Say again? AI provides the latest word in clearer audio

18 December 2020, by Molly Sharlach

Credit: Unsplash/CC0 Public Domain

If you've been listening to more podcasts while stuck at home this year, you may have noticed a side effect of the uptick in virtual conversations: a decline in audio quality. Interviews conducted by phone or video chat often include background noise, reverberation and distortion.

Now, a new method developed in part by researchers at Princeton University could improve the listening experience in the COVID era and beyond. Using an artificial intelligence (AI) approach known as deep learning, the technique can transform low-quality recordings of human speech, approaching the crispness and clarity of a studio-recorded voice.

While other AI-based methods for improving speech recordings have generally tackled a single aspect of audio quality, such as filtering out background noise or removing reverb, this method is more of an all-in-one tool. Ultimately, the researchers hope to apply their framework to enable fully automated, real-time speech enhancement.

"Previous approaches have mostly focused on improving the intelligibility of speech, but these can make the listening experience flatter, so the resulting quality is not that great for listening," said Jiaqi Su, a graduate student in computer science and lead author of a paper describing the method, which the researchers call HiFi-GAN.

HiFi-GAN uses artificial neural networks, key tools of deep learning that mimic the interconnected architecture of biological neurons. In this system, two separate networks compete to improve audio quality. One network, called a generator, produces cleaned-up recordings of speech. The other network, called a discriminator, analyzes recordings to try to determine whether they are real studio-quality recordings or audio that has been cleaned by the generator. The competition between these generative adversarial networks (GANs) improves the method's ability to produce clear audio.

The generator and discriminator networks engage in a kind of arms race. "The generator's job is to try to fool the discriminator," said coauthor Adam Finkelstein, a professor of computer science. "The two of them ratchet their way up, each becoming more and more effective during training. When that process is complete, you can throw away the discriminator and what you have is an awesome generator."

To evaluate the recordings generated by HiFi-GAN, the researchers used several objective measures of audio quality. They also turned to the crowdsourcing platform Amazon Mechanical Turk to collect subjective judgments from human listeners, who rated HiFi-GAN's results and those of other audio quality improvement algorithms. In 28,000 listener ratings of recordings on Amazon Mechanical Turk, HiFi-GAN scored higher than five other methods of improving audio quality.

"The issue we commonly observe in experiments is that objective metrics do not fully correlate with human perception, so it's very possible that your

method gets a higher score but it actually produces a worse listening experience. That's why we also conduct subjective evaluations," said Su.

In related work, Finkelstein's group and others developed an objective metric to detect and quantify subtle differences in audio recordings that are perceptible to the human ear but have been challenging for AI algorithms to handle. The metric, which is trained on roughly 55,000 human judgments collected on Amazon Mechanical Turk, could boost the performance of audio quality enhancers like HiFi-GAN, as well as more broadly aid the evaluation of deep learning methods for processing audio recordings.

The paper puts forth a new metric for machine learning tools that assess audio quality or compare audio recordings. The method builds on existing approaches to adversarial learning in which a generator and discriminator network compete to improve an algorithm's outputs. The metric can determine, for example, how close an AI-generated audio recording is to a reference, studio-quality recording.

"We wanted to find a perceptual metric that humans would relate to," said Pranay Manocha, a computer science graduate student and lead author of the research. "For example, if we play two recordings and then ask if they are exactly the same or different, our metric is able to give an answer that is correlated with judgments that humans would make."

While there are many such metrics in audio processing, the method improves on these by detecting small differences, which researchers call "just-noticeable," such as subtle changes in higher-frequency overtones that are not the main components of speech.

"Deep learning has already had a huge impact in audio processing, and we expect it to become even more profound" in the coming decade, said Finkelstein, "but there is a big problem, which is a little esoteric: For the machine to learn, it needs to know how well it's doing … it needs something called a loss function."

In designing a good loss function, "we need a fully automatic method to determine whether humans would say two audio clips sound similar to each other," said Finkelstein. "It's not practical to ask humans that question" while training a neural network, "because it would involve asking humans a gazillion questions while the algorithm searches for a good solution. So instead we are developing an automatic method to predict how humans would answer that question."

Su and Manocha presented papers describing these projects at INTERSPEECH, an international conference focused on speech processing and applications, which was held entirely virtually this October.

Both HiFi-GAN and the just-noticeable difference metric offer general approaches that can be used for a variety of audio processing tasks. The researchers are now adapting their methods toward real-time speech enhancement, which could potentially be used during a Zoom conversation or webinar.

The team is also adding a capability for bandwidth extension to HiFi-GAN. This will recreate the "sense of presence" listeners experience from recordings made at high sample rates, which is often missing from consumer-grade audio recordings and online conference calls, said Finkelstein.

Su, Finkelstein and others were coauthors of the paper "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks." Coauthors of the paper "A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences" included Manocha and Finkelstein.

More information: HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. pixl.cs.princeton.edu/pubs/Su_2020_HiFi/

A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences. pixl.cs.princeton.edu/pubs/Manocha_2020_ADP/

Provided by Princeton University

APA citation: Say again? AI provides the latest word in clearer audio (2020, December 18) retrieved 28 May 2022 from https://techxplore.com/news/2020-12-ai-latest-word-clearer-audio.html