

BÁO CÁO NGHIÊN CỨU TUẦN 12

CHỦ ĐỀ: CÔNG NGHỆ TỔNG HỢP TIẾNG NÓI (TEXT-TO-SPEECH)

Ngày 16 tháng 12 năm 2025

Tóm tắt nội dung

Báo cáo này trình bày tổng quan về bài toán Text-To-Speech (TTS), phân tích quá trình phát triển từ các hệ thống ghép nối cơ bản đến các mô hình Generative AI hiện đại (Few-shot learning). Báo cáo cũng đi sâu vào ưu nhược điểm của từng phương pháp, các trường hợp sử dụng cụ thể, và cách các nghiên cứu hiện nay tối ưu hóa pipeline để cân bằng giữa hiệu suất, tài nguyên và tính tự nhiên.

Mục lục

1	Tổng quan về bài toán và Bối cảnh nghiên cứu	2
2	Bức tranh toàn cảnh: Các cấp độ phát triển của TTS	2
2.1	Level 1: Phương pháp ghép nối (Concatenative/Parametric TTS)	2
2.2	Level 2: Deep Learning chuyên biệt (Neural TTS)	2
2.3	Level 3: Generative AI & Few-shot Learning	2
3	Phân tích chi tiết: Ưu nhược điểm và Ứng dụng	2
3.1	So sánh các phương pháp tiếp cận	2
4	Tối ưu hóa Pipeline trong các nghiên cứu hiện tại	2
4.1	Tối ưu hóa tốc độ: Non-Autoregressive Models	3
4.2	Tối ưu hóa tính tự nhiên: Variance Adaptor	3
4.3	Tối ưu hóa dữ liệu: Self-Supervised Learning (SSL)	3
5	Đạo đức nghiên cứu: Watermarking	4
6	Kết luận	4

1 Tổng quan về bài toán và Bối cảnh nghiên cứu

Bài toán **Text-To-Speech (TTS)** là quá trình chuyển đổi văn bản đầu vào thành tín hiệu âm thanh lời nói tương ứng, sao cho người nghe có thể hiểu được nội dung và cảm nhận được ngữ điệu tự nhiên.

Trong bối cảnh hiện nay, sự bùng nổ của Deep Learning và các mô hình ngôn ngữ lớn (LLMs) đã đưa TTS sang một kỷ nguyên mới. Việc nghiên cứu TTS không chỉ dừng lại ở việc "đọc đúng" mà còn phải "đọc hay", "có cảm xúc" và khả năng "sao chép giọng" (voice cloning) chỉ từ dữ liệu mẫu cực ngắn.

2 Bức tranh toàn cảnh: Các cấp độ phát triển của TTS

Dựa trên tiến trình công nghệ, có thể chia các phương pháp TTS thành 3 cấp độ (Level) chính:

2.1 Level 1: Phương pháp ghép nối (Concatenative/Parametric TTS)

Đây là phương pháp truyền thống. Hệ thống hoạt động dựa trên việc lưu trữ các đơn vị âm thanh nhỏ (từ điển âm vị) và ghép nối chúng lại theo quy tắc ngôn ngữ học, hoặc sử dụng các tham số vật lý của bộ máy phát âm để tổng hợp tiếng nói.

2.2 Level 2: Deep Learning chuyên biệt (Neural TTS)

Sử dụng các mạng nơ-ron sâu (DNN, LSTM, Transformer) để học mối quan hệ phi tuyến tính giữa văn bản và âm thanh. Các mô hình này thường bao gồm hai phần: *Acoustic Model* (tạo Mel-spectrogram từ văn bản) và *Vocoder* (tạo sóng âm từ Mel-spectrogram). Ví dụ: Tacotron 2, FastSpeech.

2.3 Level 3: Generative AI & Few-shot Learning

Đây là hướng nghiên cứu tiên tiến nhất (SOTA). Mô hình được huấn luyện trên lượng dữ liệu khổng lồ (hàng trăm ngàn giờ) và có khả năng học ngữ cảnh (in-context learning). Chỉ cần 3-10 giây âm thanh mẫu (prompt), mô hình có thể tạo ra giọng nói mới mang đặc trưng của người nói đó mà không cần huấn luyện lại (Zero-shot/Few-shot). Ví dụ: VALL-E, XTTs.

3 Phân tích chi tiết: Ưu nhược điểm và Ứng dụng

3.1 So sánh các phương pháp tiếp cận

So sánh các phương pháp ở bảng 1

4 Tối ưu hóa Pipeline trong các nghiên cứu hiện tại

Để giải quyết bài toán cân bằng giữa chất lượng và hiệu suất, các nghiên cứu hiện tại tập trung vào các kỹ thuật tối ưu hóa Pipeline sau:

Phương pháp	Ưu điểm	Nhược điểm	Trường hợp sử dụng
Level 1 (Truyền thống)	<ul style="list-style-type: none"> Tốc độ phản hồi cực nhanh (Low latency). Tốn rất ít tài nguyên tính toán. Hoạt động ổn định, không bị lỗi "ảo giác". 	<ul style="list-style-type: none"> Giọng đọc vô cảm, máy móc (robotic). Khó tùy chỉnh ngữ điệu. Ngắt nghỉ thiêu tự nhiên. 	<ul style="list-style-type: none"> Thiết bị nhúng, IoT cấu hình thấp. Hệ thống gọi số, thông báo khẩn cấp.
Level 2 (Neural TTS)	<ul style="list-style-type: none"> Âm thanh rất tự nhiên, mượt mà. Tài nguyên suy luận (inference) chấp nhận được sau khi đã train. 	<ul style="list-style-type: none"> Cần nhiều dữ liệu chất lượng cao của 1 người để train (Single-speaker). Kém linh hoạt: Muốn đổi giọng phải train lại từ đầu. 	<ul style="list-style-type: none"> Trợ lý ảo (Siri, Google Assistant). Sách nói (Audiobook). Tổng đài tự động (IVR).
Level 3 (Generative AI)	<ul style="list-style-type: none"> Linh hoạt tuyệt đối: Sao chép giọng bất kỳ chỉ với vài giây mẫu. Đa ngôn ngữ và giàu cảm xúc. 	<ul style="list-style-type: none"> Tốn nhiều tài nguyên tính toán (GPU). Tốc độ chậm hơn. Rủi ro đạo đức (Deepfake). 	<ul style="list-style-type: none"> Sáng tạo nội dung (Content Creator). Lồng tiếng phim tự động. NPC trong Game.

Bảng 1: Bảng so sánh 3 hướng tiếp cận TTS

4.1 Tối ưu hóa tốc độ: Non-Autoregressive Models

- Vấn đề:** Các mô hình cũ (như Tacotron) sinh âm thanh tuần tự (tại thời điểm t phụ thuộc vào $t - 1$), gây chậm trễ.
- Giải pháp:** Chuyển sang kiến trúc song song (Parallel) như *FastSpeech 2*. Mô hình dự đoán toàn bộ phổ âm cùng lúc thay vì từng bước.
- Kết quả:** Tăng tốc độ suy luận lên hàng chục lần, giảm độ trễ, phù hợp cho triển khai thời gian thực.

4.2 Tối ưu hóa tính tự nhiên: Variance Adaptor

- Giải pháp:** Bổ sung các module dự đoán biến số (Variance Adaptor) vào pipeline để kiểm soát:
 - Duration Predictor*: Dự đoán thời lượng phát âm của từng từ.
 - Pitch & Energy Predictor*: Dự đoán cao độ và năng lượng.
- Kết quả:** Giọng nói có nhấn nhá, cảm xúc hơn và tránh hiện tượng đọc đều đùa.

4.3 Tối ưu hóa dữ liệu: Self-Supervised Learning (SSL)

- Giải pháp:** Sử dụng các mô hình học biểu diễn âm thanh từ lượng lớn dữ liệu không nhãn (như Wav2Vec 2.0, HuBERT) làm nền tảng trước khi fine-tune cho tác vụ TTS.
- Kết quả:** Giảm thiểu lượng dữ liệu chất lượng cao cần thiết cho Level 2 và tăng khả năng zero-shot cho Level 3.

5 Đạo đức nghiên cứu: Watermarking

Với sự phát triển của Level 3 (Voice Cloning), nguy cơ Deepfake là rất lớn. Các nghiên cứu hiện đại đề xuất tích hợp **Audio Watermarking**:

- Nhúng một tín hiệu kỹ thuật số không thể nghe thấy bằng tai thường vào trong phổi âm thanh đầu ra.
- Tín hiệu này bền vững ngay cả khi file âm thanh bị nén, cắt ghép hoặc thêm nhiễu.
- Mục đích: Để phân biệt âm thanh do AI tạo ra và âm thanh thật, chống giả mạo thông tin.

6 Kết luận

Nghiên cứu TTS đã chuyển dịch từ việc cô gắng ghép nối âm thanh sang việc mô phỏng cơ chế tạo tiếng nói của con người bằng AI.

- Nếu ưu tiên **tốc độ và chi phí thấp**: Chọn Level 1 hoặc các mô hình Level 2 đã được tối ưu hóa (Distilled models).
- Nếu ưu tiên **tính linh hoạt và sáng tạo**: Chọn Level 3 (Generative TTS).

Tương lai của TTS sẽ là sự kết hợp của Level 2 (hiệu suất) và Level 3 (linh hoạt), hướng tới các hệ thống "Real-time Conversational AI" với độ trễ cực thấp và cảm xúc chân thực.