HỆ THỐNG MÔ PHỎNG GIỌNG NÓI BẰNG TRÍ TUỆ NHÂN TẠO TRONG THIẾT KẾ HỌC LIỆU ĐIỆN TỬ

Đặng Hải Đăng[1*], Quách Thị Hạnh[1], Nguyễn Văn Hoàng[2], Nguyễn Đức Tín[2]
[*]Tác giả liên hệ, email: dangdh@hou.edu.vn. ORCID: 0009-0002-5290-0556

Ngày tòa soạn nhận được bài báo: 15/01/2026
Ngày phản biện đánh giá: 17/03/2026
Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1174

Tóm tắt: *Trong bối cảnh đào tạo trực tuyến phát triển mạnh, học liệu điện tử cần được thiết kế theo hướng cá nhân hóa và dễ cập nhật. Học liệu âm thanh (giọng thuyết minh/bài giảng) là thành phần quan trọng nhưng quy trình thu âm truyền thống còn tốn thời gian, phụ thuộc nhân lực và khó bảo đảm tính nhất quán khi chỉnh sửa nội dung. Nghiên cứu này phát triển hệ thống chuyển văn bản thành giọng nói (Text-to-Speech - TTS) ứng dụng trí tuệ nhân tạo có khả năng mô phỏng giọng giảng viên nhằm rút ngắn, tiến tới giảm nhu cầu thu âm khi sản xuất học liệu. Hệ thống theo kiến trúc client-server, tích hợp mô hình viXTTS tinh chỉnh trên bộ dữ liệu viVoice để tạo âm thanh theo giọng tham chiếu. Kết quả thử nghiệm cho thấy độ tương đồng giữa giọng mô phỏng và giọng gốc đạt trên 95% và duy trì tính nhất quán giữa các mẫu giọng. Nghiên cứu góp phần đề xuất một quy trình kỹ thuật khả thi để tích hợp voice cloning vào xây dựng học liệu điện tử có âm thanh trong bối cảnh tiếng Việt, đồng thời cung cấp cơ sở cho các thử nghiệm triển khai tại cơ sở giáo dục.*

Từ khóa: *nhân bản giọng nói, chuyển văn bản thành giọng nói, trí tuệ nhân tạo, học liệu điện tử, giáo dục số*

I. Mở đầu

1.1. Bối cảnh nghiên cứu

Chuyển đổi số giáo dục làm gia tăng nhu cầu học liệu điện tử có mức độ cá nhân hóa cao, dễ cập nhật và có khả năng hỗ trợ người học trong môi trường trực tuyến. Trong các thành phần đa phương tiện, giọng nói thuyết minh/bài giảng đóng vai trò then chốt trong việc dẫn dắt nội dung và giảm tải nhận thức, song quy trình thu âm thủ công vẫn là nút thắt về thời gian, chi phí và tính đồng nhất khi cần chỉnh sửa nội dung.

Những tiến bộ gần đây trong tổng hợp tiếng nói và nhân bản giọng nói (voice cloning) cho phép mô phỏng giọng nói với

[1] Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam
[2] Sinh viên ngành Công nghệ Thông tin, Viện Đào tạo và Phát triển Học tập Suốt đời, Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

dữ liệu tham chiếu hạn chế (few-shot/zero-shot) đạt chất lượng ngày càng tự nhiên. Các tổng quan gần đây cũng đã chuẩn hóa thuật ngữ, hệ thống hóa hướng tiếp cận và thước đo đánh giá, đồng thời nhấn mạnh nhu cầu nghiên cứu về phát hiện và giảm thiểu lạm dụng giọng giả (Azzuni & Saddik, 2025; Roosadi và cộng sự, 2024).

Một số nghiên cứu tại Việt Nam cho thấy tính khả thi của việc áp dụng các mô hình học sâu vào xử lý tiếng nói với dữ liệu hạn chế, bao gồm nhận dạng tiếng nói cho ngôn ngữ/dữ liệu đặc thù và các hướng thích nghi giọng trong tổng hợp tiếng Việt (Nguyễn & Trần, 2024; Phạm, 2023). Các công trình nghiên cứu trên thế giới về voice cloning và TTS cho ngôn ngữ tài nguyên thấp tiếp tục mở rộng theo hướng giảm yêu cầu dữ liệu tham chiếu, tăng tính nhất quán và khả năng chuyển miền/ngôn ngữ (Zhu, 2025; Qiao và cộng sự, 2023). Những kết quả này tạo tiền đề để nghiên cứu tích hợp mô phỏng giọng nói vào học liệu điện tử, đặc biệt trong bối cảnh tiếng Việt.

1.2. Khoảng trống trong các nghiên cứu hiện có

Tổng quan tài liệu cho thấy, dù chất lượng mô hình TTS/voice cloning đã cải thiện rõ rệt, vẫn thiếu các nghiên cứu mô tả đầy đủ một quy trình tạo ra hệ thống hướng đích học liệu: từ xử lý văn bản bài giảng, lựa chọn cấu hình sinh giọng, hậu xử lý và quản trị tệp âm thanh đến cơ chế đánh giá phù hợp với yêu cầu sư phạm.

Thứ nhất, phần lớn công trình tập trung vào tối ưu mô hình hoặc thước đo âm học, trong khi các yêu cầu đặc thù của bài giảng số (tính nhất quán trên nội dung dài, nhịp điệu/nhấn nhá theo phong cách giảng dạy, khả năng cập nhật nhanh khi thay đổi nội dung) chưa được mô hình hóa như một bài toán hệ thống (Azzuni & Saddik, 2025; Zhu, 2025).

Thứ hai, cách đánh giá hiện hành thường ưu tiên các thang đo kỹ thuật (ví dụ MOS/CMOS, chỉ số méo phổ), nhưng còn thiếu tiêu chí phản ánh hiệu quả sư phạm và trải nghiệm người học trong E-learning (độ rõ, tốc độ phù hợp, phân đoạn hợp lý, khả năng duy trì chú ý), vốn có thể không tương quan hoàn toàn với điểm tự nhiên âm học (Marty-Dugas và cộng sự, 2024; Liew và cộng sự, 2023).

1.3. Lý do lựa chọn nghiên cứu

Xuất phát từ nhu cầu mở rộng học liệu âm thanh trong đào tạo trực tuyến và các hạn chế của thu âm thủ công, nghiên cứu lựa chọn hướng tiếp cận tích hợp TTS/voice cloning để tự động hóa khâu sản xuất giọng đọc theo giọng giảng viên. Trọng tâm của bài báo là (i) thiết kế và hiện thực hóa hệ thống tạo học liệu âm thanh theo kiến trúc client-server; (ii) lựa chọn mô hình viXTTS/viVoice cho tiếng Việt; và (iii) đề xuất quy trình đánh giá định lượng mức độ tương đồng và tính nhất quán của giọng mô phỏng, làm nền tảng cho thử nghiệm triển khai tại cơ sở đào tạo.

Về bản chất, quy trình thu âm truyền thống tồn tại nhiều hạn chế: tốn kém thời gian, chi phí nhân lực cao, phụ thuộc vào thiết bị và điều kiện thu âm, khó đảm bảo tính nhất quán khi cập nhật nội dung, cũng như thiếu khả năng cá nhân hóa theo phong cách giảng viên hoặc nhu cầu người học. Các mô hình hiện đại như VITS, Transformer, Diffusion hay Voice Conversion cho phép tạo ra giọng nói tự nhiên với dữ liệu tham chiếu hạn chế (zero-shot hoặc few-shot TTS), giúp tự động hóa việc sản xuất học liệu âm thanh mà không cần thu âm lại nhiều lần.

Từ tổng quan nghiên cứu, nhóm tác giả nhận thấy các công bố trong nước và quốc tế chứng minh tính khả thi của công nghệ này, nhưng vẫn tồn tại khoảng trống lớn: thiếu mô hình được thiết kế riêng cho bối cảnh sư phạm, đặc biệt với tiếng Việt - một ngôn ngữ tài nguyên thấp - chưa có hệ thống tích hợp nhân bản giọng nói vào học liệu điện tử và phương pháp đánh giá chất lượng chưa chú trọng đến trải nghiệm người học như tốc độ nói, độ nhấn nhá, cảm xúc phù hợp với bài giảng cần truyền tải.

II. Cơ sở lý thuyết

2.1. Tổng quan về công nghệ chuyển văn bản thành giọng nói TTS

Tổng hợp tiếng nói dựa trên học sâu đã chuyển dịch từ các luồng nhiều mô-đun sang các mô hình end-to-end, giúp cải thiện độ tự nhiên, độ rõ và khả năng điều khiển ngữ điệu. Trong bài toán học liệu, trọng tâm không chỉ là đọc đúng văn bản mà còn là tái tạo phong cách giảng dạy ổn định theo giọng tham chiếu, từ đó giảm chi phí sản xuất và hỗ trợ cập nhật nhanh nội dung.

Các hướng tiếp cận nhân bản giọng nói hiện đại cho phép tái tạo đặc trưng người nói với lượng dữ liệu tham chiếu nhỏ, tạo tiền đề xây dựng hệ thống sinh giọng phục vụ nội dung dài và cần tính nhất quán cao. Do đó, nghiên cứu tập trung vào việc lựa chọn mô hình phù hợp tiếng Việt và thiết kế hệ thống triển khai thực tiễn (Azzuni & Saddik, 2025).

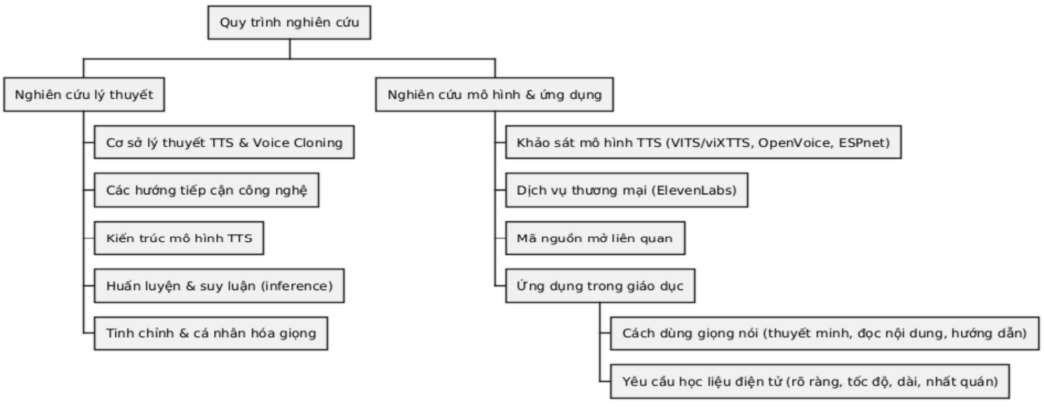
2.2. Ứng dụng AI Voice Cloning trong học liệu số

Trong học liệu số, giọng nói hỗ trợ tiếp cận nội dung, đặc biệt đối với người học từ xa hoặc người học cần hỗ trợ đọc hiểu; đồng thời có thể cải thiện tốc độ đọc, mức độ lưu giữ và tự hiệu quả khi học nếu được sử dụng như một công cụ hỗ trợ phù hợp (Raffoul & Jaber, 2023). Bên cạnh đó, đặc trưng diễn cảm trong giọng thuyết minh có liên hệ với mức độ tham gia và động cơ học tập trong bài giảng trực tuyến, dù tác động lên điểm kiểm tra có thể không rõ rệt (Marty-Dugas và cộng sự, 2024; Liew và cộng sự, 2023). Vì vậy, nhân bản giọng nói cho học liệu cần hướng tới giọng đọc rõ ràng, ổn định và phù hợp phong cách giảng dạy, thay vì chỉ tối ưu các chỉ số âm học.

III. Phương pháp nghiên cứu

3.1. Quy trình nghiên cứu

Nghiên cứu được triển khai theo ba bước: (i) khảo sát tài liệu và các mô hình TTS/voice cloning, tập trung vào khả năng hỗ trợ tiếng Việt và yêu cầu dữ liệu tham chiếu; (ii) thiết kế-hiện thực hệ thống sinh học liệu âm thanh theo kiến trúc client-server; và (iii) đánh giá thử nghiệm mức độ tương đồng và tính nhất quán của giọng mô phỏng để lựa chọn cấu hình sinh giọng nói phù hợp.



Hình 1. Quy trình nghiên cứu

3.2. Lựa chọn giải pháp

Trong các giải pháp khảo sát, viXTTS được chọn làm lõi TTS vì phù hợp tiếng Việt và hỗ trợ nhân bản giọng với mẫu tham chiếu ngắn. Mô hình được fine-tune từ XTTS-v2.0.3, mở rộng tokenizer tiếng Việt và tinh chỉnh trên bộ dữ liệu viVoice, nhờ đó cải thiện khả năng nắm bắt đặc trưng thanh điệu và cấu trúc âm tiết tiếng Việt; đồng thời có sẵn trọng số/mã nguồn mở giúp giảm chi phí triển khai và cho phép tập trung vào thiết kế hệ thống và đánh giá chất lượng (Le, 2024a; Le, 2024b).

Để phân tích tín hiệu và trích xuất đặc trưng phục vụ đánh giá, nhóm sử dụng thư viện Librosa cho các thao tác xử lý âm thanh cơ bản (McFee và cộng sự, 2015). Đặc trưng MFCC được sử dụng để so sánh tương đồng giữa giọng mẫu và giọng tổng hợp theo cách tiếp cận phổ biến trong phân tích tiếng nói (Davis & Mermelstein, 1980).

3.3. Thiết kế hệ thống

Dựa trên các phân tích về mô hình TTS, công cụ xử lý tín hiệu và đặc trưng âm thanh phù hợp cho bài toán tổng hợp giọng nói tiếng Việt, nhóm nghiên cứu tiến hành xây dựng kiến trúc hệ thống được đề xuất như sau:

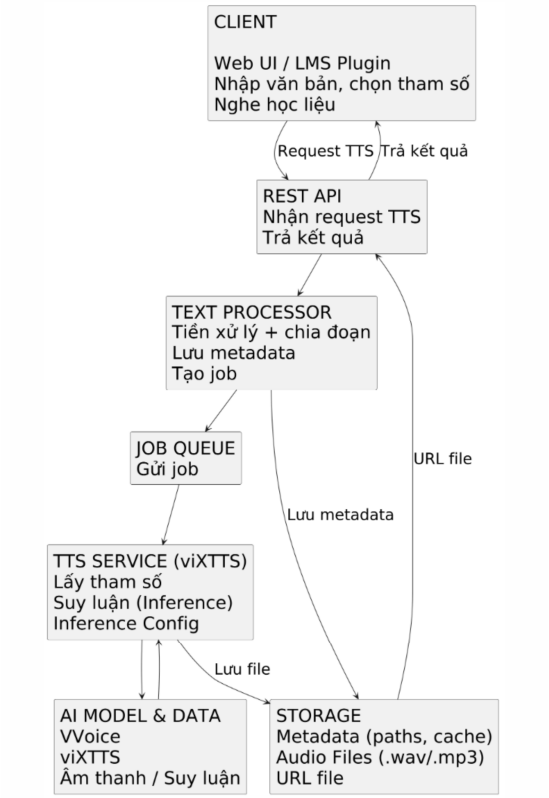
Hệ thống sử dụng kiến trúc khách - chủ (client - server), trong đó client là giao diện web để người dùng nhập văn bản và chọn tham số giọng đọc (tốc độ, cao độ, phong cách).

Python cung cấp REST API, nhận yêu cầu từ client, thực hiện tiền xử lý văn bản, chia đoạn và đưa vào hàng đợi xử lý.

Dịch vụ TTS dùng mô hình viXTTS (fine-tuned trên ViVoice) thực hiện suy luận, áp dụng cấu hình suy luận (Inference Config) để sinh sóng âm thanh tiếng Việt.

Khối hậu xử lý ghép các đoạn âm thanh, chuẩn hóa và lưu tệp .wav/.mp3 vào kho lưu trữ; metadata (đường dẫn, tham số, bộ nhớ đệm) được lưu riêng để dễ truy xuất và tái sử dụng.

API trả lại URL tệp âm thanh cho giao diện web, cho phép giảng viên/sinh viên nghe trực tiếp hoặc nhúng vào hệ thống LMS hiện có.

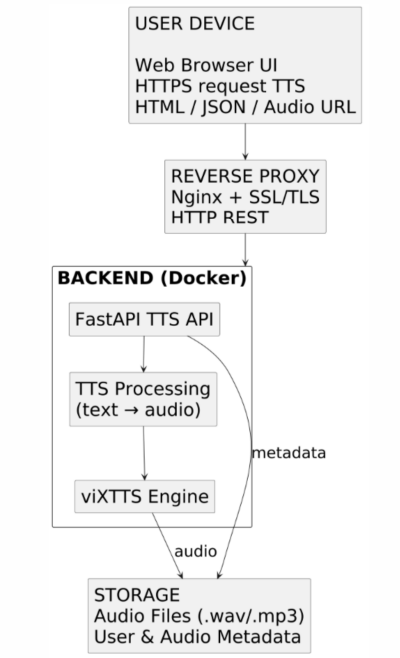


Hình 2. Kiến trúc hệ thống Voice Cloning tiếng Việt cho học liệu điện tử có âm thanh

IV. Kết quả và bàn luận

4.1. Kết quả phát triển hệ thống

Hệ thống được triển khai theo kiến trúc web client-server nhằm tách biệt giao diện và dịch vụ suy luận TTS, thuận lợi cho mở rộng và tích hợp. Người dùng nhập văn bản, tải giọng tham chiếu và chọn cấu hình; backend tiếp nhận, tiền xử lý/chia đoạn, suy luận bằng viXTTS và trả về tệp âm thanh cùng metadata để nghe lại hoặc nhúng vào LMS.

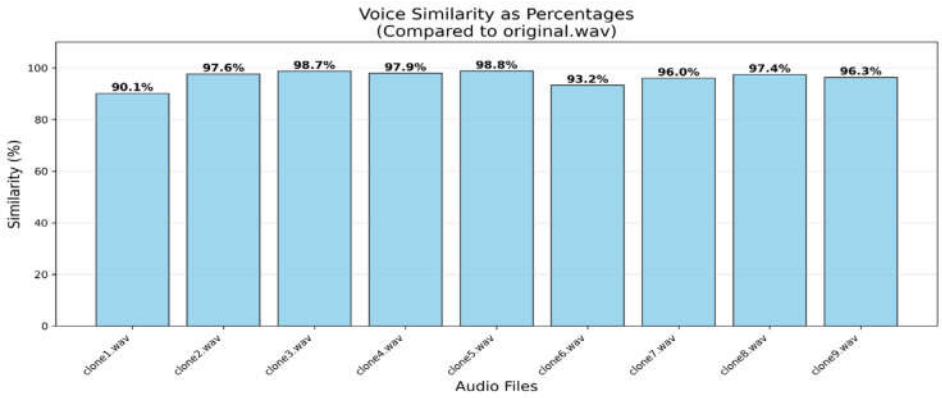


Hình 3. Triển khai AI Voice Cloning tiếng Việt cho học liệu điện tử có âm thanh

4.2. Đánh giá và thử nghiệm

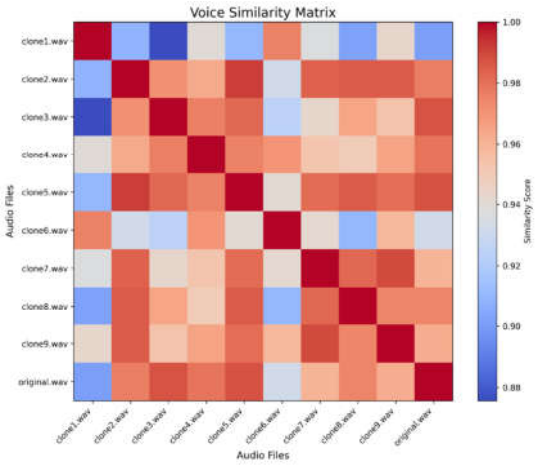
Đánh giá định lượng được thực hiện bằng cách trích xuất MFCC từ giọng gốc và giọng tổng hợp, sau đó tính độ tương đồng giữa các vector đặc trưng để phản ánh mức độ giống giọng và độ ổn định giữa các mẫu sinh (Davis & Mermelstein, 1980).

Thí nghiệm được thực hiện sinh 9 giọng và so sánh với giọng gốc. Kết quả cho thấy phần lớn mẫu đạt độ tương đồng trên 95% so với giọng gốc, đồng thời các mẫu sinh từ cùng một giọng gốc có mức tương đồng cao với nhau, phản ánh tính nhất quán của hệ thống. Các trường hợp thấp hơn chủ yếu liên quan đến cấu hình tham số suy luận và chất lượng âm thanh tham chiếu.



Hình 4. Độ tương tự giọng nói nhân tạo theo phần trăm

Hình 5 thể hiện ma trận độ tương đồng giọng nói thể hiện mức độ gần nhau giữa các mẫu âm thanh thông qua thang đo chuẩn hóa từ 0 đến 1, trong đó giá trị càng cao cho thấy mức tương đồng càng lớn. Đường chéo chính của ma trận luôn đạt giá trị gần 1, phản ánh việc mỗi mẫu có độ trùng khớp tối đa với chính nó. Các phần tử ngoài đường chéo biểu thị mức độ tương đồng giữa các bản ghi giọng nói được tạo ra (clone) và bản gốc. Sự phân bố màu sắc chủ đạo ở vùng đỏ và cam cho thấy các mẫu clone duy trì mức tương đồng cao với giọng gốc, chứng tỏ mô hình tổng hợp giọng nói tái tạo được đặc trưng âm học cốt lõi của người nói. Những biến thiên nhẹ giữa các cặp giọng nhân bản phản ánh sự khác biệt nhỏ do quá trình suy luận và điều kiện sinh âm thanh, nhưng nhìn chung vẫn nằm trong ngưỡng chấp nhận được đối với hệ thống TTS hướng tới bảo toàn đặc trưng giọng gốc.



Hình 5. Ma trận độ tương đồng giọng nói giữa các mẫu âm thanh

Phân tích cho thấy các tham số suy luận (temperature, top_k, top_p, repetition_penalty, length_penalty) ảnh hưởng trực tiếp đến độ ổn định và mức tự nhiên của giọng sinh; vì vậy nghiên cứu đề xuất một bộ cấu hình/preset theo trạng thái giọng nhằm cân bằng giữa tính rõ ràng và mức biến thiên ngữ điệu. Bảng 1 trình bày các preset đề xuất làm cấu hình khởi đầu cho ứng dụng học liệu điện tử.

Bảng 1. Tham số đề xuất cấu hình giọng đọc theo trạng thái cảm xúc

Cảm xúc	Độ ngẫu nhiên (temperature)	Hệ số lặp lại (Repetition Penalty)	Ngưỡng xác suất (Top_p)	Số lượng từ khóa chọn lọc (Top_k)	Đặc điểm giọng nói
Trung lập (neutral)	0.3	10	0.85	30	Trung tính
Bình tĩnh (calm)	0.15	15	0.6	15	Chậm, ổn định
Buồn (sad)	0.2	12	0.7	20	Trầm, cảm xúc nhẹ
Vui (happy)	0.4	8	0.9	40	Sáng, sinh động

4.3. So sánh với hệ thống TTS tiếng Việt mã nguồn mở khác

Để so sánh chất lượng của mô hình hệ thống đề xuất trong bối cảnh các giải pháp TTS tiếng Việt hiện có, nghiên cứu thực hiện so sánh trực tiếp với mô hình VieNeu-TTS - một hệ thống TTS tiếng Việt mã nguồn mở được phát triển dựa trên kiến trúc NeuTTS Air (phiên bản 0.5B tham số), huấn luyện trên bộ dữ liệu VieNeu-TTS 1000h gồm 443641 mẫu giọng đọc tiếng Việt. VieNeu-TTS hỗ trợ nhân bản giọng nói tức thì với chỉ 3-5 giây audio tham chiếu, sinh âm thanh 24 kHz và cho phép suy luận thời gian thực trên CPU.

Trong thí nghiệm, tổng cộng 11 mẫu giọng được tạo ra từ cùng một đoạn văn bản gốc và cùng tệp âm thanh tham chiếu của người nói (origin), kèm thêm một mẫu giọng nam khác đóng vai trò đối chứng. Hệ thống viXTTS được chạy 9 lần độc lập để sinh ra các mẫu clone1-clone9, nhằm đánh giá mức độ ổn định của mô hình qua nhiều lần suy luận. Ngược lại, VieNeu-TTS chỉ tạo 2 mẫu đại diện, do mô hình này có cấu hình cố định và không yêu cầu lặp lại nhiều lần để kiểm tra độ biến thiên. Tất cả 11 mẫu được đưa vào cùng một quy trình đánh giá, cho phép so sánh trực tiếp mức độ tương đồng giọng nói giữa các mẫu sinh và giọng gốc, cũng như giữa hai hệ thống TTS.

Hai hệ thống được kiểm tra dưới cùng điều kiện thử nghiệm: sử dụng chung bộ câu văn bản trích từ nội dung bài giảng và cùng giọng giảng viên gốc làm tham chiếu. Đầu ra của mỗi hệ thống được đánh giá bằng bộ chỉ số đa chiều, bao gồm: MOS dự đoán (SQUIM, thang 1-5), PESQ đánh giá chất lượng cảm nhận, STOI đo độ rõ ràng khách quan (thang 0-1), và độ

tương đồng giọng nói dựa trên embedding Resemblyzer/GE2E. Kết quả tổng hợp được trình bày trong Bảng 2, cho phép đối chiếu toàn diện giữa hai hệ thống về chất lượng âm thanh, độ rõ ràng và mức độ bảo toàn đặc trưng người nói.

Bảng 2. So sánh tổng hợp chất lượng giọng nói giữa viXTTS và VieNeu-TTS

Hệ thống	MOS (SQUIM)	PESQ	STOI	Speaker Sim
viXTTS	3,982	2,934	0,989	86,88%
VieNeu-TTS	3,985	3,204	0,988	79,30%

Từ kết quả Bảng 2, có thể rút ra một số nhận xét. Thứ nhất, về chất lượng âm thanh tổng thể (MOS), hai hệ thống đạt mức gần tương đương: viXTTS đạt 3,982 và VieNeu-TTS đạt 3,985 trên thang 5 điểm. Cả hai đều nằm trong vùng "tốt" và vượt ngưỡng chấp nhận cho ứng dụng TTS (≥ 3,5), cho thấy cả hai mô hình đều phù hợp để tạo học liệu âm thanh. Thứ hai, về chất lượng cảm nhận giọng nói (PESQ), VieNeu-TTS đạt điểm cao hơn đáng kể (3,204 so với 2,934), cho thấy giọng nói sinh ra từ VieNeu-TTS có chất lượng cảm nhận tốt hơn ở mức tín hiệu. Điều này có thể liên quan đến codec âm thanh NeuCodec và tần số lấy mẫu đầu ra 24 kHz của VieNeu-TTS so với cơ chế sinh sóng của viXTTS.

Thứ ba, về độ tương đồng giọng nói (Speaker Similarity), viXTTS vượt trội rõ rệt so với VieNeu-TTS (86,88% so với 79,30%). Điều này cho thấy viXTTS bảo toàn đặc trưng giọng giảng viên gốc tốt hơn, phù hợp với cơ chế few-shot voice cloning sử dụng mẫu tham chiếu dài hơn so với zero-shot 3 đến 5 giây của VieNeu-TTS. Trong bối cảnh học liệu điện tử, nơi tính cá nhân hóa giọng giảng viên là yêu cầu quan trọng, đây là ưu thế đáng kể của viXTTS. Thứ tư, chỉ số STOI (độ rõ ràng) của cả hai hệ thống đều rất cao (> 0,98), cho thấy giọng nói sinh ra đều đảm bảo độ rõ ràng cần thiết cho việc truyền tải nội dung học thuật.

Tổng hợp lại, viXTTS phù hợp hơn cho bài toán học liệu điện tử nhờ khả năng bảo toàn giọng giảng viên cao hơn, trong khi VieNeu-TTS có ưu thế về chất lượng cảm nhận tín hiệu và khả năng triển khai nhẹ trên thiết bị đầu cuối (hỗ trợ GGUF lượng tử hóa, suy luận trên CPU). Sự bổ trợ này gợi ý rằng trong tương lai, việc kết hợp ưu điểm của cả hai hướng tiếp cận có thể tạo ra hệ thống tối ưu hơn cho sản xuất học liệu số.

V. Kết luận

5.1. Kết luận chung

Nghiên cứu đã phát triển hệ thống mô phỏng giọng nói phục vụ xây dựng học liệu điện tử có âm thanh trên nền mô hình viXTTS, triển khai theo kiến trúc client-server và cung cấp luồng xử lý từ văn bản đến tệp âm thanh đầu ra. Thử nghiệm cho thấy hệ thống đạt độ tương đồng giọng cao (>95%) và duy trì tính nhất quán giữa các mẫu, qua đó khẳng định tính khả thi của hướng tiếp cận mô phỏng giọng nói trong sản xuất học liệu tiếng Việt.

Nghiên cứu đóng góp vào việc lấp đầy khoảng trống về ứng dụng nhân bản giọng nói trong xây dựng học liệu điện tử có âm thanh, đồng thời cung cấp nền tảng kỹ thuật và phương pháp đánh giá cho việc triển khai thực tế tại các cơ sở đào tạo. Bên cạnh đó, nghiên cứu cũng chỉ ra những thách thức còn tồn tại như cần tinh chỉnh tham số suy luận để tối ưu giọng đọc cho từng loại nội dung học liệu, cần mở rộng bộ dữ liệu giọng giảng viên để nâng cao tính cá nhân hóa.

5.2. Hạn chế và hướng phát triển

Hạn chế chính gồm quy mô thử nghiệm còn nhỏ và chưa đo lường trực tiếp tác động của giọng mô phỏng lên kết quả học tập. Trong tương lai, cần mở rộng dữ liệu giọng giảng viên, bổ sung đánh giá người dùng (người học/giảng viên) theo kịch bản học liệu thực, và nghiên cứu cơ chế tự động lựa chọn cấu hình giọng theo loại nội dung để tối ưu trải nghiệm học tập.

Ngoài các hạn chế kỹ thuật, việc triển khai nhân bản giọng nói trong thực tế đặt ra những vấn đề nghiêm túc về đạo đức và bảo mật thông tin. Công nghệ nhân bản giọng nói có thể bị lạm dụng để giả mạo danh tính, tạo nội dung sai lệch hoặc xâm phạm quyền riêng tư (Azzuni & Saddik, 2025). Trong bối cảnh học liệu điện tử, rủi ro này tuy thấp hơn so với các ứng dụng thương mại hay truyền thông, nhưng vẫn cần được quản lý chặt chẽ.

Cụ thể, hệ thống cần đảm bảo: (i) giọng giảng viên chỉ được sao chép và sử dụng với sự đồng ý rõ ràng bằng văn bản; (ii) tích hợp cơ chế đánh dấu số (audio watermarking) vào tệp âm thanh đầu ra, cho phép xác minh nguồn gốc và phát hiện giả mạo; (iii) thiết lập chính sách quản trị quyền truy cập, giới hạn đối tượng được phép tạo và phân phối giọng mô phỏng trong hệ thống LMS. Nghiên cứu trong tương lai cần kết hợp các phương pháp phát hiện giọng giả (deepfake voice detection) để tạo lớp bảo vệ bổ sung khi triển khai ở quy mô lớn.

5.3. Đóng góp của nghiên cứu

Đóng góp của nghiên cứu gồm: (1) đề xuất và hiện thực hóa một hệ thống tạo học liệu âm thanh bằng TTS; (2) trình bày quy trình lựa chọn mô hình tiếng Việt và cấu hình phù hợp với sản xuất học liệu dạng âm thanh; (3) cung cấp cách tiếp cận đánh giá định lượng về độ tương đồng và tính nhất quán của giọng nói để đưa vào triển khai thực tiễn.

Tài liệu tham khảo

Azzuni, H., & El Saddik, A. E. (2025). *Voice cloning: Comprehensive survey*. arXiv. https://arxiv.org/abs/2505.00579

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357-366. https://doi.org/10.1109/TASSP.1980.1163420

Kumari, M., Goyal, N. K., Dandotiya, M. K., & Kushwaha, V. (2024). An effective detection of voice cloning using deep learning. In *Proceedings of the 6th International Conference on Information Management & Machine Intelligence (ICIMMI 2024)*. Association for Computing Machinery. https://doi.org/10.1145/3745812.3745895

Le, T. (Thinh Le). (2024a). *viVoice: Enabling Vietnamese multi-speaker speech synthesis* [Dataset]. Hugging Face. https://huggingface.co/datasets/capleaf/viVoice

Le, T. (Thinh Le). (2024b). *viXTTS* [Model]. Hugging Face. https://huggingface.co/capleaf/viXTTS

Liew, T. W., Tan, S. M., Pang, W. M., Khan, M. T. I., & Kew, S. N. (2023). I am Alexa, your virtual tutor!: The effects of Amazon Alexa's text-to-speech voice enthusiasm in a multimedia learning environment. *Education and Information Technologies, 28*(2), 1455-1489. https://doi.org/10.1007/s10639-022-11255-6

Marty-Dugas, J., Rajasingham, M., McHardy, R. J., Kim, J., & Smilek, D. (2024). Instructor enthusiasm in online lectures: How vocal enthusiasm impacts student engagement, learning, and memory. *Frontiers in Education, 9*, 1339815. https://doi.org/10.3389/feduc.2024.1339815

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18-25). https://doi.org/10.25080/Majora-7b98e3ed-003

Nguyễn Thành Việt, & Trần Duy Linh. (2026). 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. *Tạp chí Khoa học & Công nghệ Việt Nam, 68*(1). https://doi.org/10.31276/VJST.2024.2810

Phạm, N. P. (2023). *Nghiên cứu phát triển hệ thống thích nghi giọng nói trong tổng hợp tiếng Việt và ứng dụng* (Luận án Tiến sĩ, Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam).

Qiao, Z., Yang, J., & Wang, Z. (2023). Multi-feature cross-lingual transfer learning approach for low-resource Vietnamese speech synthesis. In *Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms (AI2A '23)*. Association for Computing Machinery. https://doi.org/10.1145/3611450.3611476

Raffoul, S., & Jaber, L. (2023). Text-to-speech software and reading comprehension: The impact for students with learning disabilities. *Canadian Journal of Learning and Technology, 49*(2), 1-18. https://doi.org/10.21432/cjlt28296

Roosadi, H. R. P., Prakosa, S. W., & Lhaksmana, K. M. (2024). Indonesian voice cloning text-to-speech system with Vall-E-based model and speech enhancement. *IEEE Access, 12*, 193131-193140.

Sadik, M., Vijaya, P., Revathi, Y., Siva Naga Tanuja, V., Soudhamini, B., & Vaishnavi, R. (2025). AI-based voice cloning system: From text to speech. *International Journal of Innovative Science and Research Technology, 10*(4), 1453-1461.

Zheng, Z., Peng, P., Diwan, A., Huynh, C. P., Sun, X., Liu, Z., Bhat, V., & Harwath, D. (2025). *VoiceCraft-X: Unifying multilingual, voice-cloning speech synthesis and speech editing*. arXiv. https://arxiv.org/abs/2511.12347

AI-BASED VOICE CLONING SYSTEM FOR DIGITAL LEARNING MATERIAL DESIGN

Dang Hai Dang[1], Quach Thi Hanh[1], Nguyen Van Hoang[2], Nguyen Duc Tin[2]

Abstract: *In the context of the rapid expansion of online education, digital learning materials need to be developed using new approaches that enhance personalization, accessibility, and communicative effectiveness. Audio-based resources, particularly narrated speech, play a crucial role in modern E-learning systems. However, traditional recording workflows remain limited because they are time-consuming, heavily dependent on human resources, and lack the flexibility for rapid updates. This study focuses on developing an artificial intelligence (AI)-based text-to-speech (TTS) system capable of simulating and reproducing natural-sounding voices, aiming to shorten, or even eliminate, manual recording in the creation of learning materials. The system is implemented in a client-server architecture, leveraging the viXTTS model, fine-tuned on the viVoice dataset, to generate audio learning materials from the instructor's original voice. Experimental results show that the system achieves a high similarity to the original voice (above 95%) and maintains consistency across cloned speech samples. The study contributes to addressing the gap in the application of voice cloning and voice reproduction technologies in the Vietnamese online education context, while providing a practical testbed for deployment at the institutional level.*

Keywords: *voice cloning, text-to-speech, artificial intelligence, digital learning materials, digital education*

[1] Hanoi Open University, Hanoi, Vietnam

[2] Information Technology Student, Institute for Training and Lifelong Learning Development, Hanoi Open University, Hanoi, Vietnam