Ewan Lee

Professor Hoover

Introduction to Generative AI

November 6, 2024

**Attain Accurate Accents**

**Description: (100-150 words): A brief description of your project idea and your goals**

When it comes to acting, one of the most important parts of portraying a character is how you speak. I have found that dialect and accents are some of the key parts of convincing the audience and giving them that sense of immersion. This considered, one of the hardest parts about perfecting an accent is having material to work with to mimic. Generative AI has the ability to learn different voices such as characters using models that can learn and manipulate audio files. My project idea is to use that idea and those existing models, training them to learn all of the different accents with the intention of accurately replicating them.

**Background: (200-300 words): Explain why this project is interesting to you with at least three citations of related work (e.g., websites, research papers, blog posts).**

This project interests me because of both the use of text-to-speech models and the inclusion of accents in these text-to-speech models. First of all, text-to-speech itself is a generative model that manipulates sound files and learns to output text into the form of sound which is already very interesting. One such example is ElevenLabs, which has multiple forms of generative sound manipulation. The main ones I will focus on are its text-to-speech and voice cloning aspects which are very accurate and fairly humanlike for generative AI. I find it interesting enough that you can convert text into sound files and create voices with code, and taking it that step further for replicating accents makes the project even more interesting. Within Deep Voice 2 and Fakeyou, there are models using Tacotron in order to have different voices for text-to-speech purposes. For instance, Fakeyou can replicate Emma Stone’s voice, therefore, these text-to-speech models can be trained to have different voices for different accents. One thing that must be considered is the importance of morals and ethics when it comes to a topic such as this. AI itself has to have a very strong sense of ethics, but when it comes to replicating accents, the topic becomes even more sensitive. For this project, the chosen accents must be general accents rather than specific to smaller regions of the world to avoid insensitivity and appropriation. The key difference is the words being used to replicate the accent whereas using generalized phrases stereotypically associated with such accents is insensitive compared to trying to accurately replicate the accent. Overall, for the theatre objective, this project is very interesting and helpful for those who honestly want to learn to copy an accent to create a character from another part of the world.

**Objectives: Define 2-3 specific goals for your project.**

1. Have the code do text-to-speech in general
2. Be able to recreate accurately at least 2 different accents
3. Make it feasible to switch between different accents

**Methodology: Provide a rough plan for how you intend to achieve these objectives (e.g., describe the steps in the tutorial that you are following). This may include the data sources, tools, or frameworks you plan to use and any anticipated challenges.**

My rough plan for achieving these objectives is taking each part of the project as a mini project, and doing them one thing at a time. The first step is to get comfortable with text-to-speech in general and create a model that can do very well in converting text to speech. Then, after testing and being satisfied with the tts, I would next try to have the model learn an accent. The first would likely be something very recognizable like British or Australian. After I manage to see how that turns out, I would then choose another accent to learn, ideally a very different accent from the first such as a more Slavic accent or Southern American. I plan to follow the existing framework for AI voices and text-to-speech models. The main anticipated challenge is whether the different models will be compatible with each other and work together. Using multiple learned voices may be tricky, I may have to use different models based on selected accents or use different training data. The biggest challenge will be to get multiple accents to work, but once two or three are done, the rest will be easy. I may be a little over my head, but I am definitely interested in committing to this project.

**Timeline: Break down your project into weekly goals between the time period of Monday, November 11, 2024 and Wednesday, December 4, 2024. We will start project presentations on December 09, 2024. You can plan for your presentations to be about five minutes.**

| Goal | Deadline |
| --- | --- |
| Have research done and prepare tutorials/code to start coding the TTS accents | November 11, 2024 |
| Finish/be almost finished with at least a TTS model that turns text into speech | November 18, 2024 |
| Polish TTS with one default voice for it to read text accurately | November 22, 2024 |
| Create the first voice with an accent and figure out how to get it to work | November 25, 2024 |
| Finalize the first voice, then work on the second voice to see if the method works | November 29, 2024 |
| Continue testing and debugging the different accents, dare for a third if time permits | December 1, 2024 |
| Try to implement an app interface to easily use the accent TTS model | December 4, 2024 |

**Expected Outcomes**: **Describe what you expect to have at the end of your project. (i.e., will you have built a pipeline for generating your images?)**

By the end of the project, I expect to have at least a bare code that can convert input text into speech and also be able to mimic at least one accent. I have yet to find out how hard it is to train a model to learn different accents to replicate. I know that many different languages can be used, but I wonder how a model performs with different accents for possibly the same voice. I hope to have an interface where you can type out the text and easily switch between different-pitched voices and most importantly between different accents.

Works Cited

Awan, Abid Ali. “5 Projects Build with Generative Models and Open Source Tools.” DataCamp, DataCamp, 21 Apr. 2023, [www.datacamp.com/blog/5-projects-you-can-build-with-generative-ai-models](http://www.datacamp.com/blog/5-projects-you-can-build-with-generative-ai-models).

ElevenLabs. “Free Text to Speech & AI Voice Generator.” ElevenLabs, <https://elevenlabs.io/>. Accessed 4 Nov. 2024.

“Fakeyou Celebrity AI Voice and Video Generator.” *FakeYou. Deep Fake Text to Speech.*, [fakeyou.com/tts](http://fakeyou.com/tts). Accessed 4 Nov. 2024.

Gibiansky, Andrew, et al. “Deep Voice 2: Multi-Speaker Neural Text-to-Speech.” Advances in Neural Information Processing Systems, 1 Jan. 1970, [proceedings.neurips.cc/paper\_files/paper/2017/hash/c59b469d724f7919b7d35514184fdc0f-Abstract.html](http://proceedings.neurips.cc/paper_files/paper/2017/hash/c59b469d724f7919b7d35514184fdc0f-Abstract.html).

Manobharathi93. “Create Your Own Text-to-Speech Engine with Tacotron2 and Pytorch Lightning.” Medium, Medium, 7 July 2023, [medium.com/@immanobharathi21/create-your-own-text-to-speech-engine-with-tacotron2-and-pytorch-lightning-662f3301cdc9](http://medium.com/@immanobharathi21/create-your-own-text-to-speech-engine-with-tacotron2-and-pytorch-lightning-662f3301cdc9).

“Tacotron 2.” PyTorch, [pytorch.org/hub/nvidia\_deeplearningexamples\_tacotron2/#:~:text=The%20Tacotron%202%20and%20WaveGlow,text%20using%20encoder%2Ddecoder%20architecture](http://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/#:~:text=The%20Tacotron%202%20and%20WaveGlow,text%20using%20encoder%2Ddecoder%20architecture). Accessed 4 Nov. 2024.

tttzof351. “Build Text-to-Speech from Scratch.” Medium, Medium, 16 Aug. 2023, [medium.com/@tttzof351/build-text-to-speech-from-scratch-part-1-ba8b313a504f](http://medium.com/@tttzof351/build-text-to-speech-from-scratch-part-1-ba8b313a504f).