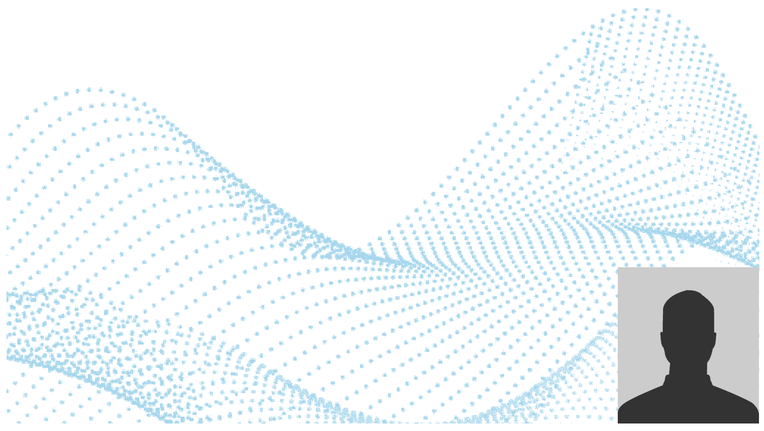
THM TECHNISCHE HOCHSCHULE MITTELHESSEN | CAMPUS GIESSEN | MNI Mathematik, Naturwissenschaften und Informatik

MerkelNet: Training a Self-Supervised Lip-to-Speech Model in German

Gian Saß



1. Introduction

The lip-to-speech task involves taking in a soundless video of a person talking as input and reconstructing the original speech. This may be useful where a noisy sound recording renders a person's speech unintelligible, or recovering speech from silent CCTV recordings, etc. Notably, this involves many challenges. First, the lip movements and speech have to be in synchronization with each other, and lip movements may contain ambiguities, as several phonemes can be recognized from the same sequence. It may not be entirely possible to reconstruct speech from only lip or face movements, since a significant portion of speech generation is done internally, for instance, in the vocal cords and pharynx.

2. Architecture

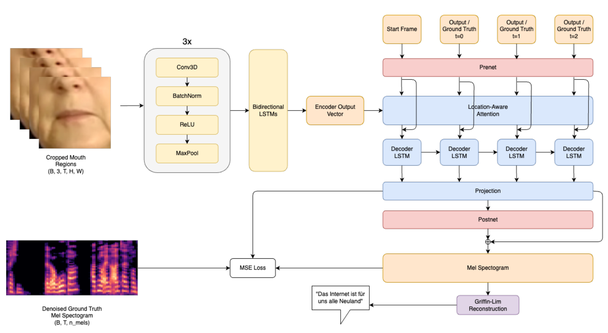


Figure 1: Overview of model architecture.

The model presented in this project report tries to leverage the natural co-occurrence of lip movement and audio without requiring any manual annotations. MerkelNet employs an **encoder-decoder** architecture which is trained end-to-end. The encoder consists of a stack of 3D convolution blocks to learn an **implicit lip feature representation** followed by a bidirectional LSTM to model long term dependencies in the temporal dimension. The decoder consists of an additive attention mechanism coupled with a single bidirectional LSTM cell, and a projection layer. The mel spectrogram is **decoded auto-regressively** by conditioning it on the previous output (during training the ground-truth is used). A pre-net is used to bottleneck the information from the previous step, as this is crucial for attention alignment. Finally, a post-net adds a residual to the resulting mel spectrogram to smoothen the transitions between individual frames.

3. Training

For training, random clips are selected from the training dataset. The Adam [2] optimizer is used with a learning rate of 10^{-3} and a batch size of 32. Training was done on the THM Slurm Cluster on a NVIDIA V100 and on the RunPod service running on a NVIDIA A100. The platform Weights & Biases is used to track training metrics. The model is trained until the average test loss plateaus for at least 50k steps, for a total of about 150k steps or 300 epochs.

During the validation phase that occurs after every training epoch, the attention weights per decoder timestep are collected and plotted as a heatmap. This is to keep insight into the model's alignment. Linear alignment between encoder and decoder starts to emerge around 15k steps and seems to firmly manifest at around 30k steps.

4. Results

Evaluation metrics are determined using the latest model checkpoint on a random validation dataset. The average Short-Time Objective Intelligibility [5] (STOI) and Extended Short-Time Objective Intelligibility [1] (ESTOI) measures are calculated between the ground truth waveform and the inferenced waveform using the pystoi [3] library, after converting both mel spectrogram representations back to their waveform representations using the Griffin-Lim reconstruction. While a **STOI measure of 0.54** is okay, the **ESTOI measure of 0.318** is rather poor, hinging on the verge of unintelligibility.

On selected random data samples the model performs rather poorly, however the model clearly **demonstrates solid alignment between lip movements and speech**, e.g. there is no speech during moments of pause. Also, some frequent words like "Deutschland" and "Bundestag" are clearly discernible.

5. Conclusions

- **Amount of Data Crucial:** While the model shows promising results, there's clearly a lack of data, as the Merkel Podcast Corpus only compiles to about 17 hours worth of usable footage (as returned by the dataset preprocessing script). Meanwhile projects like Lip2Wav [4] are trained on about 10x the same amount of data. It is assumed that a greater volume of data would yield better results.
- **Prenet crucial for learning attention:** Training experiments have shown that the density of the prenet layer is important for the model to learn attention. If the prenet does not sufficiently "compress" the previous mel vector the model starts to rely too much on the previous decoder output and not sufficiently enough on the decoder output. This can be clearly seen in the attention alignment heatmap, as the model only pays attention to the very beginning or very end of the encoder output. Notably the amount of dropout in the prenet layer is important. The model uses a prenet dropout of 0.5.

References

[1] Jesper Jensen and Cees H. Taal. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, November 2016. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[3] Pariente Manuel. mpariente/pystoi, March 2024. original-date: 2018-04-18T12:01:22Z.

[4] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C. V. Jawahar. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis, May 2020. arXiv:2005.08209 [cs, eess].

[5] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, March 2010. ISSN: 2379-190X.