CS410 Text Information System

Technology Review

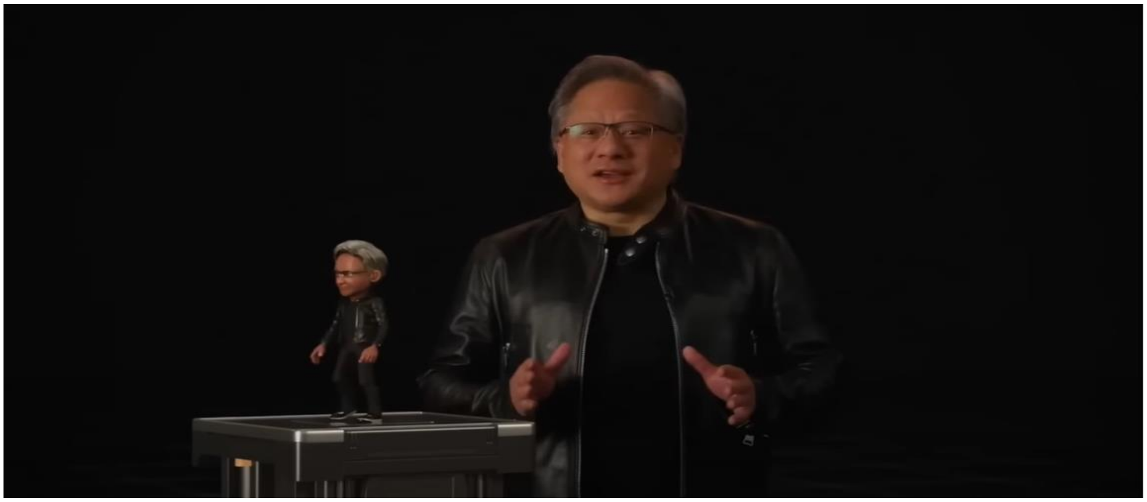
NetID: [chhavi2]
Name: Chhavi Jain

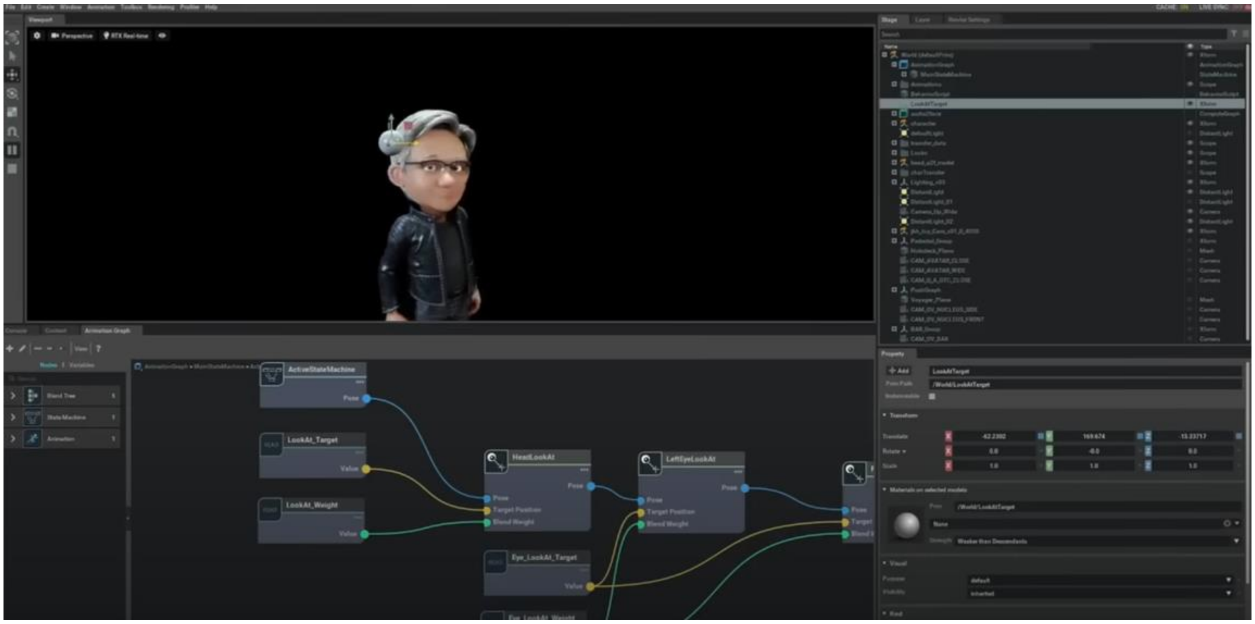
Topic- Cloning of human voice with AI, using very few samples and text-to-speech (TTS) model

Technology Review –

In "One TTS Alignment To Rule Them All" paper, released in oct 2022, Nvidia showcased how human voice can be cloned easily with very few samples. NVIDIA is offering voice cloning with only 30 minute voice recording. And using this 30 minute voice sample AI model can clone the real human voice across emotions, across languages and with very good quality.

In NVIDIA's keynote, Jensen Jr., an AI-powered virtual assistant of NVIDIA's CEO, Jensen Huang. Cloned Jensen's voice and it even generated hand gestures that can go well with explanation.





These virtual AI assistants are going to appear everywhere to help us with our daily tasks.

For instance, in our car, the promise is that AI will be able to recognize the owner of the car, recommend shows nearby, and even drive the user there.

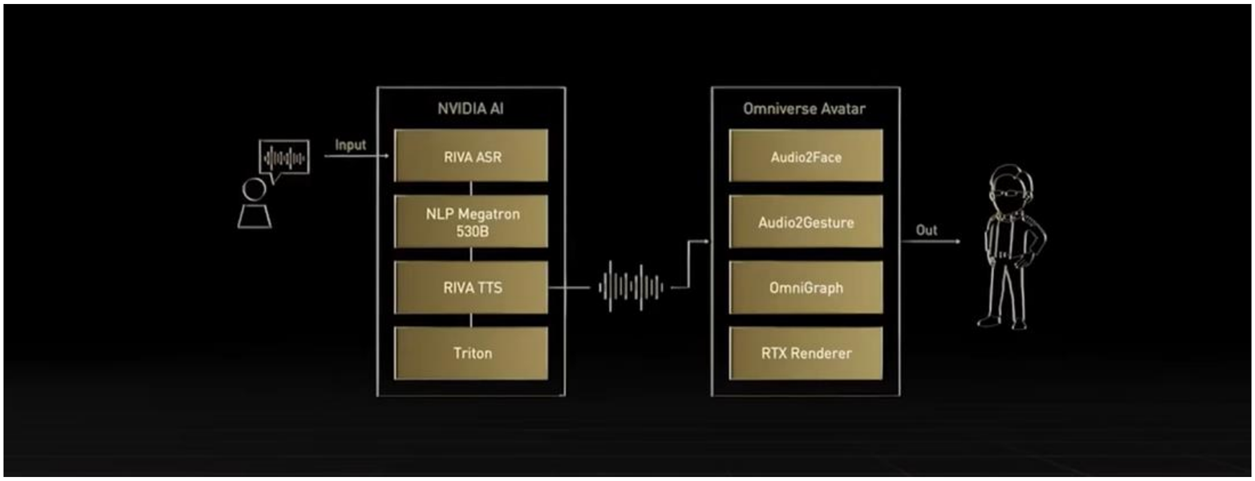
These Omniverse avatars may also help us order our favorite burgers too. And, we won't even need to push buttons on a touchscreen. We just need to say what we wish to eat and the assistant will answer and take our orders, perhaps later even in a familiar person's voice.

Onion Rings (large) $2.39

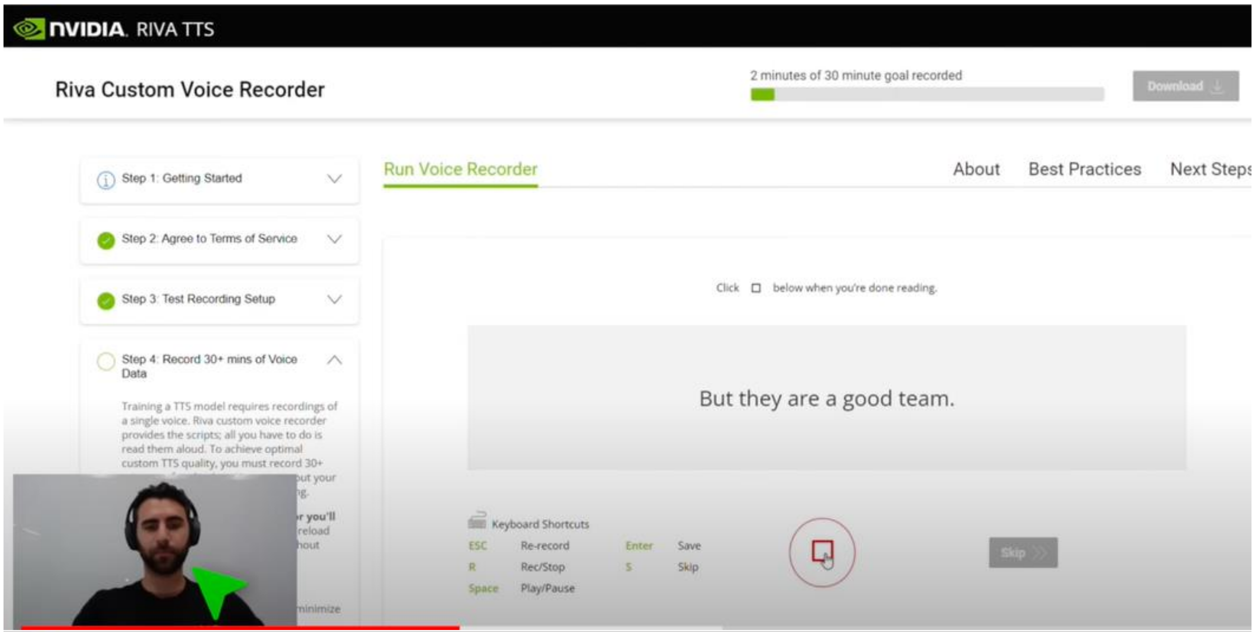
Cheeseburger (regular) $4.50

Our simple, classic cheeseburger begins with a 100% pure beef burger seasoned with just a pinch of salt and pepper.

With this new paper that is released in Oct 2022, imagine a future where we can all have our Toy Jensens or our own virtual assistants with our own voice.
A bit more about the AI that makes this voice synthesis happen.



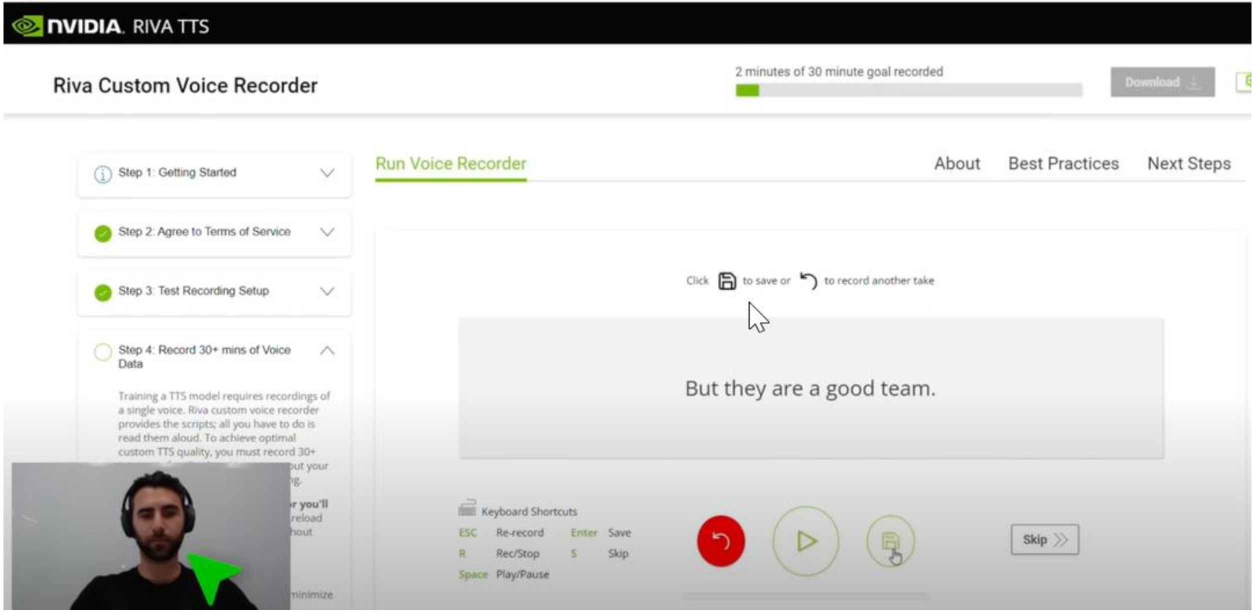
This work is an AI-based technique that takes samples of our voice, and can then clone it.



Training a TTS model requires recordings of a single voice. Riva custom voice recorder provides the scripts; all you have to do is read them aloud. To achieve optimal custom TTS quality, you must record 30+

Click □ below when you're done reading.

But they are a good team.

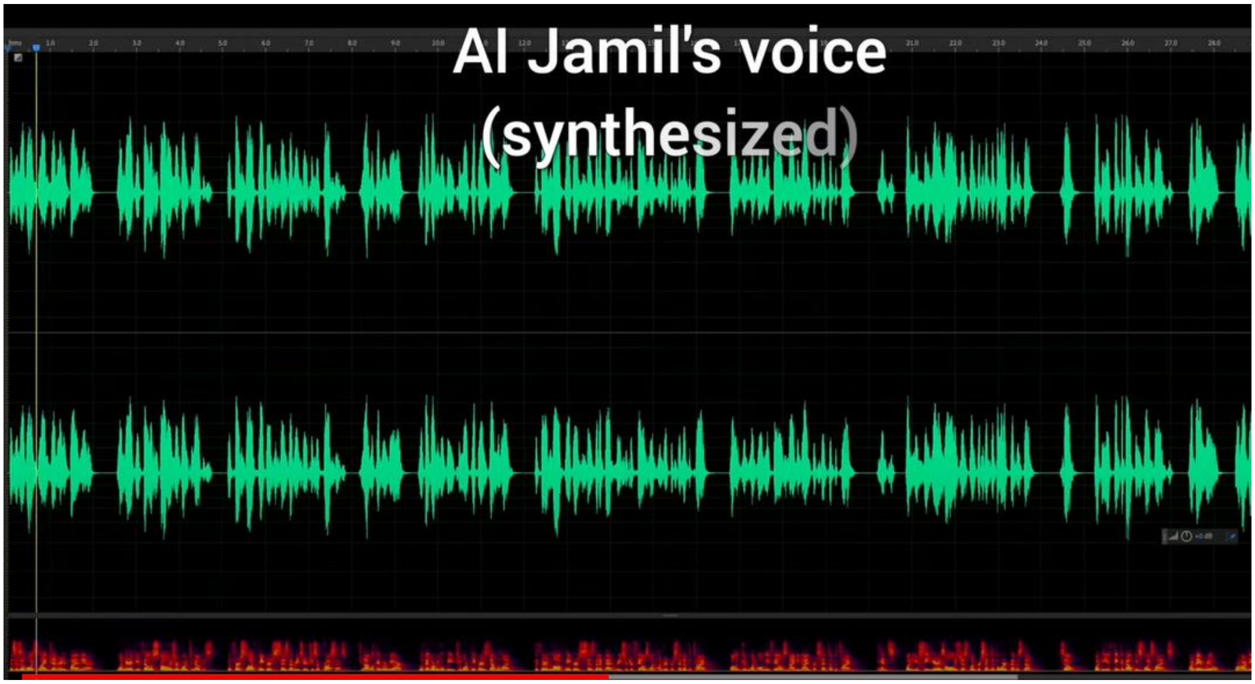


Training a TTS model requires recordings of a single voice. Riva custom voice recorder provides the scripts; all you have to do is read them aloud. To achieve optimal custom TTS quality, you must record 30+

Click 💾 to save or ↶ to record another take

But they are a good team.

Which is as simple as saying some sentences and pressing play, pause and stop buttons. And to train the AI we need less than 30 minutes of voice samples. The technique asks us to say these sentences and analyzes the timbre, prosody and the rhythm of our voice, which is quite a task for a AI model to do with so few samples. And it can clone the human voice afterwards



Although the sound is not perfect, and humans can differentiate by carefully listening that it is a synthesized voice, but in my opinion is the generated voice is good enough for a helpful humanlike virtual assistant. And it is really a great feat to be able to clone a human voice from half an hour worth of sound samples. And note that I have been really tough with them, these are some long, scholarly sentences that would give a challenge to any of these algorithms. Now, just a compare to earlier work there was a AI technique called Tacotron, which can perform voice cloning from a really short, few second-long sample. I have only heard examples of simpler, shorter sentences for that, up to 5 seconds, and what is new here is that this new technique takes more data, but in return, offers higher quality. But it does not stop there! It does more! This new method is easier to train, and it also generalizes to more languages better. And these are already good enough to be used in real products, so I wonder what a technique like this will look like just two more papers down the line. Maybe my voice could be synthesized by an AI model for some scholarly YouTube video presentations, Maybe it already is? Would that be a good thing? What do you think? Also, a virtual technology review for all fellow students to read latest advances in technology. Since NVIDIA has a great track record of putting these amazing tools into everyone's hands, interested scholars can sign up for an early access program by applying to below link.

https://developer.nvidia.com/riva/studio-early-access

Conclusion: The advances in text-to-speech and speech cloning are beyond imagination and with this new TTS model by Nvidia, human cloning and virtual assistants will become reality sooner. I hope that some of us will be able to try this and share feedback in the discussion forum. The early access link to the model is pasted above. This previous works was able to clone human voice but they were not able to generalize the model well and the earlier works were not able to add different emotion variations and language variations. But with the new model all of this is possible and even configuration interface is very intuitive. Thanks for reading the review. To access full paper please find link below https://arxiv.org/abs/2108.10447