Speemo ASR+SER HPC Integration Documentation

Prepared for: Username mk696056
Container Registry: ghcr.io/martinkhadjo/speemo_asr_ser_models_for_audio_files_and_real_time_inference:0.1.0

1. Push Your Docker Image to GHCR
   a. In PowerShell:
      $Env:GHCR_PAT = "<YOUR_TOKEN>"
      echo $Env:GHCR_PAT | docker login ghcr.io -u MartinKhadjo --password-stdin
      docker tag speemo:0.1.0 \
        ghcr.io/martinkhadjo/speemo_asr_ser_models_for_audio_files_and_real_time_inference:0.1.0
      docker push ghcr.io/martinkhadjo/speemo_asr_ser_models_for_audio_files_and_real_time_inference:0.1.0

2. SSH into RWTH HPC & Load Apptainer
   a. ssh -l mk696056 login23-1.hpc.itc.rwth-aachen.de
   b. module avail Apptainer
   c. module load Apptainer

3. Pull the SIF on the Cluster
   a. mkdir -p $HPCWORK/containers
   b. apptainer pull \
      $HPCWORK/containers/speemo.sif \
      docker://ghcr.io/martinkhadjo/speemo_asr_ser_models_for_audio_files_and_real_time_inference:0.1.0

4. Stage Data & Models
   ln -s $HPCWORK/data $HPCWORK/containers/data
   ln -s $HPCWORK/models $HPCWORK/containers/models
   ln -s $HPCWORK/Archive $HPCWORK/containers/Archive

5. Launch Flask Dashboard
   source activate fastx_env
   apptainer exec --nv \
     --bind $HPCWORK/speemo_md_0.005:/workspace \
     $HPCWORK/containers/speemo.sif \
     python run.py serve --host 0.0.0.0 --port 5000

6. Tunnel to Local Browser
   ssh -L 5000:localhost:5000 mk696056@login23-1.hpc.itc.rwth-aachen.de
   Open http://localhost:5000 in your browser

7. Submit SLURM Jobs via the UI
   Fill in form on dashboard and click Start Training
   Monitor logs & status in the web UI