Cross Lingual Speaker Adaptation for TTS Applications

Software Project-2021/2022

Rasul, Anna, Claesia, Sharmila

Outline

- Idea reminder
- Finished tasks
  - Preprocessing pipeline
  - Running Grad-TTS on Grid 5000
  - Prepare data for training
  - Web app update
- Plans
- Possible issues
- Timeline

Project idea

Grad-TTS modified for multilingual TTS:

EN text, EN embedding, FR speaker embedding ⇒ EN speech in FR speaker voice

FR text, FR embedding, EN speaker embedding ⇒ FR speech in EN speaker voice

Preprocessing pipeline (FR)

Wrapper to get a sequence of numbers that represent phonemes for French text.

```
sharmila@sharmila-OMEN-Laptop-15-ek0xxx:/mnt/classes-so/Uni-of-Lorraine-Winter-Semester/Software-Pr
Input: salut je vais bien
./get_phonemes.pl tmp/input.txt texts hts run > tmp/output.txt
Output [27, 5, 32, 9, 31, 12, 29, 4, 23, 17, 13]
```

Grad-TTS on Grid 5000

Grad-TTS is the main model we will use for training.

Goal: set up the training and inference pipeline in Grid 5000.

Result: both training and inference work in the Grid 5000 without any problems.

Prepare data for training

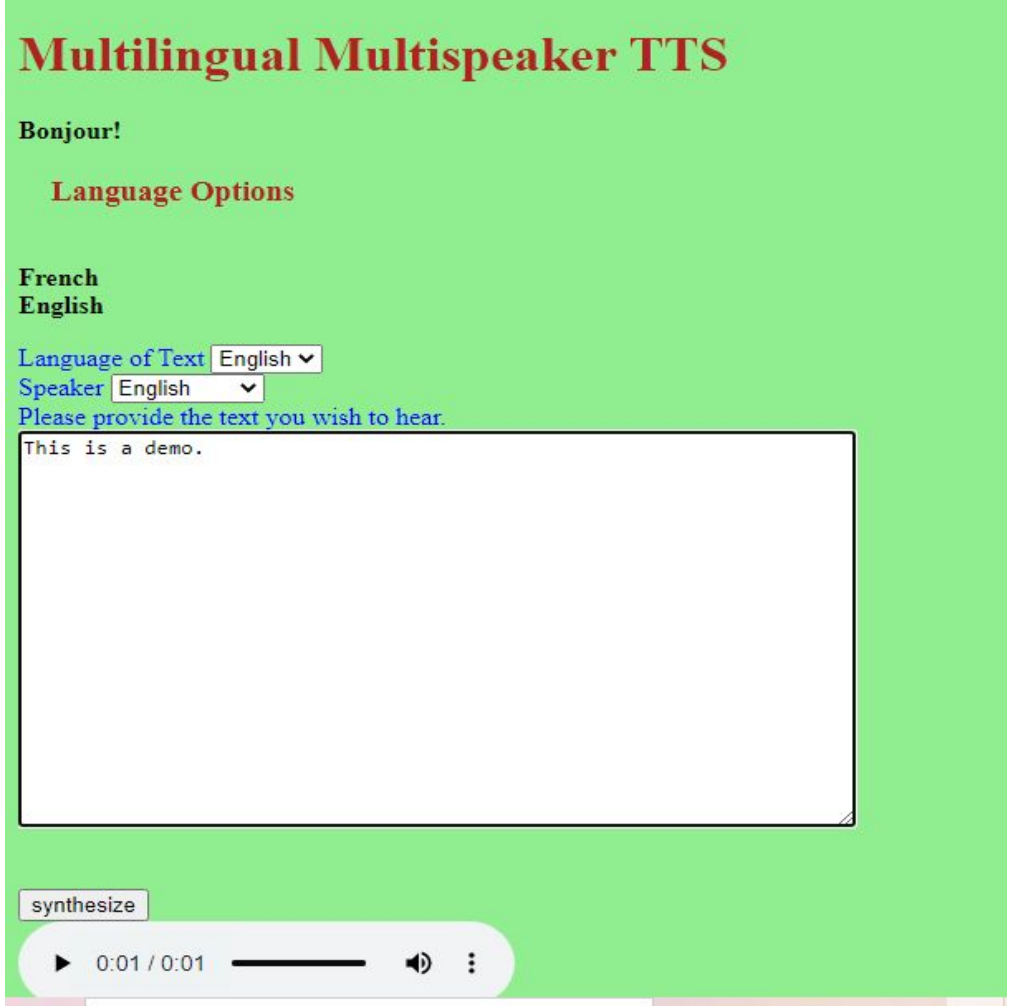
1. Feature extraction from the audio files into .npy files (MEL frequencies)
2. Creation of a text file with the defined structure: input to training

```
features_file | phoneme_integers | speaker_id | emotion_id | lang_id
```

```
11  SiwisFrenchSpeechSynthesisDatabase/wavs/part1/neut_parl_s05_0169.npy|25,2,17,16,35,12,32,5,1,1,5,20,5,30,5,20,32,2,22,5,1,12,35,16,20,32,9,1|0|0|0
12  SiwisFrenchSpeechSynthesisDatabase/wavs/part1/neut_parl_s02_0671.npy|32,5,20,5,33,6,32,3,21,5,34,12,27,17,10,20,17,4,33,5,32,13,34,9,4,1|0|0|0
13  SiwisFrenchSpeechSynthesisDatabase/wavs/part1/neut_parl_s03_0394.npy|21,8,32,12,34,16,24,27,5,1,1,15,25,2,28,1|0|0|0
14  SiwisFrenchSpeechSynthesisDatabase/wavs/part1/neut_parl_s01_0467.npy|35,8,27,6,34,24,16,22,32,6,31,2,22,31,9,27,22,5,1,1,7,23,8,1|0|0|0
```

Web App

https://lctproject.eu.pythonanywhere.com/



Plans

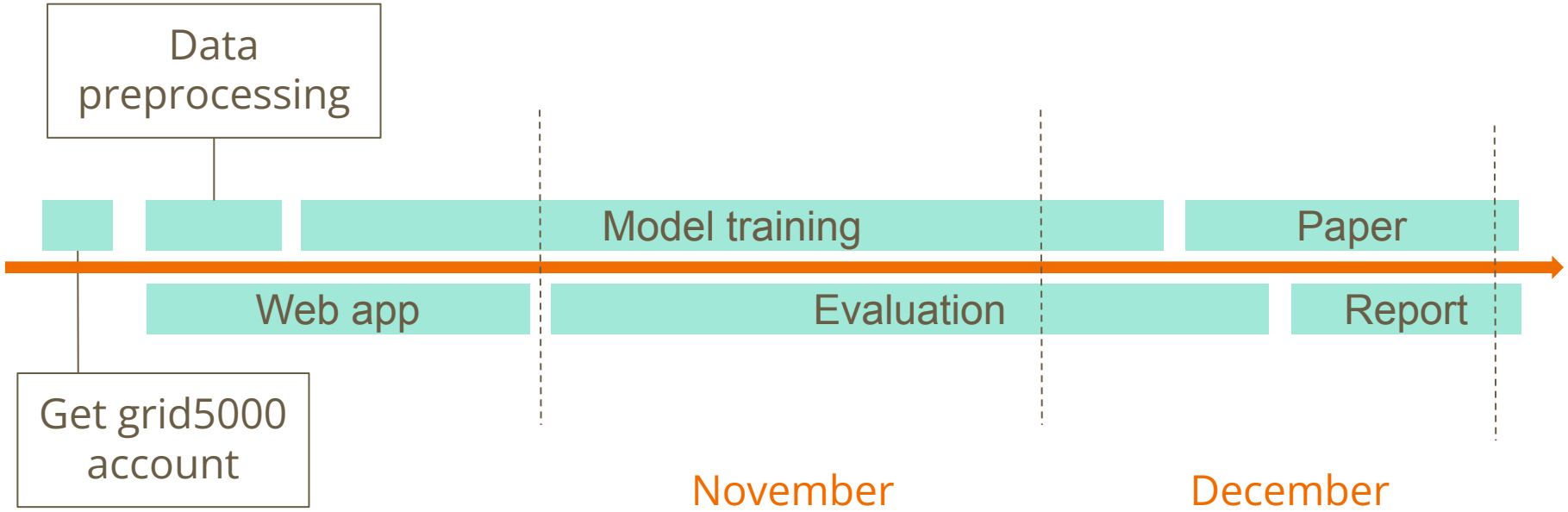
- Start the training
  - Several questions left to discuss with our supervisor before running it
- Evaluate the first results
- Improve the model
  - When we see the first results, we will be able to analyse what can be done to make it better

Possible issues

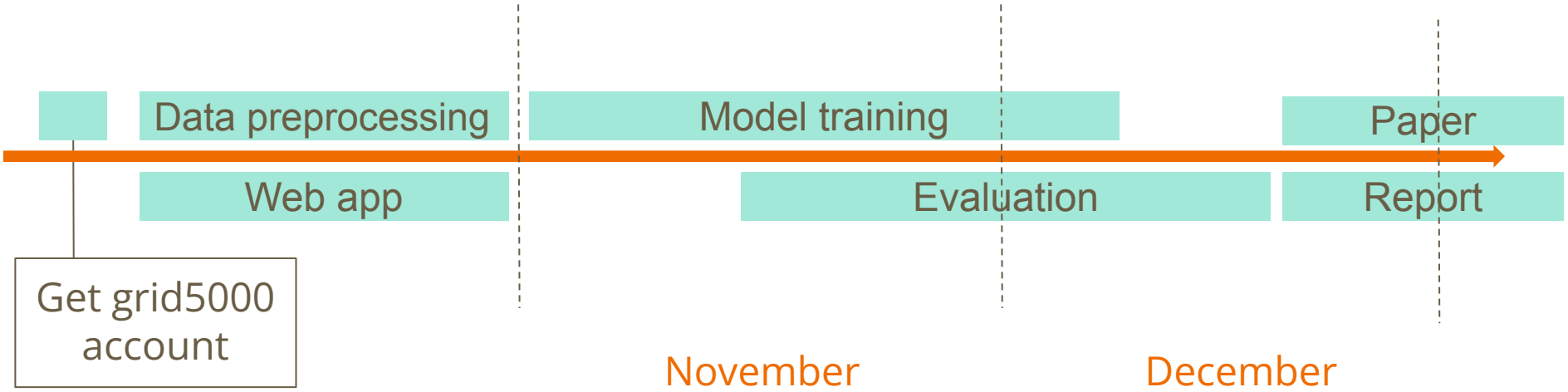
Problem: we don't know how model distinguishes the characteristics of the speaker and language from the same speaker.

Solution: use new dataset, VCTK: https://datashare.ed.ac.uk/handle/10283/2950

Timeline (previous version)



Timeline (new one)



Thank you! Any questions?