Saving Big Bucks: Finetuning Deep TTS Model for Custom Voice Cloning

By Harry Walters (19166700)

And Jonathan Ellert (19138453)

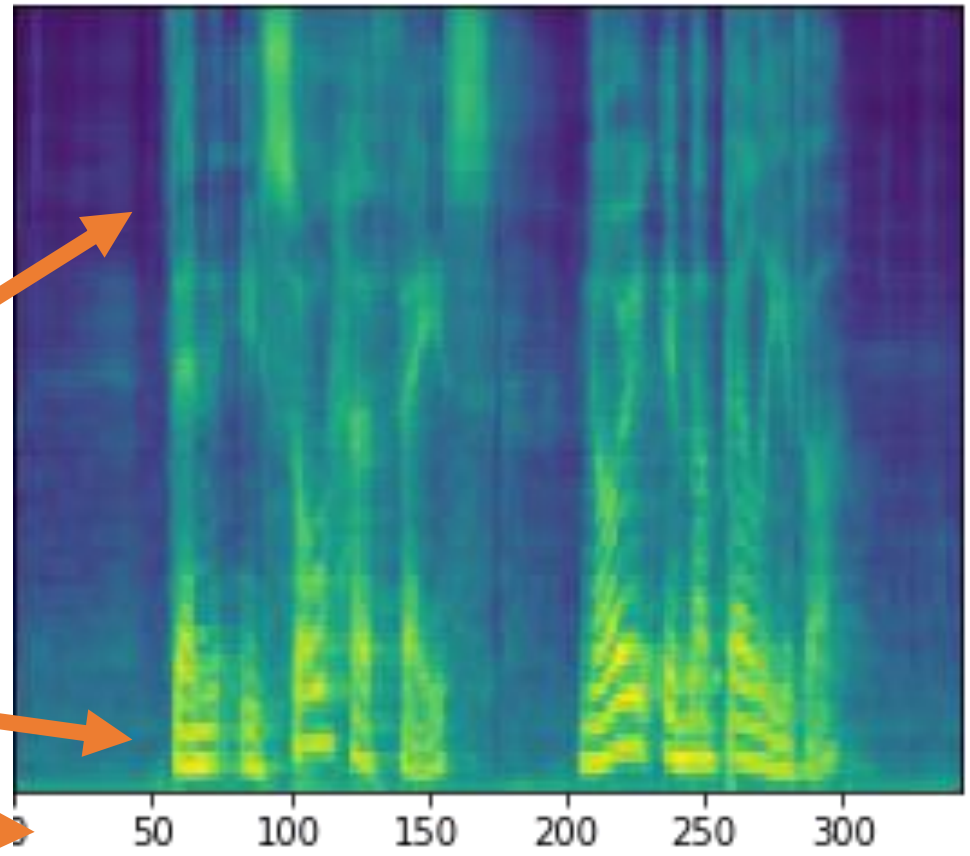
Introduction

CC: Welcome to our presentation, documenting our attempts to clone a speaker’s voice with deep neural networks.

CC: Together, we’ve spent the past several weeks on exploring the field of dimensionality reduction, as well as working with state of the art models.

What is a spectrogram?

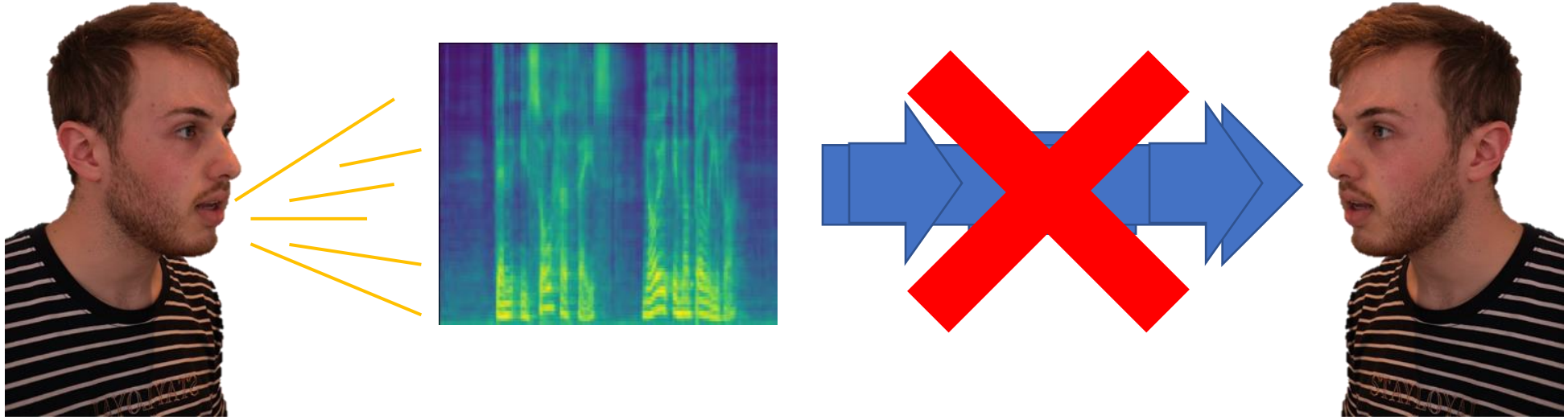
- Visual representation of audio.
- Breaks down the frequencies
- Higher frequencies at top, lower ones at bottom
- X axis is time

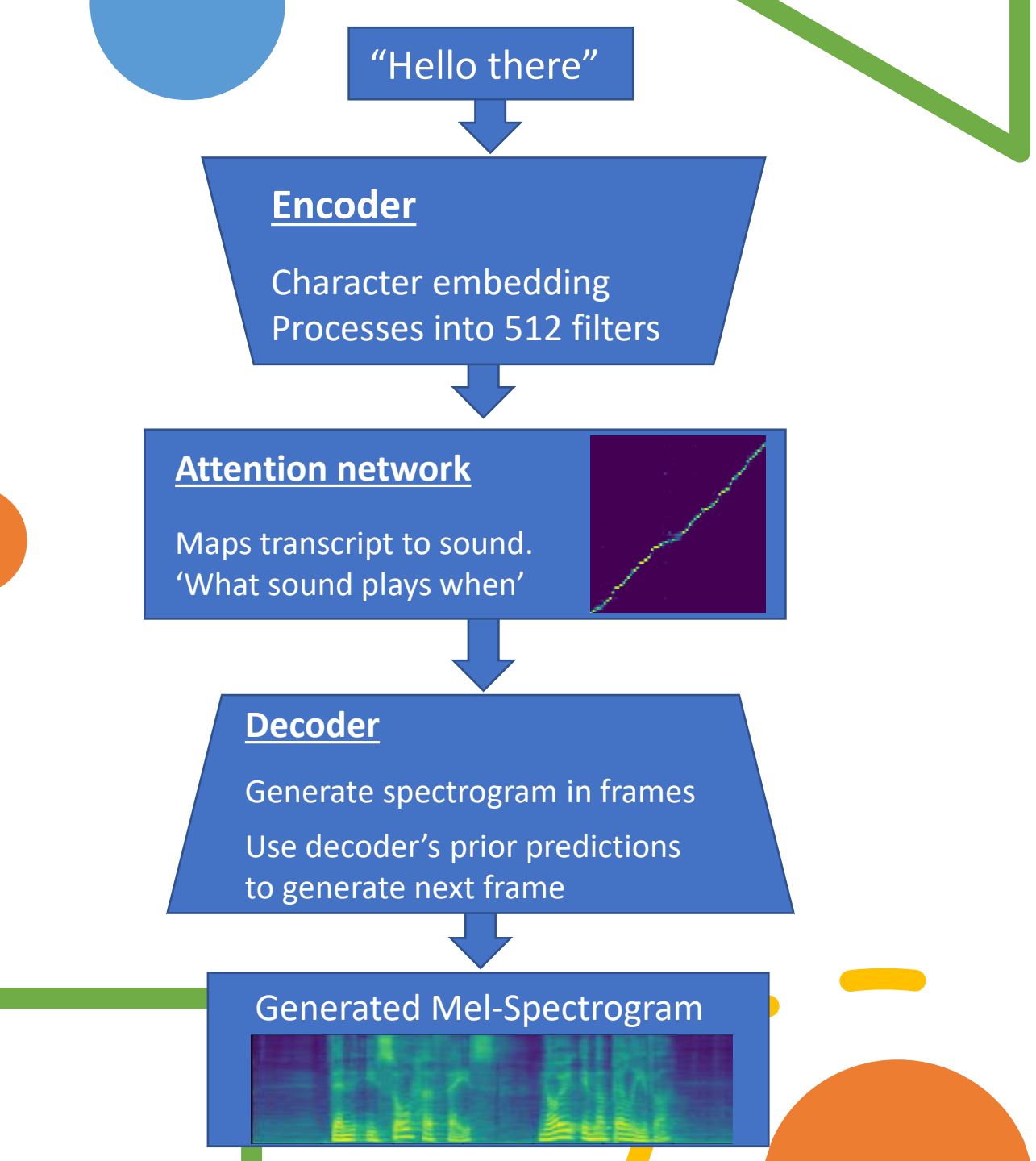


"Then I saw her face, now I'm a believer" (Jones, D. 1967)

What is a spectrogram?

- Audio converts to spectrogram, but not vice-versa.
- *Kinda*
- Algorithms estimate original audio.
  - *Pattern playback, Griffin-Lim, Wavenet, Waveglow*



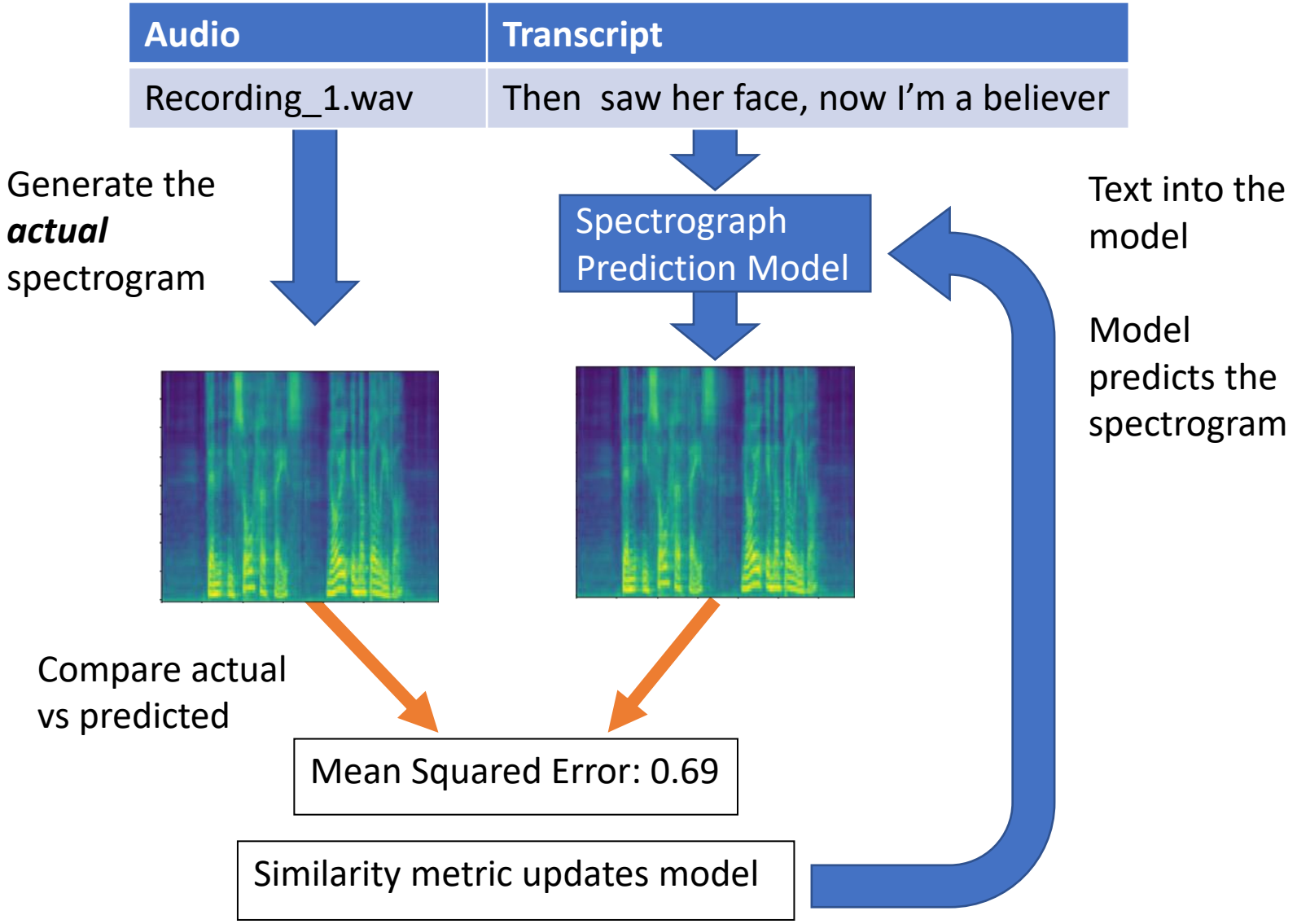


Tacotron 2, Summarised

- Encoder/decoder neural network
- Encoder: text to vector
- Decoder: vector to spectrogram
- Vocoder: spectrogram to audio

How to Train Your Tacotron 2

- Input: audio and transcript
- Generate spectrogram from audio.
- Predict spectrogram from transcript
- Prediction's accuracy updates model.



LJ Speech Dataset

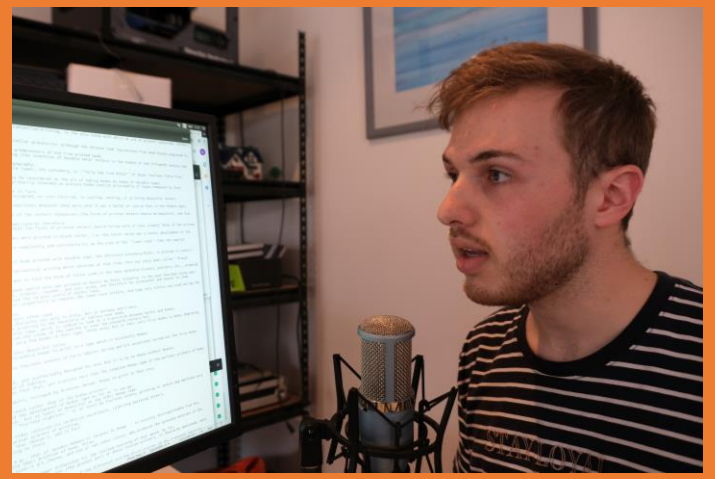
- Voice actress reading passages
- 13,100 files – 13,821 words
- 24 hours of audio
- Low sample rate – 22.05kHz

2.6
gigabytes

LibriVox

Methodology

- Download LJ Speech dataset
- Extract each audio file's transcript
- Harry read out each transcript
- Replace LJ speech audio with Harry Audio



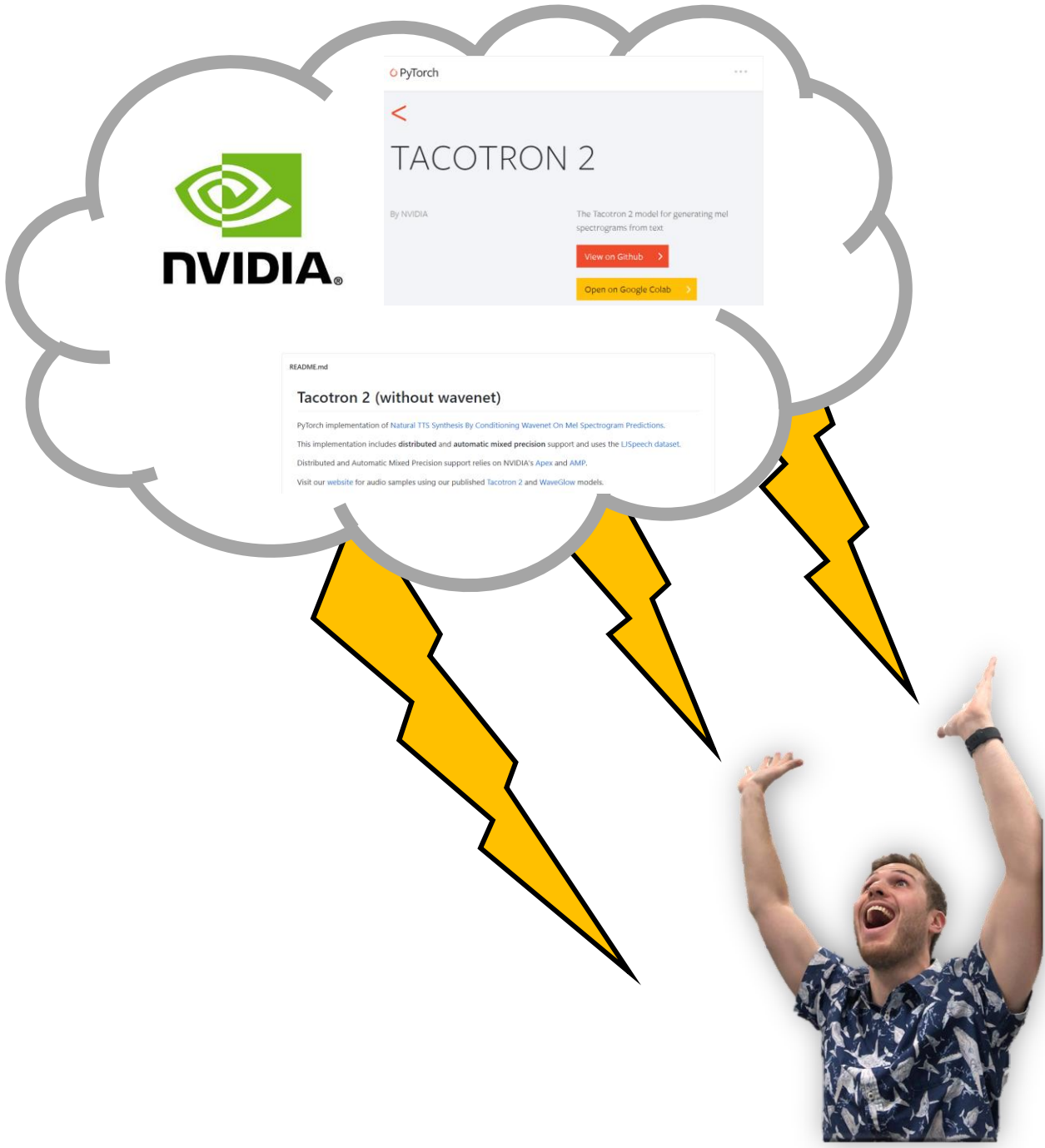
```
HarrysSoundRecordings/LJ001-0002.wav|in being comparatively modern.
HarrysSoundRecordings/LJ001-0003.wav|For although the Chinese took
relief for centuries before the woodcutters of the Netherlands, by
```

Data Creation

- Take LJ Speech Dataset
- (Audio and corresponding transcript)
- Record Harry reading transcripts
- Replace audio with Harry’s

Help from High Places

- Nvidia’s Tacotron/Waveglow implementation
- Use pretrained model weights
- Saves time and money
- Retrain Tacotron - keep Waveglow

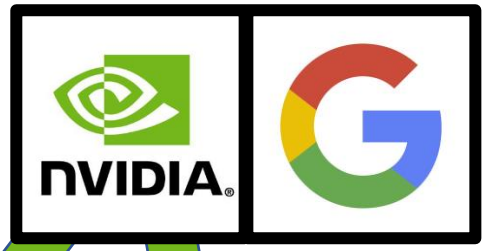


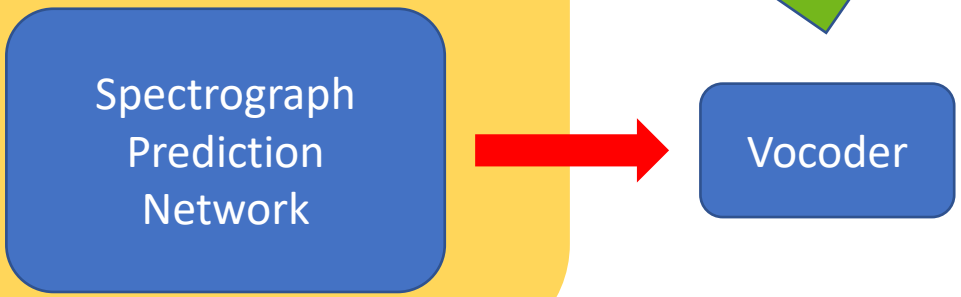
Lol nope.

- Only retrain spectrograph generator
- Keep vocoder as is

Train both parts concurrently.

It can't work otherwise





An Important discovery

- Pretrained Waveglow - Big surprise!
- Concurrent discovery in Stockholm
- Researchers agree with us.
- Won "Best Paper Award"

The Shyamalan Plot Twist

- CC: You've probably figured it already, but the speech you've been hearing is all synthesized by our model!

Tongue twisters

CC: Do you actually think the real Harry can say any of the following…?

CC: Peter Piper picked a peck of pickled peppers. A peck of pickled peppers Peter Piper picked. If Peter Piper picked a peck of pickled peppers, Where's the peck of pickled peppers Peter Piper picked?

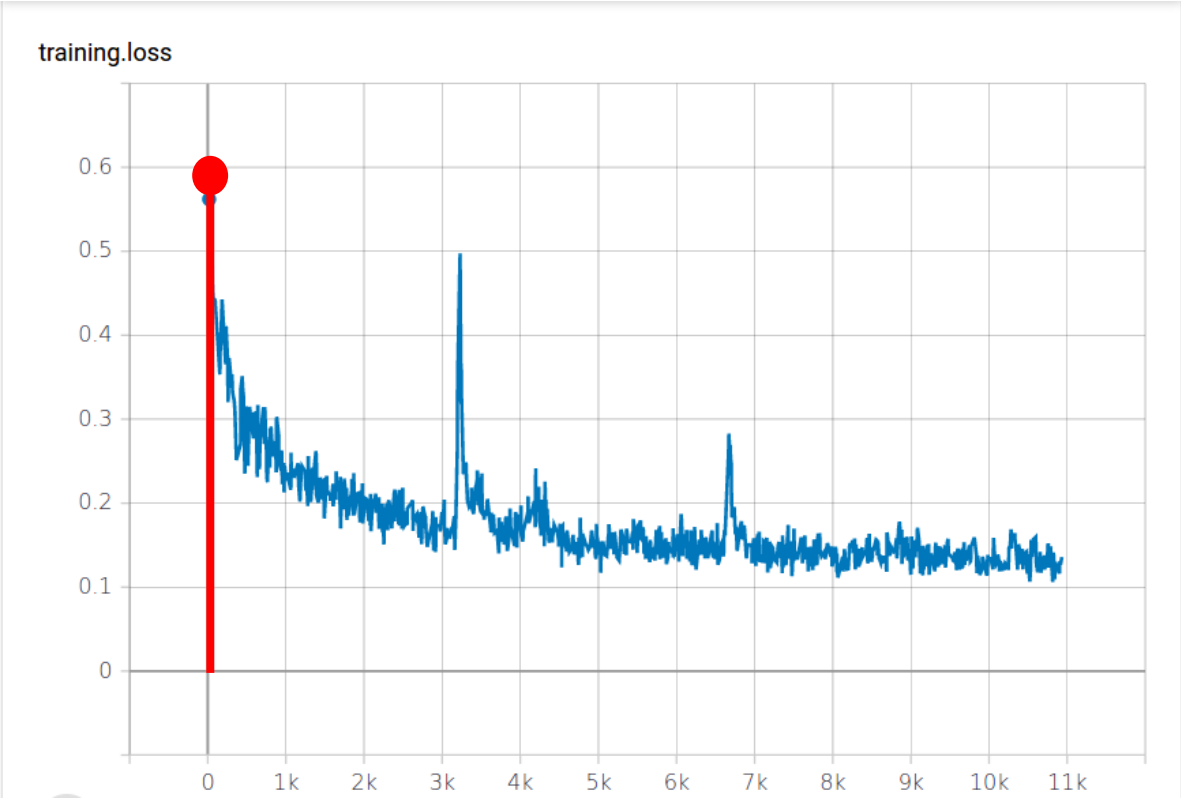
CC: She sells seashells on the seashore. The shells she sells are surely sea shells.

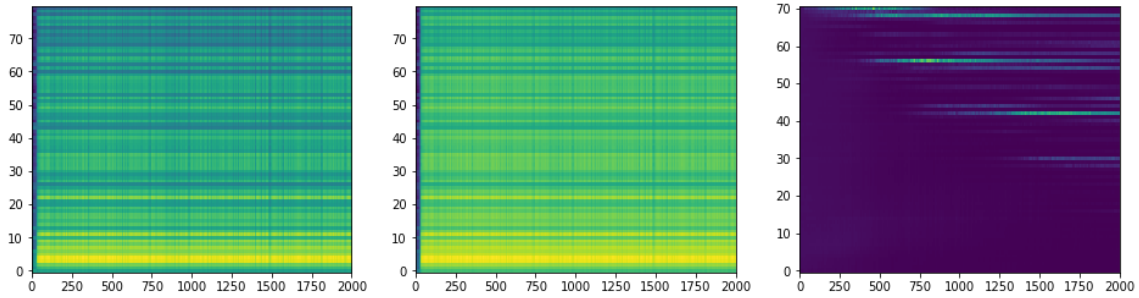
CC:Betty Botter bought a bit of butter. The butter Betty Botter bought was a bit bitter and made her batter bitter. But a bit of better butter makes better batter. So Betty Botter bought a bit of better butter making Betty Botter's bitter batter better.

Gibberish

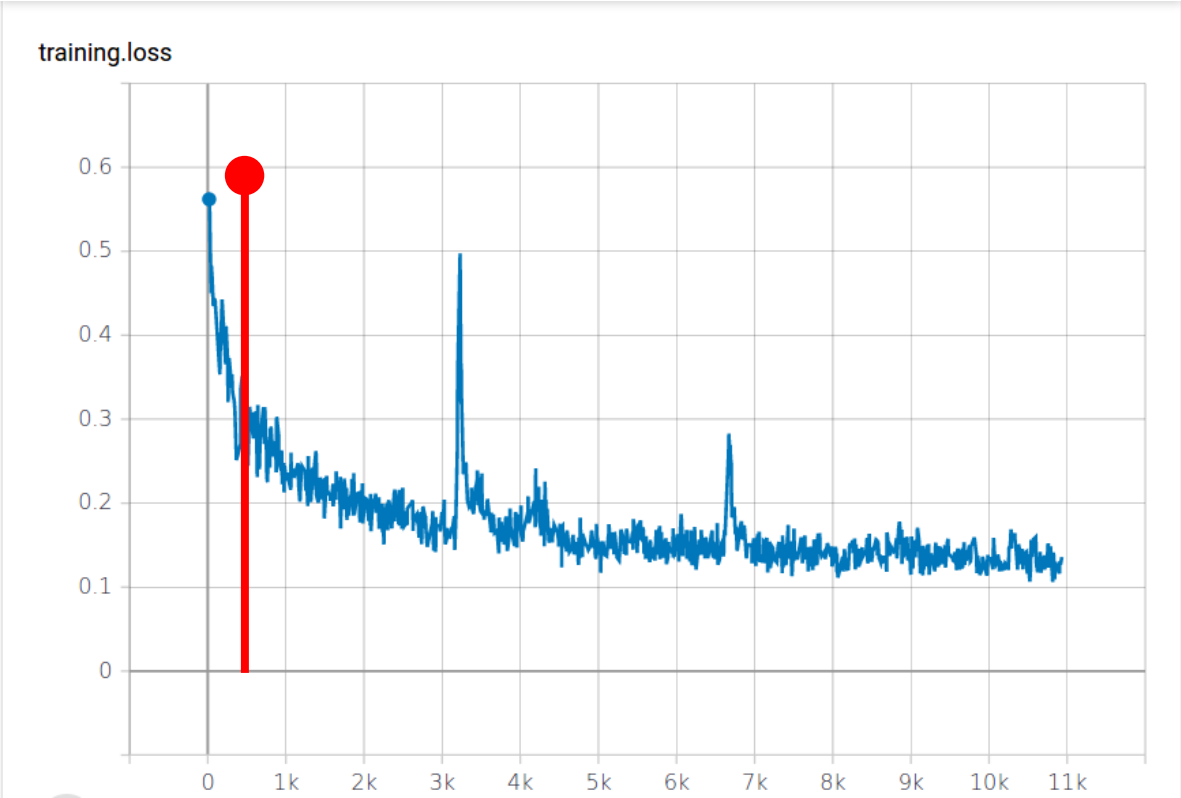
CC: Here is an example on how the model can produce gibberish. “And then…” [Gibberish]

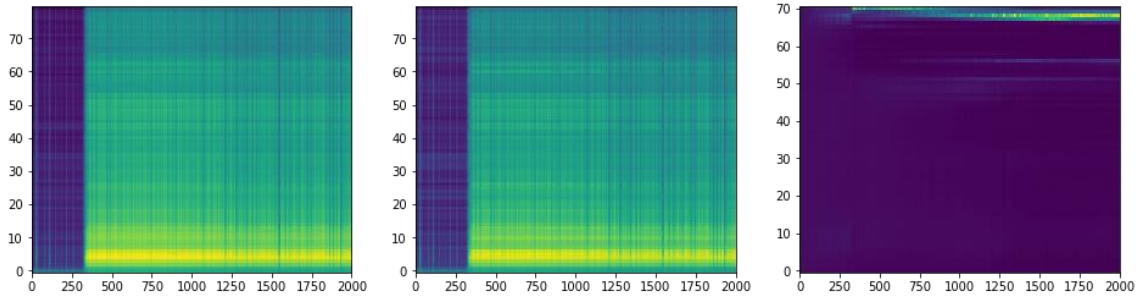
Results



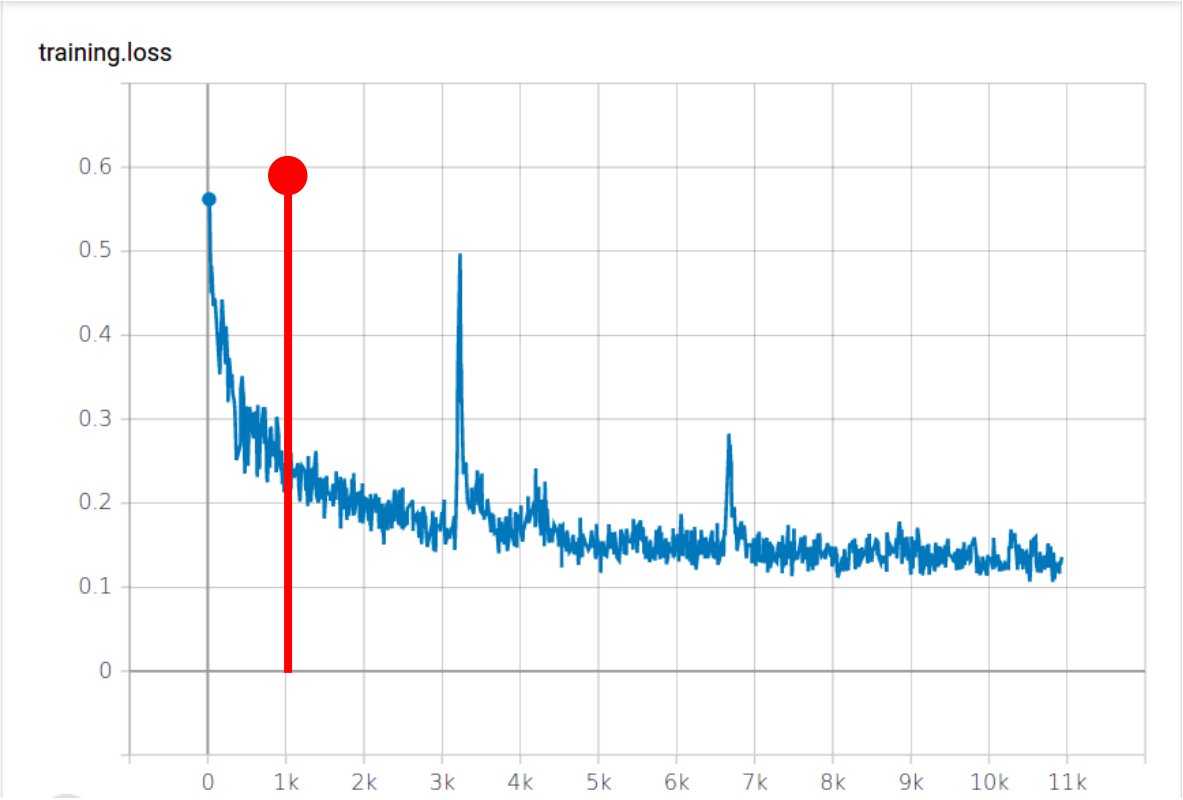


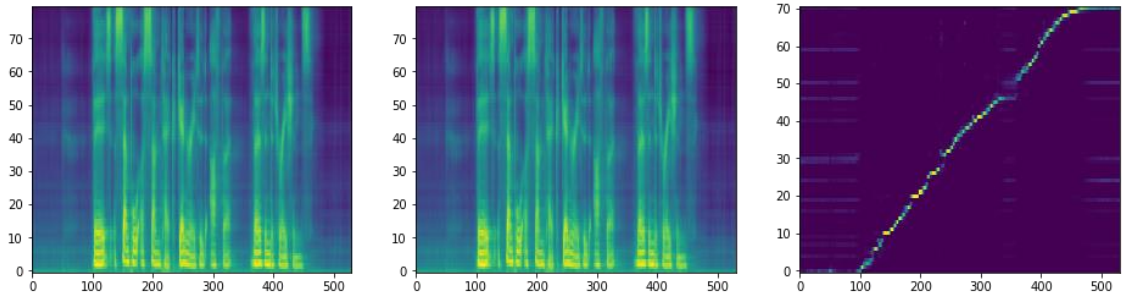
“CC: Here is a sample of some text, generated after one iteration.”



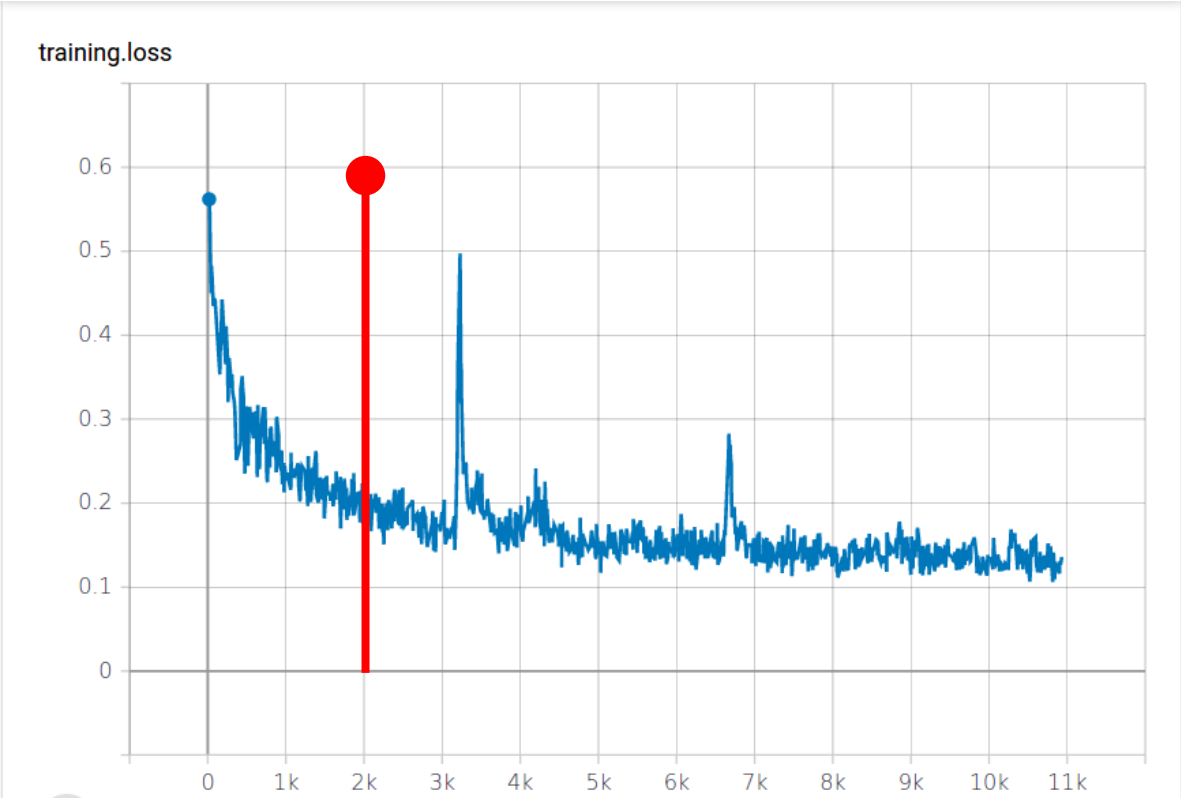


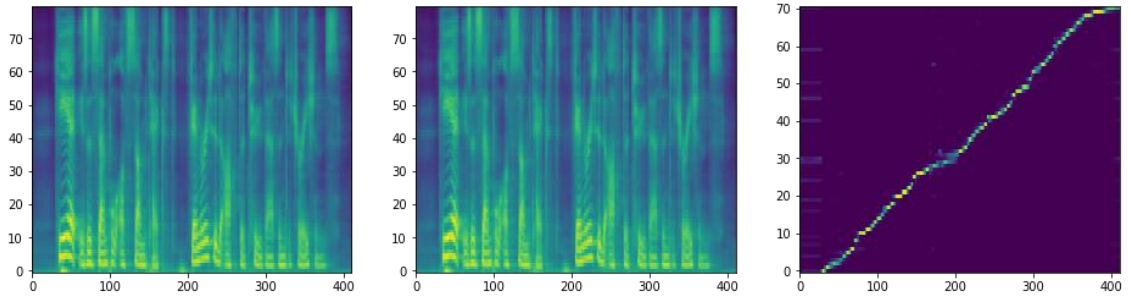
"CC: Here is a sample of some text, generated after 500 iterations."



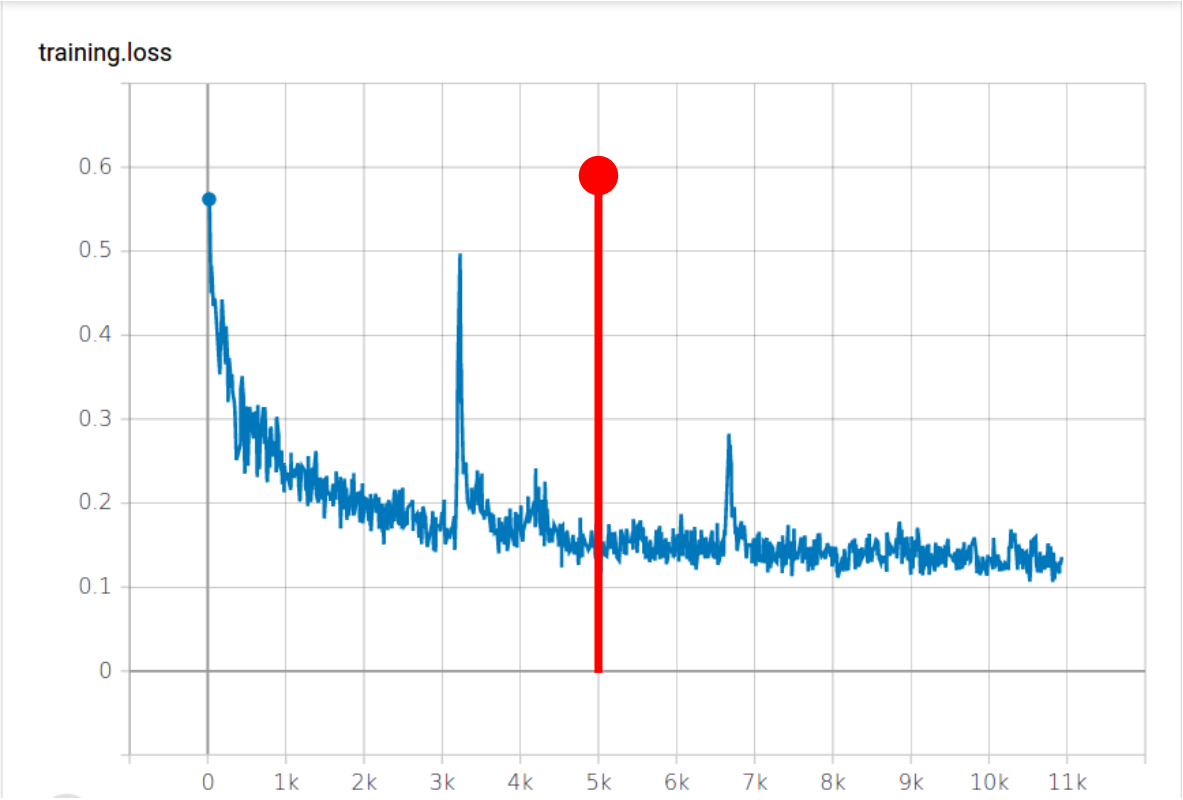


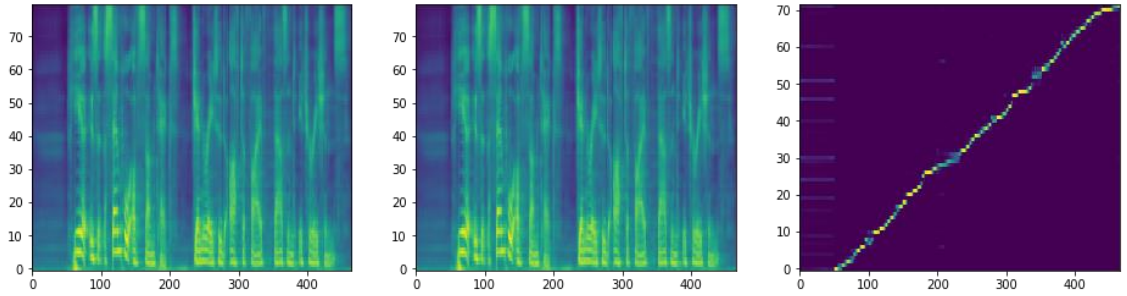
"CC: Here is a sample of some text, generated after 1000 iterations."



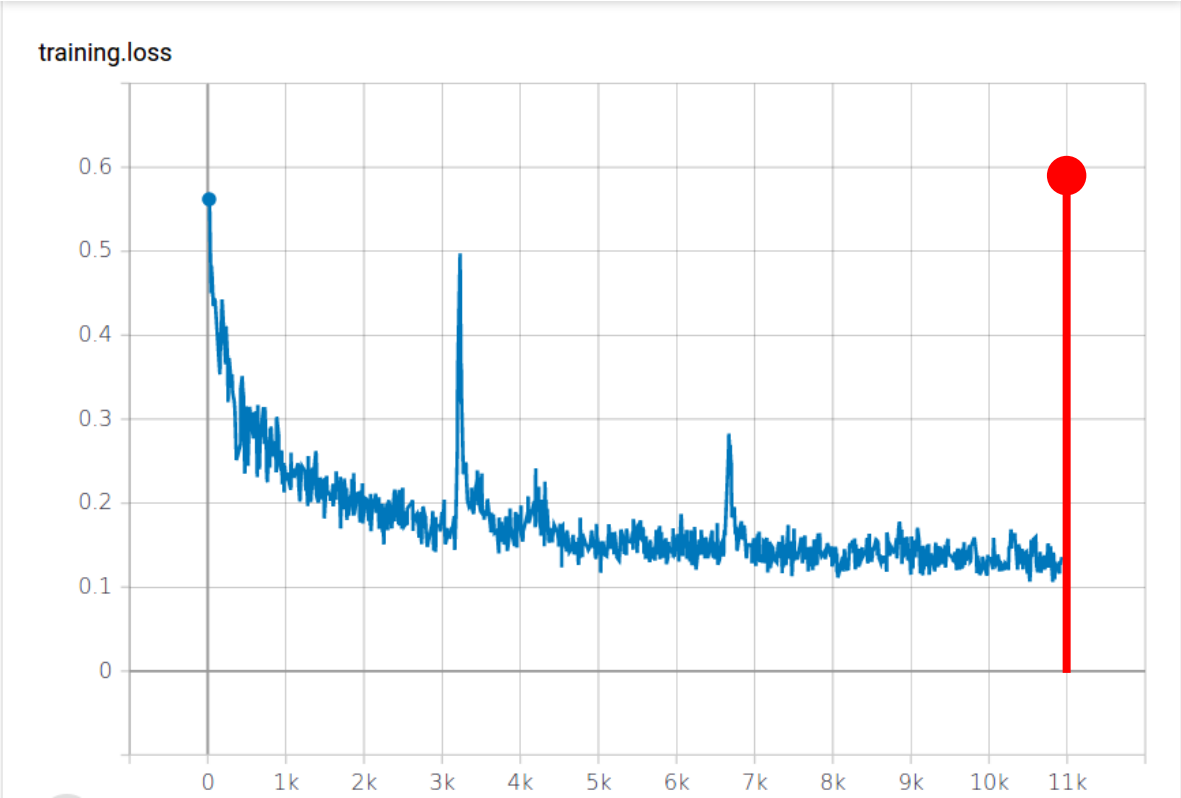


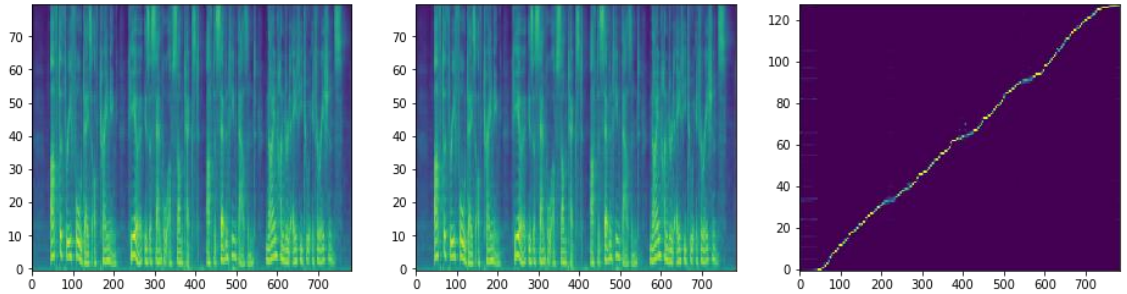
"CC: Here is a sample of some text, generated after 2000 iterations."





“CC: Here is a sample of some text, generated after 5000 iterations.”





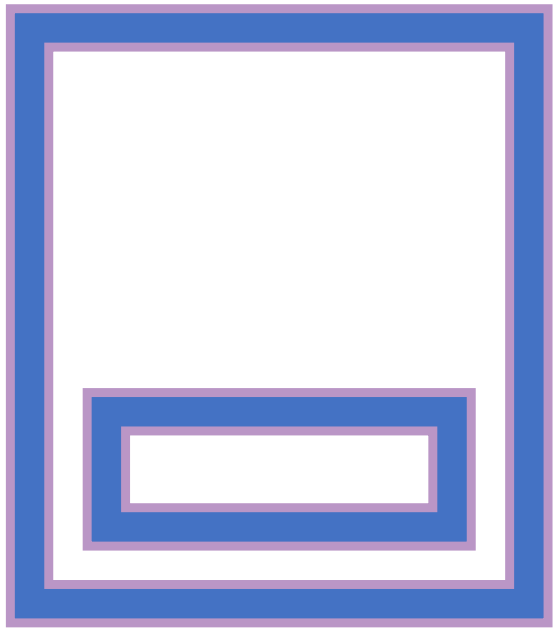
“CC: After 3 full days of training, here is a sample of some text, after 17,982 iterations”

Discussions

Conclusions

References

References



Alexanderson, Simon, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. “Generating Coherent Spontaneous Speech and Gesture from Text.” Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, October. https://doi.org/10.1145/3383652.3423874.

“Audio Samples Related to Tacotron, an End-to-End Speech Synthesis System by Google.” 2017. Github.Io. 2017. https://google.github.io/tacotron/.

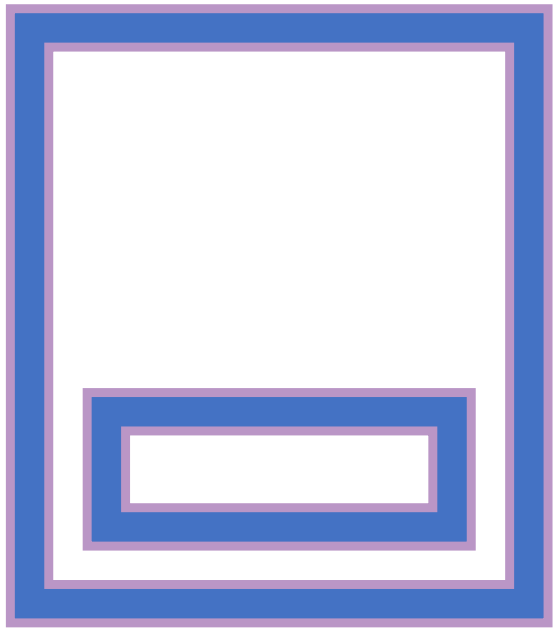
Misra, Sayak. 2020. Our TTS Architecture. Image. https://towardsdatascience.com/text-to-speech-with-tacotron-2-and-fastspeech-using-espnet-3a711131e0fa.

Misra, Sayak. 2020. "Text To Speech With Tacotron-2 And Fastspeech Using Espnet.". Towards Data Science. https://towardsdatascience.com/text-to-speech-with-tacotron-2-and-fastspeech-using-espnet-3a711131e0fa.

Liu, Yifan, and Jin Zheng. 2019. "Es-Tacotron2: Multi-Task Tacotron 2 With Pre-Trained Estimated Network For Reducing The Over-Smoothness Problem". Information 10 (4): 131. doi:10.3390/info10040131.

Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. 2018. “WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS.” NVIDIA Corporation. https://arxiv.org/pdf/1811.00002.pdf.

References



Shen, Jonathan, and Ruoming Pang. 2017. “Tacotron 2: Generating Human-like Speech from Text.” Google AI Blog. 2017. https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html.

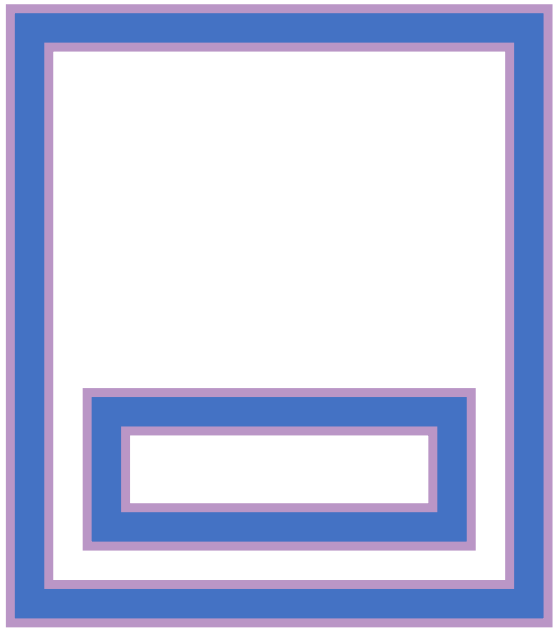
Shen, Jonathan, Ruoming Pang, Ron Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, et al. 2018. “NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS.” Berkeley: Google, Inc., University of California. https://arxiv.org/pdf/1712.05884.pdf.

Valle, Rafael. 2020a. “NVIDIA/Tacotron2.” GitHub. June 12, 2020. https://github.com/NVIDIA/tacotron2.

———. 2020b. “NVIDIA/Waveglow.” GitHub. September 3, 2020. https://github.com/NVIDIA/waveglow.

Van Der Maaten, Laurens, Eric Postma, and Jaap Van Den Herik. 2009. "Dimensionality Reduction: A Comparative Review". Tilburg Centre For Creative Computing TR 2009–005: 1-2. https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf.

References



Wang, Yuxuan, R Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron Weiss, Navdeep Jaitly, Zongheng Yang, et al. 2017. “TACOTRON: TOWARDS END-TO-END SPEECH SYN- THESIS.” Google, Inc. https://arxiv.org/pdf/1703.10135.pdf.

Wang, Yuxuan, R Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif Saurous. 2017. “Uncovering Latent Style Factors for Expressive Speech Synthesis.” Mountain View, CA: Google Research. https://arxiv.org/pdf/1711.00520.pdf.

Zhang, Ya-Jie, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. “Learning Latent Representations For Style Control and Transfer in End-To-End Speech Synthesis.” In , 1,2. Hefei, China. https://doi.org/arXiv:1812.04342v2.

Problems

LJ Speech 24 hours!

Ours' much smaller – 1hr

Leads to overfit synthesis

New words generalise badly

Future

Use phonetic pangrams?

- Sentences with every sound
- Bigger isn’t always better