Chuyển đổi văn bản thành giọng nói mang theo cảm xúc

Huỳnh Phạm Đức Lâm **Nguyễn Đình Huy**

Trường đại học công nghệ thông tin - Đại học quốc gia thành phố Hồ Chí Minh

What ?

Chúng tôi đề xuất một mô hình chuyển đổi văn bản với đầu vào là giọng nói tham chiếu không mang theo cảm xúc, một nhãn cảm xúc, và văn bản cần chuyển đổi.

- Đề xuất phương pháp học zero-shot để tổng hợp giọng nói dựa trên style-based generators và diffusion-model.
- Thực nghiệm trên bộ dataset ESD và LibriTTS
- Đánh giá bằng các phương pháp Sota

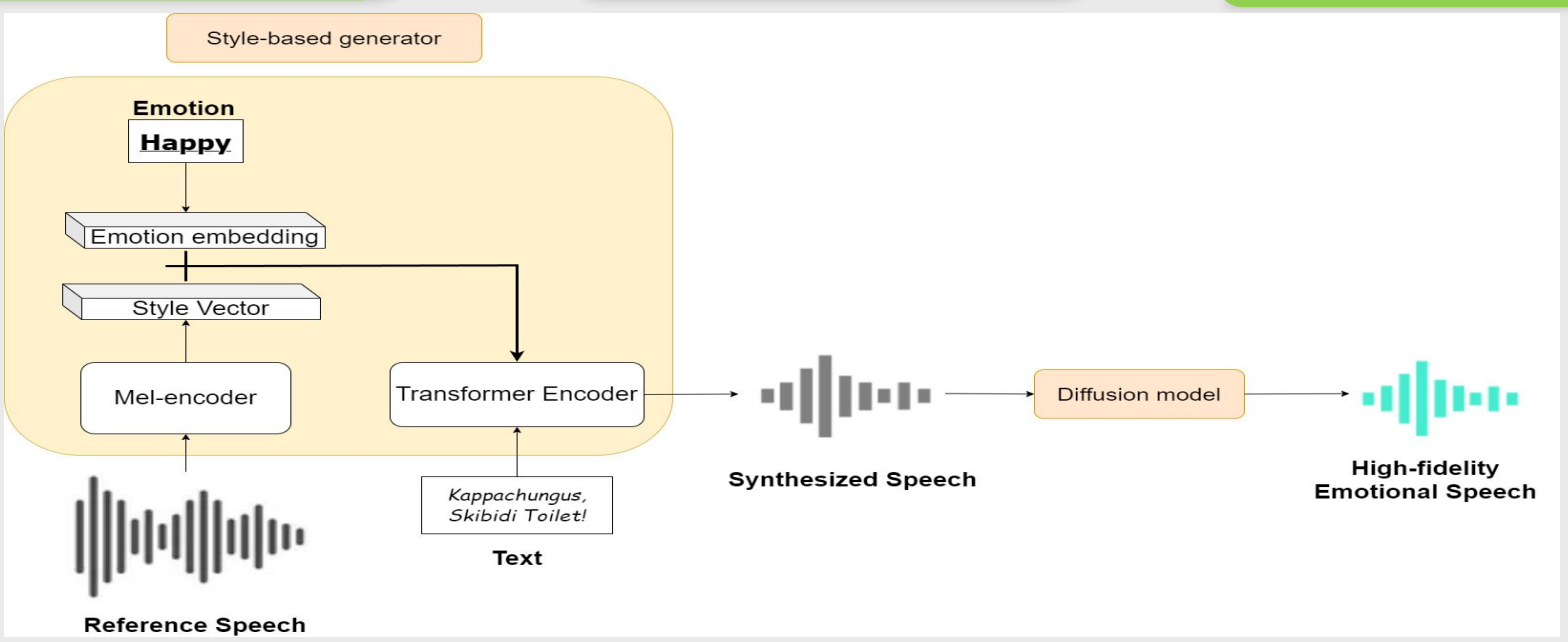
Why ?

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 giữa người và máy, trợ lý ảo và phân tích cảm xúc người dùng thông qua văn bản, sách nói,...

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. Sau đó, cần phải tăng hiệu suất của mô hình. Các phương pháp hiện có chỉ nhằm mục đích tạo ra các mô hình 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.

Overview

Tách style vector và kết hợp vector embedding cảm xúc → Tổng hợp giọng nói với văn bản đầu vào có chứa cảm xúc → Sử dụng diffusion model để sinh ra giọng nói chân thật



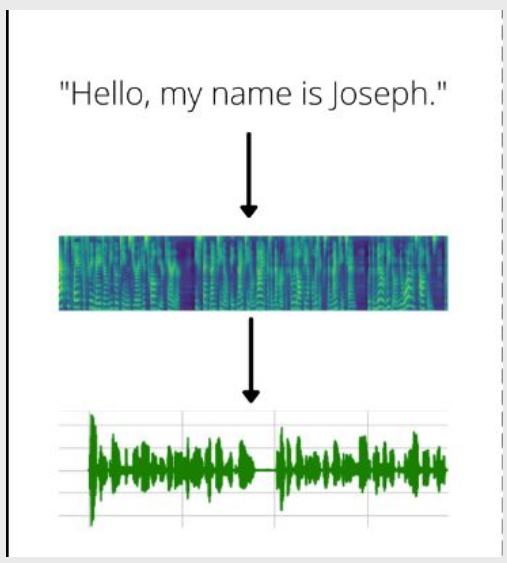
Description

1. Tách style vector và kết hợp vector embedding cảm xúc.

- Sử dụng mel-encoder để trích xuất đặc trưng trong văn bản tham chiếu.
- Kết hợp với vector embedding cảm xúc để chuẩn bị đầu vào cho bước tiếp theo
- Thử nghiệm với các kỹ thuật học đối trọng để loại bỏ các đặc trưng cảm xúc trong style vector, giúp vector tổng hợp được có ý nghĩa tốt hơn.

2. Tổng hợp giọng nói với văn bản đầu vào có chứa cảm xúc

- Từ vector đặc trưng của giọng nói và cảm xúc ở trước, đưa vào transformer encoder cùng với văn bản đầu vào để sinh ra văn bản nói.



3. Sử dụng diffusion model để sinh ra giọng nói chân thật

- Sử dụng diffusion model để tăng tính chân thật cho giọng nói được sinh ra.
- Thử nghiệm các kỹ thuật guidance cho mô hình.