CO.DESIGN
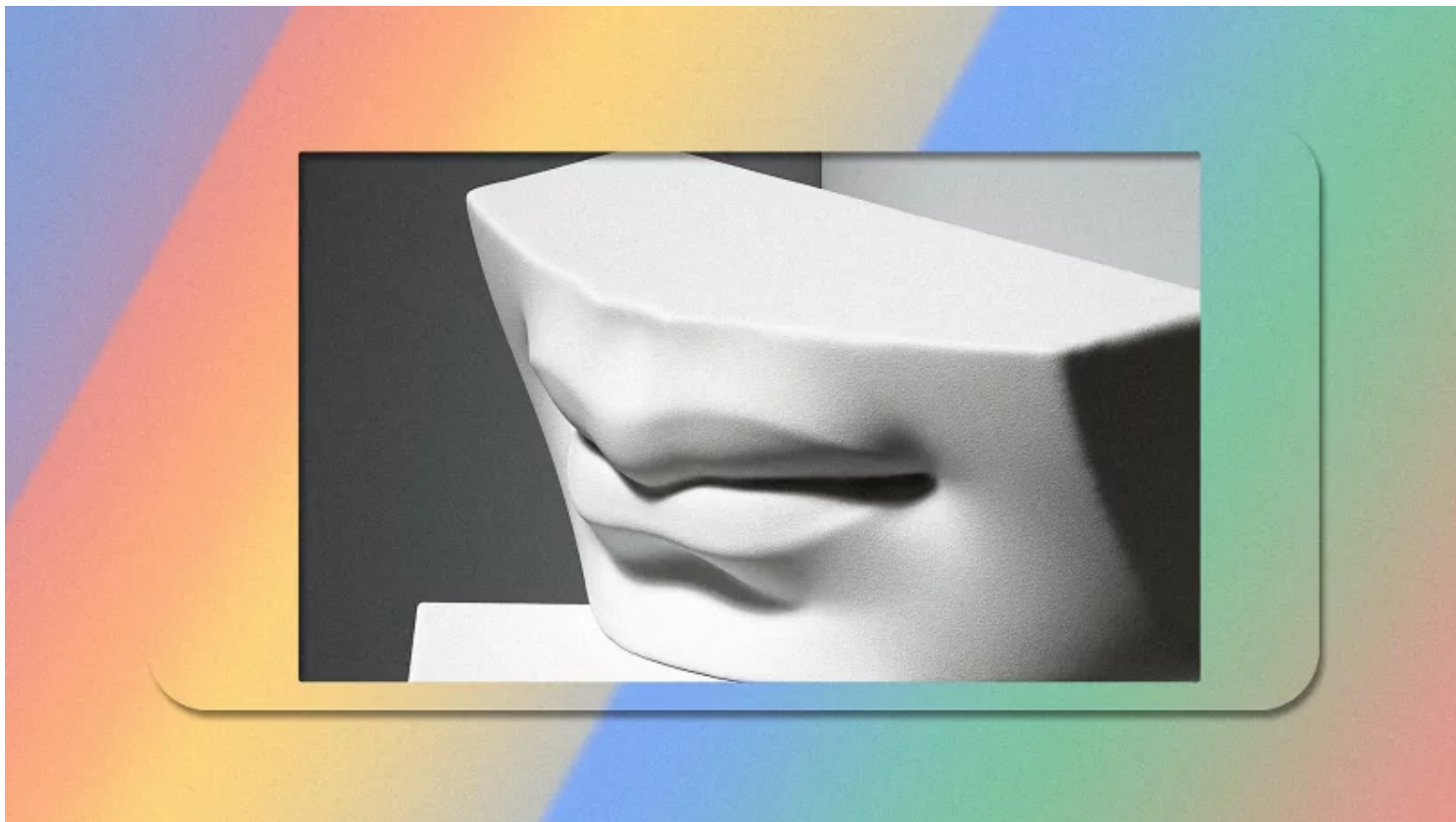
2 MINUTE READ

Google Can Finally Understand Your Drunk Texts

"Thisss isrealy awhsome."

[Photo: wacomka/iStock]

BY JOHN PAVLUS

04.21.17 | 7:30 AM

Google would love to have the voices that come out of its products sound as natural as Samantha, the near-future AI who lived in Joaquin Phoenix's earpiece in the movie *Her*. To achieve this effect, you could hire a dulcet-voiced actor to record all the words, phrases, and phonemes you'd ever need and write sophisticated software to recombine them on the fly . . . *like a chump*. Or you could automate the whole enchilada with deep learning. Google being Google, you can imagine what path they chose. Meet Tacotron: It's no ScarJo, but it can say "basilar membrane and otolaryngology are not auto-correlations" better than you ever will.

Granted, that's a phrase you've never seen before and will almost certainly never need to hear again. But that's the whole point: Tacotron has never seen the phrase before either, yet it can dance through the complicated pronunciation like a speech-synthesizing Fred Astaire. Tacotron handles prosody (the musical "tone" of speech), semantic disambiguation (e.g., saying "read" differently in the present or past tense), and drunk-text-esque spelling errors ("Thisss isrealy awhsome") with surprising aplomb.
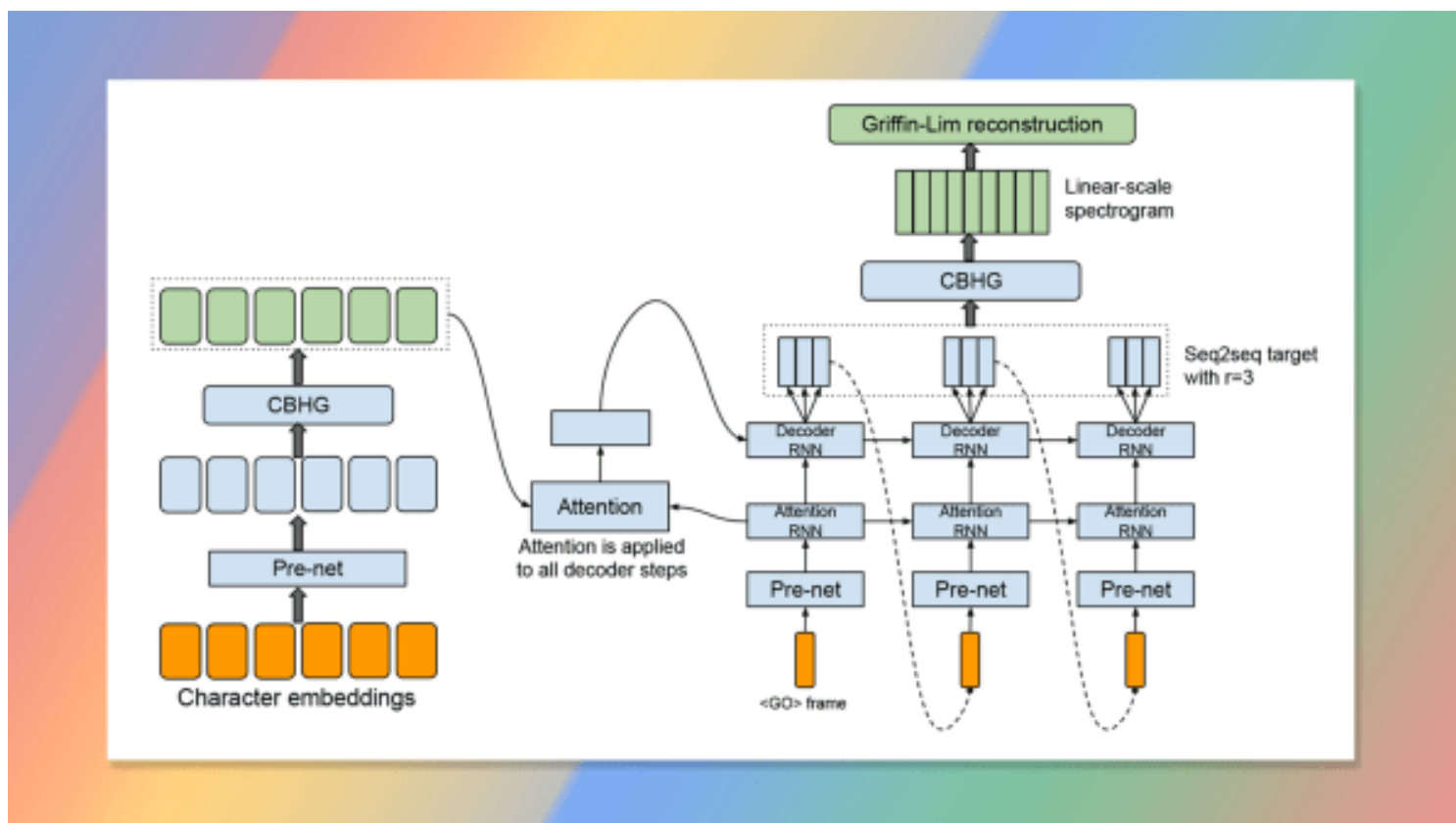
[Image: Google]

In its technical report, Google admits that concatenative speech synthesis—the aforementioned "chump" approach, employed famously by Apple's Siri—actually delivers *more* natural-sounding results than Tacotron. But concatenation is tedious, expensive, and involves "brittle design choices," to quote Google's report. It's this brittleness that Tacotron is specifically engineered to circumvent, says William Wang, a natural language processing expert at the University of California, Santa Barbara who's familiar with the research. Programming a speech synthesizer by hand involves "making many design choices that are all very arbitrary," Wang explains.

Take prosody as an example. The natural-sounding differences in pronunciation between a statement ("The quick brown fox jumps over the lazy dog.") and a question ("Does the quick brown fox jump over the lazy dog?") come down to subtle shifts in tone. How, exactly, should those shifts be encoded in software? Someone has to decide in advance—and while their choices may deliver natural-sounding prosody in certain sentences, the same choices may produce awkward results in others. There's no way to tell in advance, and no way to account for errors and edge cases except with more hand-coding across multiple components of the system.

[Image: Google]

Tacotron uses deep-learning software (commonly referred to as "AI") to capture and automate all that decision-making. It's not best-in-class for "natural-ness": Google's own WaveNet technology, which also uses deep learning to synthesize speech, claims to beat even concatenative software in terms of quality. But Tacotron is *faster* than WaveNet, and simpler too: It handles the entire text-to-speech processing pipeline "using a single neural network architecture," according to Wang. That makes Tacotron much easier to train on Google's ever-swelling galaxy of text and voice data.

Google is characteristically silent about what, if any, plans they have to apply Tacotron to its current products (the researchers did not respond to repeated interview requests, and a spokesperson declined to comment on the record). But if you're someday able to summon one of Google's self-driving cars with little more than a drunkenly typed text message—and have it speak back to you without sounding like Robby the Robot—it might just be Tacotron doing the heavy lifting.

ABOUT THE AUTHOR

John Pavlus is a writer and filmmaker focusing on science, tech, and design topics. His writing has appeared in Wired, New York, Scientific American, Technology Review, BBC Future, and other outlets. More

---

Co.Design Daily Newsletter

| YOUR EMAIL ADDRESS | SIGN UP |

☑ Receive special Fast Company offers   See All Newsletters

VIDEO

This Online Dictionary Will Help You Understand Tricky Tech Jargon

Sideways Dictionary uses analogies to help describe tech terms you hear all the time but might …

▶