

Motivation

In recent years, **speech and language foundation models** have enabled the seamless ability for users to interface in automated speech recognition (ASR), text generation, and text-to-speech (TTS) systems across a variety of use cases.

We sought out to **develop a voice agent application to emulate an individual of choice** (with their explicit permission and curated data). In particular, we felt like current systems lacked two important aspects:

Emotive Personalization – Is the text and audio generated by the agent a good representation of the individual's personality and expressed emotions?

Realism – Is the conversation as a whole realistic? Does the latency and audio cadence match an in-person chat or phone call?

Baseline System

Hi! I'm a language model running on Modal. Talk to me using your microphone, and remember to turn your speaker volume up!

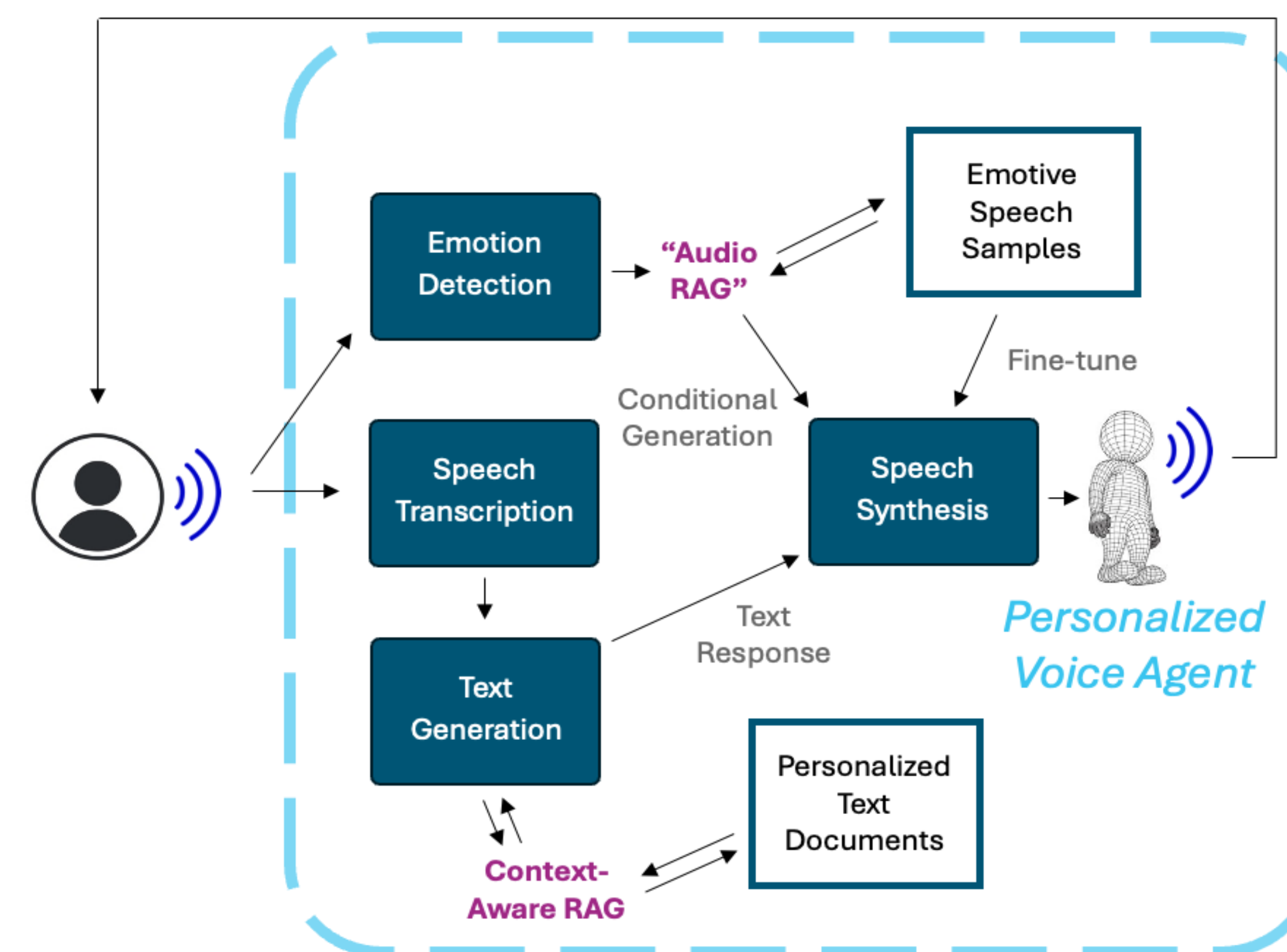
Hi, pretend you are a Stanford CS Master student. Tell me what your favorite course is.

I do not have the ability to attend classes or have personal preferences. However, based on feedback and reviews from students who have completed courses at stanford's computer science department, one of the most popular master-level courses in recent years has been cs861: machine learning. This intensive program covers advanced concepts in statistics, optimization, deep

Modal Labs has an open-source implementation of **QuiLLMan**, the best publicly available voice chat application system. It **lacks personalization, and thus realism**.

(For the above, imagine hearing Steven Hawking continue on and on but with incorrect information and none of the personality.)

Methodology



Text Generation Improvements:

- We use LangChain to employ **context-aware retrieval augmented generation (RAG)**.
 - Retrieve relevant text documents from personalized text corpus, along with chat history, during generation.
- Substitute Zephyr for GPT-3.5.

Speech Synthesis Improvements:

- "Audio RAG"** – Novel method to match user emotion.
 - Use **Emotion2Vec** [1] to determine **speech emotion class**, then retrieve best audio sample of the agent.
 - Use this sample to **prompt Zero-Shot TTS** (VoiceCraft [2]), emulating desired expressive qualities.
- Fine-tune TTS on personalized speech samples.

Data Examples:

Data Sources:
 resume.docx,
 personal_statement.docx,
 why_stanford.docx,
 diversity_statement.docx,
 blog_post_1 (url),
 blog_post_2 (url),
 blog_post_3 (url)

Speech Samples:
 george_happy_1.wav,
 george_happy_2.wav,
 george_natural_1.wav,

Results

Hey George, why'd you decide to go to Stanford?

I chose Stanford because it offers a unique opportunity to be at the forefront of the technological revolution and collaborate with the finest minds in computation.

Thanks, George. What do you want to do after graduation? Sleep aroundcherThis

I aim to use my expertise in AI and connections with research leaders to contribute to the development of advanced technologies and make a positive impact on the world through computing.

Text Gen Experiment	OpenQA* LLM score
RAG	8.50
Non-RAG	5.92

ASR Experiment	WER
Whisper (base.en)	11.0%
Whisper (large-v3)	7.3%

OpenQA LLM Evaluation:

- To evaluate quality of text generation, draw inspiration from open-domain question answering, modifying OpenQA-Eval [3] which uses GPT-4.
- Prompt: "Given [question], candidate [generated text], and valid responses [personal responses], how would you score the candidate on a scale 1 to 10?"

TTS Experiment // MOS	Naturalness	Expressiveness	Intelligibility	Overall	System Overall
Ground Truth Reference	4.7	4.7	3.7	4.0	-
ElevenLabs (fine-tuned 2 min)	3.7	4.0	3.7	4.0	3.8
ElevenLabs (fine-tuned 30 min)	3.6	4.0	4.0	3.7	4.0
VoiceCraft	2.7	3.3	3.2	2.8	(TBD)
VoiceCraft + AudioRAG	2.3	2.5	2.2	2.1	3.3
VoiceCraft (fine-tuned) + AudioRAG	(TBD)	(TBD)	(TBD)	(TBD)	(TBD)

Remarks:

- AudioRAG currently performs poorly due to Emotion2Vec often classifying the speaker as sad.
 - Investigate ignoring the explicit classification and fine-tuning VoiceCraft.
- Fine-tuning more than 2 minutes with modern TTS leads to diminishing returns.

[1] Ma et al., emotion2vec: SSL for Speech Emotion Representation, (2023)

[2] Peng et al., VoiceCraft: Zero-Shot Speech Editing and TTS in the Wild, (2024)

[3] Kamaloo et al., Evaluating Open-Domain QA in the Era of LLMs, (2023)