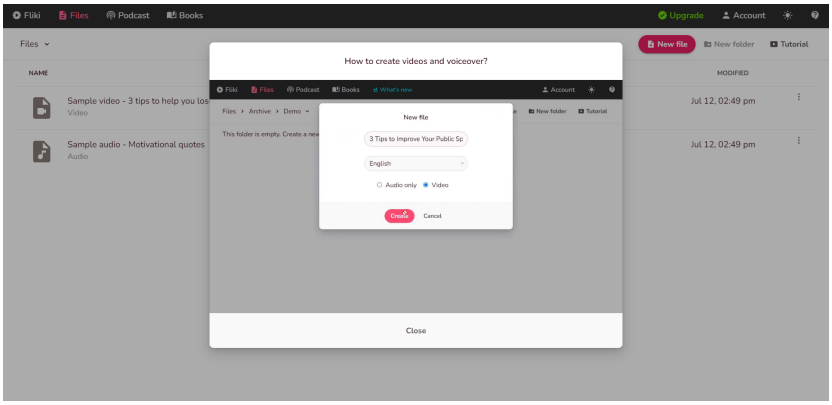
Fliki

1 Tổng quan

Fliki là một công cụ chuyển đổi văn bản thành giọng nói AI giúp người dùng có thể chuyển đổi bất kỳ dạng văn bản nào thành giọng nói ảo. Nhờ có Fliki ,ta có thể tạo video một cách nhanh chóng không cần thu giọng nói thủ công cho video. Giọng nói của Fliki cũng rất tự nhiên, ở phiên bản Tiếng Anh có trọng âm nghe rất bắt tai.

2 Giao diện và Đánh giá về Fliki

Giao diện đơn giản, thân thiện với người dùng.(có thể sử dụng những công cụ cơ bản của lập trình web như expressjs để làm giao diện này)(2)



Fliki cho phép tạo nhiều scene trong video . Mỗi Scence đó ta có thể chọn voiceover , đoạn văn bản ta muốn phát trong scene đó, ngoài ra ta cũng có thể điều chỉnh mức độ Tune (Rate và Pitch) cho các từ trong văn bản. Hỗ trợ tạo nhạc nền và có thể custom thời gian scene.

Ưu điểm:

- Dễ sử dụng: Giao diện của Fliki cho phép bất kỳ ai chuyển đổi văn bản thành video
- Lựa chọn giọng nói rộng: Fliki cung cấp hơn nhiều ngôn ngữ và phương ngữ. Âm thanh rất độc đáo và chân thực đối với các giọng nói thông dụng trên thế giới như tiếng anh bởi lượng data dồi dào.
- API của Fliki cho phép tích hợp với nhiều công cụ và nền tảng.

Nhược điểm:

- Đối với các giọng nói ít dữ liệu hơn như tiếng việt thì cách nói chuyện còn khá cứng, nói rất đều như chỉ ghép các từ vào. Thiếu sự hấp dẫn với người xem. (trong bản chưa upgrade)
- Fliki hỗ trợ nhiều giọng nói và ngôn ngữ, nhưng việc cá nhân hóa bị hạn chế(chỉ có trong bản premium vì thế mà em chưa test được). Điều chỉnh cách nói của giọng nói bị giới hạn.(mới chỉ cho điều chỉnh tune cao hay thấp thủ công và tốc độ nói).
- Chỉnh sửa video của Fliki còn tương đối cơ bản, do đó các biên tập viên chuyên nghiệp có thể cần sử dụng các ứng dụng khác.

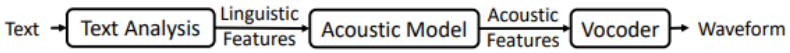
- Các gói của Fliki hạn chế thời lượng video, do đó khách hàng muốn tạo ra các bộ phim lớn hơn có thể cần nâng cấp hoặc sử dụng phần mềm khác.

3 Công nghệ Text to speech - TTS

Công nghệ chính của Fliki chính là áp dụng AI cho bài toán Text-to-speech(TTS).

3.1 Các thành phần chính và cơ chế xử lý của Fliki

Một hệ thống TTS hiện đại bao gồm ba thành phần cơ bản: một module phân tích văn bản, mô hình âm thanh (acoustic model) và bộ phát âm (vocoder) (Hình 3.1)



- Module phân tích văn bản chuyển đổi một chuỗi văn bản thành các đặc trưng ngôn ngữ
- Các mô hình âm thanh tạo ra các đặc trưng âm thanh từ đặc trưng ngôn ngữ
- Bộ phát âm tổng hợp và tạo ra dạng sóng từ các đặc trưng âm thanh

Cơ chế chuyển đổi từ văn bản sang giọng nói của Fliki gồm các bước như sau :

- Tách văn bản thành các kí tự đầu vào mang đặc trưng của ngôn ngữ đó (ví dụ như tiếng việt sẽ có các kí tự hay cách ghép từ riêng biệt thì ở bước này nó sẽ trích xuất được những thông tin đó)
- Biến đổi từ những đặc trưng đó sang thành các đặc trưng âm thanh (ở bước này sẽ biến các đặc trưng ngôn ngữ thành các spectrogram - cũng có thể coi là các âm vị mang đặc trưng của người nói-dữ liệu train)
- Sau đó sẽ chuyển đổi các spectrogram thành waveform. Dạng sóng này chính là cách mã hóa âm thanh trong thực tế của máy tính .

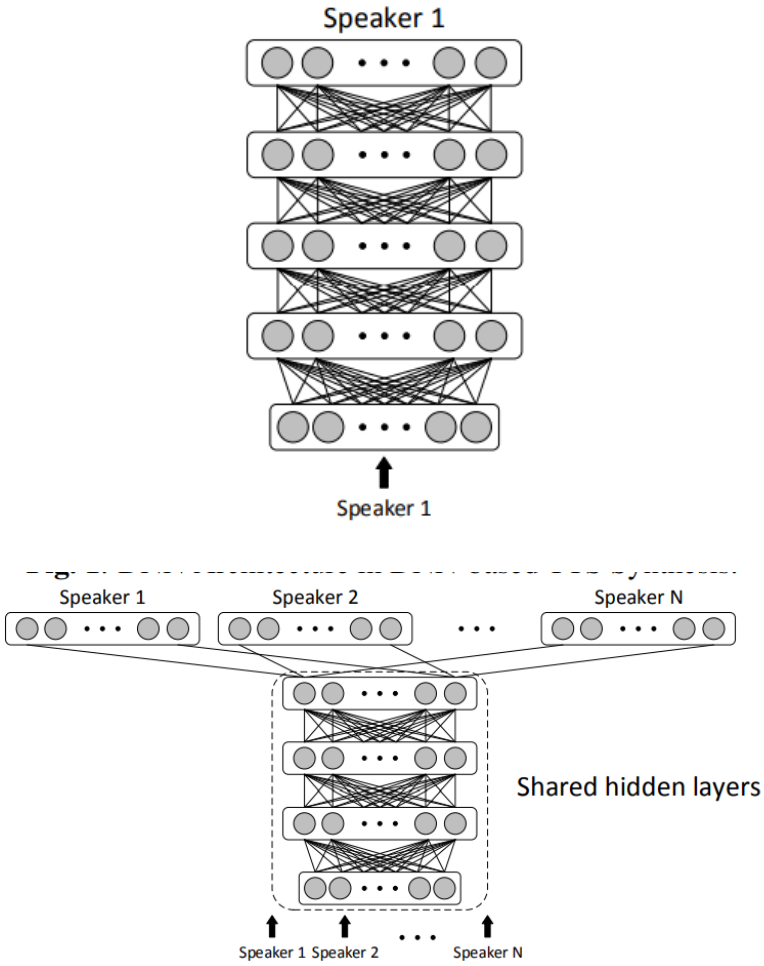
3.2 Áp dụng DeepLearning cho bài toán TTS

Từ khi deep learning phát triển, chúng đã được áp dụng rộng rãi trong các lĩnh vực .Công nghệ Text-To-Speech cũng vậy. Các model được phát triển để áp dụng giải quyết từng thành phần : Acoustic Model, Vocoder và cả những model end-to-end cũng được phát triển. Dưới đây, e trình bày model đơn giản nhất được sử dụng để thay thế Acoustic Model truyền thống.

Với model được đào tạo cho Single-speaker : Model sẽ lấy các đặc điểm ngôn ngữ được chuyển đổi làm inputs và tính năng âm học cho outputs. Sau đó học cách ánh xạ giữa ngôn ngữ và không gian âm học. Model cho mỗi speaker sẽ được huấn luyện riêng với giọng nói của chính họ. 3.2

Với model cho Multi-speaker: các lớp ẩn được chia sẻ qua tất cả các người nói trong tập huấn luyện và có thể coi là quá trình biến đổi đặc trưng ngôn ngữ toàn cục được chia sẻ bởi tất cả các người nói. Ngược lại, mỗi người nói có lớp đầu ra riêng của mình, được gọi là lớp hồi quy, để mô hình không gian âm học cụ thể của mình. So với mạng single-speaker truyền thống, mạng nơ-ron đa người nói sử dụng cùng một đặc trưng ngôn ngữ đầu vào, được chuyển đổi từ văn bản theo cùng một cách thức, và cùng một đặc trưng âm học đầu ra cho mỗi người nói.(Hình 3.2)

Hiện nay, các model Text-to-speech đã được phát triển rất nhiều có thể kể đến như : các acoustic model Tacotron 2, DeepVoice 3, TransformerTTS, FastSpeech , ...; các vocoder model WaveNet, WaveGAN,WaveFlow ,.. đã được áp dụng rất rộng rãi và cho perform tốt. Các kiến trúc end-to-end như Char2Wav, FastSpeech 2s, Wave-Tacotron,...



4 Các phương án xây hệ thống

Ở đây em chỉ đề cập đến các phương pháp giải quyết bài toán chính text to speech và sẽ không đề cập đến giao diện

4.1 Sử dụng API

4.1.1 Sử dụng API của Google Cloud

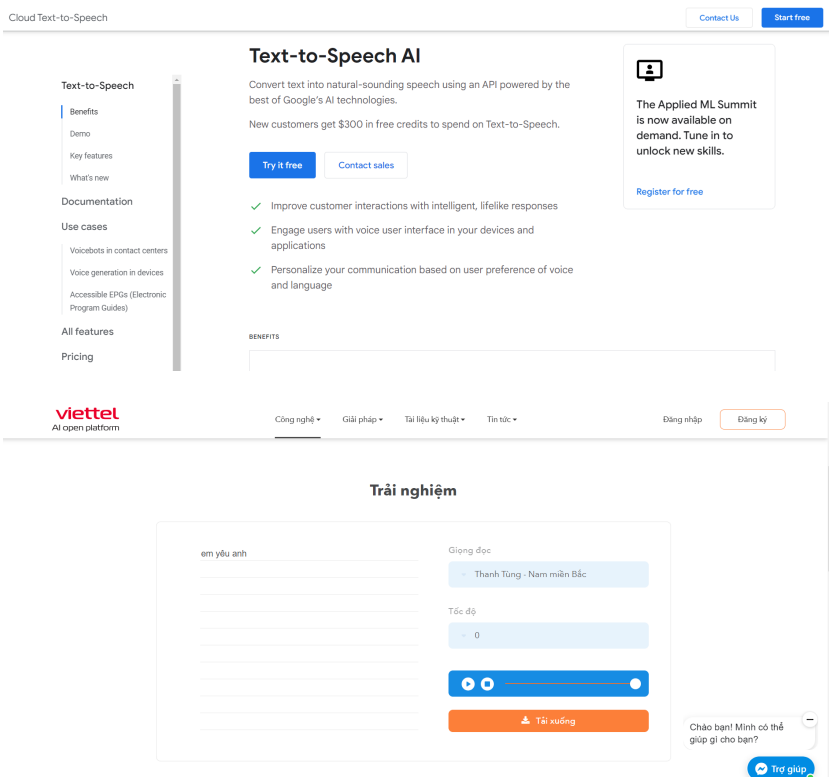
Google Cloud cung cấp dịch vụ TTS với 300$ per month. Chất lượng giọng nói Tiếng Anh thì khá tốt, đa dạng, ngữ điệu hợp lí trong các bối cảnh khác nhau. Nhưng bản Tiếng Việt thì cung cấp khá ít chất giọng, giọng nói đều đều thiếu tự nhiên. (4.1.1)

4.1.2 Sử dụng API của Viettel

Chất lượng giọng nói tốt, có ngữ điệu nhưng hầu hết là phù hợp cho tổng đài hoặc thuyết minh. Giá thì em không thấy đề cập đến. (4.1.2)

4.1.3 API của Fliki

Cũng như GG Cloud, giọng đọc của Fliki khá hay và nhiều ngữ điệu nhất là cho các Youtuber nhưng Tiếng Việt thì vẫn còn yếu. Điểm khác biệt là bản premium của FLiki chỉ mất 66$ per month

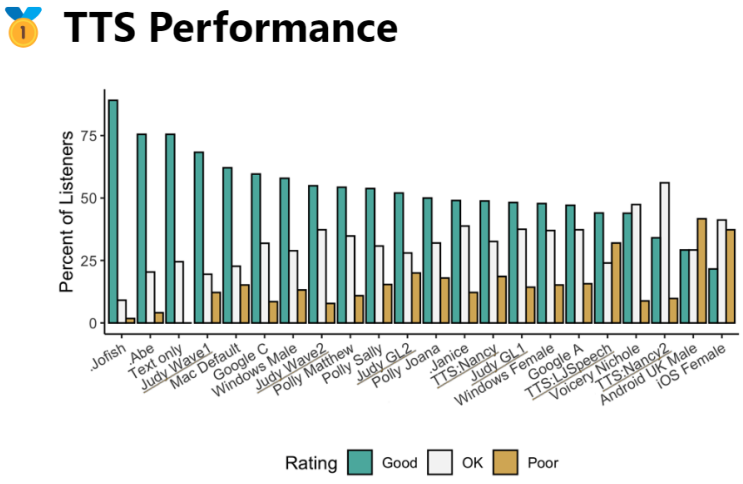


4.2 Deploy model TTS của riêng mình

4.2.1 Coqui TTS

Coqui TTS là một thư viện cho việc tạo ra giọng nói từ văn bản tiên tiến. Nó được xây dựng dựa trên các nghiên cứu mới nhất và được thiết kế để đạt được sự cân bằng tốt nhất giữa việc dễ huấn luyện, tốc độ và chất lượng. TTS đi kèm với các mô hình được huấn luyện trước, các công cụ để đo lường chất lượng tập dữ liệu và đã được sử dụng trong hơn 20 ngôn ngữ cho các sản phẩm và dự án nghiên cứu.

Hình 4.2.1 là đánh giá của người nghe



Thư viện cũng hỗ trợ đầy đủ các loại model cho tất cả các tác vụ từ Spectrogram models, Vocoders hay đến cả End-to-End Models.

Thư viện này là open-source nên ta có thể dễ dàng cải tiến, fine-tune cho model

4.3 Torchaudio và TensorflowTTS

Cả pytorch và tensorflow đã có những hỗ trợ nhất định cho các vấn đề về Speech. Tuy nhiên, PyTorch không cung cấp trực tiếp các mô hình hoàn chỉnh cho bài toán Text-to-Speech (TTS). Và TensorflowTTS không thấy cập nhật mạnh mẽ trên github như Coqui-tts

4.4 VietTTS

Là một mã nguồn mở và áp dụng Non-Attentive Tacotron (NAT) + HiFiGAN vocoder cho dữ liệu tiếng việt.

5 So Sánh

5.1 Coqui-TTS

- Là một mã nguồn mở cung cấp các model mới nhất của Text-To-Speech,(hỗ trợ cả config model cho developer).Được cập nhật liên tục với các paper mới nhất.
- Nếu muốn sử dụng mà không can thiệp code, coqui-tts cũng cung cấp các api như vậy.
- Cung cấp một số pretrain model và một vài tập dataset public như English - LJ Speech, English - Nancy , English - VCTK, Multilingual - M-AI-Labs, Spanish - thx! @carlfm01, German - Thorsten OGVD, Japanese - Kokoro, Chinese, Ukrainian - LADA.
- Một yếu điểm là nó không có dataset có sẵn hay pretrain có sẵn về Tiếng việt nhưng ta có thể tìm các nguồn dataset tiếng việt public để huấn luyện

5.2 VietTTS

- Là một mã nguồn mở cho Text-to-speech nhưng mới chỉ áp dụng sẵn Non-Attentive Tacotron (NAT) + HiFiGAN vocoder cho dữ liệu tiếng việt
- Ta có thể từ dữ liệu đó, tạo thêm các model và phát triển bộ mã nguồn này
- Viet-tts có lẽ đã không còn cập nhật từ 2 năm trước

5.3 ViettelTTS

- Chỉ cung cấp các API để người dùng có thể chuyển văn bản thành giọng nói theo ý muốn. Người dùng sẽ không thể can thiệp vào các model mà Viettel sử dụng
- Chất lượng giọng nói tốt, có ngữ điệu nhưng hầu hết các giọng nói cung cấp là phù hợp cho tổng đài hoặc thuyết minh.
- Có thể các model của vietteltts cũng được xây dựng từ thư viện coquiTTS và với dữ liệu tiếng việt của họ.