(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) **1-0053627**

(51)[2022.01] **G10L 15/00**; G10L 17/00; G10L 15/18; G10L 15/20; G10L 15/05; G10L 15/14

(13) **B**

(21) 1-2023-06532

(22) 22/09/2023

(45) 25/11/2025 452

(43) 25/01/2024 430A

(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.

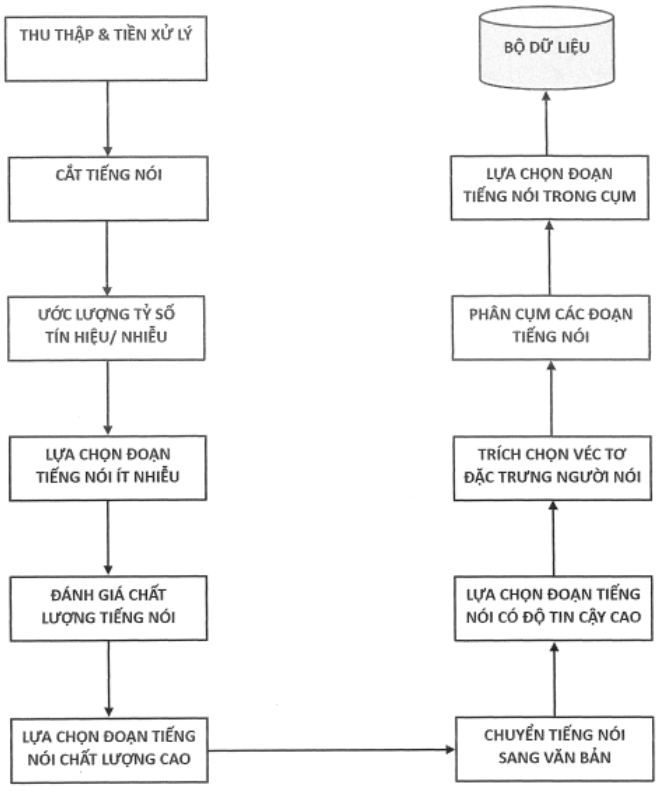
(72) Đỗ Văn Hải (VN); Lê Nhật Minh (VN); Nguyễn Tùng Lâm (VN); Nguyễn Tiến Thành (VN); Lê Đăng Linh (VN); Đặng Đình Sơn (VN); Nguyễn Thị Ngọc Ánh (VN); Phạm Minh Khang (VN); Tạ Bảo Thắng (VN); Lê Minh Tú (VN); Bùi Tiến Đạt (VN); Nguyễn Ngọc Dũng (VN); Trần Mạnh Quân (VN); Nguyễn Mạnh Quý (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP XÂY DỰNG BỘ DỮ LIỆU CHO TỔNG HỢP TIẾNG NÓI NHIỀU NGƯỜI NÓI

(21) 1-2023-06532

(57) Sáng chế đề xuất phương pháp tự động xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói giúp giảm chi phí xây dựng trong khi chất lượng tiếng nói vẫn được đảm bảo. Phương pháp sử dụng một loạt các công cụ cũng như bộ lọc nhằm giữ lại những đoạn tiếng nói đạt chất lượng đủ tốt, ít hoặc không có nhiễu, cũng như giữ được độ đa dạng về người nói trong bộ dữ liệu. Với phương pháp này ta có thể xây dựng được các bộ dữ liệu một cách nhanh chóng, với chi phí thấp từ đó có thể xây dựng được những mô hình tổng hợp tiếng nói nhiều người nói có chất lượng cao.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói. Cụ thể, phương pháp xây dựng bộ dữ liệu tự động nhằm tạo ra bộ dữ liệu với chi phí thấp trong khi chất lượng dữ liệu vẫn đảm bảo cho việc xây dựng các mô hình tổng hợp tiếng nói nhiều người nói.

Tình trạng kỹ thuật của sáng chế

Hiện nay các hệ thống tổng hợp tiếng nói nhiều người nói ngày càng được phát triển trong thực tế. Ưu điểm của các hệ thống này là có thể tạo ra giọng tổng hợp người nói đa dạng khác nhau. Tuy nhiên, để xây dựng các hệ thống như vậy chúng ta cần rất nhiều dữ liệu để huấn luyện. Yêu cầu với các bộ dữ liệu này đó là cần có chất lượng tiếng nói cao cũng như đa dạng nhiều người nói. Để xây dựng những bộ dữ liệu như vậy người ta thường tổ chức mời các phát thanh viên chuyên nghiệp đến phòng thu để đọc những văn bản có sẵn. Cách thu thập dữ liệu như vậy thường có chi phí cao cũng như thời gian thu âm, xử lý lâu, đặc biệt với bộ dữ liệu lớn. Do đó cần có một phương pháp tiếp cận mới trong việc thu thập và xử lý dữ liệu với chi phí thấp trong khi chất lượng bộ dữ liệu vẫn được đảm bảo.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp tự động xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói với chi phí thấp trong khi chất lượng bộ dữ liệu vẫn được đảm bảo. Cụ thể, phương pháp bao gồm:

Bước 1: Thu thập và tiền xử lý dữ liệu tiếng nói. Việc thu thập dữ liệu tiếng nói được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói hoặc video trực tiếp từ thiết bị lưu trữ hoặc thông qua các kết nối mạng dữ liệu. Với dữ liệu video ta sử dụng các công cụ như ffmpeg để trích xuất lấy phần tiếng nói có trong video. Sau khi thu thập dữ liệu, dùng các công cụ như ffmpeg, sox để chuyển đổi tần số lấy mẫu của các tệp tiếng nói về chung một giá trị F_S, trong đó $8kHz \leq F_S \leq 48kHz$;

Bước 2: Cắt tệp tiếng nói thành các đoạn nhỏ. Mỗi tệp tiếng nói thu được ở bước 1 được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu nhằm loại bỏ các đoạn không chứa tiếng nói đủ dài. Phương pháp phổ biến có thể được áp dụng như: dựa trên

năng lượng trung bình của tín hiệu hoặc ta có thể sử dụng trên các hệ thống học máy được huấn luyện trước cho tác vụ này.Việc cắt này chỉ được thực hiện nếu độ dài đoạn không chứa tiếng nói lớn hơn hoặc bằng một giá trị L_{MIN}, trong đó 0,1 giây $\leq L_{MIN} \leq$ 5 giây.

Bước 3: Ước lượng tỷ số tín hiệu trên nhiễu. Tất cả các đoạn tiếng nói ở bước 2 được ước lượng tỷ số tín hiệu trên nhiễu SNR theo công thức:

$$SNR[dB] = 10 \log \frac{P_{speech}}{P_{nonspeech}}$$

trong đó P_{speech} là công suất trung bình của đoạn tiếng nói, $P_{nonspeech}$ là công suất trung bình của các phần không phải là tiếng nói liền trước và liền sau của đoạn tiếng nói đó, *log* là hàm logarithm cơ số 10;

Bước 4: Lựa chọn đoạn tiếng nói ít nhiễu. Lựa chọn đoạn tiếng nói trong bước 3 thỏa mãn điều kiện có chỉ số tín hiệu trên nhiễu SNR $\geq$ SNR_{MIN}, trong đó 0dB $\leq$ $SNR_{MIN} \leq$ 50dB nhằm loại bỏ những đoạn tiếng nói có độ nhiễu lớn, giữ lại những đoạn tiếng nói sạch, ít hoặc không có nhiễu.

Bước 5: Đánh giá chất lượng tiếng nói. Tất cả các đoạn tiếng nói thu được ở bước 4 được đánh giá chất lượng tiếng nói bằng cách sử dụng các mô hình học máy đánh giá chất lượng tiếng nói không cần đầu vào tham chiếu (No-Reference Speech Quality Assessment Models), với mỗi đoạn tiếng nói thu được một điểm chất lượng Q được chuẩn hóa tuyến tính về dải từ 0 đến 5.

Bước 6: lựa chọn đoạn tiếng nói có điểm chất lượng cao. Lựa chọn đoạn tiếng nói trong bước 5 thỏa mãn điều kiện: có điểm chất lượng Q $\geq$ Q_{MIN}, trong đó 2 $\leq$ $Q_{MIN} \leq$ 5 nhằm loại bỏ những đoạn tiếng nói có chất lượng thấp, giữ lại những đoạn tiếng nói có chất lượng đủ tốt.

Bước 7: Chuyển đổi tiếng nói sang văn bản. Tất cả các đoạn tiếng nói thu được ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói. Với mỗi đoạn tiếng nói thu được một văn bản tương ứng TEXT và một chỉ số độ tin cậy nhận dạng CONFIDENCE có dải giá trị từ 0 đến 1.

Bước 8: Lựa chọn đoạn tiếng nói có độ tin cậy cao. Lựa chọn đoạn tiếng nói trong bước 7 có độ tin cậy CONFIDENCE $\geq$ $CONFIDENCE_{MIN}$, trong đó 0,4 $\leq$

$CONFIDENCE_{MIN} \leq 1$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy thấp thường là những đoạn tiếng nói có chất lượng kém hoặc có sai sót về văn bản TEXT nhận dạng.

Bước 9: trích chọn véc tơ đặc trưng người nói. Tất cả các các đoạn tiếng nói thu được ở bước 8 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng người nói được huấn luyện trước cho tác vụ nhận dạng người nói như mạng nơ rơn học sâu (DNN). Với mỗi đoạn tiếng nói sẽ thu được một véc tơ có số chiều cố định D đặc trưng người nói tương ứng, trong đó $16 \leq D \leq 2048$.

Bước 10: Phân cụm các đoạn tiếng nói. Phân cụm các đoạn tiếng nói thu được ở bước 8 thành N cụm $\{C_1, \ldots, C_N\}$ dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 9 nhằm gom những đoạn tiếng nói có đặc trưng người nói gần giống nhau lại thành các cụm. Trong đó số lượng cụm N có giá trị từ 2 đến 10^5.

Bước 11: Lựa chọn các đoạn tiếng nói trong các cụm. Với mỗi cụm C_i thu được ở bước 10, giữ lại ngẫu nhiên số lượng K'_i các đoạn tiếng nói trong cụm, K'_i được tính như sau:

$$K'_i = floor\left(min(K_i, \alpha \log K_i)\right)$$

trong đó K_i là số lượng đoạn tiếng nói trong cụm C_i thu được ở bước 10, α là tham số được lựa chọn thỏa mãn $1 \leq \alpha \leq 10^6$, *log* là hàm logarithm cơ số 10, *min* là hàm tính giá trị nhỏ nhất, *floor* là hàm trả về giá trị nguyên dưới gần nhất. Việc lựa chọn này nhằm mục đích loại bớt những đoạn tiếng nói do một số người nói quá nhiều trong bộ dữ liệu để đảm bảo độ đa dạng về người nói của bộ dữ liệu; cuối cùng ta thu thập được bộ dữ liệu gồm các đoạn tiếng nói còn lại ở bước 11 cùng với văn bản tương ứng TEXT thu thập được ở bước 7.

Mô tả vắn tắt các hình vẽ

Hình 1: Hình vẽ minh hoạ các bước thực hiện sáng chế.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây, cụ thể, phương pháp xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói bao gồm các bước:

Bước 1: Thu thập và tiền xử lý dữ liệu tiếng nói;

Bước 2: Cắt tệp tiếng nói thành các đoạn nhỏ;

Bước 3: Ước lượng tỷ số tín hiệu trên nhiễu;

Bước 4: Lựa chọn đoạn tiếng nói ít nhiễu;

Bước 5: Đánh giá chất lượng tiếng nói;

Bước 6: Lựa chọn đoạn tiếng nói có điểm chất lượng cao;

Bước 7: Chuyển đổi tiếng nói sang văn bản;

Bước 8: Lựa chọn đoạn tiếng nói có độ tin cậy cao;

Bước 9: Trích chọn véc tơ đặc trưng người nói;

Bước 10: Phân cụm các đoạn tiếng nói;

Bước 11: Lựa chọn các đoạn tiếng nói trong các cụm.

Nội dung cụ thể của các bước này như sau:

Bước 1: Thu thập và tiền xử lý dữ liệu tiếng nói. Việc thu thập dữ liệu tiếng nói được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói hoặc video trực tiếp từ thiết bị lưu trữ hoặc thông qua các kết nối mạng dữ liệu. Với dữ liệu video ta sử dụng các công cụ như ffmpeg để trích xuất lấy phần tiếng nói có trong video. Sau khi thu thập dữ liệu, dùng các công cụ như ffmpeg, sox để chuyển đổi tần số lấy mẫu của các tệp tiếng nói về chung một giá trị F_S, trong đó $8kHz \leq F_S \leq 48kHz$;

Bước 2: Cắt tệp tiếng nói thành các đoạn nhỏ. Mỗi tệp tiếng nói thu được ở bước 1 được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu nhằm loại bỏ các đoạn không chứa tiếng nói đủ dài. Phương pháp phổ biến có thể được áp dụng như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể sử dụng trên các hệ thống học máy được huấn luyện trước cho tác vụ này.Việc cắt này chỉ được thực hiện nếu độ dài đoạn không chứa tiếng nói lớn hơn hoặc bằng một giá trị L_{MIN}, trong đó 0,1 giây $\leq L_{MIN} \leq 5$ giây.

Bước 3: Ước lượng tỷ số tín hiệu trên nhiễu. Tất cả các đoạn tiếng nói ở bước 2 được ước lượng tỷ số tín hiệu trên nhiễu SNR theo công thức:

$$SNR[dB] = 10 \log \frac{P_{speech}}{P_{nonspeech}}$$

trong đó P_{speech} là công suất trung bình của đoạn tiếng nói, $P_{nonspeech}$ là công suất trung bình của các phần không phải là tiếng nói liền trước và liền sau của đoạn tiếng nói đó, *log* là hàm logarithm cơ số 10;

Bước 4: Lựa chọn đoạn tiếng nói ít nhiễu. Lựa chọn đoạn tiếng nói trong bước 3 thỏa mãn điều kiện có chỉ số tín hiệu trên nhiễu $SNR \geq SNR_{MIN}$, trong đó $0dB \leq SNR_{MIN} \leq$

50dB nhằm loại bỏ những đoạn tiếng nói có độ nhiễu lớn, giữ lại những đoạn tiếng nói sạch, ít hoặc không có nhiễu.

Bước 5: Đánh giá chất lượng tiếng nói. Tất cả các đoạn tiếng nói thu được ở bước 4 được đánh giá chất lượng tiếng nói bằng cách sử dụng các mô hình học máy đánh giá chất lượng tiếng nói không cần đầu vào tham chiếu (No-Reference Speech Quality Assessment Models), với mỗi đoạn tiếng nói thu được một điểm chất lượng Q được chuẩn hóa tuyến tính về dải từ 0 đến 5.

Bước 6: lựa chọn đoạn tiếng nói có điểm chất lượng cao. Lựa chọn đoạn tiếng nói trong bước 5 thỏa mãn điều kiện: có điểm chất lượng $Q \geq Q_{MIN}$, trong đó $2 \leq Q_{MIN} \leq 5$ nhằm loại bỏ những đoạn tiếng nói có chất lượng thấp, giữ lại những đoạn tiếng nói có chất lượng đủ tốt.

Bước 7: Chuyển đổi tiếng nói sang văn bản. Tất cả các đoạn tiếng nói thu được ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói. Với mỗi đoạn tiếng nói thu được một văn bản tương ứng TEXT và một chỉ số độ tin cậy nhận dạng CONFIDENCE có dải giá trị từ 0 đến 1.

Bước 8: Lựa chọn đoạn tiếng nói có độ tin cậy cao. Lựa chọn đoạn tiếng nói trong bước 7 có độ tin cậy $CONFIDENCE \geq CONFIDENCE_{MIN}$, trong đó $0,4 \leq CONFIDENCE_{MIN} \leq 1$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy thấp thường là những đoạn tiếng nói có chất lượng kém hoặc có sai sót về văn bản TEXT nhận dạng.

Bước 9: trích chọn véc tơ đặc trưng người nói. Tất cả các các đoạn tiếng nói thu được ở bước 8 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng người nói được huấn luyện trước cho tác vụ nhận dạng người nói như mạng nơ rơn học sâu (DNN). Với mỗi đoạn tiếng nói sẽ thu được một véc tơ có số chiều cố định D đặc trưng người nói tương ứng, trong đó $16 \leq D \leq 2048$.

Bước 10: Phân cụm các đoạn tiếng nói. Phân cụm các đoạn tiếng nói thu được ở bước 8 thành N cụm $\{C_1, \ldots, C_N\}$ dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 9 nhằm gom những đoạn tiếng nói có đặc trưng người nói gần giống nhau lại thành các cụm. Trong đó số lượng cụm N có giá trị từ 2 đến 10^5.

Bước 11: Lựa chọn các đoạn tiếng nói trong các cụm. Với mỗi cụm C_i thu được ở bước 10, giữ lại ngẫu nhiên số lượng K'_i các đoạn tiếng nói trong cụm, K'_i được tính như sau:

$$K_i' = floor\left(min(K_i, \alpha \log K_i)\right)$$

trong đó K_i là số lượng đoạn tiếng nói trong cụm C_i thu được ở bước 10, α là tham số được lựa chọn thỏa mãn $1 \leq \alpha \leq 10^6$, *log* là hàm logarithm cơ số 10, *min* là hàm tính giá trị nhỏ nhất, *floor* là hàm trả về giá trị nguyên dưới gần nhất. Việc lựa chọn này nhằm mục đích loại bớt những đoạn tiếng nói do một số người nói quá nhiều trong bộ dữ liệu để đảm bảo độ đa dạng về người nói của bộ dữ liệu; cuối cùng ta thu thập được bộ dữ liệu gồm các đoạn tiếng nói còn lại ở bước 11 cùng với văn bản tương ứng TEXT thu thập được ở bước 7.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào để xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói tại Trung tâm Không gian Mạng Viettel. Bằng việc áp dụng phương pháp này, chỉ trong vòng 2 tuần chúng tôi đã xây dựng được một bộ dữ liệu 3000 giờ tiếng nói tiếng Việt cho tổng hợp nhiều người nói. Bộ dữ liệu có chất lượng cao, tiếng nói rõ ràng, môi trường rất ít hoặc không có nhiễu, phân bố về người nói rất đa dạng đủ cả ba vùng bắc, trung, nam cũng như đa dạng về giới tính cũng như độ tuổi. So sánh với phương pháp thu âm với các phát thanh viên trong phòng thu, để thu được 3000 giờ dữ liệu ta cần thời gian thu âm và xử lý dữ liệu gấp xấp xỉ 4 lần tức xấp xỉ 12000 giờ cho thu âm và 12000 giờ cho xử lý dữ liệu. Do đó thời gian hoàn thành bộ dữ liệu có thể lên đến cả năm với hàng chục phòng thu làm việc song song. Do đó thời gian và chi phí để làm ra bộ dữ liệu là cực kì lớn. Rõ ràng với phương pháp được đề xuất trong sáng chế này đã cắt giảm được cả thời gian và chi phí đi nhiều lần trong khi chất lượng bộ dữ liệu vẫn được đảm bảo.

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là đưa ra được một phương pháp xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói một cách tự động. Phương pháp này đã được áp dụng thành công tại Trung tâm Không gian mạng Viettel giúp giảm thời gian và chi phí xây dựng bộ dữ liệu đi nhiều lần so với các phương pháp khác. Từ đó, có

thể xây dựng nhanh chóng các mô hình tổng hợp tiếng nói nhiều người nói áp dụng vào các sản phẩm thực tế như như tổng đài tự động, trợ lý ảo, báo nói, sách nói.

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. Phương pháp xây dựng bộ dữ liệu cho tổng hợp tiếng nói nhiều người nói bao gồm các bước:

bước 1: thu thập và tiền xử lý dữ liệu tiếng nói;

việc thu thập dữ liệu tiếng nói được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói hoặc video trực tiếp từ thiết bị lưu trữ hoặc thông qua các kết nối mạng dữ liệu; với dữ liệu video ta sử dụng các công cụ như ffmpeg để trích xuất lấy phần tiếng nói có trong video; sau khi thu thập dữ liệu, dùng các công cụ như ffmpeg, sox để chuyển đổi tần số lấy mẫu của các tệp tiếng nói về chung một giá trị F_S, trong đó $8kHz \leq F_S \leq 48kHz$;

bước 2: cắt tệp tiếng nói thành các đoạn nhỏ;

mỗi tệp tiếng nói thu được ở bước 1 được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu nhằm loại bỏ các đoạn không chứa tiếng nói đủ dài; phương pháp phổ biến có thể được áp dụng như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể sử dụng trên các hệ thống học máy được huấn luyện trước cho tác vụ này; việc cắt này chỉ được thực hiện nếu độ dài đoạn không chứa tiếng nói lớn hơn hoặc bằng một giá trị L_{MIN}, trong đó $0,1 \text{ giây} \leq L_{MIN} \leq 5 \text{ giây}$;

bước 3: ước lượng tỷ số tín hiệu trên nhiễu;

tất cả các đoạn tiếng nói ở bước 2 được ước lượng tỷ số tín hiệu trên nhiễu SNR theo công thức $SNR[dB] = 10log(P_{speech}/P_{nonspeech})$; trong đó P_{speech} là công suất trung bình của đoạn tiếng nói, $P_{nonspeech}$ là công suất trung bình của các phần không phải là tiếng nói liền trước và liền sau của đoạn tiếng nói đó, *log* là hàm logarithm cơ số 10;

bước 4: lựa chọn đoạn tiếng nói ít nhiễu;

lựa chọn đoạn tiếng nói trong bước 3 thỏa mãn điều kiện có chỉ số tín hiệu trên nhiễu $SNR \geq SNR_{MIN}$; trong đó $0dB \leq SNR_{MIN} \leq 50dB$ nhằm loại bỏ những đoạn tiếng nói có độ nhiễu lớn, giữ lại những đoạn tiếng nói sạch, ít hoặc không có nhiễu;

bước 5: đánh giá chất lượng tiếng nói;

tất cả các đoạn tiếng nói thu được ở bước 4 được đánh giá chất lượng tiếng nói bằng cách sử dụng các mô hình học máy đánh giá chất lượng tiếng nói không cần đầu vào tham chiếu (No-Reference Speech Quality Assessment Models), với mỗi đoạn tiếng nói thu được một điểm chất lượng Q được chuẩn hóa tuyến tính về dải từ 0 đến 5;

bước 6: lựa chọn đoạn tiếng nói có điểm chất lượng cao;

lựa chọn đoạn tiếng nói trong bước 5 thỏa mãn điều kiện: có điểm chất lượng $Q \geq Q_{MIN}$; trong đó $2 \leq Q_{MIN} \leq 5$ nhằm loại bỏ những đoạn tiếng nói có chất lượng thấp, giữ lại những đoạn tiếng nói có chất lượng đủ tốt;

bước 7: chuyển đổi tiếng nói sang văn bản;

tất cả các đoạn tiếng nói thu được ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói; với mỗi đoạn tiếng nói thu được một văn bản tương ứng TEXT và một chỉ số độ tin cậy nhận dạng CONFIDENCE có dải giá trị từ 0 đến 1;

bước 8: lựa chọn đoạn tiếng nói có độ tin cậy cao;

lựa chọn đoạn tiếng nói trong bước 7 có độ tin cậy $CONFIDENCE \geq CONFIDENCE_{MIN}$; trong đó $0,4 \leq CONFIDENCE_{MIN} \leq 1$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy thấp thường là những đoạn tiếng nói có chất lượng kém hoặc có sai sót về văn bản TEXT nhận dạng;

bước 9: trích chọn véc tơ đặc trưng người nói;

tất cả các các đoạn tiếng nói thu được ở bước 8 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng người nói được huấn luyện trước cho tác vụ nhận dạng người nói như mạng nơ rơn học sâu (DNN); với mỗi đoạn tiếng nói sẽ thu được một véc tơ có số chiều cố định D đặc trưng người nói tương ứng, trong đó $16 \leq D \leq 2048$;

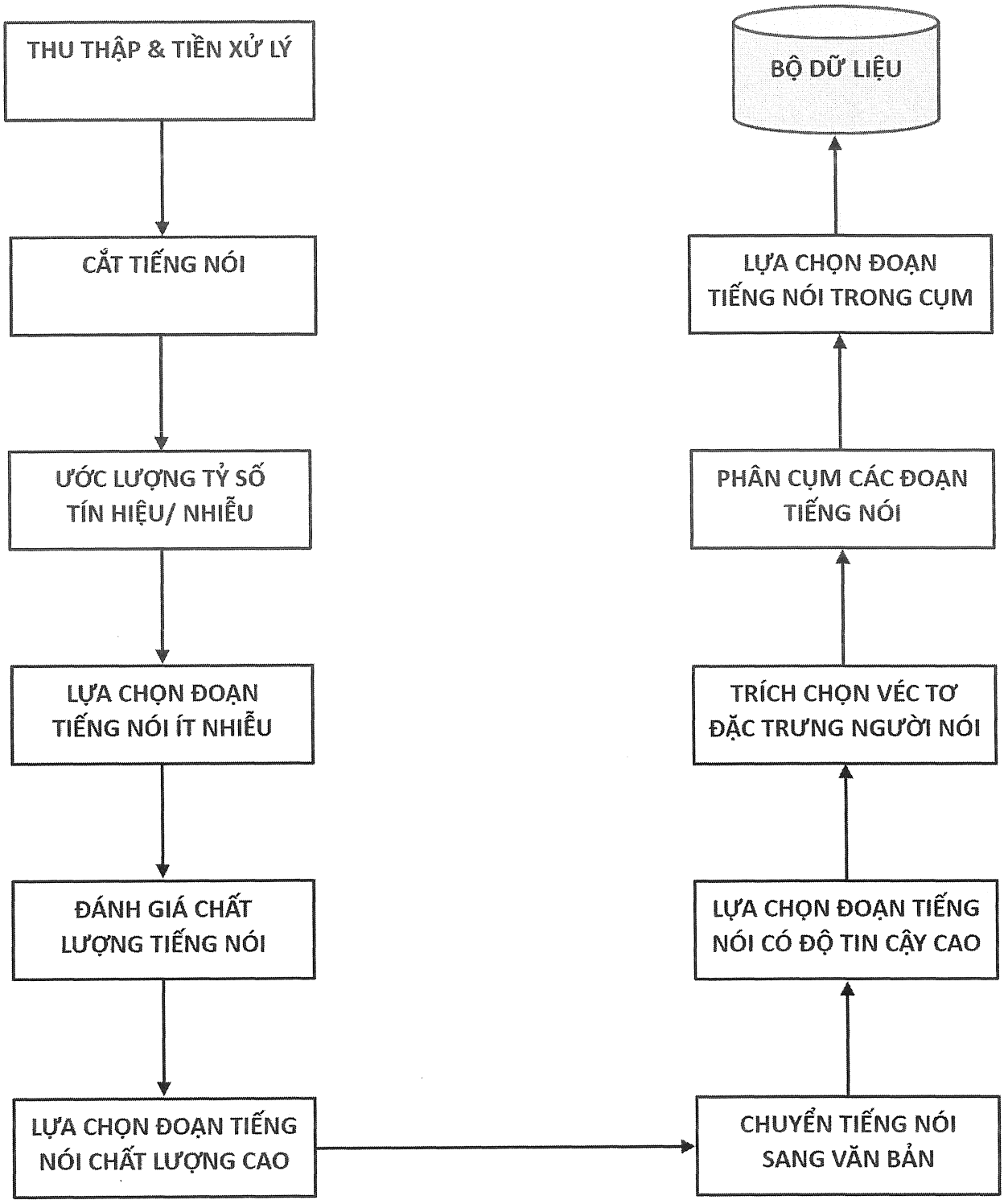
bước 10: phân cụm các đoạn tiếng nói;

phân cụm các đoạn tiếng nói thu được ở bước 8 thành N cụm $\{C_1, \ldots, C_N\}$ dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 9 nhằm gom những đoạn tiếng nói có đặc trưng người nói gần giống nhau lại thành các cụm; trong đó số lượng cụm N có giá trị từ 2 đến 10^5;

bước 11: lựa chọn các đoạn tiếng nói trong các cụm;

với mỗi cụm C_i thu được ở bước 10, giữ lại ngẫu nhiên số lượng K'_i các đoạn tiếng nói trong cụm, K'_i được tính như sau $K'_i = floor(min(K_i, \alpha*log(K_i)))$; trong đó K_i là số lượng đoạn tiếng nói trong cụm C_i thu được ở bước 10, α là tham số được lựa chọn thỏa mãn $1 \leq \alpha \leq 10^6$, *log* là hàm logarithm cơ số 10, *min* là hàm tính giá trị nhỏ nhất, *floor* là hàm trả về giá trị nguyên dưới gần nhất; việc lựa chọn này nhằm mục đích loại bớt những đoạn tiếng nói do một số người nói quá nhiều trong bộ dữ liệu để đảm bảo độ đa dạng về người nói của bộ dữ liệu; cuối cùng ta thu thập được bộ dữ liệu gồm các đoạn tiếng nói còn lại ở bước 11 cùng với văn bản tương ứng TEXT thu thập được ở bước 7.

Hình vẽ


THU THẬP & TIỀN XỬ LÝ
CẮT TIẾNG NÓI
ƯỚC LƯỢNG TỶ SỐ TÍN HIỆU/ NHIỄU
LỰA CHỌN ĐOẠN TIẾNG NÓI ÍT NHIỄU
ĐÁNH GIÁ CHẤT LƯỢNG TIẾNG NÓI
LỰA CHỌN ĐOẠN TIẾNG NÓI CHẤT LƯỢNG CAO
BỘ DỮ LIỆU
LỰA CHỌN ĐOẠN TIẾNG NÓI TRONG CỤM
PHÂN CỤM CÁC ĐOẠN TIẾNG NÓI
TRÍCH CHỌN VÉC TƠ ĐẶC TRƯNG NGƯỜI NÓI
LỰA CHỌN ĐOẠN TIẾNG NÓI CÓ ĐỘ TIN CẬY CAO
CHUYỂN TIẾNG NÓI SANG VĂN BẢN