

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT HÀ NỘI



BÁO CÁO HỌC THUẬT

**SO SÁNH CÁC CÔNG NGHỆ TTS (TEXT-TO-SPEECH)
VÀ ỨNG DỤNG TRONG SOẠN VIDEO BÀI GIẢNG.**

Gv. Ths. NGÔ HÙNG LONG

Hà Nội, tháng 09/2025

MỤC LỤC

MỤC LỤC.....	2
MỞ ĐẦU.....	3
1.1 Giới thiệu về bối cảnh giáo dục số và nhu cầu tự động hóa trong việc tạo nội dung bài giảng.....	3
1.2 Lý do chọn đề tài.....	3
CHƯƠNG 2 : KHÁI NIỆM TTS (TEXT-TO-SPEECH).....	4
2.1 Khái niệm TTS (Text-to-Speech).....	4
2.2 Các thành phần cơ bản của hệ thống TTS:.....	4
2.3 Các loại TTS phổ biến:.....	5
CHƯƠNG 3 . VAI TRÒ CỦA TTS TRONG GIÁO DỤC SỐ	6
3.1 Tạo giọng đọc tự động cho nội dung bài giảng.....	6
3.2 Tiết kiệm thời gian và chi phí so với thu âm truyền thống.....	6
3.3 Hỗ trợ cá nhân hóa bài giảng.....	6
3.4 Tăng khả năng tiếp cận cho học sinh có nhu cầu đặc biệt.....	6
CHƯƠNG 4 . SO SÁNH CÁC GIẢI PHÁP TTS HIỆN NAY	7
4.1 So sánh các giải pháp TTS hiện nay.....	7
4.2 Tiêu chí so sánh các giải pháp TTS:.....	7
4.3 Ưu – nhược điểm từng giải pháp:.....	8
CHƯƠNG 5 . ỨNG DỤNG TTS TRONG SOẠN VIDEO BÀI GIẢNG	9
5.1 Chuẩn bị nội dung bài giảng (text).....	9
5.2 Chọn giải pháp TTS và tạo file âm thanh.....	10
5.3 Tạo file âm thanh từ văn bản.....	12
5.4 Tạo nhiều phiên bản giọng đọc.....	13
5.5 Tích hợp âm thanh vào video, đồng bộ với hình ảnh/slides.....	14
5.6 Chỉnh sửa, xuất bản và phân phối.....	15
5.7 Lợi ích và bất cập:.....	15
KẾT LUẬN VÀ ĐỀ XUẤT.....	17
TÀI LIỆU THAM KHẢO.....	19

MỞ ĐẦU

1.1 Giới thiệu về bối cảnh giáo dục số và nhu cầu tự động hóa trong việc tạo nội dung bài giảng

Trong bối cảnh cách mạng công nghiệp 4.0, giáo dục số đang trở thành xu hướng chủ đạo trên toàn cầu. Các trường học, đại học và trung tâm đào tạo ngày càng chú trọng áp dụng công nghệ để nâng cao hiệu quả giảng dạy và học tập. Việc số hóa tài liệu, triển khai bài giảng trực tuyến, học tập kết hợp (blended learning) và các nền tảng học tập trực tuyến (LMS) đã làm thay đổi cách thức truyền đạt kiến thức, từ truyền thống sang tương tác đa phương tiện, linh hoạt và cá nhân hóa.

Tuy nhiên, việc tạo ra nội dung bài giảng số, đặc biệt là các video bài giảng chất lượng cao, vẫn gặp nhiều thách thức. Quá trình này thường yêu cầu giáo viên phải ghi âm giọng đọc, biên tập video và đồng bộ âm thanh với hình ảnh, slides hoặc các hiệu ứng minh họa. Đây là một công việc tốn thời gian, công sức và đôi khi còn hạn chế khả năng cập nhật nhanh chóng nội dung bài giảng khi cần chỉnh sửa. Do đó, nhu cầu tự động hóa quá trình tạo nội dung trở nên cấp thiết, giúp giáo viên tiết kiệm thời gian, tối ưu nguồn lực và nâng cao chất lượng trải nghiệm học tập cho học sinh, sinh viên.

1.2 Lý do chọn đề tài

Trong số các công nghệ hỗ trợ tự động hóa bài giảng, TTS (Text-to-Speech – chuyển văn bản thành giọng nói) ngày càng trở nên phổ biến và được ứng dụng rộng rãi. TTS cho phép chuyển đổi nội dung văn bản thành giọng đọc tự nhiên, nhấn nhá đúng ngữ điệu, giúp giáo viên tạo ra âm thanh bài giảng mà không cần tự thu âm. Công nghệ này đặc biệt hữu ích trong việc tạo video bài giảng, nơi giọng đọc đóng vai trò quan trọng trong việc truyền tải thông tin, duy trì sự tập trung và hỗ trợ học sinh tiếp nhận kiến thức hiệu quả.

Việc nghiên cứu, so sánh các giải pháp TTS hiện nay không chỉ giúp đánh giá ưu – nhược điểm của từng công nghệ mà còn đề xuất hướng ứng dụng tối ưu trong soạn video bài giảng, góp phần nâng cao chất lượng giáo dục số, hỗ trợ cá nhân hóa trải nghiệm học tập và giảm thiểu thời gian, chi phí sản xuất nội dung giảng dạy.

CHƯƠNG 2: KHÁI NIỆM TTS (TEXT-TO-SPEECH)

2.1 Khái niệm TTS (Text-to-Speech)

Định nghĩa TTS:

TTS (Text-to-Speech) là công nghệ chuyển đổi văn bản viết thành giọng nói tự nhiên, giúp máy tính hoặc thiết bị điện tử phát âm nội dung văn bản mà không cần sự tham gia trực tiếp của con người. Công nghệ này cho phép tạo ra giọng đọc nhân tạo, có ngữ điệu, nhấn nhá phù hợp, hỗ trợ nhiều ngôn ngữ và giọng đọc khác nhau. TTS đang được ứng dụng rộng rãi trong các lĩnh vực như trợ lý ảo, ứng dụng đọc sách điện tử, hệ thống thông báo, và đặc biệt là trong giáo dục số để tạo ra các video bài giảng hoặc nội dung học tập tương tác.

2.2 Các thành phần cơ bản của hệ thống TTS:

Một hệ thống TTS điển hình bao gồm ba bước chính:

- Phân tích văn bản (Text analysis):

Xử lý văn bản đầu vào để chuẩn hóa cú pháp, tách câu, từ, dấu câu.

Nhận diện các ký hiệu đặc biệt, chữ số, ngày tháng, từ viết tắt, ký hiệu toán học... và chuẩn hóa thành dạng văn bản phù hợp để chuyển đổi thành âm thanh.

Mục tiêu là giúp hệ thống hiểu được ngữ cảnh và cấu trúc câu, từ đó tạo ra phát âm chính xác và tự nhiên.

- Chuyển đổi thành ngữ âm (Phonetic conversion):

Biến văn bản đã chuẩn hóa thành các ký hiệu ngữ âm (phoneme) hoặc chuỗi âm vị.

Xác định trọng âm, ngữ điệu, nhịp điệu của câu để giọng đọc nhân tạo giống giọng người hơn.

Đây là bước quan trọng để đảm bảo tính tự nhiên và dễ nghe của giọng đọc.

- Tổng hợp giọng nói (Speech synthesis):

Chuyển chuỗi âm vị thành tín hiệu âm thanh thực tế.

Có thể sử dụng các kỹ thuật khác nhau, từ ghép nối mẫu âm thanh (concatenative) đến mô hình học sâu (neural TTS) để tạo giọng đọc mượt mà, giàu biểu cảm.

Kết quả cuối cùng là file âm thanh hoặc phát trực tiếp giọng nói từ văn bản.

2.3 Các loại TTS phổ biến:

- Dựa trên quy tắc (Rule-based TTS):

Sử dụng các quy tắc ngữ âm và ngữ điệu được lập trình sẵn để phát âm từ văn bản.

Ưu điểm: kiểm soát tốt ngữ điệu, phát âm chính xác với các từ đơn giản.

Nhược điểm: giọng đọc thường nghe máy móc, thiếu tự nhiên, khó mở rộng cho các ngôn ngữ phức tạp.

- Dựa trên mẫu (Concatenative TTS):

Ghép nối các đoạn âm thanh đã ghi sẵn (mẫu) để tạo thành câu hoàn chỉnh.

Ưu điểm: giọng đọc tự nhiên hơn so với TTS dựa trên quy tắc, thích hợp cho các ứng dụng thương mại.

Nhược điểm: cần bộ dữ liệu mẫu lớn, khó thay đổi ngữ điệu, trọng âm và phát âm các từ mới chưa có trong dữ liệu.

- Dựa trên mô hình học sâu (Deep learning / Neural TTS):

Sử dụng các mạng nơ-ron học sâu để dự đoán tín hiệu âm thanh từ văn bản.

Ưu điểm: giọng đọc tự nhiên, giàu biểu cảm, dễ tùy chỉnh tốc độ, ngữ điệu và giọng đọc.

Nhược điểm: yêu cầu tài nguyên tính toán cao, bộ dữ liệu huấn luyện lớn.

Ví dụ: Google WaveNet, Tacotron 2, YourTTS, Microsoft Neural TTS.

CHƯƠNG 3. VAI TRÒ CỦA TTS TRONG GIÁO DỤC SỐ

Vai trò của TTS trong giáo dục số: TTS (Text-to-Speech) đóng vai trò quan trọng trong việc số hóa và nâng cao hiệu quả giảng dạy, đặc biệt trong bối cảnh giáo dục trực tuyến và blended learning. Một số vai trò nổi bật bao gồm:

3.1 Tạo giọng đọc tự động cho nội dung bài giảng

TTS cho phép chuyển đổi toàn bộ văn bản bài giảng thành giọng nói một cách tự động, giúp học sinh nghe bài giảng thay vì chỉ đọc văn bản. Điều này không chỉ làm tăng tính sinh động, hấp dẫn của bài học mà còn giúp học sinh tiếp thu kiến thức dễ dàng hơn, đặc biệt với các nội dung dài hoặc phức tạp.

3.2 Tiết kiệm thời gian và chi phí so với thu âm truyền thống

Thay vì phải thu âm từng bài giảng bằng giọng người, chỉnh sửa và đồng bộ âm thanh với video, TTS cho phép tạo ra giọng đọc ngay lập tức từ văn bản. Điều này giúp giáo viên tiết kiệm thời gian, giảm chi phí thuê nhân sự thu âm và thiết bị, đồng thời dễ dàng tạo ra nhiều phiên bản giọng đọc khác nhau cho cùng một nội dung bài giảng.

3.3 Hỗ trợ cá nhân hóa bài giảng

Chọn giọng đọc, tốc độ, ngữ điệu phù hợp với học sinh: TTS hiện đại cho phép lựa chọn nhiều giọng đọc (nam, nữ, trẻ em), điều chỉnh tốc độ đọc, nhấn nhá và ngữ điệu, từ đó tạo ra trải nghiệm học tập phù hợp với từng nhóm học sinh.

Dễ dàng cập nhật, chỉnh sửa nội dung bài giảng: Khi nội dung thay đổi hoặc cần bổ sung, giáo viên chỉ cần chỉnh sửa văn bản, TTS sẽ tự động tạo giọng đọc mới mà không cần thu âm lại từ đầu, đảm bảo bài giảng luôn cập nhật và nhất quán.

3.4 Tăng khả năng tiếp cận cho học sinh có nhu cầu đặc biệt

TTS giúp học sinh gặp khó khăn trong việc đọc hoặc thị giác kém tiếp cận nội dung bài học dễ dàng hơn. Việc nghe bài giảng qua giọng đọc tự động giúp các học sinh này không bị bỏ lỡ kiến thức, đồng thời hỗ trợ các hình thức học tập đa giác quan, nâng cao khả năng học tập và trải nghiệm giáo dục toàn diện.

CHƯƠNG 4. SO SÁNH CÁC GIẢI PHÁP TTS HIỆN NAY

4.1 So sánh các giải pháp TTS hiện nay

Hiện nay, trên thị trường có nhiều công nghệ TTS phổ biến, được phát triển bởi các hãng lớn hoặc dự án mã nguồn mở. Một số giải pháp tiêu biểu bao gồm:

- **Google TTS (Google Cloud Text-to-Speech):** cung cấp giọng đọc tự nhiên, đa dạng ngôn ngữ, tích hợp dễ dàng qua API.
- **Amazon Polly:** nổi bật với khả năng tạo giọng nói tự nhiên, hỗ trợ SSML (Speech Synthesis Markup Language) để điều chỉnh nhấn nhá, ngữ điệu.
- **Microsoft Azure TTS:** cung cấp nhiều giọng đọc chuẩn, hỗ trợ neural TTS, tùy chỉnh tốc độ, nhấn nhá, dễ tích hợp vào ứng dụng.
- **YourTTS:** dựa trên mô hình học sâu, hỗ trợ tạo giọng đọc cá nhân hóa, có thể học theo giọng mẫu.
- **OpenTTS:** dự án mã nguồn mở, dễ triển khai trên nhiều nền tảng, cho phép tùy chỉnh cao nhưng giọng đọc có thể chưa tự nhiên bằng các giải pháp thương mại.

4.2 Tiêu chí so sánh các giải pháp TTS:

1. Chất lượng giọng đọc:

Đánh giá mức độ tự nhiên, nhấn nhá, ngữ điệu, biểu cảm.

Neural TTS thường cho giọng đọc mượt mà và giàu cảm xúc hơn so với TTS dựa trên mẫu hoặc quy tắc.

2. Ngôn ngữ và giọng đọc hỗ trợ:

Số lượng ngôn ngữ và giọng đọc (nam, nữ, trẻ em).

Khả năng hỗ trợ các từ địa phương, tên riêng, thuật ngữ chuyên ngành.

3. Khả năng tùy chỉnh:

Điều chỉnh tốc độ đọc, ngữ điệu, trọng âm.

Hỗ trợ SSML hoặc các tùy chọn nâng cao để tăng tính tự nhiên.

4. Tính năng bổ sung:

Chuyên đổi đa nền tảng (Windows, Linux, Web, Mobile).

Hỗ trợ tích hợp API, có thể kết hợp với các công cụ tạo video hoặc ứng dụng học tập.

5. Chi phí, điều kiện sử dụng, yêu cầu phần cứng:

Một số dịch vụ thương mại tính phí theo lượng ký tự hoặc thời lượng âm thanh.

OpenTTS miễn phí nhưng yêu cầu máy chủ đủ mạnh để chạy mô hình neural.

4.3 Ưu – nhược điểm từng giải pháp:

Giải pháp	Ưu điểm	Nhược điểm
Google TTS	Giọng đọc tự nhiên, nhiều ngôn ngữ, tích hợp API dễ dàng	Chi phí theo ký tự, hạn chế tùy chỉnh sâu
Amazon Polly	Hỗ trợ SSML, đa giọng, nhấn nhá tự nhiên	Cần đăng ký AWS, chi phí phát sinh theo thời gian sử dụng
Microsoft Azure	Neural TTS chất lượng cao, nhiều tùy chỉnh, dễ tích hợp	Miễn phí, yêu cầu kết nối Internet
YourTTS	Giọng đọc cá nhân hóa, học theo mẫu giọng	Cần dữ liệu huấn luyện, yêu cầu GPU mạnh để chạy hiệu quả
OpenTTS	Miễn phí, mã nguồn mở, tùy chỉnh cao	Giọng đọc chưa tự nhiên bằng các giải pháp thương mại, cần máy mạnh

Bảng 1: So sánh ưu nhược điểm các giải pháp TTS

CHƯƠNG 5. ỨNG DỤNG TTS TRONG SOẠN VIDEO BÀI GIẢNG

Ứng dụng TTS trong soạn video bài giảng. Công nghệ TTS đã mở ra nhiều cơ hội trong việc tạo video bài giảng số, giúp giáo viên tiết kiệm thời gian và nâng cao chất lượng giảng dạy. Quy trình tạo video bài giảng với TTS thường gồm các bước sau:

5.1 Chuẩn bị nội dung bài giảng (text)

Bước chuẩn bị nội dung bài giảng là bước quan trọng đầu tiên trong quy trình tạo video bài giảng sử dụng TTS. Giáo viên cần soạn thảo nội dung bài giảng dưới dạng văn bản, bao gồm:

Lý thuyết: trình bày các khái niệm, định nghĩa, nguyên lý cơ bản.

Hướng dẫn thực hành: các bước thực hiện, ví dụ minh họa, bài tập ứng dụng.

Câu hỏi kiểm tra: bài tập trắc nghiệm, câu hỏi ôn tập để học sinh tự đánh giá.

Nội dung văn bản phải được trình bày rõ ràng, mạch lạc, dễ đọc, vì TTS sẽ chuyển toàn bộ văn bản này thành giọng nói. Văn bản lộn xộn, thiếu dấu câu hoặc sai chính tả có thể dẫn đến giọng đọc không tự nhiên hoặc nhấn nhá sai ngữ điệu.

Chuyển đổi văn bản từ nhiều định dạng khác nhau:

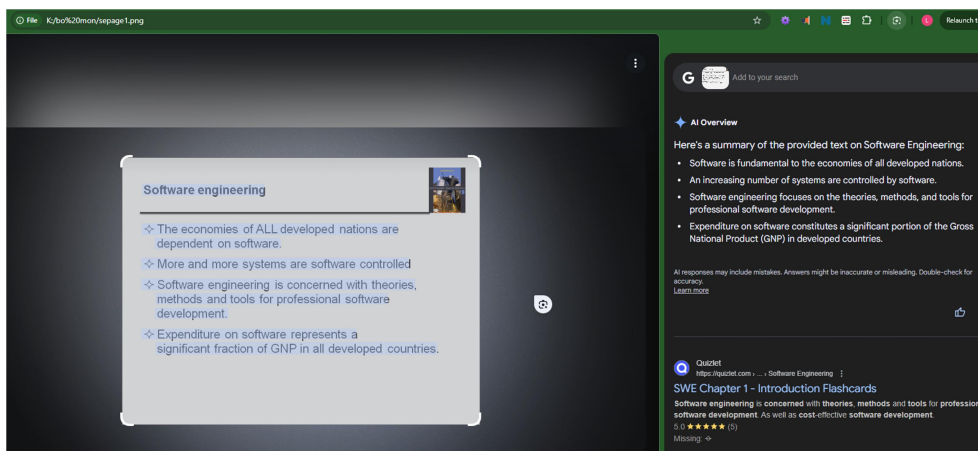
Trong thực tế, giáo viên có thể nhận hoặc soạn nội dung bài giảng ở nhiều định dạng khác nhau như DOCX, PDF, hình ảnh, HTML. Một số công cụ chuyển đổi hữu ích:

DOCX to Text: Sử dụng Python python-docx hoặc LibreOffice để xuất văn bản thuần từ file Word.

PDF to Text: Dùng PyPDF2, pdftminer.six, hoặc công cụ trực tuyến như SmallPDF để trích xuất nội dung.

Image to Text (OCR): Sử dụng Tesseract OCR, Google Vision API hoặc các công cụ trực tuyến để nhận dạng văn bản từ ảnh, slide chụp.

Hiện tại để nhận diện văn bản Tiếng Việt tốt nhất chỉ có Google Vision API, hoặc “ABBY FineReader”. Có thể tạo ứng dụng tự động nhận diện nhiều trang bằng Google Vision API. Sau khi nhận diện văn bản dùng GPT để chuẩn hóa hoặc dịch ngôn ngữ, việc dùng giọng đọc trực tiếp của GPT cũng là giải pháp tốt.



Hình 1: Dùng Google Vision API để nhận diện văn bản

HTML to Text: Dùng BeautifulSoup (Python) hoặc các thư viện tương tự để lọc thẻ HTML, lấy văn bản thuần.

Chuẩn hóa và kiểm tra văn bản trước khi đưa vào TTS:

Loại bỏ các ký tự đặc biệt, gạch đầu dòng, chú thích không cần thiết.

Kiểm tra chính tả và ngữ pháp.

Có thể sử dụng GPT hoặc các mô hình ngôn ngữ AI để:

Chuẩn hóa văn bản, đảm bảo câu mạch lạc, dễ đọc.

Tách câu, đánh dấu ngắt câu hợp lý để giọng đọc TTS tự nhiên hơn.

Chỉnh sửa lỗi chính tả, ngắt đoạn, hoặc viết lại các câu dài thành các câu ngắn, dễ nghe.

Hiện thư viện chuẩn hóa Tiếng Việt mã mở được dùng nhiều là underthesea tại <https://github.com/undertheseanlp/underthesea>. Nhưng theo kinh nghiệm thực tế thì GPT làm rất tốt.

Bằng cách kết hợp các công cụ chuyển đổi và chuẩn hóa văn bản, giáo viên sẽ có một file văn bản sạch, chuẩn hóa, sẵn sàng cho bước tạo giọng đọc TTS, giúp video bài giảng đạt chất lượng cao và giọng đọc tự nhiên..

5.2 Chọn giải pháp TTS và tạo file âm thanh

Sau khi chuẩn bị nội dung văn bản sạch và chuẩn hóa, bước tiếp theo là lựa chọn công nghệ TTS phù hợp và tạo file âm thanh cho video bài giảng.

- **Lựa chọn công nghệ TTS phù hợp**

Hiện nay, có nhiều giải pháp TTS phổ biến, mỗi giải pháp có ưu – nhược điểm riêng:

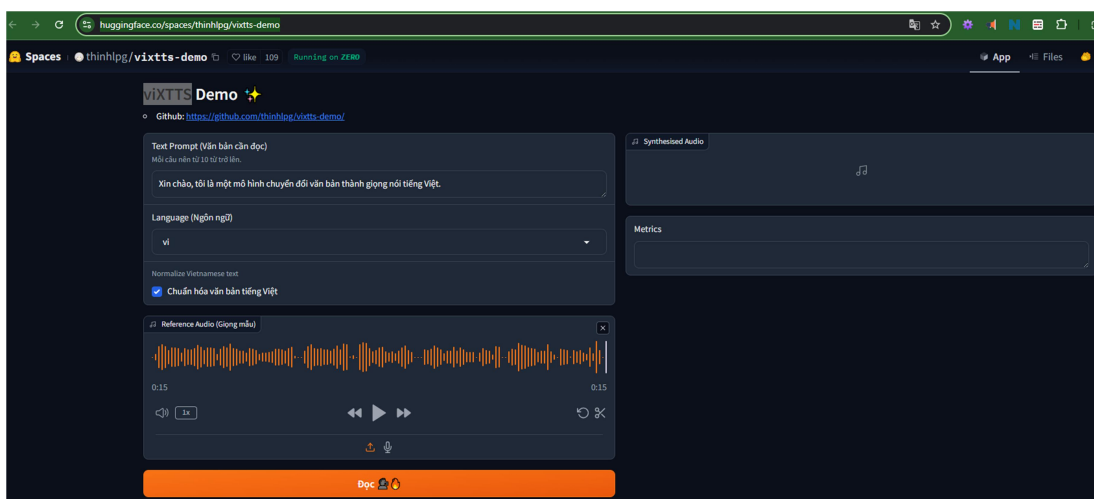
Google TTS: Giọng đọc tự nhiên, nhiều ngôn ngữ, dễ tích hợp API, hỗ trợ SSML để điều chỉnh nhấn nhá, tốc độ và ngữ điệu, **có mất phí**.

Amazon Polly: Giọng đọc mượt, hỗ trợ SSML và nhiều giọng đọc (nam/nữ, trẻ em), tích hợp dễ dàng với ứng dụng web và mobile, **có mất phí**.

Microsoft Azure TTS (MS TTS): Cung cấp giọng đọc neural chất lượng cao, khả năng điều chỉnh nhấn nhá, tốc độ, trọng âm; dễ tích hợp vào ứng dụng giáo dục hoặc tạo file âm thanh dạng WAV/MP3, **hiện có free**.

YourTTS: Hỗ trợ tạo giọng đọc cá nhân hóa, học theo giọng mẫu, thích hợp cho các bài giảng yêu cầu phong cách giọng đọc đặc thù.

Nếu lựa chọn cách tự đọc bằng mô hình AI đã được tinh chỉnh cho Tiếng Việt thì đề xuất viXTTS tại <https://huggingface.co/spaces/thinhlp/vixtts-demo> lưu ý mô hình này mới chỉ có giọng nữ đọc tốt còn giọng nam bị lỗi.



Hình 2: viXTTS mô hình AI TTS tiếng việt

OpenTTS: Mã nguồn mở, tùy chỉnh cao, miễn phí; thích hợp thử nghiệm nhưng giọng đọc có thể chưa tự nhiên bằng các giải pháp thương mại. Chất lượng kém.

ChatGPT Voice (OpenAI): Hỗ trợ tạo giọng đọc tự nhiên, giàu biểu cảm, có thể tạo nhiều phiên bản giọng khác nhau từ cùng một văn bản, tích hợp dễ dàng qua API. Đọc tốt nhưng giới hạn số ký tự mỗi lần đọc free, **nếu đọc nhiều cần API thì mất phí**.

Gợi ý lựa chọn:

Với các bài giảng chuyên nghiệp, cần giọng đọc tự nhiên, mượt mà, có thể chọn Microsoft TTS hoặc ChatGPT Voice.

Nếu cần giọng đọc cá nhân hóa theo mẫu, YourTTS thì **viXTTS** là lựa chọn tốt.

5.3 Tạo file âm thanh từ văn bản

Quy trình tạo file âm thanh với TTS gồm:

Nhập văn bản đã chuẩn hóa vào hệ thống TTS.

Điều chỉnh giọng đọc:

Chọn giọng (nam, nữ, trẻ em)

Tốc độ đọc: nhanh, vừa, chậm

Ngữ điệu, nhấn nhá theo ngữ cảnh (tăng tính tự nhiên)

Xuất file âm thanh: định dạng phổ biến như MP3 hoặc WAV, phục vụ tích hợp vào video bài giảng.

Ví dụ mã nguồn Python sử dụng thư viện pyttsx3 chuyển file txt thành mp3

```
import pyttsx3

# Khởi tạo engine TTS
engine = pyttsx3.init()

# Thiết lập giọng đọc (Windows: "Microsoft Vietnamese"
# nếu đã cài giọng)
voices = engine.getProperty('voices')
for voice in voices:
    print(voice.id) # Xem các giọng có sẵn
# Ví dụ chọn giọng Việt (thay bằng id giọng bạn muốn)
engine.setProperty('voice', 'Microsoft Vietnamese
Desktop') # hoặc voice.id tương ứng
```

```

# Điều chỉnh tốc độ đọc (mặc định ~200)
engine.setProperty('rate', 160)

# Đọc nội dung từ file TXT
txt_file = "input.txt"
with open(txt_file, "r", encoding="utf-8") as f:
    text = f.read()

# Xuất file MP3
mp3_file = "output.mp3"
engine.save_to_file(text, mp3_file)

# Thực hiện tổng hợp giọng nói
engine.runAndWait()

print(f"File MP3 đã được tạo: {mp3_file}")

```

5.4 Tạo nhiều phiên bản giọng đọc

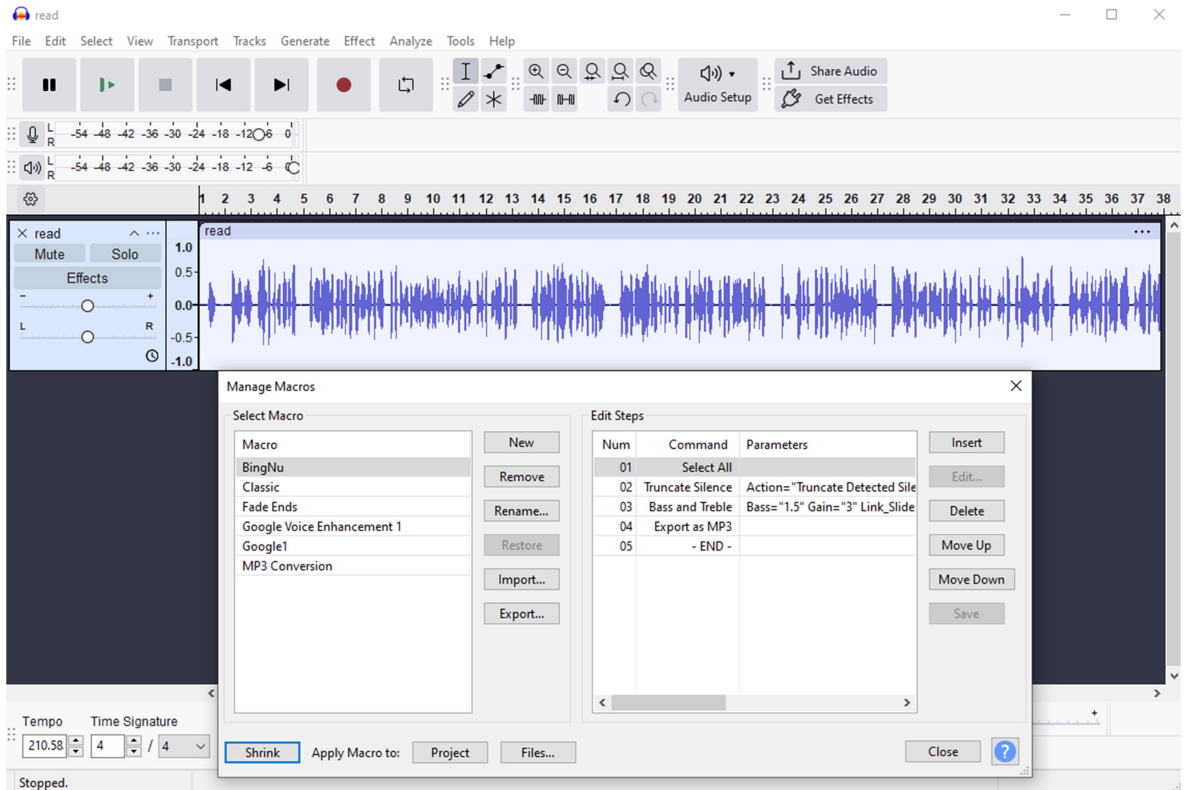
Tạo nhiều phiên bản giọng đọc khác nhau từ cùng một văn bản để so sánh chất lượng, tốc độ, nhấn nhá và độ tự nhiên.

Chọn phiên bản phù hợp nhất với đối tượng học sinh, nội dung bài giảng và phong cách giảng dạy.

Việc so sánh các phiên bản giúp đảm bảo giọng đọc cuối cùng không gây nhàm chán, dễ nghe, và phù hợp với tốc độ tiếp thu của học sinh.

Bằng cách lựa chọn đúng công nghệ TTS (MS TTS hoặc ChatGPT Voice), điều chỉnh giọng đọc và tạo nhiều phiên bản âm thanh, giáo viên có thể đảm bảo video bài giảng vừa tự nhiên, vừa dễ tiếp thu, đồng thời tiết kiệm thời gian và công sức so với thu âm truyền thống.

Hiện tại hầu hết các mô hình AI đều hỗ trợ nhiều giọng đọc: nam, nữ, theo nhân vật hay vùng miền. Thậm chí bắt chước giọng mẫu, tuy nhiên các mô hình cao cấp thường mất phí và phí tính theo số từ đọc, khó khả thi với các giáo trình lên đến hàng nghìn trang. Giải pháp miễn phí và đơn giản, có thể tự động hóa hàng trăm file là dùng phần mềm Audacity để hiệu chỉnh âm thanh. Audacity đảm bảo đáp ứng được tất cả các nhu cầu về chỉnh sửa âm thanh để có chất lượng giọng đọc tốt.



Hình 3: Giao diện quản lý Macros của Audacity

5.5 Tích hợp âm thanh vào video, đồng bộ với hình ảnh/slides

Kết hợp file âm thanh TTS với các slide, hình ảnh minh họa hoặc đoạn video ngắn.

Đồng bộ giọng đọc với các hiệu ứng, hình ảnh và chữ trên màn hình để tăng tính trực quan và sinh động.

Các công cụ phổ biến: Cupcut, Camtasia, Adobe Premiere, OBS Studio, hoặc các phần mềm tạo video bài giảng tích hợp.

Đối với số lượng cần biên soạn nhiều thì nên dùng các công cụ có thể tự động hóa như FFMPEG.

Ví dụ lệnh:

```
ffmpeg -loop 1 -i image.jpg -i audio.mp3 -c:v libx264 -c:a aac -b:a 192k -shortest -pix_fmt yuv420p output.mp4
```

Giải thích:

-loop 1: lặp ảnh vô hạn (tạo video từ ảnh tĩnh).

- i image.jpg: file ảnh đầu vào.
- i audio.mp3: file âm thanh MP3 đầu vào.
- c:v libx264: codec video H.264.
- c:a aac -b:a 192k: codec audio AAC với bitrate 192 kbps.
- shortest: kết thúc video khi audio kết thúc.
- pix_fmt yuv420p: đảm bảo video tương thích với hầu hết trình phát.

5.6 Chỉnh sửa, xuất bản và phân phối

Kiểm tra đồng bộ âm thanh – hình ảnh, hiệu chỉnh các đoạn không trùng khớp.

Xuất bản video ở định dạng phù hợp (MP4, MOV...) và phân phối trên nền tảng học tập trực tuyến hoặc LMS của nhà trường.

Video có thể cập nhật dễ dàng khi nội dung bài giảng thay đổi, chỉ cần chỉnh sửa văn bản và tạo lại âm thanh TTS.

5.7 Lợi ích và bất cập:

- Video không dùng TTS:

Giáo viên phải thu âm giọng đọc thủ công, tốn nhiều thời gian và công sức. Giọng đọc có thể không đồng đều về nhịp điệu, tốc độ và ngữ điệu giữa các phần của bài giảng.

Khó chỉnh sửa khi cần cập nhật nội dung, vì mỗi lần thay đổi văn bản đều phải thu âm lại toàn bộ hoặc phần lớn bài giảng.

- Video dùng TTS:

Giọng đọc tự động, nhịp điệu đều và ổn định, đồng bộ dễ dàng với slides, hình ảnh hoặc animation minh họa. Có thể tạo nhiều phiên bản giọng đọc khác nhau (nam, nữ, trẻ em) để so sánh và lựa chọn phiên bản phù hợp nhất với học sinh. Dễ dàng chỉnh sửa nội dung: khi cập nhật văn bản, chỉ cần tạo lại file âm thanh TTS mà không cần thu âm thủ công.

- Lợi ích khi ứng dụng TTS trong video bài giảng:

Tiết kiệm thời gian và chi phí: giảm công sức thu âm, biên tập và đồng bộ giọng đọc với video. Tăng tính tương tác và hấp dẫn: giọng đọc tự nhiên giúp học sinh tập trung hơn, nội dung bài giảng trở nên sinh động.

Dễ cập nhật nội dung: văn bản được chỉnh sửa nhanh chóng, TTS tạo lại giọng đọc chỉ trong vài phút. Hỗ trợ học tập cá nhân hóa: có thể tạo nhiều giọng đọc, điều chỉnh tốc độ, nhấn nhá và ngữ điệu phù hợp với từng đối tượng học sinh.

- Bất cập cần lưu ý:

Giọng đọc AI có thể gây nhàm chán nếu video quá dài hoặc giọng đọc không biến đổi ngữ điệu. Để duy trì sự tập trung của học sinh, thời lượng video bài giảng nên giới hạn khoảng 15 phút, kết hợp hình ảnh, slide và animation để tăng tính sinh động.

Nên tách nội dung dài thành nhiều video ngắn, mỗi video có trọng tâm rõ ràng, tránh việc học sinh mất tập trung khi nghe giọng đọc AI liên tục trong thời gian dài..

KẾT LUẬN VÀ ĐỀ XUẤT

Công nghệ TTS (Text-to-Speech) đã chứng minh là một công cụ quan trọng trong việc nâng cao chất lượng giáo dục số. Với khả năng chuyển đổi văn bản thành giọng đọc tự nhiên, TTS không chỉ giúp giáo viên tiết kiệm thời gian và chi phí so với việc thu âm truyền thống, mà còn tạo ra các bài giảng sinh động, dễ tiếp cận và cá nhân hóa. Ứng dụng TTS trong video bài giảng giúp tăng tính tương tác, giữ sự tập trung của học sinh, đồng thời hỗ trợ các học sinh có nhu cầu đặc biệt, như khuyết tật thị giác hoặc khó đọc. TTS cũng mở ra khả năng cập nhật nội dung bài giảng nhanh chóng và linh hoạt, phù hợp với bối cảnh giáo dục hiện đại, nơi nội dung cần được điều chỉnh liên tục.

Đề xuất hướng ứng dụng

Lựa chọn giải pháp TTS phù hợp với mục tiêu bài giảng

Cân nhắc chất lượng giọng đọc, khả năng tùy chỉnh, chi phí và nền tảng hỗ trợ trước khi lựa chọn công nghệ TTS.

Ví dụ: các giải pháp thương mại như Google TTS, Amazon Polly hoặc Microsoft Azure TTS phù hợp với các bài giảng yêu cầu giọng đọc tự nhiên và nhiều tùy chỉnh, trong khi các dự án mã nguồn mở như OpenTTS thích hợp cho môi trường thử nghiệm hoặc chi phí thấp.

Kết hợp TTS với video, hình ảnh minh họa, animation để tăng hiệu quả học tập

Tích hợp giọng đọc TTS vào video bài giảng, đồng bộ với slides, hình ảnh, biểu đồ hoặc animation giúp học sinh vừa nghe vừa quan sát, nâng cao khả năng tiếp thu.

Việc sử dụng các hiệu ứng hình ảnh cùng giọng đọc tự nhiên tạo ra trải nghiệm học tập đa giác quan, sinh động và hấp dẫn.

Khuyến nghị nghiên cứu thêm về cá nhân hóa giọng đọc và phát triển TTS cho ngôn ngữ ít phổ biến

Nghiên cứu các mô hình TTS có khả năng học theo giọng mẫu để tạo giọng đọc cá nhân hóa, phù hợp với từng đối tượng học sinh.

Phát triển TTS cho các ngôn ngữ ít phổ biến hoặc các biến thể địa phương giúp mở rộng khả năng tiếp cận giáo dục số đến nhiều đối tượng học sinh hơn.

Khuyến khích tích hợp các công nghệ hỗ trợ như AI để tự động điều chỉnh ngữ điệu, nhấn nhá theo nội dung và bối cảnh học tập.

Tổng thể, việc ứng dụng TTS trong giáo dục số không chỉ nâng cao chất lượng bài giảng mà còn hỗ trợ cá nhân hóa, mở rộng khả năng tiếp cận kiến thức và tối ưu hóa nguồn lực giáo dục. Đây là hướng đi tiềm năng và phù hợp với xu hướng phát triển giáo dục hiện đại.

TÀI LIỆU THAM KHẢO

- [1] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
- [2] T. Wu, Y. Zhang, H. Wang, Z. Ren, Y. Liu, X. Li, and J. Chen, “YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone,” *arXiv preprint arXiv:2209.15017*, 2022.
- [3] Google Cloud, “Text-to-Speech,” [Online]. Available: <https://cloud.google.com/text-to-speech>. [Accessed: Sep. 01, 2025].
- [4] Amazon Web Services, “Amazon Polly – Text-to-Speech Service,” [Online]. Available: <https://aws.amazon.com/polly/>. [Accessed: Sep. 01, 2025].
- [5] Microsoft, “Azure Cognitive Services – Text to Speech,” [Online]. Available: <https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/text-to-speech>. [Accessed: Sep. 01, 2025].
- [6] viXTTS, “viXTTS Demo – Vietnamese TTS,” [Online]. Available: <https://huggingface.co/spaces/thinhlp/vixtts-demo>. [Accessed: Sep. 01, 2025].
- [7] Audacity Team, Audacity (Version 3.3.2) [Computer Software], 2023. [Online]. Available: <https://www.audacityteam.org/>. [Accessed: Sep. 01, 2025].
- [8] OpenTTS, “OpenTTS – Open Source Text-to-Speech Server,” [Online]. Available: <https://opentts.github.io/>. [Accessed: Sep. 01, 2025].
- [9] Underthesea, “Underthesea – Vietnamese NLP Toolkit,” [Online]. Available: <https://github.com/undertheseanlp/underthesea>. [Accessed: Sep. 01, 2025].
- [10] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.
- [11] UNESCO, *Digital Education: Moving Towards Learning and Teaching for the 21st Century*, Paris: UNESCO Publishing, 2020.