Samuel Jackson AudioFake

Adam Ansari
Daniel Tinoco
Yusong Wang

GOAL

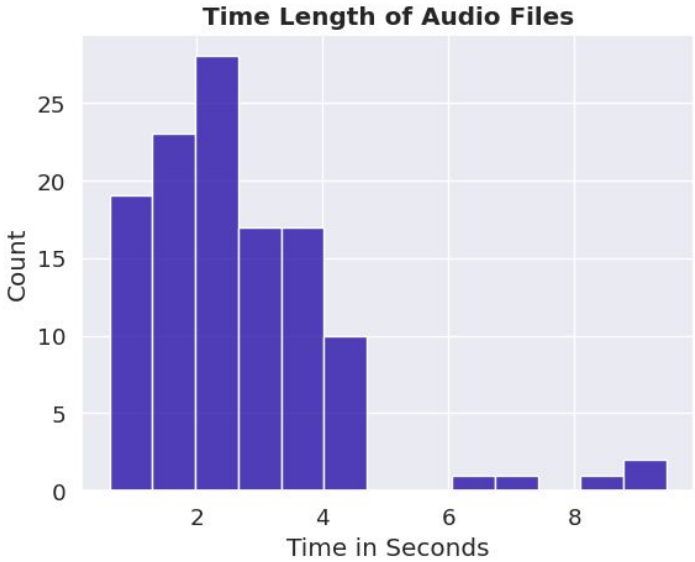
Two of the project team members have their birthdays in July. The stated goal is to recreate the actor Samuel Jackson’s voice in the “audio deep fake” to wish them a happy birthday

METHOD

Utilizing Microsoft’s pre-trained SpeechT5 model[1] and Speechbrain’s pre-trained VoxCeleb model[2], we aim to create speaker embeddings of Samuel Jackson to recreate his voice and generate audio of our choosing.

Data COLLECTION

Sound files collected from 3 separate online sound boards[3,4,5] in mp3 and wav formats. Final dataset contains 158 audio files. Time length is shown below.



Data preprocessing

1. Download/read files using Pydub[6]
2. Change sample rate to 16k Hz, channels to 1, and normalize using Pydub
3. Export audio as .wav files

```python
def mp3_2_wav(mp3_filepath, desired_duration_to_subtract=False, sr=16000):
    # Create wav filepath
    wav_filepath = mp3_filepath.replace('mp3', 'wav')

    # Load mp3 audio file using Pydub
    audio = AudioSegment.from_file(mp3_filepath)

    # For removing ending watermarks
    if desired_duration_to_subtract is not False:
        # Construct duration parameters (optional)
        total_duration = audio.duration_seconds
        subtract_duration = total_duration - desired_duration_to_subtract

        audio = audio[:subtract_duration * 1000]  # Convert to milliseconds

    # Set sample rate and channels to desired specification
    audio = audio.set_frame_rate(sr)
    audio = audio.set_channels(1)
    # Normalize audio to have mean zero
    audio = effects.normalize(audio)

    # Export audio as wav file
    audio.export(wav_filepath, format='wav')
```

Model implementation

- Two models: Speechbrain VoxCeleb, Microsoft SpeechT5

1. SpeechT5:
   - Takes either speech or text as inputs, generates spectrogram as output
2. VoxCeleb:
   - Takes speech as inputs and generates speaker embeddings

Model architecture

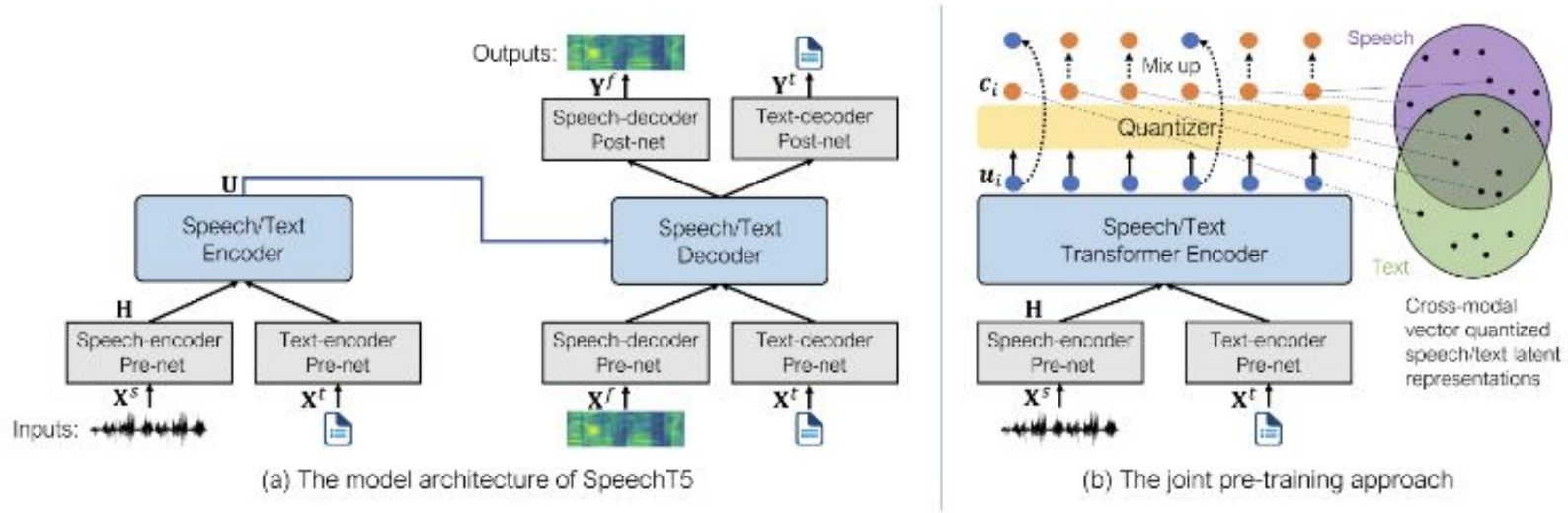


Image Source: https://arxiv.org/pdf/2110.07205.pdf

METHODOLOGY

1. Collect/clean the data
2. Store on Google Drive
3. Load all the files with their corresponding text
4. Process data for VoxCeleb
5. Create embeddings using VoxCeleb
6. Fine-tune pre-trained SpeechT5 on audio files with text and speaker embeddings
7. Create audio samples and test for authenticity in respect to the original voice

Experiments and results

- After 400 training iterations, we checkpointed our model after every 25 epochs
- Of 16 test files, we chose one embedding with the best audio output

Step	Training Loss	Validation Loss
25	1.169100	0.887768
50	0.962800	0.733065
75	0.879900	0.668853
100	0.825100	0.637679
125	0.740400	0.604342
150	0.677700	0.580739
175	0.665400	0.564774
200	0.674700	0.546634
225	0.647500	0.547437
250	0.615800	0.536354
275	0.607600	0.543518
300	0.629500	0.532287
325	0.601800	0.528874
350	0.594200	0.515765
375	0.588100	0.526614
400	0.595700	0.523525

Result (nsfw)

Audio is slightly robotic, but otherwise carries the original actor’s tone.

Conclusion

For the purpose of voice cloning, there are far better/simpler models that exist. However, SpeechT5 does a commendable job in multiple scenarios, such as speech to text, text to speech, speech to speech, and more. As a result, it generalizes to these fields but does not excel in any of them.

REMARKS

The original SpeechT5 model was trained on a dataset of Audiobooks with monotone speakers and no background noise, improving both the encoder and its understanding of language. However, it struggles to capture rougher audio with more tonal variance. We note that when trained on our voices (monotone), we received much better results.

Future work

- Obtain a larger data set for desired speakers
- Clean the data more thoroughly and/or find higher quality data
- Better post-processing of the audio to remove choppiness
- Larger training cycle to accommodate larger data set

Thank you!

REFERENCES

[1] https://huggingface.co/microsoft/speecht5_tts

[2] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[3] https://movie-sounds.org/samuel-l-jackson

[4] https://www.101soundboards.com/boards/10917-samuel-l-jackson-soundboard

[5] https://www.voicy.network/official-soundboards/movies/samuel-l-jackson

[6] https://github.com/jiaaro/pydub