TRƯỜNG ĐẠI HỌC LÂM NGHIỆP

**KHOA CƠ ĐIỆN VÀ CÔNG TRÌNH**

**----------o0o----------**



**BÁO CÁO TIỂU LUẬN**

**MÔN: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: “NHẬN DIỆN GIỌNG NÓI SỬ DỤNG TEACHABLE MACHINE”**

|  |  |
| --- | --- |
| ***Giảng viên hướng dẫn***  ***Lớp K66-HTTT*** | **: ThS. Mai Hà An**  **:** |

**HÀ NỘI - 2024**

# LỜI MỞ ĐẦU

Trong thời đại công nghệ số phát triển mạnh mẽ như hiện nay, việc tương tác giữa con người và máy tính ngày càng trở nên phổ biến và đa dạng. Một trong những phương thức tương tác hiệu quả và tự nhiên nhất chính là thông qua giọng nói. Nhận diện giọng nói - công nghệ giúp máy tính hiểu và xử lý ngôn ngữ nói của con người - đã trở thành nền tảng cho nhiều ứng dụng hiện đại như trợ lý ảo, hệ thống tổng đài thông minh, công nghệ chuyển văn bản thành lời nói,...

Với tiềm năng to lớn cùng sự phát triển mạnh mẽ của các thuật toán học máy (Machine Learning), đề tài "Nhận diện giọng nói sử dụng Machine Learning" là một hướng đi đầy triển vọng trong lĩnh vực khai phá dữ liệu. Báo cáo này sẽ trình bày chi tiết quá trình xây dựng một hệ thống nhận diện giọng nói dựa trên các thuật toán học máy, từ giai đoạn thu thập dữ liệu, tiền xử lý, đến huấn luyện và đánh giá mô hình.

Contents

[LỜI MỞ ĐẦU 2](#_Toc195220297)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 4](#_Toc195220298)

[1.1. Lý do chọn đề tài 4](#_Toc195220299)

[1.2. Mục tiêu 4](#_Toc195220300)

[1.3. Phạm vi nghiên cứu 5](#_Toc195220301)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 6](#_Toc195220302)

[2.1. Tổng quan về khai phá dữ liệu 6](#_Toc195220303)

[2.2. Tổng quan về nhận diện giọng nói 7](#_Toc195220304)

[2.3. Machine Learning là gì? 8](#_Toc195220305)

[2.4. Các kỹ thuật trong nhận diện giọng nói 10](#_Toc195220306)

[Chương 3: Phân tích và thiết kế hệ thống 12](#_Toc195220307)

[3.1. Kiến trúc hệ thống tổng thể 12](#_Toc195220308)

[3.2. Thu thập và tiền xử lý dữ liệu âm thanh 12](#_Toc195220309)

[3.3. Trích xuất đặc trưng âm thanh 13](#_Toc195220310)

[3.4. Thiết kế mô hình nhận diện 13](#_Toc195220311)

[3.5. Giao diện huấn luyện và kiểm thử 14](#_Toc195220312)

[Chương 4: Triển khai mô hình 14](#_Toc195220313)

[4.1. Công cụ và nền tảng sử dụng 14](#_Toc195220314)

[4.2. Quá trình huấn luyện mô hình 14](#_Toc195220315)

[4.3. Đánh giá mô hình 17](#_Toc195220316)

[Chương 5: Ứng dụng và đánh giá 17](#_Toc195220317)

[5.1 Ứng dụng thực tiễn 17](#_Toc195220318)

[5.2 Ưu nhược điểm 17](#_Toc195220319)

[5.3 Hướng phát triển 18](#_Toc195220320)

[Kết luận 18](#_Toc195220321)

[Tài liệu tham khảo 18](#_Toc195220322)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## 1.1. Lý do chọn đề tài

Trong xu hướng phát triển của công nghệ hiện đại, việc giao tiếp giữa người và máy tính thông qua giọng nói đang ngày càng trở nên phổ biến. Các hệ thống như Siri, Google Assistant, Alexa,... đều ứng dụng công nghệ nhận diện giọng nói nhằm mang đến trải nghiệm tiện lợi và thân thiện hơn với người dùng. Tuy nhiên, việc xây dựng hệ thống này đòi hỏi sự kết hợp của nhiều kỹ thuật trong lĩnh vực trí tuệ nhân tạo và khai phá dữ liệu. Chính vì vậy, việc nghiên cứu và ứng dụng Machine Learning trong nhận diện giọng nói là điều cần thiết và mang lại nhiều giá trị thực tiễn.

## 1.2. Mục tiêu

* **Tìm hiểu các kỹ thuật tiền xử lý và trích xuất đặc trưng trong nhận diện giọng nói**

Âm thanh thô (raw audio) là dữ liệu phi cấu trúc và thường chứa nhiễu, biến động tần số, và các yếu tố môi trường khác. Do đó, mục tiêu đầu tiên là nghiên cứu và áp dụng các phương pháp tiền xử lý dữ liệu âm thanh như chuẩn hóa tín hiệu, loại bỏ nhiễu, và cắt đoạn âm thanh hợp lý. Sau đó, trích xuất đặc trưng là bước quan trọng để chuyển đổi tín hiệu âm thanh thành dạng đặc trưng (features) mà mô hình có thể hiểu được. Các kỹ thuật được tìm hiểu bao gồm MFCC (Mel-Frequency Cepstral Coefficients), Spectrogram và Chroma Features.

* **Xây dựng mô hình nhận diện giọng nói sử dụng thuật toán học máy truyền thống**

Mục tiêu tiếp theo là áp dụng các thuật toán Machine Learning truyền thống như K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, hoặc Random Forest để huấn luyện mô hình có khả năng nhận diện giọng nói dựa trên các đặc trưng đã trích xuất. Các bước thực hiện bao gồm: phân chia tập dữ liệu, huấn luyện mô hình, tinh chỉnh tham số, và đánh giá hiệu năng.

* **Đánh giá độ chính xác và hiệu quả của mô hình**

Cuối cùng, mô hình được đánh giá dựa trên các tiêu chí như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision), và F1-score. Mục tiêu là xác định mô hình nào hoạt động hiệu quả nhất trên tập dữ liệu giọng nói đã chuẩn bị. Ngoài ra, hiệu năng xử lý (thời gian huấn luyện, tốc độ dự đoán) và khả năng mở rộng (scalability) cũng được xem xét để đưa ra nhận định toàn diện về mô hình.

## ****1.3. Phạm vi nghiên cứu****

Trong khuôn khổ đề tài, phạm vi nghiên cứu được giới hạn ở một số khía cạnh cụ thể nhằm đảm bảo tính khả thi và tập trung vào việc triển khai các mô hình học máy truyền thống cho bài toán nhận diện giọng nói tiếng Việt:

* **Dữ liệu**

Dữ liệu âm thanh được sử dụng trong đề tài chủ yếu là các tệp ghi âm tiếng Việt cơ bản, thu thập từ các nguồn dữ liệu mở như Common Voice (Mozilla), VietSpeech, hoặc một số bộ dữ liệu do cộng đồng phát triển. Các tệp âm thanh bao gồm các câu ngắn, từ vựng hoặc cụm từ thường dùng, được gán nhãn theo nội dung lời nói hoặc danh tính người nói (tùy theo mục tiêu nhận diện). Chất lượng dữ liệu được đảm bảo ở mức trung bình đến tốt, với các thao tác tiền xử lý để loại bỏ nhiễu và chuẩn hóa tín hiệu.

* **Mô hình**

Đề tài tập trung vào việc sử dụng các mô hình học máy truyền thống thay vì các mô hình học sâu (deep learning) phức tạp. Cụ thể, các thuật toán được áp dụng bao gồm:

* + **K-Nearest Neighbors (KNN)**: đơn giản, hiệu quả trong các bài toán phân loại nhỏ.
  + **Support Vector Machine (SVM)**: mạnh trong phân loại nhị phân với biên tách rõ ràng.
  + **Random Forest**: một mô hình cây quyết định tổ hợp, cho hiệu năng tốt và ít bị overfitting. Các mô hình này được huấn luyện trên tập đặc trưng được trích xuất từ tín hiệu âm thanh, và được so sánh dựa trên độ chính xác và độ ổn định trong nhận diện.
* **Phần mềm và công cụ hỗ trợ**

Đề tài được triển khai hoàn toàn bằng **ngôn ngữ lập trình Python**, với sự hỗ trợ của các thư viện chuyên dụng:

* + **Librosa**: thư viện mạnh mẽ để xử lý âm thanh và trích xuất đặc trưng như MFCC, Spectrogram, Zero Crossing Rate, v.v.
  + **Scikit-learn**: thư viện học máy phổ biến, cung cấp đầy đủ các thuật toán và công cụ đánh giá mô hình.
  + Ngoài ra, các công cụ hỗ trợ trực quan hóa như Matplotlib, Seaborn cũng được sử dụng để biểu diễn kết quả đánh giá.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Tổng quan về khai phá dữ liệu

Khai phá dữ liệu (*Data Mining*) là một nhánh quan trọng trong lĩnh vực khoa học dữ liệu, tập trung vào việc tìm kiếm các mẫu, xu hướng hoặc thông tin có giá trị từ các tập dữ liệu lớn và phức tạp. Đây là một quá trình có hệ thống nhằm chuyển đổi dữ liệu thô thành tri thức hữu ích, phục vụ cho việc ra quyết định, dự đoán hoặc khám phá các mối quan hệ tiềm ẩn giữa các yếu tố trong dữ liệu.

Quá trình khai phá dữ liệu thường bao gồm các bước cơ bản sau:

* Thu thập dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp dữ liệu, dữ liệu thời gian thực, dữ liệu âm thanh, hình ảnh,...
* Tiền xử lý dữ liệu: Làm sạch và chuẩn hóa dữ liệu, loại bỏ các giá trị thiếu, xử lý nhiễu, chuẩn hóa định dạng dữ liệu để chuẩn bị cho quá trình phân tích.
* Trích xuất đặc trưng và biến đổi dữ liệu: Biến đổi dữ liệu thành dạng phù hợp cho thuật toán khai phá như rút trích đặc trưng, giảm chiều dữ liệu, mã hóa,...
* Áp dụng thuật toán khai phá: Sử dụng các kỹ thuật học máy như phân loại, phân cụm, hồi quy, luật kết hợp,… để phát hiện ra các mẫu hoặc mối quan hệ trong dữ liệu.
* Đánh giá và diễn giải kết quả: Kiểm tra độ chính xác, hiệu quả và ý nghĩa thực tiễn của kết quả khai phá.
* Trình bày thông tin: Hiển thị kết quả khai phá bằng các biểu đồ, bảng, mô hình trực quan,... để người dùng cuối dễ hiểu và sử dụng.

Khai phá dữ liệu được ứng dụng trong nhiều lĩnh vực thực tiễn, ví dụ:

* Tài chính: Phát hiện gian lận, dự đoán rủi ro tín dụng, phân tích hành vi đầu tư.
* Y tế: Hỗ trợ chẩn đoán bệnh, phân tích gen, theo dõi tiến trình điều trị.
* Thương mại điện tử: Gợi ý sản phẩm, phân tích hành vi khách hàng, tối ưu chiến lược bán hàng.
* Giáo dục: Theo dõi tiến độ học tập, cá nhân hóa phương pháp giảng dạy.
* An ninh: Nhận diện khuôn mặt, phát hiện hành vi đáng ngờ, giám sát video.

## 2.2. Tổng quan về nhận diện giọng nói

Nhận diện giọng nói (*Speech Recognition*) là một lĩnh vực trong trí tuệ nhân tạo (AI) và xử lý tín hiệu, cho phép máy tính hoặc hệ thống tự động nhận dạng và phiên âm lời nói của con người thành văn bản. Công nghệ này đóng vai trò quan trọng trong việc giao tiếp giữa người và máy, mở ra nhiều ứng dụng trong cuộc sống như trợ lý ảo (Google Assistant, Siri), nhập liệu bằng giọng nói, điều khiển thiết bị thông minh và dịch vụ khách hàng tự động.

Mục tiêu của nhận diện giọng nói

Mục tiêu chính của hệ thống nhận diện giọng nói là phân tích tín hiệu âm thanh đầu vào, sau đó trích xuất đặc trưng quan trọng để từ đó suy đoán ra chuỗi từ hoặc câu tương ứng. Quá trình này phải xử lý được các yếu tố phức tạp như:

* Biến thể ngữ điệu, tốc độ nói
* Tiếng ồn nền
* Giọng nói của nhiều người khác nhau (speaker-independent)
* Các ngôn ngữ và phương ngữ khác nhau

Phân loại nhận diện giọng nói

Nhận diện giọng nói có thể được phân loại thành hai dạng phổ biến sau:

1. Nhận diện từ khóa (Keyword Spotting)

Đây là hình thức nhận diện các từ hoặc cụm từ cụ thể trong tín hiệu giọng nói. Thường được sử dụng trong các hệ thống trợ lý ảo hoặc điều khiển bằng giọng nói, nơi người dùng cần nói các từ khóa như "OK Google", "Hey Siri" hoặc các lệnh đơn giản như "Bật đèn", "Mở nhạc".

* Ưu điểm: Đơn giản, tốc độ xử lý nhanh, yêu cầu ít tài nguyên.
* Nhược điểm: Giới hạn phạm vi nhận diện, không nhận diện được ngôn ngữ tự nhiên phức tạp.

2. Nhận diện câu hoàn chỉnh (Full Speech-to-Text)

Đây là hình thức nhận diện toàn bộ nội dung lời nói liên tục và chuyển đổi thành văn bản hoàn chỉnh. Đây là công nghệ được ứng dụng trong ghi âm cuộc họp, phụ đề tự động, dịch thuật, và nhiều ứng dụng xử lý ngôn ngữ tự nhiên khác.

* Ưu điểm: Độ bao phủ rộng, có khả năng xử lý ngôn ngữ tự nhiên.
* Nhược điểm: Yêu cầu mô hình phức tạp, cần nhiều dữ liệu huấn luyện và tài nguyên tính toán lớn.

Các yếu tố ảnh hưởng đến độ chính xác

Một hệ thống nhận diện giọng nói hiệu quả cần chú ý đến nhiều yếu tố như:

* Chất lượng âm thanh: Tín hiệu bị nhiễu sẽ làm giảm độ chính xác nhận diện.
* Ngữ cảnh ngôn ngữ: Việc hiểu nghĩa của từ trong câu giúp tăng khả năng nhận diện chính xác.
* Giọng nói người dùng: Khác biệt giọng vùng miền, độ tuổi, giới tính, tốc độ nói,...
* Thuật toán và mô hình sử dụng: Các thuật toán học máy, đặc biệt là học sâu (deep learning), ngày càng được áp dụng để cải thiện hiệu suất nhận diện.

## 2.3. Machine Learning là gì?

**Machine Learning (Học máy)** là một lĩnh vực thuộc **Trí tuệ nhân tạo (Artificial Intelligence - AI)**, tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính học hỏi từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần được lập trình tường minh cho từng tác vụ cụ thể.

Thay vì viết mã theo cách truyền thống để xử lý từng trường hợp riêng biệt, trong học máy, các mô hình sẽ **tự động cải thiện hiệu suất** thông qua việc **học từ dữ liệu quá khứ**. Điều này đặc biệt hữu ích trong các lĩnh vực như nhận diện giọng nói, xử lý ngôn ngữ tự nhiên, thị giác máy tính, phát hiện gian lận, và nhiều ứng dụng khác.

**Quy trình học máy cơ bản**

Một hệ thống học máy thông thường sẽ trải qua các bước chính sau:

1. **Thu thập dữ liệu**
2. **Tiền xử lý dữ liệu** (chuẩn hóa, làm sạch, trích xuất đặc trưng,…)
3. **Chia tập dữ liệu** (train/test)
4. **Lựa chọn mô hình phù hợp**
5. **Huấn luyện mô hình**
6. **Đánh giá và tinh chỉnh**
7. **Dự đoán/ứng dụng thực tế**

**Phân loại các mô hình học máy**

Học máy có thể được chia thành 3 loại chính.

**1. Học có giám sát (Supervised Learning)**

Đây là loại học máy phổ biến nhất. Dữ liệu huấn luyện bao gồm cả **đầu vào (input)** và **nhãn đầu ra (label)**. Mục tiêu là để mô hình học cách ánh xạ từ đầu vào sang đầu ra đúng.

* **Ví dụ**: Nhận diện giọng nói → Đầu vào: âm thanh; Đầu ra: văn bản tương ứng.
* **Thuật toán tiêu biểu**: KNN, SVM, Decision Tree, Random Forest, Logistic Regression,…

**Ưu điểm**:

* Dễ đánh giá độ chính xác.
* Ứng dụng hiệu quả trong phân loại và hồi quy.

**Nhược điểm**:

* Cần dữ liệu gắn nhãn, tốn chi phí.

**2. Học không giám sát (Unsupervised Learning)**

Trong loại này, dữ liệu không có nhãn. Mục tiêu là tìm ra cấu trúc ẩn hoặc mẫu tiềm ẩn trong dữ liệu.

* **Ví dụ**: Nhóm các đoạn âm thanh có đặc điểm giống nhau (clustering).
* **Thuật toán tiêu biểu**: K-means, PCA, Hierarchical Clustering,…

**Ưu điểm**:

* Áp dụng cho dữ liệu chưa được gắn nhãn.
* Tốt cho khám phá dữ liệu và giảm chiều dữ liệu.

**Nhược điểm**:

* Khó đánh giá chính xác kết quả.
* Dễ bị ảnh hưởng bởi nhiễu và phân phối dữ liệu.

**3. Học tăng cường (Reinforcement Learning)**

Loại học máy này dựa trên nguyên lý **thử và sai**. Một **agent (tác nhân)** sẽ học cách hành động trong môi trường sao cho đạt được **phần thưởng tối đa** theo thời gian.

* **Ví dụ**: Robot học cách di chuyển, AI chơi cờ vua hoặc game.
* **Thuật toán tiêu biểu**: Q-Learning, Deep Q-Network (DQN), Policy Gradient,…

**Ưu điểm**:

* Phù hợp với các hệ thống ra quyết định tự động.
* Có khả năng thích nghi trong môi trường động.

**Nhược điểm**:

* Cần thời gian huấn luyện dài.
* Môi trường huấn luyện phức tạp.

**Ứng dụng của học máy trong nhận diện giọng nói**

Trong đề tài này, học máy được áp dụng để **huấn luyện mô hình nhận diện giọng nói tiếng Việt**, sử dụng các đặc trưng âm thanh như MFCC, Spectrogram,... Kết quả huấn luyện giúp mô hình có thể phân biệt và dự đoán được nội dung lời nói dựa trên tín hiệu âm thanh đầu vào.

## 2.4. Các kỹ thuật trong nhận diện giọng nói

**1. Tiền xử lý âm thanh (Audio Preprocessing)**

Trước khi đưa dữ liệu âm thanh vào mô hình học máy, dữ liệu cần được xử lý để đảm bảo chất lượng và độ đồng nhất. Một số bước tiền xử lý phổ biến bao gồm:

* **Lọc nhiễu (Noise Reduction):** Giảm tạp âm không mong muốn trong tín hiệu âm thanh để tăng độ rõ của giọng nói. Các kỹ thuật như spectral gating, Wiener filtering thường được sử dụng.
* **Chuẩn hóa âm lượng (Volume Normalization):** Đảm bảo mức âm lượng giữa các tệp âm thanh là nhất quán để mô hình không bị ảnh hưởng bởi sự chênh lệch cường độ.
* **Chuyển đổi mẫu (Resampling):** Đồng bộ tần số lấy mẫu (sample rate), ví dụ như 16kHz hoặc 44.1kHz, để thống nhất định dạng âm thanh.
* **Cắt/Chia đoạn âm thanh:** Tách đoạn có chứa giọng nói (voice activity detection) để loại bỏ khoảng lặng không cần thiết.

**2. Trích xuất đặc trưng (Feature Extraction)**

Sau khi dữ liệu âm thanh được tiền xử lý, bước tiếp theo là trích xuất các đặc trưng có ý nghĩa để mô hình học máy có thể học được các mẫu trong giọng nói. Một số kỹ thuật phổ biến:

* **MFCC (Mel-Frequency Cepstral Coefficients):** Là đặc trưng phổ biến nhất trong nhận diện giọng nói. MFCC mô phỏng cách tai người cảm nhận tần số, giúp biểu diễn các thông tin quan trọng trong tín hiệu giọng nói.
* **Spectrogram:** Là biểu diễn trực quan của tín hiệu âm thanh theo thời gian và tần số, giúp mô hình nhìn thấy cấu trúc của âm thanh dưới dạng hình ảnh (dùng nhiều trong Deep Learning).
* **Chroma Features:** Đại diện cho 12 lớp âm sắc trong một quãng tám, thường dùng cho nhận diện nhạc, nhưng đôi khi áp dụng cho giọng nói để phân biệt cao độ.
* **Zero Crossing Rate, Energy, Pitch:** Một số đặc trưng đơn giản khác cũng được sử dụng để cung cấp thông tin về âm sắc và năng lượng của âm thanh.

**3. Phân loại (Classification)**

Sau khi có đặc trưng đầu vào, bước tiếp theo là áp dụng các mô hình học máy để phân loại hoặc dự đoán nội dung lời nói. Một số thuật toán phân loại truyền thống phổ biến trong nhận diện giọng nói bao gồm:

* **K-Nearest Neighbors (KNN):** Dự đoán đầu ra dựa trên các điểm dữ liệu gần nhất trong không gian đặc trưng. Phù hợp với bài toán phân loại cơ bản, nhưng không tối ưu với dữ liệu lớn.
* **Support Vector Machine (SVM):** Mô hình mạnh mẽ trong việc phân chia các lớp bằng siêu phẳng tối ưu. SVM hoạt động hiệu quả với dữ liệu có số chiều cao như MFCC.
* **Decision Tree và Random Forest:** Dễ hiểu, dễ triển khai và có thể xử lý tốt dữ liệu không tuyến tính. Random Forest cải thiện hiệu suất bằng cách kết hợp nhiều cây quyết định.
* **Logistic Regression:** Phù hợp với bài toán phân loại nhị phân. Đơn giản và hiệu quả với các bộ dữ liệu nhỏ.

# Chương 3: Phân tích và thiết kế hệ thống

## 3.1. Kiến trúc hệ thống tổng thể

Hệ thống nhận diện giọng nói được thiết kế theo quy trình gồm các bước chính như sau:

1. **Thu thập dữ liệu âm thanh**: Ghi âm các câu lệnh hoặc từ khóa.
2. **Tiền xử lý dữ liệu**: Chuẩn hóa âm thanh, lọc nhiễu, tách từ.
3. **Trích xuất đặc trưng**: Chuyển âm thanh thành các đặc trưng như MFCC hoặc spectrogram.
4. **Huấn luyện mô hình**: Sử dụng Teachable Machine để tạo mô hình học máy nhận diện giọng nói.
5. **Đánh giá và kiểm thử**: Kiểm tra độ chính xác của mô hình bằng mẫu giọng nói mới.
6. **Triển khai mô hình**: Xuất mô hình và tích hợp vào ứng dụng (Web hoặc Android).

Sơ đồ khối minh họa kiến trúc hệ thống:

[Thu âm] → [Tiền xử lý] → [Trích xuất đặc trưng] → [Huấn luyện mô hình] → [Dự đoán & Đánh giá] → [Ứng dụng thực tiễn]

## 3.2. Thu thập và tiền xử lý dữ liệu âm thanh

* **Thu thập dữ liệu**:
  + Sử dụng micro hoặc thiết bị ghi âm để thu các câu lệnh đơn giản bằng tiếng Việt.
  + Ví dụ: “Bắt đầu”, “Tạm dừng”, “Kết thúc”, “Xin chào”...
  + Dữ liệu có thể tự thu hoặc lấy từ các nguồn mở như Mozilla Common Voice.
* **Tiền xử lý dữ liệu**:
  + **Chuẩn hóa**: Điều chỉnh tần số lấy mẫu về 16kHz hoặc 22kHz.
  + **Lọc nhiễu**: Loại bỏ tiếng ồn nền.
  + **Cắt âm thừa**: Bỏ đoạn im lặng đầu/cuối mỗi tệp âm thanh.
  + Dữ liệu sau xử lý được lưu ở định dạng WAV hoặc MP3, dùng cho huấn luyện.

## 3.3. Trích xuất đặc trưng âm thanh

Các đặc trưng phổ biến được trích xuất để đưa vào mô hình huấn luyện bao gồm:

* **MFCC (Mel-Frequency Cepstral Coefficients)**:
  + Biểu diễn đặc trưng âm thanh theo cách gần giống với cách con người nghe.
  + Thường dùng với mô hình truyền thống như SVM, KNN.
* **Spectrogram**:
  + Biểu diễn dạng hình ảnh phổ tần số theo thời gian.
  + Thích hợp cho các mô hình CNN hoặc sử dụng trực tiếp trong Teachable Machine.
* **Chroma Features, Zero Crossing Rate** (tùy chọn): Thêm các đặc trưng nâng cao.

## 3.4. Thiết kế mô hình nhận diện

Dựa vào công cụ Teachable Machine, mô hình được thiết kế đơn giản với các lớp phân loại tương ứng với từng từ/câu lệnh.

* **Input**: Tín hiệu âm thanh đã xử lý.
* **Feature Extraction**: Tự động thực hiện trong nền tảng Teachable Machine.
* **Classification Layer**: Mô hình học máy phân loại âm thanh thành các nhãn tương ứng.
* **Output**: Tên của từ/câu lệnh được nhận diện.

## 3.5. Giao diện huấn luyện và kiểm thử

Teachable Machine hỗ trợ trực quan hóa toàn bộ quá trình:

* Tạo các class (ví dụ: "Xin chào", "Dừng lại"...).
* Giao diện kéo-thả để thêm hoặc ghi âm dữ liệu.
* Hiển thị trực tiếp kết quả dự đoán theo thời gian thực.

# Chương 4: Triển khai mô hình

## 4.1. Công cụ và nền tảng sử dụng

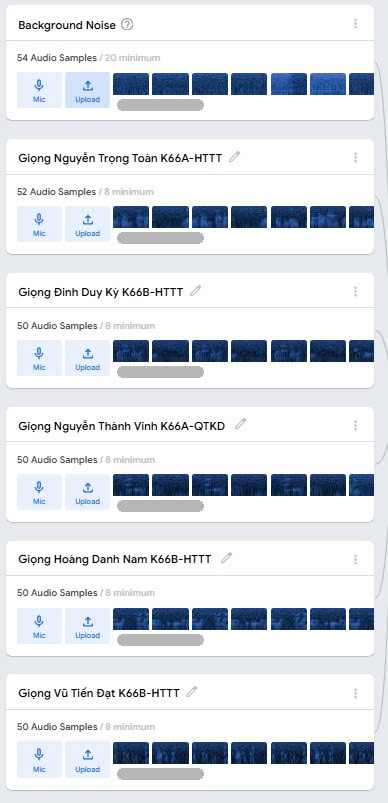
Để triển khai mô hình nhận diện giọng nói một cách trực quan, dễ sử dụng và không yêu cầu viết quá nhiều mã, đề tài sử dụng nền tảng Teachable Machine của Google. Đây là một công cụ trực tuyến cho phép người dùng huấn luyện các mô hình học máy một cách dễ dàng qua giao diện đồ họa.

**Các công cụ chính:**

* Teachable Machine: Nền tảng trực tuyến hỗ trợ tạo mô hình học máy từ dữ liệu âm thanh, hình ảnh hoặc tư thế cơ thể.
* Trình duyệt Chrome: Được khuyến nghị để thu âm, huấn luyện và xuất mô hình từ Teachable Machine.
* Visual Studio Code / Android Studio (tùy chọn): Dùng để tích hợp mô hình vào ứng dụng.

## 4.2. Quá trình huấn luyện mô hình

1. **Chuẩn bị dữ liệu âm thanh:**
   * Sử dụng micro máy tính để thu âm các câu nói đại diện cho các lớp cần phân loại (ví dụ: "Xin chào", "Tạm biệt", "Bắt đầu", "Dừng lại",...).
   * Mỗi lớp nên có ít nhất 20–30 mẫu để đảm bảo tính đa dạng và khả năng tổng quát.
2. **Tải dữ liệu vào Teachable Machine:**
   * Chọn dự án Audio Project.
   * Tạo các nhãn (class) tương ứng với từng loại giọng nói cần nhận diện.
   * Thu trực tiếp hoặc tải lên các tệp âm thanh mẫu.



**Hình ảnh minh họa**

1. **Huấn luyện mô hình:**
   * Nhấn “Train Model” để hệ thống bắt đầu học đặc trưng âm thanh.
   * Có thể tùy chỉnh các thông số như số epoch, learning rate,...
   * Sau khi huấn luyện xong, hệ thống sẽ cung cấp một mô hình có thể nhận diện được từng lớp giọng nói.

A graph of loss and loss

AI-generated content may be incorrect.

**Hình ảnh minh họa**

1. **Kiểm tra mô hình trực tiếp:**
   * Teachable Machine cung cấp giao diện test ngay trên trình duyệt.
   * Người dùng nói vào micro để xem mô hình có nhận diện đúng không.
2. **Xuất mô hình:**
   * Mô hình có thể xuất theo hai định dạng:
     + TensorFlow Lite: Dành cho tích hợp vào ứng dụng Android.
     + Web Model: Tích hợp trực tiếp vào các ứng dụng web.
   * Có thể tải xuống mã mẫu để sử dụng cùng với các thư viện như TensorFlow.js hoặc TensorFlow Lite trên Android.

## 4.3. Đánh giá mô hình

* **Phương pháp đánh giá:**
  + Sử dụng kết quả test trực tiếp trong trình duyệt.
  + Kiểm tra độ chính xác (accuracy) hiển thị khi chạy thử từng lớp.
  + Với các dự án tích hợp sâu hơn, có thể đánh giá thêm bằng confusion matrix, F1-score nếu sử dụng lại mô hình xuất ra với Python hoặc Android.
* **Nhận xét sơ bộ:**
  + Teachable Machine có thể huấn luyện nhanh, dễ dùng với dữ liệu nhỏ và vừa.
  + Không yêu cầu kiến thức lập trình chuyên sâu.
  + Tuy nhiên, độ chính xác có thể giảm nếu môi trường âm thanh quá nhiễu hoặc lớp dữ liệu không đa dạng.

# Chương 5: Ứng dụng và đánh giá

## 5.1 Ứng dụng thực tiễn

* Trợ lý ảo (Google Assistant, Siri, Alexa).
* Gọi điện bằng giọng nói.
* Nhập liệu văn bản bằng giọng nói.
* Điều khiển thiết bị IoT bằng giọng nói.

## 5.2 Ưu nhược điểm

* **Ưu điểm**:
  + Giao tiếp tự nhiên, tiện lợi.
  + Hữu ích cho người khuyết tật.
  + Tích hợp tốt trong các hệ thống thông minh.
* **Nhược điểm**:
  + Dễ bị nhiễu bởi tạp âm.
  + Phụ thuộc vào ngôn ngữ và giọng địa phương.
  + Cần nhiều dữ liệu để huấn luyện mô hình tốt.

## 5.3 Hướng phát triển

* Tăng cường khả năng nhận diện đa ngôn ngữ.
* Cải thiện độ chính xác với tiếng ồn lớn.
* Ứng dụng mô hình học sâu tiên tiến như Transformers (Wav2Vec, Whisper).

# Kết luận

Trong báo cáo này, chúng tôi đã trình bày quá trình xây dựng hệ thống nhận diện giọng nói sử dụng kỹ thuật khai phá dữ liệu và Machine Learning. Hệ thống trải qua các bước từ thu thập dữ liệu, tiền xử lý, trích xuất đặc trưng, huấn luyện mô hình đến đánh giá kết quả. Kết quả thực nghiệm cho thấy mô hình RNN (LSTM) đạt hiệu quả cao nhất. Tuy nhiên, hệ thống còn một số hạn chế như độ nhạy với tiếng ồn và sự phụ thuộc vào giọng nói người dùng.

Hướng phát triển trong tương lai sẽ tập trung vào cải tiến độ chính xác, khả năng nhận diện nhiều ngôn ngữ và tích hợp vào các ứng dụng thực tế.

# Tài liệu tham khảo

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
3. [https://librosa.org](https://librosa.org/)
4. [https://scikit-learn.org](https://scikit-learn.org/)
5. [https://www.tensorflow.org](https://www.tensorflow.org/)
6. [https://paperswithcode.com](https://paperswithcode.com/)