HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Tất Hậu

NGHIÊN CỨU XÂY DỰNG MÔ HÌNH PHÂN LOẠI GIỚI TÍNH VÀ VÙNG MIỀN CHO TIẾNG NÓI TIẾNG VIỆT DỰA TRÊN ÂM THANH

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

TÓM TẮT LUẬN VĂN THẠC SỸ

(Theo định hướng ứng dụng)

Hà Nội - 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Nguyễn Ngọc Điệp

Phản biện 1: PGS. TS. Nguyễn Đức Dũng

Phản biện 2: PGS. TS. Hoàng Hữu Hạnh

Luận văn này được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 14 giờ ngày 28 tháng 8 năm 2021

Có thể tìm hiểu luận văn này tại:

Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Trong những năm gần đây, các bước tiến trong lĩnh vực Deep Learning (Học sâu), nhất là về Thị giác Máy tính đã giải quyết rất nhiều vấn đề từ các lĩnh vực khác nhau và đóng góp vào sự cải thiện đời sống hàng ngày của con người.

Ngày nay, để tăng tính bảo mật và xác thực, nhiều kĩ thuật liên quan đến nhận diện giọng nói, giới tính đang áp dụng trong các ngân hàng, cơ quan và tổ chức doanh nghiệp, riêng trong viễn thông vẫn đề về gian lận cước đang gây ra những tổn hại về doanh thu rất lớn cho các nhà mạng, tập đoàn viễn thông. Việc tìm bắt những thuê bao lậu cước là một thử thách rất lớn vì các hệ thống lậu được lập trình một cách rất tinh vi để hành vi của chúng trở nên vô cùng khó phân biệt với các thuê bao của người dùng thông thường. Các tập đoàn viễn thông đang thử nghiệm theo dõi các yếu tố mà các nhóm làm lậu cước khó tác động vào nhằm phát hiện ra các bất thường, một trong số đó giọng nói thu được từ các cuộc gọi. Vì các thuê bao lậu cước là các thuê bao mà từ đó nhiều người dùng gọi đi, nên giọng nói sẽ thay đổi theo từng người, sự thay đổi này có thể được thấy trong giới tính và giọng vùng miền của người nói. Như vậy, nếu như có một cách để tự động chỉ ra giới tính và giọng vùng miền của người nói thì chúng ta có thể phần nào phát hiện ra các bất thường.

Hiện nay, lĩnh vực xử lý âm thanh – mà chủ yếu là các bài toán phân loại âm thanh đã tận dụng nhiều từ các kĩ thuật mà được sử dụng nhiều trong lĩnh vực Thị giác Máy tính và xử lý hình ảnh.

Xuất phát từ thực tế và mục tiêu như trên, học viên với sự giúp đỡ của TS. Nguyễn Ngọc Điệp học viên lựa chọn thực hiện đề tài luận văn tốt nghiệp chương trình đào tạo thạc sĩ có tên “Nghiên cứu xây dựng mô hình phân loại giới tính và vùng miền cho tiếng nói tiếng Việt dựa trên âm thanh”.

2. Tổng quan vấn đề cần nghiên cứu

Hiện nay đã có rất nhiều bài toán phân loại dữ liệu âm thanh sử dụng các mô hình học sâu, từ các cuộc thi Khoa học dữ liệu của Kaggle như “TensorFlow Speech

Recoginition Challenge" (2017), "Freesound Tagging" (2018, 2019) trong đó người tham gia tận dụng việc đưa dữ liệu âm thanh về dạng biểu diễn tần số (được coi là "ảnh của âm thanh") và áp dụng các kĩ thuật về xử lý ảnh, huấn luyện mạng nơ-ron để giải quyết bài toán, trong đó dữ liệu âm thanh sẽ được chuyển sang bài toán phân loại ảnh phổ.

Một số kỹ thuật phân loại âm thanh khác đã được nghiên cứu và chứng minh tính khả thi và độ chính xác cao bằng cách sử dụng mô hình GMM (Gaussian mixture models), kết hợp với việc xây dựng tập thuộc tính âm thanh (Mel-frequency cepstrum coefficients)

Vào tháng 8/2018, ZALO Inc. cũng đã tổ chức một cuộc thi phân loại giới tính và vùng miền cho tiếng nói tiếng Việt với một bộ dữ liệu khá lớn và đa dạng. Lời giải của đội chiến thắng cũng là sử dụng phương pháp kết hợp trích xuất thuộc tính của âm thanh và ứng dụng các thuật toán Deep Learning. Vì vậy nội dung của luận văn này sẽ tập trung nghiên cứu giải pháp để xây dựng được một bộ phân loại với độ chính xác cao hơn và tốc độ xử lý nhanh hơn. Để làm được việc đó, ngoài việc tận dụng các mạng nơ-ron tiên tiến có khả năng phân loại, luận văn sẽ nghiên cứu thêm về các kĩ thuật sử dụng trong từng phần của của xử lý phân loại, ví dụ như kĩ thuật xử lý và đưa âm thanh về miền tần số hay các kiến trúc mạng nơ-ron khác nhau, v.v nhằm nâng cao hơn nữa độ chính xác. Ngoài ra, luận văn cũng sẽ sử dụng thêm bộ data giọng nói VIVOS để huấn luyện mô hình chứ không chỉ bộ data của ZALO, vì bộ VIVOS này được thu trong điều kiện lý tưởng hơn, do đó dễ dàng cho mô hình hơn trong giai đoạn đầu của việc học.

3. Mục đích nghiên cứu

- Tìm hiểu về các phương pháp xử lý dữ liệu giọng nói;
- Nghiên cứu phương pháp xây dựng mô hình học sâu phân loại giới tính và vùng miền của giọng nói;
- Rèn luyện phương pháp và khả năng nghiên cứu.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Tiếng nói Tiếng Việt được thu trong môi trường có tạp âm, đa dạng về độ tuổi, cảm xúc của người nói.

- Phạm vi nghiên cứu: Mô hình phân loại giới tính (nam – nữ) và vùng miền (bắc – trung - nam) từ giọng nói.

5. Phương pháp nghiên cứu

Dựa trên cơ sở lý thuyết của xử lý dữ liệu giọng nói và các phương pháp huấn luyện mô hình học sâu phân loại âm thanh nói chung và giới tính - vùng miền của giọng nói nói riêng.

Cấu trúc nội dung luận văn gồm 3 chương với các nội dung như sau:

Chương 1: Tổng quan về các phương pháp phân loại âm thanh

Nội dung chương 1 của luận văn sẽ trình bày tổng quan về các phương pháp phân loại âm thanh ứng dụng các mô hình học từ dữ liệu, từ phương pháp dùng mô hình học máy truyền thống đến các mô hình học sâu. Ưu nhược điểm của từng phương pháp sẽ được phân tích để lựa chọn ra phương pháp phù hợp cho bài toán đang cần giải quyết.

Chương 2: Một số phương pháp về phân loại âm thanh

Nội dung chương 2 của luận văn sẽ trình bày các cơ sở lý thuyết liên quan đến phương pháp sử dụng mô hình học sâu CNN trên dữ liệu dạng biểu diễn tần số của âm thanh. Cụ thể, phần đầu tiên của chương này sẽ tóm lược các một vài phương pháp chuyển dữ liệu âm thanh từ miền thời gian sang miền tần số để thu được "ảnh của âm thanh". Tiếp đến, một số mô hình học sâu quan trọng, thường dùng trong các bài toán phân loại hình ảnh sẽ được trình bày.

Chương 3: Thực nghiệm và đánh giá

Nội dung chương 3 của luận văn sẽ trình bày các bước triển khai tiền xử lý dữ liệu và xây dựng, huấn luyện mô hình cũng như các bước hậu xử lý, sau cùng là đánh giá độ chính xác của mô hình trên dữ liệu mới.

Kết luận.

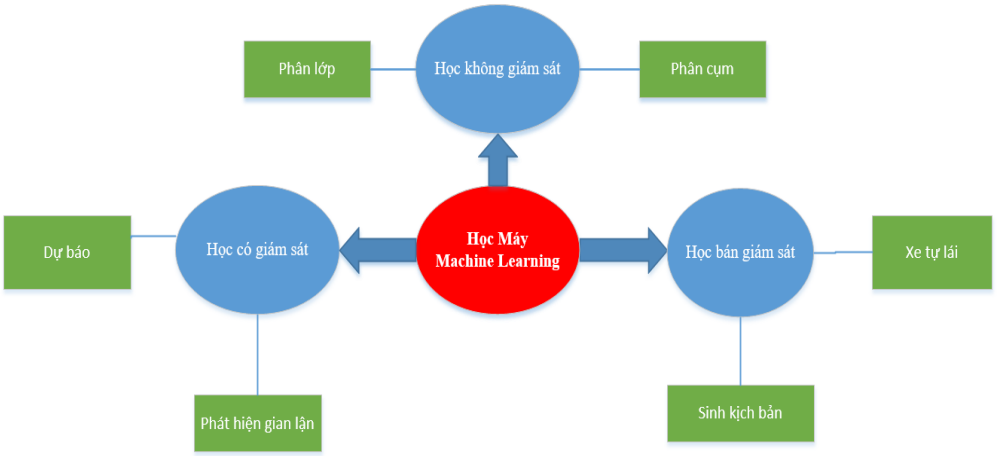
CHƯƠNG 1: TỔNG QUAN VỀ CÁC PHƯƠNG PHÁP PHÂN LOẠI ÂM THANH

Tóm tắt chương: Chương 1 trình bày tổng quan về các phương pháp phân loại âm thanh ứng dụng các mô hình học từ dữ liệu, từ phương pháp dùng mô hình học máy truyền

thống đến các mô hình học sâu. Ưu nhược điểm của từng phương pháp sẽ được phân tích để lựa chọn ra phương pháp phù hợp cho bài toán đang cần giải quyết.

1.1. Mô hình học máy truyền thống

1.1.1. Giới thiệu về học máy và các mô hình học máy truyền thống



Hình 1.1: Phân loại các thuật toán học máy

- Học máy được giám sát (*Supervised*): đây là các thuật toán áp dụng những gì đã được học trong quá khứ vào dữ liệu mới bằng cách sử dụng các ví dụ được gắn nhãn để dự đoán các sự kiện trong tương lai. Bắt đầu từ việc phân tích một tập dữ liệu huấn luyện đã biết, thuật toán học tạo ra một hàm được suy ra để đưa ra dự đoán về các giá trị đầu ra.

- Học máy không giám sát (*Unsupervised*): ngược lại, thuật toán học máy không giám sát được sử dụng khi thông tin được sử dụng để đào tạo không được phân loại cũng không được dán nhãn. Nghiên cứu học tập không giám sát làm thế nào các hệ thống có thể suy ra một chức năng để mô tả một cấu trúc ẩn từ dữ liệu không được gắn nhãn.

- Học máy bán giám sát (*Reinfocement*): các thuật toán học máy được giám sát bán nằm ở đâu đó giữa học tập có giám sát và không giám sát, vì chúng sử dụng cả dữ liệu được gắn nhãn và không nhãn cho đào tạo - thường là một lượng nhỏ dữ liệu được gắn nhãn và một lượng lớn dữ liệu không được gắn nhãn. Các hệ thống sử dụng phương pháp này có thể cải thiện đáng kể độ chính xác trong học tập.

1.1.2. Giới thiệu một số thuật toán học máy có giám sát

1.1.3. Giới thiệu về đặc trưng thủ công

❖ **Một số hand-crafted features thông dụng**

- SIFT (Scale Invariant Feature Transform)

  SIFT là một feature descriptor được sử dụng trong computer vision và xử lý hình ảnh được dùng để nhận dạng đối tượng, khớp hình ảnh, hay áp dụng cho các bài toán phân loại…

  Với đầu vào là một hình ảnh, sử dụng SIFT ta thu được các tập các đặc trưng. Mỗi đối tượng trong hình ảnh sẽ cho ra các đặc trưng khác nhau. Ứng với mỗi keypoint ta sẽ thu được: toạ độ keypoint, scale và orientation của keypoint, descriptor. Ta phân biệt các keypoint này với nhau thông qua một vector 128 chiều hay còn gọi là descriptor. Các descriptor này được sử dụng để nhận dạng đối tượng hay sử dụng cho các bài toán phân loại.

- SURF (Speeded-Up Robust Features)

  Cũng gồm các bước tương tự như SIFT nhưng ở từng bước, SURF có những cải thiện đáng kể để cải thiện tốc độ tính toán mà vẫn đảm bảo độ chính xác. SURF có tốc độ xử lý nhanh gấp nhiều lần so với SIFT.

- HOG (Histogram of Oriented Gradients)

  HOG là một feature descriptor được sử dụng trong computer vision và xử lý hình ảnh, dùng để nhận diện và mô tả hình dạng một đối tượng.

  HOG tương tự như các biểu đồ edge orientation, scale-invariant feature transform descriptors (như sift, surf, ...), shape contexts nhưng hog được tính toán trên một lưới dày đặc các cell và chuẩn hóa sự tương phản giữa các block để nâng cao độ chính xác.

1.2. Các mô hình Deep Learning: RNN và CNN

1.2.1. RNN với dữ liệu tín hiệu trên miền thời gian

1.2.2. CNN với dữ liệu "ảnh của âm thanh" (dạng biểu diễn tần số của âm thanh)

1.3. Các mô hình mô hình học sâu cho phân loại hình ảnh

1.3.1. Các mô hình học sâu tiên tiến

❖ ***AlexNet***

❖ ***ZF Net***

- ***GoogleNet (2015)***
- ***ResNet (2015)***
- ***DenseNet (2016)***

1.3.2 Phương pháp học chuyển giao mạng nơ-ron

Học chuyển tiếp là một phương pháp rất nổi tiếng trong lĩnh vực xử lý thị giác máy tính vì với học chuyển tiếp, thay vì phải bắt đầu quá trình học từ đầu, ta có thể bắt đầu từ các mẫu đã được học khi giải quyết các vấn đề khác.

Trong lĩnh vực thị giác máy tính, học chuyển tiếp thường sử dụng các mô hình được huấn luyện sẵn. Mô hình huấn luyện sẵn là mô hình được huấn luyện trên một tập dữ liệu lớn để giải quyết một vấn đề tương tự như cái mà ta muốn giải quyết. Thông thường, ta sẽ sử dụng các mô hình từ các bài nghiên cứu được công khai (ví dụ VGG, Inception, …).

Tại sao nên dùng học chuyển tiếp trong Deep Learning?

- **Không đủ dữ liệu**: DL cần rất nhiều dữ liệu, và ngốn rất nhiều tài nguyên để học trên tập dữ liệu và ứng dụng đó và việc huấn luyện DL trên tập dữ liệu ít phần nhiều là không hiệu quả. Vậy ngoài kỹ thuật data augmentation (kĩ thuật sử dụng để làm tăng số lượng dữ liệu trong tập dữ liệu huấn luyện).
- **Không đủ tài nguyên**: Việc học trên tập dữ liệu lớn rất ngốn nhiều tài nguyên. Transfer learning sẽ góp phần giảm phần nào thời lượng training.
- **Cải thiện chất lượng**. Rất nhiều trường hợp transfer learning cải thiện chất lượng dự đoán của Target Task so với việc train lại từ đầu. Lý do có thể do Source Network được train với dữ liệu lớn và học được tính khái quát hóa tốt hơn.

Transfer learning giúp giải quyết các vấn đề mà không có dữ liệu, các vấn đề mới chưa được học trước đó.

1.3.2. Phương pháp học chuyển giao mạng nơ-ron

1.3. Kết luận chương 1

Trong chương 1, luận văn đã tìm hiểu được tổng quan về học máy, giới thiệu một số thuật toán học máy truyền thống cũng như tìm hiểu về các mô hình, ứng dụng của

các mạng Deep Learning như RNN&CNN. Trong chương tiếp theo, luận văn sẽ nghiên cứu cơ sở lý thuyết về âm thanh, phương pháp chuyển đổi tín hiệu âm thanh từ miền thời gian sang miền tần số từ đó đưa ra được "ảnh phổ của âm thanh".

CHƯƠNG 2: MỘT SỐ PHƯƠNG PHÁP PHÂN LOẠI ÂM THANH

Tóm tắt chương: Trong chương 2, luận văn trình bày các cơ sở lý thuyết liên quan đến phương pháp sử dụng mô hình học sâu CNN trên dữ liệu dạng biểu diễn tần số của âm thanh. Cụ thể, phần đầu tiên của chương này sẽ tóm lược các một vài phương pháp chuyển dữ liệu âm thanh từ miền thời gian sang miền tần số để thu được "ảnh của âm thanh". Tiếp đến, một số mô hình học sâu quan trọng, thường dùng trong các bài toán phân loại hình ảnh sẽ được trình bày.

2.1. Phương pháp tiền xử lý dữ liệu âm thanh

2.1.1. Short-time Fourier Transform

Short-time Fourier Transform là phép biến đổi Fourier thời gian ngắn (STFT), nói một các đơn giản, STFT thực hiện các nội dung sau:

- Định nghĩa một cửa sổ để phân tích (ví dụ: hẹp 30ms, rộng 5ms);
- Định nghĩa một lượng trùng lên nhau giữa các cửa sổ (ví dụ: 30%);
- Định nghĩa hàm cửa sổ (ví dụ: Hann, Gaussian);
- Tạo ra phân đoạn cửa sổ (nhân tín hiệu với hàm cửa sổ);
- Áp dụng biến đổi Fourier thời gian ngắn với mỗi phân đoạn cửa sổ.

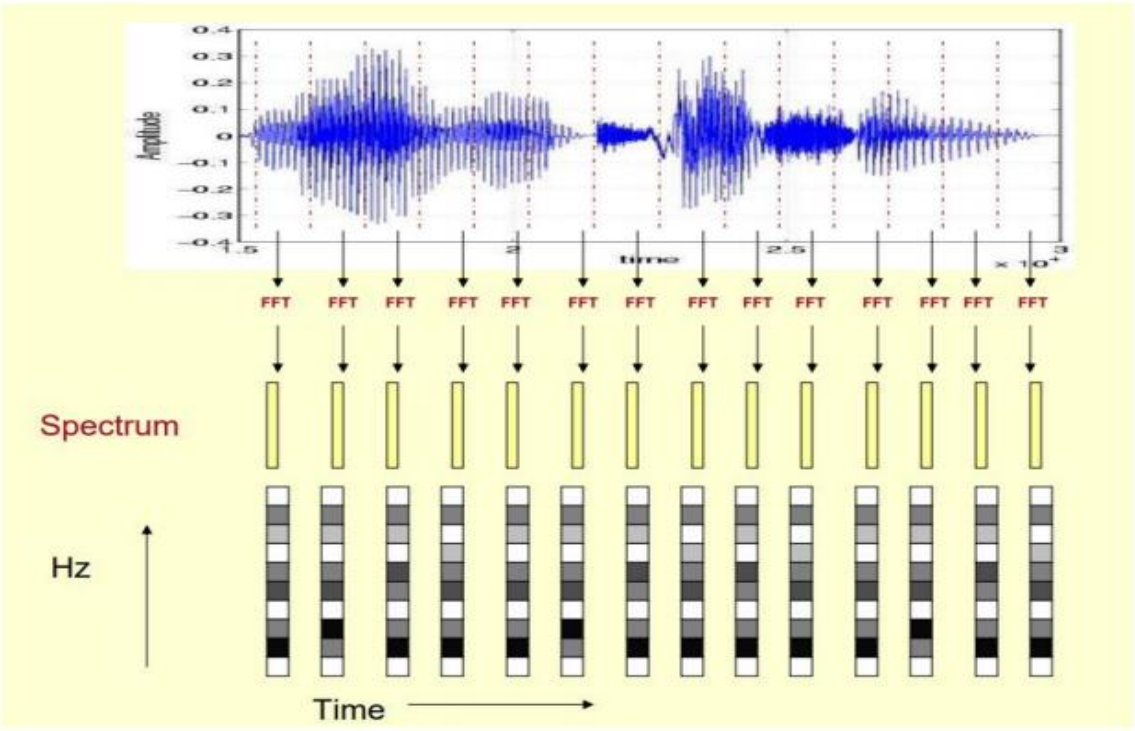
2.1.2. Spectrogram

Có thể liên tưởng spectrogram là "bức tranh của âm thanh", với một cột biểu diễn những tần số tạo nên âm thanh, từ thấp đến cao trong khi cột còn lại biểu diễn sự thay đổi theo thời gian. Nói đơn giản, spectrogram là những tín hiệu âm thanh được biểu diễn bởi một chuỗi các vector quang phổ.

Spectrogram được tạo nên bởi các bước sau:

- Chia tín hiệu thành các đoạn có độ dài bằng nhau. Các phân đoạn phải đủ ngắn để đảm bảo tần số của tín hiệu không thay đổi đáng kể trong đoạn. Đoạn có thể trùng lặp hoặc không.
- Nhân chập mỗi đoạn với hàm cửa sổ để tính STFT.

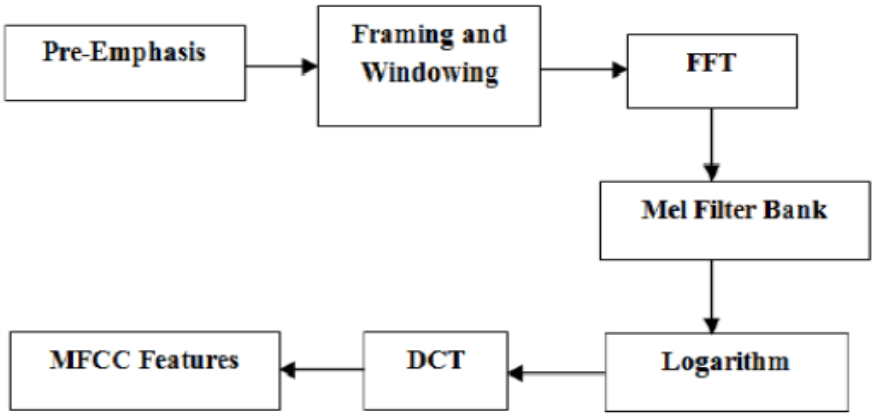
Trục x là trục thời gian (tương ứng với thứ tự các frame), trục y thể hiện dải tần số từ 0 đến 10000 Hz, giá trị cường độ tại từng tần số được thể hiện bằng màu sắc.



Hình 2.3: Spectrogram của âm thanh

(Nguồn: Tìm hiểu về xử lý ngôn ngữ tư nhiên Speeech To Text - Internet)

2.1.3. Ngân hàng bộ lọc và Mel-Frequency Cepstral Coefficients (MFCC)

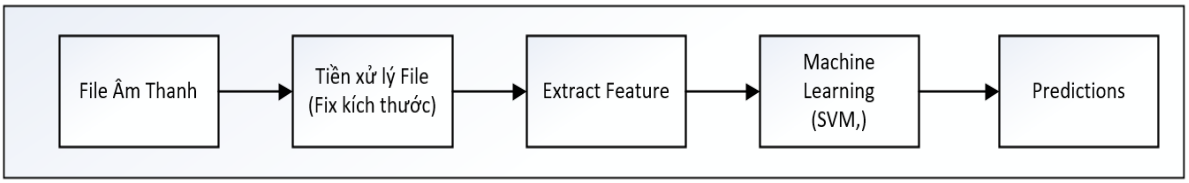


Hình 2.5: Quá trình phép biến đổi MFCC

2.2. Giải pháp thường áp dụng để xây dựng mô hình phân loại âm thanh

2.2.1. Phương pháp sử dụng học máy truyền thống

Quá trình thực thiện xây dựng mô hình phân loại âm thanh dựa trên học máy truyền thống được mô tả như hình dưới đây:



Có rất nhiều các đặc trưng có thể được sử dụng để mô tả dữ liệu âm thanh, tuy nhiên, trong phạm vi của bài luận này, các đặc trưng sẽ không được liệt kê và giải thích rõ ràng hết tất cả. Dưới đây là một số đặt trưng quan trọng được sử dụng rộng rãi:

- Các đặc trưng thống kê: tần số trung bình, độ lệch chuẩn của tần số, tần số trung vị, độ lệch, entropy của phổ, độ phẳng của phổ, mode của tần số, centroid của tần số, …
- Mô tả tông của âm thanh: hàm cao độ, FFT âm độ, HPCP, tần số lên dây (Tuning Frequency), phát hiện hợp âm (Chords Detection), …
- Các đặc trưng miền thời gian: ZCR, độ to, khoảng thời gian hiệu quả (Effective Duration), …
- Đặc trưng về nhịp điệu: Beat Tracker Degara, Beat Tracker MultiFeature, Novelty Curve, Onset Detection, Onsets, …

Đặc trưng SFX: LogAttackTime, MaxToTotal, MinToTotal

❖ ***Đánh giá***

Phương pháp tiếp cận bằng học máy cơ bản khó sử dụng, bởi các phương pháp này rất tốn tài nguyên cho bước trích xuất đặc trưng và các đặc trưng này rất nhạy cảm với độ nhiễu của dữ liệu âm thanh.

2.2.2. Phương pháp sử dụng bộ nhớ dài ngắn hạn (LSTM) với tín hiệu thô

Các kiến trúc LSTM thông thường khá tốt cho việc phát hiện ra các đặc điểm trong các tín hiệu, do đó phương pháp này có thể được dùng để tạo mô hình học máy cho tín hiệu âm thanh liên tục trên miền thời gian.

Một cách cụ thể, một tín hiệu âm thanh có thể được chia ra thành các phần trên miền thời gian, mỗi phần là mảng một chiều có độ dài cố đinh, có thể thêm phần đệm nếu không đủ độ dài. Giờ đây, mỗi tín hiệu âm thanh có thể được hiểu là một chuỗi các

vectơ phân đoạn, các vectơ này sau đó được cấp cho LSTM để mô hình có thể xuất ra một vectơ tóm tắt cho toàn bộ âm thanh.

❖ ***Đánh giá***

LSTM trở nên phổ biến vì có thể giải quyết vấn đề mất mát gradient. Nhưng thực tế, LSTM không thể xử lý vấn đề này một cách hoàn hảo. Vấn đề nằm ở chỗ, dữ liệu đầu vào vẫn phải di chuyển từ cell này sang cell khác trong quá trình tính toán. Hơn nữa, mô hình sẽ trở nên khá phức tạp với việc bổ sung thêm các đặc trưng (chẳng hạn như cổng quên – forget gate).

LSTM đòi hỏi nhiều tài nguyên, thời gian để được huấn luyện và trở nên sẵn sàng cho các ứng dụng trong thế giới thực. Về mặt kỹ thuật, LSTM cần băng thông bộ nhớ cao do các lớp tuyến tính hiện diện trong mỗi ô mà hệ thống thường không cung cấp được.

LSTM bị ảnh hưởng bởi các lần khởi tạo trọng số ngẫu nhiên khác nhau và do đó hoạt động khá giống với mạng nơ-ron truyền thẳng (feed – forward).

Các LSTM dễ bị overfitting và rất khó áp dụng thuật toán drop out để giải quyết vấn đề này.

2.2.3. Phương pháp sử dụng CNN với các đặc trưng về tần số

Theo cách tiếp cận này, dữ liệu âm thanh sẽ được trích xuất thành các đặc trưng dựa trên tần số như MFCC, biểu đồ phổ log-mel như đã thảo luận trong Chương 1. Sau đó, các đặc trưng này có thể được diễn giải dưới dạng hình ảnh và có thể được đưa vào các mô hình CNN.

Để được đưa vào các mô hình CNN, các đặc trưng dựa trên tần số có thể được giữ nguyên dưới dạng mảng 2 chiều, hoặc được chuyển đổi sang mảng 3 chiều để các mô hình CNN có thể xem chúng dưới dạng hình ảnh 3 kênh. Có một số cách phổ biến để hoàn thành nhiệm vụ này, cách thông thường là lặp lại từng phần tử, tiếp theo là các bước chuẩn hóa (các giá trị sẽ nằm trong khoảng [0, 255]); sử dụng thư viện của bên thứ ba (ví dụ: python matplotlib) để chuyển đổi và lưu chúng vào đĩa dưới dạng ảnh màu.

Về kiến trúc mạng sẽ được sử dụng, chúng ta có thể tạo một mạng CNN của riêng mình hoặc sử dụng kiến trúc CNN hiện đại như VGG, ResNet, Inception, v.v. Các mô hình này có thể được tinh chỉnh trọng số ImageNet để hội tụ nhanh hơn.

Nếu chuyển đổi các đặc trưng dựa trên tần số thành "hình ảnh màu", chúng ta có thể sử dụng các kỹ thuật nâng cao thường được áp dụng cho hình ảnh bình thường chẳng hạn như lật ngang hoặc lật dọc, cắt ngẫu nhiên, thêm nhiễu phụ gia trắng, biểu đồ chuẩn hóa đặc trưng, ...

❖ ***Đánh giá***

Mạng nơ-ron tích chập thường được sử dụng để phân loại hình ảnh, chúng thường xử lý dữ liệu nhiều chiều (hình ảnh). Mặc dù cấu trúc của ConvNet nhằm mục đích giảm thiểu sự overfitting, một lượng lớn dữ liệu vẫn là điều cần thiết để cho một mạng CNN hoạt động hiệu quả.

Vì vậy, luận văn lựa chọn phương pháp sử dụng CNN với các đặc trưng về tần số để xây dựng mô hình.

2.3. Kết luận chương 2

Trong chương 2, luận văn đã giới thiệu tổng quát cơ sở lý thuyết và các đặc trưng cơ bản của âm thanh. Đưa ra được khái niệm "ảnh chụp của âm thanh" bằng các phép biến đổi từ miền âm thanh (liên tục) sang miền tần số (rời rạc). Đưa ra hướng giải quyết chuyển từ phân loại âm thanh sang phân loại hình ảnh ứng dụng các kiến trúc mạng Deep Learning tiến tiến hiện nay. Ngoài ra nội dung chương cũng đã nghiên cứu và đề cập phương pháp học chuyển tiếp, đưa ra nhưng điểm ưu việt của phương pháp này.

Trên cơ sở các kết quả đã đạt được của chương 2, trong chương tiếp theo luận văn sẽ tiến hành thực nghiệm xây dựng mô hình phân loại giới tính và vùng miền ứng dụng phương pháp học chuyển tiếp dựa trên kiến trúc DenseNet, ResNet.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

Tóm tắt chương: Trong chương 3, luận văn sẽ trình bày các bước triển khai tiền xử lý dữ liệu và xây dựng, huấn luyện mô hình cũng như các bước tinh chỉnh xử lý sau khi xây dựng mô hình, sau cùng là đánh giá độ chính xác của mô hình trên dữ liệu mới.

Có rất nhiều cách để giải quyết nhiệm vụ phân loại âm thanh, chúng khác nhau ở cách tín hiệu âm thanh được diễn giải và các đặc trưng được trích xuất, các mô hình được huấn luyện từ các đặc trưng này, từ các mô hình học máy truyền thống đến các mô hình học sâu với các loại kiến trúc khác nhau. Trong phần này, trước tiên ta sẽ thảo luận về ba cách tiếp cận chính đối với các nhiệm vụ phân loại âm thanh, mỗi cách có ưu điểm và nhược điểm khác nhau khi được áp dụng. Sau đó, thực hiện phân tích khai thác dữ liệu để chỉ ra các thách thức gặp phải, từ đó đưa ra các phương hướng để giải quyết khi xây dựng mô hình nhận dạng âm thanh

3.1. Giới thiệu về bộ dữ liệu âm thanh

Để xây dựng một hệ thống có thể phân loại giới tính và giọng của người nói như đề xuất, ta sẽ sử dụng tập dữ liệu giọng nói từ Zalo AI Challenges. Tập dữ liệu này chứa

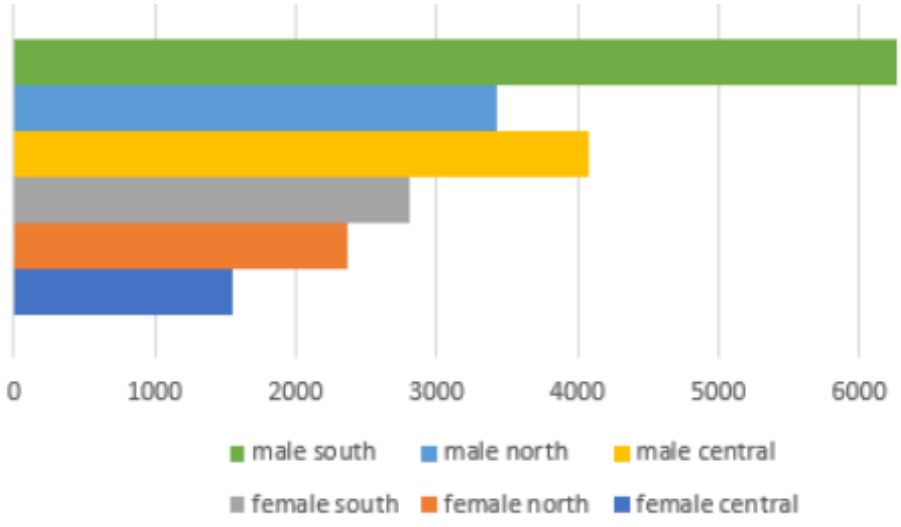
2 thư mục - "train", chứa 6 thư mục với các tệp âm thanh bên trong tương ứng với một trong 6 lớp “female_central”, “female_north”, …, “male_south”; và "test", cũng chứa các tập tin âm thanh. Tập dữ liệu cũng có thêm một tệp csv có 2 cột.

- “train”: chứa 6 thư mục “female_central”, “female_north”, …, "male_south"; mỗi tệp chứa các tệp âm thanh tương ứng với lớp.
- “public_test”: chứa các tệp âm thanh.
- “public_test_gt.csv”: chứa 2 cột, “id” - tên tệp trong public_test, “gender”: “0” (nữ) hoặc “1” (nam), “accent”: “0” (Bắc), “1” (Trung), “2” (Nam).

Dữ liệu trong tập “train” sẽ được sử dụng để huấn luyện mô hình và dữ liệu trong tập “public_test” sẽ được sử dụng để đánh giá mô hình.

Các tệp âm thanh trong tập dữ liệu có 3 định dạng: .wav, amr và .mp3 có thể thay đổi theo độ dài - các tệp trong ‘female_central’ chủ yếu dài 3 giây với tệp dài nhất là 19 giây. Trong khi đó, các tệp ở các lớp khác dao động từ dưới 1 giây đến hơn 200 giây hoặc thậm chí 500 giây như trong “female_south”.

Một vấn đề của tập dữ liệu này là nó rất mất cân bằng. Dữ liệu lớp ‘male_south’ lớn hơn gần 4 lần so với lớp các mẫu "female_central". Sự mất cân bằng này có thể được nhìn thấy trong biểu đồ dưới đây

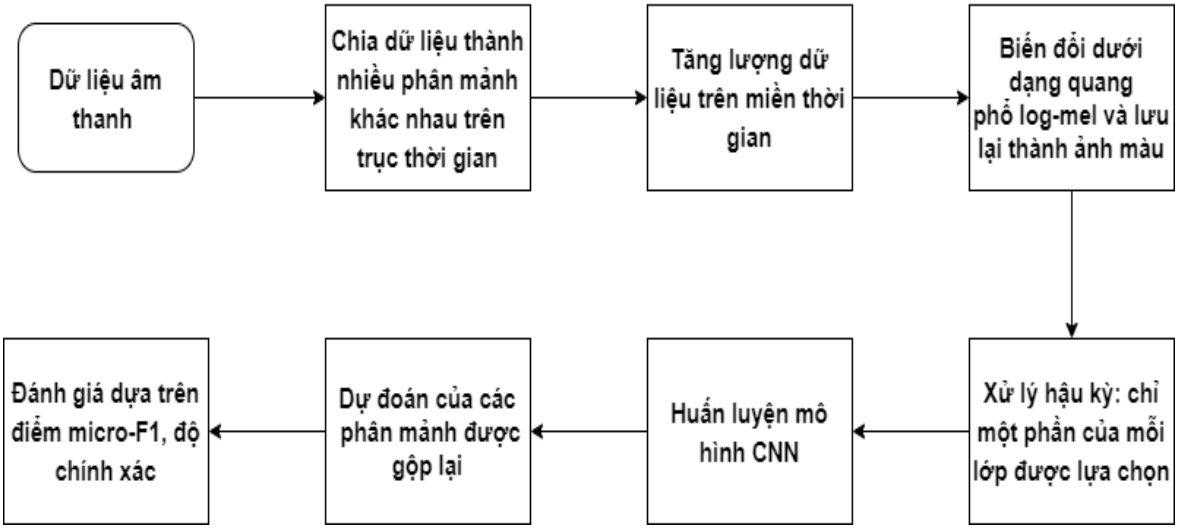


Hình 3.1: Số lượng phân bố của dữ liệu trong tập mẫu

3.2. Kịch bản xây dựng mô hình phân loại giới tính vùng miền

Điều đầu tiên cần làm là thiết lập một chiến lược đánh giá phù hợp. Đối với mỗi lớp, danh sách id người nói được lấy và chia theo tỷ lệ 8:2, đối với tập dữ liệu huấn

luyện, kết quả sẽ xác thực bằng cách lấy các tập tương ứng với mỗi người nói. Làm như vậy, có thể đảm bảo rằng các tệp âm thanh của một người nói chỉ xuất hiện trong một trong các tập: huấn luyện hoặc xác thực. Hình sau trình bày tổng quan về các bước trong phương án vừa được đề xuất:



Hình 3.3: Quá trình xây dựng mô hình

Các tệp âm thanh gốc được chia thành nhiều phần khác nhau trên trục thời gian. Các kỹ thuật để tăng lượng dữ liệu trên miền thời gian như thay đổi tốc độ, trộn tiếng ồn cho các tệp âm thanh của các lớp có ít dữ liệu sẽ được áp dụng.

Các tệp âm thanh sau khi được phân mảnh sẽ được biến đổi dưới dạng quang phổ log-mel và được lưu lại thành ảnh màu. Các bước xử lý này sẽ được được áp dụng cho tập huấn luyện, tập kiểm thử và tập đánh giá.

Trước khi huấn luyện mô hình với các hình ảnh này, xử lý hậu kỳ được thực hiện trên các lớp sử dụng để huấn luyện dữ liệu, mục đích là để làm cân bằng dữ liệu. Đối với mỗi lớp, chỉ một phần của tập dữ liệu được lựa chọn. Điều này giúp cân bằng số lượng của các dữ liệu đã được xử lý ở mỗi lớp và có càng nhiều dữ liệu càng tốt.

Sau khi mô hình CNN được huấn luyện trên tập các hình ảnh của âm thanh, dự đoán của các phần bị phân mảnh sẽ được gộp lại thành dự đoán cho cả tệp âm thanh. Điểm micro F1 và độ chính xác được tính toán trên những dự đoán gộp này.

3.3.1. Tiền xử lý dữ liệu và trích xuất đặc trưng

Biến đổi các phần âm thanh thành quang phổ log-mel với cấu hình như sau:

Bảng 3.1: Các thông số sử dụng trong phép biến đổi âm thanh

Sampling rate	8000 Hz
Manimum frequency	20 Hz
Maximum frequency	4000 Hz
Mel bands	224
Hop length	256
FFT window size	1024

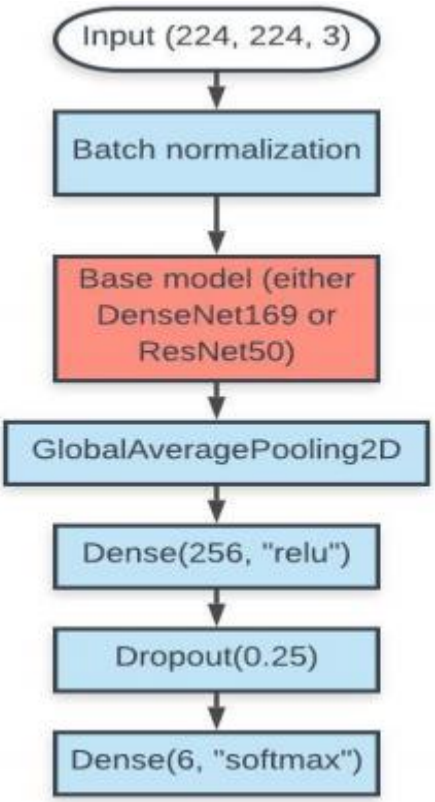
Sau quá trình xử lý hậu kỳ, số lượng các tệp hình ảnh của mỗi lớp như sau:

- “female_central”: 7914 tệp hình ảnh
- “female_north”: 8708 tệp hình ảnh
- “female_south”: 8331 tệp hình ảnh
- “male_central”: 6872 tệp hình ảnh
- “male_north”: 8589 tệp hình ảnh
- “male_south”: 8725 tệp hình ảnh

3.2.2 Làm giàu nguồn dữ liệu

3.2.3. Kiến trúc mạng áp dụng trong mô hình

Phương pháp học chuyển tiếp (transfer learning) sẽ được áp dụng. Huấn luyện một mô hình học sâu cần rất nhiều dữ liệu và tài nguyên. Vậy nên ngoài kĩ thuật tăng cường dữ liệu, học chuyển tiếp là một giải pháp cho bài toán này. Một số kiến trúc CNN hiện đại đã được thử nghiệm làm mô hình cơ sở, ví dụ như DenseNet169, ResNet50.



Hình 3.5: Kiến trúc mạng học sâu để phân loại âm thanh đề xuất

3.3. Cài đặt mô hình phân loại

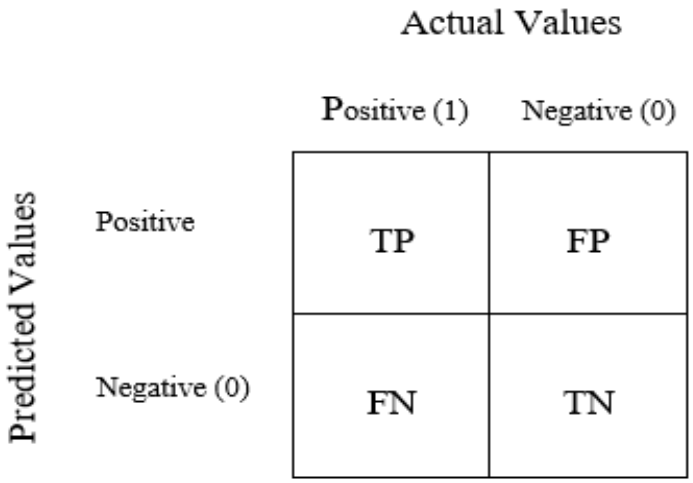
3.3.1. Một số yêu cầu về cài đặt

3.3.2. Phương pháp đánh giá

Trong luận văn này, để đánh giá độ chính xác của mô hình tác giả đã lựa chọn phương pháp đánh giá độ chính xác bằng cách sử dụng ma trận độ đo (confusion matrix) được mô tả như sau:

Confusion Matrix là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:

- **TP (True Positive)**: mẫu mang nhãn dương được phân lớp **đúng** vào lớp **dương**
- **TN (True Negative)**: mẫu mang nhãn âm được phân lớp **đúng** vào lớp **âm**.
- **FP (False Positive - Type 1 Error)**: mẫu mang nhãn âm bị phân lớp **sai** vào lớp **dương**.
- **FN (False Negative - Type 2 Error)**: mẫu mang nhãn dương bị phân lớp **sai** vào lớp **âm**.



Hình 3.6: Ma trận độ đo (Conusion matrix)

Ký hiệu TP là True Positive; TN là True Negative; FP là False Positive và FN là False Negative. Thực hiện phép đo Precision – Recall, trong đó, Precision là tỉ lệ số điểm TP trong những điểm được phân loại Positive, còn Recall là tỉ lệ số điểm TP trong số điểm thực sự là Positive. Công thức như sau:

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

Ta thấy rằng, Precision và Recall phủ càng cao thì càng tốt. Nhưng trong thực tế, hai giá này không thể đạt được cực đại cùng một lúc và thông thường phải tìm kiếm sự cân bằng. Thước đo $F1_{score}$ là trung bình hài hòa giữa Precision và Recall. Nó có xu hướng bằng không nếu hai giá trị này có xu hướng bằng không.

$$F1_{score} = 2 * \frac{precision \times recall}{precision + recall}$$

3.3.3. Kết quả của thử nghiệm

Dưới đây là kết quả ta thu được khi sử dụng mạng Densenet 169 và mạng ResNet50 sau 15 vòng huấn luyện cho mỗi mô hình.

Đánh giá dựa trên các dự đoán không được thống nhất (lựa chọn ngẫu nhiên dự đoán từ một phần của tập tin để dự đoán toàn bộ):

Bảng 3.2: Kết quả xây dựng mô hình dựa trên F1-Score

Dữ liệu	Đánh giá	DenseNet169	ResNet50
Kiểm thử	Độ chính xác tổng	0.78	0.76
	Độ chính xác giới tính	0.96	0.96
	Độ chính xác giọng	0.80	0.78
	Micro F1 tổng	0.78	0.76
	Micro F1 giới tính	0.93	0.93
	Micro F1 giọng	0.80	0.78
Đánh giá	Độ chính xác tổng	0.72	0.73
	Độ chính xác giới tính	0.95	0.96
	Độ chính xác giọng	0.75	0.75
	Micro F1 tổng	0.72	0.73
	Micro F1 giới tính	0.89	0.89
	Micro F1 giọng	0.75	0.75

Đánh giá dựa trên dự đoán được thống nhất bằng phương pháp trung bình nhân các kết quả của từng phần:

Bảng 3.3: Kết quả xây dựng mô hình dựa trên Micro - F1 Score

Dữ liệu	Đánh giá	DenseNet169	ResNet50
Kiểm thử	Độ chính xác tổng	0.81	0.80
	Độ chính xác giới tính	0.97	0.97
	Độ chính xác giọng	0.83	0.82
	Micro F1 tổng	0.81	0.80
	Micro F1 giới tính	0.95	0.95
	Micro F1 giọng	0.83	0.82
Đánh giá	Độ chính xác tổng	0.78	0.78
	Độ chính xác giới tính	0.97	0.97
	Độ chính xác giọng	0.80	0.80
	Micro F1 tổng	0.79	0.78
	Micro F1 giới tính	0.92	0.92
	Micro F1 giọng	0.80	0.80

Dựa vào kết quả của, ta có thể quan sát được rằng, với phương pháp tổng hợp các dự đoán, kết quả được cải thiện và DenseNet169 có độ chính xác cao hơn ResNet50. Cả 2 mô hình đều có điểm số cao với phân loại giới tính và điểm chấp nhận được với phân loại giọng vùng miền.

3.4. Kết luận chương 3

Chương 3 đã trình bày quá trình thu thập, xử lý dữ liệu chuyển đổi từ âm thanh sang ảnh phục vụ cho quá trình huấn luyện mô hình. Từ bộ dữ liệu này, luận văn xây dựng và huấn luyện mô hình phân loại giới tính và vùng miền pretrain model là Resnet 50 và DenseNet169.

Các kết quả thử nghiệm ban đầu cho thấy giải pháp xây dựng mô hình phân loại giới tính và vùng miền cho tiếng việt dựa trên âm thanh đề xuất có tính khả thi cao và phù hợp với các yêu cầu đề ra.

KẾT LUẬN

1. Những đóng góp của luận văn

Với mục tiêu nghiên cứu các phương pháp xử lý dữ liệu giọng nói từ đó xây dựng mô hình học sâu phân loại giới tính và vùng miền cho tiếng việt, luận văn đã đi sâu nghiên cứu các vấn đề xung quanh đề tài nghiên cứu, từ các thuật toán học máy cổ điển đến các mô hình mạng nơ-ron tiến tiến hiện nay.

Những kết quả chính đã đạt được trong luận văn:

- Nghiên cứu về học máy cổ điển và các mô hình học sâu
- Tìm hiểu về các cơ sở lý thuyết về âm thanh, từ đó đưa ra phương pháp trích chọn thuộc tính từ giọng nói
- Lựa chọn và áp dụng thành công mô hình học chuyển tiếp ứng dụng các kiến trúc mạng nâng cao như ResNet, DenseNet
- Thực nghiệm cài đặt và xây dựng mô hình phân loại giọng nói vùng miền

2. Hướng phát triển của luận văn

Một số hướng phát triển tiếp theo của luận văn:

- Hiện tại tốc độ xử lý phân loại còn chưa nhanh, cần thử nghiệm thêm về các tham số trong mô hình nhằm giảm thời gian của việc phân loại
- Mặc dù ở bước tiền xử lý dữ liệu đã nêu ra được hướng xử lý âm thanh trong môi trường chứa nhiễu, tuy nhiên trong thực tế nếu ở môi trường nhiễu thì hệ thống hoạt động không ổn định.
- Kết quả của mô hình cho thấy việc phân loại vùng miền (Bắc-Trung-Nam) có độ chính xác chưa cao, cần thu thập, làm giàu thêm dữ liệu tập huấn luyện để tăng độ chính xác của mô hình.