Họ tên: Hoàng Kim Giáp

MSV: 22001255

TÌM HIỂU VỀ BÀI TOÁN TEXT TO SPEECH

1. Giới thiệu

Text-to-Speech (TTS) là lĩnh vực nghiên cứu trọng yếu trong Xử lý Ngôn ngữ Tự nhiên (NLP) và Xử lý Tín hiệu Số, tập trung vào việc chuyển đổi văn bản viết thành tín hiệu tiếng nói tổng hợp có độ tự nhiên và dễ hiểu cao. Công nghệ này đóng vai trò nền tảng trong nhiều ứng dụng đột phá như trợ lý giọng nói thông minh (Virtual Assistants, ví dụ: Siri, Google Assistant), sản xuất sách nói tự động (Audiobook Generation), hỗ trợ tiếp cận thông tin cho người khiếm thị (qua các thiết bị như screen readers), dịch thuật song ngữ theo thời gian thực (Simultaneous Dubbing), và các hệ thống đối thoại tự động (Conversational AI, như chatbot giọng nói).

Trong thập kỷ qua, TTS đã chứng kiến sự phát triển vượt bậc, được phân chia thành ba cấp độ công nghệ chính (Level 1–3). Sự dịch chuyển này được thúc đẩy bởi sự thay đổi mạnh mẽ trong kiến trúc mô hình, nguồn dữ liệu và yêu cầu về khả năng tính toán, dẫn đến cải thiện đáng kể về độ tự nhiên, tính đa ngôn ngữ, và khả năng mô phỏng giọng nói cá nhân. Báo cáo này trình bày một bức tranh tổng quan và phân tích chi tiết về ba hướng tiếp cận cốt lõi này, đánh giá các ưu điểm, nhược điểm, và cách các nghiên cứu hiện đại tối ưu hóa pipeline nhằm nâng cao hiệu suất và chất lượng đầu ra. Ngoài ra, với sự phát triển nhanh chóng đến năm 2025, các mô hình mới như Fish Speech V1.5 và Gemini 2.5 TTS đã mang lại những cải tiến đáng kể về độ trễ thấp và biểu cảm tự nhiên hơn.

1.1 Định nghĩa bài toán

Text-to-Speech (TTS) là bài toán chuyển đổi đầu vào dạng văn bản (text) thành dạng tín hiệu âm thanh có thể nghe được. Mục tiêu của hệ thống TTS hiện đại không chỉ dừng lại ở việc “phát âm đúng”, mà còn hướng đến việc tạo ra giọng nói có tính tự nhiên, biểu cảm và gần với con người nhất có thể. Cụ thể, TTS phải xử lý các yếu tố như ngữ điệu (prosody), nhấn nhá (stress), và tốc độ nói để giọng nói không bị máy móc. Ví dụ, trong tiếng Việt, TTS cần xử lý đúng thanh điệu và âm vị để tránh hiểu lầm (như phân biệt "ma" và "má").

1.2 Các thành phần cấu trúc cốt lõi của một hệ thống TTS

Trước kỷ nguyên Deep Learning, các hệ thống TTS thường được chia thành nhiều module xử lý nối tiếp. Với sự xuất hiện của End-to-End TTS, nhiều module đã được hợp nhất. Tuy nhiên, về mặt chức năng, một hệ thống TTS hiện đại thường bao gồm hai thành phần chính: Mô hình Âm học (Acoustic Model) và Bộ Mã hóa Giọng nói (Vocoder). Pipeline cơ bản có thể được minh họa như sau: Text → Text Analysis → Linguistic Features → Acoustic Model → Acoustic Features → Vocoder → Waveform.

1.2.1 Phân tích Ngôn ngữ và Tiền xử lý (Text Analysis and Preprocessing)

Đây là bước đầu tiên chuyển đổi văn bản thô thành một dạng dễ dàng được mô hình học sâu xử lý.

- Chuẩn hóa Văn bản (Text Normalization): Chuyển đổi các ký hiệu phi chữ cái (số, ngày, tiền tệ, từ viết tắt) thành dạng chữ cái đầy đủ (ví dụ: "10%" thành "mười phần trăm", "USD 100" thành "một trăm đô la Mỹ"). Bước này rất quan trọng để tránh lỗi phát âm, đặc biệt trong các ngôn ngữ như tiếng Việt với nhiều từ viết tắt (e.g., "TP.HCM" thành "Thành phố Hồ Chí Minh").
- Phân tích Ngữ pháp/Ngữ âm (Linguistic Analysis): ○ Grapheme-to-Phoneme (G2P): Chuyển đổi các chữ cái thành chuỗi các âm vị (phonemes) tương ứng. Đây là bước quan trọng để đảm bảo phát âm chính xác. Ví dụ, trong tiếng Anh, "knight" được chuyển thành /naɪt/; trong tiếng Việt, "nguyễn" thành /ŋwiən/. Các mô hình G2P hiện đại sử dụng neural networks để cải thiện độ chính xác. ○ Dự đoán ngữ điệu (Prosody Prediction): Phân tích vị trí dấu câu, cấu trúc câu để dự đoán cao độ (pitch), độ dài âm (duration), và cường độ (energy) thích hợp cho giọng nói. Ví dụ, câu hỏi thường có pitch tăng ở cuối câu để tạo cảm giác tự nhiên.

1.2.2 Mô hình Âm học (Acoustic Model)

Mô hình âm học đóng vai trò cốt lõi trong việc ánh xạ các đơn vị ngôn ngữ (âm vị hoặc ký tự đã được xử lý) thành các đặc trưng âm học (Acoustic Features).

● Đầu vào: Chuỗi âm vị/ký tự đã được xử lý.
● Đầu ra: Các đặc trưng âm học ở dạng phổ tần số thấp, phổ biến nhất là Mel-spectrogram hoặc Mel-Frequency Cepstral Coefficients (MFCCs).

● Vai trò: Kiểm soát nội dung (content) và ngữ điệu (prosody) của tiếng nói được tổng hợp. Trong các mô hình hiện đại, acoustic model có thể tích hợp attention mechanisms để tập trung vào các phần quan trọng của văn bản.

1.2.3 Bộ mã hóa giọng nói (Vocoder)

Bộ mã hóa giọng nói chịu trách nhiệm chuyển đổi các đặc trưng âm học tần số thấp do Mô hình âm học tạo ra trở lại thành tín hiệu âm thanh thô (raw audio waveform).

● Đầu vào: Mel-spectrogram hoặc các đặc trưng âm học.
● Đầu ra: Tín hiệu âm thanh liên tục Y\mathbf{Y}Y.

● Vai trò: Đảm bảo chất lượng âm thanh (audio quality), độ rõ ràng và độ tự nhiên ở cấp độ sóng âm (waveform level). Các vocoder hiện đại như WaveNet hoặc HiFi-GAN sử dụng generative adversarial networks (GANs) để tạo sóng âm mượt mà hơn, giảm nhiễu và tăng tốc độ suy luận.

Trong các hệ thống End-to-End TTS, Mô hình âm học và Vocoder được huấn luyện chung trong một mạng lưới duy nhất để tối ưu hóa trực tiếp chất lượng sóng âm, giúp loại bỏ lỗi tích lũy giữa các module.

2. Bức tranh toàn cảnh về TTS

Sự tiến hóa của TTS được đánh dấu bằng sự chuyển dịch từ các phương pháp dựa trên luật (Rule-based) sang các mô hình Học Sâu và gần đây là các Mô hình cơ sở (Foundation Models).

2.1. Level 1 - TTS dựa trên Luật âm vị và ghép nối đơn vị

Đây là giai đoạn khởi đầu, nơi việc tổng hợp tiếng nói được thực hiện thông qua việc áp dụng các quy tắc ngữ âm học đã được định nghĩa trước hoặc ghép nối các đơn vị âm thanh (phones, diphones, syllables) được trích xuất từ một cơ sở dữ liệu tiếng nói lớn (Hunt & Black, 1996; Tokuda et al., 2000). Ví dụ, hệ thống concatenative TTS ghép các đoạn ghi âm ngắn (diphones như "ba" + "at" để tạo "bat") để hình thành câu hoàn chỉnh.

● Đặc trưng chính: Mô hình dựa trên các luật ngôn ngữ thống kê hoặc thiết kế thủ công.
● Ưu điểm: Có hiệu suất tính toán rất nhanh, độ trễ thấp, và chi phí tính toán thấp, dễ triển khai trên phần cứng cũ hoặc thiết bị nhúng. Công nghệ này cũng có khả năng xử lý đa ngôn ngữ cao do quy tắc được thiết kế độc lập theo từng ngôn ngữ.
● Nhược điểm: Độ tự nhiên thấp, chất lượng âm thanh thường khô, đều, thiếu cảm xúc. Công nghệ này khó mô phỏng đặc trưng giọng nói cá nhân và khó mở rộng để đạt chất lượng cao, vì các điểm nối giữa đơn vị âm có thể gây ra âm thanh gián đoạn.

● Ứng dụng phù hợp: Thiết bị nhúng, hệ thống có tài nguyên hạn chế, và các kịch bản chỉ cần tiếng nói hiệu quả mà không đòi hỏi tự nhiên cao, như hệ thống thông báo công cộng.

2.2. Level 2 - TTS học sâu cổ điển với cá nhân hóa bằng fine-tuning

Đây là thế hệ TTS sử dụng kiến trúc mạng nơ-ron sâu (Neural TTS - NTTS) như Tacotron 2 (Shen et al., 2018) – một mô hình autoregressive dự đoán spectrogram từ text qua encoder-decoder với attention; FastSpeech (Ren et al., 2019) – non-autoregressive để tăng tốc độ bằng cách dự đoán duration riêng; VITS (Kim et al., 2021) – kết hợp VAE và GAN cho end-to-end TTS; và các bộ mã hóa giọng nói (Vocoders) hiệu suất cao.

● Đặc trưng chính: Cốt lõi là kiến trúc NTTS (Autoregressive hoặc Non-autoregressive). Mỗi người dùng có một tập trọng số (weights) được fine-tune riêng.
● Ưu điểm: Đạt độ tự nhiên rất cao, gần đạt chất lượng tiếng nói người thật. Dễ dàng điều khiển ngữ điệu (prosody), cảm xúc, và tốc độ nói thông qua các biến tiềm ẩn (Latent Variables). Mô hình nhẹ hơn Cấp độ 3, thích hợp để triển khai trên môi trường có tài nguyên trung bình.
● Nhược điểm: Không phải Zero-shot. Mô hình phụ thuộc vào dữ liệu của người dùng, yêu cầu vài phút dữ liệu giọng nói chất lượng cao để thực hiện fine-tuning cho mỗi giọng mới.

● Ứng dụng phù hợp: Tạo giọng đọc sách nói, hệ thống thông báo tự động, và các hệ thống cần giọng cá nhân hóa nhưng chấp nhận quy trình thu âm và huấn luyện cá biệt.

2.3. Level 3 - Zero-shot / Few-shot TTS dựa trên mô hình ngôn ngữ âm thanh

Đây là xu hướng tiên tiến nhất, chuyển bài toán tổng hợp tiếng nói thành một bài toán dự đoán chuỗi token rời rạc, tương tự cách thức hoạt động của các Mô hình Ngôn ngữ Lớn (LLMs). Các mô hình tiên phong bao gồm VALL-E (Microsoft, 2023) – sử dụng neural codec để mã hóa audio thành tokens và LM để dự đoán; YourTTS (Casanova et al., 2022); và CosyVoice (Alibaba ModelScope, 2024) – hỗ trợ đa ngôn ngữ và cảm xúc.

● Kiến trúc Lõi: Mô hình sử dụng Audio Codec để chuyển đổi tín hiệu âm thanh liên tục thành các token rời rạc. Một Language Model được huấn luyện để dự đoán chuỗi token âm thanh đầu ra, được điều kiện hóa bởi văn bản đầu vào và một mẫu giọng nói ngắn (Voice Prompt).
● Ưu điểm: Khả năng Zero-shot / Few-shot Learning, cho phép mô phỏng giọng nói cá nhân ngay lập tức mà không cần training riêng. Giữ được đặc trưng giọng nói rất tốt, tính tự nhiên cao nhất hiện nay. Tiềm năng hỗ trợ mạnh mẽ đa ngôn ngữ nhờ pretraining trên tập dữ liệu quy mô lớn.
● Nhược điểm: Chi phí huấn luyện cực lớn (Foundation Models) với tập dữ liệu lên đến hàng nghìn giờ. Chi phí suy luận cao hơn đáng kể so với Cấp độ 1 và 2. Gây ra rủi ro Deepfake cao nếu không tích hợp cơ chế bảo vệ.

● Ứng dụng phù hợp: Dubbing tự động, hệ thống Hội thoại AI, sản xuất media chất lượng cao, các ứng dụng yêu cầu cá nhân hóa tức thời.

2.4. So sánh tổng hợp

Tiêu chí So Sánh	Level 1	Level 2	Level 3
Độ Tự Nhiên	Thấp (Robotic)	Cao (Near-Human Quality)	Rất Cao (Human-like)
Chi phí Tính toán (Suy luận)	Thấp (Rất nhanh)	Trung bình (Real-time trên GPU/CPU)	Cao (Độ trễ cao hơn Level 2)
Đa Ngôn Ngữ	Cao (Tách biệt ngôn ngữ)	Trung bình (Cần dữ liệu multi-speaker)	Rất Cao (Thông qua Pretraining lớn)
Cá Nhân Hóa Giọng	Không	Cần Fine-tuning (Vài phút data)	Zero-shot/Few-shot (Vài giây data)
Dữ liệu Yêu cầu	Không đáng kể	Dữ liệu Huấn luyện Lớn + Vài phút Fine-tuning	Dữ liệu Huấn luyện Rất Lớn + Vài giây Prompt
Khả năng Biểu cảm	Thấp	Cao (Có thể điều khiển)	Rất Cao (Mô phỏng từ Prompt)

3. Tình hình nghiên cứu

3.1. Xu hướng hiện tại

Nghiên cứu TTS hiện đại tập trung giải quyết các thách thức còn tồn đọng của Cấp độ 2 và 3, cụ thể là tối ưu hóa tốc độ, giảm chi phí tính toán và tăng cường tính biểu cảm.

● Tăng tính tự nhiên và biểu cảm: Phát triển các mô hình dựa trên mạng khuếch tán (Diffusion Models) như Grad-TTS và DiffSpeech để tạo ra phân phối âm thanh đầu ra chính xác hơn. Các nghiên cứu cũng tập trung vào Multi-Modal TTS (kết hợp video khuôn mặt) và kiểm soát Prosody bằng VAE hoặc Flow-based Models.
● Giảm chi phí tính toán: Áp dụng kiến trúc Non-Autoregressive Parallel TTS như FastSpeech 2 và các bộ mã hóa giọng nói hiệu quả (Lightweight Vocoder) như HiFi-GAN và UnivNet, cho phép tổng hợp theo thời gian thực.
● Hợp nhất các bài toán âm thanh: Xu hướng hợp nhất các nhiệm vụ Xử lý Tiếng nói (TTS, ASR, VC) vào một mô hình duy nhất (ví dụ: VITS, SpeechLM, CosyVoice 2.0) để học được các biểu diễn âm thanh đa chiều và linh hoạt.

● Zero-shot TTS hỗ trợ đa ngôn ngữ: Sử dụng tập dữ liệu đa ngôn ngữ quy mô lớn và kiến trúc LM-based để tạo ra các hệ thống TTS toàn cầu có khả năng tổng hợp giọng nói của một người bằng nhiều ngôn ngữ.

3.2. Tối ưu hóa Pipeline

Các công trình nghiên cứu đã tập trung giảm thiểu nhược điểm cố hữu của từng cấp độ công nghệ.

● Tối ưu hóa Level 1: Các hệ thống Hybrid TTS kết hợp phương pháp dựa trên luật với mô hình thống kê để giảm tính "robotic". Kỹ thuật Cost Minimization (Hunt & Black, 1996) được dùng để cải thiện nối ghép âm bằng cách lựa chọn đơn vị âm thanh tối ưu tại điểm nối.
● Tối ưu hóa Level 2: Sử dụng Speaker Encoder (d-vector, x-vector) để trích xuất đặc trưng giọng nói thành một vector embedding, cho phép Few-shot Fine-tuning với lượng dữ liệu người dùng tối thiểu. Triển khai Non-Autoregressive TTS cùng với Vocoder tốc độ cao (HiFi-GAN) để tăng tốc độ.

● Tối ưu hóa Level 3: Huấn luyện Language Model trên Discrete Audio Tokens thay vì Mel-spectrogram để bảo toàn đặc trưng giọng nói. Các kỹ thuật nén token codec và Streaming Decoding được nghiên cứu để giảm độ trễ và chi phí suy luận.

3.3. Các tiến bộ mới nhất năm 2025 (Bổ sung)

Năm 2025 chứng kiến sự ra mắt của nhiều mô hình TTS mã nguồn mở và thương mại tiên tiến, tập trung vào độ trễ thấp, đa ngôn ngữ và kiểm soát biểu cảm. Dưới đây là một số mô hình nổi bật:

- **Fish Speech V1.5**: Kiến trúc DualAR với hai transformer autoregressive, huấn luyện trên hơn 300.000 giờ dữ liệu tiếng Anh/Trung và 100.000 giờ tiếng Nhật. Ưu điểm: Đa ngôn ngữ mạnh mẽ, điểm ELO 1339 trên TTS Arena, tỷ lệ lỗi từ (WER) chỉ 3.5% cho tiếng Anh. Lý tưởng cho ứng dụng thời gian thực với chất lượng cao.
- **CosyVoice2-0.5B**: Mô hình streaming dựa trên LLM với quantization FSQ và chunk-aware matching, giảm độ trễ xuống 150ms. Hỗ trợ tiếng Trung (gồm phương ngữ), Anh, Nhật, Hàn và cross-lingual. Cải thiện MOS từ 5.4 lên 5.53, giảm lỗi phát âm 30-50%. Phù hợp cho hội thoại trực tiếp.

- **IndexTTS-2**: Mô hình zero-shot autoregressive với ba giai đoạn huấn luyện, sử dụng GPT latents và soft instruction từ Qwen3. Cho phép kiểm soát duration chính xác và tách biệt cảm xúc/timbre. Vượt trội về WER, speaker similarity và emotional fidelity so với các mô hình zero-shot khác.
- **Gemini 2.5 TTS (Google)**: Cập nhật cho mô hình Flash và Pro, cải thiện biểu cảm (tone đa dạng hơn, tuân thủ style prompts), pacing (tốc độ linh hoạt theo ngữ cảnh), và dialogue (giữ giọng nhân vật nhất quán). Hỗ trợ 24 ngôn ngữ, phù hợp cho audiobook và e-learning.

Những tiến bộ này giúp TTS gần hơn với giọng người thật, nhưng vẫn cần tối ưu hóa cho tài nguyên hạn chế.

4. Thách thức và đạo đức nghiên cứu

4.1. Các thách thức trong nghiên cứu TTS

Các thách thức chung mà cộng đồng nghiên cứu hướng đến: (1) Hiệu suất nhanh và độ trễ thấp như Level 1; (2) Tốn ít tài nguyên tính toán; (3) Tính tự nhiên cao như Level 2–3; (4) Hỗ trợ đa ngôn ngữ và chuyển giọng linh hoạt; (5) Thêm cảm xúc, prosody linh hoạt; và (6) Giảm công sức người dùng (chỉ cần vài giây audio). Thêm vào đó, xử lý dữ liệu lớn mà không vi phạm quyền riêng tư là vấn đề lớn.

4.2. Khía cạnh đạo đức và rủi ro Deepfake

Sự ra đời của Zero-shot TTS Cấp độ 3 đã làm gia tăng đáng kể các mối đe dọa về Deepfake và giả mạo danh tính giọng nói. Các nhà nghiên cứu đang xây dựng các cơ chế bảo vệ:

● Watermark và phát hiện giả mạo: Nhúng một tín hiệu nhận dạng (Watermark) không nghe được vào tín hiệu tổng hợp để phân biệt âm thanh thật và âm thanh giả, đồng thời truy xuất nguồn gốc của nội dung.
● Bảo vệ dữ liệu giọng nói: Thiết lập các chính sách Ẩn danh Dữ liệu (Anonymization) và quy tắc sử dụng giọng nói rõ ràng.

● Điều khiển quyền truy cập: Yêu cầu các quy trình Xác thực Giọng nói (Voice Biometric Authentication) và hạn chế cung cấp API Zero-shot Voice Cloning đại trà.

5. Kết luận

Quá trình phát triển của công nghệ Text-to-Speech phản ánh sự tiến hóa từ các phương pháp dựa trên luật, qua giai đoạn học sâu với pipeline tối ưu, đến thế hệ mô hình cơ sở hiện đại. Mỗi cấp độ công nghệ mang theo những ưu điểm và hạn chế riêng, đồng thời phục vụ cho các mục tiêu ứng dụng khác nhau:

● Level 1: tối ưu tốc độ và chi phí, phù hợp với hệ thống nhúng và đa ngôn ngữ nhưng thiếu tự nhiên.

● Level 2: cân bằng giữa độ tự nhiên, khả năng cá nhân hóa và chi phí tính toán, thích hợp cho audiobook, trợ lý ảo và các ứng dụng cần giọng nói giàu cảm xúc.

● Level 3: mở ra kỷ nguyên zero-shot/few-shot TTS, đạt độ tự nhiên cao nhất, hỗ trợ đa ngôn ngữ và cá nhân hóa tức thời, song đi kèm thách thức về chi phí huấn luyện và rủi ro đạo đức.

Xu hướng nghiên cứu hiện nay tập trung vào việc kết hợp ưu điểm của cả ba cấp độ, nhằm xây dựng hệ thống TTS vừa nhanh và nhẹ như các phương pháp ở Level 1, vừa đạt độ tự nhiên và giàu cảm xúc như các mô hình học sâu ở Level 2–3, đồng thời có khả năng đa ngôn ngữ và cá nhân hóa tức thời như các mô hình cơ sở (Foundation Models). Bên cạnh đó, các nghiên cứu cũng nhấn mạnh yêu cầu đảm bảo an toàn và đạo đức trong sử dụng, đặc biệt trước nguy cơ giả mạo giọng nói và deepfake.

Có thể khẳng định rằng, TTS đang trở thành một thành phần cốt lõi trong hệ sinh thái trí tuệ nhân tạo, với tiềm năng ứng dụng rộng rãi từ giáo dục, y tế, truyền thông, giải trí cho đến giao tiếp người–máy. Tương lai của TTS sẽ là sự hội tụ giữa tính hiệu quả, độ tự nhiên, khả năng mở rộng và trách nhiệm xã hội, đưa công nghệ giọng nói nhân tạo trở thành cầu nối quan trọng giữa con người và máy tính.

6. Tài liệu tham khảo

1. Hunt, A. J., & Black, A. W. (1996). Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. ICASSP.
2. Tokuda, K. et al. (2000). HMM-based Speech Synthesis System (HTS).
3. Wang, Y. et al. (2017). Tacotron: Towards End-to-End Speech Synthesis. Interspeech.
4. Shen, J. et al. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. ICASSP.
5. Ren, Y. et al. (2019). FastSpeech: Fast, Robust and Controllable Text-to-Speech. NeurIPS.
6. Kim, J. et al. (2021). VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. ICML.
7. Popov, V. et al. (2021). Grad-TTS: A Diffusion Probabilistic Model for TTS. ICML.
8. Casanova, E. et al. (2022). YourTTS: Towards Zero-shot Multi-speaker TTS. ICASSP.
9. Microsoft Research (2023). VALL-E: Neural Codec Language Models are Zero-Shot TTS Synthesizers.
10. Alibaba ModelScope (2024). CosyVoice: A Comprehensive Speech Generation Framework.
11. SiliconFlow (2025). Ultimate Guide - The Best Open Source Text-to-Speech Models in 2025.
12. Google Blog (2025). Improving Gemini Text-to-Speech models for better control and capabilities.