Analysis of the Authenticity of Tacotron2 and Talknet Synthesis

Felix Peng, Erick Amaro-Hernandez

Supplementary Materials:

https://drive.google.com/file/d/137AEjOd9xAfQslnhPu_qCmJ6p7FTCoGz/view?usp=sharing

(Link to corpora used and sentences generated for the survey)

Feedback: No detailed feedback desired.

Self Assessment

Cover Page and Self- Assessment: Masterful. The cover page is present and includes all of the desired components. The self assessment is complete as well.

Scope of Writeup: Masterful-to-Acceptable. All ten elements of a project 2 writeup are included and explained, some sections are shorter since they do not apply directly to our project which is explained in the sections. However some sections may lack a little info which may be under other sections.

Demonstration of Knowledge: Masterful. Our project demonstrates proficiency with our tools on real natural language data which we collected. With the explicit consent of professor Styler, the tool is run using Google Colab instead of installing it locally.

Richness: Acceptable. Our project applies the tools in a way which is demonstrative of the NLP concepts learned in class. This includes talking about strengths and weaknesses of the tools, and

also the ethical implications of TTS systems. These concepts are discussed in the paper, however may not cover the full scope of topics discussed in class.

Formatting and Length: Masterful. The paper is formatted in a reasonable and easily understandable manner. The paper meets the 5-7 single space page length requirement.

Structure and Organization: Masterful-Acceptable. All of the sections are clearly labeled and follow the structure guidelines. Some of the sections may lack transitions, however since the format is followed, the reader should be able to have a clear understanding of each section and its relevance.

Language and Argumentation: Masterful. The paper is in academic prose and clearly and easily understandable with few grammatical errors or typos.

Academic Integrity: Masterful. All works are cited properly, and the resources and tools we used are linked to within the appropriate sections where they are discussed.

Proposed Project Grade: 90% (about the average of the different categories' scores). We have put in a good amount of effort to create a corpora and create a robust TTS system. We gathered a variety of responses pertaining to the quality of our TTS model, and used the responses to talk in depth about the strengths and weaknesses of the tools we used. Many aspects of the systems are discussed in the paper although some topics from class may have been left out.

Task

Artificial Intelligence text-to-speech systems have become scarily accurate and increasingly accessible and easier to make. They are able to imitate voices using minimal amounts of data, from as low as 5 minutes as shown by websites like 15.ai and fakeyou.com. The issue is so prevalent that voice actors and artists have specifically requested TTS modeling websites to stop any users from trying to submit models of their voice for public use, an example being Uberduck.ai's (another TTS model host) blacklisted voices spreadsheet.

In awe of the power of modern voice synthesis, and combined with various ethical concerns we had about generated, malicious audio passed off as real, we looked into creating our own text to speech model. We were curious whether we could convince students that a TTS generated voice was real and how effective and accurate modern TTS systems were. To trick our peers, we used tacotron2 and TalkNet voice synthesis to create various models of professor Will Styler's voice, each using different amounts of training data. To determine the accuracy of models created, we tested whether these easily accessible training methods would fall victim to tricky problems in voice synthesis like correct pronunciation of special words like singular letters and acronyms.

Finally, we test the limits and have fun with the more advanced TalkNet. When TalkNet is given an audio to inference and a transcript of that audio, it can make the TTS model mimic the given audio. These results varied from our TTS model achieving world record rapping speeds to the model giving it its best shot to hold a note in a song.

Toolkit

For this project, we'll be using tacotron2 and TalkNet models to synthesize speech. Both voice synthesizers use a corpus of audio files of individual sentences spoken by the person whose voice is being cloned, as well as a text file containing transcriptions of each audio file. TalkNet adds an extra layer to non-prosodic voice synthesis like tacotron2. Instead of going from Text To Speech, it uses reference audio and text to generate a representation of how long each word should be as well as the pitch of each word and finally synthesizes audio using that representation [1].

Some of the tools that power this project are only available on Linux operating systems which unfortunately no one in our group uses, so we have taken advantage of Google Colab Notebooks. They allow us to use a remote server from google to host the tools and train models and synthesize audio from there. These notebooks also have pre-written code which leavescorpus data collection and annotation as our only job.

The following Google Colab Notebooks were used:

TalkNet training and Synthesis notebooks by Justinjohn0306 on Github
https://github.com/bycloudai/TalkNET-colab/blob/main/README.md

Tacotron2 training and synthesis notebooks by Uberduck.ai

Training:

https://colab.research.google.com/drive/1WTilMdm9Vf7KE79gzkeeTBigAN6iv3Bg?usp=sharing#scrollTo=iQBUZTuMogui

Synthesis:
https://colab.research.google.com/drive/1NVA3ndxhYWsKn-zwh3NnzMMgoVdJ5xUx

In order to collect audio of Will Styler's speech, we downloaded SP22 LIGN 6 lectures from podcast.ucsd.edu. Then we converted them to an audio-only format and listened for clear sentences with little to no pauses. These audio files were processed using Audacity, which has a feature to mark segments in the audio which can be exported as their own individual audio files. These segments were exported as .wav files with 22050 Khz sampling rate and mono audio. This audio file format is the only specific format that we used to train TalkNet and tacotron2.

Once sufficient data about Will Styler's speech was collected, Erick wrote a python script to measure data of tacotron/TalkNet corpora collecting details such as total words and unique words trained on, average length of each sentence included in the corpus in words as well as average length of the audio, and standard deviation of the aforementioned two.

Finally, a survey was conducted using google forms that showed participants audio files which were hosted on Google Drive.

Install Process

As we're using Google Colab notebooks written by people online, a lot of the installations are set up for us using scripts written to install the many dependencies these tools run on. The only things we needed to "install" in order to execute this project are Google Accounts to save the outputted models into our Google Drive.

We gave Will Styler access to his own TTS model, and supplied the following tutorials on how to synthesize. Both guides are written by Erick Amaro.

TalkNet Synthesis Guide (~10-15 minute setup)

https://docs.google.com/document/d/14WnH8F5NUQ1Jgz_4HgKyrH5bfM7vW_uwpdG-7LQC4NQ/edit?usp=sharing

Tacotron Synthesis Guide (3-5 minute setup)

https://www.youtube.com/watch?v=rP9uGO7Ki-8

Corpora Used

The Will Styler Audio corpus uses 204 sentences spoken by Will Styler through four different LIGN 6 lectures, summing up to eleven minutes and thirty-five seconds worth of non-stop Will Styler speech audio.

The link below contains the corpora used in this paper, it contains:

- Training Data for the 25% , 50%, 75%, and 100% data models
- The sentences used in the survey conducted for this paper

https://drive.google.com/file/d/137AEjOd9xAfQslnhPu_qCmJ6p7FTCoGz/view?usp=sharing

```
wavs/88.wav|The idea that a word can have different meaning.
wavs/89.wav|In a given situation I usually only mean one sense of a word.
wavs/90.wav|If I say the chair was unhappy I mean the ranking faculty member of an academic department.
wavs/91.wav|All of these are words that have the same form and the same pronounciation but different meanings.
wavs/92.wav|This is not the same thing as on the low down.
wavs/93.wav|None of this can be discovered from the written form alone.
wavs/94.wav|We have a situation where the same words can have massively divergent meanings.
wavs/95.wav|I could have a massive library but that doesn't mean I know the things in it.
wavs/96.wav|We can think about probability as the degree of certainty that the value of a variable is one thing and not another.
wavs/97.wav|That is impossible that will never happen.
wavs/98.wav|We have intuitions about this very much as humans but probability is the quantification.
```

Example annotation file in the format of

To accomplish our task of tricking students into thinking synthesized audio was real audio, we wanted to see what fraction of our audio data data would be needed to create speech students would believe is organically produced. Four sets of tacotron2 synthesized audio were presented to students with each set being from models trained on 25%, 50%, 75%, and 100% of the original training data in the Will Styler Audio corpus. Below is a table listing the different variables that each of these training data sets had.

25% data	50% data	75% data	100% data	
51	102	153	204	Total Audios*
0:02:32	0:05:25	0:08:25	0:11:35	Total Audio Length h:mm:ss
565	1203	1882	2614	Total Words
307	512	694	883	Unique Words
1.84	2.35	2.71	2.96	Total/Unique ratio
11.08	11.79	12.3	12.81	Avg. Sentence

				Length (words)
3.66	3.97	4.04	4.09	StdDeviation of Sentence Len.
2.99	3.19	3.3	3.41	Avg. Audio Length(seconds)
0.89	1.0	1.02	1.03	StdDeviation of Audio Length

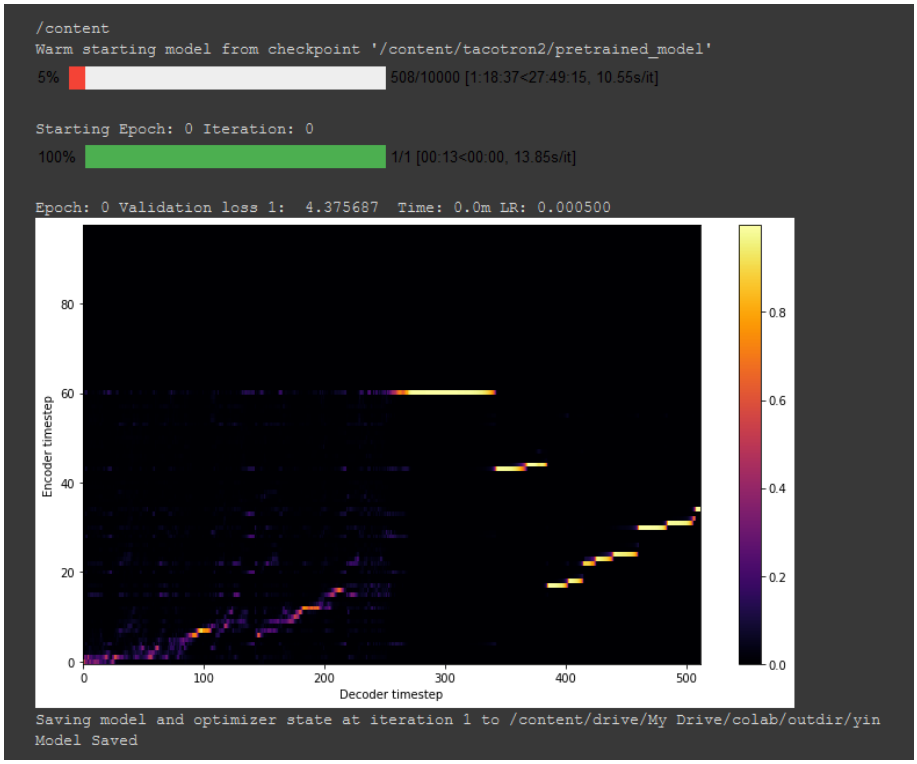
Model Training

All we need to train a tacotron2 or TalkNet model is some audio files of sentences spoken by the voice we want to clone, and a transcription of those files, but we can choose to train a model for 10 minutes to one hour, which is going to give different progress depending on how big our corpus is.

Fortunately, the Google Colab Notebook gives some measure of how much a model has been trained, which is given in the form of Epochs and Validation Loss (example below).

Validation Loss is a measurement of confidence that a model will perform good synthesis of audio based on its inputed corpus.

An epoch is completed when the TTS model analyzes a batch of audios selected from a corpus, and decreases validation loss when one is completed as the analysis gives the model more to work with. All models worked with in this paper were trained to a validation loss of about 0.115, which is a little over what the uberduck.ai training guide deems good enough, so these are well trained.



Code Used to Run the Tool

In order to synthesize audio, tacotron and TalkNet have their own respective Google Colab notebooks to synthesize from.

Tacotron simply prompts for a text string to synthesize audio from, while TalkNet requires reference audio to infer talking speed and pitch from, as well as a transcript of that reference audio. Reference audio can be disabled to revert TalkNet into non-prosodic speech synthesis but was not used in our experiments.

Experimentation and Results

In order to present synthesized audio to students, we made a survey on Google Forms which presented them with five sets of questions and audio clips pertaining to each question. Each set of questions asked them to rate how close a given speech synthesis audio was to sounding like real human speech with a 1 rating being “Not at all” and a 10 rating being “practically indistinguishable”. At the end of each question set, we asked the student if they believed any of the presented audio clips were from a real human, or if they had just been shown only synthesized speech. This survey was posted in the Linguistics Undergraduate Association discord chat as well as the group discord for classes being taught by Will Styler in SP22. In total, seventeen responses were gathered.

As discussed in “Corpora Used”, the TTS models were trained on 25%, 50%, 75%, and 100% of data. Additionally, the first set of audio was generated using TalkNet, which was trained

with 100% data. The following 4 sets of audio were generated from tacotron2 models, with the clips in the order of least data (25%) to most data (100%).

TalkNet was made to synthesize 3 sentences that were not used to train the model, but had been said by Will Styler in the first LIGN 6 lecture of SP22. TalkNet was given 3 audios to reference prosody directly from lecture and Will Styler himself. This was done to apply Professor Styler's prosody to the generated speech and test the synthesis quality of TalkNet at the same time. If the model was able to properly replicate prosody, it would create the most authentic sounding audios users would be asked about.

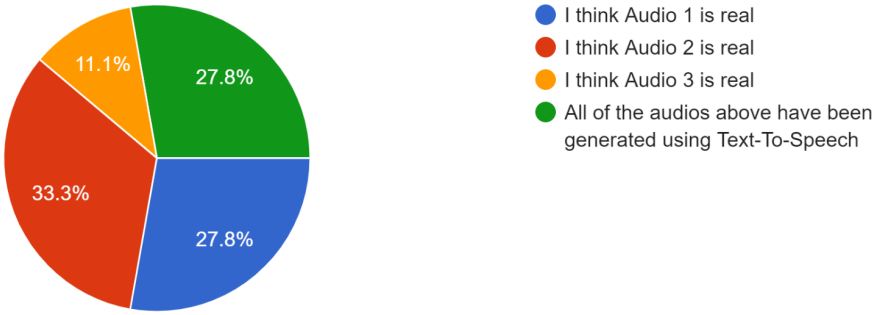
The sentences synthesized by tacotron2 models were chosen from uberduck.ai's Voice Cloning script which is given to users who want to clone their own voice. This is a list of what uberduck deems phonetically diverse sentences which a user can read and record to supply the model. The model learns most from these sentences, which makes sense because the TTS model should have a sense of how a speaker produces a wide variety of phones in different contexts. Sentences used are listed below.

```
TalkNet Sentences, referencing from authentic Will Styler Prosody

"I try to be really kind but the way you can screw that up is by blowing off the class"
"One of the big questions that we can start with is well, why dont we just make computers programmable in english
     why don't we just use human language with them generally."
"But again you have no excuse to come to class sick, just dont, don't even think about it."


Sentences generated by tacotron using uberduck.ai script

An array is given at the end to signify which portion of the survey they were used in
[25, 100] would mean it was used in sections where the model was trained with 25 and
100 percent of data.

List 6 Sentence 2. The crooked maze failed to fool the mouse.
List 7 Sentence 4. The clock struck to mark the third period [50]
List 8 Sentence 1. A yacht slid around the point into the bay [25,50,75,100]
List 12 Sentence 1. The bark of the pine tree was shiny and dark. [25,75]
List 14 Sentence 5. The glow deepened in the eyes of the sweet girl.
List 17 Sentence 1. The jacket hung on the back of the wide chair.
List 19 Sentence 8. He carved a head from the round block of marble.  [100]
List 21 Sentence 1. The brown house was on fire to the attic
List 21 Sentence 2. The lure is used to catch trout and flounder. [25,50,75,100]
List 22 Sentence 10. The wreck occurred by the bank on Main Street.
List 24 Sentence 4. They floated on the raft to sun their white backs.
```

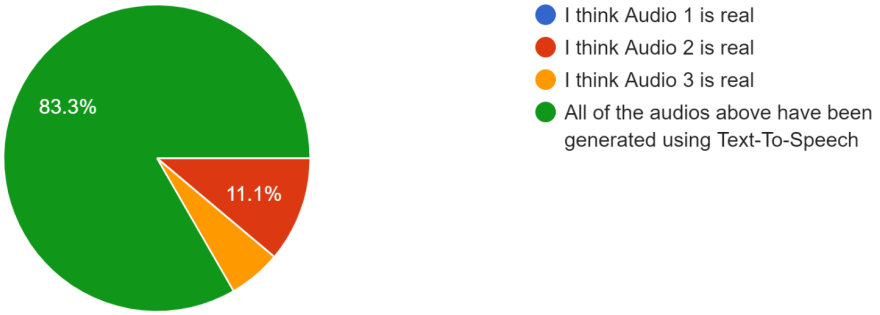
*Not all sentences picked for the survey were actually used for it, however you can still listen to them as they are included with the corpus download linked in "Corpora Used"

Now looking at the results:

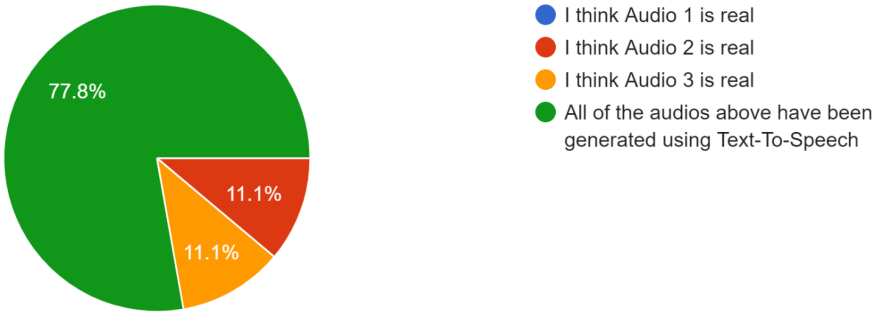
Do you have any suspicion that any of these audios come from a real organic human speaker?
18 responses



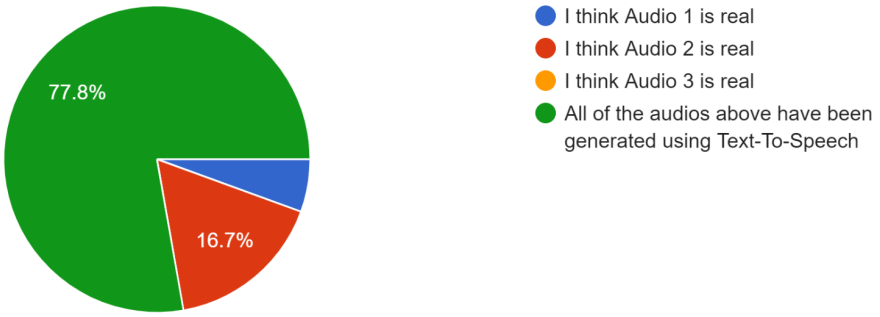
Talknet Model Using 100% Data and Prosody Reference



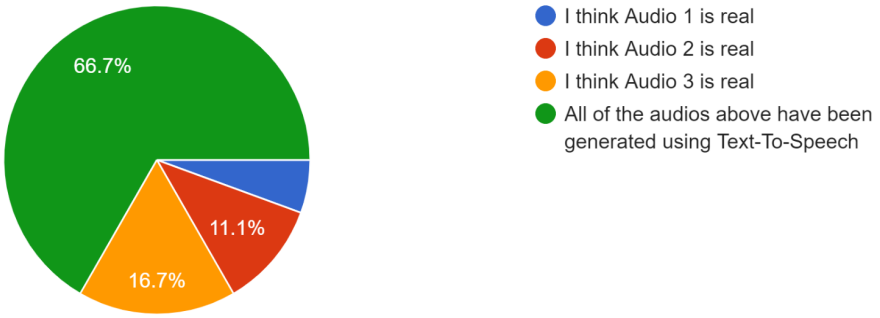
Tacotron2 Model Using 25% Data



Tacotron2 Model Using 50% Data



Tacotron2 Model Using 75% Data



Tacotron2 Model Using 100% Data

As shown in the diagrams, increasing the amount of training data used for the tacotron2 does slightly increase the percentage of people who believed that one of the audio clips was a real person. This is inline with the idea that more training data produces a better TTS model. However, none of the tacotron2 models managed to convince a majority of people that a real person generated the audio. Listening to the audio on the 100% model explains why this is the case. While the words are intelligible, the TTS's pitch and volume is inconsistent and the voice sounds very "robo-fied". These problems are exacerbated in the models that have less data.

The truly surprising result is the effectiveness of the TalkNet model. The 3 audio clips managed to fool a whopping 72.2 percent of people into thinking one of the clips was from a real person. The reason is clear, the problems with the tacotron2 models are all lessened, with the voice sounding natural and more closely resembling Professor Styler's. However, what sets the TalkNet audio apart is the added prosody, which seems like it perfectly matches the prosody of the reference clips from lecture. Because of this, the audio goes from sounding like a synthetic voice, to sounding like a human voice with a robo filter.

Listening to the audios provided in the TalkNet section of the survey, it has all the prosody we would never expect from a normal text-to-speech system.

Strengths of the Tool

<u>Tacotron</u>

Despite describing it as “non-prosodic” earlier in the paper, the voice that is synthesized sometimes has qualities that the original speaker has in their own speech. We don’t believe it's something tacotron does super fancily but picks up on based on the corpus it was trained on. For example, if all the “but”s in a corpus have the pause after that usually comes with the word in spoken English, we theorize that tacotron will also pick up on this small habit.

Talknet

Prosody in text to speech is something that we had never heard of before encountering TalkNet. And transferring prosody from a speaker to TTS was something we had never even thought was possible, the only exposure remotely similar to this were memes like of Barack Obama singing “ Call me Maybe” which used concatenative synthesis that were entertaining but not too technologically impressive. What took this small youtube channel probably a whole days worth of work can now be done by a bored college student with a model and ~20 minutes to spare.

What concatenative synthesis falls victim to when it comes to having tempo added to the voice is that the duration of a word varies wildly when it comes to music. Even in some of the words held for shorter duration in the video linked above showcase this where instead of it being a smooth human voice, the audio is often stretched or shrunk to fit how long a word is held, destroying the human-sounding properties of it.

TalkNet however is not concatenative synthesis, and does an incredible job with both having a smooth flowing human-like voice at world record speeds. When we think of fast speaking we think of rapping, and who else raps faster than Eminem? We used a version of “Godzilla” by Eminem that has the vocals isolated as reference audio as that is what works best as reference audio, no instrumentals for the model to try and ignore. What was generated was Will Styler matching the world record 224 words in 31 seconds (7.24 words/second) by eminem, smooth unlike the spliced audio that would have resulted with concatenative synthesis. You can click here to listen to the results.

Weaknesses of the Tool

Tacotron

A strength of this tool is the simple prosody that comes with it from the data it has been trained on, this is also a weakness. Some of the sample sentences were poorly said in ways that we would not expect a human to say them.

It also doesn’t handle many problems graphemes introduce when trying to synthesize audio from text. We’ve shown the results of the tacotron model to Will Styler, and had implemented it on his own computer to send back the message “You are going to get an A on this project”, but when mentioning the letter grade “A”, tacotron had synthesized it as /ɑ/ instead of /aj/. Tacotron is not state-of-the-art and doesn’t handle similar things like Acronyms, which also tried to pronounce NSA as a word instead of its individual characters.

Another problem that gets minimalized as more audio files are included into the corpus its trained on is metallic static in the background of the synthesized audio. Although it still synthesizes audio that sounds like the speaker, any authenticity goes out the window and lowers the quality of the output.

TalkNet

Making cloned voices sing isn't quite perfect yet. We spoke about having to stretch/shrink the duration of a word and it worked perfectly when we wanted to speak extremely fast. However, when it comes to mimicking singers who hold their notes longer than one or two seconds, it sounds a bit awkward.

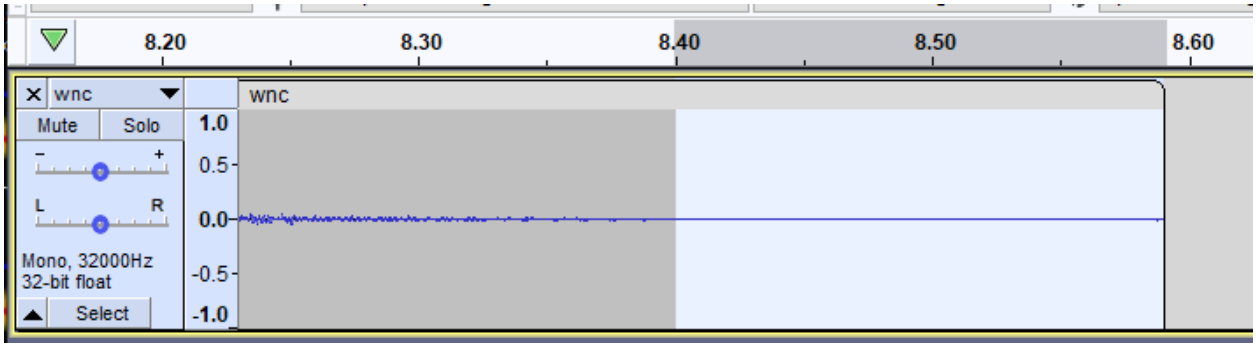
It also has some trouble pronouncing some words, this audio fails to pronounce the word "don't" and instead says it as "daunt". And in an attempt to make it say "around" it became a mess noise at that moment. It seems that of the two layers of Machine Learning this is built (one to process prosody into something that the second layer can synthesize into speech), it lacks in being able to pronounce things.

TalkNet can also have metallic noise in the background of audio as well, but as discussed in "Possible Improvements", this can be remedied in post-processing of the audio.

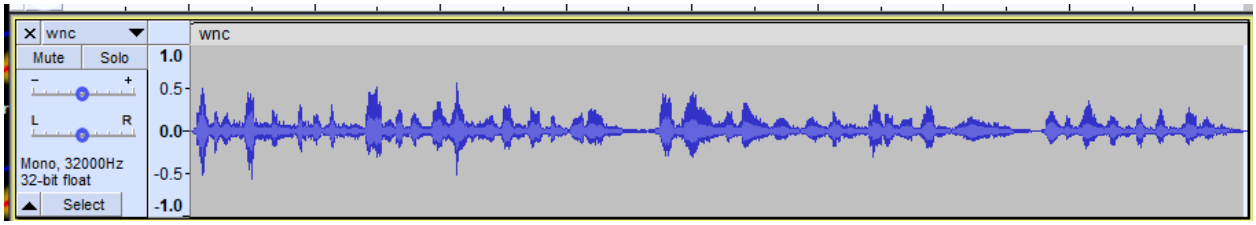
Possible Improvements

When it comes to tacotron2 voice synthesis, between the spaces of words is where you can most clearly hear the metallic static we talked about in the Weaknesses section of this paper. Since its a steady wave of noise, I believed it would be incredibly easy to get rid of it post-synthesis by simply subtracting that wave from the entirety of the file.

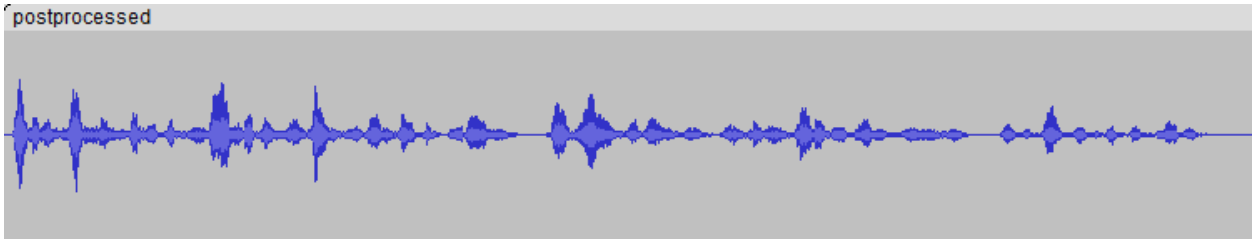
Audacity has a "Noise Reduction" tool which does exactly this. It surprised me since my hypothesis was drawn from what we learned about digital audio in class. We can simply select a section with a pause in the speech (sometimes the synthesized audio has a half-second silent portion at the very end) which isolates the waves audacity will remove from the file. An example is shown below.



The audio highlighted in this picture is the noise to be reduced from the entire file, not visible, but make a lot of difference in the final audio.



Track before noise reduction



And after.

Unprocessed audio:
https://drive.google.com/file/d/1wJWKbshwRO6wBAXg2ANWmUSXnYf15EjL/view?usp=sharing

Noise Reduced Audio:
https://drive.google.com/file/d/1m9HFvvtqEu6QRcN5sOoOquMu0oSWbnqS/view?usp=sharing

A detail that we notice after the noise reduction is that Will's voice sounds like it does in the lecture videos. Like it was spoken through a microphone and echoed through the room. We believe had we used audio recorded directly from a microphone, it would sound identical.

More data would be beneficial to the Will Styler model as well, as better models have around 20-30 minutes of audio and over three hundred audio files in total. Although we have a good chunk of that already, the audio is taken from a lecture where along with Will's voice we have ambience and echo in the training data as well. This could eliminate our static problem, where we do not have to use post-processing to make it sound as best as it can.

Implications of the Advantages and Disadvantages

Tacotron and TalkNet are far from perfect, but they do their job pretty darn well. Good enough to where students falsely believed some of the audios presented to them could have come from a real recording of a person's voice.

This brings us to a heavy topic when it comes to speech synthesis, which is the ethical concerns of it, especially consent of the person whose voice a TTS model is based on and misuse of TTS systems. For this project, after drafting our idea, we made sure to receive consent from Professor Styler to create TTS models based on his voice. Creating these TTS models has shown us that it is easy for anyone with a couple of hours on their hands to create TTS that passes as

authentic, giving anyone the ability to put words in others' mouths. In a world where fake news is already rampant (look to COVID-19 misinformation as an example), TTS systems represent a dangerous new source of misinformation. For example, TTS models could be used to generate fake statements from politicians, framing them for saying certain things. This is a worldwide problem, as models have been trained on different languages besides English. On a more personal scale, this could affect elderly relatives, who are the primary targets of scams such as fake calls. Scam callers could easily pose as a relative, where the disadvantage of TTS audio being produced with static is lost over the poor audio quality of phone calls.

As far as we know, the most that has been done with this synthesis is create memes on the internet. Characters from shows/video games and public figures are rampant on the websites that host TTS for the public like 15.ai mentioned at the beginning of the article. Some of the more paranoid figures who have had their voice cloned requesting their voices to be taken down, but others accept the memes. With regards to that, we end this paper with one last audio we've produced. Listen to it here.

References

Beliaev, S., Rebryk, Y., & Ginsburg, B. (2020). TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model. *arXiv: Audio and Speech Processing*.

Barack Obama Singing Call Me Maybe by Carly Rae Jepsen. (2012, June 4). [Video]. YouTube. https://www.youtube.com/watch?v=hX1YVzdnpEc

Eminem - Godzilla (Vocals Only - Acapella). (2020, January 19). [Video]. YouTube. https://www.youtube.com/watch?v=1qhoTvE118I

Thanks for signing up for Uberduck Voice Cloning. (n.d.). Retrieved June 7, 2022, from https://app.uberduck.ai/uberduck_voice_cloning.pdf