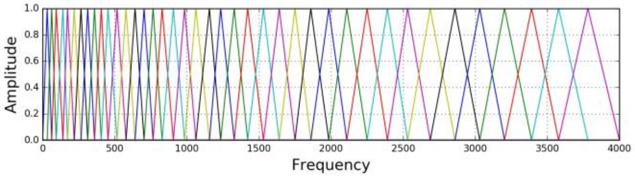
Adaptive Filterbanks using Autoencoders

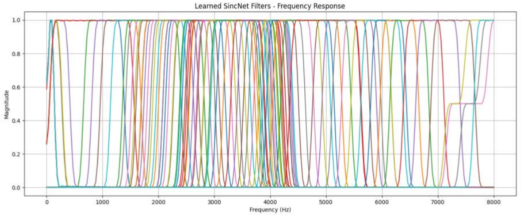
Robert Viehweg

Fixed vs Adaptive Filterbanks

- Fixed filterbanks for example Mel-Filterbank
  - Drawback: not specialized for holistic audio tasks



- Adaptive filterbanks, trained on specific data
  - Advantage: highly adabdable for specific tasks

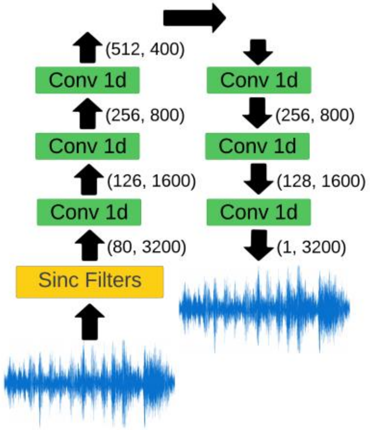


Project setup

1. Link SincNet adaptive filterbank to autoencoder architecture
2. Train model on speech data
3. Evaluate recreated audio
4. Train the same model on fixed Mel-Filterbank
5. Compare fixed and learned Filterbanks

Models

- Autoencoder:



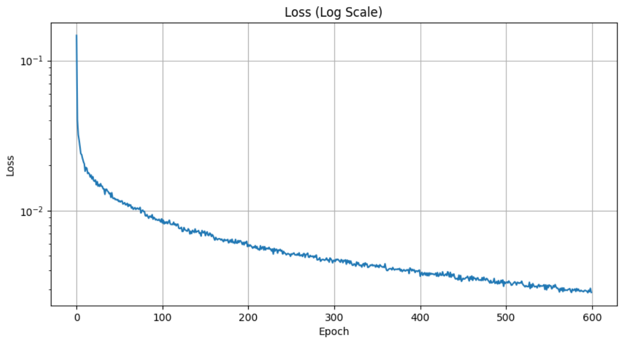
- SincNet Filters (80 filters of length 251 samples):
  - Low and high cutoff frequencies learned during training

$$g[n, f_1, f_2] = 2f_2 \operatorname{sinc}(2\pi f_2 n) - 2f_1 \operatorname{sinc}(2\pi f_1 n)$$

Training

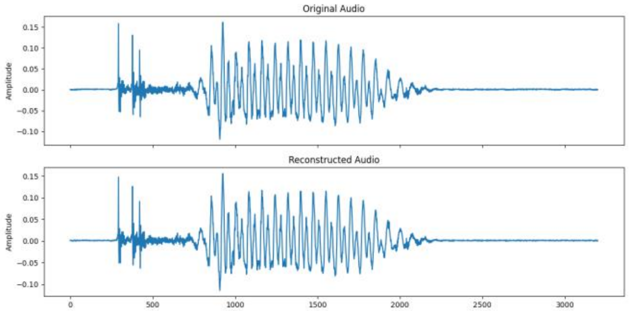
- Dataset: DAPRA TIMIT Acoustic-Phonetic Continuous Speech Corpus
  - Random speech audio snippets of length 3200 samples

- RMSprop optimizer trains both models to recreate the input audio snippet
  - Learning rate scheduler
  - MAE loss
  - 600 epochs (1h and 12 seconds)

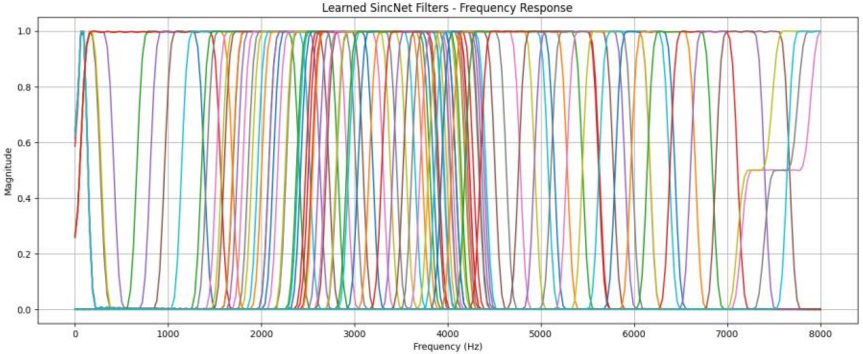


Audio Reproduction

- Trained model achieved good audio reproduction

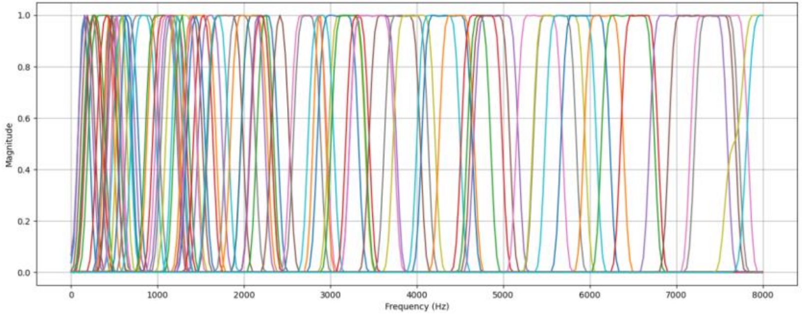


- Learned Filterbank concentraded in the center frequencies (most prominent in speech)



Comparison Fixed and Learned Filters

- Second model trained with fixed Mel-scale Filterbank



- Comparison by computing loss of output and input audio for both models 1000 times

Results

- Average MAE fixed Filterbank: 0.00259
- Average MAE adaptive Filterbank: 0.00215

- The adaptive Filterbank achieved better audio reconstruction measured with MAE
  - The reason could be highly specialized filters for speech data, while Mel-scale filters are more holistic

Discussion

- MAE potentially not sufficient for capturing differences in audio as it is a purely time-domain approach
  - Better alternative: Spectral Loss or Log-Magnitude Spectral Loss

- Worth experimenting with other model architectures or modifications
  - Upscaling of layers
  - Increasing training time

- Adaptive Filterbanks could also be trained on more holistic audio data to compare with fixed Mel-scale filterbank, which is adaptive to a wide field of audio tasks