(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0053653**

(51)[2020.01] **G01L 13/00**

(13) **B**

(21) 1-2021-00141

(22) 12/01/2021

(45) 25/11/2025 452

(43) 26/04/2021 397A

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

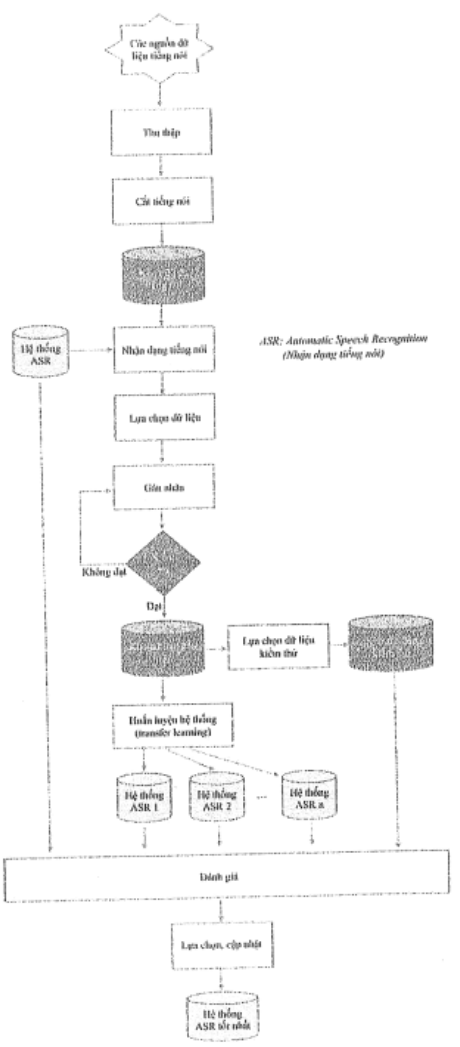
(72) Đỗ Văn Hải (VN); Lê Nhật Minh (VN); Nguyễn Tùng Lâm (VN); Lê Đăng Linh (VN); Nguyễn Tiến Thành (VN); Lê Quang Trung (VN); Mai Văn Tuấn (VN); Nguyễn Ngọc Dũng (VN); Trần Mạnh Quân (VN); Nguyễn Mạnh Quý (VN).

(74) Công ty TNHH Tư vấn Quốc Dân (NACILAW)

(54) QUY TRÌNH XÂY DỰNG DỮ LIỆU VÀ HUẤN LUYỆN LIÊN TỤC HỆ THỐNG NHẬN DẠNG TIẾNG NÓI

(21) 1-2021-00141

(57) Sáng chế đề xuất quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói. Sáng chế khắc phục được nhược điểm của các phương pháp hiện có bằng cách đề xuất phương pháp xây dựng dữ liệu và huấn luyện liên tục, giúp giảm thời gian và chi phí huấn luyện. Từ đó người dùng có thể cập nhập hệ thống nhận dạng tiếng nói một cách liên tục và đều đặn dựa vào dữ liệu thực tế của chính họ. Sáng chế bao gồm các bước: bước 1: thu thập dữ liệu tiếng nói; bước 2: tự động cắt tệp tiếng nói thành các đoạn nhỏ; bước 3: chuyển đổi tiếng nói sang văn bản; bước 4: lựa chọn đoạn tiếng nói thỏa mãn điều kiện; bước 5: gán nhãn và chỉnh sửa lại văn bản; bước 6: kiểm tra chất lượng gán nhãn; bước 7: tạo các tập kiểm thử; bước 8: lựa chọn thời điểm huấn luyện hệ thống; bước 9: huấn luyện hệ thống nhận dạng; bước 10: đánh giá các hệ thống nhận dạng với các tập kiểm thử; bước 11: lựa chọn cập nhật hệ thống nhận dạng.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói. Cụ thể, sáng chế đề cập đến quy trình xây dựng dữ liệu và huấn luyện liên tục cho một hệ thống nhận dạng tiếng nói, giúp giảm thời gian và chi phí huấn luyện, từ đó có thể cập nhập hệ thống nhận dạng tiếng nói một cách liên tục và đều đặn.

Tình trạng kỹ thuật của sáng chế

Hiện nay các ứng dụng nhận dạng tiếng nói đã trở nên rất phổ biến. Ví dụ ta có thể nhập liệu, tìm kiếm bằng tiếng nói thay vì gõ vào bàn phím thông qua các ứng dụng của hệ điều hành iOS, Android, Windows,…

Để xây dựng được những hệ thống nhận dạng tiếng nói ta cần có quá trình huấn luyện để máy tính học được mối quan hệ giữa tiếng nói ở đầu vào và văn bản ở đầu ra. Về nguyên tắc khi ta càng có nhiều dữ liệu huấn luyện thì hệ thống nhận dạng càng có khả năng nhận dạng chính xác hơn. Tuy nhiên để huấn luyện một hệ thống nhận dạng tiếng nói với một lượng dữ liệu lớn ta cần rất nhiều thời gian, cùng với đó là một hệ thống máy tính mạnh. Do đó, việc huấn luyện mô hình nhận dạng tiếng nói thường chỉ có thể được thực hiện tại các công ty cung cấp dịch vụ nhận dạng tiếng nói, mà khó có thể triển khai ở phía người dùng sử dụng dịch vụ.

Trong khi đó trong quá trình sử dụng dịch vụ nhận dạng tiếng nói, phía người dùng thu thập được nhiều dữ liệu thực tế, cùng với đó là việc phát hiện những trường hợp máy nhận dạng nhầm. Mong muốn của người dùng đó là làm sao có thể huấn luyện cho máy biết những lỗi sai trong quá trình sử dụng như vậy và khắc phục ngay trong các lần sau. Ngoài ra, để đáp ứng các yêu cầu về bảo mật dữ liệu, người sử dụng mong muốn toàn bộ quá trình xử lý được thực hiện tại chỗ, tránh việc truyền tải dữ liệu ra bên ngoài hệ thống.

Do đó cần thiết có một phương pháp có thể huấn luyện hệ thống nhận dạng tiếng nói liên tục từ đó có thể cập nhật nhanh nhất hệ thống nhận dạng với những dữ liệu được thực hiện ở phía người dùng.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế này là đưa ra quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói nhằm liên tục nâng cao chất lượng của hệ thống nhận dạng tiếng nói phù hợp với chính nhu cầu, dữ liệu của người dùng.

Cụ thể, sáng chế đề xuất quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói bao gồm:

- Bước 1: thu thập dữ liệu tiếng nói; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ hoặc thông qua các kết nối mạng dữ liệu;

- Bước 2: tự động cắt tệp tiếng nói thành các đoạn nhỏ; bước này được thực hiện bằng cách dựa vào đặc tính tín hiệu của tiếng nói nhằm loại bỏ những đoạn không chứa tiếng nói như các khoảng lặng hoặc các đoạn chỉ chứa âm thanh nhiễu, ồn;

- Bước 3: chuyển đổi tiếng nói sang văn bản; tại bước này, tất cả các đoạn tiếng nói ở bước 2 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói sau khi qua hệ thống nhận dạng thu được một văn bản tương ứng có số từ là N và một chỉ số độ tin cậy nhận dạng DTC có dải giá trị từ 0 đến 1;

- Bước 4: lựa chọn đoạn tiếng nói thỏa mãn điều kiện; tại bước này, lựa chọn các đoạn tiếng nói trong bước 2 thỏa mãn hai điều kiện: một là có độ tin cậy ở bước 3 nằm trong ngưỡng cho phép, tức là $DTC \geq DTC_{min}$ và $DTC \leq DTC_{max}$; hai là: có số từ nhận dạng trong văn bản ở bước 3 cũng nằm trong ngưỡng cho phép, tức là: $N \geq N_{min}$ và $N \leq N_{max}$. Trong đó DTC_{min} có giá trị từ 0,4 đến 0,8 nhằm loại bỏ nhưng đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu; DTC_{max} có giá trị từ 0,8 đến 1,0 nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá cao, nếu bổ sung vào dữ liệu huấn luyện sẽ không mang lại nhiều giá trị; N_{min} có giá trị từ 1 đến 10 nhằm loại bỏ những đoạn tiếng nói quá ngắn không chứa nhiều thông tin; N_{max} có giá trị từ 10 đến 40 nhằm loại bỏ những đoạn tiếng nói quá dài gây khó khăn trong việc nghe và làm dữ liệu;

- Bước 5: gán nhãn và chỉnh sửa lại văn bản; tại bước này, đưa các đoạn tiếng nói được lựa chọn ở bước 4 cùng với văn bản tương ứng được nhận dạng ở bước 3 lên hệ thống gán nhãn để người gán nhãn nghe và chỉnh sửa lại văn bản cho đúng với nội dung tương ứng của đoạn tiếng nói;

- Bước 6: kiểm tra chất lượng gán nhãn; tại bước này, người kiểm tra đánh giá chất lượng nhãn văn bản được gán ở bước 5, với các đoạn tiếng nói không đạt sẽ yêu cầu người gán nhãn chỉnh sửa lại, nếu đạt cho đoạn tiếng nói cùng văn bản tương ứng vào kho dữ liệu được gán nhãn;

- Bước 7: tạo các tập kiểm thử; theo đó, người quản trị quyết định lựa chọn một số đoạn tiếng nói trong kho dữ liệu được gán nhãn ở bước 6 để tạo các tập kiểm thử với yêu cầu kích thước mỗi tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu để đảm bảo tập kiểm thử đủ lớn

và tin cậy, trong đó $H_{test\_min} \geq 0,5$ giờ; với những đoạn tiếng nói được lựa chọn làm tập kiểm thử sẽ được xóa khỏi kho dữ liệu được gán nhãn;

- Bước 8: lựa chọn thời điểm huấn luyện hệ thống; là thời điểm khi dữ liệu huấn luyện trong kho lớn hơn một ngưỡng H_{train_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{train\_min} \geq 1$ giờ;

- Bước 9: huấn luyện hệ thống nhận dạng; tại bước này, bằng cách áp dụng học chuyển giao (transfer learning) với tốc độ học khởi tạo α, trong đó hệ thống đầu vào là hệ thống nhận dạng hiện tại, dữ liệu huấn luyện để học chuyển giao là dữ liệu tiếng nói trong kho dữ liệu được gán nhãn; trong đó $0,001 \geq \alpha \geq 0,00001$; sau khi kết thúc mỗi lần duyệt dữ liệu huấn luyện (epoch) ta sẽ lưu ra một hệ thống để thực hiện kiểm thử trong bước tiếp theo;

- Bước 10: đánh giá các hệ thống nhận dạng với các tập kiểm thử; tại bước này, bằng cách sử dụng hệ thống hiện thời và các hệ thống được tạo ra từ bước 9 để nhận dạng các đoạn tiếng nói trong các tập kiểm thử và sử dụng công cụ để tự động so sánh văn bản được nhận dạng với văn bản do người gán nhãn dữ liệu đã nhập để đưa ra bảng các chỉ số sai số từ (word error rate) của các hệ thống với các tập kiểm thử;

- Bước 11: lựa chọn cập nhật hệ thống nhận dạng; từ kết quả ở bước 10, người quản trị sẽ quyết định lựa chọn hệ thống nhận dạng nào có sai số trung bình thấp nhất để cập nhật hoặc giữ nguyên hệ thống hiện thời.

Mô tả vắn tắt hình vẽ

Hình 1 là quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây có dựa vào các hình vẽ kèm theo, các hình vẽ này nhằm mục đích minh họa các phương án của sáng chế mà không làm giới hạn phạm vi bảo hộ sáng chế. Cụ thể, quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói bao gồm các bước:

- Bước 1: thu thập dữ liệu tiếng nói;

- Bước 2: tự động cắt tệp tiếng nói thành các đoạn nhỏ;

- Bước 3: chuyển đổi tiếng nói sang văn bản;

- Bước 4: lựa chọn đoạn tiếng nói thỏa mãn điều kiện;

- Bước 5: gán nhãn và chỉnh sửa lại văn bản;

- Bước 6: kiểm tra chất lượng gán nhãn;

- Bước 7: tạo các tập kiểm thử;

- Bước 8: lựa chọn thời điểm huấn luyện hệ thống;

- Bước 9: huấn luyện hệ thống nhận dạng;

- Bước 10: đánh giá các hệ thống nhận dạng với các tập kiểm thử;

- Bước 11: lựa chọn cập nhật hệ thống nhận dạng.

Tham chiếu Hình 1, nội dung cụ thể của các bước như sau:

Bước 1: thu thập dữ liệu tiếng nói; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như băng từ, ổ đĩa cứng, thẻ nhớ,... hoặc thông qua các giao thức truyền tệp qua mạng dữ liệu như FTP. Sau khi thu thập dữ liệu, sử dụng các công cụ xử lý âm thanh như sox để chuyển tần số lấy mẫu của các tệp tiếng nói về chung một giá trị mà hệ thống nhận dạng tiếng nói hoạt động được, ví dụ 8kHz hoặc 16kHz.

Bước 2: tự động cắt tệp tiếng nói thành các đoạn nhỏ. Với mỗi tệp tiếng nói thu được ở bước 1, tiếng nói được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu nhằm loại bỏ những đoạn không chứa tiếng nói như các khoảng lặng hoặc các đoạn chỉ chứa âm thanh nhiễu, ồn. Ta có thể dựa vào các phương pháp phổ biến như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể dựa trên các hệ thống nhận dạng tiếng nói ví dụ như WebRTC.

Bước 3: chuyển đổi tiếng nói sang văn bản; tất cả các đoạn tiếng nói ở bước 2 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói sau khi qua hệ thống nhận dạng thu được một văn bản tương ứng có số từ là N và một chỉ số độ tin cậy nhận dạng DTC có dải giá trị từ 0 đến 1.

Bước 4: lựa chọn đoạn tiếng nói thỏa mãn điều kiện; có hai điều kiện mà đoạn tiếng nói được lựa chọn cần thỏa mãn, cụ thể:

Một là: có độ tin cậy DTC ở bước 3 nằm trong khoảng $DTC \geq DTC_{min}$ và $DTC \leq DTC_{max}$, trong đó DTC_{min} có giá trị từ 0,4 đến 0,8; tốt hơn là DTC_{min} có giá trị từ 0,5 đến 0,8 nhằm loại bỏ nhưng đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu; DTC_{max} có giá trị từ 0,8 đến 1,0; tốt hơn là DTC_{max} có giá trị từ 0,85 đến 0,95 nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá cao, nếu bổ sung vào dữ liệu huấn luyện sẽ không mang lại nhiều giá trị. Những đoạn tiếng nói có DTC quá cao này, hệ thống nhận dạng đã đủ tự tin để nhận dạng, do đó không cần bổ sung vào dữ liệu huấn luyện.

Hai là: có số từ nhận dạng trong văn bản ở bước 3 nằm trong khoảng $N \geq N_{min}$ và $N \leq N_{max}$, trong đó; N_{min} có giá trị từ 1 đến 10; tốt hơn là N_{min} có giá trị từ 2 đến 8 nhằm loại bỏ những đoạn tiếng nói quá ngắn không chứa nhiều thông tin; N_{max} có giá trị từ 10 đến 40; tốt hơn là N_{max} có giá trị từ 15 đến 30 nhằm loại bỏ những đoạn tiếng nói quá dài gây khó khăn trong việc nghe và làm dữ liệu ở bước 5.

Cách thức lựa chọn dữ liệu như trên giúp lựa chọn được những dữ liệu hữu ích nhất cho hệ thống nhận dạng hiện tại. Từ đó làm giảm chi phí gán nhãn dữ liệu ở bước 5 cũng như giảm tài nguyên và thời gian tính toán để huấn luyện hệ thống nhận dạng ở bước 9.

Bước 5: gán nhãn và chỉnh sửa lại văn bản; tại bước này, ta đưa các đoạn tiếng nói được lựa chọn ở bước 4 cùng với văn bản tương ứng được nhận dạng lên hệ thống gán nhãn để người gán nhãn nghe và chỉnh sửa lại văn bản cho đúng với nội dung tương ứng của đoạn tiếng nói.

Bước 6: kiểm tra chất lượng gán nhãn; tại bước này, người kiểm tra đánh giá thủ công chất lượng nhãn văn bản được gán ở bước 5 bằng cách so sánh văn bản với nội dung thực tế có trong đoạn tiếng nói, với các đoạn tiếng nói không đạt sẽ yêu cầu người gán nhãn làm lại, nếu đạt cho đoạn tiếng nói cùng vào văn bản tương ứng vào kho dữ liệu được gán nhãn. Sử dụng người kiểm tra để đánh giá chất lượng văn bản được gán với đoạn tiếng nói ở bước 5 nhằm tăng cường chất lượng của dữ liệu trước khi đưa vào huấn luyện hệ thống nhận dạng. Việc này giúp hệ thống nhận dạng có thể được huấn luyện với dữ liệu được gán nhãn chính xác, tránh trường hợp học với dữ liệu sai sẽ làm suy giảm chất lượng nhận dạng của hệ thống.

Bước 7: tạo các tập kiểm thử; theo đó, người quản trị có thể quyết định lựa chọn một số đoạn tiếng nói trong kho dữ liệu được gán nhãn ở bước 6 để tạo các tập kiểm thử với yêu cầu kích thước mỗi tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu để đảm bảo tập kiểm thử đủ lớn và tin cậy, trong đó $H_{test\_min} \geq 0,5$ giờ; tốt hơn là $H_{test\_min} \geq 1,0$ giờ hoặc tốt hơn nữa là $H_{test\_min} \geq 2,0$ giờ để đảm bảo tập kiểm thử đủ lớn và tin cậy khi đánh giá hệ thống nhận dạng ở các bước tiếp theo. Với những đoạn tiếng nói được lựa chọn làm tập kiểm thử sẽ được xóa khỏi kho dữ liệu được gán nhãn.

Bước 8: lựa chọn thời điểm huấn luyện hệ thống; là khi dữ liệu huấn luyện trong kho lớn hơn một ngưỡng H_{train_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{train\_min} \geq 1$ giờ. Tốt hơn là $H_{train\_min} \geq 5$ giờ hoặc tốt hơn nữa là $H_{train\_min} \geq 10$ giờ nhằm đảm bảo lượng dữ liệu huấn luyện để cập nhập hệ thống đủ lớn, tránh được việc hệ thống nhận dạng bị ảnh hưởng bởi một tập dữ liệu nhỏ và không điển hình.

Bước 9: huấn luyện hệ thống nhận dạng; bước này được thực hiện bằng cách áp dụng học chuyển giao (transfer learning) với tốc độ học khởi tạo α, trong đó hệ thống đầu vào là hệ thống nhận dạng hiện tại, dữ liệu huấn luyện để cho học chuyển giao là dữ liệu tiếng nói trong kho dữ liệu được gán nhãn; trong đó $0,001 \geq \alpha \geq 0,00001$; tốt hơn là $0,0004 \geq \alpha \geq 0,00003$ hoặc tốt hơn nữa là $0,0002 \geq \alpha \geq 0,00005$ nhằm giúp hệ thống được học chuyển giao với dữ liệu mới một cách từ từ mà không bị quên đi những tri thức đã học trước đó; sau khi kết thúc mỗi lần duyệt dữ liệu huấn luyện (epoch) ta sẽ lưu ra một hệ thống để thực hiện kiểm thử trong bước tiếp theo. Việc huấn luyện có thể thực hiện trên các công cụ xây dựng các hệ thống tiếng nói như Kaldi, WeNet.

Bước 10: đánh giá các hệ thống nhận dạng với các tập kiểm thử; theo đó, bằng cách sử dụng hệ thống hiện thời và các hệ thống được tạo ra từ bước 9 để nhận dạng các đoạn tiếng nói trong các tập kiểm thử và sử dụng công cụ để tự động so sánh văn bản được nhận dạng với văn bản do người gán nhãn dữ liệu đã nhập để đưa ra bảng các chỉ số sai số từ (word error rate - WER) của các hệ thống với các tập kiểm thử. Từ đó tạo ra một bảng gồm m x n chỉ số sai số từ WER, trong đó m là số lượng hệ thống cần đánh giá, n là số lượng các tập kiểm thử; từ thông tin này người quản trị có thể đưa ra quyết định trong bước sau. Việc tính toán WER có thể được thực hiện trên các công cụ phổ biến như Kaldi, WeNet.

Bước 11: lựa chọn cập nhật hệ thống nhận dạng; từ kết quả ở bước 10, người quản trị sẽ căn cứ vào sai số trên các tập kiểm thử khác nhau để lựa chọn ra hệ thống có sai số trung bình thấp nhất để cập nhật. Nếu sai số trung bình của hệ thống hiện thời thấp hơn sai số trung bình thấp nhất của các hệ thống được tạo ra ở bước 9, thì ta giữ nguyên hệ thống cũ. Ngược lại, ta chọn cập nhật hệ thống hiện thời bằng hệ thống mới có sai số trung bình thấp nhất.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào hoạt động để xây dựng quy trình làm dữ liệu và huấn luyện liên tục cho hệ thống nhận dạng tiếng nói của tổng đài chăm sóc khách hàng của Viettel.

Tại trung tâm chăm sóc khách hàng (CSKH) Viettel, chúng tôi xây dựng hệ thống nhận dạng tiếng nói để chuyển đổi toàn bộ các cuộc gọi chăm sóc khách hàng của đầu số 18008119 sang văn bản. Từ đó có thể giám sát, thống kê được nội dung của các cuộc gọi một cách tự động và nhanh chóng. Ngoài ra, ta còn có thể biết được tâm tư, bức xúc của khách hàng cũng như việc trả lời khách hàng của điện thoại viên.

Hệ thống nhận dạng tiếng nói này được huấn luyện ban đầu tại Trung tâm Không gian mạng Viettel sử dụng 1000 giờ dữ liệu. Thời gian để huấn luyện hệ thống là 83 giờ.

Khi bắt đầu triển khai tại trung tâm CSKH, hệ thống nhận dạng tiếng nói còn có tỷ lệ lỗi cao, tỷ lệ lỗi từ (word error rate) = 22,1%, tức trung bình 1000 từ thì có 221 từ bị nhận dạng sai. Sau khi áp dụng quy trình làm dữ liệu và huấn luyện liên tục, thu được kết quả sau:

Lần cập nhật	Số lượng dữ liệu làm (giờ)	Thời gian huấn luyện (giờ)	Kết quả sai số từ (%)
Hệ thống gốc	-	-	*22,1*
1	30	2,5	21,4
2	30	2,5	20,9
3	30	2,5	20,4
4	30	2,5	19,7
5	30	2,5	19,4
6	30	2,5	18,9
7	30	2,5	18,2
8	30	2,5	17,4
9	30	2,5	16,9
10	30	2,5	16,5

Bên CSKH thực hiện quy trình làm dữ liệu theo phương pháp đề xuất. Sau đó mỗi khi làm được 30 giờ dữ liệu thì hệ thống nhận dạng lại được huấn luyện bổ sung, thời gian huấn luyện là 2,5 giờ. Ta có thể thấy sai số của hệ thống nhận dạng giảm khá ổn định khi được bổ sung dữ liệu và huấn luyện liên tục. Sau mười lần cập nhật sai số giảm từ 22,1% xuống 16,5%. Với phương pháp đề xuất này ta có thể huấn luyện liên tục theo các mức dữ liệu khác nhau, giảm thời gian huấn luyện và chờ đợi. Để so sánh, nếu ta dùng phương pháp huấn luyện từ đầu tức mỗi lần thêm 30 giờ dữ liệu ta lại gộp vào 1000 giờ dữ liệu gốc và huấn luyện thời gian huấn luyện sẽ cần ít nhất là 83 giờ thay vì chỉ 2,5 giờ như phương pháp đề xuất. Ngoài ra, toàn bộ dữ liệu được xử lý và huấn luyện nội bộ tại trung tâm CSKH, không cần truyền ra ngoài, qua đó đáp ứng đầy đủ các yêu cầu về bảo mật dữ liệu.

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là xây dựng được một quy trình làm dữ liệu và huấn luyện liên tục được hệ thống nhận dạng tiếng nói. Với phương pháp đề xuất này, hệ thống nhận dạng đươc cập nhật liên tục và nhanh chóng mỗi khi có dữ liệu từ phía người sử dụng. Điều này giúp hệ thống nhận dạng hoạt động càng ngày càng tốt hơn với bài toán của chính người sử dụng.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Quy trình xây dựng dữ liệu và huấn luyện liên tục hệ thống nhận dạng tiếng nói bao gồm:

bước 1: thu thập dữ liệu tiếng nói; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như băng từ, ổ đĩa cứng, thẻ nhớ,... hoặc thông qua các giao thức truyền tệp qua mạng dữ liệu như FTP; sau đó sử dụng các công cụ xử lý âm thanh như sox để chuyển tần số lấy mẫu của các tệp tiếng nói về chung một giá trị mà hệ thống nhận dạng tiếng nói hoạt động được, ví dụ 8kHz hoặc 16kHz;

bước 2: tự động cắt tệp tiếng nói thành các đoạn nhỏ; với mỗi tệp tiếng nói thu được ở bước 1, tiếng nói được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu nhằm loại bỏ những đoạn không chứa tiếng nói như các khoảng lặng hoặc các đoạn chỉ chứa âm thanh nhiễu, ồn; ta có thể dựa vào các phương pháp phổ biến như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể dựa trên các hệ thống nhận dạng tiếng nói ví dụ như WebRTC;

bước 3: chuyển đổi tiếng nói sang văn bản; tất cả các đoạn tiếng nói ở bước 2 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói sau khi qua hệ thống nhận dạng thu được một văn bản tương ứng có số từ là N và một chỉ số độ tin cậy nhận dạng DTC có dải giá trị từ 0 đến 1;

bước 4: lựa chọn đoạn tiếng nói trong bước 2 thỏa mãn cả hai điều kiện: một là có độ tin cậy ở bước 3 nằm trong ngưỡng cho phép, tức là $DTC \geq DTC_{min}$ và $DTC \leq DTC_{max}$; hai là: có số từ nhận dạng trong văn bản ở bước 3 cũng nằm trong ngưỡng cho phép, tức là: $N \geq N_{min}$ và $N \leq N_{max}$; trong đó DTC_{min} có giá trị từ 0,4 đến 0,8 nhằm loại bỏ nhưng đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu; DTC_{max} có giá trị từ 0,8 đến 1,0 nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá cao, nếu bổ sung vào dữ liệu huấn luyện sẽ không mang lại nhiều giá trị; N_{min} có giá trị từ 1 đến 10 nhằm loại bỏ những đoạn tiếng nói quá ngắn không chứa nhiều thông tin; N_{max} có giá trị từ 10 đến 40 nhằm loại bỏ những đoạn tiếng nói quá dài gây khó khăn trong việc nghe và làm dữ liệu;

bước 5: gán nhãn và chỉnh sửa lại văn bản; tại bước này, đưa các đoạn tiếng nói được lựa chọn ở bước 4 cùng với văn bản tương ứng được nhận dạng ở bước 3 lên hệ thống gán nhãn để người gán nhãn nghe và chỉnh sửa lại văn bản cho đúng với nội dung tương ứng của đoạn tiếng nói;

bước 6: kiểm tra chất lượng gán nhãn; tại bước này, người kiểm tra đánh giá thủ công chất lượng nhãn văn bản được gán ở bước 5 bằng cách so sánh văn bản với nội dung thực tế

có trong đoạn tiếng nói, với các đoạn tiếng nói không đạt sẽ yêu cầu người gán nhãn làm lại, nếu đạt cho đoạn tiếng nói cùng vào văn bản tương ứng vào kho dữ liệu được gán nhãn; sử dụng người kiểm tra để đánh giá chất lượng văn bản được gán với đoạn tiếng nói ở bước 5 nhằm tăng cường chất lượng của dữ liệu trước khi đưa vào huấn luyện hệ thống nhận dạng; việc này giúp hệ thống nhận dạng có thể được huấn luyện với dữ liệu được gán nhãn chính xác, tránh trường hợp học với dữ liệu sai sẽ làm suy giảm chất lượng nhận dạng của hệ thống;

bước 7: tạo các tập kiểm thử; theo đó, người quản trị có thể quyết định lựa chọn một số đoạn tiếng nói trong kho dữ liệu được gán nhãn ở bước 6 để tạo các tập kiểm thử với yêu cầu kích thước mỗi tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu để đảm bảo tập kiểm thử đủ lớn và tin cậy, trong đó $H_{test\_min} \geq 0,5$ giờ; tốt hơn là $H_{test\_min} \geq 1,0$ giờ hoặc tốt hơn nữa là $H_{test\_min} \geq 2,0$ giờ để đảm bảo tập kiểm thử đủ lớn và tin cậy khi đánh giá hệ thống nhận dạng ở các bước tiếp theo; với những đoạn tiếng nói được lựa chọn làm tập kiểm thử sẽ được xóa khỏi kho dữ liệu được gán nhãn;

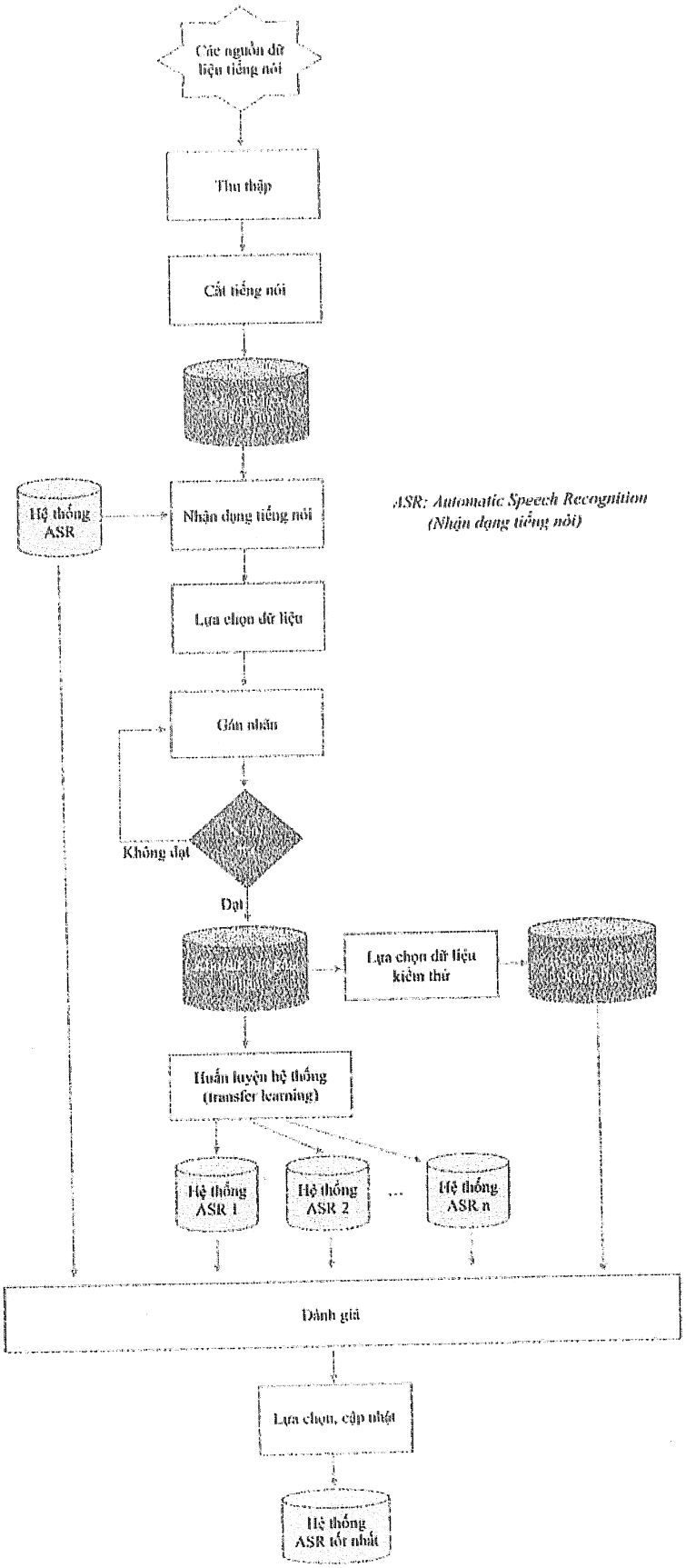
bước 8: lựa chọn thời điểm huấn luyện hệ thống; là khi dữ liệu huấn luyện trong kho lớn hơn một ngưỡng H_{train_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{train\_min} \geq 1$ giờ; tốt hơn là $H_{train\_min} \geq 5$ giờ hoặc tốt hơn nữa là $H_{train\_min} \geq 10$ giờ nhằm đảm bảo lượng dữ liệu huấn luyện để cập nhập hệ thống đủ lớn, tránh được việc hệ thống nhận dạng bị ảnh hưởng bởi một tập dữ liệu nhỏ và không điển hình;

bước 9: huấn luyện hệ thống nhận dạng; bước này được thực hiện bằng cách áp dụng học chuyển giao (transfer learning) với tốc độ học khởi tạo α, trong đó hệ thống đầu vào là hệ thống nhận dạng hiện tại, dữ liệu huấn luyện để cho học chuyển giao là dữ liệu tiếng nói trong kho dữ liệu được gán nhãn; trong đó $0,001 \geq \alpha \geq 0,00001$; tốt hơn là $0,0004 \geq \alpha \geq 0,00003$ hoặc tốt hơn nữa là $0,0002 \geq \alpha \geq 0,00005$ nhằm giúp hệ thống được học chuyển giao với dữ liệu mới một cách từ từ mà không bị quên đi những tri thức đã học trước đó; sau khi kết thúc mỗi lần duyệt dữ liệu huấn luyện (epoch) ta sẽ lưu ra một hệ thống để thực hiện kiểm thử trong bước tiếp theo; việc huấn luyện có thể thực hiện trên các công cụ xây dựng các hệ thống tiếng nói như Kaldi, WeNet;

bước 10: đánh giá các hệ thống nhận dạng với các tập kiểm thử; theo đó, bằng cách sử dụng hệ thống hiện thời và các hệ thống được tạo ra từ bước 9 để nhận dạng các đoạn tiếng nói trong các tập kiểm thử và sử dụng công cụ để tự động so sánh văn bản được nhận dạng với văn bản do người gán nhãn dữ liệu đã nhập để đưa ra bảng các chỉ số sai số từ (word error rate - WER) của các hệ thống với các tập kiểm thử; từ đó tạo ra một bảng gồm m x n chỉ số sai số từ WER, trong đó m là số lượng hệ thống cần đánh giá, n là số lượng các tập kiểm thử;

từ thông tin này người quản trị có thể đưa ra quyết định trong bước sau; việc tính toán WER có thể được thực hiện trên các công cụ phổ biến như Kaldi, WeNet;

bước 11: lựa chọn cập nhật hệ thống nhận dạng; từ kết quả ở bước 10, người quản trị sẽ căn cứ vào sai số trên các tập kiểm thử khác nhau để lựa chọn ra hệ thống có sai số trung bình thấp nhất để cập nhật; nếu sai số trung bình của hệ thống hiện thời thấp hơn sai số trung bình thấp nhất của các hệ thống được tạo ra ở bước 9, thì ta giữ nguyên hệ thống cũ; ngược lại, ta chọn cập nhật hệ thống hiện thời bằng hệ thống mới có sai số trung bình thấp nhất.


Các nguồn dữ liệu tiếng nói
Thu thập
Cắt tiếng nói
Hệ thống ASR
Nhận dạng tiếng nói
ASR: Automatic Speech Recognition
(Nhận dạng tiếng nói)
Lựa chọn dữ liệu
Gán nhãn
Không đạt
Đạt
Lựa chọn dữ liệu kiểm thử
Huấn luyện hệ thống (transfer learning)
Hệ thống ASR 1
Hệ thống ASR 2
...
Hệ thống ASR n
Đánh giá
Lựa chọn, cập nhật
Hệ thống ASR tốt nhất

Hình 1