Sử dụng mô hình Hidden Markov Model để dự đoán phân loại 1 trong 5 từ tiếng việt.

Bùi Tiến Đạt – 17021230

Trần Quang Huy - 17021268

1. Dữ liệu

- Bộ phân chia dữ liệu:
  - Tập train: 80%
  - Tập test: 20%
- Tập dữ liệu bao gồm:
  - "toi" : 79 tập train, 20 tập test
  - "ban" : 101 tập train, 20 tập test
  - "khong" : 80 tập train, 20 tập test
  - "khach" : 79 tập train, 20 tập test
  - "vietnam" : 80 tập train, 20 tập test
- Định dạng dữ liệu: file âm thanh định dạng .wav

2. Phương Pháp

2.1 Sử dụng mô hình Multinomial HMM:

- Tôi:

```
toi_model.startprob_ = np.array([0.5,0.2,0.1,0.1,0.1,0.0,0.0,0.0,0.0,0.0])
toi_model.transmat_ = np.array([
    [0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.5],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],
])
```

- Bạn:

```
ban_model.startprob_ = np.array([0.5,0.2,0.1,0.1,0.1,0.0,0.0,0.0,0.0,0.0])
ban_model.transmat_ = np.array([
    [0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.5],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],
])
```

- Không:

```
khong_model.startprob_ = np.array([0.5,0.2,0.1,0.1,0.1,0.0,0.0,0.0,0.0,0.0])
khong_model.transmat_ = np.array([
    [0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.5],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],
])
```

- Khách:

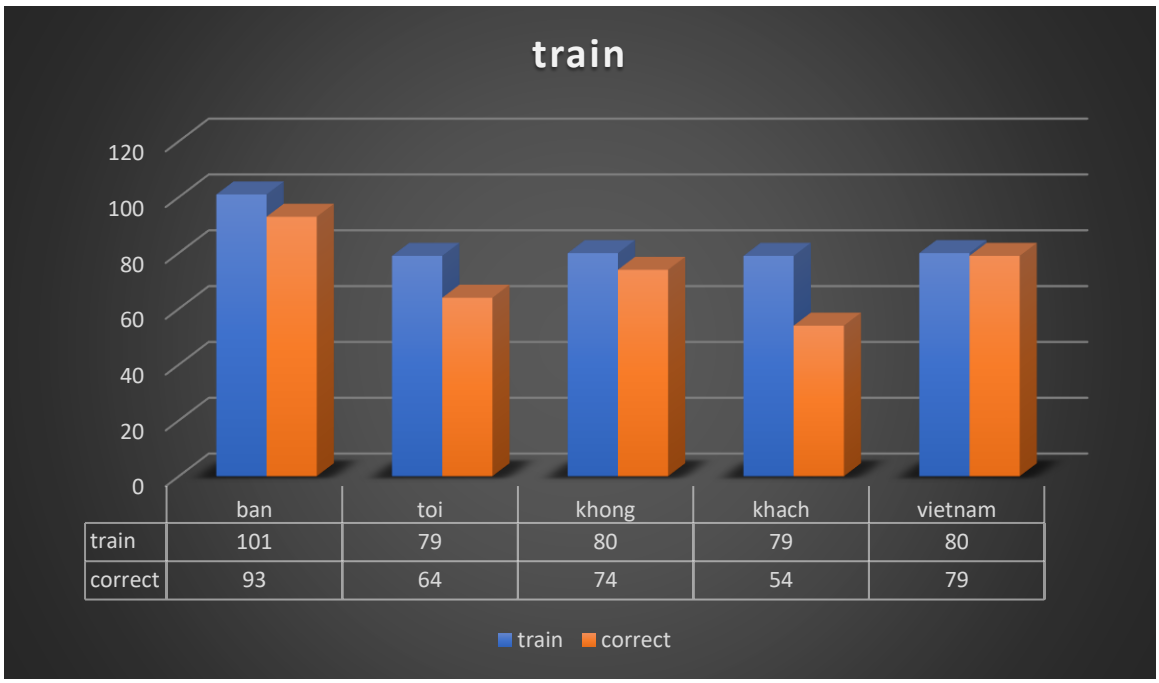
```
khach_model.startprob_ = np.array([0.5,0.2,0.1,0.1,0.1,0.0,0.0,0.0,0.0,0.0])
khach_model.transmat_ = np.array([
    [0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.5],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],
])
```

- Việt Nam:

```
vietnam_model.startprob_ = np.array([0.5,0.2,0.1,0.1,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0])
vietnam_model.transmat_ = np.array([
    [0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.2,0.1,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7,0.3,0.0],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.5],
    [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0],
])
```

2.2 Kết Quả:

- Tập train



	ban	toi	khong	khach	vietnam
train	101	79	80	79	80
correct	93	64	74	54	79

- Tập test

3. Thiết kế mô hình

3.1. Thiết kế mô hình:

Nhóm thiết kế mô hình phân loại âm thanh được nói vào 1 trong 5 từ: tôi, bạn, không, khách, việt nam, sử dụng Gaussian HMM cho mô hình acoustic, thống kê tần suất từ cho mô hình ngôn ngữ

- Tách dữ liệu
    - Dữ liệu là các file âm thanh khi nói 5 từ: tôi, bạn, không, khách, việt nam, được tách bằng phần mềm Audacity từ dữ liệu ghi âm khi đọc các bài báo của cả lớp.
- Tăng số lượng dữ liệu
    - Lượng dữ liệu có được khi tách từ dữ liệu ghi âm còn ít và chưa đủ tính đa dạng. Nhóm đã dùng phương pháp SpecAugment để thêm độ đa dạng cho dữ liệu, tăng gấp đôi số lượng file. Do phương pháp SpecAugment thực hiện trên melspectrogram chứ không phải trên dạng sóng của âm thanh, nên có ý nghĩa cho việc nhận dạng
- Trích xuất đặc trưng
    - Từ dạng sóng của mỗi file âm thanh, ta chia thành các frame và biến đổi thành các vector mfcc với 12 thành phần. Ta áp dụng phương pháp chuẩn hóa mean và variance cho cepstral để giảm ảnh hưởng của nhiễu do kênh truyền (multiplicative noise), và nhiễu do môi trường (additive noise).

    - Với mỗi frame, ta thêm giá trị năng lượng của frame đó, được vector 13 thành phần. Sau đó, áp dụng phương pháp delta và delta-delta để được vector 39 thành phần.
- Kiến trúc mô hình
    - Nhóm sử dụng phương pháp Gaussian HMM, dùng gói hmmlearn để triển khai mô hình.
    - Từ việc phân tích âm vị của 5 từ, nhóm thấy các từ tôi, bạn, không, khách, việt nam, có lần lượt 3, 3, 3, 6 âm vị chính. Mỗi âm vị có thể mô hình hóa bởi 3 thành phần, do đó nhóm sử dụng các mô hình HMM với 9, 9, 9, 15 thành phần cho từng từ.
    - Mô hình HMM được sử dụng cho mô hình acoustic là left-to-right HMM, ma trận chuyển 1 trạng thái khởi tạo có các phần tử trên đường chéo chính và ngay trên đường chéo chính bằng 0.5 để kì vọng sẽ gần với giá trị tối ưu (hmmlearn không hỗ trợ cụ thể mô hình left- to-right).
    - Mô hình HMM do hmmlearn cung cấp trả về log likelihood của trạng thái quan sát được với mô hình HMM, trong trường hợp này là log likelihood của đặc trưng file âm thanh ứng với mô hình đó
- Mô hình ngôn ngữ
    - Mô hình ngôn ngữ ở đây đơn giản là thống kê số lần xuất hiện của các từ trong toàn bộ dữ liệu ghi âm.
    - Với đặc trưng dữ liệu của mỗi file âm thanh, ta tính được log likelihood của đặc trưng với 5 mô hình HMM cộng với log của số lần xuất hiện từ ứng với mô hình HMM đó, file âm thanh sẽ được dự đoán vào từ có giá trị này lớn nhất.

3.2. Thí nghiệm:

- Do dữ liệu huấn luyện được ghi âm khi đọc các bài báo, nên các từ được đọc nhanh, số frame cho mỗi từ ít, còn với âm thanh khi thu âm trực tiếp khi đọc 1 từ lại dài, có nhiều frame thừa, đọc chậm hơn. Do đó, khi kiểm tra với dữ liệu thu âm trực tiếp, nhóm đã bỏ đi 1 nửa đầu số frame, và tăng tần số để được dữ liệu âm thanh nhanh hơn
    - Thử nghiệm cho thấy với 3 từ ’bạn’, ’khong’, ’việt nam’, mô hình dự đoán chính xác. Tuy nhiên khi phát âm từ ’tôi’, mô hình không đoán được và thường nhầm sang từ ’khach’. Điều này có thể do đặc trưng khi phát âm 2 từ này giống nhau, hoặc ảnh hưởng của việc tăng tốc độ âm thanh khi kiểm tra trực tiếp.