SDS PODCAST EPISODE 648

FIVE MINUTE FRIDAY: VALL-E: UNCANNILY REALISTIC VOICE IMITATION FROM A 3-SECOND CLIP

(00:03):

This is Five-Minute Friday on VALL-E, the incredible and frightening new voice-imitation model.

(00:19):

Text-to-speech models are models that take in text as an input, e.g., a sentence that you type and provide the model, and then these text to speech models output an audio waveform that sounds like a human reading out the sentence you provided as an input. So you provided the text and it outputs an audio waveform in human voice. Text to speech systems like this have been around for decades, but until the past few years the quality of the audio was not compellingly human-like. Five years ago, in 2018, Google stunned attendees at its Google I/O conference with an algorithm called Google Duplex that marked a step change in the quality of TTS: Initially capable of making restaurant reservations, Duplex sounded compelling human-like because of its capacity to "um" and "uh" and stammer like humans do when they engage in natural, unscripted conversation. We've included a link in the show notes if you'd like to listen to examples of the high-quality, human-like Google Duplex audio.

(01:22):

But Duplex is not the focus of the show. The focus of today's episode is VALL-E, which was released earlier this month, by Microsoft. It's another "text-to-speech" model and this VALL-E term is spelt the same was as OpenAI's popular DALL-E series of text-to-image models, so it's VALL-E, in all caps, it's supposed to look like WALL-E, the Pixar robot in terms of spelling. I couldn't find an explanation of why they called it VALL-E specifically. The DALL-E model from OpenAI makes sense because it sounds like Salvador Dali the artist, and it generates art, but I'm not sure about VALL-E, I have this idea that could be like Frankie Valli in the Four Seasons but I think probably more likely it's related to the "uncanny valley"

concept that refers to the unpleasant reaction that humans have to machines that closely mimic human capabilities, so I don't know, that's my guess.

(02:22):

Anyway, it's called VALL-E, spelled V-A-L-L-E in all caps. Relative to baseline Text to Speech models, VALL-E doesn't produce game-changing audio quality of human voices — as Duplex illustrated, we've had human-level Text to Speech for five years, so what's the big deal with VALL-E? Well, what VALL-E does do is that in addition to a text prompt, you can also provide it with just three seconds of a recording of someone's voice and it will generate audio that is compellingly in the style of that recorded-person's voice. Just three seconds! That's all it takes. So that should maybe already have your brain cells worrying about why that could be worrying.

(03:06):

Before we get to the concerns of this, to illustrate how cool and effective this new VALL-E model is, here are some examples of VALL-E outputting a sentence from the classic historical-romance novel by James Fenimore Cooper called The Last of the Mohicans. So the quote that we provide as input, is text for VALL-E to emulate is "Notwithstanding the high resolution of hawkeye, he fully comprehended all the difficulties and danger he was about to incur." Ok, so that's the sentence. "Notwithstanding the high resolution of hawkeye, he fully comprehended all the difficulties and danger he was about to incur."

(03:42):

So we provide that as text into the model and alongside that typed text that we provide as an input, we also provide three seconds of audio of someone speaking. So here's one example of an input prompt: "The lodge in which Uncas was confined was in the very center of the"

(04:00):

And here's VALL-E's imitation of that speaker's style, but outputting the natural language of the Last of the Mohicans quote instead of the input, the audio's input of whatever it was saying: "Notwithstanding the high resolution of hawkeye, he fully comprehended all the difficulties and danger he was about to incur."

(04:16):

Pretty amazing, right? Here's a second speaker's style:

(04:24):

And now here's VALL-E's imitation of that style, again outputting the Mohicans quote: "Notwithstanding the high resolution of hawkeye, he fully comprehended all the difficulties and danger he was about to incur."

(04:37):

Ha! And third time's the charm. Here's one final speaker style input: "Just like at babies. And she has the same three frackles on her..."

(04:46):

And here again is VALL-E's output of the Mohicans quote in the third speaker's style: "Notwithstanding the high resolution of hawkeye, he fully comprehended all the difficulties and danger he was about to incur."

(04:59):

Wow! Pretty cool, right? To get examples of how VALL-E performs relative to previous state-of-the-art baselines, you can refer to VALL-E demo GitHub link in the show notes. So that has examples of other kinds of models that were trying to do the same kind of thing in the past, taking in text that you want to output as well as the speaker's style audio and then outputting

your input text in the style of whatever speaker style. So yeah, people have been trying to do that before, and you can see from the link that's in the show notes that previous models previous state-of-the-art was nowhere near as good as VALL-E is.

(05:42):

Having heard now how amazingly realistic and accurate VALL-E's outputs are based on just a three-second sample of someone speaking, perhaps your next was how scary that is. If a scam artist has access to VALL-E and just a three-second clip of you speaking, they could use it to send recordings or perhaps even generate responses in close-to-real-time with a loved one or colleague of yours, convincing them that it's you that they're dealing with. As an example, if you got a voicemail from your boss telling you to buy gift cards from an electronics store and to provide her with the unique gift card code on the back, would you do it? Well, maybe you'd be suspicious and phone them because you're up to date on the state-of-the-art in A.I., but a lot of people out there could be had by such a scam. As another example, if you received a voicemail from a loved one saying they were in prison and you need to wire bail to a specific crypto wallet… you yourself might be suspicious but a lot of folks out there could be conned.

(06:42):

So certainly there are ethical concerns here, but this is a world we're going to have to get used to. Generative A.I. capabilities across images, video, and audio are becoming increasingly compelling, increasingly realistic. Thankfully, there are solutions, technology actually offers solutions. For example, while I wouldn't recommend that you purchase NFT art, non-fungible tokens could be used to helpfully verify that a media file was genuinely created by a trusted source.

(07:11):

Ok, now that you know what VALL-E is as well as its potentially dangerous implications, for you SuperDataScience super nerds out there, there are a few key points on how Microsoft trained and architected their VALL-E model. They used a hybrid model training approach that blended supervised learning on 960 hours of labeled speech data, they blended that with unsupervised learning on a much much larger data set — more than 60x larger — of 60,000 hours of unlabelled training data from around 7000 different human speakers. So huge unlabeled training data set, more than 60,000 hours and then a much smaller training data set that is labeled. So, this kind of hybrid machine learning approach that blends supervised learning on labeled speech data with unsupervised learning on unlabeled training data allows us to take advantage of large unlabeled training data sets such as this, and incorporate some of that information, some of the detail and diversity of that much larger data set than if we just had the smaller supervised leaning set.

(08:20):

In terms of model architecture, the VALL-E creators used a transformer architecture with 12 layers, 16 attention heads, and a 1024-dimensional embedding space. To train this large language model efficiently on the huge amounts of data at their disposal, they used 16 NVIDIA Tesla V100 GPUs with 32 GB of memory each. For more details than that, you can check out the full paper via ArXiV — we've got a link for you in the show notes.

(08:49):

And, if you'd like to learn more about transformer architectures and attention like VALL-E takes advantage of, coming up on March 1st, I'll be hosting a virtual conference on natural language processing with large language models like BERT and the GPT series architectures and yes, VALL-E. It'll be interactive, practical, and it'll feature some of the most influential scientists and instructors in the large natural language model space as speakers. It'll be live in the O'Reilly platform, which many

employers and universities provide access to; otherwise you can grab a free 30-day trial of O'Reilly using our special code SDSPOD23. We've got a link to that code ready for you in the show notes as well.

(09:30):

All right. This has been a great one, a fun little technical episode Five-Minute Friday here, I hope you enjoyed and learned a lot. Until next time, keep on rockin' it out there, folks, and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.