**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**HỌC SÂU**

*Người thực hiện*: **NGUYỄN TRƯỜNG AN – 21010151**

**TRƯƠNG THÀNH THẢO – 21002721**

Lớp **: 420300371101**

Khoá  **: 17**

*Người hướng dẫn*: **TS ĐẶNG THỊ PHÚC**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**HỌC SÂU**

Người thực hiện: **NGUYỄN TRƯỜNG AN- 21010151**

**TRƯƠNG THÀNH THẢO – 21002721**

Lớp **: 420300371101**

Khoá  **: 17**

Người hướng dẫn: **TS. ĐẶNG THỊ PHÚC**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành đến cô Đặng Thị Phúc đã tạo điều kiện cho em thực hiện dự án này.

Cô Phúc đã không chỉ chia sẻ kiến thức chuyên môn mà còn tạo điều kiện cho chúng em tiếp cận và nắm vững các khía cạnh quan trọng của lĩnh vực Deep Learning. Chúng em xin chân thành cảm ơn cô đã luôn sẵn sàng giải đáp mọi câu hỏi và hỗ trợ trong việc xác định và giải quyết các khó khăn trong quá trình học tập.

Dự án này đã là cơ hội thú vị và quý báu để chúng em áp dụng kiến thức của mình và phát triển kỹ năng trong lĩnh vực Deep Learning. Chúng em rất biết ơn và trân trọng sự hướng dẫn và cơ hội này.

Chúng em hy vọng rằng dự án này sẽ đánh giá cao và phản ánh tốt sự cống hiến của chúng em trong việc nghiên cứu và phát triển trong lĩnh vực này. Chúng em sẽ tiếp tục học hỏi và phát triển khả năng của mình dưới sự hướng dẫn của thầy.

Một lần nữa, chúng em xin chân thành cảm ơn cô Phúc và mong rằng sẽ có nhiều cơ hội hợp tác khác trong tương lai!

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

1. Vấn đề nghiên cứu:

Nhận diện giọng nói vùng miền là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói. Với sự đa dạng về ngôn ngữ và văn hóa vùng miền, việc phân loại giọng nói theo ba vùng chính của Việt Nam (Bắc, Trung, Nam) là một thách thức. Nghiên cứu này nhằm phát triển các mô hình học sâu để nhận diện và phân loại giọng nói vùng miền Việt Nam.

1. Các hướng tiếp cận:

* **Thu thập dữ và tiền xử lý dữ liệu:** Sử dụng bộ dữ liệu **“VIVOS”** đa dạng.
* **Trích xuất đặc trưng:** Sử dụng MFCC để nhận dạng các đặc trưng âm thanh của từng vùng miền
* **Xây dựng mô hình học sâu:** Áp dụng các mô hình như KNN, MLP, CNN
* **Huấn luyện và đánh giá mô hình:** Dữ liệu được chia thành tập huấn luyện và tập kiểm tra để đánh giá hiệu quả của các mô hình.

1. **Cách giải quyết vấn đề:**

* Tiền xử lý dữ liệu: Loại bỏ nhiễu, chuẩn hóa và trích xuất đặc trưng MFCC
* Xây dựng và huấn luyện mô hình: Tối ưu hoá mô hình
* Đánh giá mô hình: Sử dụng các chỉ số đánh giá để so sánh và chọn mô hình tốt nhất

1. **Kết quả đạt được:**

* **Hiệu quả của MFCC**: Phương pháp trích xuất đặc trưng hiệu quả.
* **Mô hình Random Forest:** Cho kết quả tốt nhất với độ chính xác và F1-score cao.
* **Sự khác biệt vùng miền rõ rệt:** Các đặc trưng giọng nói giúp phân biệt tương đối chính xác.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc184201831)

[TÓM TẮT iii](#_Toc184201832)

[1. Vấn đề nghiên cứu: iii](#_Toc184201833)

[2. Các hướng tiếp cận: iii](#_Toc184201834)

[3. Cách giải quyết vấn đề: iii](#_Toc184201835)

[4. Kết quả đạt được: iii](#_Toc184201836)

[DANH MỤC KÍ HIỆU VÀ CÁC CHỮ VIẾT TẮT 3](#_Toc184201838)

[NHẬN DẠNG TIẾNG NÓI VÙNG MIỀN VIỆT NAM BẰNG CÁCH PHƯƠNG PHÁP HỌC SÂU 4](#_Toc184201839)

[1.1 Giới thiệu về bài toán 4](#_Toc184201840)

[1.2 Phân tích yêu cầu của bài toán 5](#_Toc184201841)

[1.2.1 Yêu cầu của bài toán 5](#_Toc184201842)

[1.2.2 Các phương pháp giải quyết bài toán 6](#_Toc184201843)

[1.2.3 Phương pháp đề xuất giải quyết bài toán 6](#_Toc184201844)

[1.3 Phương pháp giải quyết bài toán 7](#_Toc184201845)

[1.3.1 Mô hình tổng quát 7](#_Toc184201846)

[1.3.2 Đặc trưng của mô hình đề xuất 7](#_Toc184201847)

[1.4 Thực nghiệm 8](#_Toc184201848)

[1.4.1 Dữ liệu 8](#_Toc184201849)

[1.4.2 Xử lý dữ liệu 9](#_Toc184201850)

[1.4.3 Công nghệ sử dụng 9](#_Toc184201851)

[1.4.4 Cách đánh giá 9](#_Toc184201852)

[1.5 Kết quả đạt được 11](#_Toc184201853)

[1.6 Kết luận 11](#_Toc184201854)

[TÀI LIỆU THAM KHẢO 12](#_Toc184201855)

# DANH MỤC KÍ HIỆU VÀ CÁC CHỮ VIẾT TẮT

* **CÁC CHỮ VIẾT TẮT**

MLP: Multi – Layer Perceptron

HMM: Hidden Markov Model

GMM: Gaussian Mixture Model

LTSTM: Long Short-Term Memory

ANN: Artificial Neural Network

CNN: Convolutional Neural Network

LSTM: Long Short-term Memory

MFCC: Mel-Frequency Cepstral Coefficients

FFT: Fast Fourier Transform

MSE: Mean Squared Error

RMSE: Root Mean Squared Error

MAP: Mean Average Precision

# NHẬN DẠNG TIẾNG NÓI VÙNG MIỀN VIỆT NAM BẰNG CÁCH PHƯƠNG PHÁP HỌC SÂU

1.1 Giới thiệu về bài toán

*Trong nghiên cứu này, chúng tôi tiếp cận bài toán phân loại giọng nói vùng miền Việt Nam thông qua việc áp dụng các phương pháp học sâu. Mục tiêu của chúng tôi là xác định nguồn gốc của một đoạn giọng nói thuộc ba vùng miền chính của Việt Nam: Bắc* – *Trung – Nam. Đây là một bài toán phân loại quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) và nhận dạng giọng nói (Speech Recognition). Bằng cách tiếp cận này, nhóm hy vọng có thể cải thiện khả năng nhận diện và phân loại giọng nói theo vùng miền một cách chính xác và hiệu quả.*

Khi nói về bài toán này thì có nhiều ý nghĩa quan trọng như:

* Cải thiện hệ thống trợ lý ảo và nhận dạng giọng nói: Khả năng nhận diện chính xác vùng miền của giọng nói có thể nâng cao độ chính xác và tính cá nhân hóa của các hệ thống trợ lý ảo như Google Assistant, Siri, hay các dịch vụ chăm sóc khách hàng tự động. Điều này góp phần mang lại trải nghiệm tốt hơn cho người dùng.
* Bảo tồn và nghiên cứu ngôn ngữ địa phương: Với sự đa dạng về giọng nói ở các vùng miền tại Việt Nam, việc nghiên cứu và phân loại giọng nói không chỉ giúp bảo tồn các đặc điểm ngôn ngữ mà còn cung cấp thêm hiểu biết sâu sắc về văn hóa địa phương.
* Ứng dụng trong giáo dục và đào tạo: Các công cụ phân loại giọng nói có thể hỗ trợ trong việc giảng dạy ngôn ngữ, giúp học viên phát âm chuẩn hơn theo từng vùng miền, đồng thời hỗ trợ các ứng dụng giáo dục ngôn ngữ.
* Phát triển các ứng dụng thương mại: Trong lĩnh vực kinh doanh, nhận diện giọng nói vùng miền giúp các doanh nghiệp hiểu rõ hơn về khách hàng, từ đó cung cấp dịch vụ phù hợp và tối ưu hóa trải nghiệm người dùng, đặc biệt trong các dịch vụ chăm sóc khách hàng.
* Nghiên cứu học thuật: Bài toán phân loại giọng nói vùng miền mở ra nhiều cơ hội mới trong lĩnh vực học sâu và xử lý ngôn ngữ tự nhiên. Đặc biệt, việc áp dụng các mô hình học sâu (deep learning) giúp giải quyết hiệu quả các bài toán phức tạp liên quan đến ngôn ngữ và giọng nói.

1.2 Phân tích yêu cầu của bài toán

1.2.1 Yêu cầu của bài toán

* Xác định đặc trưng giọng nói vùng miền: Hệ thống cần nhận diện chính xác giọng nói thuộc ba miền Bắc, Trung, Nam của Việt Nam dựa trên các đặc trưng ngữ âm và ngữ điệu.
* Xử lý dữ liệu âm thanh: Dữ liệu giọng nói cần được thu thập, làm sạch, và xử lý trước khi đưa vào hệ thống. Quá trình này bao gồm việc chuẩn hóa âm thanh, cắt xén nhiễu, và trích xuất đặc trưng bằng các phương pháp như MFCC hoặc Spectrogram.
* Thiết kế và huấn luyện mô hình học sâu: Hệ thống yêu cầu sử dụng các mô hình học sâu như CNN, MLP, và LSTM để học và phân loại các đặc trưng từ dữ liệu giọng nói. Mô hình phải được tối ưu hóa để đạt độ chính xác cao.
* Đánh giá hiệu quả mô hình: Hệ thống cần có tiêu chí đánh giá hiệu quả, chẳng hạn như độ chính xác (accuracy), độ nhạy (recall), và độ đặc hiệu (specificity) để so sánh và lựa chọn mô hình phù hợp.
* Khả năng tổng quát hóa: Mô hình phải đảm bảo khả năng áp dụng với các giọng nói khác nhau không nằm trong tập huấn luyện, đảm bảo tính chính xác khi phân loại giọng nói thực tế.
* Ứng dụng thực tiễn: Kết quả của hệ thống phải khả thi để tích hợp vào các ứng dụng như trợ lý ảo, dịch vụ chăm sóc khách hàng, hoặc nghiên cứu văn hóa và ngôn ngữ địa phương.

1.2.2 Các phương pháp giải quyết bài toán

* Mạng nơ-ron tích chập (CNN): Đây là phương pháp mạnh mẽ để xử lý dữ liệu có cấu trúc, đặc biệt là âm thanh. CNN giúp trích xuất tự động các đặc trưng từ giọng nói mà không cần phải thiết kế thủ công.
* Mạng Perceptron đa lớp (MLP): Là một phương pháp học sâu cơ bản nhưng hiệu quả trong các bài toán phân loại nhờ khả năng kết nối các đặc trưng phi tuyến tính.
* Long Short-Term Memory (LSTM): Một biến thể của mạng nơ-ron hồi tiếp (RNN), được thiết kế để xử lý và học các chuỗi dữ liệu dài hạn. LSTM rất phù hợp cho việc phân loại giọng nói, nhờ khả năng ghi nhớ và khai thác thông tin từ các mối quan hệ ngữ âm trong chuỗi âm thanh theo thời gian.

1.2.3 Phương pháp đề xuất giải quyết bài toán

Nhận dạng giọng nói vùng miền Việt Nam là một lĩnh vực đầy tiềm năng nhưng cũng đầy thử thách bởi sự đa dạng về phương ngữ và giọng điệu trên khắp đất nước. Để chinh phục thử thách này, chúng tôi đã dày công nghiên cứu và suy nghĩ nhiều phương pháp để mở ra những cánh cửa mới việc giải quyết vấn đề này.

Từ những nghiên cứu trong thời gian qua, nhóm đề xuất phương pháp giải quyết bài toán như sau:

* MLP (Multilayer Perceptron): MLP là một mạng nơ-ron nhân tạo được sử dụng để phân loại dữ liệu.
* LSTM (Long Short-Term Memory): LSTM là mạng nơ-ron hồi tiếp, chuyên xử lý dữ liệu chuỗi và phụ thuộc theo thời gian, thường được dùng trong phân loại giọng nói và văn bản.
* CNN (Convolutional Neural Networks): CNN là mạng nơ-ron chuyên nhận diện các đặc trưng không gian trong hình ảnh hoặc phổ âm thanh, dùng để phân loại giọng nói và hình ảnh.

Thu thập dữ liệu giọng nói Việt Nam và label vùng miền chính xác sau đó xử lý data và trích xuất đặc trưng bằng phương pháp MFCC, cuối cùng sẽ dùng nhiều Model Deep learning để phân loại bao gồm Multi-Layer Perceptron, CNN, LSTM, sau đó dùng Accuracy để đánh giá độ chính xác.

1.3 Phương pháp giải quyết bài toán

1.3.1 Mô hình tổng quát

A diagram of a model

Description automatically generated

Hình 1: Mô hình tổng quát của bài toán

* Tệp âm thanh đầu vào đã được làm sạch, định dạng file .wav khoảng 10s.
* Được trích xuất đặc trưng bằng MFCC để lấy các feature cần thiết cho việc phân loại
* Sử dụng các thuật toán, mô hình deep learning để phân loại các tệp âm thanh
* Đánh giá và lưu mô hình, và có thể predict ra được giọng từ một người tới từ miền nào.

1.3.2 Đặc trưng của mô hình đề xuất

Mô hình đề xuất trong hình ảnh là một hệ thống nhận dạng giọng nói dựa trên các đặc trưng MFCC và các thuật toán học sâu. Hệ thống bao gồm các thành phần chính sau:

Dữ liệu đầu vào: Dữ liệu đầu vào là các tệp âm thanh định dạng .wav có độ dài khoảng 10 giây. Các tệp âm thanh này cần được làm sạch trước khi đưa vào hệ thống để loại bỏ nhiễu và tạp âm.

Trích xuất đặc trưng MFCC: Các tệp âm thanh đầu vào được trích xuất các đặc trưng MFCC (Mel-Frequency Cepstral Coefficients). MFCC là một tập hợp các đặc trưng tần số được sử dụng rộng rãi trong xử lý tiếng nói . Các đặc trưng MFCC được tính toán bằng cách chia tín hiệu âm thanh thành các khung thời gian ngắn, sau đó áp dụng phép biến đổi Fourier nhanh (FFT) để tính toán phổ tần số của mỗi khung. Cuối cùng, các đặc trưng MFCC được tính toán từ phổ tần số bằng cách sử dụng bộ lọc Mel và phép biến đổi cepstrum.

Phân loại: Các đặc trưng MFCC được sử dụng để huấn luyện các mô hình học sâu để phân loại các tệp âm thanh. Hệ thống đề xuất sử dụng các thuật toán học sâu sau:

Đánh giá và lưu mô hình: Các mô hình học sâu được đánh giá dựa trên độ chính xác của chúng trên một tập dữ liệu kiểm tra. Mô hình có độ chính xác cao nhất được lưu lại để sử dụng cho việc dự đoán.

Dự đoán: Mô hình lưu được sử dụng để dự đoán giọng nói của người nói trong một tệp âm thanh mới.

1.4 Thực nghiệm

1.4.1 Dữ liệu

Đối với bài toán này, chúng tôi không tìm được dữ liệu sẵn có để giải quyết vấn đề trên, vậy nên phương pháp đề xuất của chúng tôi là tìm tập dữ liệu tiếng việt đã có sẵn và đánh label lại vùng miền cho các tập dữ liệu đó.

Chúng tôi sử dụng tập dữ liệu VIVOS là kho ngữ liệu tiếng Việt miễn phí bao gồm 15 giờ ghi âm giọng nói được chuẩn bị cho nhiệm vụ Nhận dạng giọng nói tự động. Ngữ liệu được biên soạn bởi AILAB, phòng thí nghiệm khoa học máy tính của Đại học Khoa học Tự nhiên - ĐHQG-HCM, do GS. Vũ Hải Quân làm trưởng phòng. Nhóm đã sử dụng bộ dữ liệu này và label lại các tệp âm thanh theo vùng miền Việt Nam, tập dữ liệu của nhóm bao gồm 46 speaker với 6,297 tệp âm thanh.

1.4.2 Xử lý dữ liệu

Đối với tập dữ liệu này, một lợi ích là không có nhiễu vậy nên nhóm không cần phải xử lý dữ liệu nhiễu, chỉ cần xử lý mất cân bằng dữ liệu và xử lý độ dài của một tệp âm thanh không quá dài

Đối với vấn đề mất cân bằng dữ liệu nhóm sử dụng phương pháp Under-sampling là một kỹ thuật xử lý dữ liệu được sử dụng để giải quyết vấn đề mất cân bằng dữ liệu trong học sâu. Under-sampling hoạt động bằng cách loại bỏ một số mẫu từ lớp đa số để giảm kích thước của nó và làm cho nó cân bằng hơn với các lớp thiểu số. Có hai phương pháp chính để thực hiện under-sampling: là loại bỏ ngẫu nhiên (Random Under-sampling) và loại bỏ dựa trên mật độ (Density-based Under-sampling) nhóm đã chọn phương pháp ngẫu nhiên.

Đối với vấn đề độ dài của một tệp âm thanh quá dài, nhóm đã chỉ đơn giản là cắt tệp âm thanh tối đa 10s và sau đó encoding data.

1.4.3 Công nghệ sử dụng

Đối với bài toán này nhóm đã sử dụng ngôn ngữ chính là Python, ngoài ra có các thư viện phụ trợ như Numpy, Pandas, matplotlib, librosa (để trích xuất đặc trưng với MFCC) và sklearn (để xây dựng các mô hình).

1.4.4 Cách đánh giá

Để đánh giá mô hình các đơn giản nhất là nhóm đã sử dụng độ đo Accuracy và F1 Score. Accuracy là một trong những độ đo phổ biến nhất cho các bài toán phân loại, thể hiện tỷ lệ phần trăm các trường hợp được dự đoán đúng so với tổng số trường hợp.

Công thức: *Accuracy = (Số lượng dự đoán đúng / Tổng số trường hợp) \* 100%*

Ưu điểm:

Dễ hiểu, dễ tính toán.

Phù hợp với các bài toán phân loại hai lớp (nhị phân).

Nhược điểm:

Có thể bị ảnh hưởng bởi tỷ lệ các lớp trong tập dữ liệu: Nếu một lớp có tỷ lệ cao hơn nhiều so với các lớp khác, mô hình có thể đạt độ chính xác cao mặc dù khả năng phân biệt các lớp còn yếu.

Không cung cấp thông tin chi tiết về hiệu suất của mô hình trên từng lớp: Ví dụ, mô hình có thể đạt độ chính xác cao nhưng lại dự đoán sai nhiều trường hợp thuộc lớp thiểu số.

F1 score là một thước đo thống kê được sử dụng để đánh giá hiệu suất của mô hình phân loại nhị phân. Nó là trung bình điều hòa giữa độ chính xác (precision) và độ bao quát (recall), hai thước đo quan trọng để đánh giá hiệu quả của mô hình.

Công thức tính F1 score:

F1 = 2 \* (precision \* recall) / (precision + recall)

Trong đó:

precision là tỷ lệ các dự đoán dương tính chính xác so với tổng số dự đoán dương tính.

recall là tỷ lệ các mẫu dương tính thực sự được dự đoán chính xác so với tổng số mẫu dương tính thực sự.

Giải thích:

F1 score nằm trong khoảng từ 0 đến 1, với 1 là giá trị tốt nhất.

Giá trị F1 score cao cho thấy mô hình có cả độ chính xác cao và độ bao quát cao.

## 1.5 Kết quả đạt được

A screenshot of a computer

Description automatically generated

*Hình 2. Kết quả đạt được*

1.6 Kết luận

Kết quả nghiên cứu đã chứng minh hiệu quả ứng dụng của các thuật toán Deep Learning trong xử lý giọng nói, với độ chính xác đạt gần 97%. Điều này thể hiện tiềm năng lớn của Deep Learning trong lĩnh vực này. Tuy nhiên, các thuật toán Deep Learning hiện tại vẫn còn một số hạn chế, đặc biệt khi dữ liệu huấn luyện còn thiếu. Do đó, nghiên cứu cần tiếp tục được đẩy mạnh để mở rộng bộ dữ liệu, từ đó cải thiện độ chính xác.

Để tiến xa hơn, cần tập trung vào các hướng nghiên cứu sau:

* Khám phá các phương pháp mới để xử lý giọng nói phức tạp, chẳng hạn như tiếng ồn nền và giọng nói đa ngôn ngữ.
* Phát triển các hệ thống xử lý giọng nói hiệu quả và thân thiện hơn với người dùng.

Để cải thiện kết luận, có thể:

* Cung cấp thông tin chi tiết về các thuật toán Deep Learning được sử dụng và kết quả cụ thể của từng thuật toán.
* Phân tích kỹ lưỡng hơn những hạn chế của các thuật toán Deep Learning hiện tại trong xử lý giọng nói.
* Đề xuất rõ ràng các hướng nghiên cứu tiềm năng trong tương lai.

# TÀI LIỆU THAM KHẢO

1. Othman O. Khalifa1\* , M. I. Alhamada, Aisha H. Abdalla (2020), “Emotion Speech Recognition using Deep Learning”, *Article in Majlesi Journal of Electrical Engineering*.
2. S. G. Shaila(B) , A. Sindhu, L. Monish, D. Shivamma, and B. Vaishali (2023), “Speech Emotion Recognition Using Machine Learning Approach”, ICAMIDA 2022, ACSR 105, pp. 592–599, 2023.
3. Prof. Vu Hai Quan (2016), “VIVOS: Vietnamese Speech Corpus for ASR”, WLSI/OIAF4HLT2016.
4. ZRAR KH. ABDUL 1,2 AND ABDULBASIT K. AL-TALABANI 3 (2022), “Mel Frequency Cepstral Coefficient and Its Applications: A Review”, IEEE
5. Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi (2010), *“*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques” *,* JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617.