CHUYỂN ĐỔI VĂN BẢN THÀNH GIỌNG NÓI MANG THEO CẢM XÚC

Nguyễn Đình Huy - 22520558

Huỳnh Phạm Đức Lâm - 21521050

Thông tin nhóm

Họ và tên: Nguyễn Đình Huy

MSSV: 22520558

Họ và tên: Huỳnh Phạm Đức Lâm

MSSV: 21521050

- Link Github của nhóm: https://github.com/ReichelHuy/NguyenDinhHuy-CS519.O21.KHTN
- Link YouTube video: https://youtu.be/Z_1vYBup8DU

Tóm tắt

Tổng hợp giọng nói cảm xúc từ văn bản (Emotional Text-to-Speech Synthesis - ETTS) là một lĩnh vực nghiên cứu quan trọng trong các ứng dụng công nghệ ngôn ngữ tự nhiên, như hội thoại, trợ lý ảo và phân tích cảm xúc người dùng. Nhiệm vụ chính của lĩnh vực này là chuyển đổi một văn bản đầu vào thành một đoạn giọng nói mang cảm xúc tương ứng, giúp tăng tính tự nhiên và hấp dẫn của các ứng dụng nói trên.

Tuy nhiên, việc tổng hợp giọng nói cảm xúc từ văn bản gặp phải một số thách thức chính. Trước hết, cần phải tích hợp thành công thông tin ngôn ngữ (văn bản) và phi ngôn ngữ (nhịp điệu, âm sắc) để tạo ra giọng nói mang cảm xúc phù hợp. Các phương pháp hiện tại chỉ nhằm mục đích tạo ra các mô hình ETTS cho các giọng nói đã được nhận biết trước trong quá trình đào tạo, mà không xem xét đến việc khái quát cho các giọng nói chưa được huấn luyện từ trước.

Giới thiệu

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.

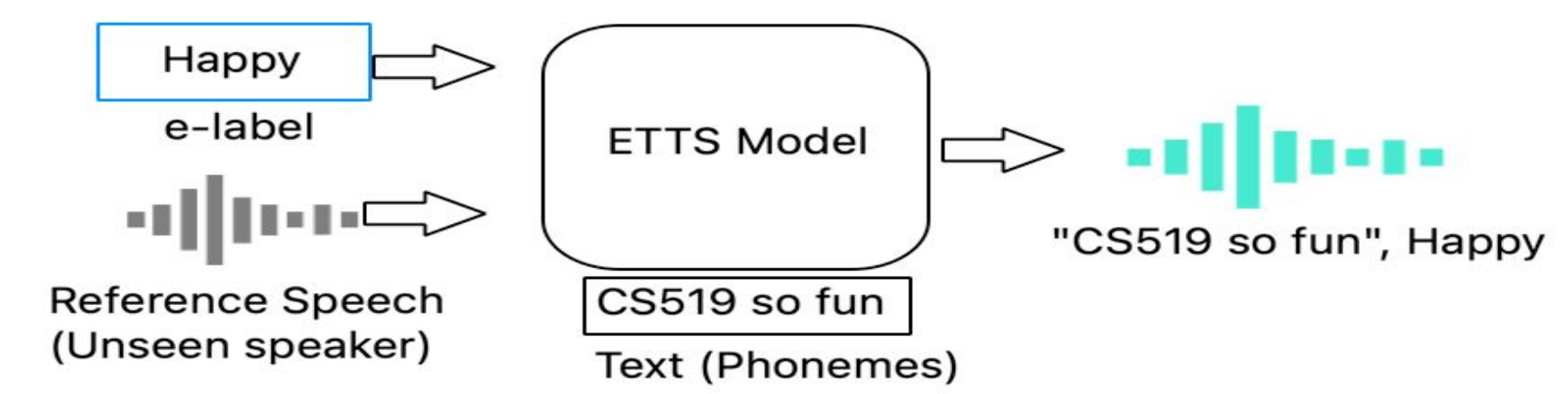
Giới thiệu

Input của bài toán:

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

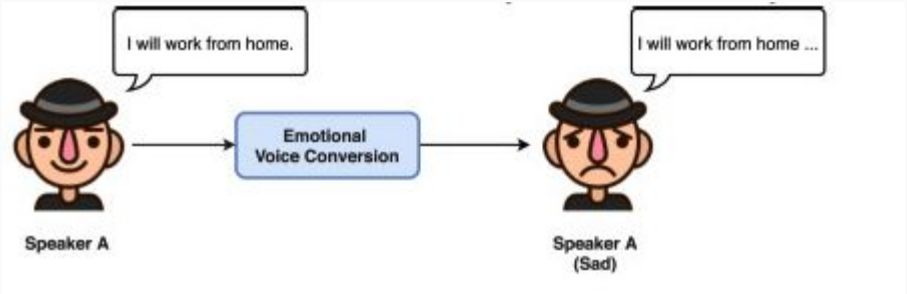
Output của bài toán:

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

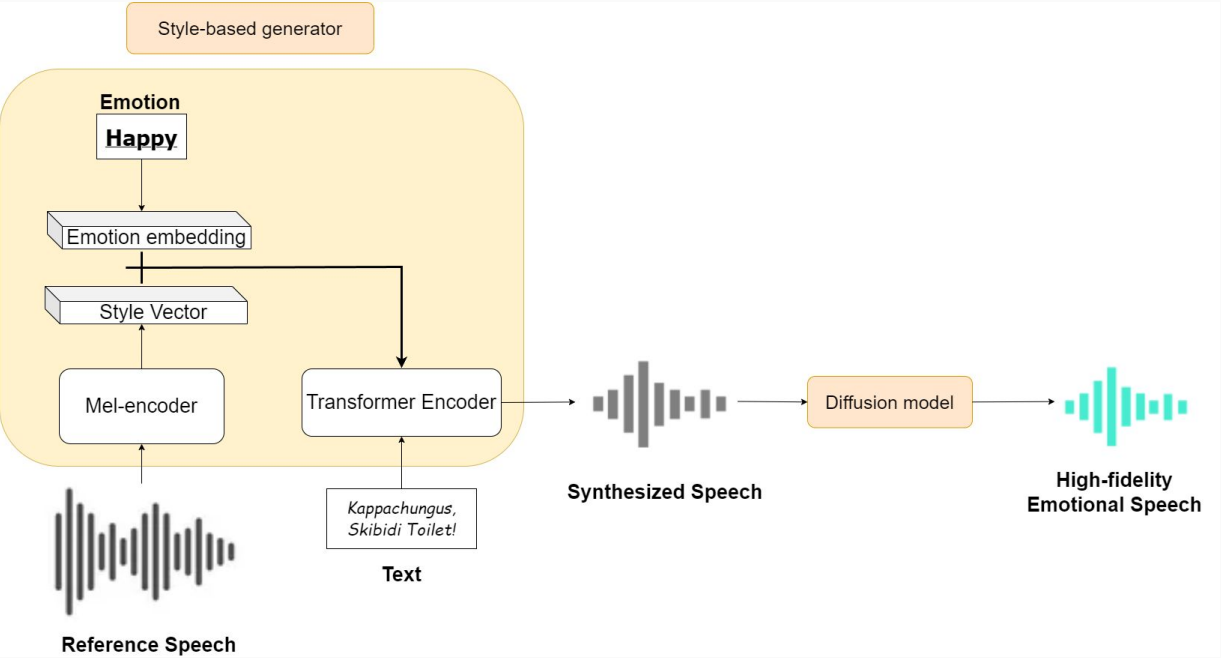


Mục tiêu

- 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

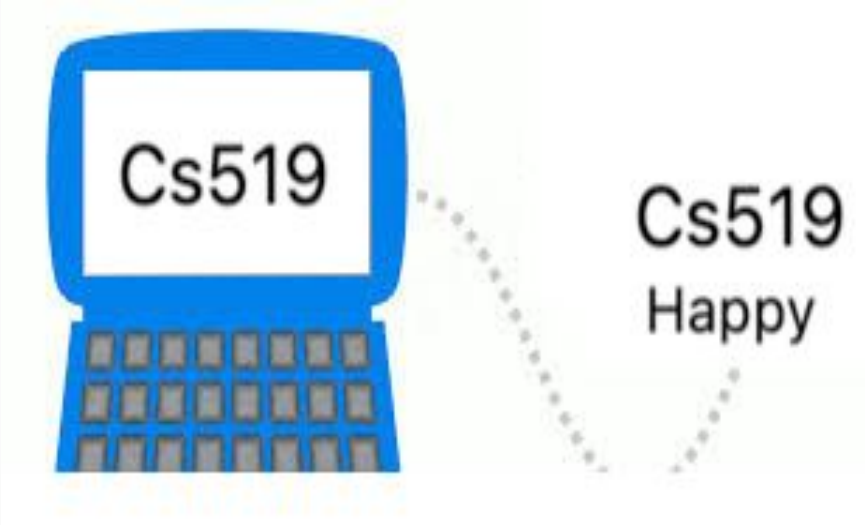


Nội dung và Phương pháp

- Kiến trúc dựa theo Grad-Style Speech: gồm 2 giai đoạn kết hợp giữa style-based generator và diffusion model.
  - Trích xuất đặc trưng từ giọng nói tham chiếu để tổng hợp style vector và kết hợp với vector embedding cảm xúc.
  - Đưa vào vector đặc trưng vào Transformer Encoder cùng với văn bản cần chuyển đổi thành giọng nói để sinh ra giọng nói có cảm xúc.
  - Áp dụng diffusion model để biến đổi giọng nói trở nên chân thật.
- Thử nghiệm các kỹ thuật trên style-based generator và diffusion model để tăng hiệu quả mô hình.
- Sử dụng dataset tiếng Anh ESD trong training và LibriTTS để đánh giá.
- Sử dụng độ đo đánh giá ECA và SECS.

Kết quả dự kiến

- 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 lí 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

[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.