ĐỀ XUẤT KIẾN TRÚC HỆ THỐNG AI HỖ TRỢ LUYỆN PHÁT ÂM TIẾNG ANH CHO SINH VIÊN

Huỳnh Phước Danh, Nguyễn Mạnh Hoàng Nguyên, Nguyễn Tuấn Đạt
Khoa Kỹ thuật Cơ – Điện và Máy tính, Trường Đại học Văn Lang, TP. Hồ Chí Minh
Email: danh.hp@vlu.edu.vn

Tóm tắt: *Nghiên cứu này đề xuất kiến trúc và quy trình vận hành của một hệ thống hỗ trợ luyện phát âm tiếng Anh cho sinh viên dựa trên học sâu và xử lý tín hiệu giọng nói. Xuất phát từ thực tế lớp học đại học có số lượng người học đông, giảng viên khó theo dõi thường xuyên từng lượt luyện phát âm của sinh viên, nghiên cứu định hướng xây dựng một mô hình phần mềm gồm ứng dụng người dùng, backend/API và AI Worker. Hệ thống tiếp nhận bản ghi âm của sinh viên, thực hiện tiền xử lý âm thanh, kiểm tra nội dung đọc, trích xuất đặc trưng, căn chỉnh âm vị và tạo phản hồi bước đầu về điểm luyện tập hoặc âm vị cần chú ý. Khác với cách tiếp cận xem AI như công cụ chấm điểm tự động hoàn toàn, bài viết nhấn mạnh vai trò của giảng viên trong giao bài, xem kết quả, nhận xét và định hướng luyện tập. Bài báo tập trung vào khung thiết kế, quy trình kỹ thuật và bộ tiêu chí đánh giá dự kiến, làm cơ sở cho giai đoạn triển khai thực nghiệm tiếp theo trong môi trường đại học Việt Nam.*

Từ khóa: *trí tuệ nhân tạo; học sâu; luyện phát âm tiếng Anh; công nghệ giáo dục; căn chỉnh âm vị.*

PROPOSING AN AI-BASED SYSTEM ARCHITECTURE FOR ENGLISH PRONUNCIATION PRACTICE IN HIGHER EDUCATION

Abstract: *This paper proposes an architecture and operating workflow for a deep-learning-based system that supports English pronunciation practice in higher education. The system is designed to connect student practice data, automatic speech processing, and teacher feedback through three main components: a user application, a backend/API layer, and an AI Worker. The proposed workflow includes audio submission, preprocessing, reading-content verification, acoustic feature extraction, phoneme alignment, and initial feedback generation. Rather than positioning AI as a fully automatic pronunciation scoring tool, the paper emphasizes its role as a pedagogical assistant that supports teachers in assigning practice tasks, monitoring student progress, and providing follow-up guidance. The contribution of the paper lies in the proposed system architecture, technical workflow, and projected evaluation framework for future empirical implementation in Vietnamese higher education.*

Keywords: *artificial intelligence; deep learning; English pronunciation practice; educational technology; phoneme alignment.*

Nhận bài: 15/05/2026 Phản biện: 15/06/2026 Duyệt đăng: 19/06/2026

I. ĐẶT VẤN ĐỀ

Phát âm là một thành tố quan trọng trong năng lực giao tiếp bằng tiếng Anh. Đối với sinh viên Việt Nam, việc luyện phát âm thường gặp khó khăn do sự khác biệt giữa hệ âm vị tiếng Việt và tiếng Anh, đặc biệt ở phụ âm cuối, nguyên âm, tổ hợp phụ âm và một số âm vị không quen thuộc. Trong lớp học đông, giảng viên khó nghe và phản hồi chi tiết cho từng sinh viên sau mỗi lượt luyện tập. Điều này làm giảm cơ hội cá nhân hóa phản hồi, trong khi phát âm lại là kỹ năng cần luyện lặp lại và điều chỉnh liên tục.

Các ứng dụng học tiếng Anh hiện nay đã sử dụng nhiều công nghệ nhận dạng giọng nói, nhưng không phải hệ thống nào cũng cung cấp phản hồi sư phạm phù hợp với hoạt động dạy học ở bậc đại học. Một số công cụ thiên về chấm điểm tức thời cho cá nhân người học, ít chú trọng dữ liệu lớp học, vai trò nhận xét của giảng viên và khả năng giao lại bài luyện theo lỗi thường gặp. Vì vậy, cần có một hướng tiếp cận đặt AI vào vị trí trợ lý học tập và trợ lý sư phạm, hỗ trợ cả sinh viên lẫn giảng viên.

Từ bối cảnh trên, bài báo đề xuất kiến trúc hệ thống AI hỗ trợ luyện phát âm tiếng Anh cho sinh viên. Đóng góp của bài viết nằm ở ba điểm: xác lập mô hình tổ chức hệ thống phù hợp với môi trường giáo dục đại học; mô tả quy trình xử lý bản ghi âm từ tiếp nhận dữ liệu đến phản hồi học tập; và đề xuất bộ tiêu chí đánh giá dự kiến để làm cơ sở cho giai đoạn thực nghiệm tiếp theo.

II. CƠ SỞ KHOA HỌC VÀ PHƯƠNG PHÁP TIẾP CẬN

Đánh giá phát âm tự động là lĩnh vực kết hợp xử lý tín hiệu giọng nói, mô hình học máy và công nghệ giáo dục. Phương pháp Goodness of Pronunciation (GOP) được xem là một trong những hướng truyền thống để ước lượng mức phù hợp giữa phát âm của người học và âm vị mục tiêu. Tuy nhiên, GOP phụ thuộc nhiều vào mô hình âm học và bước căn chỉnh âm vị, do đó trong hệ thống đề xuất, phương pháp này được xem như cơ sở lý thuyết hoặc baseline tham khảo, không phải thành phần duy nhất quyết định phản hồi.

Các nghiên cứu gần đây cho thấy mô hình học sâu có khả năng khai thác biểu diễn âm thanh phức tạp, phục vụ phát hiện lỗi phát âm ở cấp độ âm vị hoặc mức từ, câu. Các mô hình học biểu diễn âm thanh như wav2vec 2.0 có thể hỗ trợ kiểm tra nội dung đọc hoặc trích xuất đặc trưng. Tuy vậy, kết quả nhận dạng văn bản không được đồng nhất với điểm phát âm, bởi người học có thể đọc đúng nội dung nhưng vẫn phát âm chưa chuẩn ở một số âm vị.

Về dữ liệu, các bộ ngữ liệu phát âm của người học tiếng Anh không bản ngữ như L2-ARCTIC có giá trị cho bài toán phát hiện lỗi phát âm. CMU ARCTIC có thể được dùng làm nguồn tham chiếu phát âm bản ngữ khi cần đối chiếu. Common Voice và LibriSpeech có ích cho thử nghiệm nhận dạng giọng nói hoặc học biểu diễn âm thanh, nhưng không phải nguồn chính cho lỗi phát âm ở cấp độ âm vị.

Bài viết sử dụng phương pháp nghiên cứu tài liệu, phân tích thiết kế hệ thống và mô hình hóa quy trình xử lý dữ liệu âm thanh. Trên cơ sở tổng hợp các nghiên cứu về đánh giá phát âm tự động, mô hình học sâu trong xử lý giọng nói và ứng dụng công nghệ giáo dục, bài báo đề xuất kiến trúc hệ thống thử nghiệm phù hợp với bối cảnh luyện phát âm tiếng Anh trong môi trường đại học. Phương pháp tiếp cận của bài viết không nhằm công bố một mô hình thuật toán mới, mà tập trung làm rõ cách tổ chức các thành phần kỹ thuật, dữ liệu và vai trò sư phạm của giảng viên trong hệ thống hỗ trợ luyện phát âm.

Từ các nghiên cứu liên quan có thể thấy, nhiều hướng tiếp cận hiện nay tập trung vào nâng cao độ chính xác của mô hình phát hiện lỗi phát âm hoặc chấm điểm phát âm tự động. Tuy nhiên, các nghiên cứu này chưa nhấn mạnh đầy đủ đến bài toán tổ chức dữ liệu luyện tập trong lớp học đông, nơi giảng viên cần theo dõi nhiều sinh viên, nhận biết lỗi phổ biến và giao lại bài luyện phù hợp. Đây là điểm khác biệt trong hướng tiếp cận của bài viết: hệ thống không chỉ xử lý bản ghi âm của từng cá nhân, mà còn được thiết kế để kết nối dữ liệu luyện tập với hoạt động quản lý lớp học và phản hồi sư phạm của giảng viên.

III. KIẾN TRÚC VÀ QUY TRÌNH HỆ THỐNG ĐỀ XUẤT

3.1. Yêu cầu thiết kế trong môi trường đại học

Trong môi trường đại học, hệ thống hỗ trợ luyện phát âm cần đáp ứng đồng thời hai nhóm người dùng là sinh viên và giảng viên. Sinh viên cần giao diện đơn giản để nhận bài luyện, ghi âm, nộp bài, xem phản hồi và luyện lại. Giảng viên cần công cụ tạo bài luyện, quản lý lớp, xem bài nộp, nhận biết lỗi phổ biến và đưa ra nhận xét bổ sung. Vì vậy, hệ thống không chỉ xử lý âm thanh mà còn phải lưu trữ dữ liệu học tập theo người học, lớp học và bài luyện.

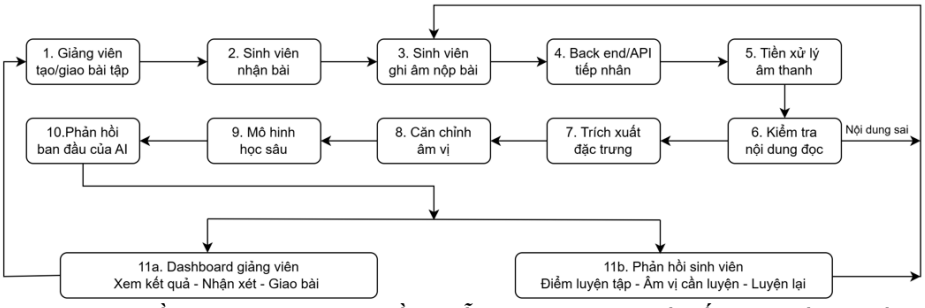
Một yêu cầu quan trọng khác là khả năng thử nghiệm và thay thế mô hình AI. Trong giai đoạn nghiên cứu, nhóm phát triển có thể cần thay đổi cách tiền xử lý, dạng đặc trưng âm học, mô hình học sâu hoặc công cụ căn chỉnh âm vị. Do đó, kiến trúc hệ thống nên tách riêng thành phần AI Worker khỏi backend nghiệp vụ để dễ bảo trì và nâng cấp.

3.2. Thành phần hệ thống

Hệ thống được tổ chức theo mô hình client-server với ba thành phần chính: ứng dụng người dùng, backend/API và AI Worker. Ứng dụng người dùng phục vụ thao tác của sinh viên và giảng viên. Backend/API quản lý tài khoản, lớp học, bài luyện, file ghi âm, lịch sử nộp bài và kết quả phản hồi. AI Worker tiếp nhận yêu cầu phân tích âm thanh, thực hiện tiền xử lý, trích xuất đặc trưng, căn chỉnh âm vị và gọi mô hình học sâu để tạo kết quả phản hồi.

Bảng 1. Vai trò các thành phần trong hệ thống

Thành phần	Vai trò chính	Ý nghĩa trong dạy học
Ứng dụng người dùng	Cung cấp giao diện ghi âm, nộp bài, xem kết quả cho sinh viên và dashboard quản lý lớp học cho giảng viên.	Tạo môi trường luyện tập liên tục cho sinh viên và công cụ theo dõi, đánh giá, giao lại bài cho giảng viên.
Backend/API	Quản lý dữ liệu người dùng, bài luyện, file âm thanh, kết quả phân tích và lịch sử học tập.	Bảo đảm dữ liệu luyện tập được tổ chức theo lớp, bài và từng sinh viên.
AI Worker	Xử lý âm thanh, trích xuất đặc trưng, căn chỉnh âm vị, gọi mô hình học sâu và tạo phản hồi.	Cung cấp gợi ý ban đầu để sinh viên luyện lại và dữ liệu học tập để giảng viên tham khảo khi nhận xét.



Hình 1. Lưu đồ quy trình xử lý, phản hồi và hỗ trợ giảng dạy của hệ thống AI luyện phát âm

3.3. Quy trình xử lý bản ghi âm và tạo phản hồi ban đầu

Quy trình vận hành bắt đầu từ bài luyện do giảng viên tạo hoặc lựa chọn. Sinh viên ghi âm từ, cụm từ hoặc câu ngắn theo yêu cầu và gửi file lên hệ thống. Backend/API kiểm tra thông tin bài luyện, định dạng file và chuyển yêu cầu sang AI Worker. Tại AI Worker, bản ghi được chuẩn hóa định dạng, tần số lấy mẫu và âm lượng; các đoạn âm thanh không hợp lệ, quá ngắn hoặc quá nhiễu được đánh dấu để hạn chế phản hồi sai.

Sau tiền xử lý, hệ thống thực hiện kiểm tra nội dung đọc bằng mô hình nhận dạng giọng nói hoặc mô hình biểu diễn âm thanh. Mục đích của bước này là xác định người học có đọc đúng hoặc gần đúng nội dung bài luyện trước khi phân tích phát âm ở cấp độ âm vị. Các đặc trưng âm học phục vụ cho kiểm tra nội dung và phân tích phát âm có thể được trích xuất theo từng cấu hình thử nghiệm, chẳng hạn MFCC, spectrogram hoặc biểu diễn học được từ mô hình tiền huấn luyện.

Tại bước kiểm tra nội dung đọc, hệ thống cần xác định mức độ phù hợp giữa nội dung yêu cầu và nội dung sinh viên thực sự đọc. Trường hợp bản ghi sai hoàn toàn so với bài luyện, chẳng hạn bài yêu cầu đọc một từ hoặc câu cụ thể nhưng sinh viên đọc nội dung khác, hệ thống không nên tiếp tục đưa bản ghi vào bước căn chỉnh âm vị và phân tích phát âm. Thay vào đó, AI Worker trả về trạng thái "nội dung đọc không phù hợp", đồng thời gợi ý sinh viên ghi âm lại đúng bài luyện. Cách xử lý này giúp hạn chế phản hồi sai, giảm lãng phí tài nguyên tính toán và bảo đảm kết quả phân tích chỉ được tạo ra khi dữ liệu đầu vào đạt điều kiện tối thiểu.

Ở bước phân tích phát âm, căn chỉnh âm vị giúp xác định vùng âm thanh tương ứng với âm vị mục tiêu trong từ hoặc câu luyện tập. Kết quả căn chỉnh được đưa vào mô hình học sâu để tạo điểm luyện tập hoặc gợi ý nhóm âm vị cần chú ý. Phản hồi được trả về cho sinh viên dưới dạng gợi ý luyện lại; đồng thời dữ liệu kết quả được lưu trữ để tổng hợp trên dashboard giảng viên, phục vụ theo dõi tiến độ, nhận biết lỗi phổ biến và giao lại bài luyện khi cần.

3.4. Chức năng giáo dục dành cho sinh viên và giảng viên

Đối với sinh viên, hệ thống hình thành chu trình học tập gồm nhận bài, ghi âm, gửi bài, xem phản hồi và luyện lại. Phản hồi nên thể hiện theo hướng gợi ý: âm nào cần chú ý, từ, câu nào cần luyện lại hoặc tiến độ luyện tập qua các lần nộp. Cách trình bày này giúp người học nhận biết mục tiêu luyện tập cụ thể thay vì chỉ nhìn một điểm số tổng quát.

Đối với giảng viên, hệ thống đóng vai trò thu thập, tổ chức và trực quan hóa dữ liệu học tập thông qua dashboard. Giảng viên có thể xem sinh viên đã luyện tập hay chưa, kết quả từng lần nộp, tiến độ luyện tập, nhóm lỗi phát âm phổ biến trong lớp và mức độ cải thiện của từng người học. Từ đó, giảng viên có thêm cơ sở để đánh giá, nhận xét, giao lại bài hoặc điều chỉnh hoạt động trên lớp.

Về mặt sư phạm, AI không thay thế vai trò đánh giá chuyên môn của giảng viên. Kết quả tự động chỉ là gợi ý ban đầu, giúp giảm tải thao tác kỹ thuật và mở rộng cơ hội luyện tập ngoài lớp. Quyết định cuối cùng về nhận xét, điều chỉnh bài học và định hướng luyện tập vẫn thuộc về giảng viên.

3.5. Dữ liệu và tiêu chí đánh giá dự kiến

Để chuẩn bị cho giai đoạn thực nghiệm, dữ liệu cần được phân nhóm theo vai trò sử dụng. Nguồn dữ liệu phát âm của người học tiếng Anh không bản ngữ được ưu tiên cho huấn luyện và đánh giá phát hiện lỗi phát âm. Nguồn phát âm bản ngữ được dùng để tham chiếu khi cần. Dữ liệu phụ như Common Voice hoặc LibriSpeech hỗ trợ kiểm tra nội dung đọc hoặc học biểu diễn âm thanh. Ngoài ra, cần thu thập một tập ghi âm quy mô nhỏ từ sinh viên Việt Nam để đánh giá tính phù hợp của hệ thống trong môi trường sử dụng thực tế.

Bảng 2. Đề xuất phân nhóm dữ liệu phục vụ nghiên cứu

Nhóm dữ liệu	Ví dụ nguồn dữ liệu	Vai trò dự kiến
Dữ liệu phát âm tiếng Anh của người học không bản ngữ	L2-ARCTIC	Huấn luyện, thử nghiệm và đánh giá phát hiện lỗi phát âm ở cấp độ âm vị.
Dữ liệu phát âm bản ngữ	CMU ARCTIC	Tham chiếu phát âm khi cần đối chiếu hoặc kiểm thử.
Dữ liệu phụ	Common Voice, LibriSpeech	Hỗ trợ kiểm tra nội dung đọc hoặc học biểu diễn âm thanh; không dùng làm nguồn nhãn chính cho lỗi âm vị.
Dữ liệu sinh viên Việt Nam	Bản ghi do sinh viên thực hiện trong thử nghiệm nhỏ	Đánh giá khả năng vận hành, độ phù hợp phản hồi và trải nghiệm người dùng.

Bộ tiêu chí đánh giá dự kiến gồm hai nhóm. Nhóm kỹ thuật có thể bao gồm Accuracy, Precision, Recall, F1-score, confusion matrix, Kappa, MCC, thời gian xử lý trung bình và tỷ lệ file xử lý thành công. Nhóm giáo dục tập trung vào mức độ dễ sử dụng, tính dễ hiểu của phản hồi, khả năng hỗ trợ sinh viên tự luyện và khả năng hỗ trợ giảng viên theo dõi lớp học. Các tiêu chí này sẽ là cơ sở để triển khai thử nghiệm trong giai đoạn tiếp theo.

Các chỉ số đánh giá kỹ thuật cần được gắn với từng thành phần cụ thể của AI Worker. Đối với bước kiểm tra nội dung đọc, có thể đánh giá tỷ lệ nhận diện đúng nội dung, tỷ lệ từ chối hợp lý đối với bản ghi sai nội dung và tỷ lệ bản ghi cần yêu cầu ghi âm lại. Đối với bước phân tích phát âm hoặc phát hiện lỗi âm vị, các chỉ số như Precision, Recall, F1-score, confusion matrix, Kappa và MCC được dùng để so sánh kết quả dự đoán của mô hình với nhãn tham chiếu do giảng viên hoặc chuyên gia xác nhận. Đối với toàn bộ quy trình xử lý, các chỉ số như thời gian xử lý trung bình, tỷ lệ file xử lý thành công và tỷ lệ file bị loại do nhiễu hoặc sai định dạng được sử dụng để đánh giá khả năng vận hành của hệ thống.

IV. THẢO LUẬN

Giải pháp đề xuất cho thấy tiềm năng kết nối hoạt động tự luyện phát âm của sinh viên với quy trình tổ chức, theo dõi và hỗ trợ học tập của giảng viên. Thay vì chỉ xem trí tuệ nhân tạo như một công cụ chấm điểm tự động, hệ thống được định hướng như một thành phần hỗ trợ sư phạm, cung cấp dữ liệu luyện tập ban đầu để giảng viên có thêm căn cứ nhận xét, giao bài bổ sung và điều chỉnh hoạt động dạy học. Cách tiếp cận này phù hợp với bối cảnh lớp học có số lượng sinh viên đông, nơi giảng viên khó có đủ thời gian để nghe và sửa lỗi phát âm chi tiết cho từng người học.

Về mặt kỹ thuật, kiến trúc tách biệt giữa ứng dụng người dùng, backend/API và AI Worker giúp hệ thống dễ mở rộng, thuận lợi cho việc thay đổi hoặc nâng cấp mô hình phân tích phát âm mà không ảnh hưởng lớn đến toàn bộ phần mềm. Quy trình xử lý gồm tiếp nhận bản ghi, kiểm tra chất lượng âm thanh, tiền xử lý tín hiệu, kiểm tra nội dung đọc, căn chỉnh âm vị và sinh phản hồi được thiết kế nhằm bảo đảm dữ liệu đầu vào có đủ điều kiện trước khi đưa vào mô hình phân tích. Điều này đặc biệt cần thiết trong môi trường học tập thực tế, nơi chất lượng ghi âm có thể chịu ảnh hưởng bởi thiết bị, nhiễu nền, âm lượng và sự khác biệt giọng nói giữa các sinh viên.

Trong phạm vi bài viết này, nghiên cứu tập trung vào việc đề xuất kiến trúc hệ thống và quy trình vận hành ở mức thử nghiệm. Việc đánh giá định lượng về độ chính xác của mô hình, thời gian xử lý, tỷ lệ bản ghi lỗi và mức độ hài lòng của người dùng cần được tiếp tục triển khai trong các nghiên cứu tiếp theo. Giai đoạn thực nghiệm sau có thể thực hiện trên tập ghi âm của sinh viên Việt Nam, kết hợp đối chiếu kết quả phân tích của hệ thống với nhận xét của giảng viên hoặc chuyên gia phát âm, đồng thời khảo sát trải nghiệm sử dụng của sinh viên và giảng viên trong lớp học.

Một vấn đề cần cân nhắc khi triển khai hệ thống là chi phí hạ tầng tính toán. Các mô hình học sâu trong xử lý giọng nói có thể đòi hỏi tài nguyên cao, đặc biệt khi xử lý nhiều bản ghi cùng lúc. Để phù hợp với điều kiện triển khai tại các cơ sở giáo dục, hệ thống nên ưu tiên kiến trúc xử lý bất đồng bộ (Asynchronous), trong đó bản ghi được đưa vào hàng đợi và AI Worker xử lý theo lượt thay vì yêu cầu phản hồi tức thời cho mọi tác

vụ. Bên cạnh đó, có thể xem xét sử dụng mô hình nhẹ, mô hình rút gọn hoặc cơ chế xử lý theo lô (batch processing) để giảm chi phí máy chủ, đồng thời vẫn bảo đảm trải nghiệm sử dụng ở mức chấp nhận được cho sinh viên và giảng viên.

V. KẾT LUẬN

Bài báo đã đề xuất kiến trúc và quy trình hệ thống AI hỗ trợ luyện phát âm tiếng Anh cho sinh viên. Hệ thống được thiết kế theo hướng kết hợp ứng dụng người dùng, backend/API và AI Worker, trong đó AI tạo phản hồi ban đầu còn giảng viên giữ vai trò định hướng sư phạm. Đóng góp chính của bài viết là làm rõ cách tổ chức hệ thống, quy trình xử lý bản ghi âm và bộ tiêu chí đánh giá dự kiến cho giai đoạn thực nghiệm.

Trong phạm vi nghiên cứu này, kết quả chính của bài báo là khung kiến trúc và quy trình vận hành hệ thống. Các nghiên cứu tiếp theo cần triển khai thử nghiệm với sinh viên Việt Nam, bổ sung số liệu kỹ thuật và khảo sát người dùng để đánh giá mức độ khả thi, độ tin cậy và giá trị sư phạm của hệ thống trong môi trường đại học.

TÀI LIỆU THAM KHẢO

Lê Thanh Hòa, "Lỗi phát âm phụ âm cuối và tổ hợp phụ âm cuối tiếng Anh của sinh viên Trường Đại học Đồng Nai," Tạp chí Ngôn ngữ và Đời sống, số 10, tr. 75-81, 2014.

Bùi Thị Ngọc Lan, "Nghiên cứu về phát âm tiếng Anh của sinh viên Đại học Phan Thiết," Tạp chí Khoa học Đại học Phan Thiết, tập 2, số 2, 2024.

Vũ Thị Hải Hà và Hạp Thu Hà, "Đối chiếu nguyên âm tiếng Việt và tiếng Anh nhằm khắc phục những lỗi phát âm tiếng Anh thường gặp ở người Việt học tiếng Anh," Tạp chí Khoa học Đại học Hạ Long, số 08, tr. 75-85, 2023.

S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, vol. 30, no. 2-3, pp. 95-108, 2000.

W.-K. Lin and H.-M. Wang, "Deep learning based mispronunciation detection and diagnosis using wav2vec 2.0," in Proceedings of Interspeech 2022, 2022, pp. 1404-1408.

W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in Proceedings of ICASSP 2019, 2019, pp. 8132-8136.

W. Hu, Y. Qian, and F. K. Soong, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, vol. 67, pp. 154-166, 2015.

I. Shahin and B. Ahmed, "Phonological-Level Mispronunciation Detection and Diagnosis for Non-Native English Speakers," in Proceedings of Interspeech 2024, 2024, pp. 307-311.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 12449-12460.

G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-Native English Speech Corpus," in Proceedings of Interspeech 2018, 2018, pp. 2783-2787.

J. Kominek and A. W. Black, "The CMU Arctic Speech Databases," in Proceedings of the ISCA Speech Synthesis Workshop, 2004, pp. 223-224.

R. Ardila et al., "Common Voice: A Massively-Multilingual Speech Corpus," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4218-4222.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in Proceedings of ICASSP 2015, 2015, pp. 5206-5210.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Proceedings of Interspeech 2017, 2017, pp. 498-502.