Phát triển mô hình nhận dạng tiếng nói dân tộc thiểu số Hrê, Co sang tiếng Việt dạng văn bản sử dụng trí tuệ nhân tạo

Nguyễn Thành Việt[1*], Trần Duy Linh[2]

[1]Trường Đại học Phạm Văn Đồng, 509 Phan Đình Phùng, phường Cẩm Thành, tỉnh Quảng Ngãi, Việt Nam
[2]Trung tâm Chuyển đổi số và Đổi mới sáng tạo, 118 Hùng Vương, phường Nghĩa Lộ, tỉnh Quảng Ngãi, Việt Nam

Ngày nhận bài 26/7/2024; ngày chuyển phản biện 29/7/2024; ngày nhận phản biện 19/8/2024; ngày chấp nhận đăng 23/8/2024

Tóm tắt:

Việc xây dựng cơ sở dữ liệu (CSDL) điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co là hết sức cần thiết nhằm thu hẹp khoảng cách giao tiếp và ngôn ngữ giữa người Hrê, Co và người Kinh. Hiện nay, công nghệ nhận diện lời nói bằng trí tuệ nhân tạo (AI) đạt độ chính xác cao với tiếng Việt và nhiều ngôn ngữ khác, cho phép ứng dụng nhận diện tiếng Việt trong chiều dịch Việt - Hrê và Việt - Co của CSDL điện tử. Tuy nhiên, chiều ngược lại, nhận dạng và dịch tiếng Hrê, Co sang tiếng Việt hiện chưa có nghiên cứu. Trong bài báo này, các cặp từ tương ứng để tạo lập kho dữ liệu Việt - Hrê, Việt - Co đã được tiến hành số hóa đồng nhất. Nhờ vào kho dữ liệu đã xây dựng, nhóm tác giả đã phát triển thành công bộ phần mềm CSDL điện tử cho phép tra cứu ngữ nghĩa giữa các ngôn ngữ Việt - Hrê, Việt - Co với hai phiên bản web và thiết bị di động. Đặc biệt, nhóm tác giả cũng đã nghiên cứu và xây dựng thành công mô hình nhận dạng lời nói tiếng Hrê và Co, mở rộng khả năng ứng dụng cho CSDL điện tử này.

Từ khóa: **Co, cơ sở dữ liệu điện tử, Hrê, mô hình nhận dạng lời nói, tiếng đồng bào dân tộc thiểu số.**

Chỉ số phân loại: **1.2, 1.9**

Developing a speech recognition model of Hre and Co ethnic minority languages into Vietnamese text using artificial intelligence

Thanh Viet Nguyen[1*], Duy Linh Tran[2]

[1]Pham Van Dong University, 509 Phan Dinh Phung Street, Cam Thanh Ward, Quang Ngai Province, Vietnam
[2]Center for Digital Transformation and Innovation, 118 Hung Vuong Street, Nghia Lo Ward, Quang Ngai Province, Vietnam

Received 26 July 2024; revised 19 August 2024; accepted 23 August 2024

Abstract:

Building an electronic database of the ethnic minority languages of Viet - Hre, Viet - Co is extremely necessary to narrow the communication and language gap between the Hre, Co and Kinh people. Currently, AI speech recognition technology has achieved high accuracy with Vietnamese and many other languages, allowing the application of Vietnamese recognition in the Vietnamese - Hre and Vietnamese - Co translation direction of the electronic database. However, the opposite direction - recognition and translation of Hre, Co into Vietnamese - has never been studied before. In this article, the corresponding word pairs to create the Vietnamese - Hre, Vietnamese - Co data warehouse have been digitized uniformly. Thanks to the built data warehouse, the authors have successfully developed an electronic database software package that allows semantic lookup between the Vietnamese - Hre, Vietnamese - Co languages. This software package has both a web version and a mobile application "Hre - Co - Viet". In particular, the authors also researched and successfully built a speech recognition model for Hre and Co languages, expanding the application capabilities for this electronic database.

Keywords: **Co, electronic database, ethnic minority languages, Hre, speech recognition model.**

Classification numbers: **1.2, 1.9**

[*]Tác giả liên hệ: Email: ntviet@pdu.edu.vn

1. Đặt vấn đề

Theo thống kê năm 2019, dân số trên địa bàn tỉnh Quảng Ngãi khoảng 1,2 triệu người với hơn 30 dân tộc sinh sống, trong đó, dân tộc Hrê chiếm khoảng 10,8%, dân tộc Co chiếm khoảng 1,6% tổng dân số. Hầu hết các xã, phường và đặc khu của tỉnh đều có đồng bào dân tộc thiểu số cư trú [1, 2].

Trong những năm qua, với sự quan tâm của các cấp ngành có liên quan, Sở Nội vụ tỉnh Quảng Ngãi phối hợp với Trường Đại học Phạm Văn Đồng hằng năm đều tổ chức các lớp giảng dạy tiếng đồng bào dân tộc thiểu số trong đó có tiếng Hrê và Co cho đối tượng cán bộ công chức, viên chức, giáo viên. Tuy nhiên, chưa có CSDL tiếng đồng bào dân tộc thiểu số nào được biên soạn, xây dựng để phục vụ việc tra cứu ngôn ngữ trong quá trình học mà chỉ có các tài liệu giáo trình giảng dạy, tài liệu nghiên cứu của các nhóm tác giả ngôn ngữ học.

Vì vậy, việc nghiên cứu, xây dựng CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co là hết sức cần thiết nhằm thu hẹp khoảng cách giao tiếp, ngôn ngữ, giúp cho người đồng bào dân tộc thiểu số học hỏi, nâng cao kiến thức, hòa nhập với sự phát triển chung của tỉnh, của đất nước; đồng thời cũng giúp cho các cá nhân người Kinh đang làm việc với người đồng bào có thể hiểu và giao tiếp thuận lợi hơn, qua đó truyền đạt các kiến thức, các chủ trương, chính sách của chính quyền đến người đồng bào, tiếp thu kinh nghiệm và hiểu rõ các vấn đề của người đồng bào được sâu sắc hơn. Bên cạnh đó, việc xây dựng CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co cũng là một hình thức số hóa tiếng đồng bào dân tộc thiểu số của tỉnh Quảng Ngãi. Việc thu thập, lưu trữ để xây dựng kho dữ liệu ngữ vựng dạng số của tiếng Hrê, Co không chỉ giúp cho người bản ngữ có ý thức bảo tồn, phát triển tiếng mẹ đẻ, mà còn hình thành một kho tài nguyên thông tin đầy đủ, chân thực, giúp cho các thế hệ nghiên cứu hiện tại và sau này có tư liệu chính xác. Đồng thời, góp phần gìn giữ cho các thế hệ sau không chỉ ngôn ngữ mà còn những giá trị văn hoá truyền thống đặc sắc của người đồng bào dân tộc tỉnh Quảng Ngãi.

Trên cơ sở hai bộ tài liệu tiếng Hrê và tiếng Co đã được UBND tỉnh Quảng Ngãi phê duyệt: "Tài liệu Đào tạo, bồi dưỡng tiếng Hrê" (dùng cho cán bộ công chức tại miền núi và công tác dân tộc tỉnh Quảng Ngãi) của UBND tỉnh Quảng Ngãi (ban hành tạm thời kèm theo Quyết định số 893/QĐ-UBND ngày 27/4/2007 của UBND tỉnh Quảng Ngãi) [3] và tài liệu "Bài học tiếng Co" (là một công trình thuộc đề tài khoa học và công nghệ cấp tỉnh có tên "Nghiên cứu, biên soạn tài liệu dạy - học tiếng Co cho cán bộ, công chức (người Kinh) công tác tại huyện Trà Bồng và huyện Tây Trà, tỉnh Quảng Ngãi", chủ trì đề tài ThS. Nguyễn Minh Trí, thực hiện năm 2016) [4], nhóm tác giả đã tiếp cận và nghiên cứu vốn từ Việt, Hrê và Co cơ bản và thường xuyên sử dụng trong cuộc sống. Nhóm đã thực hiện việc số hóa đồng nhất giữa các từ tương ứng để thành lập một kho dữ liệu Việt - Hrê, Việt - Co. Cơ sở dữ liệu gồm: kho ngữ vựng song ngữ Việt - Hrê và ngược lại; kho ngữ vựng song ngữ Việt - Co và ngược lại.

Nhờ vào kho dữ liệu đã xây dựng, nhóm tác giả đã phát triển thành công bộ phần mềm CSDL điện tử có thể sử dụng được trên các thiết bị điện tử như: máy tính, điện thoại thông minh… nhằm mục đích tra ngữ nghĩa giữa các ngôn ngữ Việt - Hrê, Việt - Co. Bên cạnh đó, phần mềm trang bị thêm một số công cụ để người dùng có thể đề xuất cập nhật thêm một số từ chưa có trong cơ sở dữ liệu. Bộ phần mềm CSDL điện tử Việt - Hrê, Việt - Co gồm các phiên bản chạy trên web [5] và ứng dụng (app) "Hrê - Co - Việt" chạy trên thiết bị di động.

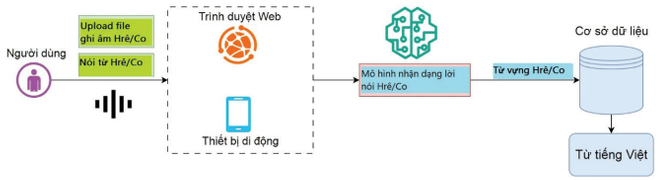
Hiện nay, các kỹ thuật AI đang phát triển mạnh mẽ cùng với những tiến bộ đồng thời về sức mạnh máy tính, dữ liệu lớn và hiểu biết lý thuyết. AI đã trở thành một phần thiết yếu của nhiều ngành và công nghệ, giúp giải quyết nhiều vấn đề thách thức trong học máy, công nghệ phần mềm, nghiên cứu vận hành, đặc biệt trong ngôn ngữ và nhận dạng lời nói.

Trong đó, đối với tiếng Việt và nhiều ngôn ngữ khác trên thế giới, công nghệ nhận diện giọng nói (ASR) bằng AI đã đạt được những kết quả với độ chính xác rất cao, cho phép nhận dạng tiếng nói và tự động chuyển tiếng nói thành văn bản điện tử. Trên cơ sở đó, có thể ứng dụng công nghệ này trong việc nhận diện tiếng Việt trong chiều dịch Việt - Hrê và Việt - Co của CSDL điện tử. Tuy nhiên, đối với chiều ngược lại - nhận dạng lời nói tiếng đồng bào dân tộc thiểu số, cụ thể là Hrê và Co rồi dịch sang tiếng Việt vẫn chưa ghi nhận nghiên cứu công bố trong nước. Với những thành tựu và xu hướng phát triển của công nghệ AI hiện nay, đặc biệt là khả năng huấn luyện thông qua mô hình học sâu (Deep learning), các tác giả đã nghiên cứu phương pháp và thực hiện xây dựng mô hình nhận dạng lời nói tiếng Hrê và Co để ứng dụng vào CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co.

2. Nội dung nghiên cứu

2.1. Mô hình ứng dụng công nghệ nhận dạng lời nói vào cơ sở dữ liệu điện tử

Các mô hình nhận dạng lời nói bằng AI và Deep learning sau khi được nghiên cứu thành công sẽ ứng dụng vào CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co theo mô hình như sau (hình 1): trong đó, người sử dụng nói trực tiếp hoặc gửi file âm thanh lên. Hệ thống nhận diện và chuyển lời nói sang văn bản, sau đó chuyển từ cần truy vấn đến CSDL. CSDL truy vấn lấy kết quả và trả kết quả (nếu tìm thấy) cho người dùng ở dạng văn bản. Nếu không tìm thấy từ vựng tương ứng, hệ thống sẽ đề xuất các từ vựng tương tự, gần giống nhất với từ được nhận diện bởi mô hình.



Hình 1. Mô hình ứng dụng nhận dạng lời nói vào cơ sở dữ liệu điện tử.

Ứng dụng nhận dạng lời nói vào CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co sẽ giúp người sử dụng có thể tra cứu bằng cách "nói từ" bên cạnh cách "nhập từ", mang lại nhiều hiệu quả thiết thực cho người sử dụng như: khả năng truy cập: Đây là một thuận lợi đối với người biết phát âm nhưng không biết cách viết, người không rành công nghệ, hoặc người khuyết tật khi không thể dùng chuột hay bàn phím, nhưng có thể dùng giọng nói để hệ thống chuyển thành văn bản, giúp nhập liệu hay thao tác một cách dễ dàng; tốc độ nhanh: phần mềm nhận dạng giọng nói có thể nắm bắt giọng nói của người dùng với tốc độ nhanh hơn so với khi nhập liệu bằng bàn phím, vì vậy tốc độ khi nhập liệu bằng giọng nói sẽ cải thiện đáng kể, từ đó giúp việc tra cứu từ vựng trở nên thuận tiện hơn.

2.2. Các mô hình học sâu phổ biến cho tác vụ công nghệ nhận diện giọng nói

Deep learning bao gồm các thuật toán học từ nhiều cấp độ để tạo ra các mô hình biểu diễn các mối quan hệ phức tạp giữa các dữ liệu. Deep learning sử dụng một hệ thống phân cấp các đặc trưng, trong đó các đặc trưng cấp cao hơn được xác định theo các đặc trưng cấp thấp hơn. Sự phổ biến của Deep learning có thể là do tăng khả năng xử lý của máy tính, xử lý lượng lớn dữ liệu huấn luyện và từ đó thúc đẩy những tiến bộ gần đây trong học máy, đặc biệt là trong xử lý thông tin và tín hiệu.

Xử lý tiếng nói là một lĩnh vực chuyên biệt của xử lý tín hiệu số liên quan đến việc thu nhận, thao tác và xuất tín hiệu lời nói. Nó bao gồm nhiều nhiệm vụ khác nhau như nhận dạng tiếng nói, nhận dạng người nói, nhận dạng cảm xúc, nhận dạng sức khỏe, hiểu ngôn ngữ, phát hiện giọng nói, nhận dạng tuổi và nhận dạng giới tính. Mục tiêu cuối cùng của các công nghệ nhận dạng tiếng nói là cho phép máy tính diễn giải và hiểu thông tin được nói để thực hiện các hành động phù hợp [6].

Trong những năm qua, lĩnh vực này đã có thành tựu đáng kể nhất là với sự phát triển của các mô hình Deep learning như: mạng nơ-ron sâu (Deep Neural Networks - DNN), mạng nơ-ron tích chập (Convolutional Neural Network - CNN) [7] đã cải tiến quá trình xử lý giọng nói bằng cách tự động học các đặc trưng có ý nghĩa từ tín hiệu giọng nói thô, cải thiện độ chính xác trong các ứng dụng khác nhau [8].

Tuy nhiên, trong thời gian gần đây, các mô hình nhận dạng lời nói tự động dựa trên kiến trúc Transformer, về bản chất cũng là các DNN xếp chồng lên nhau theo phương pháp của Deep learning [9] đã đạt được những bước đột phá trong việc xây dựng các mô hình đào tạo trước để tinh chỉnh (fine-tune) cho các tác vụ nhận dạng lời nói. Kiến trúc Transformer ban đầu được thiết kế để dịch văn bản viết từ ngôn ngữ này sang ngôn ngữ khác, trong đó gồm bộ mã hóa (Encoder) và bộ giải mã (Decoder), có nguyên tắc hoạt động như sau:

- Bộ mã hóa nhận đầu vào, trong trường hợp này là một chuỗi mã văn bản và tạo biểu diễn về nó (các thuộc tính - features). Phần này của mô hình được đào tạo để thu được sự hiểu biết từ dữ liệu đầu vào.

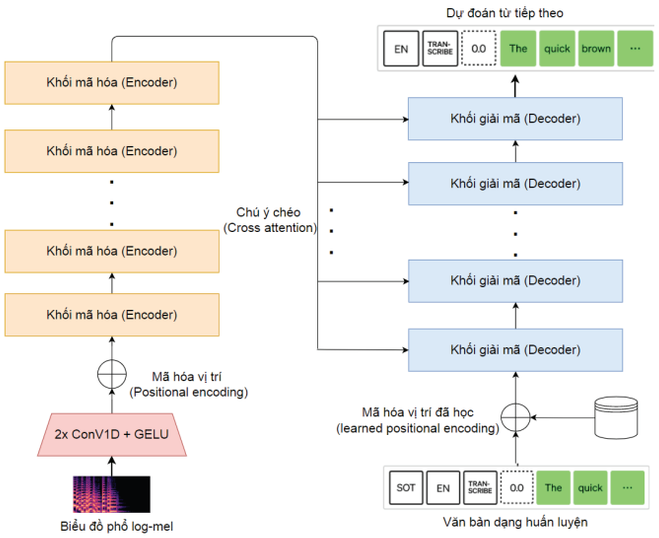
- Bộ giải mã sử dụng biểu diễn của bộ mã hóa (các thuộc tính) cùng với các đầu vào khác (các mã token được dự đoán trước đó) để tạo chuỗi mục tiêu. Phần này của mô hình được đào tạo để tạo ra kết quả đầu ra. Trong thiết kế ban đầu của Transformer, chuỗi đầu ra bao gồm các token văn bản.

Trong đó, CTC (Connectionist Temporal Classification) là một kỹ thuật được sử dụng với các kiến trúc Transformer chỉ có bộ mã hóa để nhận dạng lời nói tự động. Ví dụ về các mô hình như vậy là Wav2Vec2 [10], HuBERT [11] và M-CTC-T [12].

Tính đến trước năm 2022, CTC là kiến trúc phổ biến trong tác vụ ASR, với các mô hình chỉ dành cho bộ mã hóa như Wav2Vec2, HuBERT và XLSR [13] đã đạt được những bước đột phá trong các mô hình đào tạo trước và tinh chỉnh cho ASR. Các tập đoàn lớn, chẳng hạn như Meta và Microsoft, đã đào tạo trước bộ mã hóa Encoder lượng lớn dữ liệu âm thanh chưa được gắn nhãn trong nhiều ngày hoặc nhiều tuần. Sau đó, người dùng có thể sử dụng một điểm kiểm tra (checkpoint) được đào tạo trước và tinh chỉnh bằng CTC head trên dữ liệu giọng nói được gắn nhãn chỉ trong 10 phút để đạt được hiệu suất cao trong tác vụ ASR của mình.

Tuy nhiên, các mô hình CTC có những nhược điểm do chỉ sử dụng phần mã hóa của kiến trúc Transformer. Với các kiến trúc Transformer chỉ có bộ mã hóa, mô hình sẽ đưa ra dự đoán cho từng phần tử trong chuỗi đầu vào. Do đó, cả chuỗi đầu vào và đầu ra sẽ luôn có cùng độ dài.

Do đó, việc thêm bộ giải mã để tạo ra mô hình bộ mã hóa-giải mã (Encoder-Decoder), còn được gọi là mô hình chuỗi-sang-chuỗi (sequence-to-sequence, viết tắt là Seq2Seq). Mô hình này ánh xạ một chuỗi của một loại dữ liệu này sang một chuỗi của một loại dữ liệu khác. Với mô hình Seq2Seq, không có sự tương ứng một-một như vậy và các chuỗi đầu vào và đầu ra có thể có độ dài khác nhau. Điều đó làm cho các mô hình Seq2Seq phù hợp với các tác vụ như tóm tắt văn bản hoặc dịch giữa các ngôn ngữ khác nhau - cũng như cho các tác vụ âm thanh như nhận dạng lời nói. Trong đó, kiến trúc của bộ giải mã rất giống với kiến trúc của bộ mã hóa, cả hai đều sử dụng các lớp tương tự với tính năng tự chú ý (self-attention) làm tính năng chính.



Hình 2. Kiến trúc của mô hình Whisper.

Để xem cách thức hoạt động của tính năng này, có thể xem xét cách mô hình Whisper [14] thuộc kiến trúc Seq2Seq thực hiện nhận dạng lời nói tự động. Kiến trúc của mô hình Whisper được mô tả trong hình 2 [15]:

Trong thiết kế này, bộ giải mã đóng vai trò là mô hình ngôn ngữ, xử lý các biểu diễn trạng thái ẩn từ bộ mã hóa và tạo ra các bản phiên âm văn bản tương ứng. Đây là một cách tiếp cận hiệu quả hơn CTC, ngay cả khi mô hình CTC được kết hợp với mô hình ngôn ngữ ngoài, vì hệ thống Seq2Seq có thể được huấn luyện từ đầu đến cuối với cùng dữ liệu huấn luyện và hàm mất mát, mang lại tính linh hoạt cao hơn và hiệu suất nói chung vượt trội hơn.

Đặc biệt, nhu cầu về lượng lớn dữ liệu đào tạo đã là trở ngại trong quá trình phát triển kiến trúc Seq2Seq cho tác vụ ASR. Rất khó để có được dữ liệu giọng nói được gắn nhãn, với bộ dữ liệu có gắn nhãn lớn nhất vào thời điểm đó chỉ đạt 10.000 giờ. Tất cả điều này đã thay đổi vào tháng 9 năm 2022 khi Alec Radford và cộng sự từ OpenAI công bố Whisper - mô hình được đào tạo trước để nhận dạng giọng nói [14]. Không giống như các phiên bản CTC tiền nhiệm, vốn được huấn luyện hoàn toàn bằng dữ liệu âm thanh chưa được gắn nhãn, Whisper được đào tạo trước với một lượng lớn dữ liệu âm thanh được gắn nhãn với 680.000 giờ.

Đây là lượng dữ liệu lớn hơn nhiều so với dữ liệu âm thanh chưa được gắn nhãn được sử dụng để huấn luyện Wav2Vec 2.0 (60.000 giờ). Hơn nữa, 117.000 giờ dữ liệu đào tạo trước này là dữ liệu đa ngôn ngữ (hoặc "không phải tiếng Anh" - non-English). Điều này dẫn đến các điểm kiểm tra có thể được áp dụng cho hơn 96 ngôn ngữ, nhiều ngôn ngữ trong số đó được coi là có nguồn tài nguyên thấp, nghĩa là ngôn ngữ đó thiếu kho dữ liệu lớn phù hợp cho việc đào tạo.

Khi được mở rộng thành 680.000 giờ dữ liệu đào tạo trước được gắn nhãn, mô hình Whisper thể hiện khả năng khái quát hóa mạnh mẽ cho nhiều bộ dữ liệu và lĩnh vực. Các điểm kiểm tra được đào tạo trước đạt được kết quả cạnh tranh với các mô hình hiện đại, với tỷ lệ lỗi từ (WER) gần 3% trên các tập hợp con đã được kiểm tra của LibriSpeech pipe và TED-LIUM với 4,7% WER [14].

Điều đặc biệt quan trọng là khả năng của Whisper trong việc xử lý các mẫu âm thanh dạng dài, khả năng chống nhiễu đầu vào và khả năng dự đoán các văn bản được viết hoa và dấu câu. Do vậy, nhóm tác giả quyết định sử dụng Whisper để bắt đầu huấn luyện cho các mô hình nhận dạng lời nói trong công trình của mình.

3. Kết quả và bàn luận

3.1. Thu thập và khảo sát tập dữ liệu âm thanh

Trên cơ sở danh mục từ tiếng Hrê và tiếng Co đã chuẩn hóa, tiến hành ghi âm phát âm từ theo danh mục và đặt tên cho file ghi âm (định dạng .mp3) theo mã số của từ trong danh mục từ. Sau khi hoàn thành, đã thu thập được bộ file âm thanh thô của danh mục từ tiếng Hrê gồm 3868 file (tương ứng với 1934 từ Hrê do giọng nam và giọng nữ đọc) và tiếng Co gồm 1050 file (tương ứng 1050 từ giọng nam).

Tiếp theo, nghiên cứu sử dụng thư viện Datasets, cung cấp khả năng truy cập dễ dàng vào tuyển tập các bộ dữ liệu học máy có sẵn công khai trên nền tảng Hugging Face Hub [16]. Hơn nữa, Datasets bao gồm nhiều tính năng được điều chỉnh cho phù hợp với bộ dữ liệu âm thanh giúp đơn giản hóa cho việc nghiên cứu và thử nghiệm các mô hình học máy. Cụ thể, tạo kho lưu trữ tập dữ liệu trên Hugging Face Hub và tải lên thư mục tập dữ liệu theo cấu trúc của AudioFolder. Trong thư mục data chứa 2 thư mục con là train (gồm dữ liệu huấn luyện) và test (gồm dữ liệu kiểm tra). Trong đó, tệp siêu dữ liệu metadata.csv phải bao gồm cột file_name (chứa đường dẫn tới file âm thanh) để liên kết đến tệp âm thanh với từ vựng (transcription) tương ứng.

Ngoài ra, tăng cường dữ liệu là một kỹ thuật thường được sử dụng trong học máy và thị giác máy tính để tăng kích thước và tính đa dạng của một tập dữ liệu đào tạo một cách nhân tạo. Tác vụ này bao gồm việc áp dụng một số phép biến đổi hoặc sửa đổi cho các mẫu dữ liệu hiện có, tạo ra các mẫu mới giữ nguyên nhãn hoặc lớp như dữ liệu gốc. Dữ liệu tăng cường giúp cải thiện hiệu suất mô hình bằng cách giảm tình trạng quá khớp, cải thiện khả năng khái quát hóa và tăng độ mạnh mẽ của mô hình. Vì vậy, để cải thiện độ tin cậy và chính xác của mô hình huấn luyện, tác giả đã làm giàu dữ liệu huấn luyện bằng việc thực hiện tăng cường dữ liệu âm thanh (audio data augmentation) với 2 kỹ thuật:

- *Pitch Shifting:* Thay đổi ngẫu nhiên các tần số thành phần của Audio, trong đó cao độ của âm thanh được thay đổi trong khi vẫn duy trì độ dài của nó. Kỹ thuật này có thể mô phỏng các biến thể của giọng nói.

Bảng 1. Phân chia bộ dữ liệu từ Hrê để huấn luyện và kiểm tra.

	Từ bắt đầu	Từ kết thúc	Số lượng audio
Giọng nam	1	1934	1934
Giọng nữ	1	1934	1934
Nhiễu	1	200	200
Tăng cao độ	1	200	200
Tổng cộng			**4268**
Tập dữ liệu huấn luyện			
Giọng nam	1	1934	1934
Giọng nữ	601	1934	1334
Nhiễu	1	100	100
Tăng cao độ	1	100	100
Tổng cộng			**3468**
Tập dữ liệu kiểm tra			
Giọng nam	1	55	55
Giọng nữ	1	600	600
Nhiễu	101	200	100
Tăng cao độ	101	200	100
Tổng cộng			**855**

- *Noise Addition:* nhiễu ngẫu nhiên được thêm vào tín hiệu âm thanh. Điều này có thể giúp mô hình được huấn luyện linh hoạt hơn trước tiếng ồn môi trường hoặc sự thay đổi trong điều kiện ghi âm.

Cụ thể, lần lượt thực hiện Pitch shifting (tăng 3 tone giọng) và Noise addition (nhiễu Gauss với trung bình 0, độ lệch chuẩn 0,01) cho 200 file audio giọng nữ có số thứ tự từ 1 đến 200. Trong đó 100 từ đầu tiên dùng cho huấn luyện, 100 từ tiếp theo được đưa vào tập kiểm tra. Dữ liệu phân chia cho huấn luyện và kiểm tra được cụ thể hóa trong bảng 1.

Như vậy có trong tổng số 4268 file âm thanh, sử dụng 3468 file cho việc huấn luyện, còn lại 855 file (chiếm khoảng 20%) được dùng để kiểm tra, đánh giá chất lượng mô hình.

3.2. Huấn luyện mô hình nhận dạng lời nói và đánh giá chất lượng mô hình

Tiếp theo, tải mô hình được đào tạo trước của Whisper trên môi trường Google Colab, trong đó có sẵn trình trích xuất thuộc tính để chuẩn bị cho bước tiền xử lý trên tập dữ liệu âm thanh. Ở đây, khi muốn tinh chỉnh trên một ngôn ngữ mới (nằm ngoài danh sách các ngôn ngữ được huấn luyện trước của Whisper) như Hrê và Co, Whisper có khả năng tận dụng kiến thức về 96 ngôn ngữ khác mà nó được đào tạo trước. Bởi lẽ, tất cả các ngôn ngữ hiện đại sẽ có sự tương đồng với ít nhất một trong 96 ngôn ngữ mà Whisper đã được huấn luyện trước dựa trên mô hình biểu diễn kiến thức đa ngôn ngữ. Do vậy, để tinh chỉnh Whisper trên một ngôn ngữ mới (như Hrê và Co) là tìm ngôn ngữ giống nhất mà Whisper đã được đào tạo trước, mà cụ thể chính là tiếng Việt.

Sau khi chuẩn bị xong dữ liệu, có thể thực hiện huấn luyện mô hình nhận dạng tiếng Hrê và mô hình nhận dạng tiếng Co. Những gì cần chuẩn bị gồm:

- Xác định bộ đối chiếu dữ liệu (data collator): lấy dữ liệu đã qua tiền xử lý và chuẩn bị các tensor PyTorch sẵn sàng cho mô hình.

- Thước đo đánh giá: trong quá trình đánh giá, sử dụng tỷ lệ lỗi ký tự (character error rate - CER).

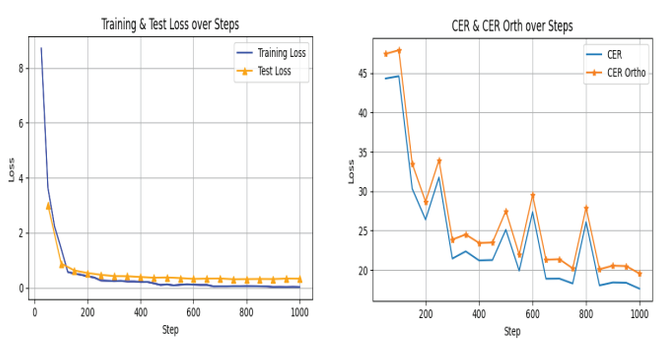
- Tải điểm kiểm tra được đào tạo trước (pre-trained checkpoint) Whisper small và thiết lập cấu hình chính xác để đào tạo.

- Xác định các đối số cho huấn luyện: những đối số này sẽ được đối tượng Trainer sử dụng trong việc xây dựng lịch trình huấn luyện. Ở đây, thiết lập số bước huấn luyện max_steps là 1000, và các siêu tham số khác như hiển thị trong hình 3.

```
from transformers import Seq2SeqTrainingArguments
training_args = Seq2SeqTrainingArguments(
    output_dir = "./ntviet/whisper-small-hre4.4",
    per_device_train_batch_size=16, gradient_accumulation_steps=1, learning_rate=1e-5,
    lr_scheduler_type="constant_with_warmup", warmup_steps=50, max_steps=1000,
    gradient_checkpointing=True, fp16=True, fp16_full_eval=True, evaluation_strategy="steps",
    per_device_eval_batch_size=16, predict_with_generate=True, generation_max_length=225,
    save_steps=1000, eval_steps=1000, logging_steps=25, report_to=["tensorboard"],
    load_best_model_at_end=True, metric_for_best_model="cer", greater_is_better=False,
    push_to_hub=True)
```

Hình 3. Thiết lập các siêu tham số để huấn luyện mô hình nhận dạng tiếng Hrê.

Sau cùng, có thể chuyển tiếp các đối số huấn luyện tới đối tượng Trainer cùng với mô hình, tập dữ liệu, trình đối chiếu dữ liệu để bắt đầu huấn luyện mô hình nhận dạng tiếng Hrê và mô hình nhận dạng tiếng Co (thực hiện hoàn toàn tương tự). Có thể thấy các mô hình nhận dạng đã huấn luyện cải thiện được độ chính xác qua từng bước và kết quả đạt khá tốt trên tập kiểm tra, với giá trị mất mát Loss sau cùng là 0,32, tỷ lệ lỗi ký tự CER sau cùng là 17,6% đối với mô hình nhận dạng tiếng Hrê, được hiển thị trong hình 4.



Hình 4. Đánh giá quá trình huấn luyện mô hình nhận diện tiếng Hrê.

Hình 5. Lớp tập huấn sử dụng phần mềm tại xã Minh Long, Quảng Ngãi (cũ).

3.3. Thử nghiệm và đánh giá kết quả trong thực tế ở địa phương

Để triển khai thử nghiệm và phát huy hiệu quả của sản phẩm, nhóm nghiên cứu đã thực hiện đưa sản phẩm vào sử dụng trong thực tế, trong đó: hướng dẫn khai thác CSDL điện tử tiếng Hrê, Co thông qua các lớp tập huấn tại 5 xã miền núi của tỉnh Quảng Ngãi (hình 5).

Nhóm tác giả đã phối hợp với Ban Dân tộc tỉnh tổ chức 5 lớp tập huấn hướng dẫn quản lý và hướng dẫn khai thác CSDL điện tử cho các đối tượng sử dụng, gồm: cán bộ công chức tại phòng Dân tộc các xã, cán bộ thôn bản, cán bộ Đoàn thanh niên, phụ nữ, hội nông dân xã miền núi; giáo viên và học sinh miền núi của tỉnh tại các xã: Ba Tơ, Minh Long, Trà Bồng, Sơn Hà, Sơn Tây; 1 lớp hướng dẫn quản trị, vận hành hệ thống cho Ban Dân tộc tỉnh và phòng Dân tộc các xã. Các lớp tập huấn đã có gần 300 học viên được giới thiệu về sản phẩm (gồm phiên bản web và app), được hướng dẫn các thao tác để khai thác, sử dụng các tính năng của CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co, biết cách gửi đóng góp về CSDL điện tử. Qua các lớp tập huấn, nhóm thực hiện cũng được tiếp xúc, trao đổi trực tiếp với đa dạng các đối tượng sử dụng, qua đó nắm bắt thêm những ý kiến, góp ý từ các học viên để hoàn thiện thêm sản phẩm, giúp sản phẩm ngày càng dễ dùng hơn, hiệu quả hơn với đông đảo người dân, từ đó giúp chia sẻ, lan truyền ứng dụng này rộng rãi đến cộng đồng.

- Thực hiện tuyên truyền, quảng bá sản phẩm trên các phương tiện truyền thông, nền tảng số: đơn vị chủ trì chủ động trong việc quảng bá và tuyên truyền sản phẩm CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co trên các phương tiện truyền thông, nền tảng số như Zalo, Facebook, các trang thông tin điện tử và các nền tảng số khác. Nhờ vào đó, thu hút được sự quan tâm của đông đảo cộng đồng mạng và các nhóm quan tâm đến bảo tồn và phát triển ngôn ngữ dân tộc.

- Giới thiệu đến tổ công nghệ số cộng đồng: Trên địa bàn tỉnh Quảng Ngãi (cũ), đã thành lập tổ công nghệ số cộng đồng đến cấp thôn, với tổng cộng hơn 1.140 tổ công nghệ số cộng đồng với hơn 7.500 thành viên tham gia. Thành viên tổ công nghệ số cộng đồng là những hạt nhân tích cực, có sứ mệnh đưa công nghệ số đến mọi ngõ ngách của cuộc sống, bằng cách tiếp cận với các công nghệ số, ứng dụng số hữu ích và triển khai, hướng dẫn lại cho người dân tại thôn, bản, tổ dân phố mình nắm bắt, sử dụng.

- Kết hợp giới thiệu, triển khai trong các lớp bồi dưỡng tiếng Hrê, Co: hằng năm, Sở Nội vụ tỉnh Quảng Ngãi phối hợp Trường Đại học Phạm Văn Đồng tổ chức các lớp bồi dưỡng tiếng đồng bào dân tộc thiểu số Hrê, Co cho các đối tượng cán bộ công chức, viên chức, giáo viên làm việc tại các xã miền núi.

- Đã tổ chức Hội thảo cấp tỉnh vào ngày 14/08/2024 để giới thiệu về kết quả sản phẩm nghiên cứu [17]. Hội thảo đã mời các chuyên gia trong lĩnh vực, các đại biểu có liên quan để tăng cường giới thiệu kết quả của công trình đến phạm vi bên ngoài tỉnh.

Thông qua việc triển khai các mô hình ứng dụng sản phẩm, đến cuối tháng 4/2025, web "Cơ sở dữ liệu điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co" [3] (hình 6) đã được khai thác sử dụng bởi hơn 617.000 lượt truy cập, và app Hrê - Co - Việt đã có hơn 1000 lượt tải và cài đặt ứng dụng.

Hình 6. Cơ sở dữ liệu điện tử tiếng Việt - Hrê, Việt - Co phiên bản web.

Sau khi triển khai thử nghiệm, thông qua các hình thức trực tiếp (lớp tập huấn) và gián tiếp (qua nền tảng số), đông đảo người sử dụng cho biết sản phẩm rất dễ sử dụng và hữu ích, giúp việc học tiếng Hrê, Co trở nên thuận tiện và dễ dàng hơn bao giờ hết. Đồng thời, người sử dụng cũng góp ý thêm để hoàn thiện sản phẩm như: có thể bổ sung các từ ghép từ các từ đơn, bổ sung chức năng cho tra cứu từ theo lĩnh vực để phục vụ tốt hơn nữa nhu cầu sử dụng của người dùng.

4. Kết luận

Hiện nay, AI đã có nhiều ứng dụng rộng rãi trong nhiều lĩnh vực, trong đó có lĩnh vực từ điển điện tử với công nghệ nhận diện lời nói. Tuy nhiên, việc nhận diện lời nói tiếng đồng bào dân tộc thiểu số Hrê, Co vẫn chưa ghi nhận nghiên cứu công bố trong nước, nên kết quả trong bài báo này là một nghiên cứu mới, vừa mang giá trị học thuật, vừa mang

lại hiệu quả thiết thực với sản phẩm ứng dụng thực tế, và sẽ là nền tảng cho các nghiên cứu khác liên quan đến ứng dụng công nghệ vào lĩnh vực xử lý ngôn ngữ tự nhiên, nhất là các ngôn ngữ tài nguyên thấp như tiếng Hrê và Co.

Hầu hết các mô hình dựa trên kiến trúc Transformer cho xử lý âm thanh đều khá giống nhau - đều được xây dựng trên nguyên tắc của các lớp Attention, mặc dù một số mô hình sẽ chỉ sử dụng bộ mã hóa, trong khi những mô hình khác sử dụng cả bộ mã hóa và bộ giải mã. Trước năm 2022, CTC là kiến trúc phổ biến trong tác vụ ASR, với các mô hình chỉ dùng bộ mã hóa, đã đạt được những bước đột phá trong mô hình đào tạo trước/tinh chỉnh cho giọng nói. Tuy nhiên, các mô hình CTC có những nhược điểm do thêm một lớp tuyến tính đơn giản vào bộ mã hóa dẫn đến dễ mắc lỗi chính tả về ngữ âm.

Trong khi đó kiến trúc Seq2Seq, tiêu biểu là mô hình Whisper, mạnh hơn hẳn mô hình chỉ sử dụng bộ mã hóa. Bằng cách tách mã hóa chuỗi đầu vào khỏi giải mã chuỗi đầu ra, việc căn chỉnh âm thanh và văn bản sẽ ít gặp vấn đề hơn. Do vậy, chúng tôi đã sử dụng mô hình Whisper small để thực hiện tiền xử lý bằng thư viện Datasets trên bộ dữ liệu đầy đủ tiếng Hrê và tiếng Co. Sau đó thực hiện tinh chỉnh huấn luyện được mô hình nhận dạng tiếng Hrê (trải nghiệm mô hình tại https://csdlhreco.nuian.vn/aihre.aspx và app "Hrê - Co - Việt"), và mô hình nhận dạng tiếng Co (trải nghiệm mô hình tại địa chỉ https://csdlhreco.nuian.vn/aico.aspx và app "Hrê - Co - Việt").

Ngoài ra, để triển khai thử nghiệm và phát huy hiệu quả của sản phẩm, nhóm tác giả đã đưa sản phẩm đi vào sử dụng trong thực tế. Mô hình ứng dụng bao gồm nhiều hoạt động và giải pháp để có thể giới thiệu sản phẩm đến đông đảo các đối tượng người dùng, đồng thời tiếp thu, hoàn thiện thêm sản phẩm, làm giàu thêm CSDL tiếng đồng bào dân tộc thiểu số Hrê, Co. Các hoạt động này bao gồm: hướng dẫn khai thác CSDL điện tử Hrê, Co thông qua các lớp tập huấn tại các khu vực miền núi của tỉnh Quảng Ngãi; tuyên truyền, quảng bá sản phẩm trên các phương tiện truyền thông, nền tảng số; giới thiệu đến tổ công nghệ số cộng đồng trên địa bàn tỉnh Quảng Ngãi; kết hợp giới thiệu, triển khai trong các lớp bồi dưỡng tiếng Hrê, Co; tổ chức hội thảo cấp tỉnh để giới thiệu về kết quả sản phẩm nghiên cứu.

Các mô hình ASR cho tiếng Hrê và tiếng Co giờ đây đã có thể ứng dụng vào CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co, giúp mang lại nhiều hiệu quả thiết thực cho người sử dụng CSDL điện tử. Việc nghiên cứu, xây dựng CSDL điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co là cần thiết và mang lại nhiều lợi ích, giá trị to lớn cho cộng đồng. Đây sẽ là nơi lưu giữ ngôn ngữ truyền thống của các dân tộc thiểu số trên địa bàn tỉnh, đồng thời qua đó cũng sẽ lưu giữ các giá trị văn hóa truyền thống, hình ảnh, phong tục, tập quán và văn hoá của người Hrê, Co, góp phần bảo tồn và phát huy các giá trị truyền thống độc đáo của mỗi dân tộc, cùng hòa vào dòng chảy phát triển chung của tỉnh Quảng Ngãi nói riêng và đất nước nói chung, bởi trong quá trình chuyển đổi số nhân văn và rộng khắp, sẽ không có ai bị bỏ lại phía sau.

LỜI CẢM ƠN

Bài báo được thực hiện trong khuôn khổ đề tài KH&CN cấp tỉnh "Xây dựng Cơ sở dữ liệu điện tử tiếng đồng bào dân tộc thiểu số Việt - Hrê, Việt - Co" do Trung tâm Chuyển đổi số và Đổi mới sáng tạo - Sở Khoa học và Công nghệ tỉnh Quảng Ngãi phối hợp cùng Sở Dân tộc và Tôn giáo tỉnh Quảng Ngãi chủ trì thực hiện, mã số đề tài: 03/2023/HĐ-ĐTKHCN.

TÀI LIỆU THAM KHẢO

[1] United Nations and the General Statistics Office (2024a), "Hre ethnic population in Vietnam", https://danso.info/dan-so-dan-toc-hre-o-viet-nam, accessed 29 July 2025 (in Vietnamese).

[2] United Nations and the General Statistics Office (2024b), "Co ethnic population in Vietnam", https://danso.info/dan-so-dan-toc-co-o-viet-nam, accessed 29 July 2025 (in Vietnamese).

[3] People's Committee of Quang Ngai Province (2007), *Temporarily Issued with Decision No. 893/QD-UBND dated April 27, 2007 on Training and Fostering Materials in H're Language* (in Vietnamese).

[4] N.M. Tri (2016), *Lessons in Co language,* Part of the provincial-level science and technology project: Research and compilation of teaching materials for Co language for Kinh public servants working in Tra Bong and Tay Tra districts, Quang Ngai province, Quang Ngai Department of Science and Technology (in Vietnamese).

[5] People's Committee of Quang Ngai Province (2024), "Database of ethnic minority languages Viet - Hre, Viet - Co", https://csdlhreco.nuian.vn, accessed 30 April 2025 (in Vietnamese).

[6] H. Purwins, B. Li, T. Virtanen, et al. (2019), "Deep learning for audio signal processing", *IEEE Journal of Selected Topics in Signal Processing*, **13(2)**, pp.206-219, DOI: 10.1109/JSTSP.2019.2908700.

[7] A. Mehrish, N. Majumder, R. Bharadwaj, et al. (2023), "A review of deep learning techniques for speech processing", *Information Fusion*, **99**, DOI: 10.1016/j.inffus.2023.101869.

[8] A. Alsobhani, H.M.A. ALabboodi, H. Mahdi (2021), "Speech recognition using convolution deep neural networks", *Journal of Physics: Conference Series*, **1973(1)**, DOI: 10.1088/1742-6596/1973/1/012166.

[9] A. Vaswani, N. Shazeer, N. Parmar, et al. (2017), "Attention is all you need", *Advances in Neural Information Processing Systems*, 11pp, DOI: 10.48550/arXiv.1706.03762.

[10] A. Baevski, H. Zhou, A. Mohamed, et al. (2020), "Wav2vec 2.0: A framework for self-supervised learning of speech representations", *Advances in Neural Information Processing Systems*, 19pp, DOI: 10.48550/arXiv.2006.11477.

[11] W.N. Hsu, B. Bolte, Y.H.H. Tsai, et al. (2021), "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units", *IEEE*, **29**, pp.3451-3460, DOI: 10.1109/TASLP.2021.3122291.

[12] L. Lugosch, T. Likhomanenko, G. Synnaeve, et al. (2022), "Pseudo-labeling for massively multilingual speech recognition", *IEEE*, pp.7687-7691, DOI: 10.1109/ICASSP43922.2022.9746832.

[13] A. Conneau, A. Baevski, R. Collobert, et al. (2021), "Unsupervised cross-lingual representation learning for speech recognition", *ISCA*, pp.2426-2430, DOI: 10.21437/Interspeech.2021-329.

[14] A. Radford, J.W. Kim, T. Xu, et al. (2022), "Robust speech recognition via large-scale weak supervision", *Proceedings of Machine Learning Research*, 28pp, DOI: 10.48550/arXiv.2212.04356.

[15] OpenAI (2022), "Introducing Whisper", https://openai.com/index/whisper/, accessed 25 December 2024.

[16] Hugging Face (2024), "Create a new dataset repository", https://huggingface.co/new-dataset, accessed 25 December 2024.

[17] Quang Ngai Newspaper (2024), "Introducing the electronic database of ethnic languages", https://baoquangngai.vn/van-hoa/202408/gioi-thieu-co-so-du-lieu-dien-tu-tieng-dong-bao-dan-toc-c0a2512, accessed 25 December 2024 (in Vietnamese).