MY VOICE IS NO LONGER MY PASSWORD —

Microsoft’s new AI can simulate anyone’s voice with 3 seconds of audio

Text-to-speech model can preserve speaker's emotional tone and acoustic environment.

BENJ EDWARDS - 1/9/2023, 11:15 PM

Ars Technica

Enlarge / An AI-generated image of a person's silhouette.

On Thursday, Microsoft researchers announced a new text-to-speech AI model called VALL-E that can closely simulate a person's voice when given a three-second audio sample. Once it learns a specific voice, VALL-E can synthesize audio of that person saying anything—and do it in a way that attempts to preserve the speaker's emotional tone.

FURTHER READING
Meta’s AI-powered audio codec promises 10x compression over MP3

Its creators speculate that VALL-E could be used for high-quality text-to-speech applications, speech editing where a recording of a person could be edited and changed from a text transcript (making them say something they originally didn't), and audio content creation when combined with other generative AI models like GPT-3.

Microsoft calls VALL-E a "neural codec language model," and it builds off of a technology called EnCodec, which Meta announced in October 2022. Unlike other text-to-speech methods that typically synthesize speech by manipulating waveforms, VALL-E generates discrete audio codec codes from text and acoustic prompts. It basically analyzes how a person sounds, breaks that information into

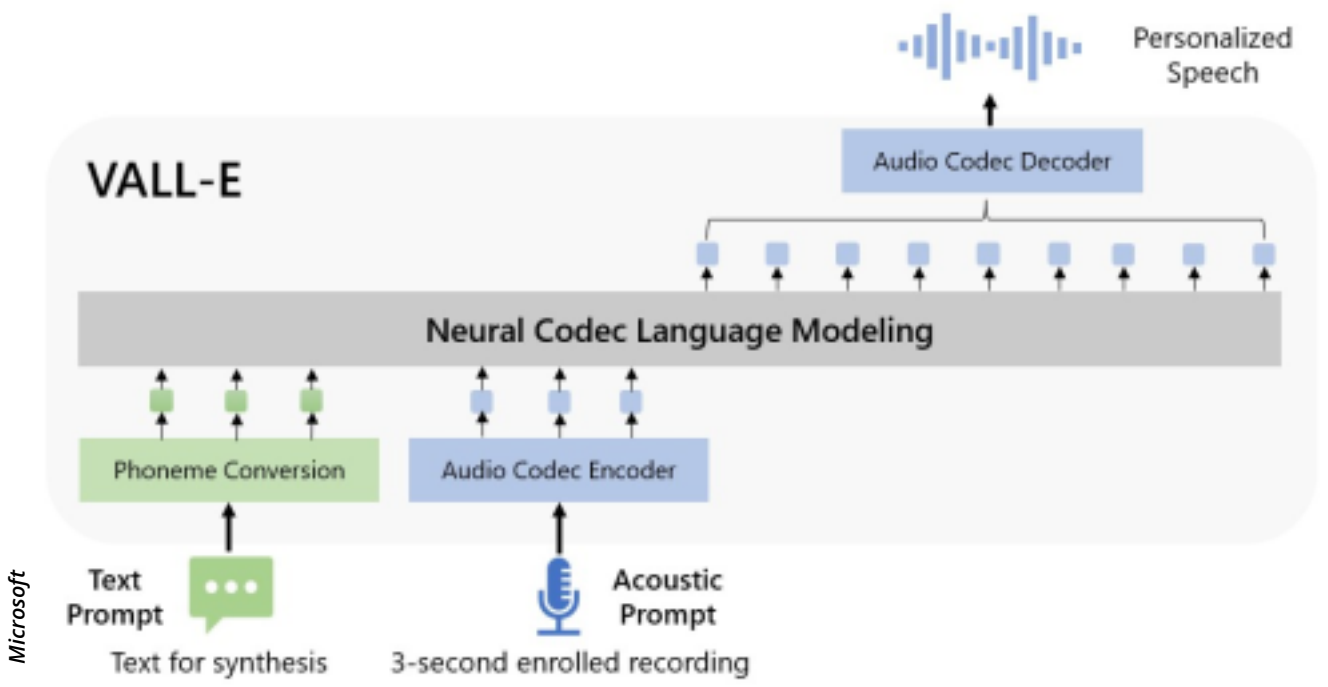
discrete components (called "tokens") thanks to EnCodec, and uses training data to match what it "knows" about how that voice would sound if it spoke other phrases outside of the three-second sample. Or, as Microsoft puts it in the VALL-E paper:

> To synthesize personalized speech (e.g., zero-shot TTS), VALL-E generates the corresponding acoustic tokens conditioned on the acoustic tokens of the 3-second enrolled recording and the phoneme prompt, which constrain the speaker and content information respectively. Finally, the generated acoustic tokens are used to synthesize the final waveform with the corresponding neural codec decoder.

Microsoft trained VALL-E's speech-synthesis capabilities on an audio library, assembled by Meta, called LibriLight. It contains 60,000 hours of English language speech from more than 7,000 speakers, mostly pulled from LibriVox public domain audiobooks. For VALL-E to generate a good result, the voice in the three-second sample must closely match a voice in the training data.

FURTHER READING
With Koe Recast, you can change your voice as easily as your clothing

On the VALL-E example website, Microsoft provides dozens of audio examples of the AI model in action. Among the samples, the "Speaker Prompt" is the three-second audio provided to VALL-E that it must imitate. The "Ground Truth" is a pre-existing recording of that same speaker saying a particular phrase for comparison purposes (sort of like the "control" in the experiment). The "Baseline" is an example of synthesis provided by a conventional text-to-speech synthesis method, and the "VALL-E" sample is the output from the VALL-E model.



Microsoft

Enlarge / A block diagram of VALL-E provided by Microsoft researchers.

While using VALL-E to generate those results, the researchers only fed the three-second "Speaker Prompt" sample and a text string (what they wanted the voice to say) into VALL-E. So compare the "Ground Truth" sample to the "VALL-E" sample. In some cases, the two samples are very close. Some VALL-E results seem computer-generated, but others could potentially be mistaken for a human's speech, which is the goal of the model.

In addition to preserving a speaker's vocal timbre and emotional tone, VALL-E can also imitate the "acoustic environment" of the sample audio. For example, if the sample came from a telephone call, the audio output will simulate the acoustic and frequency properties of a telephone call in its synthesized output (that's a fancy way of saying it will sound like a telephone call, too). And Microsoft's samples (in the "Synthesis of Diversity" section) demonstrate that VALL-E can generate variations in voice tone by changing the random seed used in the generation process.

FURTHER READING
Darth Vader's voice will be AI-generated from now on

Perhaps owing to VALL-E's ability to potentially fuel mischief and deception, Microsoft has not provided VALL-E code for others to experiment with, so we could not test VALL-E's capabilities. The researchers seem aware of the potential social harm that this technology could bring. For the paper's conclusion, they write:

"Since VALL-E could synthesize speech that maintains speaker identity, it may carry potential risks in misuse of the model, such as spoofing voice identification or impersonating a specific speaker. To mitigate such risks, it is possible to build a detection model to discriminate whether an audio clip was synthesized by VALL-E. We will also put Microsoft AI Principles into practice when further developing the models."

READER COMMENTS 154

BENJ EDWARDS
Benj Edwards is an AI and Machine Learning Reporter for Ars Technica. In his free time, he writes and records music, collects vintage computers, and enjoys nature. He lives in Raleigh, NC.

between all those time periods and people or did he simply imagine it all? What do people in the waiting room do while Sam is in their bodies? What happens to Sam's loyal ally Al? 30 years following the series finale, answers to these mysteries and more await.

← PREVIOUS STORY

NEXT STORY →

Related Stories

Today on Ars

STORE
SUBSCRIBE
ABOUT US
RSS FEEDS
VIEW MOBILE SITE

CONTACT US
STAFF
ADVERTISE WITH US
REPRINTS

NEWSLETTER SIGNUP

Join the Ars Orbital Transmission mailing list to get weekly updates delivered to your inbox. Sign me up →