

Making Machines Speak Khmer

Challenges, Solutions, and Implementation

Seanghay Yath / យ៉ាត់ សៀងហៃ

June 8, 2025

seanghay.dev@gmail.com

Seanghay Yath / យ៉ាត់ សៀងហៃ

AI Researcher, Product Developer

Frontend Lead at the Digital Government Committee



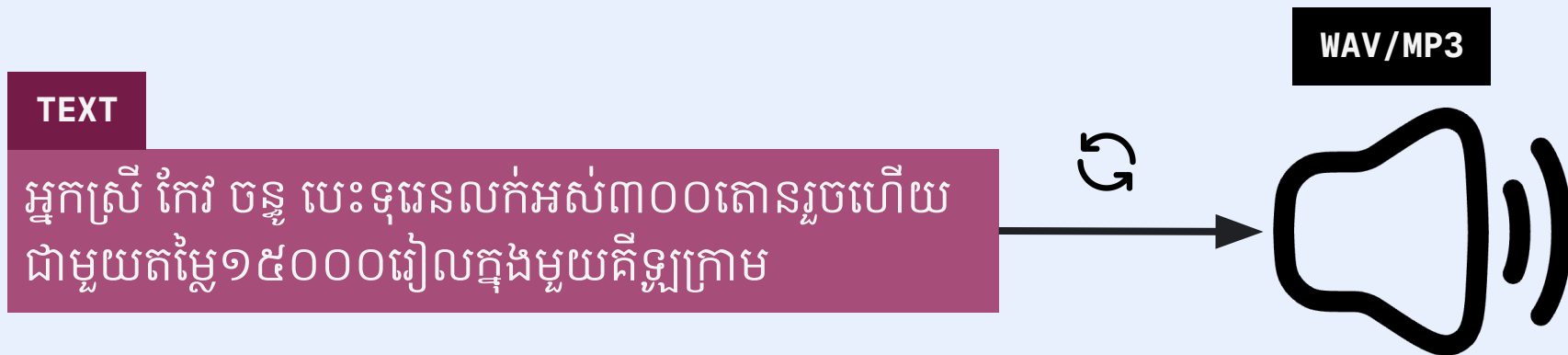
I passionate about building products for people. With over 10 years of programming experience, I've contributed to impactful projects including Pi Pay, Koh Santepheap App, JOON, StopCOVID, VERIFY.GOV.KH, SARIKA.GOV.KH, KhmerDict.com, and KhmerScan.com.



The views and opinions expressed herein are solely my own and do not reflect, represent, or constitute the official position, policy, or endorsement of any organization, company, institution, or employer with which I am or have been affiliated.

What is Text to Speech?

A technology that converts written text into spoken audio. It's a core component of speech synthesis systems that enables computers to "read aloud" digital text in a human-like voice.



Why Text to Speech? / Practical Use



Content/News Creation



Video Narration



Visual Impairment Support



Public Announcements



Virtual Assistants / Robotics



Audiobook Production

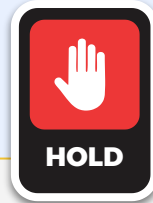


Navigation Systems



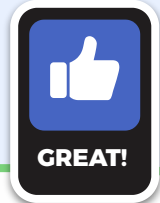
Customer Service

HUMAN



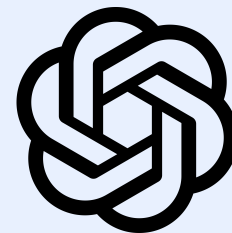
- ❑ Rich emotional expression and natural intonation
- ❑ Perfect contextual understanding
- ❑ Variable quality (fatigue, mood, health affects performance)
- ❑ High cost and time investment
- ❑ Requires scheduling and coordination
- ❑ Excellent for creative content and storytelling
- ❑ Difficult to modify after recording
- ❑ Limited availability and scalability
- ❑ មនុស្សពេលវេលានិយាយច្រើនលឺក

AI



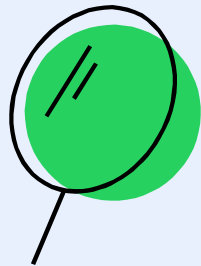
- ❑ Consistent quality every time
- ❑ Available 24/7 instantly
- ❑ Low cost after initial setup
- ❑ Easy to update and modify
- ❑ Unlimited scalability
- ❑ May sound robotic or monotone
- ❑ Can struggle with complex pronunciations
- ❑ Perfect for technical documentation and accessibility
- ❑ Keep improving rapidly with AI technology

Popular Text-to-Speech Systems (Proprietary)



ElevenLabs





How to Build One from Scratch?

Research

- ❑ Finding existing open source projects
- ❑ Read many TTS papers
- ❑ A lot of trials and errors
- ❑ Reaching out to experts

<https://github.com/seanghay/awesome-khmer-language>



Collect Requirements

1. Knowledge of Python/ML Ecosystems
2. A Fairly Powerful GPU
3. A Decent Voice Dataset
4. Khmer Natural Language Toolkit

Learning

- ❑ Python / Virtualenv / Anaconda
- ❑ Machine Learning Concepts
- ❑ PyTorch
- ❑ Scipy
- ❑ Numpy
- ❑ C/C++ for binding
- ❑ etc.

Voice Dataset

Text-to-speech datasets consist of paired audio and text samples, where each audio clip has a limited duration and corresponds to specific written content. To effectively utilize these datasets, we need techniques that can precisely align and segment the text-speech pairs.

Audio

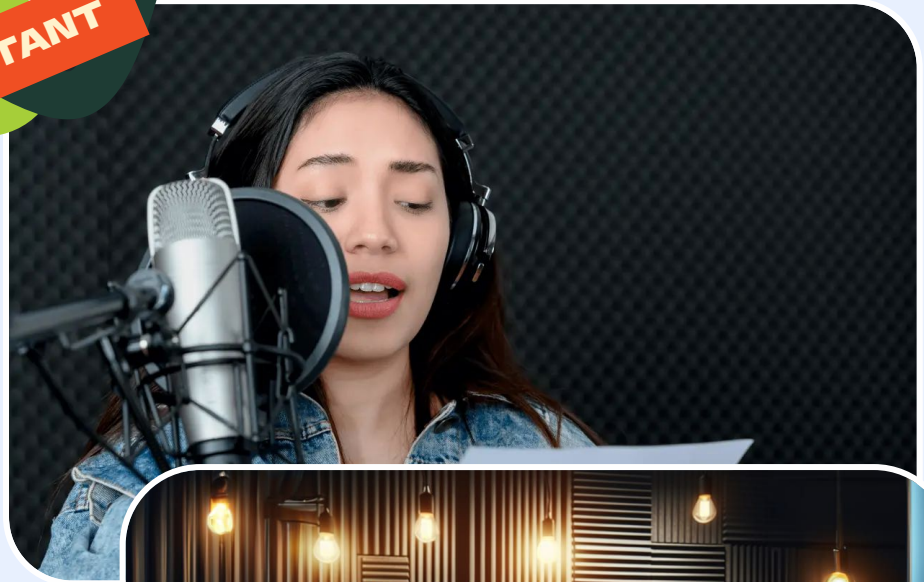


Text អ្នកស្រី កែវ ចន្ទ បេះទុរេនលក់អស់៣០០តោនរួចហើយ ជាមួយតម្លៃ១៥០០០រៀលក្នុងមួយគីឡូក្រាម

Professional Recording Setup

IMPORTANT

- ❑ Voice Actor
- ❑ Quiet Room / Less Noise
- ❑ Good Microphones
- ❑ Consistent Reading Style
- ❑ Quality Control



Why Khmer TTS is hard?

- ❑ Khmer Character Encoding
- ❑ Khmer Normalization
- ❑ Khmer Word Boundary Tokenization
- ❑ Khmer Grapheme to Phoneme (Linguistic Processing)
- ❑ Khmer Pronunciation Toolkit (Linguistic Processing)

Build the Foundation!

- ❑ Khmer Character Encoding → Makara Sok, Marc Durdin et al.
- ❑ Khmer Normalization ([khnormal.py](#) by sil.org) → **khmernormalizer**
- ❑ Khmer Word Boundary Tokenization → **khmercut**
- ❑ Khmer Grapheme to Phoneme → **sosap**
- ❑ Khmer Pronunciation Toolkit → **tha**

Reference

<https://github.com/seanghay/awesome-khmer-language>

Build the Foundation!



Khmer language is unique and it's awesome.

Building its toolkit is challenging and exciting at the same.

Dataset Postprocessing

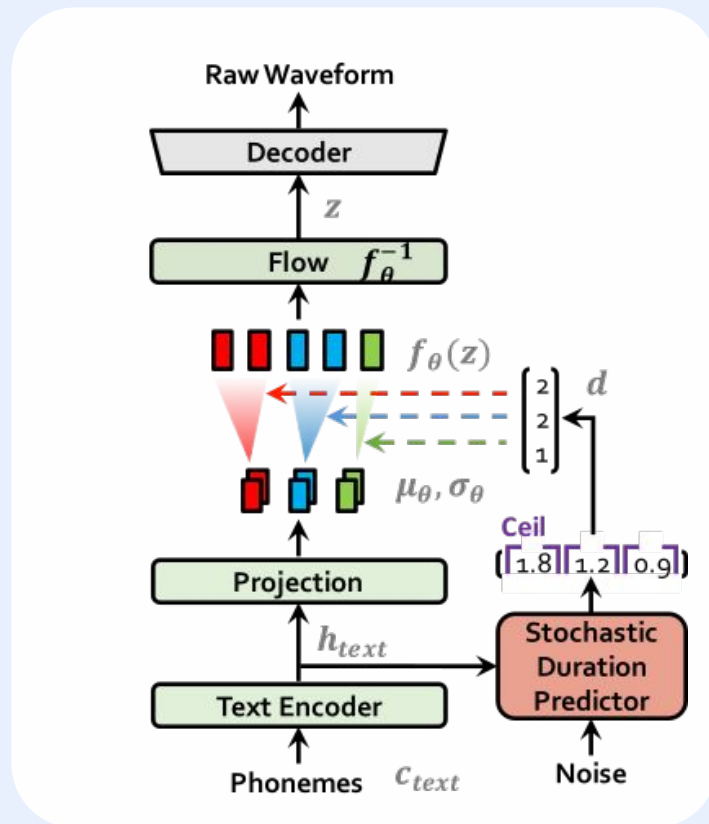
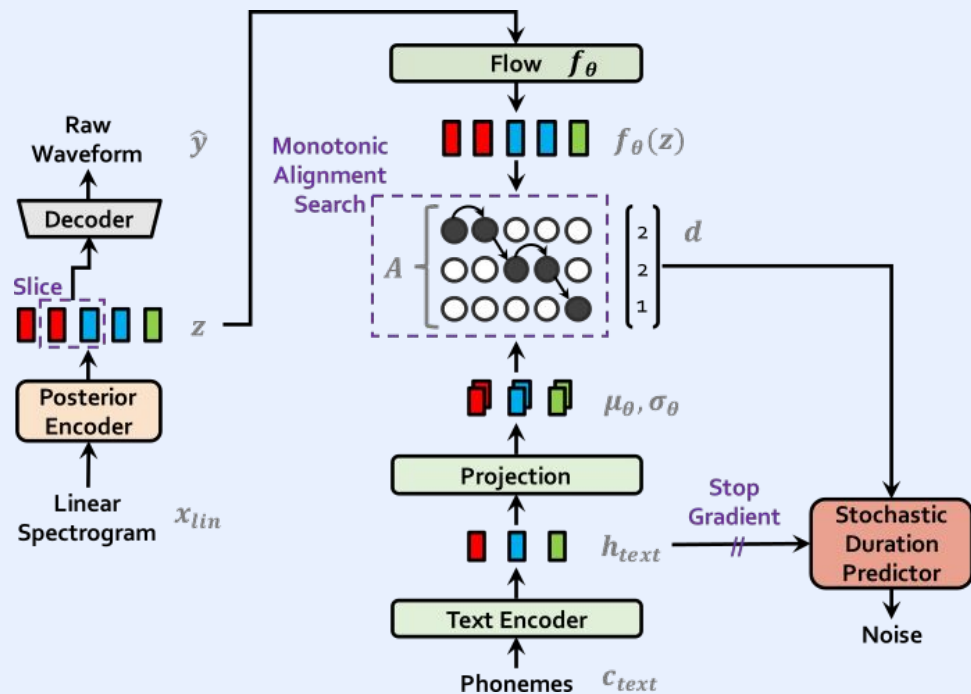
- ❑ Chunking / Alignment → **Khmer Forced Aligner (kfa)**
- ❑ Text Normalization (regex, pynini, openfst, thrax)
- ❑ Sample Rate Conversion (ffmpeg, librosa)
- ❑ Noise Reduction (scipy, pydub, librosa)
- ❑ Loudness Normalization (scipy, pydub, librosa)
- ❑ Sound Correction (scipy, pydub, librosa)

Model Architecture Selection

- ❑ Tacotron
- ❑ WaveGlow
- ❑ FastSpeech
- ❑ **VITS (<https://github.com/jaywalnut310/vits>)**
- ❑ TortoiseTTS
- ❑ StyleTTS



VITS Architecture (Jaehyeon Kim, et al.)



GPU Rental / Hosting



vast.ai



Setup a Machine

-\$3,000



Train



- ❑ Take nearly a week
- ❑ Many errors and crashes
- ❑ Out of Memory / Reconfiguration
- ❑ Learn to Resume Training
- ❑ Monitoring
- ❑ Evaluate
- ❑ Stop the training

Inference & Hosting

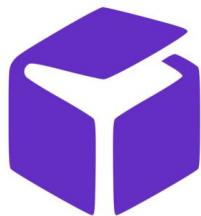
Hosts the application somewhere so that it can serve real users with efficiency and speed. There are a couple of inference engines out there such as

- ❑ **ONNXRuntime** by Microsoft
- ❑ **MNN** by Alibaba
- ❑ **NCNN** by Tencent
- ❑ **TorchServe** by PyTorch
- ❑ **Triton Inference Server** by NVIDIA

Serverless GPU



vast.ai



RunPod

eplicate

The Replicate logo, featuring a stylized icon of three stacked horizontal bars of increasing height, followed by the word "eplicate" in a bold, black, sans-serif font.



**It would be
possible without
the support and
the work from
amazing people**

- ☐ H.E. Chanty Sothy
- ☐ Dr. Riny Bouy
- ☐ Makara Sok
- ☐ Marc Durdin
- ☐ Dr. Soky Kak
- ☐ Vitou Phy
- ☐ Kyle Gorman
- ☐ Richard Sproat

References



<https://github.com/seanghay/awesome-khmer-language>

<https://github.com/AdolfVonKleist/Phonetisaurus>

<https://github.com/seanghay/khmernormalizer>

<https://github.com/seanghay/khmercut>

<https://github.com/seanghay/tha>

<https://github.com/seanghay/phonetisaurus-js>

<https://github.com/sillsdev/khmer-character-specification>

<https://github.com/seanghay/automatic-phonemic-and-phonetic-transcription>

Question



Questions? Email me at
seanghay.dev@gmail.com



Website

seanghay.com

GitHub

seanghay

Twitter

seanghay_yath

HuggingFace

seanghay

Email

seanghay.dev@gmail.com



[Download the Slide](#)

That's a wrap

Thank you!

