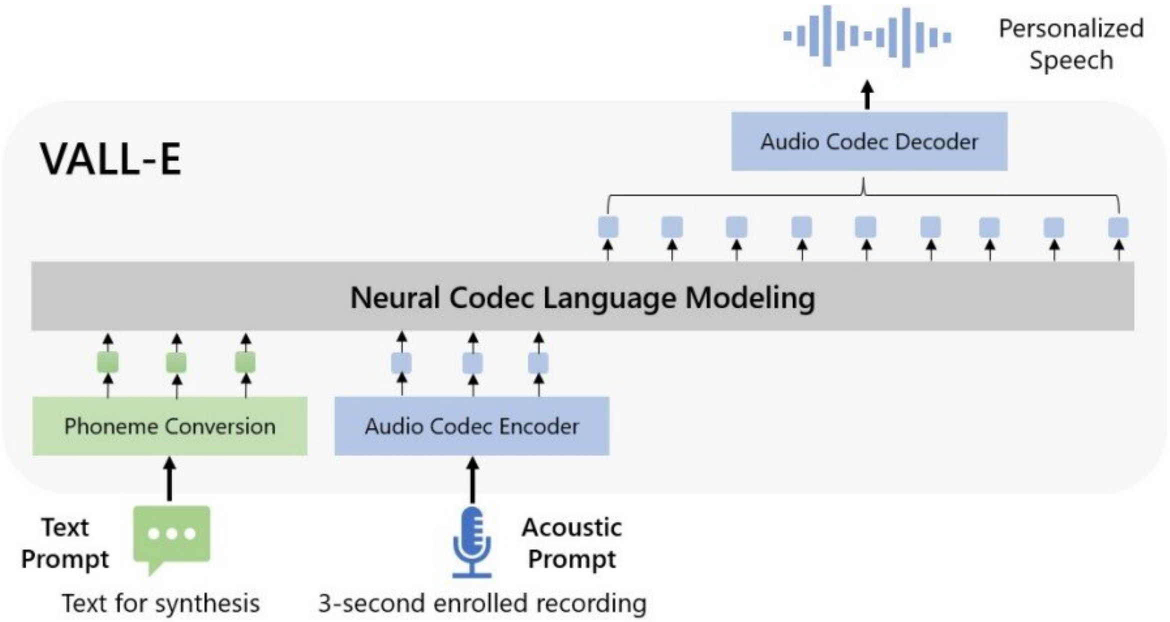
Microsoft's VALL-E can faithfully reproduce a voice after listening to a three second recording

January 11 2023, by Bob Yirka



The overview of VALL-E. Unlike the previous pipeline (e.g., phoneme → mel-spectrogram → waveform), the pipeline of VALL-E is phoneme → discrete code → waveform. VALL-E generates the discrete audio codec codes based on phoneme and acoustic code prompts, corresponding to the target content and the speaker’s voice. VALL-E directly enables various speech synthesis applications, such as zero-shot TTS, speech editing, and content creation combined with other generative AI models like GPT-3 [Brown et al., 2020]. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2301.02111

A team of researchers at Microsoft has demonstrated a new AI system that is capable of mimicking a person's voice after training with a recording just three seconds long. The team explains developing the new app in a paper published on the *arXiv* preprint server. They have also posted a webpage demonstrating the app's capabilities.

Artificial intelligence applications require training on massive amounts of data. But in this new endeavor, the team at Microsoft has shown that does not always have to be the case.

The new app was built using Meta's EnCodec audio compression technology, and was originally intended as a way to improve the quality of phone conversations. Subsequent work showed that it is capable of far more—not only can it mimic a voice, it can also simulate tone and even the acoustics of the environment in which the original recording was made.

Microsoft did not do away with the need for a massive data set, of course; instead, the researchers shifted where it was used. The app was taught to "listen" to a string of words and then to replicate its sound using Meta's Libri-light dataset, which has over 60,000 hours of recordings made by 7,000 people speaking in English.

The examples Microsoft has provided demonstrate that the system works much better for some voices than others, and it has trouble with accents. But because the app is still in its early stages, it is likely its functionality will improve over time.

Microsoft has not made the source code for VALL-E public and likely will not do so, noting that it could be used in less than responsible ways—hoax recordings of politicians, for example. When combined with deepfake video, the results could take "fake news" to new heights. Microsoft's example has shown what is possible; thus, it would seem likely that similar systems by others will appear soon.

More information: Chengyi Wang et al, Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, *arXiv* (2023). DOI: 10.48550/arxiv.2301.02111

Citation: Microsoft's VALL-E can faithfully reproduce a voice after listening to a three second recording (2023, January 11) retrieved 25 April 2024 from https://techxplore.com/news/2023-01-microsoft-vall-e-faithfully-voice.html