THM | CAMPUS GIESSEN | MNI
Mathematik, Naturwissenschaften und Informatik
TECHNISCHE HOCHSCHULE MITTELHESSEN

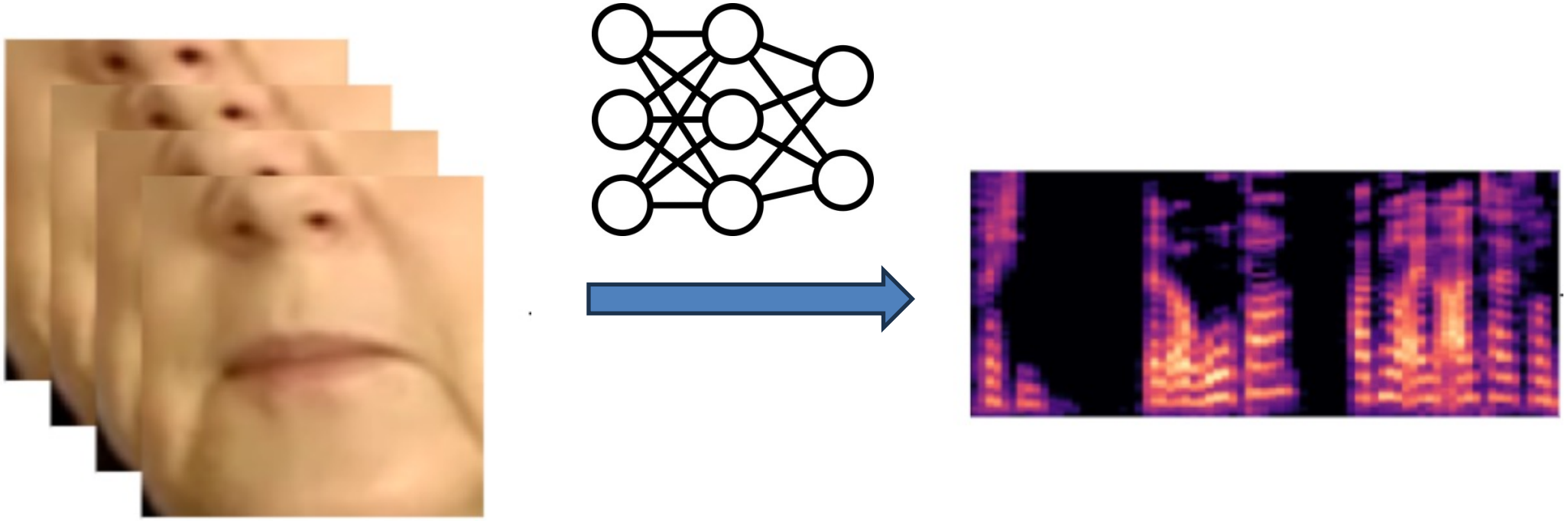
DL4CV Projekt: Training eines Lip-to-Speech Modells in Deutscher Sprache

Gian Saß

Agenda

- Problemstellung
- Datensatz-Erstellung
- Modell-Architektur
- Training
- Probleme
- Fazit

Problemstellung

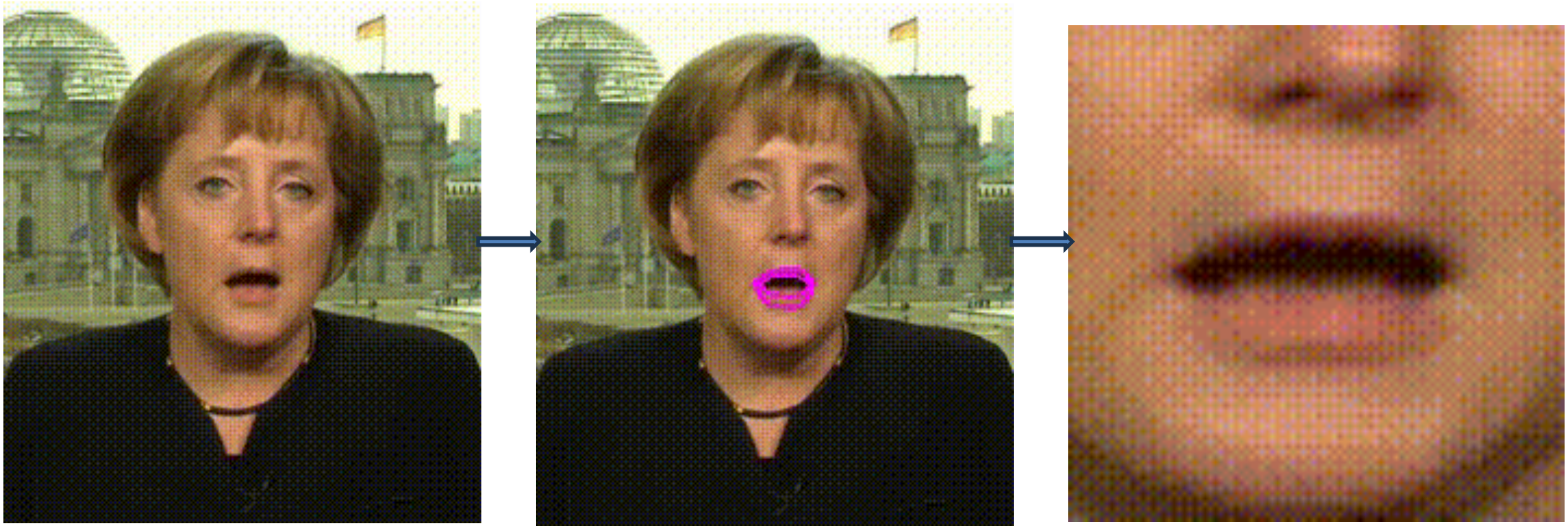


Eingabe: Videofeed einer sprechenden Person
Ganzes Gesicht bzw. **Mundregion**

Ausgabe: Sprache
Audio-Zeitsignal oder **Mel-Spektrum**

Datensatz-Erstellung

Video-Preprocessing mit moviepy und MediaPipe



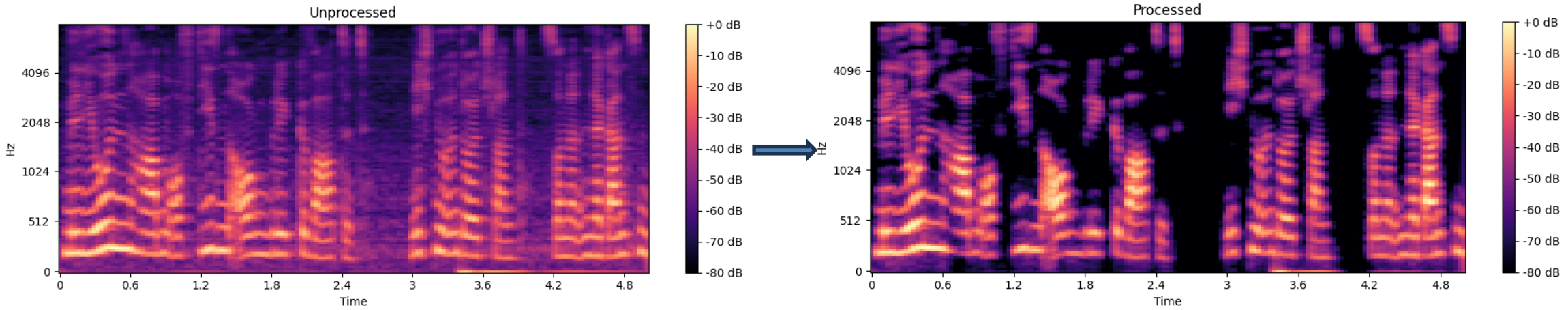
Clip selection

Landmark detection

Cropping

Datensatz-Erstellung

- Audio-Preprocessing mit librosa, noisereduce und Normalisierung

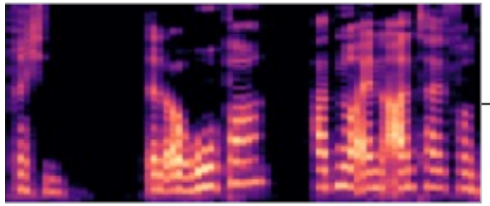


Mel Spectrogram ohne Denoising

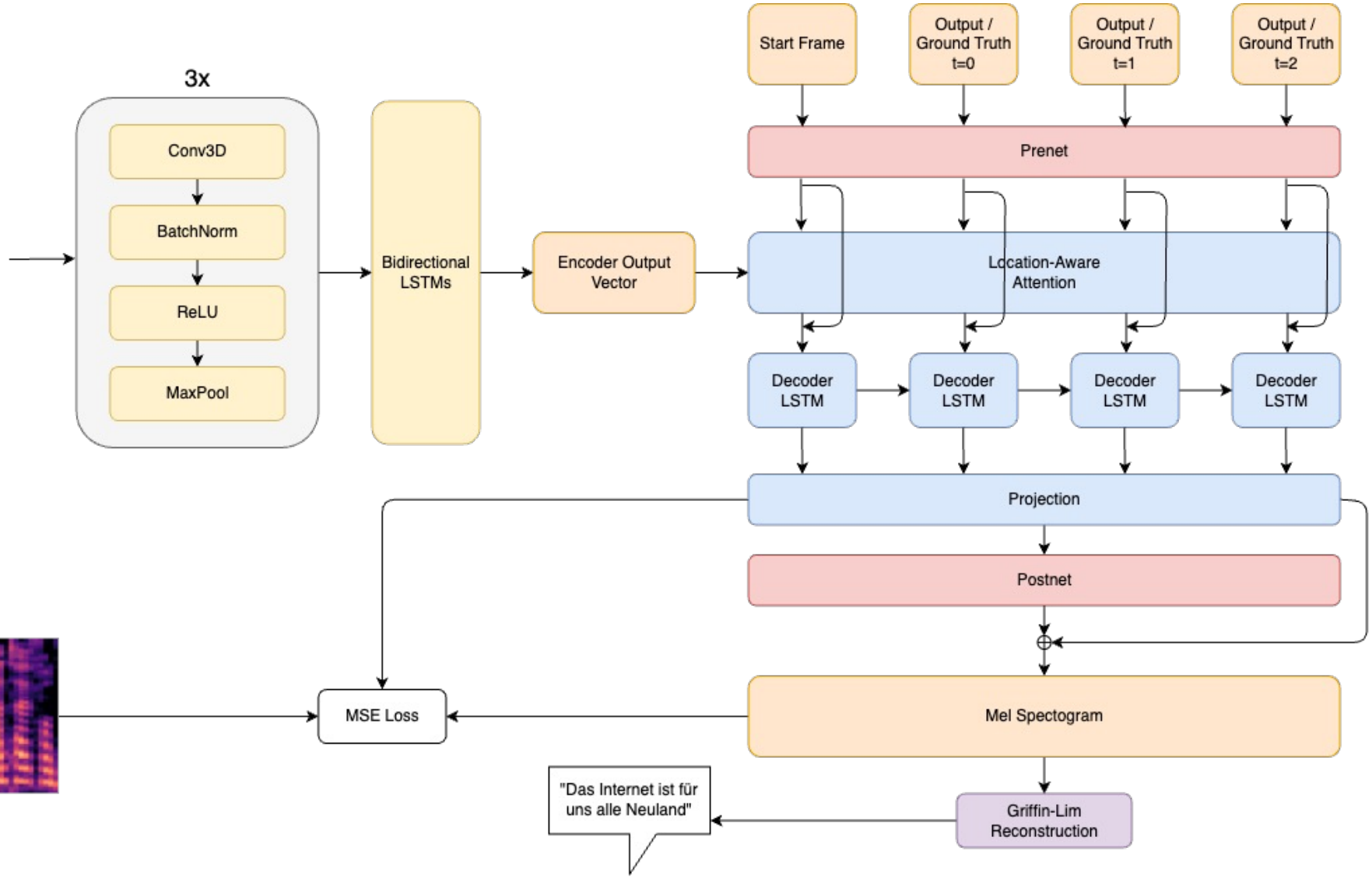
Mel Spectrogram mit Denoising

Modellarchitektur

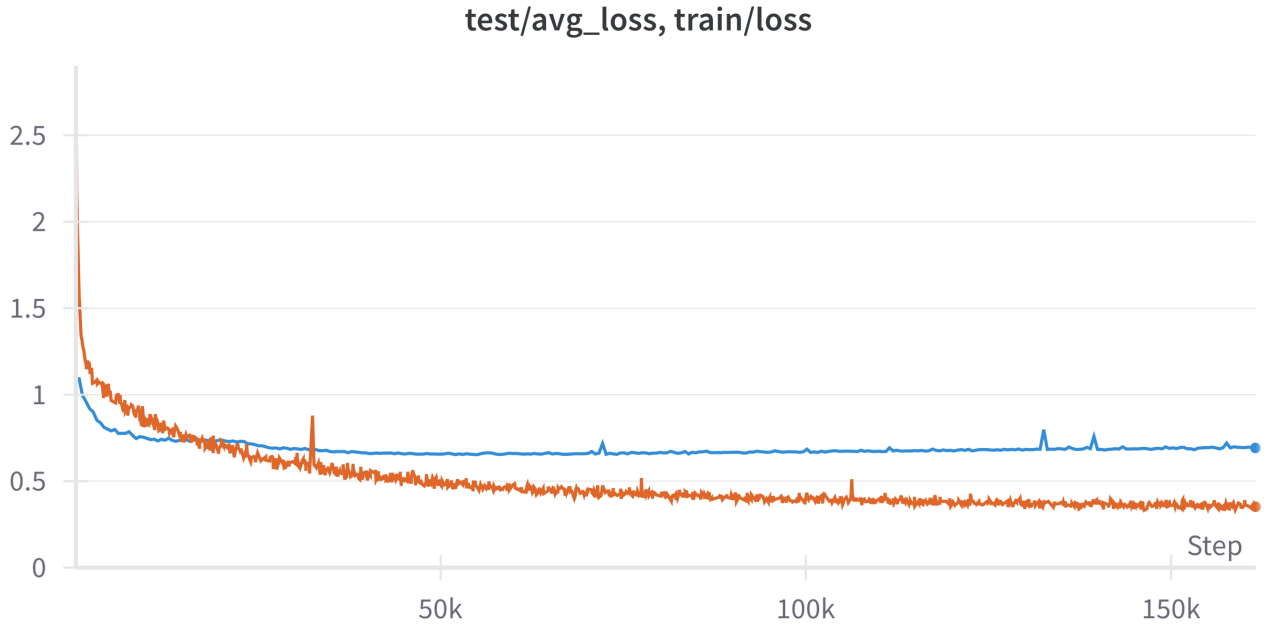
Cropped Mouth Regions
(B, 3, T, H, W)



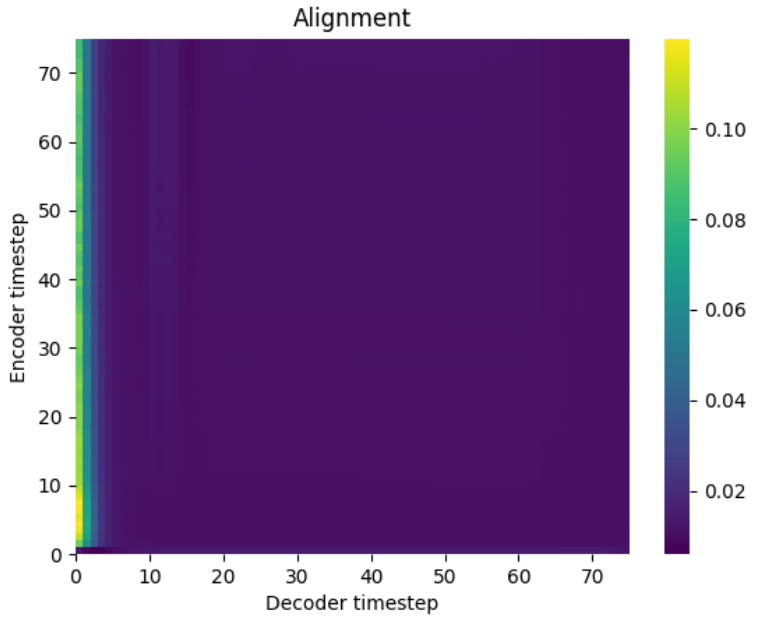
Denoised Ground Truth Mel Spectogram
(B, T, n_mels)



Training



Rot: train loss, blau: validation loss



Attention Alignment in ~30k steps

Metriken

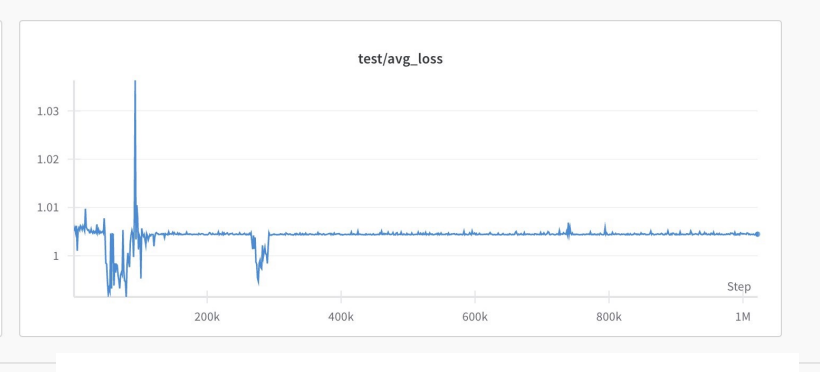
- eval.py: Berechnung von
  - STOI: Short-Time Objective Intelligibility
  - ESTOI: Extended Short-Time Objective Intelligibility

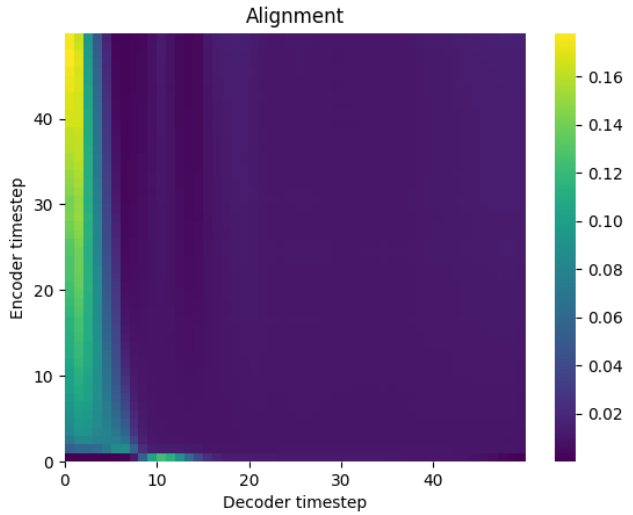
Datensatz	STOI	ESTOI
Merkel Podcast Corpus	0.54	0.318

Demo

Probleme

- Manchmal nicht sicher gewesen, ob Code oder Daten das Problem waren
- Preprocessing dauert sehr lange (~8 Stunden)
- Modell schafft es nicht, Alignement zu lernen
  - Lösung: Prenet anpassen
- Nicht genug Daten gehabt





Fazit

- Lip-to-speech ist schwer und stark abhängig vom “Speaker” selbst
- Datenvorbereitung braucht viel Geduld
- Auswahl des Datensatzes sehr wichtig
- Schwer neuronale Netzwerke zu ”debuggen”
- Einiges dazu gelernt