Final Report

29.03.2025

General

My task was to extract new training data for Paule from the Common Voice dataset (cv-corpus-18.020240614). More specifically:
mnt/Restricted/Corpora/CommonVoiceVTL/corpus_as_df_mp_folder_{language} | en or de

I did not complete the training of the models, but there is a repository for synthesis with some documentation, so it should be possible to use it without problems for new data in a similar style: https://github.com/quantling/create_vtl_corpus

Data

Data was generated for both German and English, and the German data was split approximately 80/10/10. Each word should definitely appear in the test set, which means that this might be more difficult than validation and training. For the English data, storage space could become an issue, as it has not been split yet.

Furthermore, the data was cleaned, but additional errors may occur. The code for training and splitting can be found here: https://github.com/JoJoBarthold2/training

Training

I also tried to build training for the Embedder, the Inverse, and the Forward Model. These seem to work reasonably well, at least the loss for the German mini-corpus is decreasing. Training time for the entire training data takes very long, approximately 13 minutes per dataset for the Forward Model, faster for the Embedder. When I was unsure, I simply adopted the specifications from Paul's script, which is also in the training repository mentioned above.

TODOs

One should run the training, check if it works, and then adjust and improve the scripts. This is currently happening for all models. It will probably take half a year for 200 epochs, so really long. Each epoch is saved. There is the possibility to reload the optimizer and model.

The final step would be training the GANs and ultimately combining the components to validate the complete model for Paule.

Contact

Although at this point I am (probably) no longer working for Quantling, you can contact me with questions via Github or my website.

signed, Valentin Schmidt

http://valentinschmidt.eu