


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
https://youtu.be/Z_1vYBup8DU
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/ReichelHuy/NguyenDinhHuy-CS519.O21.KHTN/blob/main/ZERO-SHOT%20EMOTIONAL%20TEXT-TO-SPEECH%20SYNTHESIS.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Huỳnh Phạm Đức Lâm• MSSV: 21521050 	<ul style="list-style-type: none">• Lớp: CS519.O21.KHTN• Tự đánh giá (điểm tổng kết môn): 8.5/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 8• Số câu hỏi QT của cả nhóm: 1• Link Github: https://github.com/darklul03/• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Làm poster○ Làm phương pháp và nội dung nghiên cứu○ Làm video YouTube
<ul style="list-style-type: none">• Họ và Tên: Nguyễn Đình Huy• MSSV: 22520558	<ul style="list-style-type: none">• Lớp: CS519.O21.KHTN• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 9• Số câu hỏi QT của cả nhóm: 1



- Link Github:
<https://github.com/ReichelHuy/NguyenDinhHuy-CS519.O21.KHTN>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
 - Làm đề cương, làm slide
 - Viết mô tả, giới thiệu.
 - Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CHUYỂN ĐỔI VĂN BẢN THÀNH GIỌNG NÓI MANG THEO CẢM XÚC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ZERO-SHOT EMOTIONAL TEXT-TO-SPEECH SYNTHESIS

TÓM TẮT (Tối đa 400 từ)

Tổng hợp giọng nói cảm xúc từ văn bản (Emotional Text-To-Speech - ETTS) là một phần quan trọng trong việc phát triển các hệ thống (ví dụ như các hệ thống tự động trả lời, tele sale,...) yêu cầu giọng nói tự nhiên và cảm xúc. Tuy nhiên, các phương pháp hiện tại chỉ nhằm mục đích tạo ra các mô hình ETTS cho các nhân vật đã được nhận biết trong quá trình đào tạo, mà không xem xét đến việc khái quát cho các nhân vật chưa được huấn luyện trước.

Trong bài toán này, chúng tôi đề xuất nghiên cứu một phương pháp zero-shot để chuyển đổi văn bản thành giọng nói có cảm xúc, cho phép người dùng tổng hợp giọng nói cảm xúc của bất kỳ nhân vật nào chỉ bằng cách sử dụng một đoạn ngôn ngữ không có cảm xúc ngắn và nhãn cảm xúc tương ứng. Cụ thể, chúng tôi đề xuất phương pháp học zero-shot dựa trên mô hình khuếch tán (Diffusion model) để tổng hợp giọng nói dựa trên cảm xúc được yêu cầu.

GIỚI THIỆU (Tối đa 1 trang A4)

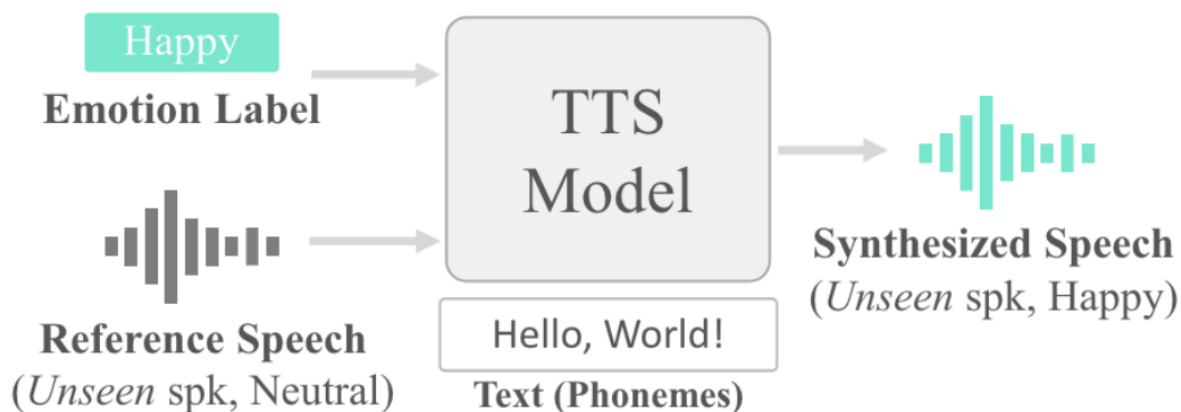
Trong bài toán này, chúng tôi đề xuất nghiên cứu một phương pháp học zero-shot để chuyển đổi văn bản thành giọng nói có cảm xúc, cho phép người dùng tổng hợp giọng nói cảm xúc của bất kỳ nhân vật nào chỉ bằng cách sử dụng một đoạn ngôn ngữ không có cảm xúc ngắn và nhãn cảm xúc tương ứng. Cụ thể, chúng tôi đề xuất phương pháp học zero-shot dựa trên mô hình khuếch tán (Diffusion model) để tổng hợp giọng nói dựa trên cảm xúc được yêu cầu.

Input của bài toán:

- Một đoạn văn bản (phonemes) bất kỳ.
- Một giọng nói tham chiếu Y (Mel-spectrogram)
- Nhãn cảm xúc e

Output của bài toán:

- Giọng nói tham chiếu (Mel-spectrogram) tổng hợp Y' với cảm xúc mong muốn.



MỤC TIÊU (Viết trong vòng 3 mục tiêu)

- Thành công trong việc tạo ra mô hình chuyển đổi văn bản thành giọng nói tương ứng với giọng nói tham chiếu.
- Giọng nói ở đầu ra rõ ràng, rành mạch, lưu loát, gần nhất với giọng nói tham chiếu ở đầu vào.
- Giọng nói ở đầu ra cần mang cảm xúc gần nhất với cảm xúc được ghi trên nhãn đầu vào.

NỘI DUNG VÀ PHƯƠNG PHÁP

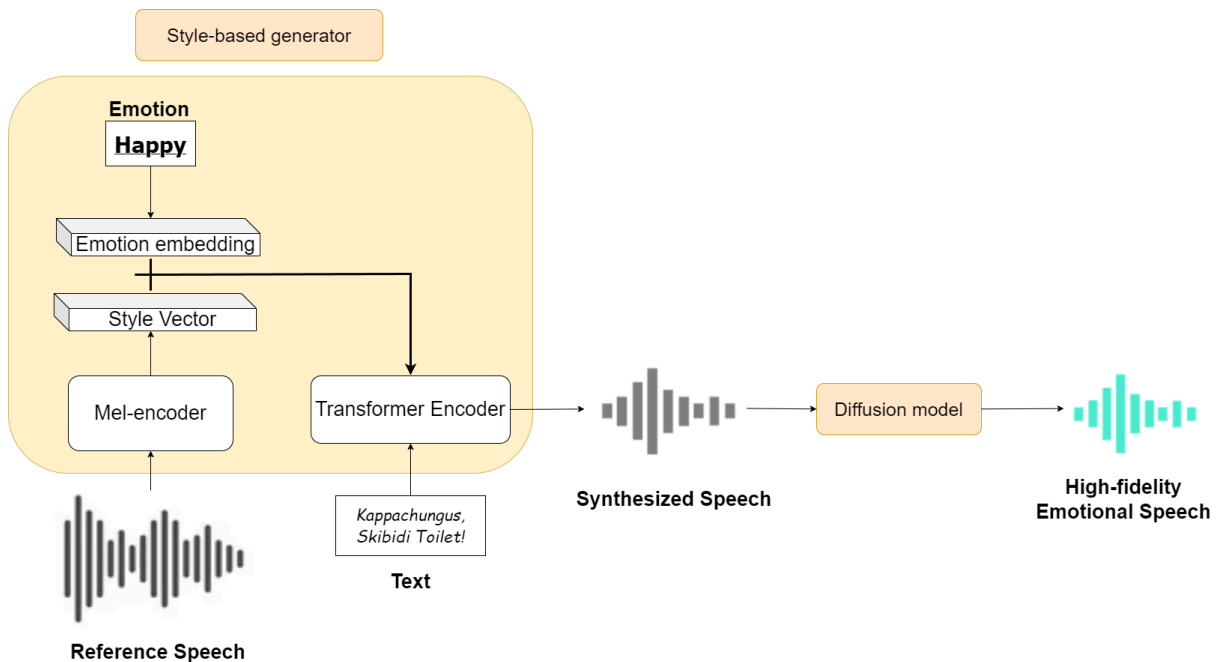
Nội dung:

- Khảo sát các phương pháp chuyển đổi văn bản thành giọng nói (TTS) hiện nay. Các hệ thống này có thể sinh ra các lời nói có chất lượng cao, nhiều chất giọng.
- Khảo sát các giải pháp TTS có thể kiểm soát được cảm xúc của giọng nói
- Nghiên cứu các kỹ thuật Domain Adaptation, Adversarial Training, Guidance cho diffusion model

- Nghiên cứu cách áp dụng diffusion model và style-based generators trong bài toán multi-speaker adaptive TTS, sử dụng các giọng nói không nằm trong training data. Từ đó đặt ra giả thuyết: có thể áp dụng mô hình này trong trường hợp TTS có bao gồm thêm cảm xúc hay không (trong bối cảnh zero-shot)?

Phương pháp:

Chúng tôi đề xuất kiến trúc gồm 2 mô hình được kết hợp với nhau để tạo ra văn bản nói: style-based generator [4] và diffusion model [5].



Mô hình style-based generator được cấu tạo từ 2 thành phần: Mel-style encoder lấy đầu vào là một giọng nói và sinh ra style vector đóng vai trò embedding giọng nói của bất kì người nào. Sau đó, transformer encode sẽ lấy style vector cùng với văn bản đầu vào để sinh ra văn bản nói ứng với đầu vào

Các mô hình Grad-tts [1] và Grad-StyleSpeech [6] đã đề xuất thêm mô hình diffusion để xử sinh ra các giọng nói giống nguyên bản. Để áp dụng trong bài toán giọng nói cảm xúc, chúng tôi thêm vào vector cảm xúc vào trong transformer encoder.

Đánh giá mô hình:

- Để có thể đạt được đúng giọng nói mang theo cảm xúc tương ứng ở đầu vào, chúng tôi sử dụng thang đo Emotional Classifier Accuracy (ECA).
- Để có thể chuyển được văn bản thành giọng nói tham chiếu, chúng tôi sử dụng thang đo Character Error Rate (CER).
- Cuối cùng, để đo độ tương đồng giữa âm thanh đầu ra và đầu vào, chúng tôi sử dụng thang đo Speaker Embedding Cosine Similarity (SECS).

KẾT QUẢ MONG ĐỢI

- Mô hình đề xuất đạt được hiệu quả tốt trên các thang đo: Emotional Classifier Accuracy (ECA), Character Error Rate (CER) và Embedding Cosine Similarity (SECS) trên các bộ dataset khác nhau.
- Cải thiện độ chính xác của mô hình với các kỹ thuật xử lý trong các mô hình trong kiến trúc.
- Bảng kết quả so sánh độ hiệu quả của các kỹ thuật và chi phí tính toán phát sinh.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, “Grad-tts: A diffusion probabilistic model for text-to-speech,” in ICML, 2021.
- [2] M. Kang, D. Min, and S. J. Hwang, “Any-speaker adaptive text-to-speech synthesis with diffusion models,” arXiv, vol. abs/2211.09383, 2022.
- [3] S. Kim, H. Kim, and S. Yoon, “Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data,” arXiv, vol. abs/2205.15370, 2022.
- [4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in CVPR 2019, Computer Vision Foundation / IEEE, pp. 4401–4410.
- [5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in ICLR, 2021.

[6] M. Kang, D. Min, and S. J. Hwang, “Any-speaker adaptive text-to-speech synthesis with diffusion models,” arXiv, vol. abs/2211.09383, 2022.