

# New Era of Speech Generation

---

BETTER THAN HUMAN

# Nội Dung

## 1. Những hướng tiếp cận trước đây dựa trên Deep Learning

1. Autoregressive Text To Speech Model
2. Non-AutoRegressive Text to Speech Model
3. Neural Vocoder.

## 2. Những hướng tiếp cận mới

1. Hướng tiếp cận dựa trên Diffusion và Flow Generative Models
2. Sử dụng Neural Codecs thay thế Vocoder
3. Hướng tiếp cận dựa trên Large Language Model Based Speech Generation





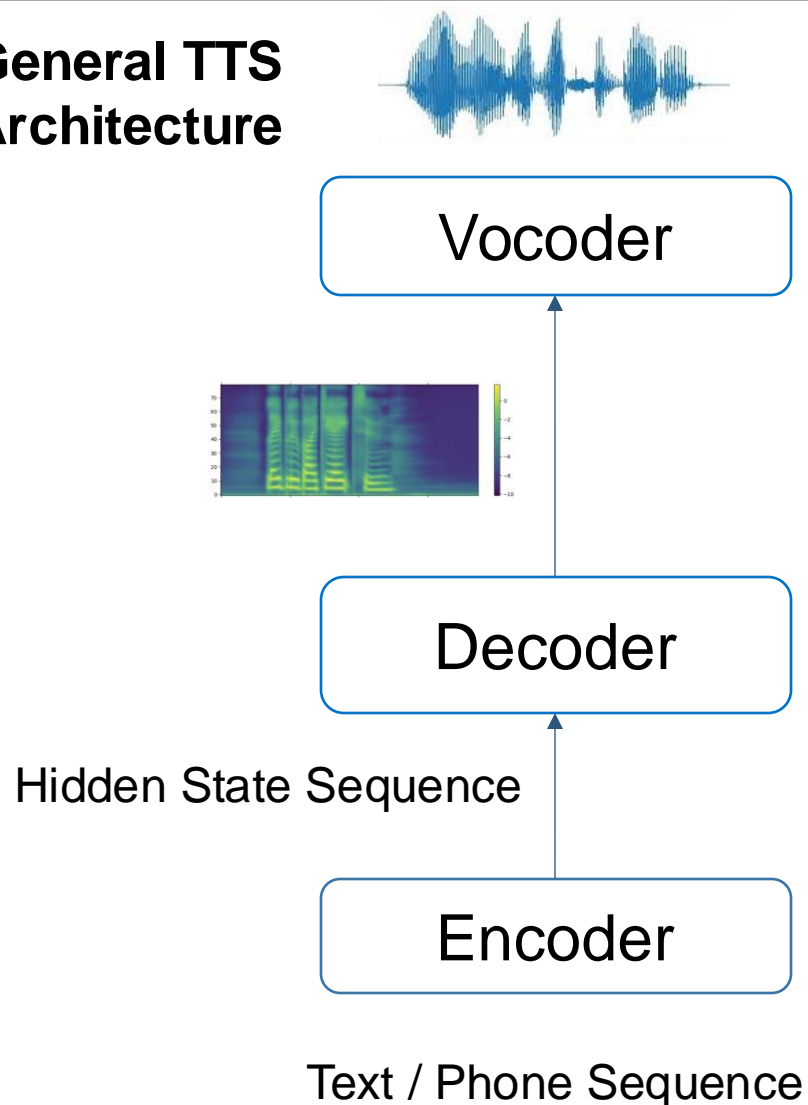
# Những Hướng Tiếp Cận Cổ Dựa Trên DeepLearning

---

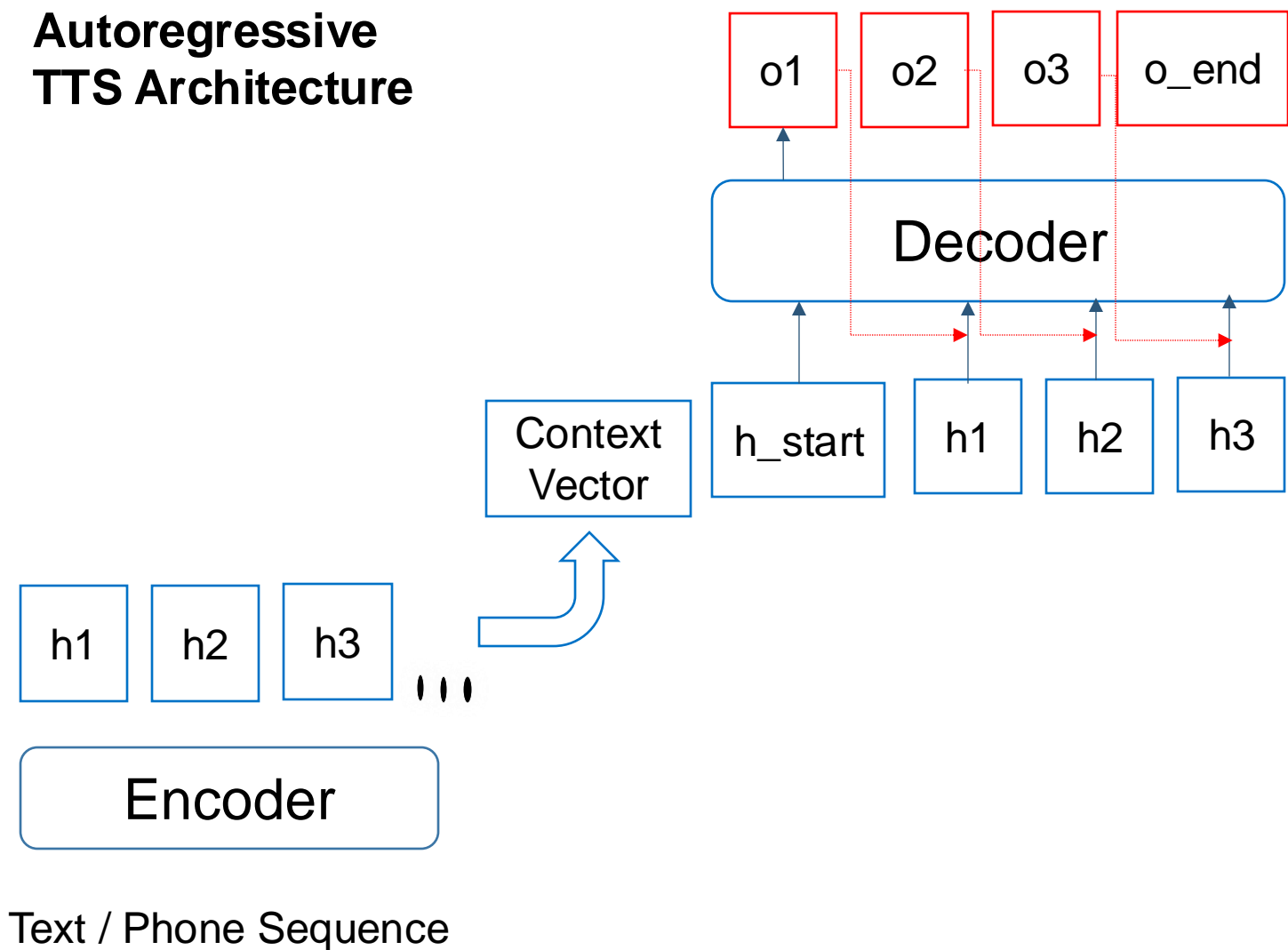
Thịnh Nguyễn/SpeechWorld

# Autoregressive TTS (Tacotron2, VITS)

## General TTS Architecture

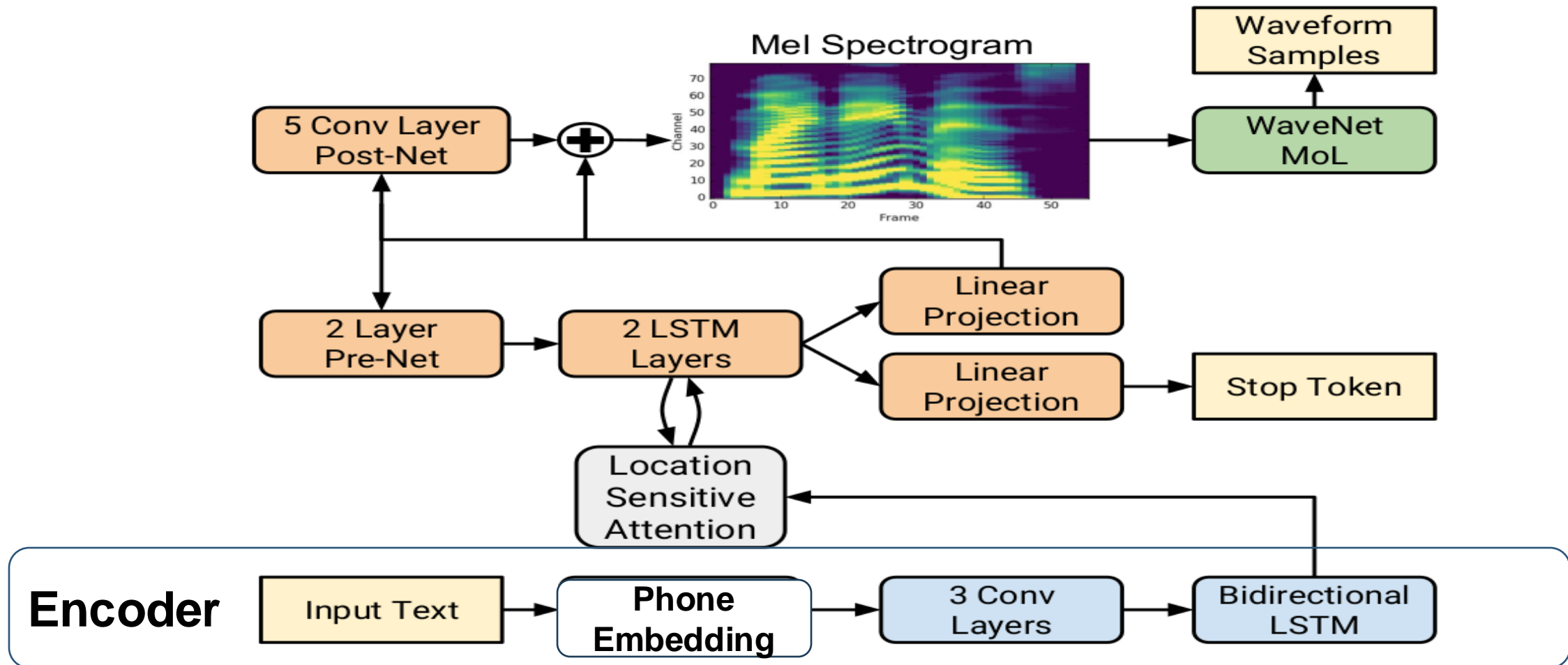


## Autoregressive TTS Architecture



# Autoregressive TTS (Tacotron2, VITS)

## Tacotron2





# Autoregressive TTS (Tacotron2, VITS)

## ***Ưu và Nhược Điểm***

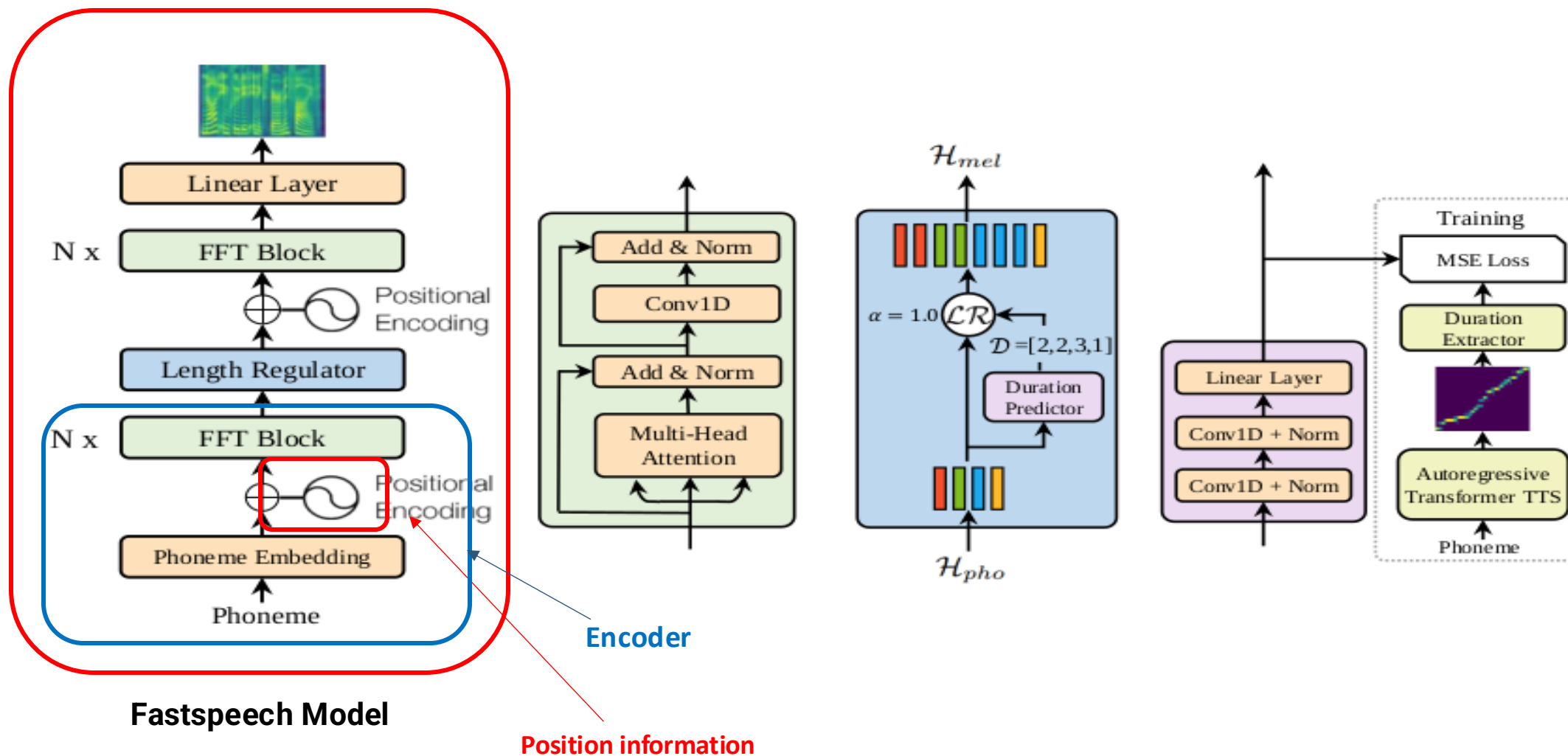
### Ưu Điểm

1. Chất lượng tiếng nói cao, tự nhiên.
2. Đôi khi không phân biệt được với người thật.

### Nhược Điểm

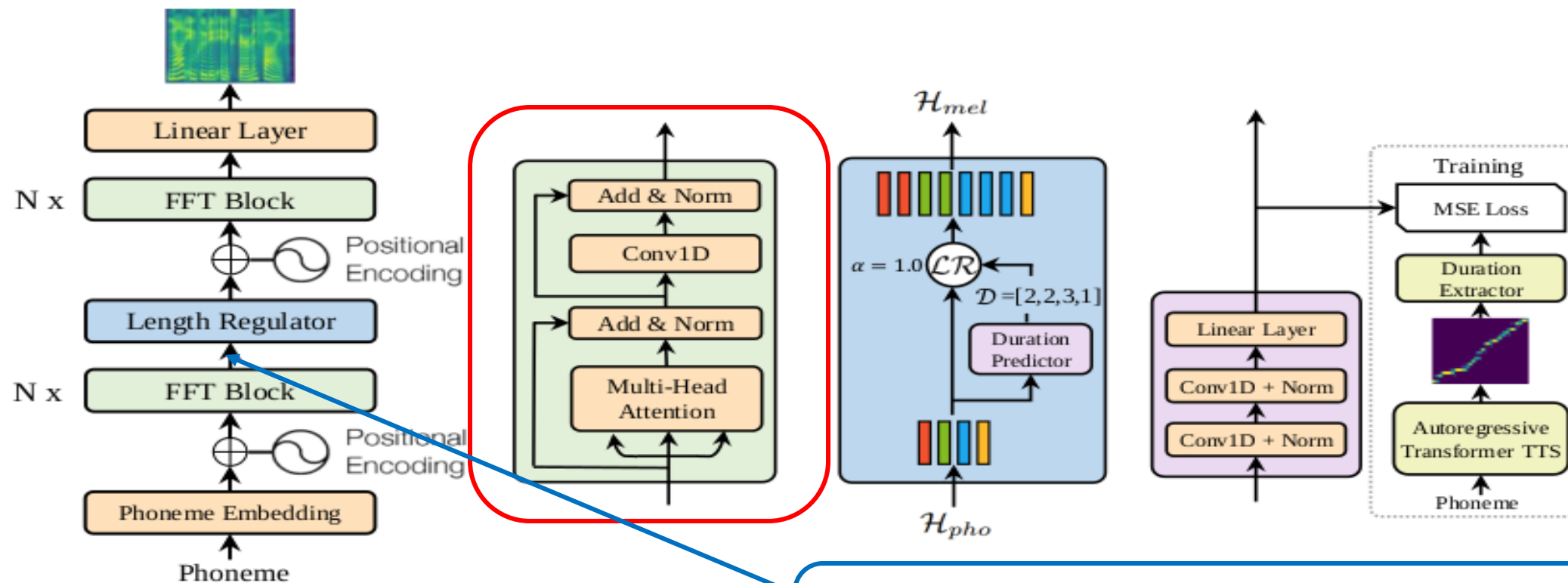
1. Tốc độ xử lý chậm do phải đợi các kết quả của step trước để predict step hiện tại
2. Đôi khi còn hiện tượng đọc sai, lặp từ, ngắt nghỉ lung tung do Alignment học không chuẩn.

# Non-Autoregressive TTS (FastSpeech / FastSpeech2)





# Non-Autoregressive TTS (FastSpeech / FastSpeech2)



FastSpeech Model

FFT Block

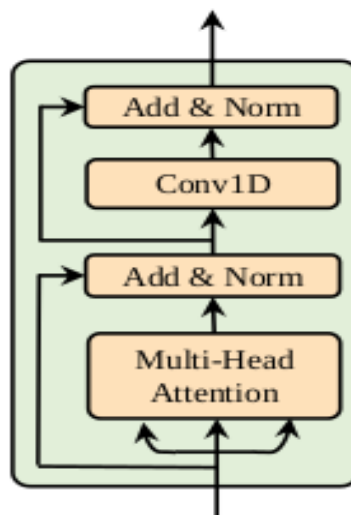
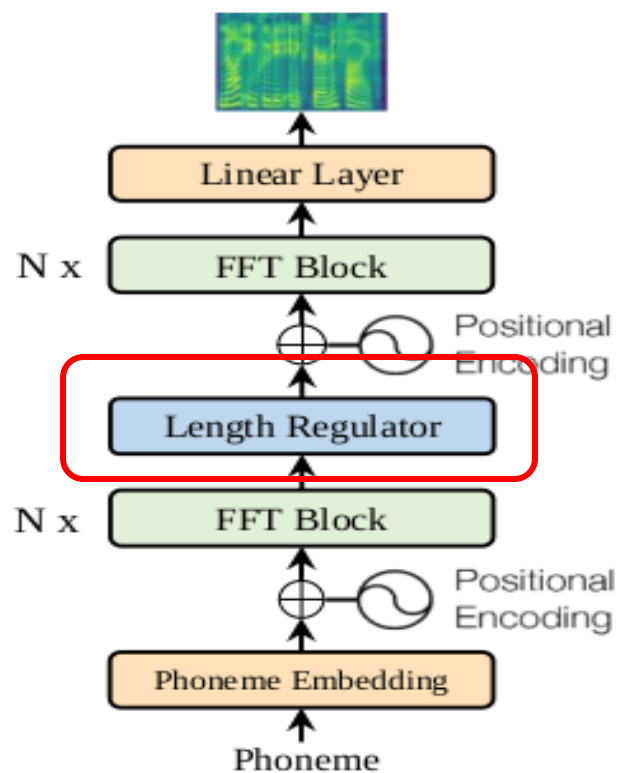
Information:

1. Meaning of source sequence.
2. source sequence structure.

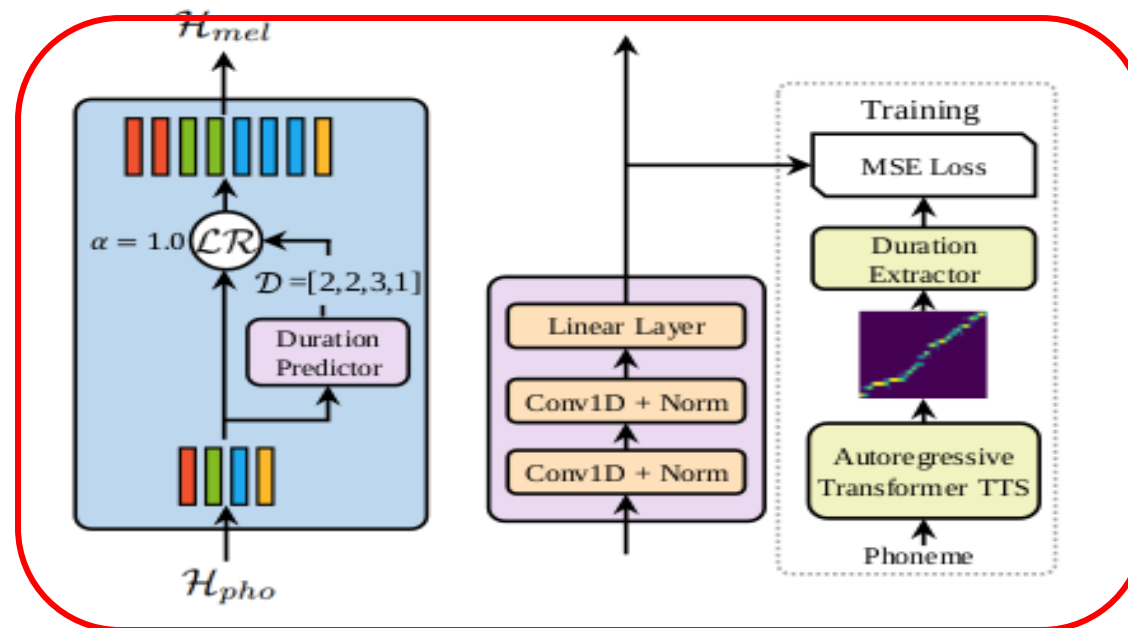


# Non-Autoregressive TTS (FastSpeech / FastSpeech2)

## FastSpeech: Length Regulator



### Key Improvement

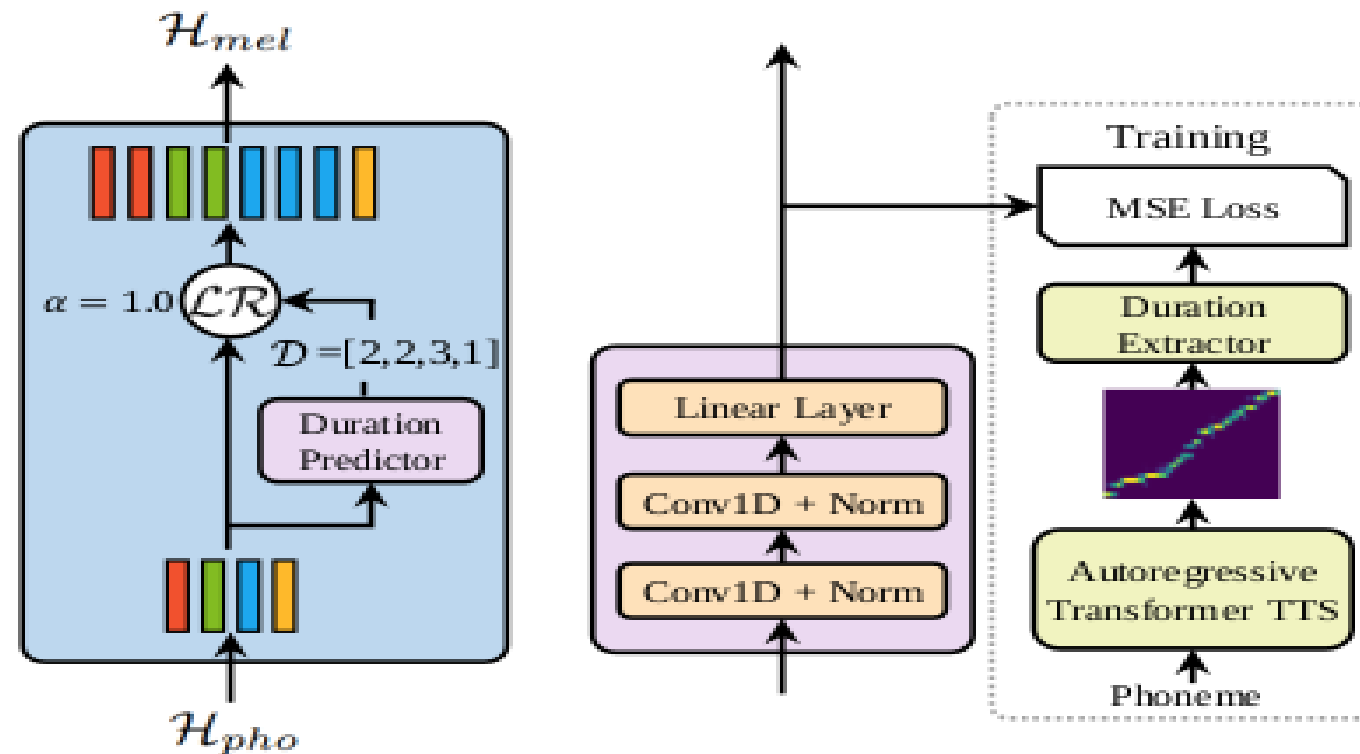


# Non-Autoregressive TTS (FastSpeech / FastSpeech2)

## FastSpeech: Length Regulator

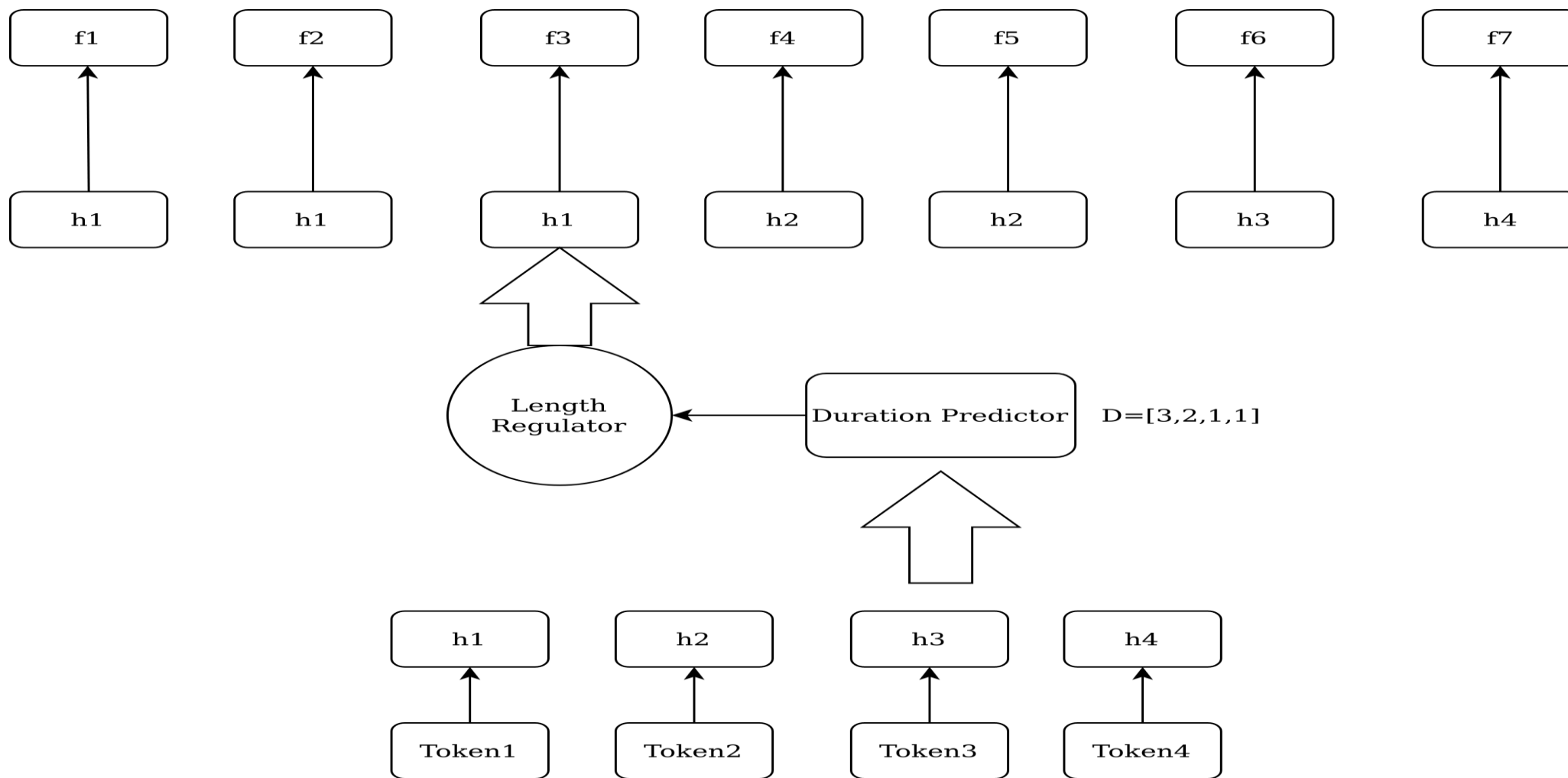
### Duration Predictor:

1. Sử dụng Forced Alignment Tools (như là Kaldi, Montreal Forced Aligner) để xác định duration từng phone
2. Tính số frame audio/mel tương ứng mỗi phone.
3. Duration predictor được train để dự đoán số frame tương ứng với phone hiện tại, sử dụng MSE Loss.



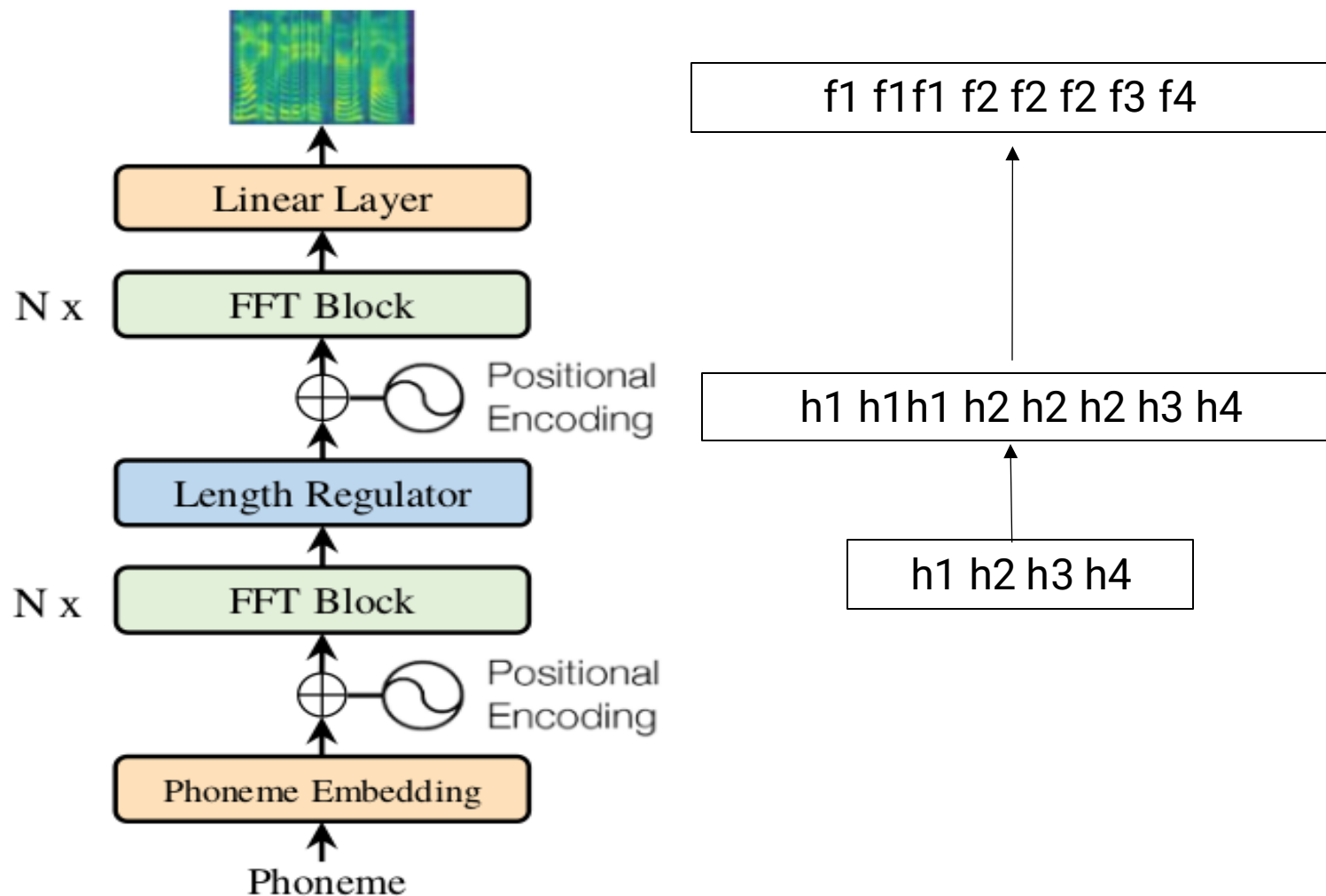
# Non-Autoregressive TTS (FastSpeech / FastSpeech2)

## *FastSpeech: Length Regulator*



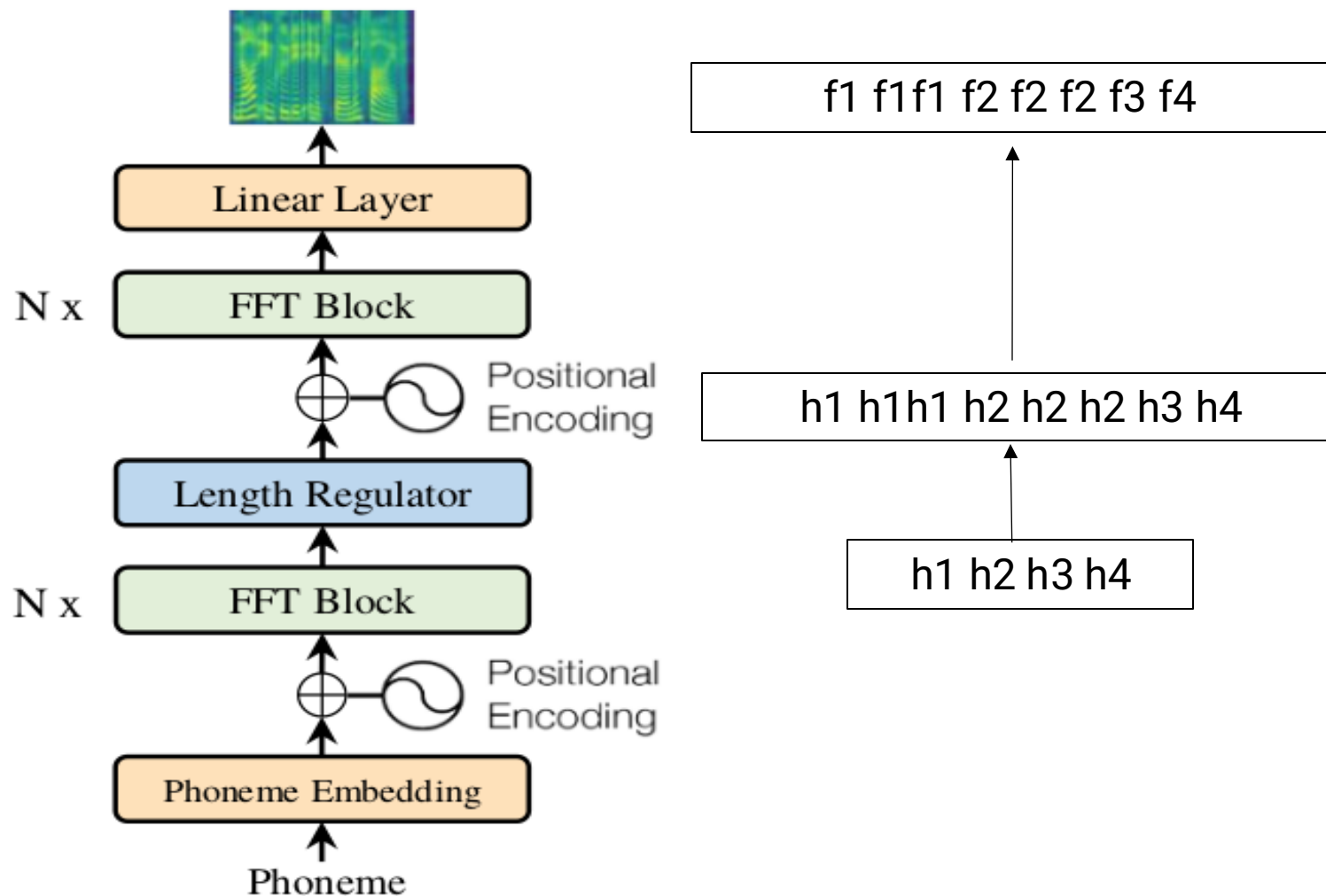
# Non-Autoregressive TTS (FastSpeech / FastSpeech2)

## *FastSpeech: Length Regulator*



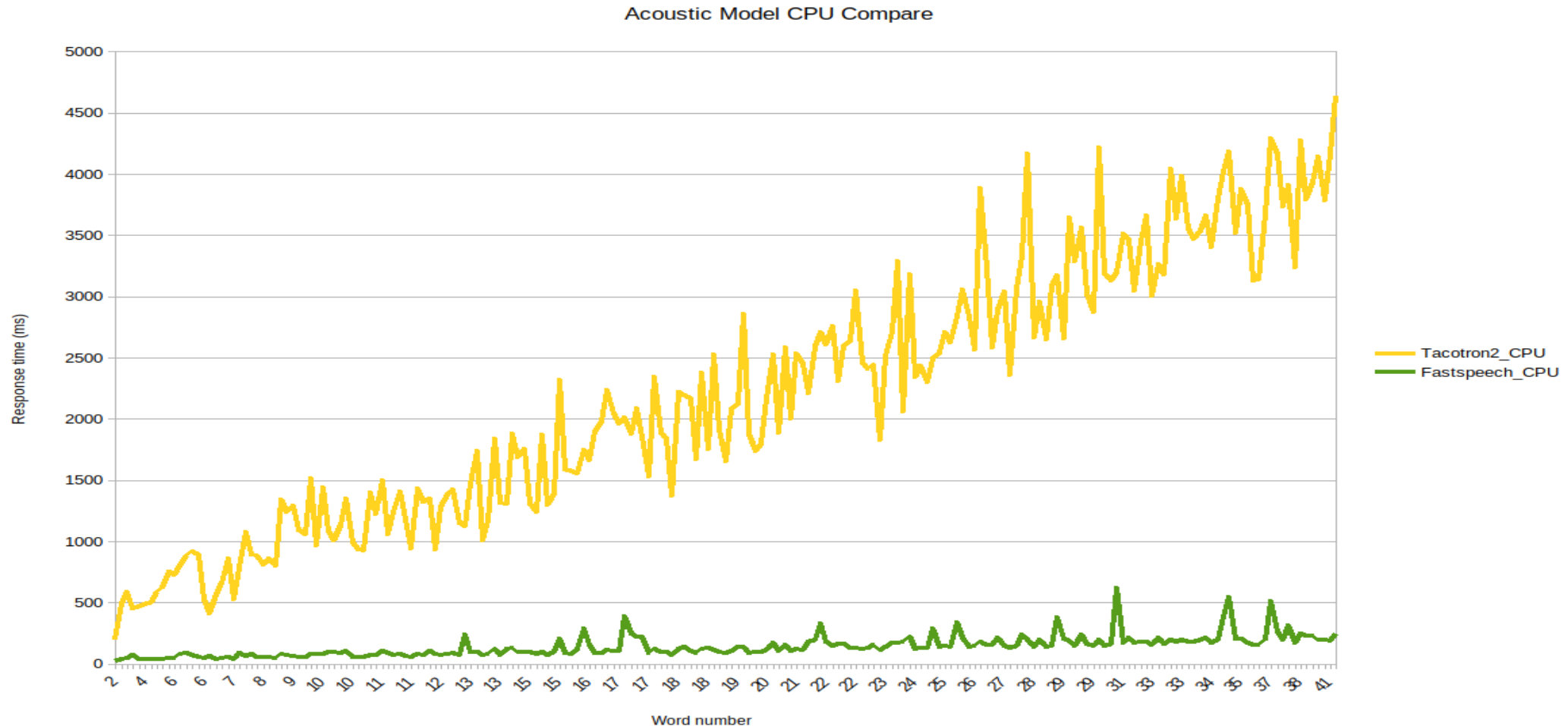
# Non-Autoregressive TTS (FastSpeech / FastSpeech2)

## *FastSpeech: Length Regulator*



# Non-Autoregressive TTS (Fastspeech / Fastspeech2)

## *Performance Comparision of Tacotron2 vs Fastspeech*





# Non-Autoregressive TTS (Fastspeech / Fastspeech2)

## ***Ưu nhược điểm của model Fastspeech / Fastspeech2***

### Ưu Điểm

1. Tốc độ xử lý nhanh do Non-Autoregressive
2. Độ chính xác phát âm rất cao do được học từ một tools Forced Alignment độc lập

### Nhược Điểm

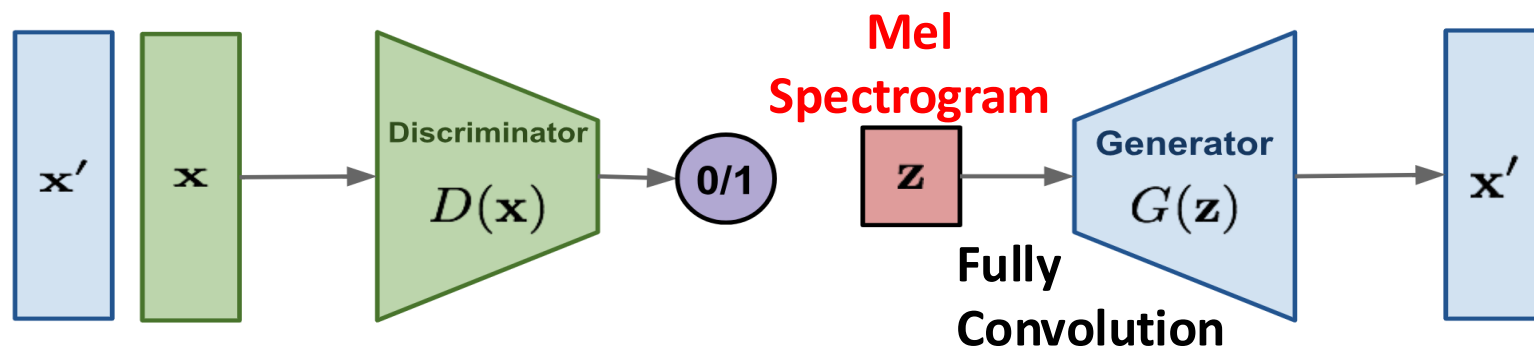
1. Độ tự nhiên không hay bằng Autoregressive model, đôi khi hơi đều đều.



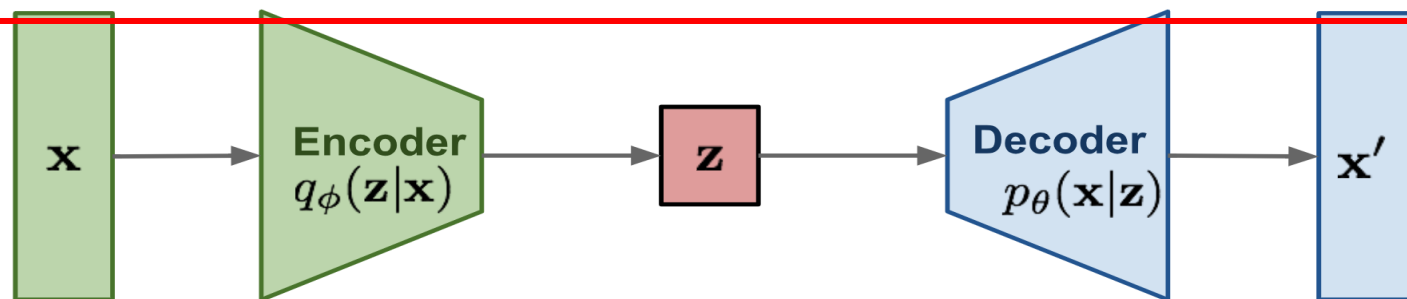
# Neural Vocoders

## Neural Vocoder: Flow based mode to Gan based Vocoder

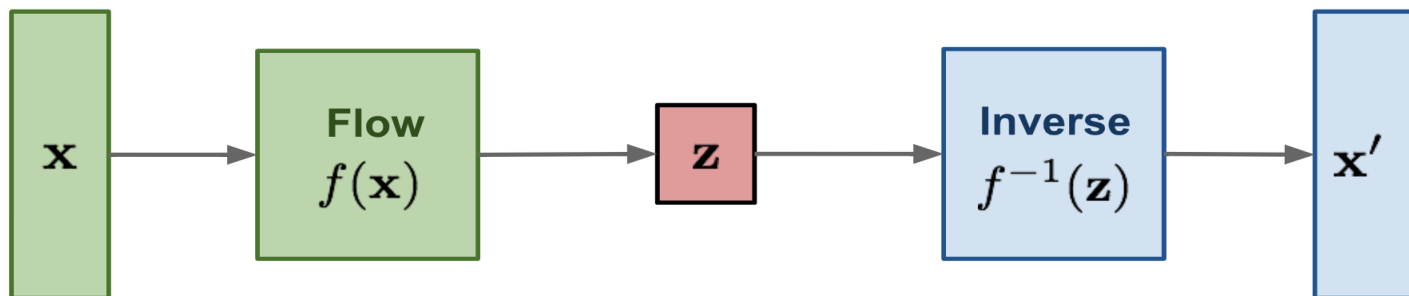
**GAN:** minimax the classification error loss.



**VAE:** maximize ELBO.



**Flow-based generative models:** minimize the negative log-likelihood

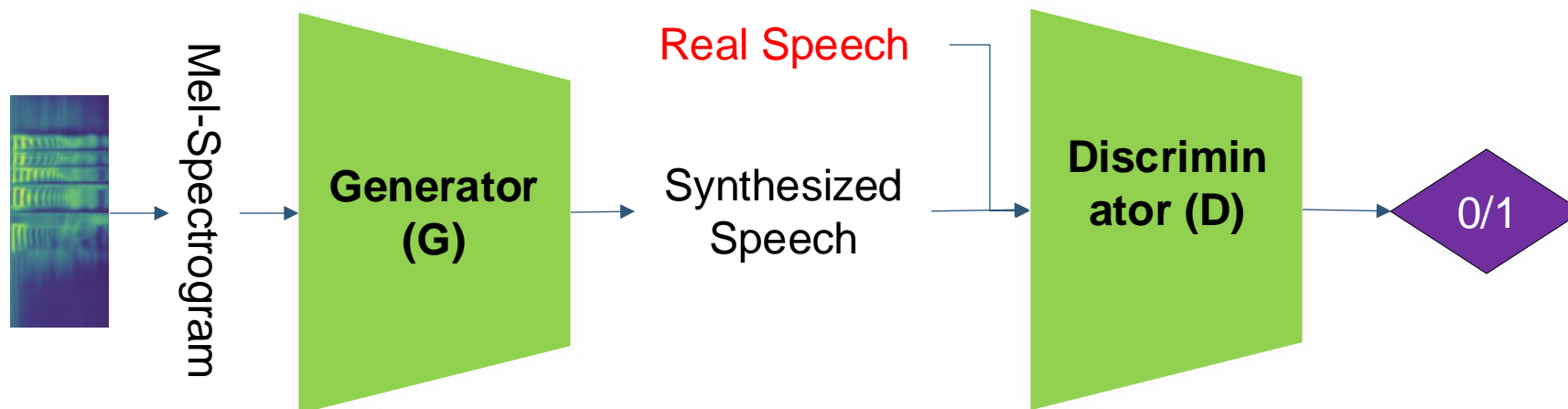


# Neural Vocoder

## *Gan Vocoder*




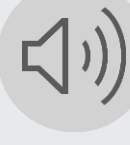

### Ưu Điểm

1. Tốc độ xử lý nhanh do chỉ cần sử dụng Generator khi inference.
2. Chất lượng âm thanh rất cao, đặc biệt với các model sau này như Hifi-gan



# Non-Autoregressive TTS (Fastspeech / Fastspeech2)

## *Đánh giá chất lượng model Fastspeech*

Text	Fastspeech + MB Melgan	Tacotron2 + Waveglow
Đoàn đã được tổng bí thư nông đức mạnh , chủ tịch nước trần đức lương , bộ trưởng quốc phòng phạm văn trà tiếp thân mật .		
Việc điều chỉnh tăng mức lương tối thiểu tập trung chủ yếu vào các doanh nghiệp thuộc ngành dệt may , giày da , chế biến gỗ .		
Hai nhà lãnh đạo đánh giá cao những bước phát triển tốt đẹp trong quan hệ hữu nghị và hợp tác việt nam na uy .		



# New Advanced Speech Generation Technique

---

Thịnh Nguyễn/SpeechWorld



# Generative Speech Generation based on Diffusion / Flow Matching

---

Thịnh Nguyễn/TTS



# Neural Codec for Speech Compression

---

Thịnh Nguyễn/TTS



# Neural Codec

1. Được phát triển để nén audio và có thể dùng thay thế Vocoder
2. Giảm dung lượng lưu trữ audio do đã được nén.
3. Giữ được nhiều thông tin về Acoustic hơn Mel-Spectrogram

## ***Residual Vector Quantization***

Các Neural Codec model dựa trên ý tưởng của Residual Vector Quantization (RVQ) để nén audio

1. Mục tiêu là quantize output của Encoder về một số lượng bit (= số Codebook) xác định trước.
2. Chiếu input vector về vector Codebook (Trung tâm của các cụm data định trước)
3. Trừ input vector cho codebook vector, lấy phần dư để làm input cho step tiếp theo với codebook tiếp theo.
4. Lặp lại cho đến hết codebook và thu được vector nén gồm các quantized vector (codebook vector) hay sau này sẽ gọi là acoustic tokens.



# Neural Codec

Aim: transform waveform from continuous time domain -> discrete frequency-based domain

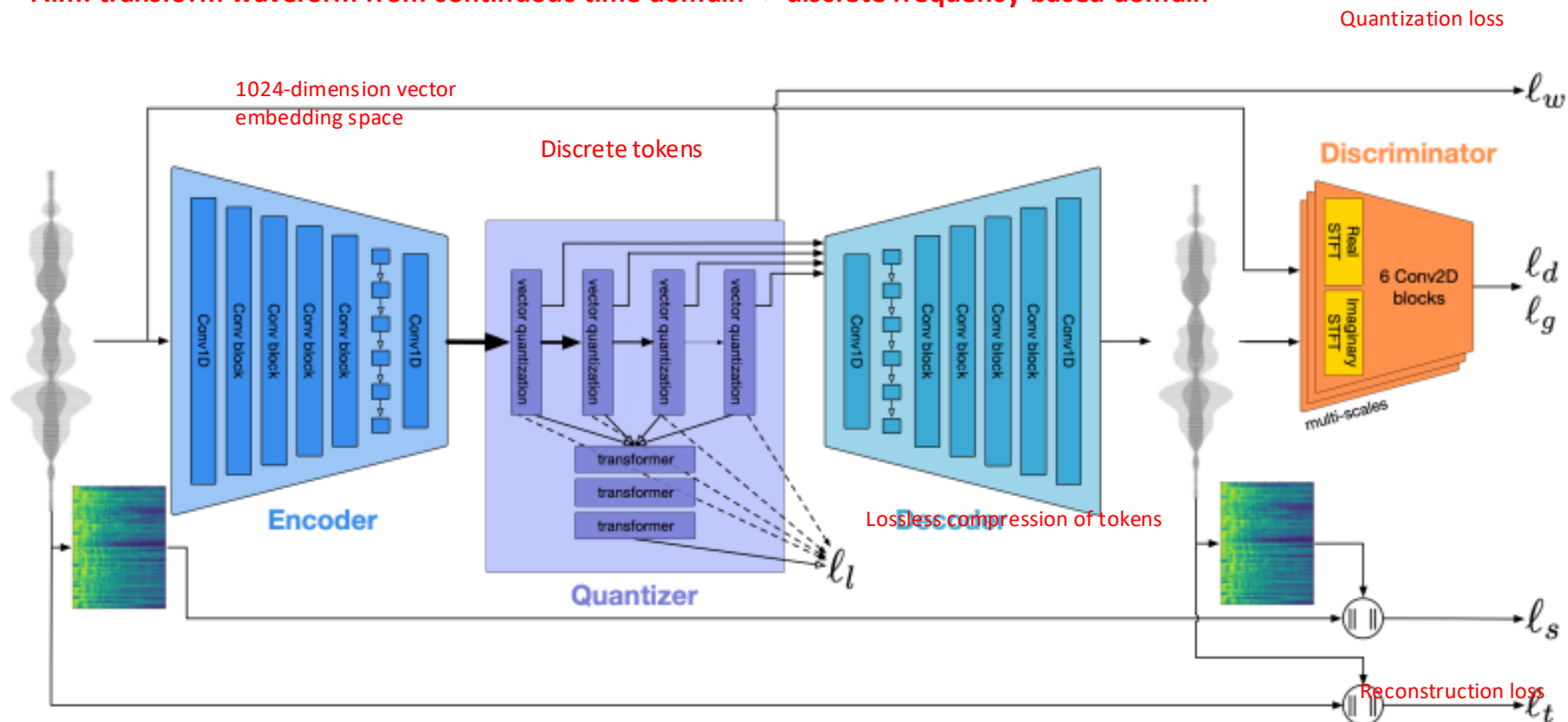


Figure 1: ENCODEC : an encoder decoder codec architecture which is trained with reconstruction ( $\ell_f$  and  $\ell_t$ ) as well as adversarial losses ( $\ell_g$  for the generator and  $\ell_d$  for the discriminator). The residual vector quantization commitment loss ( $\ell_w$ ) applies only to the encoder. Optionally, we train a small Transformer language model for entropy coding over the quantized units with  $\ell_l$ , which reduces bandwidth even further.



# Large Language Model Based Speech Generation

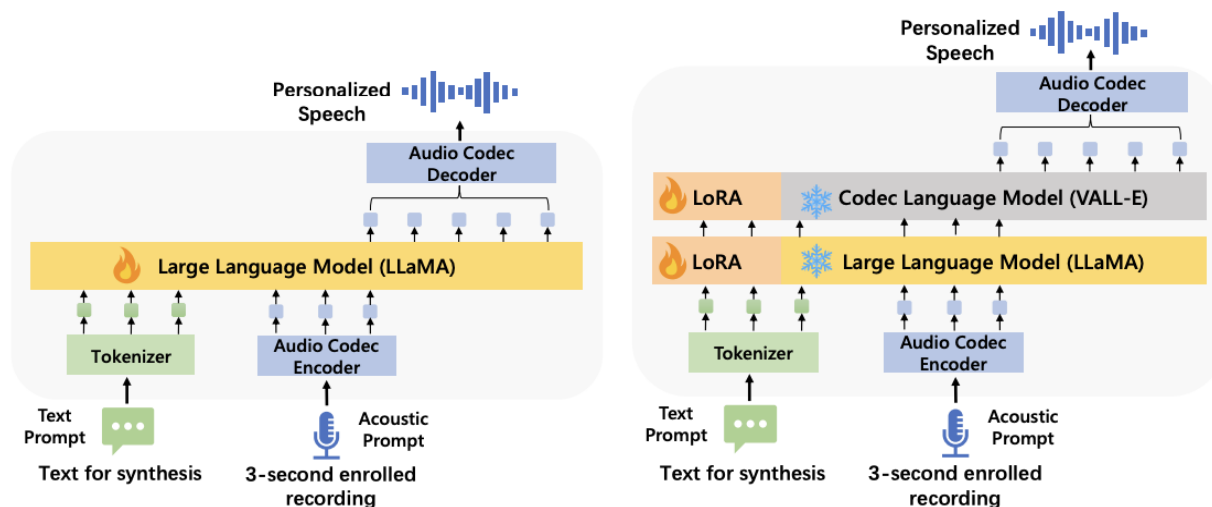
---

Thịnh Nguyễn/TTS

# Large Language Model Based Speech Generation

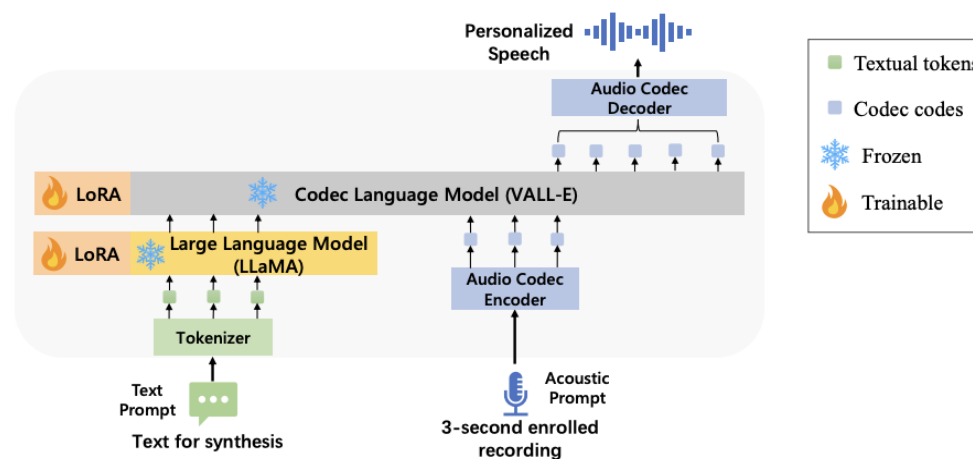
Mục tiêu của việc sử dụng Large Language Model cho Text to Speech:

1. Tận dụng khả năng hiểu ngữ cảnh của LLM, để có thể thay đổi ngữ điệu theo ngữ cảnh
2. Tận dụng khả năng Generalize của LLM để tăng sự đa dạng cho Speech



(a) Method A: Directly Fine-tuned LLM

(b) Method B: Superposed LLM and VALL-E



(c) Method C: Coupled LLM and VALL-E

# LLM Based Speech Generation: COSYVOICE

Xây dựng AutoRegressive Text to Speech model tận dụng LLM backbone

Chỉ cần thay đổi cấu trúc input là có thể đáp ứng các bài toán sau:

1. Text to Speech
2. Voice Clone
3. Multi-Lingual

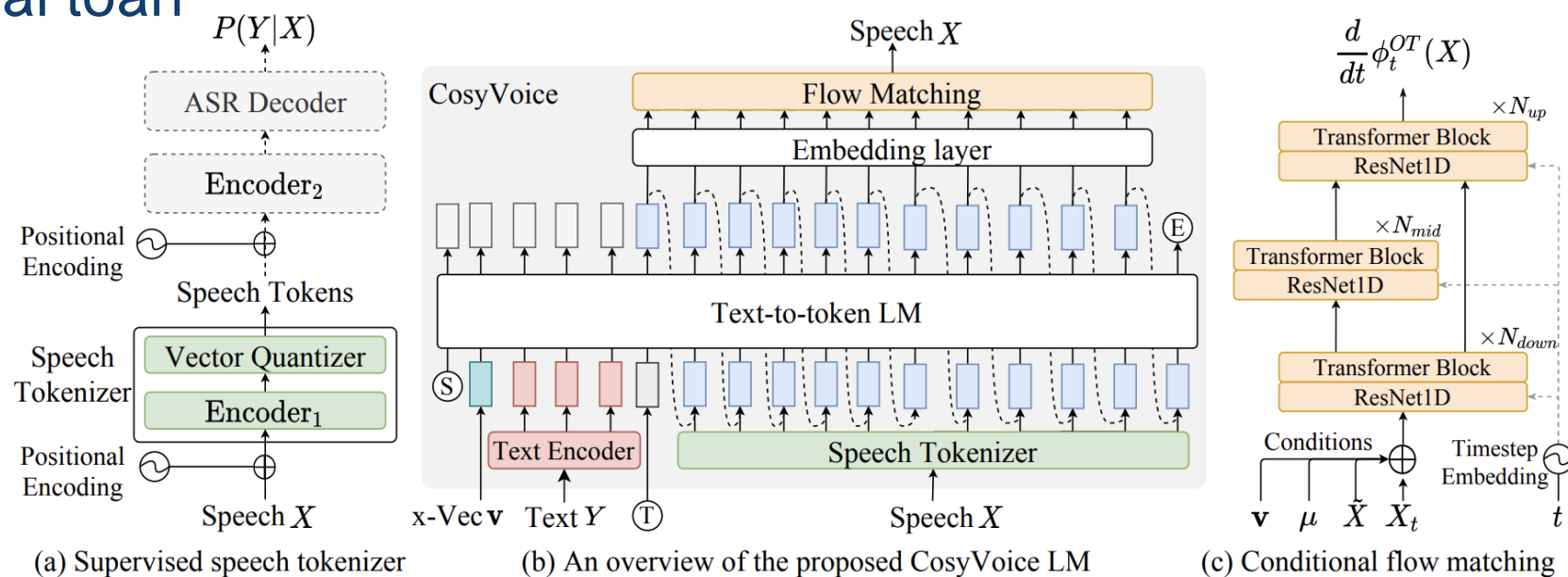


Figure 1: An overview of the proposed CosyVoice model. (a) demonstrates the  $\mathcal{S}^3$  tokenizer, where dashed modules are only used at the training stage. (b) is a schematic diagram of CosyVoice, consisting of a text-to-token LLM and a token-to-speech flow matching model.  $\textcircled{S}$ ,  $\textcircled{E}$  and  $\textcircled{T}$  denote the “start of sequence”, “end of sequence” and “turn of speech” tokens. Dashed lines indicate the autoregressive decoding at the inference stage. (c) provides an enlarged view of our flow matching model conditioning on a speaker embedding  $\mathbf{v}$ , semantic tokens  $\mu$ , masked speech features  $\tilde{X}$  and intermediate state  $X_t$  at timestep  $t$  on the probabilistic density path.

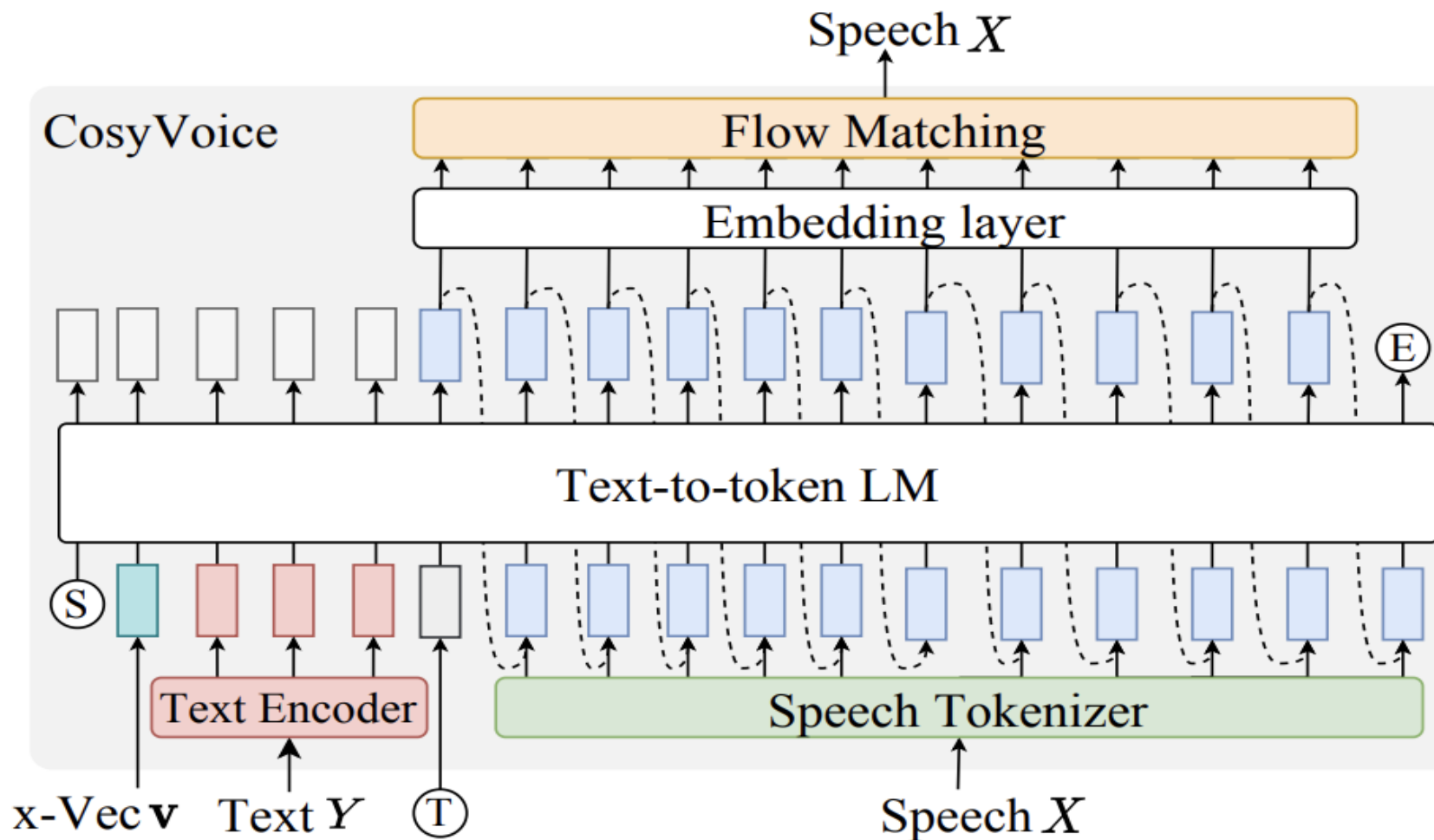
# LLM Based Speech Generation: COSYVOICE

## *Text to Speech*

1. Text được convert thành text tokens thông qua text encoder.
2. Speech được convert thành semantic speech tokens thông qua speech tokenizers.
3. Có thêm tokens đặc biệt **T** để ngăn cách giữa text và speech. Có thêm hai tokens đặc biệt **S** và **E** để xác định điểm bắt đầu và kết thúc của câu.
4. Có speaker embedding từ model Speaker recognition làm token đầu của chuỗi input tokens
5. sử dụng cơ chế auto-regressive + teacher forcing khi inference để sinh ra output semantic tokens.
  1. Trong đó input của LLM gồm có **S** + text token + **T** + left-shift speech token của ground truth khi training và khi infer thì là token của đầu ra của step trước đó.
  2. LLM sẽ lần lượt predict chuỗi các output semantics tokens từng bước một. và output hoặc groudtruth token của bước trước được đưa vào làm input tokens của bước sau tùy theo là inference hay training stage.
6. Sử dụng flow matching models để predict mel-spectrogram từ semantic tokens đầu ra của LLMs.

# LLM Based Speech Generation: COSYVOICE

## *Text to Speech*



(b) An overview of the proposed CosyVoice LM

# LLM Based Speech Generation: COSYVOICE

## Voice Clone

Với voice clone cơ chế hoạt động tương tự như Text to Speech tuy nhiên input của LLMs sẽ có thêm Prompt text tokens và prompt speech tokens. Trong đó:

- Prompt speech tokens được tính từ prompt audio của giọng muốn clone thông qua speech.
- Prompt text tokens là tokens của text tương ứng là transcript của prompt audio.



(a) Zero-shot In-context Learning



# LLM Based Speech Generation: COSYVOICE

## *Prompt, Non-Verbal, Fine-grained Control Text to Speech*

Chỉ cần đưa text vào training với cấu trúc instruction tương tự như Bảng 1:

1. Ví dụ điều khiển Speaker, style bằng prompts: “A happy girl with high tone and quick speech.<endofprompt>” Trong đó <endofprompt> để đánh dấu việc kết thúc prompts và chuyển sang text.
2. Ví dụ để tổng hợp non-verbal (cười, khóc,...) cần đưa các tags vào đúng vị trí của Non-verbal: *Well that’s kind of scary [laughter].*

---

### Speaker Identity

1. Selene 'Moonshade', is a **mysterious, elegant dancer** with a connection to the night. Her movements are both **mesmerizing** and **deadly**.<endofprompt>Hope is a good thing.
2. Theo 'Crimson', is a **fiery, passionate** rebel leader. Fights with fervor for justice, but struggles with **impulsiveness**.<endofprompt>You don't know about real loss.

---

### Speaking Style

1. A **happy girl** with **high tone** and **quick speech**.<endofprompt>The sun is shining brightly today.
2. A **sad woman** with **normal tone** and **slow speaking speed**.<endofprompt>I failed my important exam.

---

### Fine-grained Paralinguistics

1. Well that's kind of scary **[laughter]**.
  2. I don't think I over eat yeah **[breath]** and um I do exercise regularly.
  3. Well that pretty much covers **<laughter>the subject</laughter>** well thanks for calling me.
  4. The team's **<strong>unity</strong>** and **<strong>resilience</strong>** helped them win the championship.
- 

Table 1: Examples of speaker identity, speaking style, and fine-grained paralinguistics.



# LLM Based Speech Generation: COSYVOICE

## Prompt, Non-Verbal, Fine-grained Control Text to Speech

Chỉ cần đưa text vào training với cấu trúc instruction tương tự như Bảng 1:

- 1. Ví dụ điều khiển Speaker, style bằng prompts: “A happy girl with high tone and quick speech.<endofprompt>” Trong đó <endofprompt> để đánh dấu việc kết thúc prompts và chuyển sang text.
- 2. Ví dụ để tổng hợp non-verbal (cười, khóc,...) cần đưa các tags vào đúng vị trí của Non-verbal: *Well that’s kind of scary [laughter].*

---

### Speaker Identity

- 1. Selene 'Moonshade', is a **mysterious, elegant dancer** with a connection to the night. Her movements are both **mesmerizing** and **deadly**.<endofprompt>Hope is a good thing.
- 2. Theo 'Crimson', is a **fiery, passionate** rebel leader. Fights with fervor for justice, but struggles with **impulsiveness**.<endofprompt>You don't know about real loss.

---

### Speaking Style

- 1. A **happy girl** with **high tone** and **quick speech**.<endofprompt>The sun is shining brightly today.
- 2. A **sad woman** with **normal tone** and **slow speaking speed**.<endofprompt>I failed my important exam.

---

### Fine-grained Paralinguistics

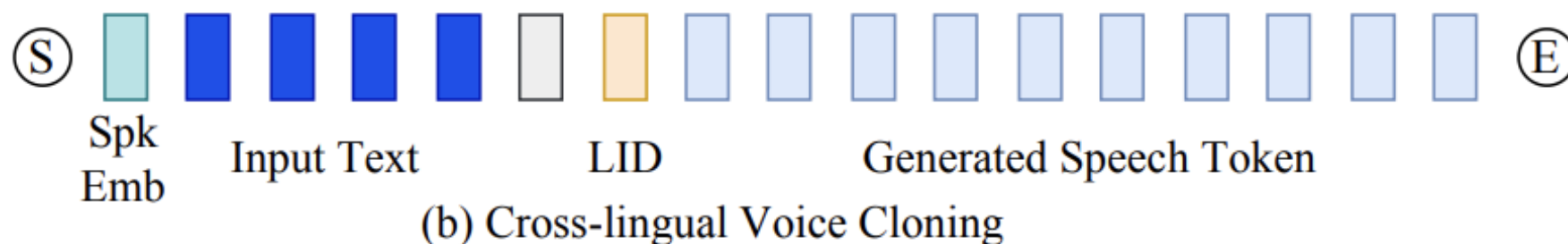
- 1. Well that’s kind of scary **[laughter]**.
  - 2. I don’t think I over eat yeah **[breath]** and um I do exercise regularly.
  - 3. Well that pretty much covers **<laughter>the subject</laughter>** well thanks for calling me.
  - 4. The team’s **<strong>unity</strong>** and **<strong>resilience</strong>** helped them win the championship.
- 

Table 1: Examples of speaker identity, speaking style, and fine-grained paralinguistics.

# LLM Based Speech Generation: COSYVOICE

## *Multi-Lingual*

Chỉ cần thêm language ID tokens như hình dưới vào sau input text tokens là được.





# LLM Based Speech Generation: COSYVOICE

***CosyVoice Demo: [Live demo](#)***



# References

---

1. Tacotron2: [paper](#), [source code](#)
2. FastSpeech2: [paper](#), [source code](#)
3. Hifigan: [paper](#), [source code](#)
4. NeuralSpeech3: [Demo page & Paper](#)
5. Audiobox: [Demo & Paper](#)
6. CosyVoice: [paper](#), [source code](#)
7. Encodec: [paper](#), [source code](#)
8. Flow Matching: [Tutorial](#)



# Thank you

Century Tower, Times City,  
458 Minh Khai, Hai Bà Trưng, Hà Nội

[v.thinhnv13@vinbigdata.com](mailto:v.thinhnv13@vinbigdata.com)

[www.vinbigdata.org](http://www.vinbigdata.org)

