1 Lab Book S5

In this lab book you will work with gTTS (Google Text-to-Speech), a Python library and CLI tool to interface with Google Translate's text-to-speech API. You can find the documentation here.

1.1 Assignment 1

[10pts] *gTTS: Synthesizing a dialogue with two speakers*:

Installing and trying out GTTS

GTTS can be installed out of the box. For Debian-based Linuxes you can use

```
apt install python3-gtts
```

(if you are using Google colab, prefix the shell commands with an exclamation mark, e.g. `!apt install python3-gtts` or `!pip install gtts`).

Task:

Let's imagine you need to synthesize a phone call dialogue between two speakers and at the same time you were asked to do some pre-processing and censor some words by substituting them with some other things.

For this task you will need to generate a transcript using chatGPT. Ask it to generate a dialogue between two friends. These two speakers should take turns in their conversation several times. Decide on the topic of the dialogue yourself: it could be pets, school subjects, food items - anything you want, but make sure to specify that pets/subject/food items/whatever would have strange names. It should look something like:

transcript = "'Friend 1: Hey! How's it going?

Friend 2: Hey! I'm good, thanks. Just got back from the park with Zoozoo. He had a blast chasing squirrels.

Friend 1: Zoozoo always has so much energy! I wish Chancellor Zorg had half his enthusiasm.

Friend 2: Well, you know how it is with labs. They're like perpetual motion machines. How's Chancellor Zorg doing?"'

Below is the step-by-step task description with some suggestions:

1. Given your conversation transcript, you will need to synthesize spoken dialogue using gTTS with alternating speaker utterances. First step could be to split your transcripts by speaker turns and run TTS on each utterance.

   However, before generating speech, you need to do some customization on your transcript using the gTTS' Speech corrections (word substitution). You will need to substitute those strange names of the pets/subjects/foods/whatever with nice things of your choice (e.g., "fluffball", "sunshine", or whatever you want whatever brings you joy).

2. The next step could be creating the temporary sequential audio fragments by iterating through enumerated list of utterances and generating one of the two different variants of English voices depending on the speaker. For example, you could take Australian English speaker voice for friend 1 and UK English speaker voice for friend 2 (or choose two other voices from the ones that described in the documentation). Something like:

```
if speaker == "friend_1":
  tts = gtts.gTTS(utterance, lang="en", tld="com.au")
  tts.save(outputfile)
else:
```

```
tts = gtts.gTTS(utterance, lang="en", tld="co.uk")
tts.save(outputfile)
```

3. Then you will need to combine your audio files of utterances into one recording of a dialogue (e.g., you can use librosa and pandas for the task).
4. Bonus part [2bpts]. If you want to make the task a bit more challenging, you can imitate a phone conversation[a] and make each speaker's recordings channel specific.

Submission

Alongside with this LaTeX file with your reflections (if any), commit the complete code you used for the task as well as the transcript of the dialogue if you used a separate file to store it. You can also commit your resulting audio file, but I expect it be generated by your code when I run it. Make sure the code works "as is" and doesn't require editing paths, providing commandline argument and so on.

[a] this would be more common in call centers

Answer

I reflect on the synthesized speech from the following aspects:

- Quality of the Synthesized Audio
  - Intonation:
    - Incorrect Pitch on Question Intonation ("Moonlight Mushrooms?"): The pitch for "Moonlight Mushrooms?" did not rise sufficiently at the end of the sentence, which is typical for a question in English. Instead of indicating a question with a rising intonation, the pitch fell flat, particularly on the "rooms" syllable of "mushrooms," making it sound less natural.
  - Pause:
    - Unnatural Pause ("They're nothing like Spaghetti Squirgle, though."): The word "though" was pronounced too distinctly and separated from the rest of the sentence, which made it sound unnatural and stilted. In natural speech, "though" often flows more smoothly and is connected closely with the preceding words.
    - Incorrect Phrasing and Pausing ("made of moonberries and then grilled"): The synthesized voice did not correctly phrase "moonberries and then grilled." Instead of pausing after "moonberries," gTTS ran "moonberries and" together and then paused before "then," which is against the usual rhythm and phrasing of English speech. This led to a misunderstanding of the sentence structure and a less natural flow.
- Improvements and Future Work
  - Weak Forms and Stress Patterns: Some common words like "and" and "them" should be pronounced as weak forms in connected speech to reflect natural English pronunciation patterns. However, gTTS system used did not seem to implement this feature, leading to an overly precise articulation of these words where a more subdued, natural pronunciation would be expected. Future work could include implementing or improving the TTS system's handling of weak forms and stress patterns to enhance the overall naturalness and intelligibility of the synthesized speech.