CÁC ĐẶC TRƯNG ÂM THANH SỬ DỤNG TRONG MÔ HÌNH NHẬN DẠNG GIỌNG NÓI

Nguyễn Huy Thế, Nguyễn Tuấn Anh
Trường Đại học Thủy lợi, email: nguyenhuythe@tlu.edu.vn

1. GIỚI THIỆU CHUNG

Nhận dạng giọng nói ngày càng được áp dụng rộng rãi, đặc biệt là trong các lĩnh vực tương tác người - máy bởi sự đa dạng và linh hoạt trong ngôn ngữ giao tiếp. Các phương pháp nhận dạng giọng nói phổ biến dựa trên việc trích xuất thông tin đặc trưng từ giọng nói và sử dụng để huấn luyện các mô hình nhận dạng. Trích xuất các đặc trưng âm thanh là bước rất quan trọng, quyết định độ chính xác và hiệu quả của mô hình nhận dạng, cần được thực hiện đảm bảo yêu cầu hạn chế tối đa hoặc không mất mát thông tin.

Hiện nay, có rất nhiều kỹ thuật trích xuất đặc trưng giọng nói đã được phát triển. Nghiên cứu này tập trung vào một số kỹ thuật được sử dụng phổ biến nhất bao gồm Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC). Các dữ liệu đặc trưng này được sử dụng để xây dựng và huấn luyện mô hình học máy nhận dạng sự có mặt của các từ khóa trong giọng nói thu âm được. Việc tính toán các bộ dữ liệu và huấn luyện mô hình nhận dạng được thực hiện với ngôn ngữ Python.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Quy trình xây dựng và huấn luyện mô hình nhận dạng từ khóa trong giọng nói bao gồm ba bước: thu thập dữ liệu âm thanh; trích xuất đặc trưng; huấn luyện và kiểm tra mô hình.

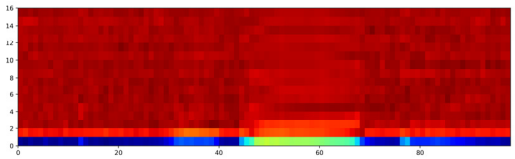
2.1. Thu thập dữ liệu âm thanh

Dữ liệu âm thanh được sử dụng trong nghiên cứu này là tập dữ liệu sẵn có Google Speech Command datasets [1]. Tập dữ liệu này bao gồm hơn 105.000 tệp thu âm ở định dạng .wav của hơn 30 từ tiếng Anh khác nhau với thời lượng khoảng 1s. Để đơn giản quá trình tính toán, nghiên cứu này chỉ sử dụng các file âm thanh tương ứng với tám từ khóa ‘yes’, ‘up’, ‘down’, ‘left’, ‘right’, ‘stop’, ‘go’, ‘off’.

2.2. Trích xuất đặc trưng âm thanh

2.2.1. Kỹ thuật MFCC

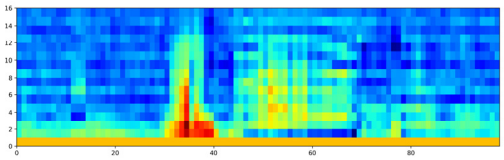
MFCC là một kỹ thuật phổ biến hàng đầu trong việc xử lý và nhận dạng giọng nói. Quá trình tính toán đặc trưng MFCC dựa trên thang đo Mel có nguyên lý tương tự như cách cảm nhận âm thanh của tai người. Các bộ lọc tần số được đặt cách đều nhau tại các tần số thấp và được bố trí theo thang logarit tại các tần số cao, từ đó thu được các đặc tính quan trọng về mặt ngữ âm của tín hiệu giọng nói. Bước đầu tiên của quá trình tính toán là chia nhỏ tệp tín hiệu âm thanh thu được thành các khung dữ liệu. Sau đó là quá trình kích hoạt các mức tần số cao để tránh làm mất mát thông tin. Phép biến đổi Fast Fourier Transform (FFT) được áp dụng cho các khung dữ liệu này để tìm phổ công suất và được đưa qua thang đo Mel. Cuối cùng, qua phép biến đổi Discrete Cosine Transform (DCT) thu được các hệ số MFCC [2]. Lặp lại các bước tính toán trên cho các khung dữ liệu tiếp theo và liên kết các kết quả tương ứng sẽ nhận được đặc trưng MFCC của tín hiệu âm thanh ban đầu là một bộ dữ liệu hai chiều, minh họa trong hình 1.



Hình 1. *Đặc trưng MFCC của từ ‘stop’*

2.2.2. Kỹ thuật LPC

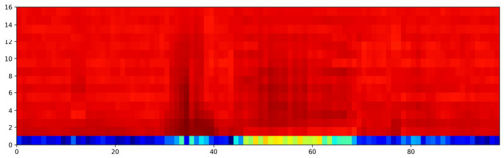
Kỹ thuật LPC dựa trên giả thiết tín hiệu âm thanh hiện tại được dự đoán thông qua một tổ hợp tuyến tính của các mẫu âm thanh trước đó. Khi đó, mô hình tự hồi quy (autoregressive) được sử dụng để ước tính các hệ số dự đoán tuyến tính LPC đặc trưng cho hình thái của tín hiệu âm thanh. Bước đầu tiên của quá trình tính toán cũng là chia tệp tín hiệu âm thanh thành các khung dữ liệu. Hàm cửa sổ (ví dụ cửa sổ Hamming) thường được áp dụng để giảm thiểu sự không liên tục của tín hiệu. Tiếp theo, mỗi khung dữ liệu này sẽ được tự tương quan. Các khung tự tương quan được biến đổi thành các hệ số LPC nhờ phương pháp Durbins [3]. Lặp lại các bước tính toán trên cho các khung dữ liệu tiếp theo, thu được bộ dữ liệu hai chiều, minh họa trong hình 2.



Hình 2. *Đặc trưng LPC của từ 'stop'*

2.2.3. Kỹ thuật LPCC

Kỹ thuật LPCC kết hợp hai phương pháp dự đoán tuyến tính và phân tích cepstral. Trước hết, các hệ số dự đoán tuyến tính mô tả đặc điểm hình thái của tín hiệu giọng nói được xác định. Thông qua biến đổi cepstral trích xuất các hệ số đặc trưng trong phổ của tín hiệu âm thanh. Các bước đầu tiên trong quá trình tính toán đặc trưng LPCC tương tự như khi tính toán đặc trưng LPC. Sau khi tính được các hệ số LPC, thực hiện biến đổi cepstral để chuyển dữ liệu từ miền tần số sang miền cepstral. Các hệ số thu được chính là các đặc trưng LPCC của tín hiệu âm thanh. Mỗi tệp tín hiệu âm thanh tương ứng với bộ dữ liệu hai chiều, minh họa trong hình 3.



Hình 3. *Đặc trưng LPCC của từ 'stop'*

Các đặc trưng âm thanh MFCC, LPC, LPCC trong nghiên cứu này được tính toán bởi thư viện SPAFE trong Python [4].

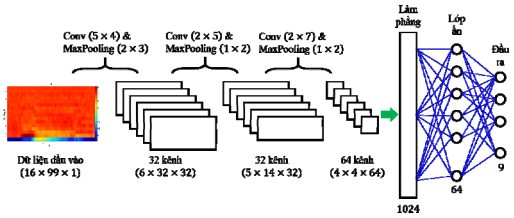
2.3. Mô hình nhận dạng giọng nói

Kết quả tính toán đặc trưng âm thanh sử dụng các kỹ thuật nêu trên có đặc điểm chung là một bộ dữ liệu hai chiều. Do đó, các bộ dữ liệu tương ứng với các tệp âm thanh có thể được nhận dạng thông qua mô hình mạng nơ-ron tích chập (CNN). Cấu trúc của mô hình CNN gồm hai lớp chính: lớp trích xuất thông tin và lớp phân loại. Lớp trích xuất thông tin sử dụng để tính toán các đặc tính của dữ liệu đầu vào thông qua phép tích chập từng phần của dữ liệu với một bộ lọc. Dữ liệu sau khi đi qua lớp tích chập có thể được dàn phẳng để đưa vào lớp phân loại. Về bản chất, lớp phân loại là một mạng nơron suy luận tiến. Quá trình huấn luyện được áp dụng phương pháp lan truyền ngược.

Mô hình CNN trong nghiên cứu này được xây dựng nhờ sử dụng thư viện Tensorflow. Đây là thư viện phổ biến hỗ trợ quá trình tính toán, xây dựng và huấn luyện mô hình học máy.

3. KẾT QUẢ NGHIÊN CỨU

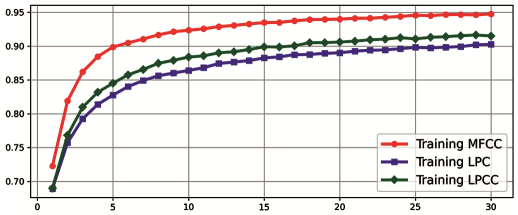
Các mô hình nhận dạng từ khóa trong giọng nói sử dụng các đặc trưng âm thanh MFCC, LPC và LPCC đều có chung một cấu trúc được thể hiện trên hình 4.



Hình 4. *Cấu trúc mô hình CNN*

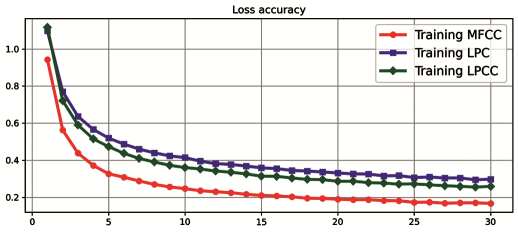
Dữ liệu đầu vào của mô hình là đặc trưng âm thanh có kích thước là $16 \times 99 \times 1$. Qua lớp trích xuất thông tin, bao gồm ba lớp tích chập kết hợp đồng thời pooling, kích thước của dữ liệu được giảm xuống $4 \times 4 \times 64$. Sau đó, dữ liệu được làm phẳng và đưa vào lớp phân loại có 1024 nút.

Tiến hành huấn luyện mô hình sử dụng dữ liệu đầu vào là các bộ dữ liệu đặc trưng MFCC, LPC và LPCC. Kết quả huấn luyện mô hình được thể hiện trong các hình 5 và 6.



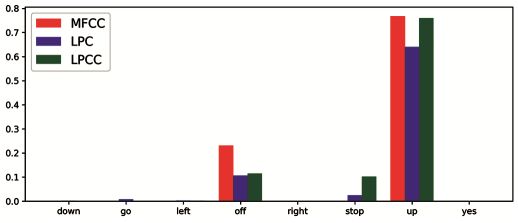
Hình 5. *Độ chính xác của mô hình*

Hình 5 mô tả độ chính xác khi huấn luyện với các bộ dữ liệu đặc trưng khác nhau. Độ chính xác của mô hình đều đạt rất tốt (trên 90%).



Hình 6. *Sai lệch của mô hình*

Hình 6 thể hiện sự thay đổi của sai lệch khi huấn luyện mô hình đối với các tập dữ liệu đặc trưng nêu trên, giá trị sai lệch giảm dần theo từng chu kỳ huấn luyện.



Hình 7. *Kết quả nhận dạng từ 'up'*

Kết quả áp dụng các mô hình đã được huấn luyện với các bộ dữ liệu đặc trưng đầu vào tương ứng để nhận dạng đối với tệp thu âm từ 'up' được biểu diễn trong hình 7. Các mô hình đều có khả năng nhận dạng tốt từ khóa cho trước, ngay cả khi từ 'up' và từ 'off' được phát âm khá giống nhau. Đối với các từ khóa khác có phát âm rõ rệt là 'yes', 'down', 'left', 'right', 'stop', 'go', kết quả nhận dạng đều đạt độ chính xác trên 90%.

Có thể thấy, chất lượng của mô hình sử dụng đặc trưng MFCC là tốt nhất, trong khi các mô hình sử dụng hai đặc trưng LPC và LPCC có chất lượng khá tương đồng. Thực tế, kỹ thuật MFCC đã được sử dụng rộng rãi hơn trong các mô hình nhận dạng giọng nói. Tuy nhiên, các đỉnh cộng hưởng trong dải tần số trên 1 kHz không được trích xuất hiệu quả do các bộ lọc tam giác phân bố càng thưa hơn trong dải tần số càng cao. Các đặc trưng MFCC kém chính xác khi có nhiễu. Kỹ thuật LPC trích xuất hiệu quả đặc trưng giọng nói với tốc độ tính toán và độ chính xác cao. Tuy nhiên, giả thiết tín hiệu giọng nói dựa trên thang đo tuyến tính chưa hoàn toàn hợp lý. Vì vậy, kỹ thuật LPC cho độ chính xác thấp hơn so với kỹ thuật kết hợp LPCC.

4. KẾT LUẬN

Dựa trên quá trình tính toán các hệ số đặc trưng của âm thanh và kết quả huấn luyện mô hình nhận dạng cho thấy tính khả thi của việc sử dụng các đặc trưng MFCC, LPC và LPCC trong xây dựng và huấn luyện mô hình nhận dạng từ khóa trong giọng nói. Kỹ thuật trích xuất đặc trưng MFCC cho kết quả nhận dạng tốt nhất, nhưng do phải qua nhiều bước nên thời gian tính toán dài hơn. Mặc dù các kỹ thuật trích xuất đặc trưng LPC và LPCC có thời gian tính nhanh hơn nhưng đôi khi cũng xảy ra lỗi nên các dữ liệu này sẽ không thể sử dụng. Như vậy, nghiên cứu cần tiếp tục theo hướng cải tiến thuật toán tính các đặc trưng âm thanh hoặc kết hợp các thuật toán tối ưu. Nhờ đó có thể nâng cao độ chính xác của mô hình nhận dạng, đồng thời giảm độ phức tạp tính toán, đặc biệt là trong điều kiện có nhiễu.

5. TÀI LIỆU THAM KHẢO

[1] P. Warden. (2018). A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209.

[2] Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms (pp. 2-19). London, UK: IntechOpen.