Báo cáo kỹ thuật
Module nhận dạng tiếng nói tiếng Việt trong underthesea

Vũ Anh
underthesea
anhv.ict91@gmail.com

Lê Phi Hùng
underthesea
lephihungch@gmail.com

Abstract

Trong báo cáo này, trong chúng mô tả hệ thống nhận dạng tiếng nói tiếng Việt trong underthesea. Trong đó, hệ thống sử dụng công cụ Kaldi để xây dựng module nhận dạng, kết quả được đánh giá trên tập dữ liệu test của VLSP 2018. Toàn bộ mã nguồn và tài liệu của dự án được phát hiện dưới dạng mở nguồn mở tại địa chỉ https://github.com/undertheseanlp/automatic_speech_recognition

1 Giới thiệu

2 Mô tả hệ thống

Các thử nghiệm được thực hiện trên bộ công cụ nhận dạng tiếng nói được viết trên C++ Kaldi. [1]

Mô hình xây dựng hệ thống nhận dạng tiếng nói

2.1 Chuẩn bị dữ liệu và các tài nguyên ngôn ngữ

Việc đầu tiên cần làm là chuẩn bị dữ liệu huấn luyện âm thanh - phụ đề. Gồm có các tập tin âm thanh (thường để ở định dạng wav) chứa các tiếng nói của người và các tập tin phụ đề tương ứng.

Việc tiếp theo là xây dựng từ điển phát âm. Hình dung một cách đơn giản, từ điển phát âm sẽ chứa cách phát âm (cách phân chia các âm) tương ứng với từng tiếng. Ngoài ra trong hệ thống còn cần các âm câm (silence_phones), các từ ngoài từ điển (out-of-vocabulary hay oov).

Cuối cùng là chuẩn bị dữ liệu cho việc huấn luyện mô hình ngôn ngữ. Mô hình ngôn ngữ giúp cải thiện chất lượng của hệ thống nhận dạng tiếng nói, bằng cách đưa ra những khả năng có thể nhất trong một cụm từ. Hãy xem xét ví dụ hệ thống đang phải quyết định từ con thiếu trong câu *Tôi đi Hà _ mấy ngày*. Nếu hệ thống sử dụng mô hình ngôn ngữ, có thể dễ dàng nhận ra từ *Nội* là từ có khả năng còn thiếu nhất trong câu này.

2.2 Huấn luyện mô hình Gaussian Mixture Model

Bước đầu tiên là huấn luyện mô hình âm học, là thành phần chuyển các tín hiệu âm thanh thành dữ liệu văn bản. Mô hình huấn luyện thường sử dụng thuật toán Gaussian Mixture Model trên các tập đặc trưng phổ biến của âm thanh như MFCC (Mel-frequency cepstral coefficients) [2]. Ngoài ra còn có các đặc trưng delta, lda, mltt hay sat.

Bước thứ hai là huấn luyện mô hình ngôn ngữ

2.3 Quá trình giải mã

- Tạo ra một đồ thị giải mã
- Tính điểm lại Lattice

3 Đánh giá

3.1 Tập dữ liệu

Có hai tập dữ liệu được sử dụng. Tập dữ liệu VIVOS và tập dữ liệu VLSP 2018. Trong đó, tập dữ liệu VIVOS được dùng để huấn luyện, tập dữ liệu VLSP 2018 được sử dụng để đánh giá kết quả mô hình.

3.2 Kết quả

4 Conclusion

5 Lời cảm ơn

Vì kiến thức còn hạn chế, trong phần mô tả kỹ thuật, tác giả có tham khảo các tài liệu *Building*

[1] http://kaldi-asr.org/

[2] Để biết thêm về đặc trưng này, xin tìm đọc tài liệu So sánh hai phương pháp trích chọn đặc trưng âm thanh: Đường bao phổ (MFCC) và cao độ Pitch trong việc tìm kiếm âm nhạc theo nội dung

References

[3] https://engineering.jhu.edu/clsp/wp-content/uploads/sites/75/2016/06/Building-Speech-Recognition-Systems-with-the-Kaldi-Toolkit.pdf