

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

Nguyễn Hùng Quang

Lê Mai An

Nguyễn Chiến Thắng

May 23, 2020

## 1 Thiết kế mô hình

Nhóm thiết kế mô hình phân loại âm thanh được nói vào 1 trong 5 từ: ngày, nhiều, những, gia đình, dịch bệnh, sử dụng Hidden Markov Model cho mô hình acoustic, thống kê tần suất từ cho mô hình ngôn ngữ

### 1.1 Tách dữ liệu

Dữ liệu là các file âm thanh khi nói 5 từ: ngày, nhiều, những, gia đình, dịch bệnh, được tách bằng phần mềm Audacity từ dữ liệu ghi âm khi đọc các bài báo của cả lớp.

### 1.2 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 [2] 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

### 1.3 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) [1].

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.

### 1.4 Kiến trúc mô hình

Nhóm sử dụng phương pháp Gaussian Mixture model HMM, dùng gói hmmlearn để triển khai mô hình.

Từ việc phân tích âm vị của 5 từ, nhóm thấy các từ ngày, nhiều, những, gia đình, dịch bệnh, có lần lượt 3, 3, 3, 5, 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, 18 thành phần cho từng từ.

Mô hình HMM được sử dụng cho mô hình acoustic là left-to-right HMM, ma trận chuyển

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

## 1.5 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.

## 2 Thí nghiệm

Mô hình được huấn luyện trên file âm thanh của 193 từ 'ngày', 237 từ 'nhiều', 342 từ 'những', 85 từ 'gia đình', 69 từ 'dịch bệnh', mỗi file âm thanh này được biến đổi với SpecAugment, ta được số lượng dữ liệu huấn luyện gấp đôi dữ liệu ban đầu.

### 2.1 Kiểm tra với dữ liệu âm thanh ghi âm

Mô hình được kiểm tra trên tập dữ liệu gồm 100 file âm thanh của 5 từ, mỗi từ 20 file, các file này chưa được huấn luyện với mô hình.

Kết quả cho thấy mô hình dự đoán đúng 93 file trên tổng số 100 file. Có 2 file 'ngày', 3 file 'những', 2 file 'gia đình' bị dự đoán sai.

### 2.2 Kiểm tra với dữ liệu thu âm trực tiếp

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 4 từ 'ngày', 'nhiều', 'gia đình', 'dịch bệnh', mô hình dự đoán chính xác. Tuy nhiên khi phát âm từ 'những', mô hình không đoán được và thường nhầm sang từ 'nhiều'. 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.

## References

- [1] Xuedong Huang et al. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. 1st. USA: Prentice Hall PTR, 2001. ISBN: 0130226165.
- [2] Daniel S. Park et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." In: *Proc. Interspeech 2019*. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680. URL: <http://dx.doi.org/10.21437/Interspeech.2019-2680>.