An Initial Investigation into the Real-Time Conversion of Facial Surface EMG Signals to Audible Speech

Lorenz Diener, Christian Herff, Matthias Janke, Tanja Schultz

lorenz.diener@uni-bremen.de

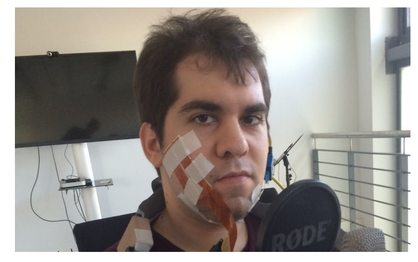
Cognitive Systems Lab

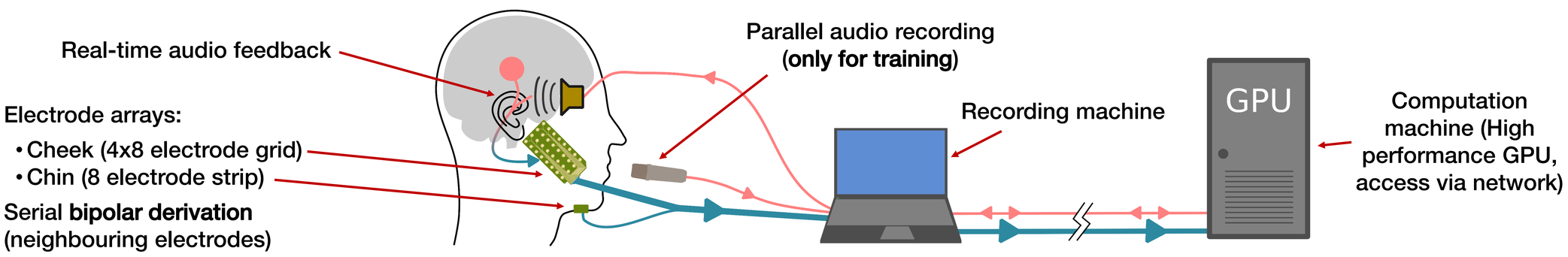
http://csl.uni-bremen.de/cms/de/staff/m-sc-lorenz-diener.html

Motivation

- **Silent Speech Interfaces** (SSIs): Speech interfaces that can be operated with **no audible speech signal**
  → Applications in silent communication, loud environments, speech prostheses
- SSIs based on **electromyography** (EMG) work only with muscle movement
  → **Fully silent operation**

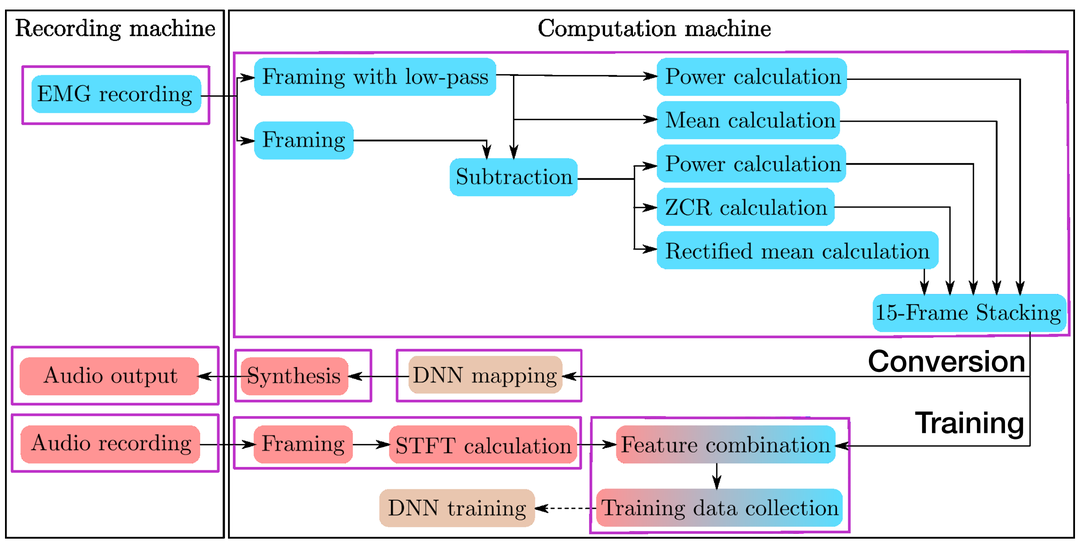
- Previous EMG SSIs: Offline conversion, not real-time low latency capable
  → Real applications and research into feedback / coadaptation effects require this
  → **Need different system structure and algorithms**



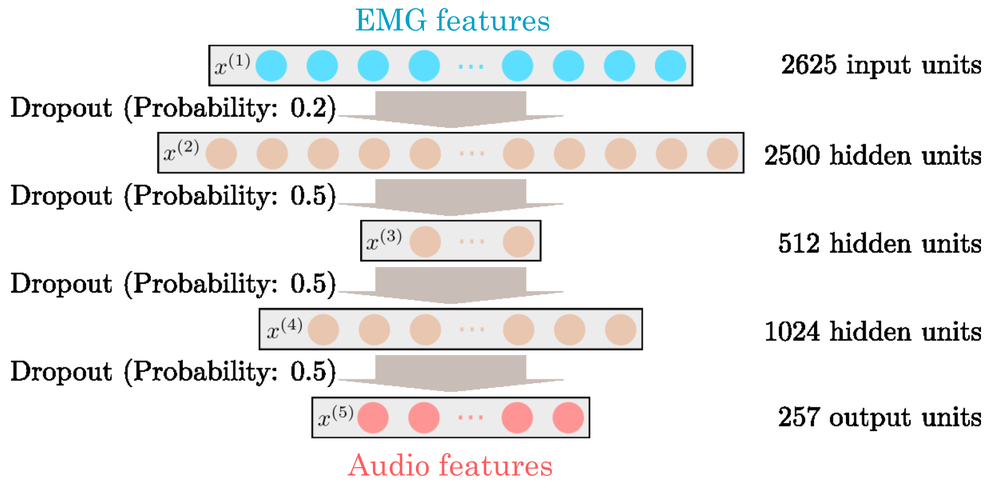


System Structure

- Broken down into **modules** that can run **in their own processes** to achieve real-time performance

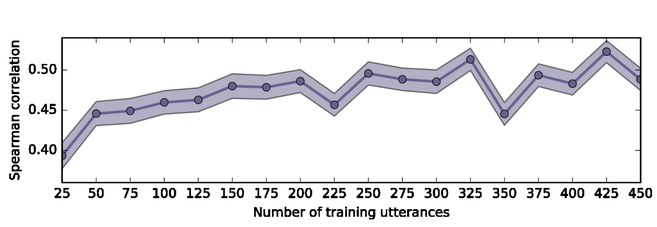


- Actual mapping: Uses **deep neural networks**
  → Proven to work well in EMG-to-Speech conversion
  → Training reasonably fast on modern GPU hardware
  → ReL activation

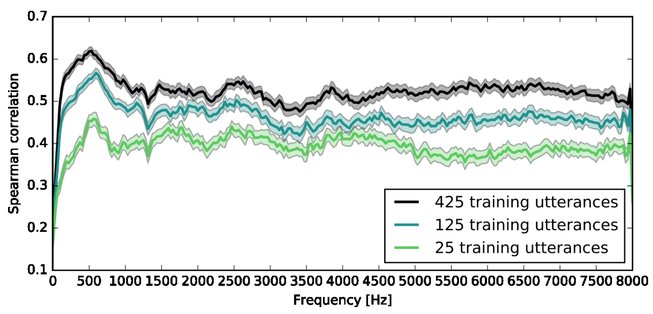


Evaluation

- **Session-dependent systems:** Less training data required means less recording / training time
  → Evaluate training set size on pre-recorded data
  → 125 to 150 utterances (~10 minutes) sufficient, diminishing returns afterwards

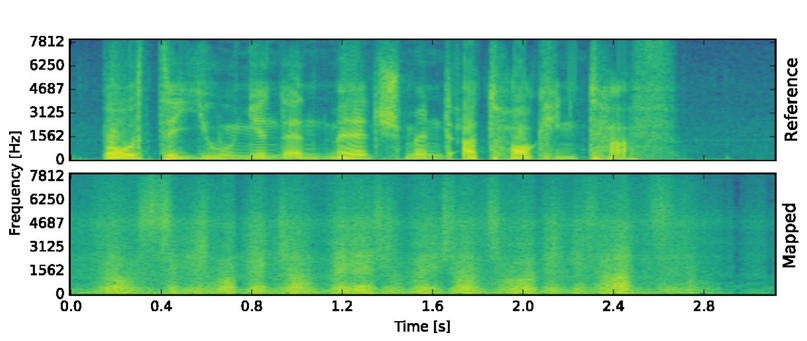


Nb. training utterances versus **test set spectral correlation.** Coloured area: standard error.



Spectrogram bandwise correlations on test set for three systems with different training set sizes. Coloured area: standard error.

- **Latency:** Four distinct categories of latency
  → **Network latency:** Measured, sub-ms
  → **Buffer latency:** Size of largest buffer, here: 40ms
  → **Computation latency:** Measured, per 10ms frame: 9.34ms (mean)
- In practice: higher latencies due to **hardware latency** (EMG amplifier, sound card)
- **Next steps:** Try to **adapt existing systems** to get closer to ideal quality, first **feedback experiments**



Spectrogram comparison: Reference (above) vs synthesized (with high-quality / ideal reference system trained using 450 utterances, below)

Universität Bremen