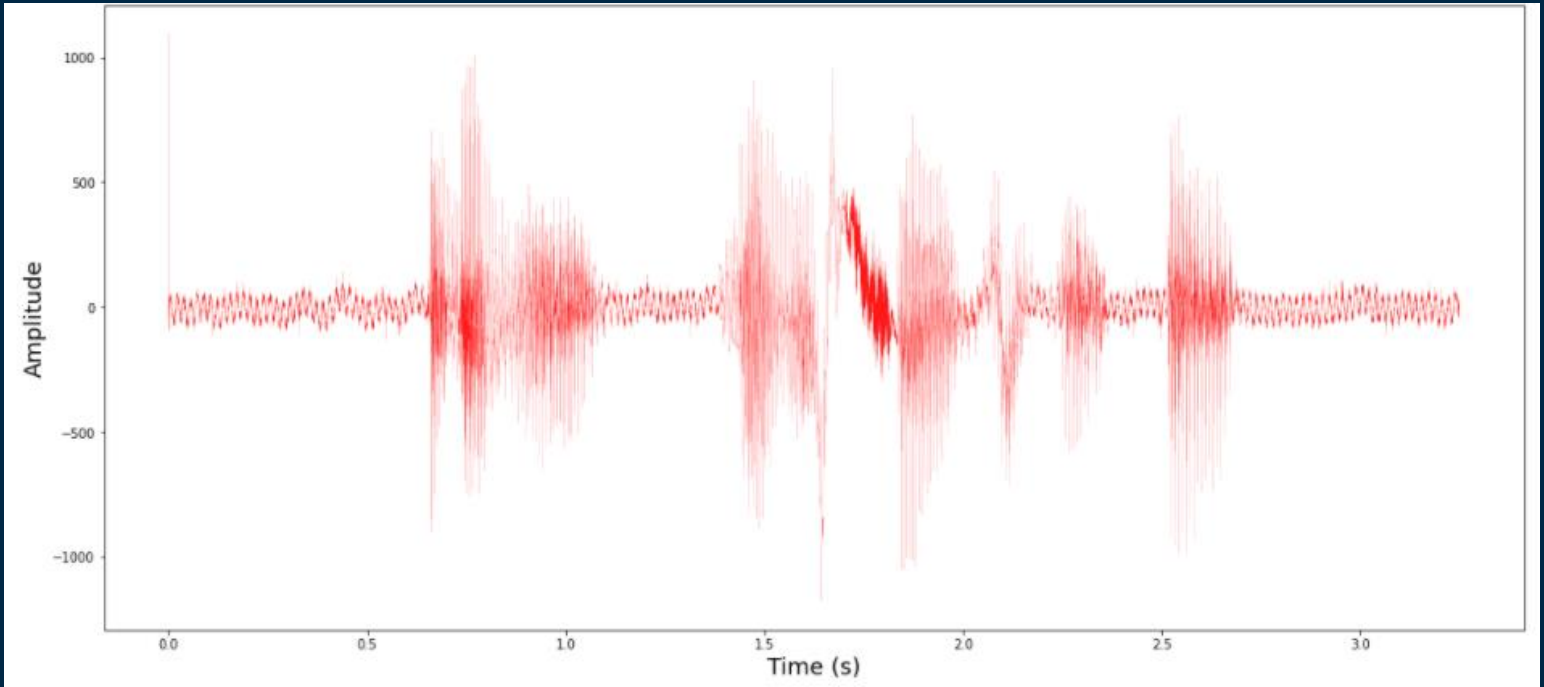
	Columns	Acoustic properties
0	nobs	Number of observations
1	mean	average of frequency measured across acoustic signal
2	skewness	skewness
3	kurtosis	kurtosis
4	median	median frequency(in Hz)
5	mode	mode frequency
6	std	standard deviation of frequency
7	low	frequency with lowest energy
8	peak	frequency with highest energy
9	q25	first quantile(in Hz)
10	q75	third quantile(in Hz)
11	iqr	interquantile range(in Hz)
12	lable	Male or Female

2. Đưa ra câu hỏi cần trả lời

- Câu hỏi sẽ có dạng input là các dữ liệu thu được từ giọng nói và output là giới tính của giọng nói đó
- Trả lời câu hỏi sẽ giúp ích cho việc điều tra tội phạm, các ngành dịch vụ mang tính chất phân lớp khách hàng nam nữ. Ngoài ra nó còn là một phần của Voice AI
- Nguồn cảm hứng: từ việc xác định giới tính bằng hình ảnh

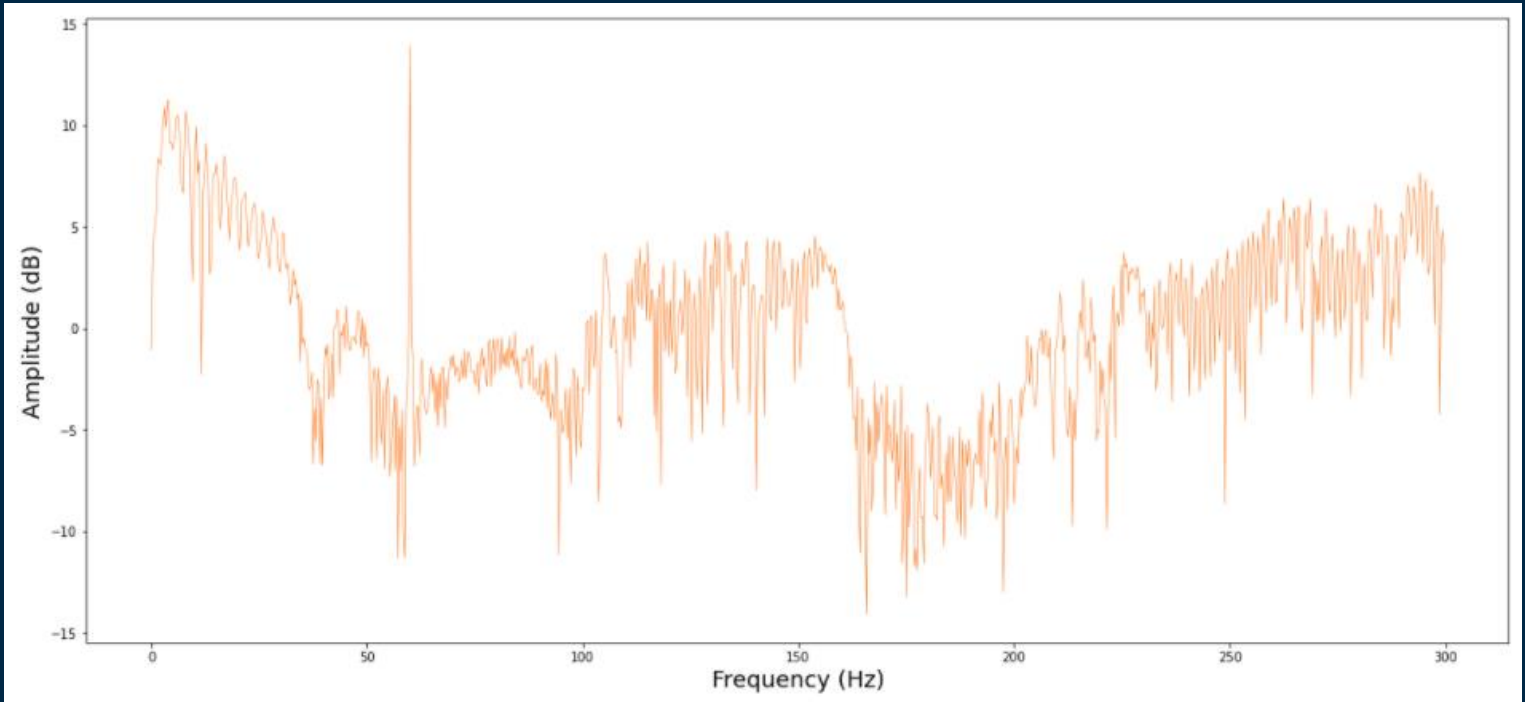
3.Khám phá dữ liệu

Biên độ âm theo thời gian



3.Khám phá dữ liệu

Biên độ âm theo tần số



3.Khám phá dữ liệu

Dữ liệu có giá trị thiếu hay không

```
1  data_df.isna().sum()
```

```
nobs        0
mean        0
skewness    0
kurtosis    0
median      0
mode        0
std         0
low         0
peak        0
q25         0
q75         0
iqr         0
lable      17
dtype: int64
```

Dữ liệu ở cột lable có giá trị "Unknow" mang ý nghĩa như giá trị thiếu.

```
1  data_df['lable'].value_counts()
```

```
Male       4764
Female      379
Unknow       49
Name: lable, dtype: int64
```

3.Khám phá dữ liệu

Kiểu dữ liệu của mỗi cột

```
nobs        float64
mean        float64
skewness    float64
kurtosis    float64
median      float64
mode        float64
std         float64
low         float64
peak        float64
q25         float64
q75         float64
iqr         float64
lable       object
dtype: object
```

Tỉ lệ giá trị nam/nữ trong cột output

```
Male      92.63076
Female    7.36924
Name: lable, dtype: float64
```

4.Tiền xử lý

Tách các tập

- Tỉ lệ tập train_val và tập test là 80%:20%
- Từ tập train_val ở trên tách thành 2 tập test và validation theo tỉ lệ 80%:20%

4.Tiền xử lý

Tiền xử lý tập input

- Ở bước khám phá dữ liệu ta đã thấy dữ liệu input của chúng ta đều là dạng số và các giá trị ở cột input đều không bị thiếu giá trị nên ở bước này ta chỉ cần tiến hành chuẩn hóa input để các thuật toán cực tiểu hóa hội tụ nhanh hơn

4.Tiền xử lý

Tiền xử lý tập output

- Ở trên khi "Khám phá dữ liệu" ta đã thấy kiểu dữ liệu output là object nên ta cần chuyển về dữ liệu dạng số
- Đồng thời do dữ liệu bị lệch khá nghiêm trọng nên ta cũng cần xử lý trường hợp này bằng cách upsamples và downsamples

5.Mô hình hóa

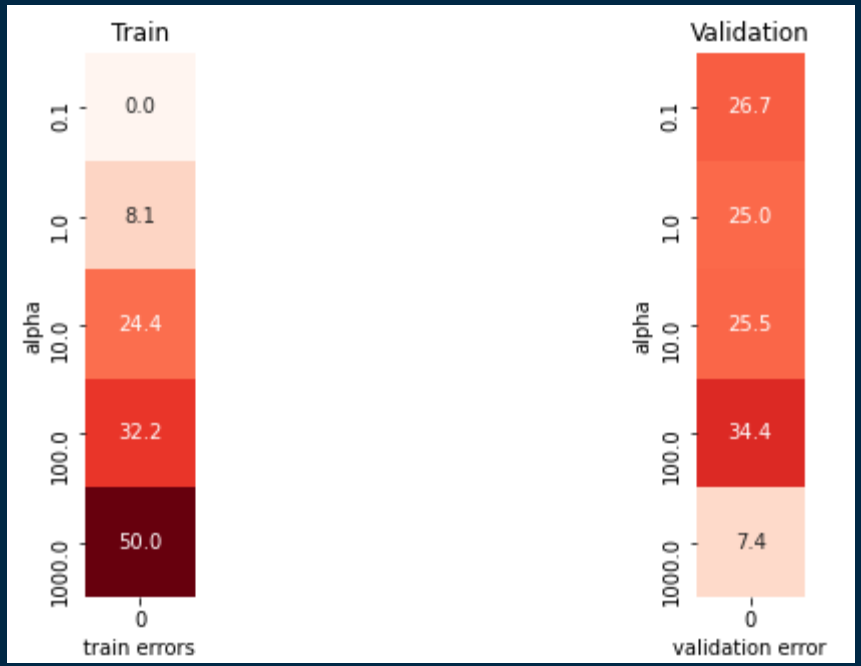
Tìm mô hình tốt nhất

- Sử dụng mô hình Neural Net để phân lớp. Thử nghiệm với các giá trị alpha với 5 giá trị khác nhau: 0.1, 1, 10, 100, 1000 để tìm alpha tốt nhất
- Đồng thời ta còn thử nghiệm mô hình với các bước tiền xử lý khác nhau ở cột output là umsamples và downsamples để tìm ra phương pháp nào của tiền xử lý cột output mang lại độ lỗi nhỏ nhất

5.Mô hình hóa

Umsamples

Downsamples



5.Mô hình hòa

- Ta có thể thấy “Downsamples” mang lại độ lỗi nhỏ nhất là 7.4 với alpha = 1000 nhưng kết quả dự đoán lại mang toàn bộ giá trị là nam điều này không hợp lý vì vậy ta chọn “Upsamples” làm phương pháp tiền xử lý cho output

5.Mô hình hóa

Đánh giá mô hình

- Ta tiền xử lý output tập train_val với phương pháp upsamples và huấn luyện lại mô hình với alpha tốt nhất bằng tập train_val
- So sánh với tập test ta thu được độ lỗi là: 13.6%

6.Nhìn lại quá trình làm đồ án

Những khó khăn đã gặp

- Khó khăn trong việc chọn đề tài
- Dữ liệu thu thập được bị lệch nghiêm trọng
- Khó khăn trong việc rút trích dữ liệu

6.Nhìn lại quá trình làm đồ án

Những điều học được thông qua đồ án

- Cần có sự phân bố thời gian hợp lý hơn
- Hiểu biết về việc xử lý dữ liệu audio
- Học được cách trình bày, làm việc nhóm một cách hiệu quả hơn
- Hiểu rõ hơn về quy trình của một dự án khoa học dữ liệu
- Học thêm được cách xử lý dữ liệu đối với trường hợp dữ liệu bị lệch
- Học được cách đánh giá mô hình hợp lý (không vì con số độ lỗi thấp mà lựa chọn nó mà phải để ý thêm các giá trị dự đoán)
- Đồng thời ý thức thêm về việc đóng góp tham gia các cuộc khảo sát để đóng góp cho cộng động

6.Nhìn lại quá trình làm đồ án

Nếu nhóm có thêm thời gian

- Tìm nguồn dữ liệu phù hợp hơn (ít bị lệch hơn)
- Thử thêm nhiều mô hình machine learning để cải thiện độ lỗi
- Viết một api đơn gian để có thể test trực tiệp mô hình trong đồ án này

Quá trình thực hiện

Thông tin nhóm

Nguyễn Chiêu Bản

Email:
18120283@student.hcmus.edu.vn

Nguyễn Duy Vũ

Email:
18120264@student.hcmus.edu.vn

CẢM ƠN THẦY ĐÃ LẮNG NGHE

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon, and infographics & images by Freepik
Please keep this slide for attribution