VoxCPM
Tokenizer-Free TTS & Voice Cloning

Mì AI

Warning

- #1: Không đủ thời gian và chuyên môn đi sâu vào kiến trúc, mã nguồn của VoxCPM
- #2: Không chịu trách nhiệm khi sử dụng Voice Cloning vào mục đích xấu: lừa đảo, deepfake… Chỉ chia sẻ với mục đích học tập, nghiên cứu.

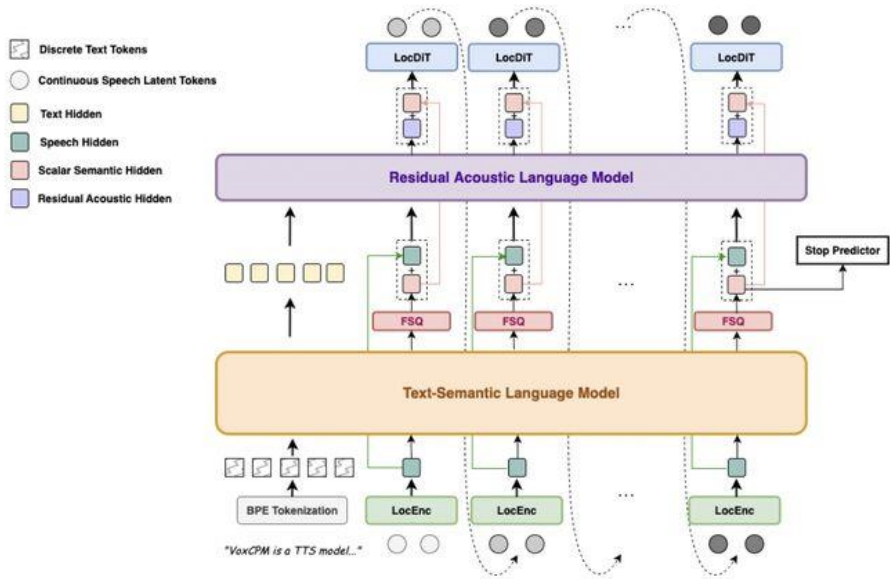
VoxCPM

Hugging Face OpenBMB | Technical Report Arxiv | Live PlayGround Demo | Audio Samples Page

Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning

Overview

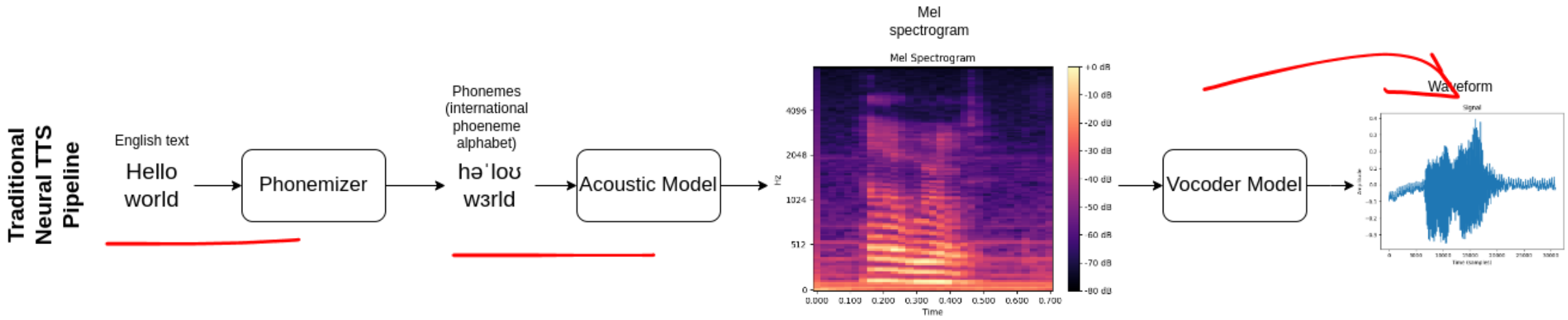
VoxCPM is a novel tokenizer-free Text-to-Speech (TTS) system that redefines realism in speech synthesis. By modeling speech in a continuous space, it overcomes the limitations of discrete tokenization and enables two flagship capabilities: context-aware speech generation and true-to-life zero-shot voice cloning.



https://github.com/OpenBMB/VoxCPM

Traditional TTS

- Flow truyền thống : Text normalization → Tokenization → Phoneme → Acoustic → Vocoder
- Tokenizer giới hạn tính biểu cảm & tự nhiên của giọng nói



Why Token-Free Matters

Natural speech

Preserves prosody

Adapts to context

The Backbone: MiniCPM-4

MiniCPM-4

Half billion parameters

Hierarchical modeling

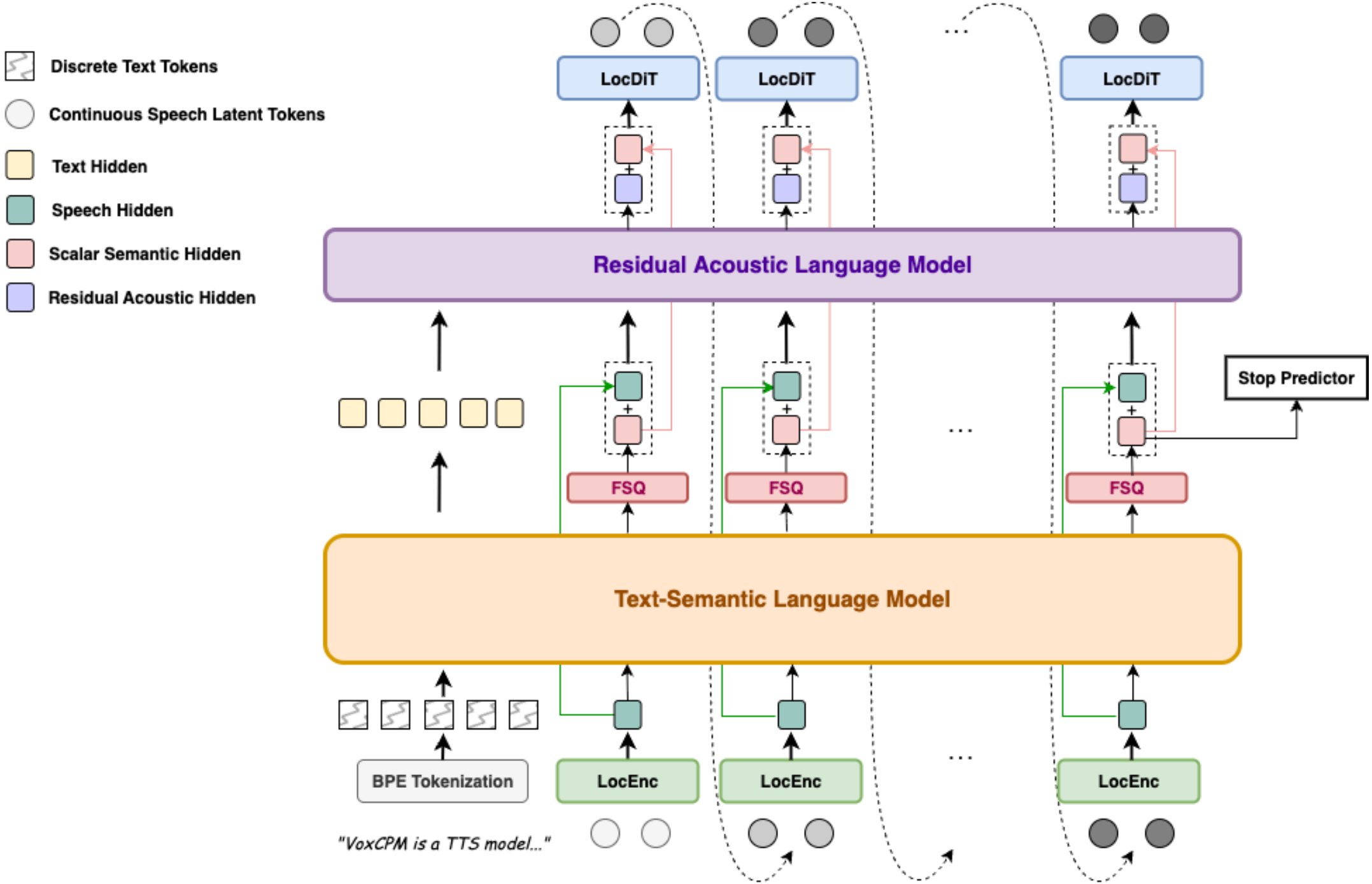
FSQ technique

Performance: Fast Enough for Real Time

0.17 Real-Time Factor (RTF)

6× Faster than playback

Viable for **live scenarios**


Discrete Text Tokens
Continuous Speech Latent Tokens
Text Hidden
Speech Hidden
Scalar Semantic Hidden
Residual Acoustic Hidden
LocDiT
LocDiT
LocDiT
Residual Acoustic Language Model
Stop Predictor
FSQ
FSQ
FSQ
Text-Semantic Language Model
BPE Tokenization
LocEnc
LocEnc
LocEnc
"VoxCPM is a TTS model..."

🔷 VoxCPM

VoxCPM hiểu **ngữ cảnh văn bản** và tự động điều chỉnh prosody (nhịp, trọng âm, cao độ):

- dấu hỏi → giọng lên ở cuối câu,
- câu nhấn mạnh → giọng mạnh hơn,
- câu mô tả → nhịp thoải mái, tự nhiên hơn. Medium

📌 Đây là kết quả của:

- **MiniCPM-4 backbone** giúp mô hình nắm được ngữ nghĩa và ngữ điệu,
- **Hierarchical semantic-acoustic modeling** để tách biệt rõ ràng ý nghĩa và cách diễn đạt. arXiv

👉 Kết quả: giọng nói trong các đoạn dài, phức tạp nghe **mượt hơn, tự nhiên hơn** so với các mô hình chỉ tập trung vào token → acoustic mapping.

✨ Context-Aware Speech Generation

✨ True-to-Life Voice Cloning

✨ Streaming TTS & Real-Time Factor ~0.15+

✨ Support Fine-tuning (SFT & LoRA)

✨ Dễ sử dụng via Python / CLI / API GitHub

Known Problems

Long Text Issues

Very long inputs may **destabilize** output

Basic Emotional Control

Cannot precisely **adjust emotion levels**

Limited Language Support

Training data mainly **English & Chinese**

Research frontier, not consumer-safe product

Hugging Face

Search models, datasets, u

65.1 TFLOPS

2

Phạm Phú Ngọc Trai

JayLL13

Follow

JayLL13/**VoxCPM-1.5-VN** like 7

Text-to-Speech | Safetensors | dolly-vn/dolly-audio-1000h-vietnamese | Vietnamese

Model card | Files and versions xet | Community

VoxCPM-1.5B-VN

Model TTS đã được training với dataset 1000h tiếng Việt dolly-vn/dolly-audio-1000h-vietname

Checkpoint at itter 50000 - epoch 0.6.

Device: NVIDIA RTX5090

How-tos

Finetune a LoRa:

Follow the original guide here

Inference:

https://huggingface.co/JayLL13/VoxCPM-1.5-VN

Hands-on

- Load Finetune model to support Vietnamese
- Try model on CLI
- Try model on prebuilt UI with Gradio
- Try to build new UI with Gradio