

Speaker Recognition using LSTM Implementation on Raspberry Pi 4 (Part 2)

I. Nêu lại vấn đề nhóm đang gặp phải tuần trước:

- Mô hình đưa ra kết quả chính xác khi triển khai trên laptop nhưng khi cài đặt trên Raspberry thì kết quả luôn mặc định trả về "Trần Đức Trí" với bất kỳ người nói nào.

⇒ Những câu hỏi nhóm cần phải trả lời:

- MFCC có trích xuất được các đặc trưng của tín hiệu âm thanh dựa trên bản chất của cách người nói tạo ra tiếng nói hay không?
- Nội dung nói có ảnh hưởng đến embedding vector mà MFCC trả về hay không?
- Vấn đề là nằm ở mặt **kỹ thuật** hay là sai sót về mặt **lý thuyết**?

II. Đi tìm hướng giải quyết vấn đề:

Nhóm đi tìm câu trả lời cho câu hỏi liệu rằng MFCC có thể trích xuất đặc trưng của người nói hay không thông qua paper: [MFCC and Prosodic Feature Extraction Techniques](#)

Nhóm rút ra được một số kết luận cơ bản từ paper trên:

- **MFCC được sử dụng rộng rãi trong bài toán Speaker recognition, kể cả những model SOTA**
- Theo paper đề xuất thì có cả phương pháp Prosodic để nhận dạng người nói nhưng hiệu quả kém xa so với MFCC

- Phần cứng, thiết bị ghi âm đóng vai trò cực kỳ quan trọng trong việc nhận dạng người nói, đặc biệt thiết bị phải đồng bộ ở cả tập train và tập test:

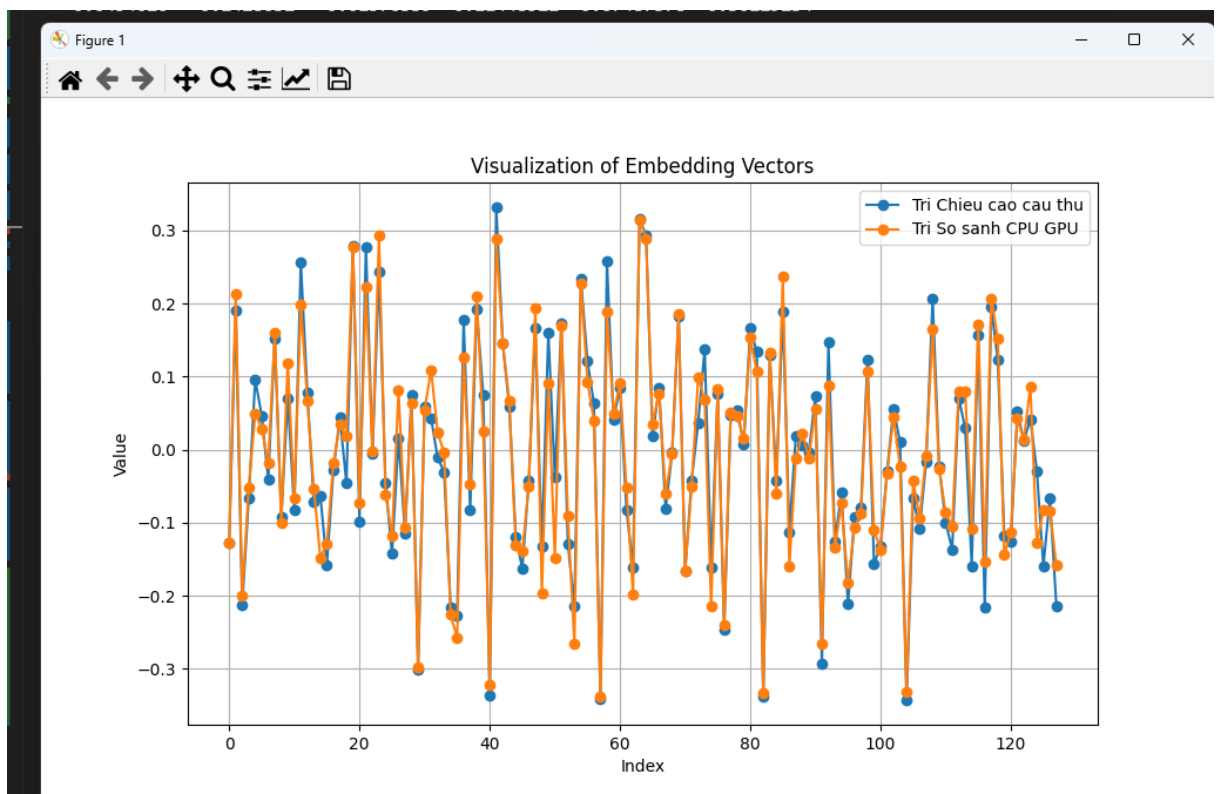
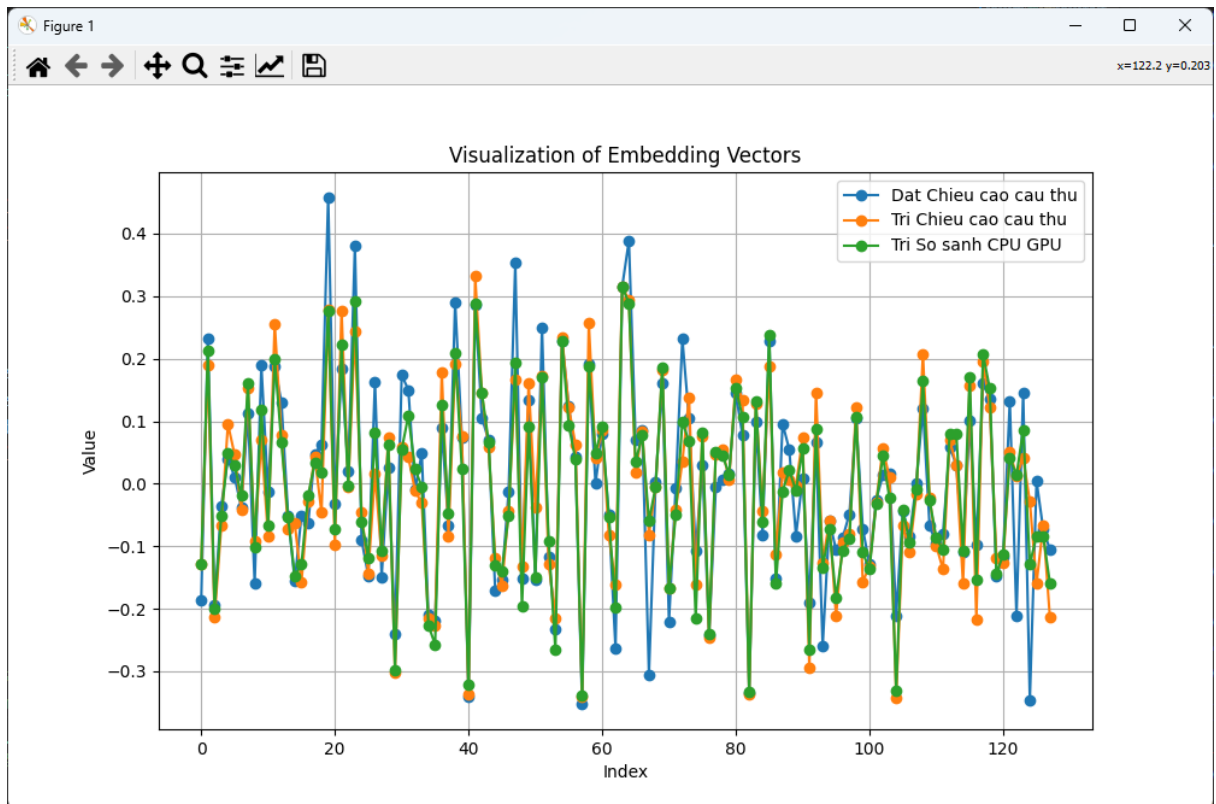
suất của quá trình nhận dạng người nói. Nghiên cứu cũng chỉ ra rằng có một số tham số khác ảnh hưởng đến hiệu suất của quá trình nhận dạng người nói, bao gồm:

- Ngôn ngữ được sử dụng trong dữ liệu huấn luyện và kiểm tra.
- Chất lượng của dữ liệu mẫu giọng nói trong quá trình huấn luyện và kiểm tra.
- Chất lượng của microphone và khoảng cách giữa microphone và người nói.
- Tiếng ồn trong quá trình ghi âm mẫu giọng nói và kiểm tra giọng nói.
- Độ dài của mẫu giọng nói được sử dụng trong quá trình huấn luyện và kiểm tra.
- Bảo vệ microphone hoặc loa vào thời gian huấn luyện và kiểm tra.
- Phụ thuộc vào văn bản, tức là dữ liệu huấn luyện và kiểm tra là giống nhau.
- Sự biến đổi trong giọng nói của người nói.

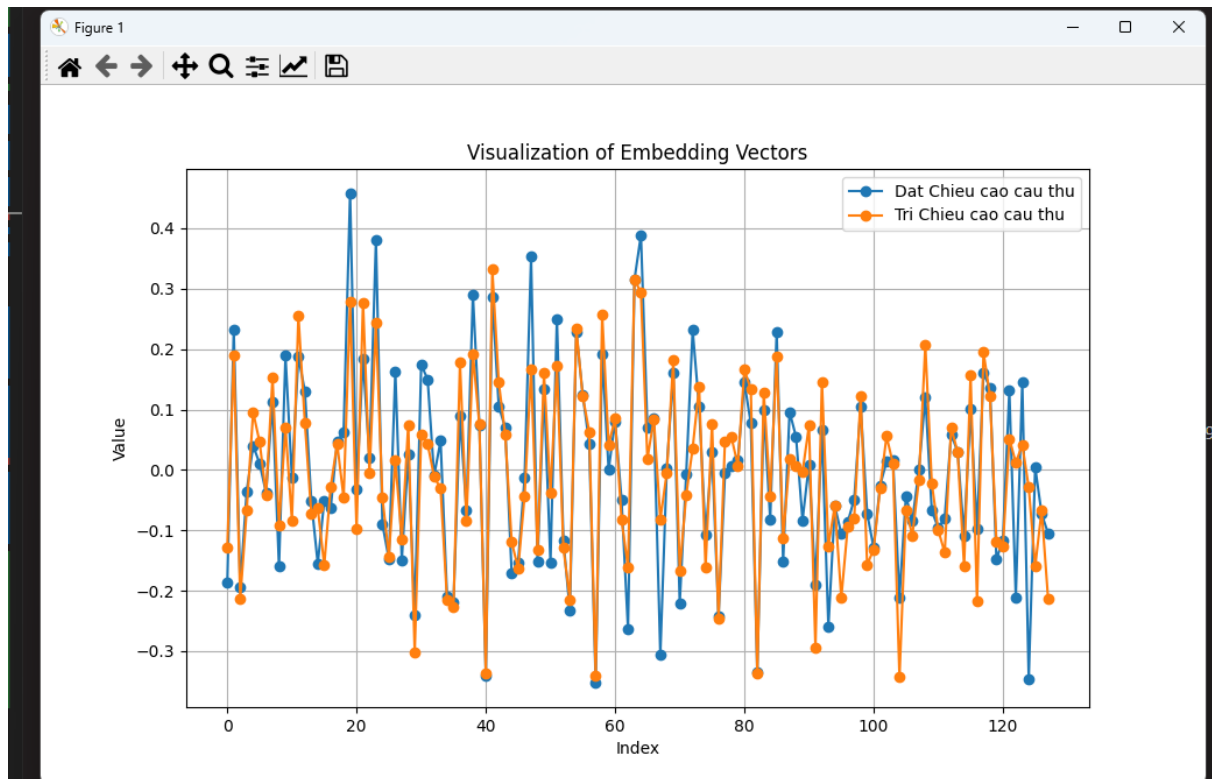
⇒ Có vẻ như về mặt lý thuyết nhóm đang làm tương đối ổn, tuy nhiên **sự khác nhau giữa thiết bị thu âm tập train và tập test là nguyên nhân cốt lõi của vấn đề nêu trên.**

III. Giải quyết vấn đề:

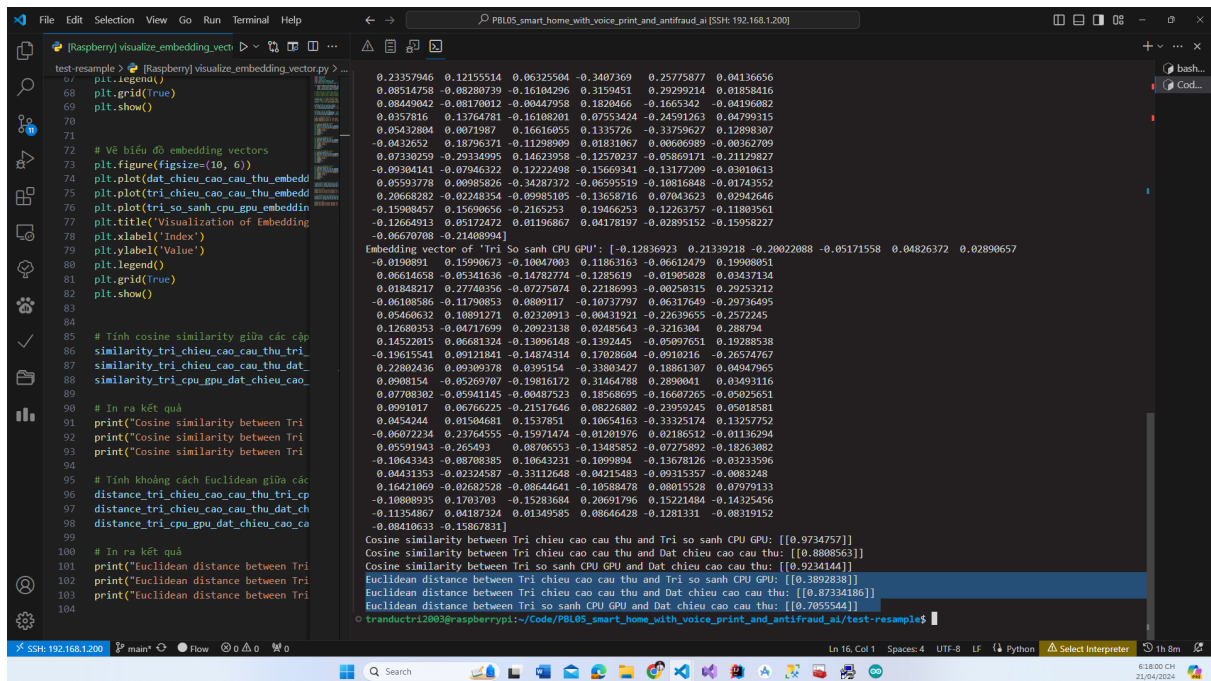
Nhóm đã thực hiện **Visualize embedding vector** để xác minh phỏng đoán của mình:



⇒ Cùng 1 người nói thì cho dù 2 nội dung khác nhau vẫn có sự tương đồng vector nhất định



⇒ Cùng một nội dung nói nhưng 2 người nói khác nhau thì vector khác nhau hoàn toàn



⇒ Nhóm khảo sát định lượng để đưa ra kết luận: MFCC trích được đặc trưng người nói và không bị ảnh hưởng bởi nội dung nói

```
Cosine similarity between Laptop Tri chieu cao cau thu and Raspberry Tri chieu cao cau thu: [[0.6846342]]
Cosine similarity between Laptop Tri so sanh CPU GPU and Raspberry Tri chieu cao cau thu: [[0.6482775]]
Cosine similarity between Laptop Dat chieu cao cau thu and Raspberry Tri chieu cao cau thu: [[0.50800574]]
Cosine similarity between Raspberry Tri so sanh CPU GPU and Raspberry Tri chieu cao cau thu: [[0.9565914]]
Euclidean distance between Laptop Tri chieu cao cau thu and Raspberry Tri chieu cao cau thu: [[1.2661959]]
Euclidean distance between Laptop Tri so sanh CPU GPU and Raspberry Tri chieu cao cau thu: [[1.3123955]]
Euclidean distance between Laptop Dat chieu cao cau thu and Raspberry Tri chieu cao cau thu: [[1.6554686]]
Euclidean distance between Raspberry Tri so sanh CPU GPU and Raspberry Tri chieu cao cau thu: [[0.4222939]]
```

⇒ Cùng 1 người nói, 1 nội dung nói nhưng thiết bị thu âm khác nhau thì 2 vector khác nhau hoàn toàn, trong khi cùng 1 người nói, nội dung nói khác nhau nhưng cùng là thiết bị thu âm thì vector tương tự nhau

⇒ Vấn đề kỹ thuật (Đừng tưởng hoa nở mà ngỡ xuân về)

IV. Triển khai phần cứng

Nhóm triển khai phần cứng lên raspberry, mặc dù gặp nhiều vấn đề kỹ thuật điện phức tạp và cực kỳ khó đoán và mất thời gian nhưng nhóm đã tìm được hướng giải quyết thành công.

Vì thời gian gấp rút (AI4LIFE, thầy đổi lịch thuyết trình) nên nhóm vẫn chưa ráp nhà thành công nhưng đã chạy thành công các module rồi.

Triển khai module Speech Recognition trên Raspberry chạy tốn nhiều thời gian hơn so với Laptop nên **nhóm đã sử dụng gọi API google** thay thế và tốc độ đã được cải thiện rất nhiều. Bên cạnh đó, để tăng tốc độ xử lý Speaker Recognition, **nhóm đã chỉnh sửa BE để cải thiện hệ thống**, lưu thông tin hiệu quả hơn.

V. Vấn đề tồn đọng

Nhóm gặp vấn đề rất lớn chỗ nguồn điện. Nhóm đã cố gắng nghiên cứu, thử nghiệm, theo học các bạn giỏi nhất ở ngành tự động hóa nhưng hiện vẫn chưa giải quyết được **câu chuyện nguồn điện pin tổ ong và phải ngậm ngùi sử dụng Adapter thay thế.**

Mạch điện của nhóm hiện tại là cực kỳ phức tạp và rối rắm, nhóm dự định sẽ thi AI4LIFE xong sẽ hàn các dây và ráp nhà.

Vấn đề còn tồn đọng lớn nhất là model nhận diện chưa ổn định ở **Lê Anh Tuấn**, khi nhóm xuất Vector để kiểm tra thì đúng là hình dạng vector của Tuấn và Trí khá tương đồng.

Nhóm phỏng đoán: hàm **triplet loss** nhóm train model sử dụng hàm **cosine**, trong khi inference lại sử dụng **euclidian distance**. Nhóm sẽ tiến hành train lại và nghiên cứu sâu hơn vấn đề này.

VI. Định hướng dự định sắp tới của nhóm

Nhóm đã và đang triển khai nghiên cứu và thực hành song song bài toán **deepvoice: Real-Time Detection of AI-Generated Speech For Deepfake Voice Conversion**

Lê Văn Tiến Đạt đang triển khai màn TFT và ESP32 cũng như truyền nhận thông tin trong nhà thông minh. Tuấn đang nghiên cứu paper và repo của ASVSpooof còn Trí và Phát vẫn tiếp tục triển khai phần cứng và mạch điện.

Dự định nhóm sau khi thi AI4Life xong sẽ tập trung hơn để hoàn thiện đồ án!

Dạ em xin chân thành cảm ơn thầy Duy đã đọc đến đây ạ!