Khmer Text-to-Speech Evolutions

Seanghay Yath
y.seanghay@dgc.gov.kh
Senior ML Engineer @ DGC, MPTC

Introduction

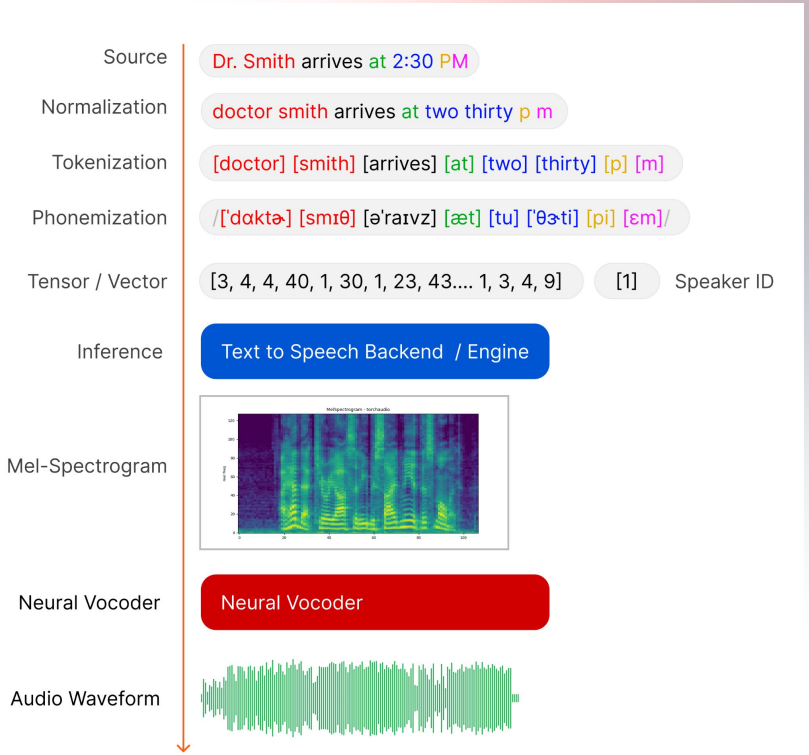
We will explore the evolution of Text-to-Speech (TTS) technology, tracking its journey from traditional methods to modern AI. By analyzing the pros and cons of key architectures—including HMM, Tacotron, VITS, OmniVoice, and VoxCPM2—we'll help you determine the best model for your specific needs, with a special focus on the Khmer language.

Text-to-Speech Pipeline

Text Processing to handle linguistics and phonetics.

Acoustic Modeling to generate frequency patterns (Mel-Spectrograms).

Neural Vocoding to synthesize the final audible waveform.



Khmer Challenge

- No standard word boundary. ម៉ាស៊ីនកិនស្រូវនៅស្រែ
- Stacking consonants, encoding orders and complex ligatures. ស្ត្រីនៅស្រុកស្រែឈ្មោះស្រីង
- Native Khmer mixed with Pali-Sanskrit. E.g លោកមកមកហើយ
- Acoustic nuance.
- Limited high-quality parallel speech-text corpora.
- Various accents. (Official, Phnom Penh, Siemreap, Banteay Meanchey, etc)

Concatenative Synthesis

Synthesizing sounds (often speech or musical instruments) by slicing up recorded audio into small segments and stringing them back together in a new order.

មួយ.mp3

ពីរ.mp3

បី.mp3

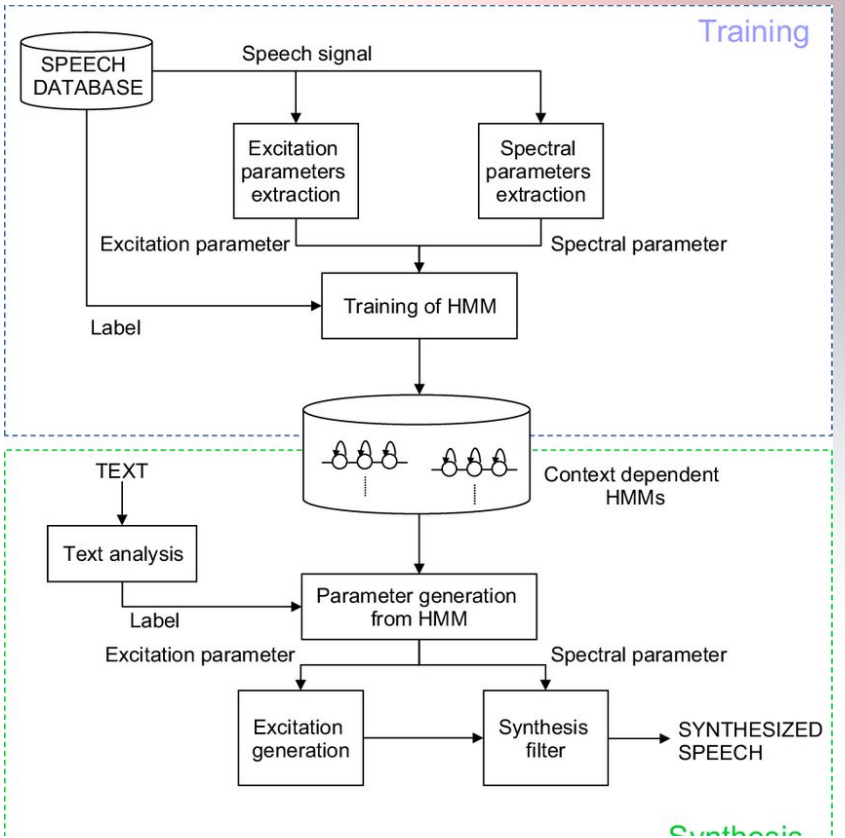
បួន.mp3

...

ដប់.mp3

HMM-based Framework

Instead of cutting and pasting raw audio waveforms like older concatenative systems, an HMM framework models the statistical properties of speech and generates smooth, mathematically generated parameters.



Tacotron / 13M

A neural network architecture developed by Google for direct text-to-speech synthesis. It simplifies the traditional, complex TTS pipeline by generating human-like speech directly from text using two primary components.

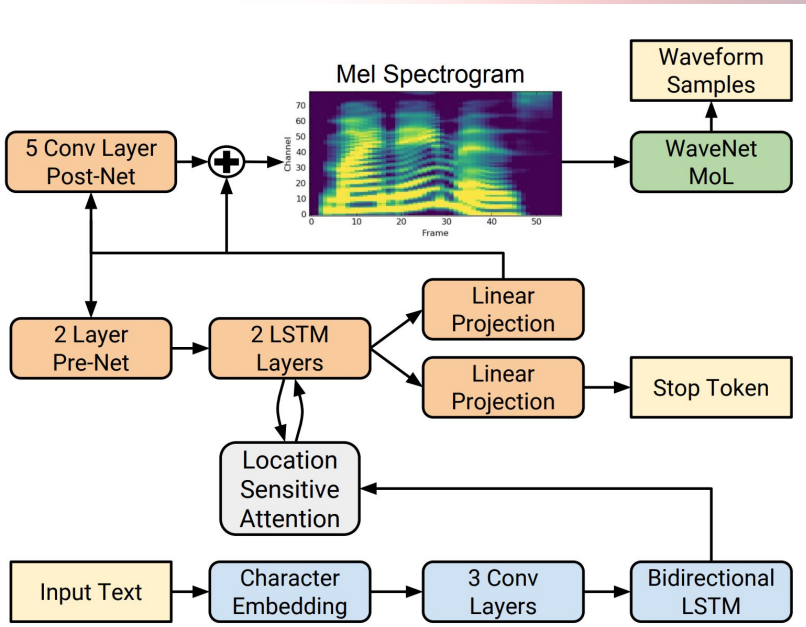
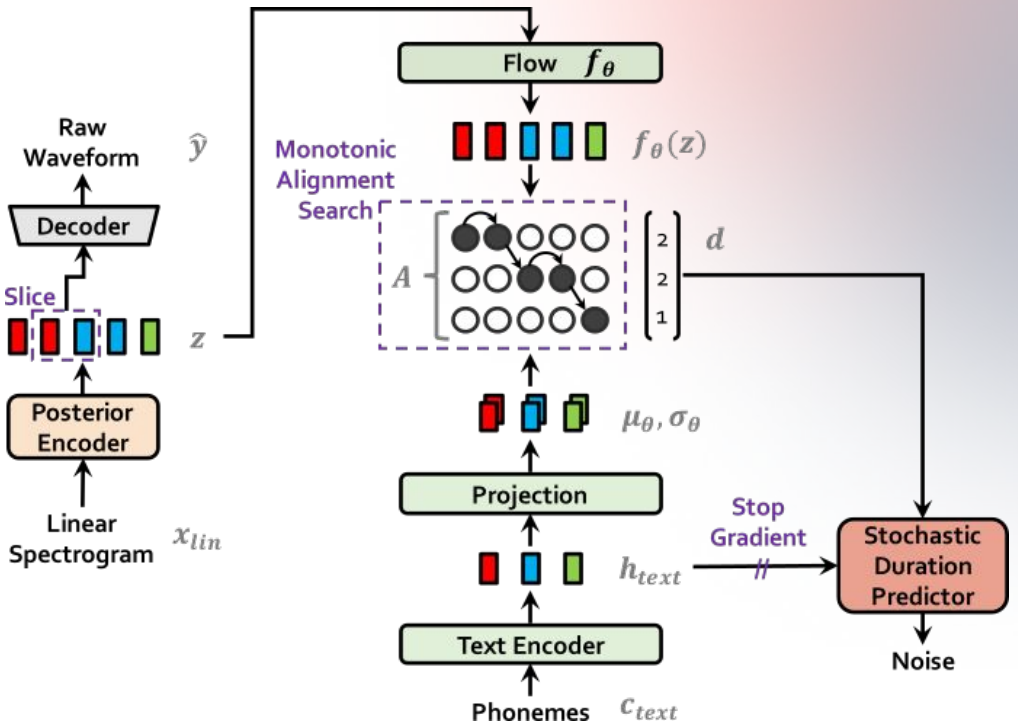


Fig. 1. Block diagram of the Tacotron 2 system architecture.

VITS / ~0.1B

Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.

While Tacotron 2 requires a two-stage process (Text → Mel-spectrogram → Waveform), VITS is a **true fully end-to-end model** that generates raw audio directly from text in a single step.



OmniVoice / 0.8B

An open-source, massively multilingual zero-shot Text-to-Speech (TTS) model developed by Xiaomi AI Lab's Next-gen Kaldi team.

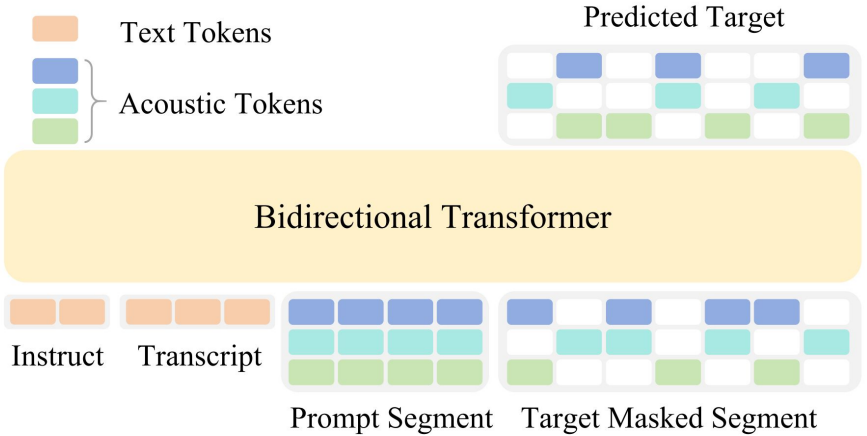
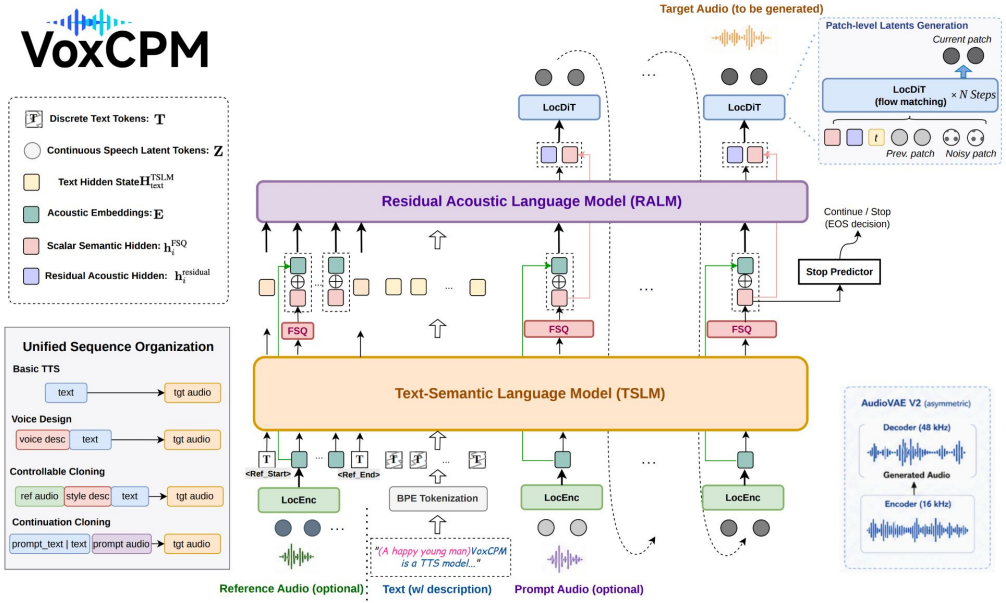


Figure 1: Illustration of OmniVoice architecture.

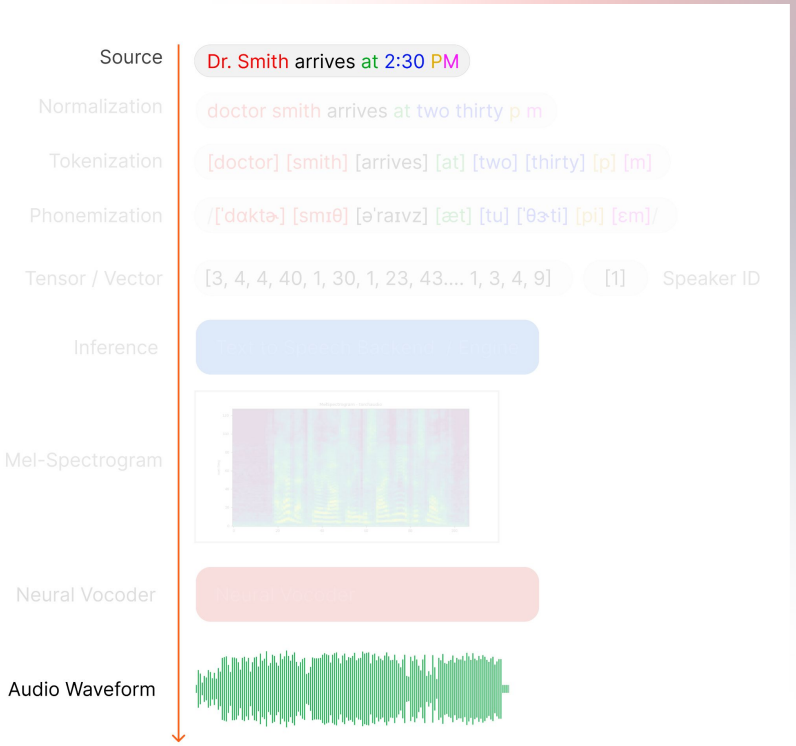
VoxCPM2 / 2B

A recently released (mid-2026) open-source tokenizer-free Text-to-Speech (TTS) foundation model developed by OpenBMB



VoxCPM2 / 2B

Completely bypasses discrete tokenization using a Diffusion-Autoregressive architecture. It leverages an asymmetric AudioVAE V2 that maps text and audio context directly into a continuous latent space, natively upsampling audio inputs to a 48kHz studio-quality output



Comparative Analysis

Model	Speed	Naturalness	Hardware	Voice Clone
HMM	Very Fast	Low	Very Low / CPU	NO
Tacotron	Slow to Moderate	High	High / CPU / GPU	NO
VITS	Fast	Very High	Moderate / CPU / GPU	NO
Omnivoice	Moderate	Excellent	High / GPU Only	YES
VoxCPM2	Moderate / Slow	State-of-the-Art	High / GPU Only	YES

What’s best for your need?

Scenario	Recommended Model	Key Reason
Mobile Apps	VITS	Best balance of speed and quality.
Global Platforms	OmniVoice	Massive multilingual support.
Audiobooks / Ads	VoxCPM2	Studio-grade realism and emotion.
IoT / Legacy devices	HMM	Minimal resource usage.

Ethical Landscape

1. Identity Security

Voice is a biometric identifier. Cloning without consent risks deepfake fraud and identity theft.

2. Explicit Consent

Voice owners must give informed, revocable, and documented permission for their likeness to be used.

3. Legislation

The EU AI Act and TN's ELVIS Act are establishing legal protections for vocal personas.

Q&A

Slide will be available at https://seanghay.com