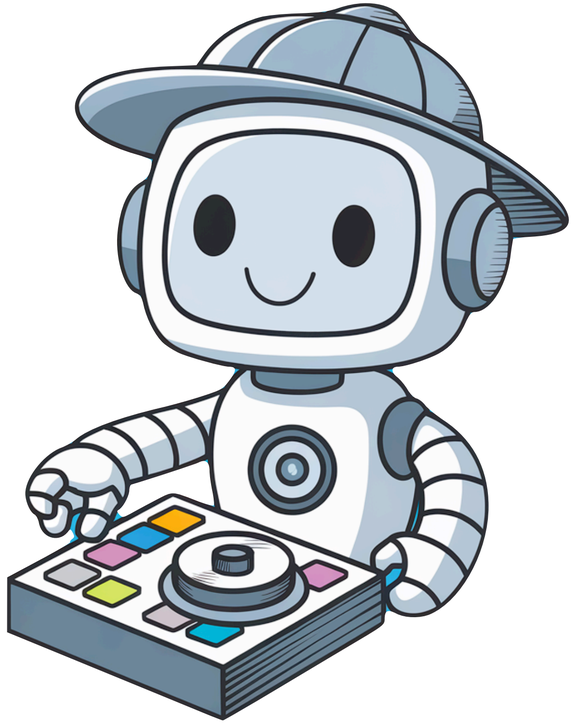
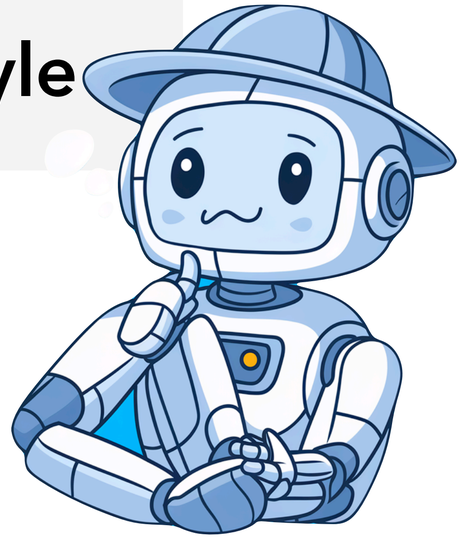
🎛️ EmoKnob Enhance Voice Cloning with Fine-Grained Emotion Control

Haozhe Chen, Run Chen, Julia Hirschberg
Columbia University



Text-to-Speech (TTS) has gotten very good.

But they usually don’t allow control of emotion/style



Traditional TTS pipeline synthesizes speech whose emotion is entirely decided by input text

Text Input

"This is great"

→

Traditional TTS Models

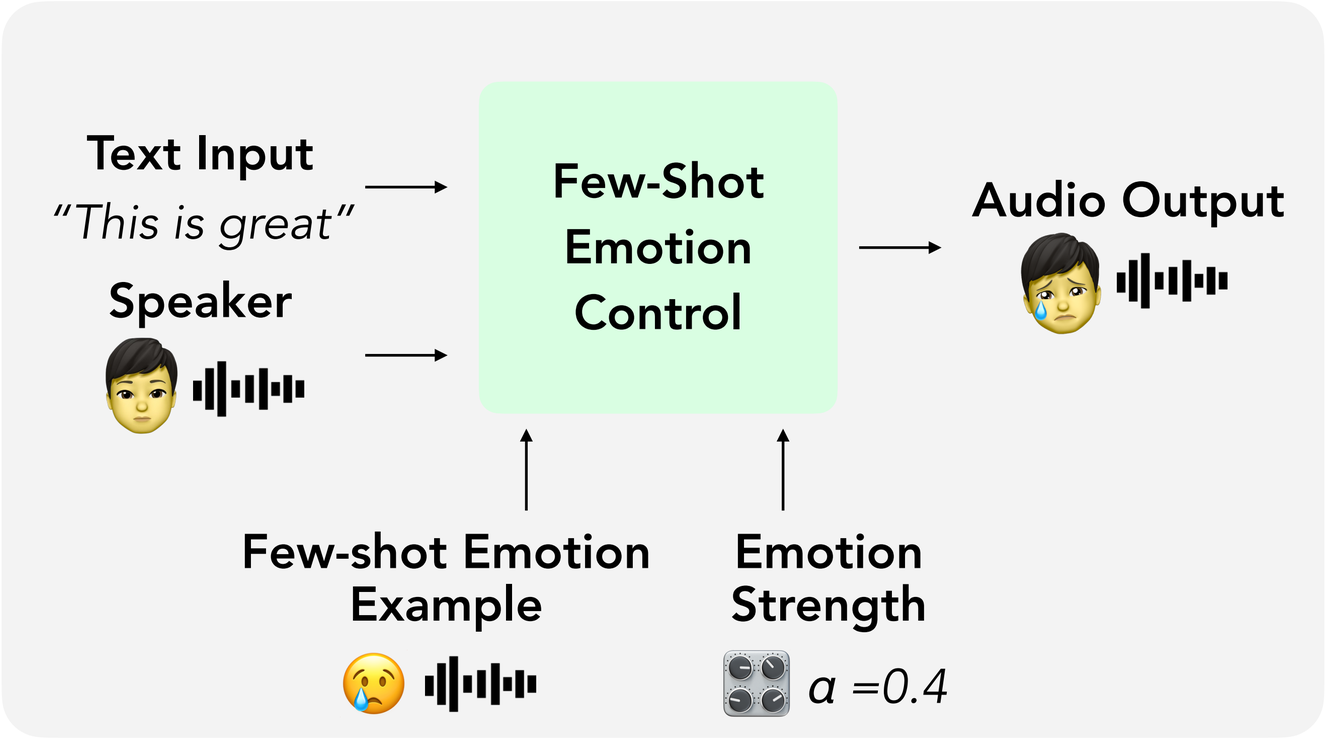
→

Audio Output

😊

Introducing 🎛️ **EmoKnob**

Fine-grained control of emotion specified by few-shot samples



Consider Shakespeare's famous line

To be, or not to be, that is the question.

Most TTS services can only convey it in one way with a given voice

🔊 (ElevenLabs)

Consider Shakespeare's famous line

To be, or not to be, that is the question.

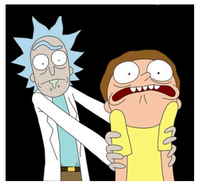
But it can be happy, sad, surprise, contempt …
Each conveys a different message

🎛️ Generated with EmoKnob

🎛️ EmoKnob allows users to specify and control a wide range of emotions

😐 → 😡

Neutral → Angry

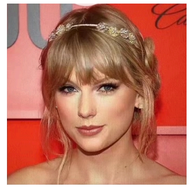


Aggresive → 

Empathetic

😐 → 💪

Neutral → Charismatic



Young Taylor Swift → 

Old Taylor Swift

😐 → 😍

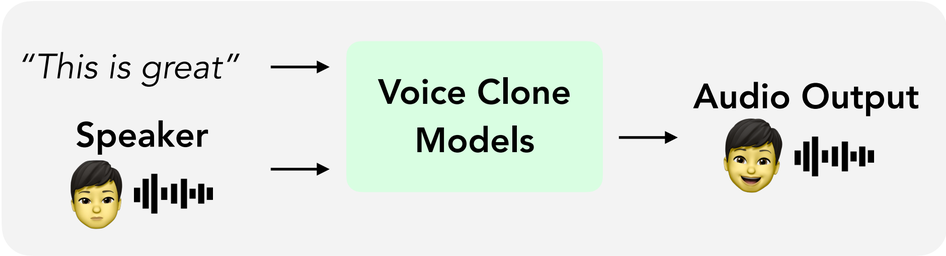
Neutral → Desire

Joe Biden → 

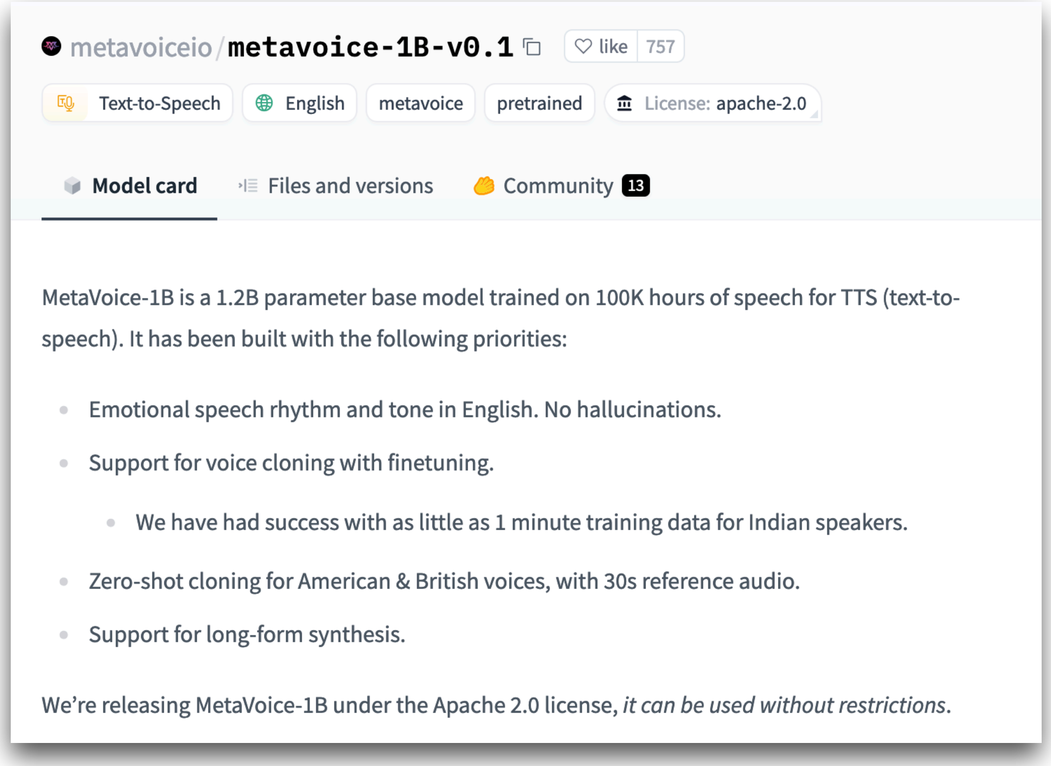
Joe Biden

With Trump speaking style

🎛️ EmoKnob is built on recent developments of voice cloning models



Voice clone models replicate speaker but speech emotion is still decided by input text



metavoiceio / **metavoice-1B-v0.1** like 757

Text-to-Speech | English | metavoice | pretrained | License: apache-2.0

Model card | Files and versions | Community 13

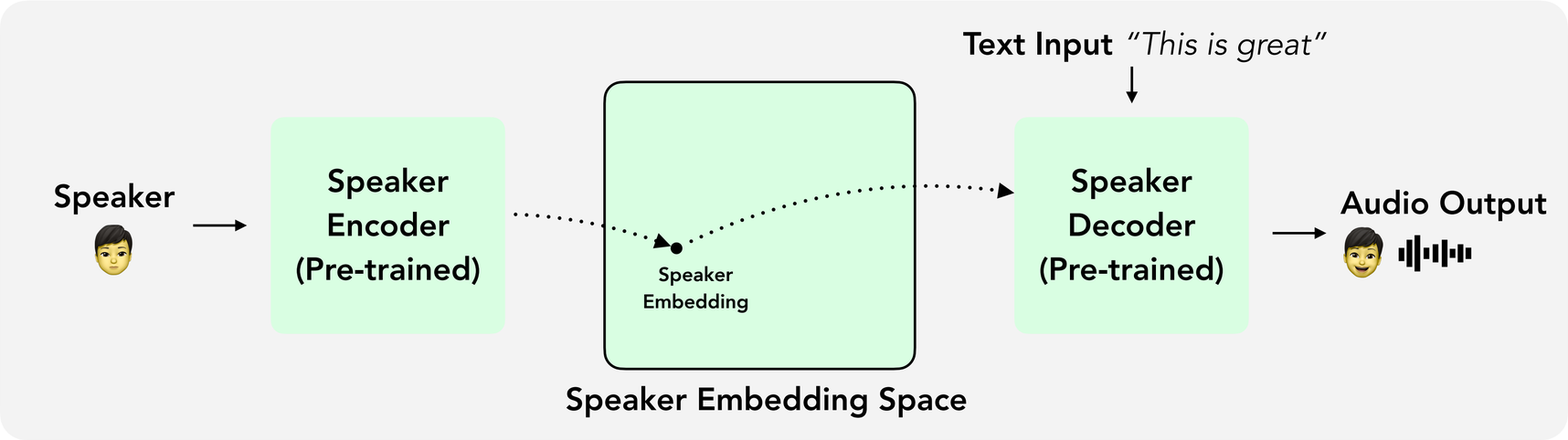
MetaVoice-1B is a 1.2B parameter base model trained on 100K hours of speech for TTS (text-to-speech). It has been built with the following priorities:

- Emotional speech rhythm and tone in English. No hallucinations.
- Support for voice cloning with finetuning.
  - We have had success with as little as 1 minute training data for Indian speakers.
- Zero-shot cloning for American & British voices, with 30s reference audio.
- Support for long-form synthesis.

We're releasing MetaVoice-1B under the Apache 2.0 license, *it can be used without restrictions*.

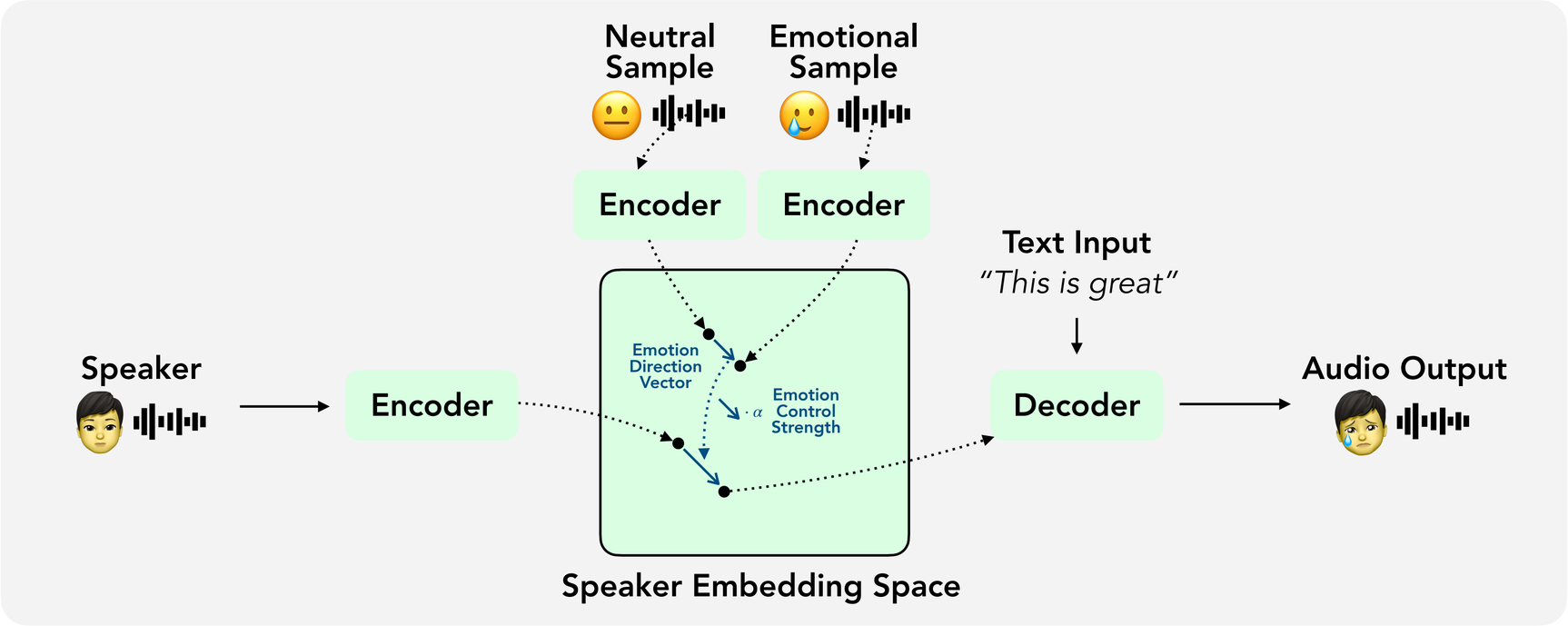
We base our work on MetaVoice (2024)

🎛️ EmoKnob is built on recent developments of voice cloning models



Voice clone models first encode speaker into an embedding and then decode speech with the speaker embedding

🎛️ EmoKnob repurposes voice cloning models for emotion control



EmoKnob takes embedding difference between emotional sample and neutral sample in speaker embedding space to extract emotion representation

🎛️ EmoKnob allows emotion strength control

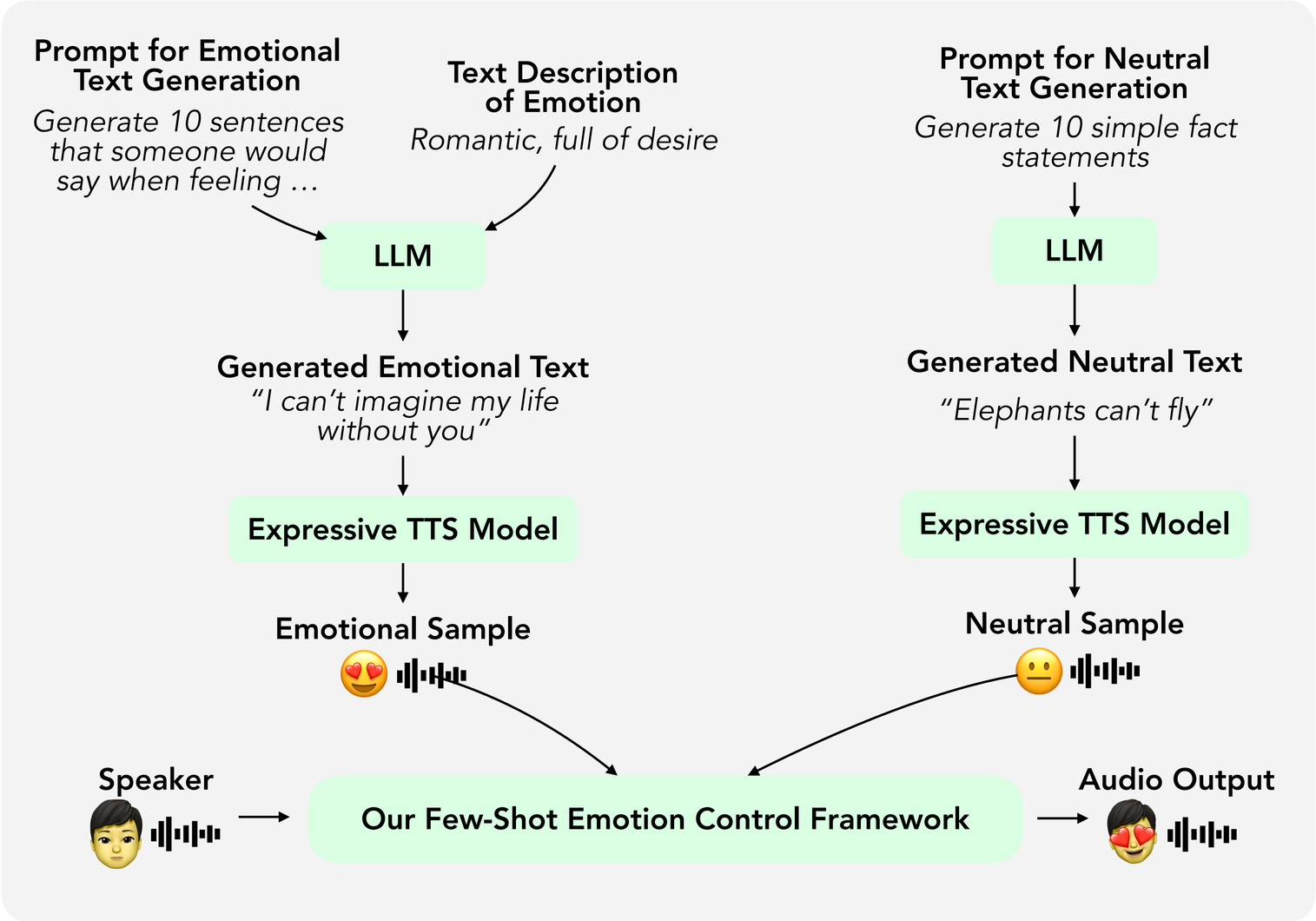
	Original Clone	Angry (0.1)	Angry (0.2)	Angry (0.3)	Angry (0.4)	Angry (0.5)
Sentence 1	▶	▶	▶	▶	▶	▶
Sentence 2	▶	▶	▶	▶	▶	▶
Sentence 3	▶	▶	▶	▶	▶	▶
Sentence 4	▶	▶	▶	▶	▶	▶

🎛️ EmoKnob allows few-shot emotion extraction

Complex emotions often lack large datasets. We can control these emotions with as few as two pairs of samples.

Controlling emotion with synthetic data

While existing TTS models do not allow emotion control, we can extract emotion from them with emotion-matching text



Controlling emotion with synthetic data

Allow us to control a wide range of emotions that lack existing data

	Original Clone	Desire (0.1)	Desire (0.2)	Desire (0.3)	Desire (0.4)	Desire (0.5)	Desire (0.6)
Sentence 1	▶	▶	▶	▶	▶	▶	▶
Sentence 2	▶	▶	▶	▶	▶	▶	▶
Sentence 3	▶	▶	▶	▶	▶	▶	▶
Sentence 4	▶	▶	▶	▶	▶	▶	▶

Controlling emotion with synthetic data

Allow us to control a wide range of emotions that lack existing data

	Original Clone	Sarcasm (0.1)	Sarcasm (0.2)	Sarcasm (0.3)	Sarcasm (0.4)	Sarcasm (0.5)	Sarcasm (0.6)
Sentence 1	▶	▶	▶	▶	▶	▶	▶
Sentence 2	▶	▶	▶	▶	▶	▶	▶
Sentence 3	▶	▶	▶	▶	▶	▶	▶
Sentence 4	▶	▶	▶	▶	▶	▶	▶

Future work

How can we control emotions without neutral sample?

🎛️ **EmoKnob** Enhance Voice Cloning with Fine-Grained Emotion Control

See more details at emoknob.cs.columbia.edu