Deepfakes

Deepfake - Audio

3 min

While impersonation was an art form left to comedians and actors, AI has quickly evolved to include audio creation (and impersonation) within its repertoire in recent years. Take, for example, hearing or seeing a clip where notable individuals, maybe musical artists or politicians, are having a discussion that would never actually occur. While humorous, this is a perfect example of a deepfake audio creation.

While generating audio is complex, we can take a high-level approach to break down these algorithms and gain a basic understanding of how they work.

When we discuss speech-generating algorithms, many of these algorithms fall under the category of a Text-to-Speech (TTS) system. That is, a system capable of converting text into a realistic human voice. While this process may sound simple, it's quite complex. Nvidia does a great job explaining the process in a bit more detail.

When we, as humans, read content or speak aloud, there is far more present than just the words spoken. Instead, we use careful timing, inflections, volume changes, and other essential characteristics to help emphasize and explain our ideas. Training a machine to do this takes a lot of work. Developing a TTS system is quite complex, but various open-source tools and datasets are designed to help with this process. These include LJ Speech and LibriTTS datasets and AI-focused libraries such as TensorFlow. Of course, even with these tools, creating a successful TTS model is time-consuming and costly.

The exact structure of any given AI model will vary, but most are built using either a "two-stage pipeline" or an "end-to-end pipeline."

In the "two-stage pipeline," data is first passed to a neural network capable of generating a spectrogram. The portion of the model will take in text and normalize it. Following this, the generated spectrogram is passed to a Vocoder neural network, which converts the spectrogram to the appropriate wavelengths, creating a realistic-sounding voice.

While the training and production of any AI algorithm is complex, it is not outside the grasp of malicious users. While rare, attacks of this nature have been increasing, especially against financial institutions. The exact complexity of these attacks varies, of course. While some tools are more complex, such as Microsoft's VALL-E, many synthetic voice algorithms must be revised. Due to either a lack of data samples or insufficient training, many models attackers use are less likely to fool an actual person.

Of course, there are successful cases of these attacks. For example 2019, a more sophisticated model was used to target a UK-based energy company. By leveraging a synthetic voice algorithm trained with the parent companies, CEO fraudsters convinced a company executive to send over 243,000 dollars to a Hungarian-based bank account.

By creating a careful ploy and combining that with a believable, deep fake of the CEO's voice, executives felt secure in sending money to a third party under the guise that repayment would be made swiftly by the energy firm's parent company. When the fraudsters called back a second time in an attempt to steal more funding, the organization realized this advanced vishing attack had targeted them.