Voice chat

Description

The latest incarnation of large language models (e.g. ChatGPT) generates convincing spoken responses to spoken input. You can talk (as well as type) to your AI, as if on a smartphone. Itâ??s called â??voice chat,â?• and I used OpenAIâ??s new voice chat app on my smartphone to enter into a conversation clarifying certain aspects of the technology.

I also wanted to explore what the technology says about the philosopher Jacques Derridaâ??s views of the relationship between text and speech, i.e. words as written and words as spoken. Iâ??ll leave that till my next post.

Voice chat work flow

Voice chat draws on several technologies, melded to produce near seamless interaction. You speak into the app, which then translates, invisibly, your words into text. The text is fed into ChatGPT which processes the text in the usual way to generated a text response. That text is then converted by a high quality text-to-speech (TTS) algorithm to produce spoken output.

This flow happens quickly enough to deliver conversation-style interaction. The content of the responses from this voice chat app are equivalent to the responses you get when you type text-based prompts. The system also retains a transcript for further reference and editing.

Of note, the speech-to-text (TTS) component of the voice chat model is trained on voice data.

> Modern TTS systems, such as those using deep learning techniques, are capable of generating speech from text on the fly. These systems are trained on large datasets of recorded speech and their corresponding text transcriptions. The neural networks learn to map text inputs to audio outputs, capturing nuances like intonation, stress, and rhythm to produce natural-sounding speech.

> For real-time applications, optimized models and efficient hardware processing ensure that TTS can generate audio quickly enough to be used in interactive systems, such as virtual assistants and conversational AI platforms.

Intonation, emphasis, and rhythm

After a few interactions it became clear that, at least in default mode, voice chat pays little attention to the to-and-fro dynamics of voice-based conversation. Replies to prompts from the system are highly articulate, delivered with expert detail and alacrity. Thatâ??s somewhat intimidating as I would hesitate while seeking the best word sequence for my next prompt. The app recognised this shortcoming:

> [t]he text-to-speech (TTS) algorithm does not modify the content of the output. It converts the generated text into spoken words while preserving the original content. The TTS system focuses on delivering the text in a natural and intelligible manner, adding appropriate intonation, emphasis, and rhythm, but it does not change the wording or meaning of the text provided by the language model.

The use of pauses, the occasional â??erâ?• and the appâ??s use of tone and emphasis helps reduce the intimidation a little.

Attention

The transformer model underlying ChatGPT claims substantial benefit from its so-called *attention* mechanisms which informâ??s how the model places its semantic emphasis in both training data, and its real-time interactions. See post Multi-head attention. But this mechanism apparently plays no part in the text to voice processing.

> The emphasis in the voice responses isnâ??t directly derived from the attention mechanisms in the transformer model. The attention mechanism in transformers is used for understanding and generating text based on contextual importance, but it doesnâ??t dictate how the text is spoken.
>
> The emphasis in spoken responses typically comes from the text-to-speech (TTS) system, which uses its own algorithms to make the speech sound more natural. The TTS system interprets punctuation, word choice, and sentence structure to add appropriate intonation and emphasis.

Nor will voice chat change its method of intonation, emphasis and rhythm on request. It would not read a made-up Dr Zeuss story as if reading to a small child. Nor would it speak like an Australian!

> Text-to-speech systems can sometimes be customized to reflect specific regional accents and intonations, but it depends on the capabilities and settings of the particular TTS system in use. For a more personalized experience, some advanced TTS systems allow for adjustments to better mimic specific accents and speech patterns, including the distinctive

rising intonation of Australian English.

Speech training

Of note, large language models such as ChatGPT are trained not on spoken words, but on written texts. That sourcing is represented in the content of the voice chat responses.

As mentioned above, TTS systems also undergo training from voice data, but that is independent of the LLM training. The app and I discussed what it would be like to train an LLM just on voice data rather than text, analogous to the way infants acquire language competence. Thatâ??s beyond the current capabilities of LLMs, and is for a later discussion.

Note

The heavily cropped featured image is a response to my voice chat request to illustrate this blog. â??Hereâ??s the revised photorealistic steampunk banner image for your blog about ChatGPTâ??s voice feature, incorporating themes from Jacques Derridaâ??s philosophy.â?•

Category

1. Artificial Intelligence

Tags

1. conversation
2. Derrida
3. text to speech
4. voice chat

Date Created
May 25, 2024
Author
rcoyne99