Building a Vietnamese Text to Speech Model

1. Giới thiệu

Text to Speech giúp chuyển đổi văn bản viết thành ngôn ngữ nói. Trong báo cáo này, em đề xuất phương pháp xây dựng mô hình Text to Speech ở Việt Nam, đồng thời đưa ra các ý tưởng và thách thức khi xây dựng.

2. Phương pháp đề xuất

2.1 Thu nhập dữ liệu

Để xây dựng một mô hình TTS hiệu quả cần có bộ dữ liệu chất lượng cao:
- Nguồn thu thập: trên báo chí, youtube, phim ảnh, …
- Định dạng dữ liệu: audio và văn bản tương ứng.
- Tính đa dạng: bao gồm nhiều âm giọng khác nhau (Bắc, Trung, Nam) để dữ liệu được đa dạng.

Các bộ dữ liệu nổi bật:

- Được tạo ra từ tiểu thuyết, truyện ngắn: https://github.com/NTT123/Vietnamese-Text-To-Speech-Dataset
- Bộ dữ liệu có phân biệt theo từng vùng miền: https://huggingface.co/datasets/nguyendv02/ViMD_Dataset

2.2 Tiền xử lí dữ liệu

Quá trình tiền xử lí dữ liệu giúp chuyển đổi sang một định dạng nhất quán, bao gồm các bước:
- Chuẩn hóa văn bản bằng cách mở rộng các chữ viết tắt, ký hiệu và số thành từ đầy đủ. Ví dụ:
    - ksao -> không sao,
    - % -> phần trăm
    - 2024 -> hai nghìn không trăm hai tư
- Mã hóa câu thành từ và âm tiết trong khi vẫn duy trì các điểm đánh dấu âm điệu.
- Lọc âm thanh để loại bỏ tiếng ồn xung quanh và cân bằng tập dữ liệu theo tông giọng và âm điệu.
- Chuyển chữ hoa thành chữ thường để dữ liệu đồng nhất.

2.3 Chuyển đổi văn bản sang âm vị

Trong mô hình TTS, tiếng nói được tổng hợp ở mức ký tự, nghĩa là mỗi ký tự trong văn bản sẽ được mô hình xử lý trực tiếp để tạo ra âm thanh tương ứng. Ví dụ, các ký tự như a, ă, b, c,… sẽ được xử lý để phát âm. Tuy nhiên, trong tiếng Việt, có những ký tự có cách phát âm khác nhau tùy thuộc vào ngữ cảnh của từ. Ví dụ:

- Kí tự ‘c’ trong ‘cá’ phát âm khác với kí tự ‘c’ trong ‘cháu’
- Nếu mô hình chỉ học từ các ký tự mà không hiểu ngữ cảnh, thì sẽ dễ gây ra nhầm lẫn trong cách phát âm khi sinh ra giọng nói.

Vì vậy việc chuyển văn bản đầu vào sang dạng âm vị là thiết yếu vì âm vị là phiên âm chi tiết của mỗi từ, mô tả cách đọc cụ thể hơn so với mức kí tự.

2.4 Mô hình âm học

Mô hình âm học là thành phần giúp mô hình hóa giọng nói dưới dạng các tham số âm học đại diện cho những đặc trưng giọng nói của con người. Sau đó có thể tái tạo lại giọng nói từ những giá trị âm học đó. Để có được những đặc trưng âm học từ văn bản đầu vào, cần phải thực hiện các bước trung gian:

- **Vectơ hóa âm vị**: giúp biến đổi chuỗi âm vị đầu vào thành một vectơ, trong đó mỗi âm vị được đại diện bởi số duy nhất. Quá trình này giúp các phép tính toán có thể thực hiện được trên máy tính.
- **Dự đoán đặc trưng**: sử dụng mô hình để có đặc trưng âm học cho một văn bản đầu vào. Sau quá trình huấn luyện sẽ có được mô hình dự đoán những đặc trưng âm học từ văn bản đầu vào và có thể tổng hợp âm thanh từ đặc trưng âm học được dự đoán.

2.5 Mô hình tạo tiếng nói

Dữ liệu trung gian được dự đoán ở mô hình âm học sẽ được đưa vào mô hình sinh tiếng nói gọi là Vocoder. Mô hình này sẽ chuyển đổi những thông tin âm học sang dạng tín hiệu âm thanh (waveform), từ đó có thể tạo tiếng nói.

2.6 Huấn luyện mô hình

- Huấn luyện mô hình trên tác vụ dự đoán âm học (MSE hoặc MAE loss).
- Huấn luyện mô hình trên tác vụ tạo tiếng nói.

⇨ Có thể huấn luyện đồng thời mô hình âm học và mô hình tiếng nói giúp cải thiện sự liên kết và tự nhiên hơn.

2.7 Độ đo đánh giá

- **MOS (Điểm ý kiến trung bình)**: Người nghe đánh giá giọng nói được tạo sau đó đưa ra điểm số theo thang điểm từ 1-5.
- **WER (Word Error Rate)**: Độ chính xác của của tiếng nói được tạo.

3. Các thách thức và hướng giải quyết

	Thách thức	Giải quyết
Dữ liệu	Chất lượng và số lượng dữ liệu	Tiền xử lí dữ liệu tốt, tang cường dữ liệu. Thu nhập thêm dữ liệu để bảo đảm tính đa dạng.
Khả năng tổng quát hóa	Khác miền dữ liệu, khác cao độ và tông giọng	Tăng cường sự đa dạng dữ liệu, sử dụng các mô hình nhận diện tốt các cao độ và tông giọng.
Overfitting	Mô hình có thể bị overfit	Sử dụng điều chuẩn (weight decay hoặc dropout), sử dụng tập dev đa dạng hơn.
Inference	Latency, kích thước mô hình	Sử dụng mô hình nhẹ và phù hợp.