ỨNG DỤNG MÔ HÌNH FASTPITCH TRONG BÀI TOÁN CHUYỂN ĐỔI VĂN BẢN TIẾNG VIỆT THÀNH GIỌNG NÓI

Trần Thị Dung[1], Lê Nhật Tùng[2]*, Nguyễn Trần Phong[3],
Nguyễn Khắc Anh[3], Lưu Toàn Định[4]

[1]*Trường Đại học Giao thông vận tải, Phân hiệu tại Thành phố Hồ Chí Minh*
[2]*Trường Đại học Công nghệ Đồng Nai*
[3]*Công ty TNHH Codelink*
[4]*Đại học Kinh tế Thành phố Hồ Chí Minh*
*Tác giả liên hệ: *Lê Nhật Tùng, email: lenhattung@dntu.edu.vn*

THÔNG TIN CHUNG

Ngày nhận bài: 17/10/2023

Ngày nhận bài sửa: 14/12/2023

Ngày duyệt đăng: 02/01/2024

TỪ KHOÁ

Text to speech;
Fastpitch;
Transformer;
Báo nói;
Nhận dạng lời nói

TÓM TẮT

Bài báo này giới thiệu một ứng dụng thực nghiệm của mô hình FastPitch, một mô hình học sâu mạnh mẽ cho bài toán chuyển đổi văn bản thành giọng nói (TTS). FastPitch được xây dựng trên kiến trúc Trans-former và mạng đồng tham chiếu, cho phép tạo ra giọng nói tổng hợp tự nhiên, mượt mà và chính xác. Trong bài báo này, các tác giả đã sử dụng mô hình FastPitch để tạo ra giọng nói tổng hợp cho các đoạn văn bản tiếng Việt mô tả các nội dung thông báo. Các tác giả đã đánh giá chất lượng của giọng nói tổng hợp bằng cách thu thập phản hồi từ người dùng. Kết quả cho thấy giọng nói tổng hợp do FastPitch tạo ra được người dùng đánh giá cao về độ tự nhiên, trôi chảy và khả năng truyền tải thông tin tốt. Bài báo này đóng góp cho lĩnh vực nghiên cứu TTS bằng cách cung cấp một ví dụ về cách sử dụng mô hình FastPitch cho các ứng dụng thực tế. Kết quả trong bài báo cho thấy FastPitch có tiềm năng được sử dụng trong nhiều ứng dụng khác nhau.

1. GIỚI THIỆU

1.1. Tổng quan

Trong thời đại chuyển đổi số ngày nay, chuyển đổi văn bản thành giọng nói đã trở nên quan trọng trong nhiều lĩnh vực, từ công nghệ trợ lý ảo đến đọc sách báo tự động và hỗ trợ giảng dạy trực tuyến. Bài toán này đòi hỏi khả năng chính xác, tự nhiên và trôi chảy của giọng nói tổng hợp, đặc biệt là khi xử lý các đoạn văn bản dài và phức tạp. Để nỗ lực nâng cao chất lượng và hiệu suất của hệ thống chuyển đổi văn bản thành giọng nói (text-to-speech), mô hình FastPitch đã trở thành một lựa chọn hàng đầu cho cộng đồng nghiên cứu và ứng dụng các bài toán chuyển đổi văn bản thành giọng nói.

Trong bài báo này, chúng tôi giới thiệu và ứng dụng mô hình FastPitch trong bài toán chuyển đổi văn bản thành giọng nói. Mô hình FastPitch là một mô hình tổng hợp giọng nói mới nhất dựa vào kiến trúc Transformer và mạng đồng tham chiếu. Với kiến trúc tiên tiến này, FastPitch đã chứng minh khả năng tạo ra giọng nói tổng hợp tự nhiên, trôi chảy và chính xác.

Chúng tôi tiến hành đánh giá mô hình FastPitch trên các bộ dữ liệu tự xây dựng theo thực tế, bao gồm cả văn bản thông thường và các

văn bản chứa các ngôn ngữ chuyên ngành đa dạng. Kết quả cho thấy giọng nói tổng hợp do FastPitch tạo ra được người dùng đánh giá cao về độ tự nhiên, trôi chảy và khả năng truyền tải thông tin.

Bên cạnh đó, chúng tôi cũng trình bày các kỹ thuật tiền xử lý và tối ưu hóa hiệu suất mà chúng tôi đã áp dụng để cải thiện khả năng chuyển đổi văn bản thành giọng nói với tốc độ cao và chất lượng ổn định. Những kỹ thuật này không chỉ tối ưu hóa mô hình FastPitch mà còn giúp tăng cường trải nghiệm người dùng trong việc sử dụng ứng dụng chuyển đổi văn bản thành giọng nói.

Bài báo này không chỉ tập trung vào việc trình bày mô hình FastPitch và kết quả thực nghiệm, mà còn đóng góp vào lĩnh vực nghiên cứu chuyển đổi văn bản thành giọng nói bằng việc cung cấp một cái nhìn tổng quan về ứng dụng mô hình FastPitch và tiềm năng phát triển trong tương lai. Các kết quả và kinh nghiệm trong bài báo này hy vọng sẽ đóng góp vào việc phát triển hệ thống chuyển đổi văn bản tiếng Việt thành giọng nói hiệu quả và chất lượng trong các ứng dụng thực tế.

1.2. Các công trình liên quan

Lĩnh vực tổng hợp giọng nói từ văn bản đã chứng kiến sự xuất hiện của nhiều mô hình quan trọng và đáng chú ý. Dưới đây là một danh sách các mô hình nổi trội trong lĩnh vực này:

Tacotron 2, được giới thiệu bởi Jonathan Shen et al. vào năm 2018, là một mô hình tổng hợp giọng nói sử dụng mạng neural dựa trên kiến trúc encoder-decoder và mô hình attention. Mô hình này đã đóng góp quan trọng trong việc tạo ra giọng nói tự nhiên từ văn bản (Shen et al., 2018).

Deep Voice, công bố vào năm 2017, là một mô hình tổng hợp giọng nói dựa trên deep learning sử dụng mạng neural mạng recurrent neural network (RNN) để ánh xạ từ văn bản đến âm thanh tổng hợp (Arık et al., 2017).

WaveNet là một mô hình tổng hợp giọng nói dựa trên mạng neural sâu, sử dụng mạng neural convolutional xuyên tầng để tạo ra âm thanh tổng hợp với chất lượng cao và chi tiết (Oord et al., 2016).

Transformer-TTS (Skerry-Ryan et al., 2018) là một mô hình tổng hợp giọng nói dựa trên kiến trúc transformer và attention mechanism. Mô hình này sử dụng kiến trúc transformer để tạo ra âm thanh tổng hợp tự nhiên và chất lượng cao.

DeepSpeech (Hannun et al., 2014) là một mô hình nhận dạng giọng nói tự nhiên. Mô hình này sử dụng mạng neural hồi quy đơn (unidirectional recurrent neural network) để chuyển đổi âm thanh thành văn bản.

Các công trình nghiên cứu đã đóng góp đáng kể vào phát triển mô hình tổng hợp giọng nói từ văn bản. Tuy nhiên, trong phạm vi đề tài nghiên cứu này, chúng tôi sẽ sử dụng Mô hình FastPitch cho thực nghiệm, các nguyên lý và cách xây dựng mô hình sẽ được trình bày trong phần tiếp theo.

2. MÔ HÌNH FASTPITCH

2.1. Tổng quan

Mô hình FastPitch là một mô hình tổng hợp giọng nói từ văn bản dựa trên kiến trúc transformer và mạng đồng tham chiếu. Mô hình được huấn luyện trên dữ liệu song song giữa văn bản và âm thanh để học cách dự đoán pitch và tạo ra âm thanh tổng hợp tự nhiên. Mô hình và nguyên lý hoạt động của mô hình FastPitch, như được trình bày trong bài nghiên cứu "FastPitch: Parallel Text-to-Speech with Pitch Prediction" của Adrian Łancucki (Łańcucki, 2021).

2.2. Các bước tổng hợp giọng nói

Mô hình FastPitch tổng hợp giọng nói từ văn bản đầu vào theo các bước sau:

Bước 1: Văn bản đầu vào được mã hóa thành các vectơ biểu diễn bằng bộ mã hóa.

Bước 2: Pitch được dự đoán cho mỗi khung âm thanh bằng Pitch Prediction Network.

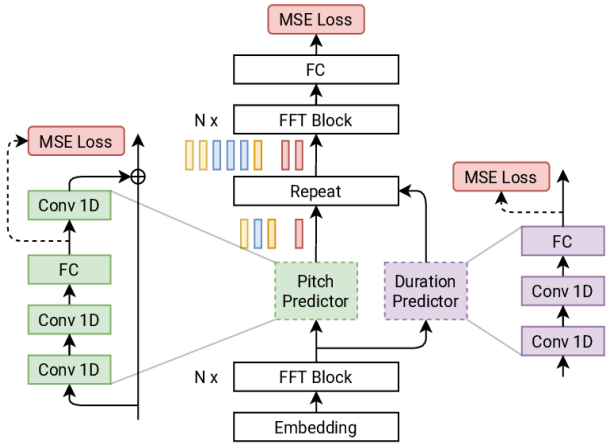
Bước 3: Các vectơ biểu diễn và pitch được sử dụng để tạo ra âm thanh tổng hợp bằng bộ giải mã.

Bước 4: Postnet được sử dụng để cải thiện chất lượng âm thanh tổng hợp.

2.3. Kiến trúc chi tiết của mô hình

Mô hình FastPitch có một kiến trúc tổng quan được thiết kế để tổng hợp giọng nói tự nhiên từ văn bản, với khả năng dự đoán pitch.

Kiến trúc mô hình FastPitch bao gồm các thành phần chính sau:



Hình 1. Mô hình Fastpitch (Łańcucki, 2021)

Bộ mã hóa: Đầu tiên, văn bản đầu vào được mã hóa thành các vectơ biểu diễn bằng một mô hình encoder. Mô hình encoder trong FastPitch sử dụng kiến trúc transformer, giúp ánh xạ từ văn bản sang các đại diện vector thông tin. Mỗi từ trong văn bản đầu vào sẽ được biểu diễn bằng một vector.

Bộ giải mã: Sau khi văn bản được mã hóa, mô hình FastPitch sử dụng một mô hình decoder để tạo ra âm thanh tổng hợp. Bộ giải mã cũng sử dụng kiến trúc transformer, tuy nhiên, có một số khác biệt so với bộ mã hóa.

Pitch Prediction Network: Đây là một phần quan trọng của mô hình FastPitch. Nó có nhiệm vụ dự đoán pitch cho âm thanh tổng hợp. Mạng dự đoán pitch sử dụng kiến trúc mạng feed-forward neural network với một số lớp fully connected và hàm kích hoạt để dự đoán pitch cho mỗi khung âm thanh.

Postnet: Sau khi âm thanh đã được tổng hợp bởi decoder, một mạng postnet được áp dụng để cải thiện chất lượng âm thanh tổng hợp. Mạng postnet trong FastPitch là một mạng convolutional neural network (CNN), được thiết kế để tạo ra âm thanh tổng hợp với chất lượng và chi tiết cao.

Bảng 1. Bảng so sánh các mô hình TTS

	FastPitch	Tacotron	WaveNet
Kiến trúc	Mạng LSTM và mạng chú ý	Mạng RNN	Mạng nơ-ron tích chập
Hiệu suất	Cao	Trung bình	Thấp
Độ phức tạp tính toán	Cao	Trung bình	Cao
Tạo giọng nói	Tự nhiên	Tự nhiên	Tự nhiên
Tốc độ	Đa dạng	Đa dạng	Chậm
Dễ huấn luyện	Khá dễ	Đòi hỏi nhiều sự điều chỉnh	Đòi hỏi nhiều sự điều chỉnh
Dung lượng	Thấp	Trung bình	Cao
Ứng dụng rộng rãi	Có	Có	Hạn chế

3. THỰC NGHIỆM

3.1. Xây dựng mô hình

3.1.1. Thu thập dữ liệu

Quá trình thu thập dữ liệu là một bước quan trọng trong việc xây dựng mô hình FastPitch và đảm bảo chất lượng âm thanh tổng hợp. Trong dự án nghiên cứu trên dữ liệu các bài thông báo trên trang chủ của trường đại học Giao thông vận tải phân hiệu tại thành phố Hồ Chí Minh (UTC2), nhóm nghiên cứu đã quyết định tự thu âm để tạo ra một dataset mang tính chất riêng của trường để dễ dàng khảo sát và thực nghiệm. Dữ liệu dùng để huấn luyện nhóm đã sử dụng các bài viết thông báo trên website UTC2, các bài viết thường có độ dài tương tự nhau với nhiều loại chủ đề như thông báo về thời khóa biểu, các thông báo liên quan đến công tác sinh viên của Nhà trường.

Dưới đây là mô tả chi tiết về quá trình thu âm từng bước thực hiện:

Chuẩn bị môi trường thu âm:

Quá trình thu âm được thực hiện ở phòng họp nhằm đảm bảo một không gian yên tĩnh và tách biệt. Với không gian này, có thể loại bỏ tiếng ồn từ bên ngoài và tạo ra một môi trường tĩnh lặng, không có sự xao lạc hay tiếng động xung quanh.

Điều này rất quan trọng trong việc đảm bảo chất lượng âm thanh thu được là tốt nhất có thể. Một không gian yên tĩnh và tách biệt giúp tối đa hóa việc ghi lại giọng nói mà không bị ảnh hưởng bởi các yếu tố ngoại vi không mong muốn như tiếng động từ bên ngoài hoặc tiếng vọng trong phòng.

Với không gian cá nhân và yên tĩnh, có thể tạo ra một môi trường thu âm chuyên nghiệp và tối ưu hóa chất lượng âm thanh. Điều này đảm bảo rằng dữ liệu thu thập được là chính xác và đáng tin cậy để sử dụng trong việc xây dựng mô hình FastPitch cho website nói UTC2.

Sử dụng thiết bị thu âm chất lượng:

Quá trình thu âm sử dụng một thiết bị thu âm chất lượng cao để đảm bảo giọng nói được ghi lại một cách rõ ràng và chân thực. Thiết bị chính là một micro thu âm.

Micro thu âm được chọn lựa kỹ càng để đáp ứng các yêu cầu của quá trình thu âm. Nó có khả năng nhận biết và ghi lại âm thanh một cách chính xác, đảm bảo rằng mọi chi tiết trong giọng nói được tái tạo một cách tốt nhất.

Micro thu âm cũng được thiết kế để giảm tiếng ồn và tiếng vọng không mong muốn trong quá trình thu âm. Điều này giúp tăng cường chất lượng âm thanh thu được và giảm thiểu các yếu tố nhiễu gây ảnh hưởng đến quá trình tiền xử lý và huấn luyện mô hình FastPitch.

Việc sử dụng một micro thu âm chất lượng cao đóng vai trò quan trọng trong việc đảm bảo chất lượng âm thanh thu thập được là tốt nhất có thể. Giúp tạo ra dữ liệu đầu vào chính xác và chất lượng để sử dụng trong quá trình xây dựng mô hình FastPitch cho website nói UTC2.

Chuẩn bị văn bản:

Quá trình lựa chọn văn bản sẽ chọn những văn bản phù hợp để thu âm trên website nói UTC2. Văn bản bao gồm các tin tức, bài viết, các câu văn hoặc đoạn văn bản ngắn. Việc lựa chọn văn bản đảm bảo có đủ dữ liệu để huấn luyện mô hình FastPitch và tạo ra âm thanh tổng hợp chất lượng.

Trước khi tiến hành thu âm, tiến hành chuẩn bị văn bản bằng cách loại bỏ các định dạng đặc biệt, các ký hiệu hay thẻ HTML không cần thiết và chỉ giữ lại nội dung chính. Điều này giúp đảm bảo rằng văn bản được đưa vào mô hình FastPitch là trong định dạng đúng và không gây ảnh hưởng đến quá trình tổng hợp giọng nói.

```
12251.wav|anh muốn uống gì không|anh muốn uống g
12252.wav|cách tốt nhất để vào trung tâm thành p
12253.wav|trái này là trái sầu riêng|trái này là
12254.wav|tôi vẫn ghét việc phải ra khỏi bếp và
12255.wav|anh có thứ gì từ thời không|anh có thứ
12256.wav|anh đánh giá sai à|anh đánh giá sai à
12257.wav|Thông báo lịch thi vấn đáp trường đại
12258.wav|tôi -N chuyến ra sân bay|tôi -N chuyến
```

Hình 2. Văn bản đã chuẩn bị để thu âm

Thu âm và lưu trữ dữ liệu:

Tiến hành quá trình thu âm và lưu trữ dữ liệu một cách cẩn thận để đảm bảo tính toàn vẹn và khả năng truy cập dễ dàng.

Sử dụng thiết bị thu âm để ghi lại giọng nói theo từng đoạn văn bản. Mỗi đoạn thu âm đã được lưu trữ thành các file âm thanh có định dạng phù hợp WAV. Điều này giúp dễ dàng quản lý và xử lý các file thu âm theo nhu cầu của dự án.

Để đảm bảo an toàn và tránh mất dữ liệu, việc sao lưu các file thu âm trên các thiết bị lưu trữ khác nhau là vô cùng cần thiết. Điều này bao gồm việc sao lưu dữ liệu lên máy tính cá nhân và các dịch vụ lưu trữ trực tuyến. Quá trình sao lưu đảm bảo rằng dữ liệu thu âm được bảo vệ và có sẵn khi cần thiết.

Vì đặc thù sinh viên tại phân hiệu trường đại học Giao thông vận tải tại thành phố Hồ Chí Minh đa phần là sinh viên thuộc khu vực miền Nam, chính vì thế nhóm tác giả đã chọn một bạn sinh viên nam với chất giọng miền Nam để phù hợp với đối tượng khảo sát. Trong tương lai nhóm sẽ bổ sung thêm đầy đủ giọng đọc của cả nam và nữ ở các vùng miền khác nhau tại Việt Nam.

3.1.2. Tiền xử lý dự liệu

Sau quá trình thu thập dữ liệu âm thanh, tiền xử lý dữ liệu được thực hiện để làm sạch và cải thiện chất lượng. Trong quá trình này, một công cụ mạnh mẽ được sử dụng là Denoiser, được phát triển bởi Facebook, để giảm thiểu nhiễu trong dữ liệu âm thanh thu thập.

**Tổng quan về Denoiser*

Mô hình Denoiser là một phần quan trọng trong quá trình xử lý âm thanh, được sử dụng để giảm tiếng ồn và cải thiện chất lượng âm thanh. Mục tiêu của mô hình Denoiser là loại bỏ tiếng ồn không mong muốn từ một tín hiệu âm thanh đã được ghi âm.

Mô hình Denoiser của Facebook là một mô hình học sâu (deep learning) được phát triển bởi Facebook AI Research (FAIR) để giảm tiếng ồn trong tín hiệu âm thanh.

Mô hình này được huấn luyện trên một lượng lớn dữ liệu âm thanh chứa tiếng ồn và âm thanh sạch để học cách loại bỏ tiếng ồn từ tín hiệu âm thanh đã cho.

** Chi tiết quá trình tiền xử lý dữ liệu:*

Sử dụng công cụ Denoiser: Sử dụng Denoiser, một công cụ xử lý tín hiệu âm thanh được phát triển bởi Facebook, để loại bỏ nhiễu trong dữ liệu thu âm của mình. Denoiser sử dụng mô hình học máy để phân tích và lọc bỏ các thành phần nhiễu không mong muốn, như tiếng ồn hạt, tiếng gió hoặc tiếng nền. Quá trình này giúp tăng cường chất lượng âm thanh và làm sạch dữ liệu thu âm trước khi đưa vào quá trình huấn luyện mô hình FastPitch.

Thực hiện xử lý Denoiser: Áp dụng công cụ Denoiser vào từng đoạn âm thanh thu âm. Công cụ này sẽ xử lý các tín hiệu âm thanh và loại bỏ các thành phần nhiễu không mong muốn. Quá trình xử lý này được thực hiện tự động và không yêu cầu sự can thiệp thủ công từ phía người dùng. Kết quả là dữ liệu âm thanh sau khi xử lý trở nên sạch hơn và tổng quát hơn.

Kiểm tra chất lượng sau xử lý: Sau khi áp dụng Denoiser, việc kiểm tra lại chất lượng của dữ liệu âm thanh là vô cùng cần thiết. Bằng cách so sánh trước và sau khi xử lý, đảm bảo rằng các thành phần nhiễu không mong muốn đã được loại bỏ một cách hiệu quả mà không làm mất đi thông tin quan trọng trong dữ liệu âm thanh. Điều này giúp đảm bảo rằng dữ liệu được sử dụng cho quá trình huấn luyện mô hình FastPitch là chất lượng và đáng tin cậy.

Tổng quan, quá trình tiền xử lý dữ liệu đã được thực hiện bằng cách sử dụng công cụ Denoiser của Facebook để loại bỏ nhiễu và cải thiện chất lượng âm thanh. Quá trình này giúp đảm bảo rằng dữ liệu thu thập được sẽ được sử dụng hiệu quả trong quá trình huấn luyện mô hình FastPitch trên website nói UTC2.

3.2. Kết quả thực nghiệm và đánh giá

Sau khi đã tích hợp mô hình vào website nói UTC2, việc tiếp theo là thực hiện thử nghiệm và đánh giá hiệu suất của mô hình đã tích hợp. Quá

trình này giúp đảm bảo rằng mô hình hoạt động một cách chính xác và đáng tin cậy trong môi trường sản xuất.

Các bài viết nhóm dùng để thực nghiệm được lấy trên website UTC2. Các bài viết thường có chủ đề về đào tạo, công tác sinh viên, thư viện, học phí,… Các bài viết thường có độ dài từ 100-300 từ, sau khi xử lý thì sẽ phát một đoạn thông báo nói tầm 2-4 phút cho mỗi bài viết.

Trong phần thực nghiệm, nhóm tác giả sử dụng độ đo Mean Opinion Scores. Mean Opinion Scores (MOS) là phương pháp đánh giá dựa theo giá trị trung bình chủ quan do chính con người đánh giá. Phương phương pháp này phù hợp cho những mô hình không thể đánh giá được bằng một công thức cụ thể mà phải dựa vào cảm quan của con người để đánh giá. Như ở trong bài báo này, để đánh giá được giọng đọc tự nhiên, mượt mà và chính xác khi chuyển văn bản thành âm thanh chúng ta sử dụng phương pháp MOS.

Phương pháp này chia thang điểm đánh giá từ 1 cho tới 5.

Bảng 2. Thang điểm phương pháp đánh giá MOS

Điểm	Nhãn
1	Quá tệ (Bad)
2	Tệ (Poor)
3	Trung bình (Fair)
4	Tốt (Good)
5	Xuất sắc (Excellent)

Sau khi đánh giá, chúng ta sẽ lấy giá trị trung bình trung của tất cả điểm đánh giá (phương trình 1).

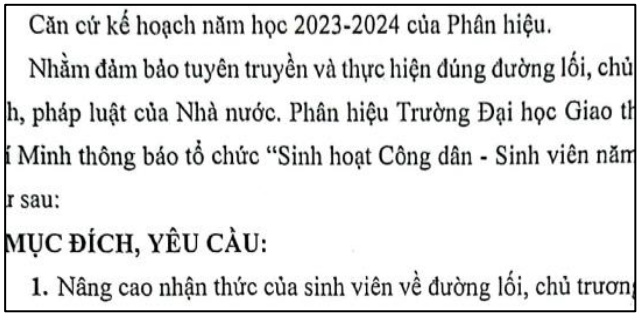
$$MOS = \sum_{n=1}^{N} R_n \quad (1)$$

Trong phương trình trên thì R là đánh giá đơn lẻ cho một hệ thống bởi N người.

Điểm mạnh của mô hình: đọc khá rõ và ổn định nhưng vẫn có một vài điểm chưa khắc phục như: khả năng đọc tiếng anh chưa tốt (dataset chưa có nhiều dữ liệu tiếng anh, giọng đọc chưa chuẩn)

Đánh giá người dùng: Tiến hành cho 100 người là sinh viên Công nghệ thông tin tại Phân hiệu Trường Đại học Giao Thông Vận Tải tại Thành phố Hồ Chí Minh nghe và đánh giá các bài thông báo trên website chính thức của trường. Để tránh tình trạng người khảo sát nghe liên tục các bài viết sẽ khó đánh giá chính xác hiệu quả của mô hình, nhóm nghiên cứu đã tiến hành triển khai website chạy trên server, khi đó sinh viên có thể khảo sát ở bất cứ lúc nào và có thời gian nhiều hơn để nghe từng bài viết theo thời gian rãnh. Sau khi khảo sát kết quả nhóm thu lại như sau:

Một số bài thông báo trên website thực nghiệm:

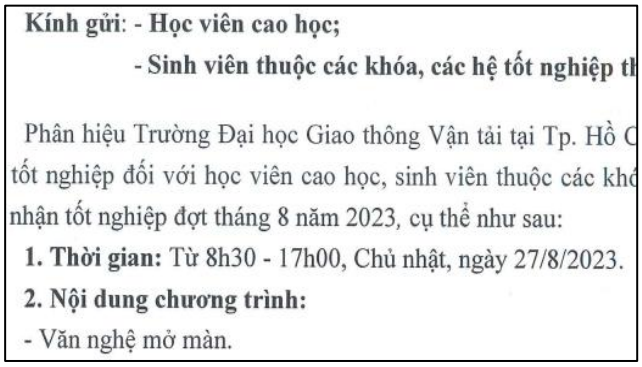

Căn cứ kế hoạch năm học 2023-2024 của Phân hiệu.

Nhằm đảm bảo tuyên truyền và thực hiện đúng đường lối, chủ
h, pháp luật của Nhà nước. Phân hiệu Trường Đại học Giao th
í Minh thông báo tổ chức "Sinh hoạt Công dân - Sinh viên năm
r sau:

MỤC ĐÍCH, YÊU CẦU:

1. Nâng cao nhận thức của sinh viên về đường lối, chủ trươn

Hình 3. Bài viết khảo sát 1


Kính gửi: - Học viên cao học;

- Sinh viên thuộc các khóa, các hệ tốt nghiệp th

Phân hiệu Trường Đại học Giao thông Vận tải tại Tp. Hồ C
tốt nghiệp đối với học viên cao học, sinh viên thuộc các khó
nhận tốt nghiệp đợt tháng 8 năm 2023, cụ thể như sau:

1. Thời gian: Từ 8h30 - 17h00, Chủ nhật, ngày 27/8/2023.

2. Nội dung chương trình:

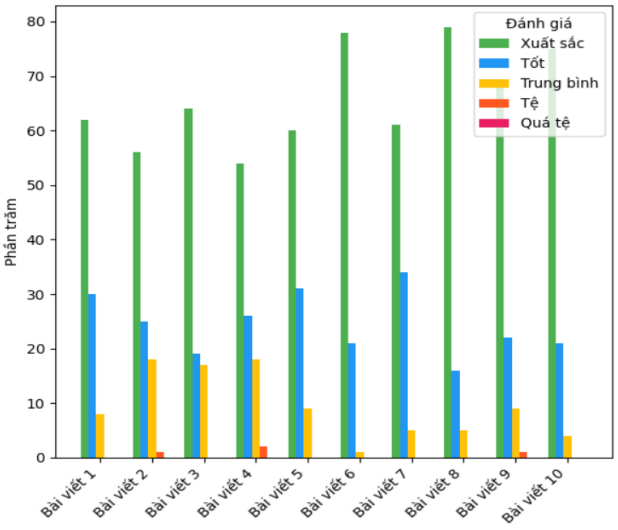
- Văn nghệ mở màn.

Hình 4. Bài viết khảo sát 2

Hình 3 và Hình 4 là một số bài viết trên website chính thức của trường dùng để khảo sát. Kết quả khảo sát được thể hiện trong Bảng 3 ở dưới.

Bảng 3. Bảng kết quả khảo sát

	Xuất sắc	Tốt	Trung bình	Tệ	Quá tệ
Bài viết 1	62%	30%	8%	0%	0%
Bài viết 2	56%	25%	18%	1%	0%
Bài viết 3	64%	19%	17%	0%	0%
Bài viết 4	54%	26%	18%	2%	0%
Bài viết 5	60%	31%	9%	0%	0%
Bài viết 6	78%	21%	1%	0%	0%
Bài viết 7	61%	34%	5%	0%	0%
Bài viết 8	79%	16%	5%	0%	0%
Bài viết 9	68%	22%	9%	1%	0%
Bài viết 10	75%	21%	4%	0%	0%



Hình 5. So sánh kết quả khảo sát

4. KẾT LUẬN

Bài báo trình bày về mô hình FastPitch, một mô hình được xây dựng để chuyển đổi văn bản thành giọng nói tự nhiên. Mô hình sử dụng bộ dữ liệu âm thanh và văn bản để huấn luyện và sử dụng kiến trúc Transformer và cơ chế attention để tạo ra âm thanh chất lượng cao. Kỹ thuật huấn luyện theo sự chú ý được áp dụng để cải thiện chất lượng giọng nói. Mô hình FastPitch đã đạt được kết quả tốt và được tích hợp vào website nói UTC2 để tạo ra âm thanh từ bài viết. Trong tương lai, nhóm tác giả sẽ tiến hành xây dựng bộ dữ liệu đa dạng hơn, cải tiến them mô hình FastPitch để giọng đọc được tự nhiên và có cảm xúc hơn.

TÀI LIỆU THAM KHẢO

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., … Shoeybi, M. (2017). Deep Voice: Real-time Neural Text-to-Speech. *Proceedings of the 34th International Conference on Machine Learning*, 195–204. PMLR. Retrieved from https://proceedings.mlr.press/v70/arik17a.html

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., … Ng, A. Y. (2014, December 19). *Deep Speech: Scaling up end-to-end speech recognition*. arXiv. https://doi.org/10.48550/arXiv.1412.5567

Łańcucki, A. (2021). Fastpitch: Parallel Text-to-Speech with Pitch Prediction. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6588–6592. https://doi.org/10.1109/ICASSP39728.2021.9413889

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016, September 19). *WaveNet: A Generative Model for Raw Audio*. arXiv. https://doi.org/10.48550/arXiv.1609.03499

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., … Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., … Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *Proceedings of the 35th International Conference on Machine Learning*, 4693–4702. PMLR. Retrieved from https://proceedings.mlr.press/v80/skerry-ryan18a.html

APPLYING THE FASTPITCH MODEL IN THE PROBLEM OF CONVERTING VIETNAMESE TEXT INTO SPEECH

Tran Thi Dung[1], Le Nhat Tung[2]*, Nguyen Tran Phong[3], Nguyen Khac Anh[3], Luu Toan Dinh[4]

[1]*University of Transport and Communications, Campus in Ho Chi Minh City*
[2]*Dong Nai Technology University*
[3]*Codelink Company Limited*
[4]*University of Economics Ho Chi Minh City*
*Corresponding author: *Le Nhat Tung, email: lenhattung@dntu.edu.vn*

GENERAL INFORMATION

Received date: 17/10/2023

Revised date: 14/12/2023

Published date: 02/01/2024

KEYWORD

Text to speech;
Fastpitch;
Transformer;
According to newspaper;
Recognize speech form

ABSTRACT

This paper presents an experimental application of the FastPitch model, a powerful deep learning model for text-to-speech (TTS). FastPitch is built on the Transformer architecture and reference network, which enables natural, fluent, and accurate speech synthesis. In this paper, the authors use FastPitch to generate synthetic speech for text descriptions of announcements. The authors evaluate the quality of the synthetic speech by collecting user feedback. The results show that the synthetic speech generated by FastPitch is highly rated by users for naturalness, fluency, and information delivery. This paper contributes to the field of TTS research by providing an example of how FastPitch can be used for real-world applications. The research results have suggested that FastPitch has the potential to be used in a variety of applications.