VOICE ASSISTANT

# OVERVIEW

Trợ lý giọng nói là một hệ thống có khả năng tiếp nhận, xử lý và phản hồi thông tin bằng giọng nói. Trên nền tảng ESP32, trợ lý giọng nói có thể giúp người dùng điều khiển thiết bị thông minh, thực hiện tìm kiếm thông tin hoặc phản hồi theo yêu cầu. Để triển khai thành công hệ thống này, cần nắm vững các khái niệm về xử lý tín hiệu giọng nói, nhận diện giọng nói, xử lý ngôn ngữ tự nhiên và tổng hợp giọng nói.

## 1 Yêu cầu của chủ đề :

1.1 Xử lý tín hiệu giọng nói (Speech Processing) :

Một trong những kỹ thuật quan trọng trong xử lý tín hiệu giọng nói là trích xuất đặc trưng Mel-Frequency Cepstral Coefficients (MFCC). MFCC giúp chuyển đổi tín hiệu giọng nói thành một tập hợp đặc trưng phản ánh tần số quan trọng của âm thanh, hỗ trợ tốt cho các thuật toán nhận diện giọng nói.

1.2 Nhận diện giọng nói (ASR):

ASR chuyển đổi lời nói thành văn bản bằng cách sử dụng các mô hình học máy. Quá trình này bao gồm trích xuất đặc trưng, mô hình âm học và mô hình ngôn ngữ. Các mô hình ASR nhẹ như TensorFlow Lite hoặc PocketSphinx thường được sử dụng trong các ứng dụng nhúng như ESP32.

1.3 Xử lý ngôn ngữ tự nhiên (NLP)

NLP giúp trợ lý hiểu và diễn giải các lệnh của người dùng. Quá trình này bao gồm phân tách từ, gốc hóa từ và nhận dạng ý định bằng cách sử dụng học máy hoặc phương pháp dựa trên quy tắc. Các mô hình NLP nhẹ hoặc dịch vụ NLP dựa trên đám mây có thể được tích hợp để cải thiện hiệu suất.

1.4 Chuyển văn bản thành giọng nói (TTS)

TTS chuyển đổi văn bản phản hồi thành giọng nói tự nhiên. Các kỹ thuật bao gồm tổng hợp dạng ghép, tổng hợp tần số và mô hình học sâu như Tacotron và WaveNet. Trên ESP32, thường sử dụng các công cụ TTS đơn giản hoặc giải pháp dựa trên đám mây.

## 2 Nghiên cứu các kỹ thuật/giải pháp hiện có và các sản phẩm thương mại

2.1 Các kỹ thuật và giải pháp hiện có

Các hệ thống nhận dạng giọng nói đã phát triển mạnh mẽ, từ phương pháp truyền thống dựa trên Hidden Markov Models (HMM) đến các mô hình học sâu tiên tiến như Transformer. Một số kỹ thuật nổi bật bao gồm:

* DeepSpeech: Được phát triển bởi Mozilla, đây là một hệ thống nhận dạng giọng nói mã nguồn mở dựa trên kiến trúc mạng nơ-ron hồi tiếp (RNN), có khả năng chuyển đổi giọng nói thành văn bản với độ chính xác cao.
* Wav2Vec: Một mô hình học tự giám sát do Facebook AI phát triển, cho phép nhận dạng giọng nói hiệu quả ngay cả khi dữ liệu huấn luyện bị giới hạn.
* Edge AI: Các mô hình nhận dạng giọng nói tối ưu hóa cho thiết bị biên, như TensorFlow Lite, giúp giảm độ trễ trong xử lý và đảm bảo quyền riêng tư cao hơn khi xử lý dữ liệu trực tiếp trên thiết bị.

2.2 Các sản phẩm thương mại liên quan

Nhiều sản phẩm thương mại đã tích hợp công nghệ nhận dạng giọng nói, mang lại trải nghiệm người dùng tiên tiến. Một số sản phẩm phổ biến bao gồm:

* Amazon Alexa: Trợ lý giọng nói của Amazon, sử dụng công nghệ xử lý ngôn ngữ tự nhiên (NLP) và học sâu để hiểu và phản hồi các lệnh thoại của người dùng.
* Google Assistant: Được tích hợp trên nhiều thiết bị Android và IoT, hỗ trợ nhận dạng giọng nói với nhiều ngôn ngữ và khả năng tương tác với hệ sinh thái Google.
* Apple Siri: Trợ lý giọng nói của Apple, tích hợp sâu với hệ sinh thái iOS, cho phép thực hiện các tác vụ như gửi tin nhắn, tìm kiếm thông tin, điều khiển thiết bị thông minh.
* Snips.ai: Một trợ lý giọng nói tập trung vào quyền riêng tư, hoạt động hoàn toàn trên thiết bị biên mà không cần gửi dữ liệu lên đám mây, giúp bảo vệ thông tin cá nhân của người dùng.

# PRELIMINARY SOLUTION