Adaptation of an EMG-Based Speech Recognizer via Meta-Learning

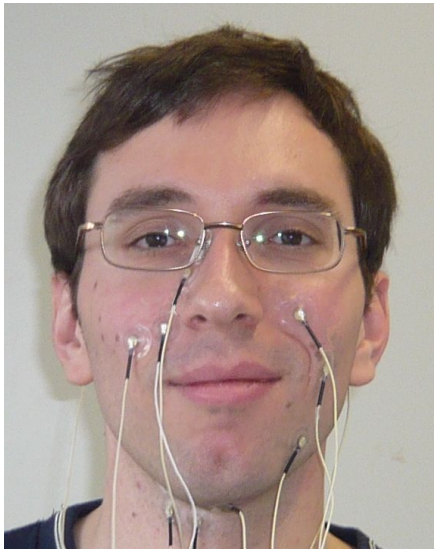
Krsto Proroković, Michael Wand, Tanja Schultz, and Jürgen Schmidhuber

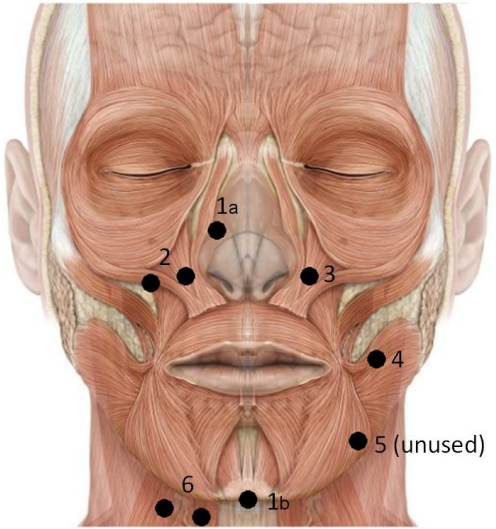
The Swiss AI Lab IDSIA

krsto@idsia.ch

Introduction

- Speech recognition from EMG signal of facial muscles
- Small muscle-generated electrical currents are recorded by electrodes





Sampling and Preprocessing

- Sampling
  - 6 EMG channels, 5 of which are used
  - 600 Hz sampling rate
- Preprocessing
  - Window length 27 ms, window shift 10 ms
  - From each of 5 channels 5 features are extracted
  - We use a context of 11 frames (±5 frames)
  - Therefore, total dimension of input is 5×5×11 = 275

Methods

- Neural network frontend
  - We use Bundled Phonetic Features (BDPFs) [1]
  - For each of 4 BDPFs we use a separate neural network [2]
    - 3 hidden layers, 600 neurons, tanh nonlinearity, 50% dropout
- Decoding performed using HMM-style time-synchronous beam search
  - We assume that phone level alignments are present
  - Frame level scores of the BDPF trees are averaged and combined with dictionary (108 words) and statistical language model
- Hybrid system - ensemble of neural networks combined with beam search

[1] Tanja Schultz and Michael Wand. Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition. Speech Communication.

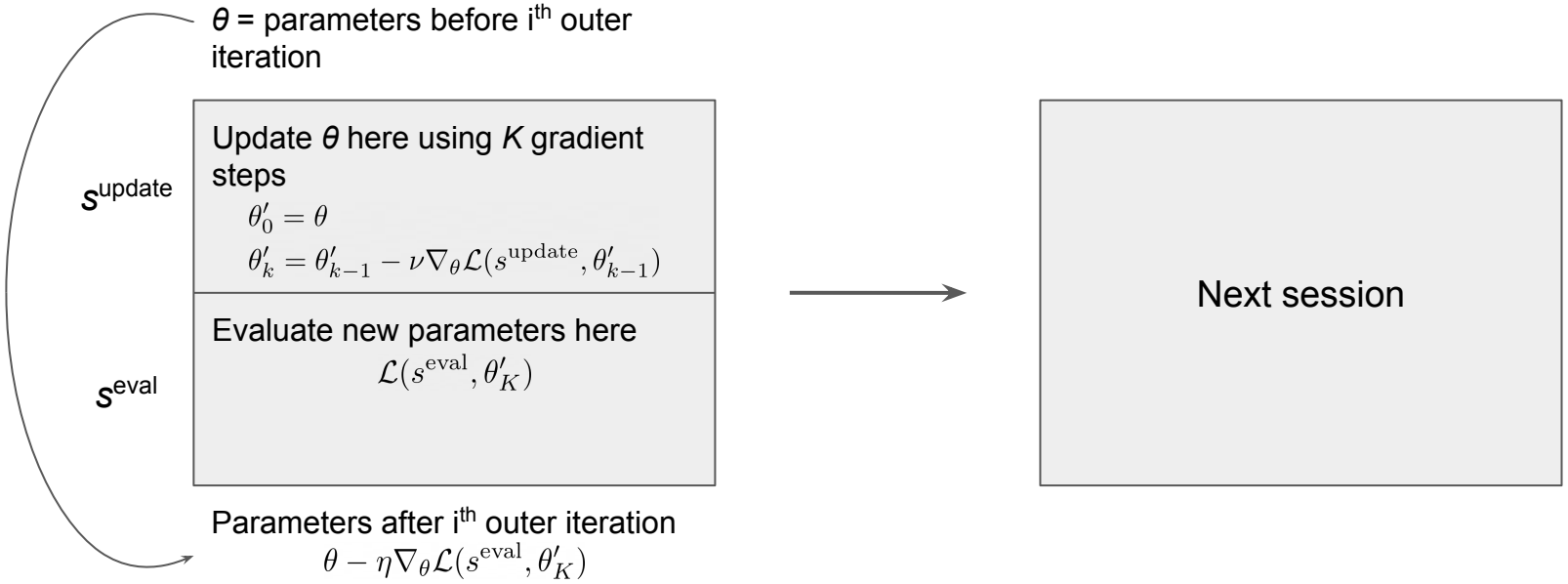
[2] Michael Wand and Jürgen Schmidhuber. Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition. Interspeech 2016.

Problem of Adaptation to Incoming Session

- Speaker has to remove the electrodes and put them again
  - Session = recording between putting and removing the electrodes
- Differences between sessions
  - Electrode displacement, sweating, environmental artefacts, etc.
  - Neural networks trained on existing sessions won't perform equally well on the incoming one
- There needs to be some adaptation
  - Of course, using little data
- Conventional pretraining and fine-tuning
  - Take neural networks trained on existing sessions and apply gradient descent on them with data from incoming session
- Meta-learning
  - This work!

Model-Agnostic Meta-Learning (MAML)

- Idea: optimize the model to be optimized with *K* gradient steps [3]



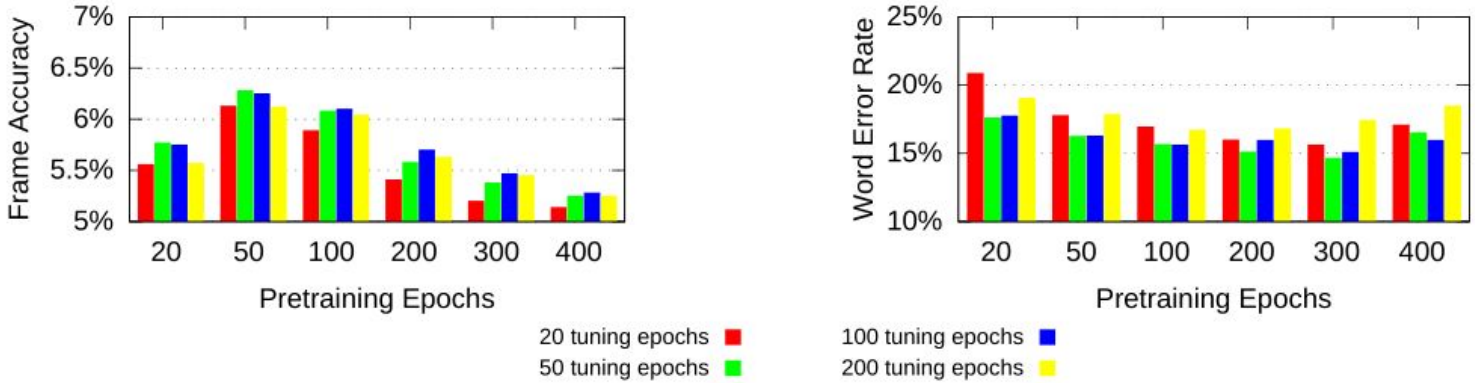
[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.

Data

- EMG-UKA Corpus
  - 7:32 hours of English language EMG and acoustic speech recordings
  - We use only audible mode
- We use 48 out of 61 sessions
  - Sessions 1 - 32 by speaker 2 = development set
    - Two blocks of 16 sessions
  - Sessions 1 - 16 by speaker 8 = evaluation set
- Each session consists of 50 sentences
  - 40 are used for pretraining or adaptation
  - 10 are used for testing
- 15 sessions are used for pretraining, 1 for adaptation and testing

Experiments: Baseline

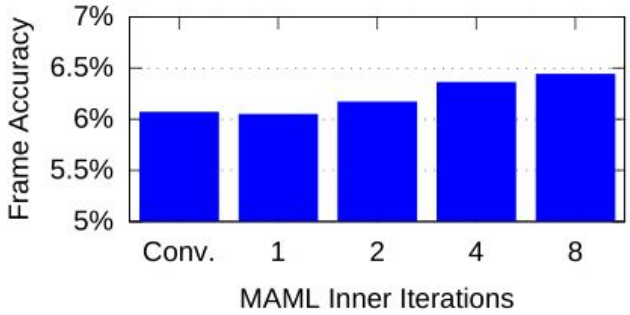
- Baseline: conventional pretraining with fine-tuning
  - We use ADAM optimizer for pretraining and SGD for fine-tuning
  - Higher frame accuracy does not necessarily correspond to lower word error rate [4]
  - Max. accuracy = 6.28% with 50 epochs of pretraining and 50 epochs of fine-tuning
  - Min. WER = 14.7% with 300 epochs of pretraining and 50 epochs of fine-tuning

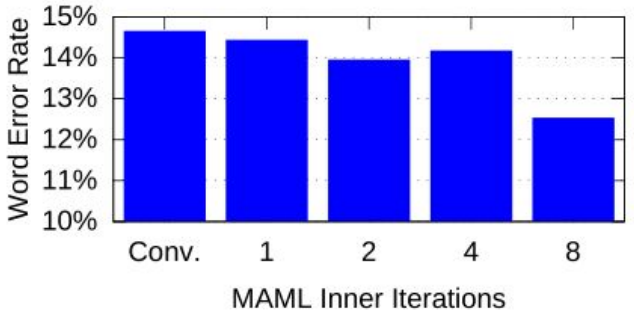


[4] Michael L. Shire. Relating Frame Accuracy with Word Error in Hybrid ANN-HMM ASR. Eurospeech 2001.

Experiments: Meta-Learning

- Meta-Learning: MAML with fine-tuning
  - We use the same learning rate for inner loop and fine-tuning
  - Pretraining with 200 meta-epochs and fine-tuning with 50 epochs
  - Adding more inner iterations (i.e. gradient steps *K*) improves performance
  - Accuracy 5.38% → 6.32% (17.5% improvement); WER 14.7% → 12.5% (15% improvement)





Evaluation

- Meta-learning with fine-tuning achieves almost the same accuracy as conventional pretraining and fine-tuning with twice as much data

Number of Sequences	Method	Frame Accuracy	Word Error Rate
10	Conventional	6.00%	13.32%
	Meta-Learning	6.57%*	12.75%
20	Conventional	6.44%	8.96%
	Meta-Learning	7.26%*	8.08%
40	Conventional	7.20%	6.19%
	Meta-Learning	8.17%*	4.92%*

* significant result

Conclusion

- We saw how meta-learning can be used for adaptation of an EMG-based speech recognizer substantially outperforming conventional pretraining with fine-tuning
- In future we might try adapting the system for different speakers rather than for different sessions of the same speaker
- We hope this work inspires more applications of recent machine learning advances in biosignal processing