Báo cáo đề xuất xây dựng hệ thống Text-to-Speech (TTS) tiếng Việt

Người thực hiện: Trung Trương

23/08/2025

1 Giới thiệu

Text-to-Speech (TTS) là công nghệ chuyển đổi văn bản thành giọng nói tự nhiên, đóng vai trò quan trọng trong nhiều ứng dụng như trợ lý ảo, đọc báo tự động, hỗ trợ người khiếm thị, audiobook hay chatbot. Với tiếng Việt – một ngôn ngữ có thanh điệu phức tạp và nhiều đặc trưng ngữ âm riêng – việc xây dựng hệ thống TTS đạt chất lượng cao là một thách thức lớn nhưng cần thiết để nâng cao trải nghiệm người dùng và khả năng tiếp cận thông tin.

1.1 Tiền xử lý văn bản

Bước đầu tiên trong pipeline là chuẩn hóa văn bản. Quá trình này bao gồm việc chuyển các viết tắt sang dạng đầy đủ (ví dụ: "TP.HCM" thành "Thành phố Hồ Chí Minh"), tách câu dựa vào dấu câu và ngữ cảnh, cũng như chuyển đổi số và ký hiệu thành dạng đọc được (ví dụ: "2025" thành "hai nghìn không trăm hai mươi lăm"). Đặc biệt, khi văn bản chứa các từ tiếng nước ngoài hoặc tiếng Anh, hệ thống cần nhận diện các phần code-switch để xử lý phát âm chính xác cho từng ngôn ngữ.

1.2 Phân tích ngôn ngữ

Sau khi chuẩn hóa, văn bản được phân tích ngôn ngữ, bao gồm tách từ tiếng Việt bằng các công cụ như VnCoreNLP hoặc RDRSegmenter. Tiếp đó, hệ thống chuyển đổi từ văn bản sang chuỗi âm vị (phoneme) theo bảng IPA và gán thanh điệu phù hợp cho tiếng Việt. Đối với các từ ngoại ngữ, lexicon riêng như CMU Pronouncing Dictionary hoặc mô hình grapheme-to-phoneme (G2P) được sử dụng để đảm bảo phát âm đúng. Quá trình này cũng bao gồm nhận diện ngôn ngữ theo token để mô hình có thể xử lý tốt câu trộn nhiều ngôn ngữ.

1.3 Mô hình sinh giọng nói

Phần sinh giọng nói (Acoustic Model) là bước trung tâm của pipeline. Các mô hình deep learning như Tacotron 2 hoặc FastSpeech 2 được đề xuất. Tacotron 2 là mô hình Seq2Seq với cơ chế attention, cho phép chuyển từ chuỗi phoneme sang spectrogram một cách tự nhiên. FastSpeech 2, dựa trên kiến trúc Transformer, mang lại tốc độ nhanh hơn, giảm lỗi nhịp điệu và dễ dàng mở rộng cho các trường hợp câu trộn nhiều ngôn ngữ

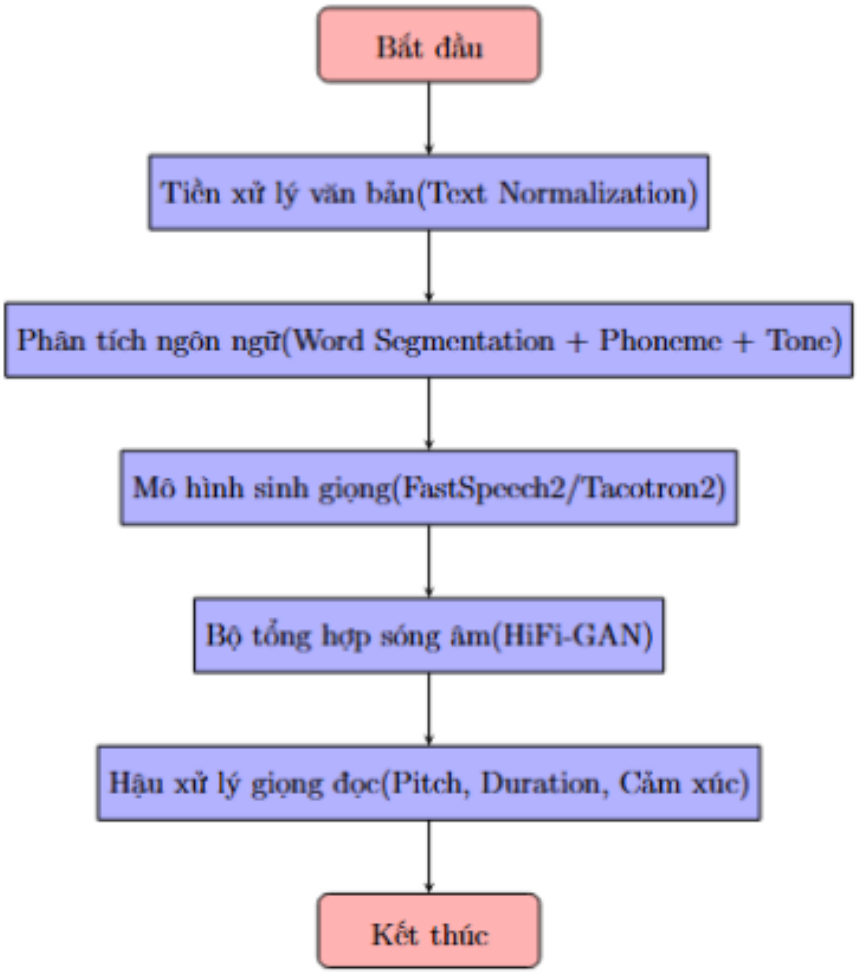
(multilingual/code-switch TTS). Các mô hình này được fine-tune với dữ liệu giọng đọc tiếng Việt chuẩn để nâng cao chất lượng phát âm.

1.4 Tổng hợp sóng âm (Vocoder)

Spectrogram từ mô hình sinh giọng được chuyển thành sóng âm (waveform) thông qua vocoder. Hiện nay, HiFi-GAN là lựa chọn ưu tiên nhờ khả năng sinh giọng tự nhiên, tốc độ nhanh và ổn định. Các vocoder khác như WaveNet hay WaveGlow có thể được sử dụng nhưng tốc độ chậm hơn.

1.5 Hậu xử lý

Sau khi tạo sóng âm, bước hậu xử lý giúp điều chỉnh cao độ, tốc độ đọc và cảm xúc (vui, buồn, trang trọng). Việc áp dụng prosody modeling giúp giọng đọc trở nên tự nhiên hơn, phù hợp với ngữ cảnh và nội dung văn bản.



Hình 1: Pipeline tổng quan của hệ thống TTS

2 Các khó khăn và giải pháp

Xây dựng TTS tiếng Việt gặp một số khó khăn chính. Thứ nhất là thanh điệu tiếng Việt, với sáu thanh điệu riêng biệt, có thể bị phát âm sai nếu mô hình không học tốt ngữ cảnh. Giải pháp là gắn nhãn rõ ràng thanh điệu khi xây dựng phoneme và thu thập dữ liệu đa dạng theo vùng miền. Thứ hai là vấn đề dữ liệu huấn luyện, do TTS cần lượng dữ liệu giọng đọc chuẩn lớn, trong khi các dataset tiếng Việt hạn chế. Có thể khắc phục bằng việc thu thập dữ liệu giọng đọc thực tế từ 20-50 giờ và sử dụng các dataset công khai như VIVOS, VLSP, Common Voice. Thứ ba là tốc độ suy luận, bởi một số vocoder truyền thống như WaveNet chậm khi tạo audio; giải pháp là kết hợp FastSpeech2 với HiFi-GAN để tăng tốc mà vẫn giữ chất lượng.

Một khó khăn quan trọng khác là câu văn trộn tiếng Anh hoặc ngoại ngữ. Nếu không xử lý đúng, mô hình có thể phát âm sai hoặc gán thanh điệu tiếng Việt vào từ nước ngoài. Giải pháp là áp dụng language ID theo token, lexicon ngoại ngữ và xây dựng hệ thống code-switch TTS. Thêm vào đó, việc đa dạng giọng đọc và cảm xúc cũng là thách thức, có thể giải quyết bằng fine-tune nhiều giọng và áp dụng multi-speaker TTS với speaker embedding. Cuối cùng, với các từ không có trong lexicon, hệ thống có thể fallback đọc gần đúng hoặc cho phép người dùng tùy chỉnh lexicon.

Hơn nữa, việc giọng đọc tự nhiên cũng là một khó khăn với TTS Tiếng Việt. Giọng điệu trong câu hỏi hoặc câu cảm thán khác với câu khẳng định. Nếu hệ thống TTS không có cơ chế mô phỏng ngữ điệu (prosody modeling), giọng đọc sẽ trở nên máy móc, thiếu tự nhiên. Giải pháp là áp dụng mô hình học ngữ điệu hoặc đánh dấu (tag) các câu hỏi, câu cảm thán để điều chỉnh cao độ (pitch) và độ dài âm (duration) phù hợp, giúp giọng đọc tự nhiên hơn.

3 Ví dụ pipeline xử lý câu trộn ngôn ngữ

Ví dụ, câu **"Tôi đang học Machine Learning ở trường"** sẽ được xử lý như sau: bước tiền xử lý giữ nguyên "Machine Learning", language ID phân biệt các từ tiếng Việt và tiếng Anh, phoneme mapping sử dụng IPA + tone cho tiếng Việt và CMUdict cho tiếng Anh, mô hình FastSpeech2 multilingual tạo spectrogram, và vocoder HiFi-GAN chuyển thành waveform. Kết quả là giọng đọc tự nhiên, phát âm đúng cả tiếng Việt và tiếng Anh.

4 Kết luận

Hệ thống TTS tiếng Việt đề xuất gồm các bước: **chuẩn hóa văn bản → tách từ và phoneme + gán thanh điệu → mô hình sinh giọng FastSpeech2/Tacotron2 → vocoder HiFi-GAN → hậu xử lý giọng đọc**. Pipeline này cho phép phát âm chính xác, xử lý câu trộn ngoại ngữ, đa dạng giọng đọc và cảm xúc, đồng thời khắc phục các vấn đề về thanh điệu, dữ liệu, tốc độ và từ không có trong lexicon. Với hướng tiếp cận này, hệ thống TTS có thể được triển khai trong nhiều ứng dụng thực tế, nâng cao trải nghiệm người dùng và khả năng tiếp cận thông tin cho mọi đối tượng.