Speech Synthesis with Densely Connected 3D Convolutional Neural Networks from ECoG

Miguel Angrick, Christian Herff, Emily Mugler, Marc Slutzky, Dean Krusienski, Tanja Schultz

Cognitive Systems Lab

OLD DOMINION UNIVERSITY

Northwestern University

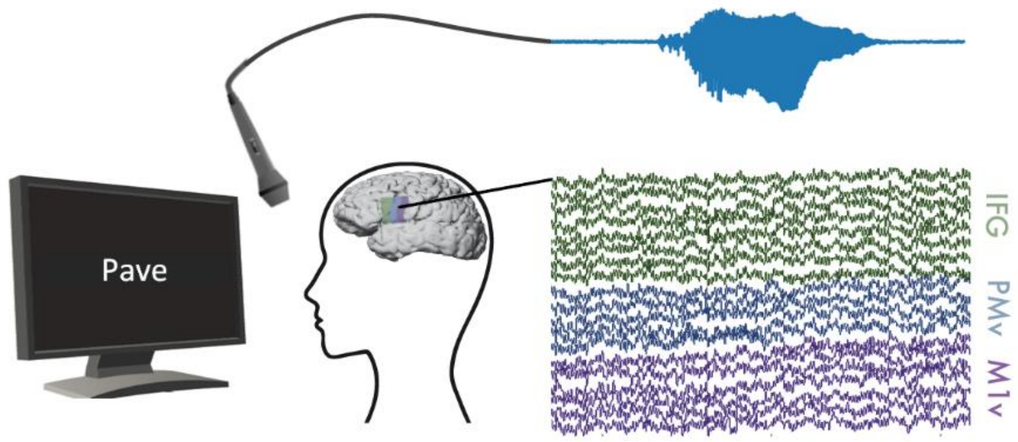
miguel.angrick@uni-bremen.de

Motivation

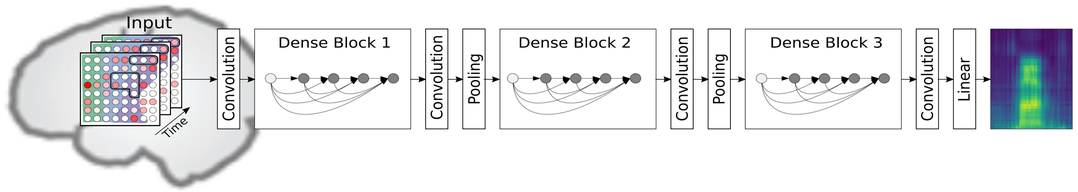
- Mute users and speech impairments
  → Reconstruct speech from brain activity data
- Underlying processes during speech production
  → Complex spatio-temporal dynamics
  → Linear mapping not applicable
- Convolutional Neural Networks
  → Non-linear model
  → Capture spatio-temporal patterns
- ECoG as suitable method
- Utilize Wavenet as vocoder
  → Trained on separate single speaker dataset
  → Locally conditioned on spectral features (LogMels)

Experiment Design

- Six patients undergoing awake craniotomy for glioma removal
  → ECoG grid with 8x8 electrodes
- Covered regions
  → Premotor cortex (PMv)
  → Motor cortex (M1v)
  → Inferior temporal gyrus (IFG)
- Speech production task
  → Read displayed words aloud
- Simultaneous recording of audio and ECoG
- Acoustic data preprocessing
  → Transformed into 40 LogMels
- Neural data preprocessing
  → Extraction of high gamma activity
  → Feature stacking for temporal informations



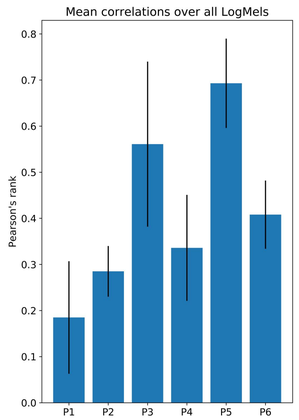
Convolutional Neural Network for Speech Synthesis

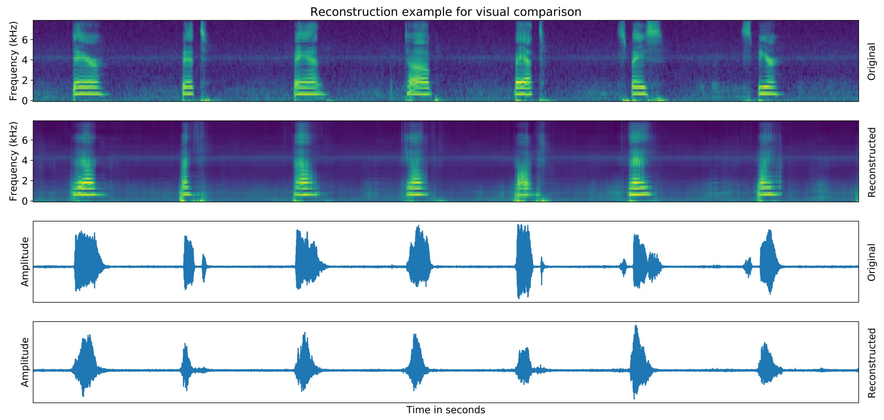


- Feature extraction of neural patterns
  → Exploit grid alignment of electrodes
  → Temporal dynamics during continuous speech
- 3D convolutional operation to extract spatio-temporal features
- CNN network architecture
  → Densely connected (DenseNet)
- Extracted feature maps passed to all subsequent layers
- Consider all found patterns in regressor
- Regression output on 40 LogMels
  → Fully-connected linear layer

Results

- Network training in 5-fold cross validation to reconstruct complete spectrogram
- Wavenet vocoder for generation of audible waveform from spectral coefficients
- **High quality audio from ECoG signals using convolutional neural networks**





Universität Bremen