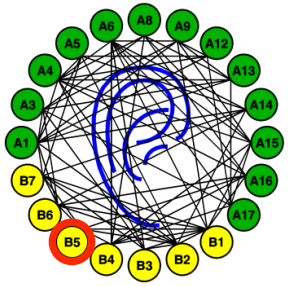


Assessment of human speech intelligibility based on machine listening

Bernd T Meyer

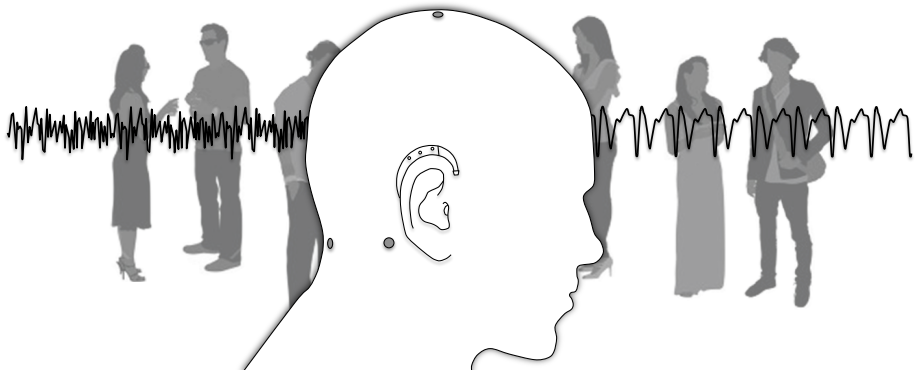
Automatic speech and audio processing (ASAP)

Medical Physics

Carl von Ossietzky Universität Oldenburg

Jan 8th 2015, SpiN Workshop

Automatic speech and audio processing

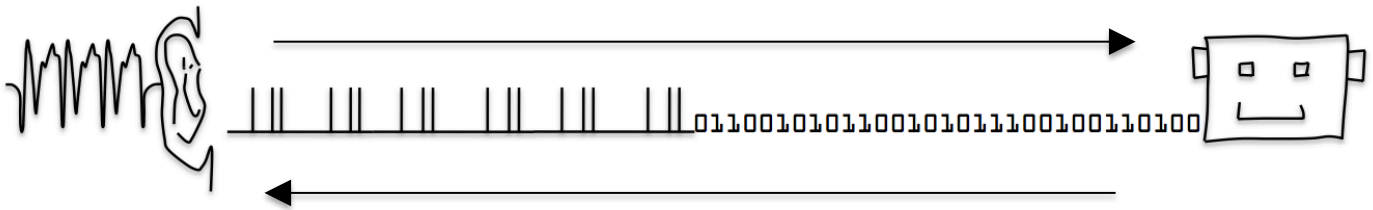


Aims of our research:

- To improve speech processing systems (e.g., for assistive devices)
- To better understand (and/or model) our auditory system

Research is based on

- relation between the auditory system and machine listening

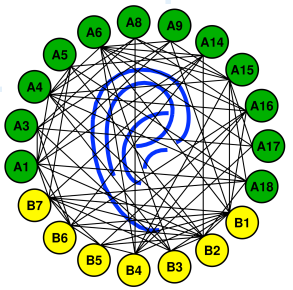


Overview

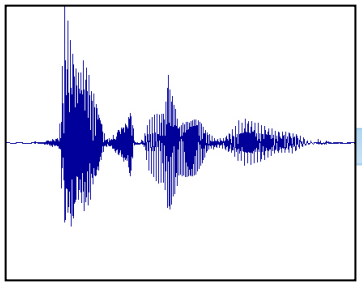
1 Improving automatic speech processing based on "auditory inspiration"

2 Models of speech intelligibility

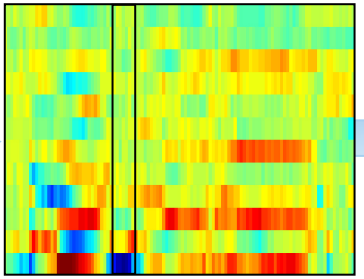
3 Models of speech perception and cortical correlates



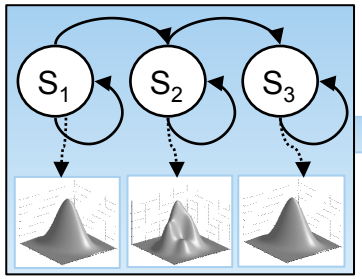
Automatic speech recognition (ASR)



Speech signal



Feature extraction



Classification

„How to wreck a nice beach“

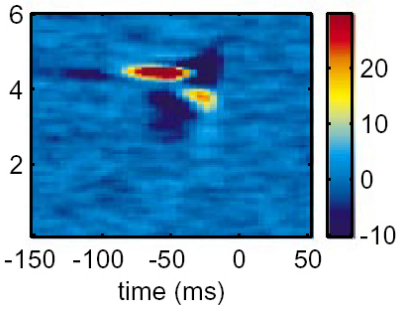
Transcript

- Feature: Carries information relevant for recognition
- Should be invariant to noise, reverberation (but often isn‘t)
- Standard features: Mel-frequency cepstral coefficients (MFCCs)

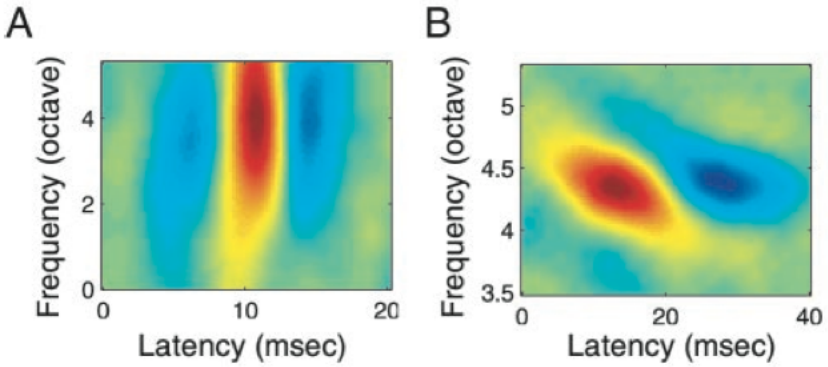
- Classification: Which word / phoneme was produced?
- Training: Models for words and sub-word units
- Standard recognizers: Hidden Markov Modelle (HMMs), neural networks

New features: Models of spectro-temporal receptive fields

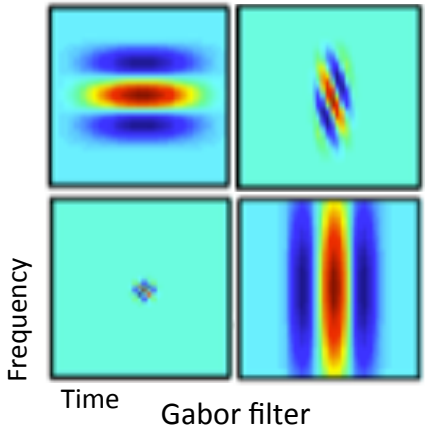
- Neurons in the primary auditory cortex of mammals are sensitive to specific spectro-temporal stimuli
- Spectro-temporal Gabor filters serve as model for spectro-temporal receptive fields (STRFs)*



deCharms et al. (1998)



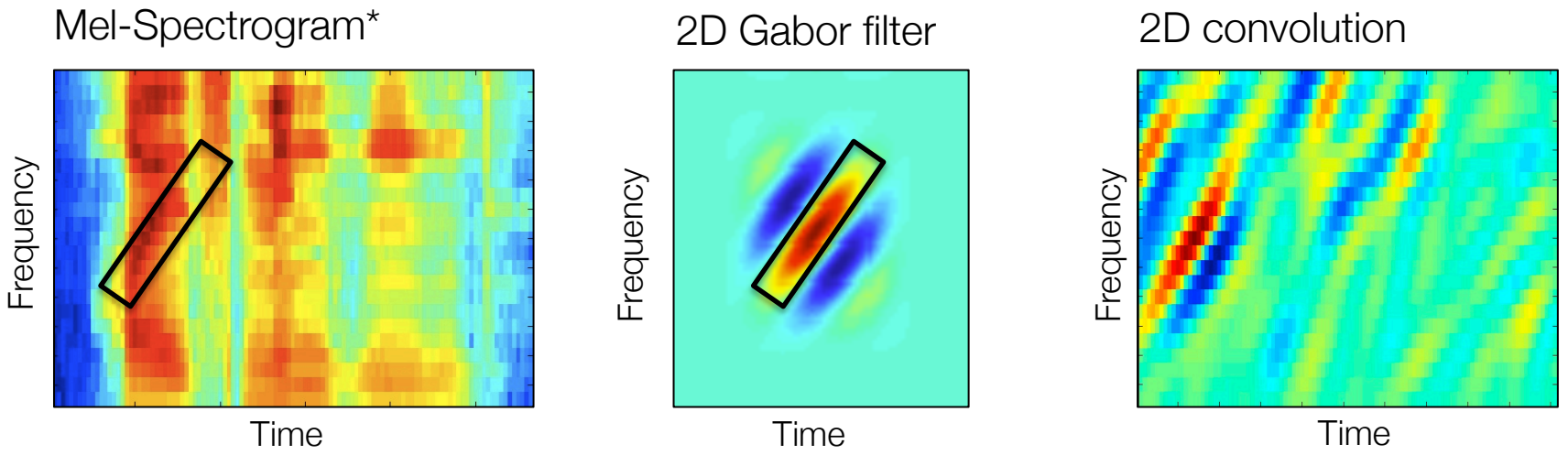
Qui, Schreiner, Escabi (2003)



Qiu, A., Schreiner, C. & Escabi, M., 2003. Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of Neurophysiology*, 90 (1), pp.456–476.

deCharms, C., Blake, D. Merzenich, M.M., 1998. Optimizing sound features for cortical neurons. *Science*, 280 (5368), pp.1439–1444.

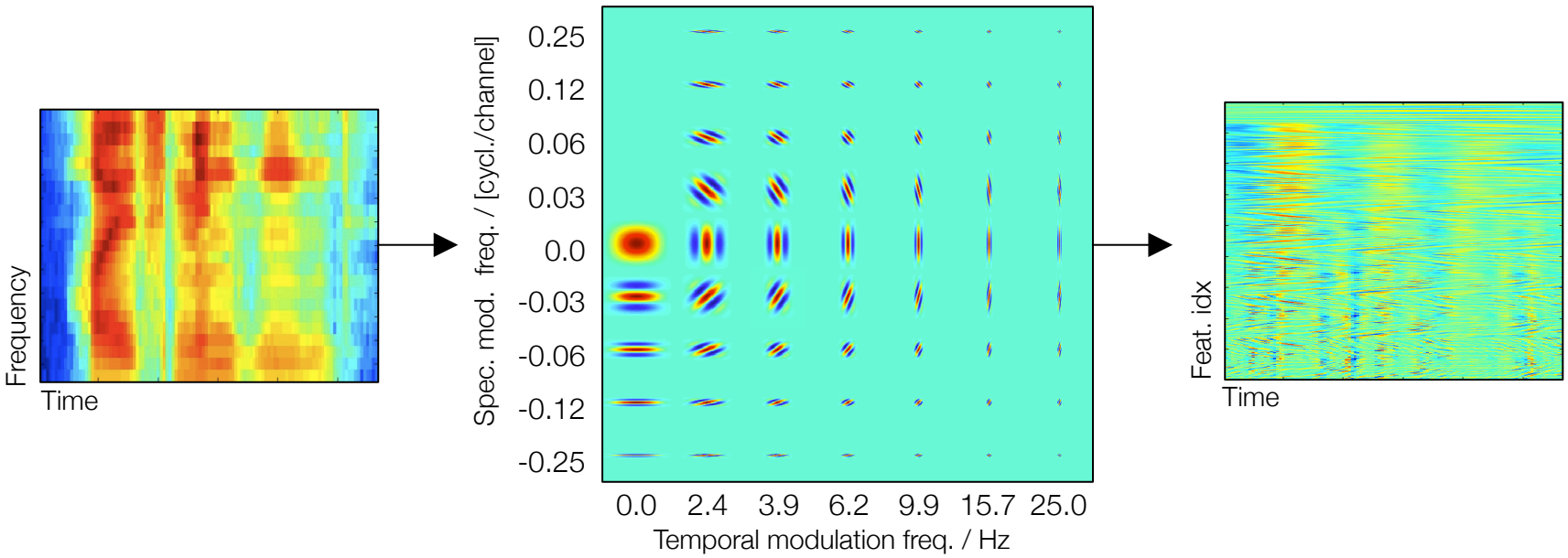
Pattern-matching with spectro-temporal Gabor filters



Meyer, Kollmeier (2011). "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition", Speech Communication 53 (5).

Spectro-temporal features

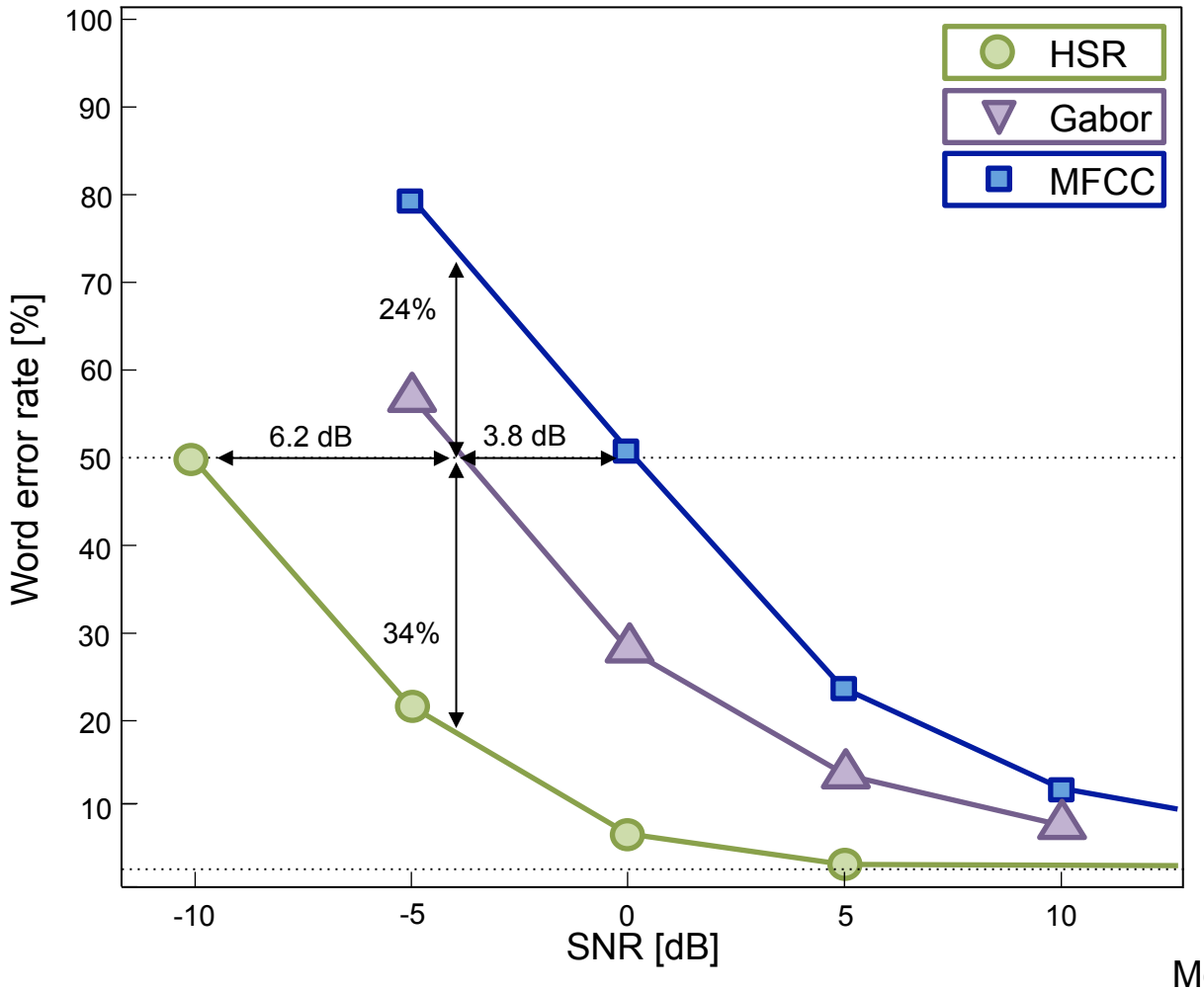
Filter bank of Gabor features: Evenly cover physiologically relevant modulation frequencies



Schädler, Meyer, Kollmeier, (2011). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition“, JASA 131.

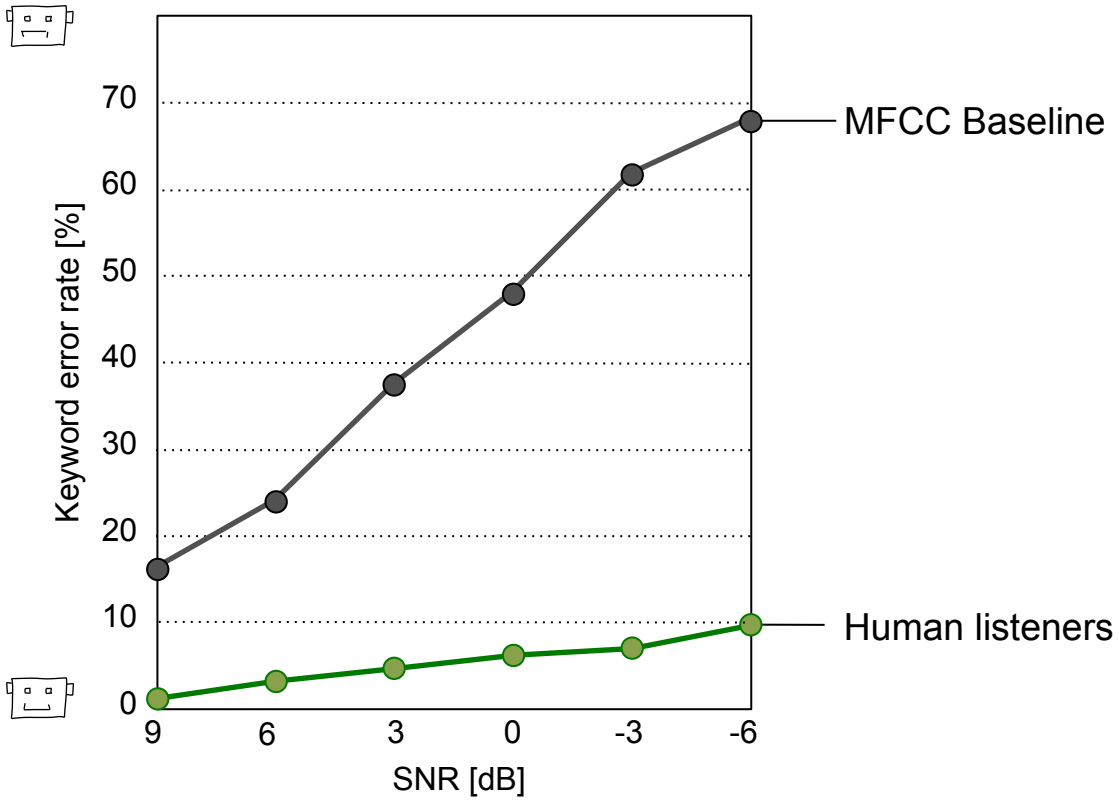
Meyer, Spille, Kollmeier, Morgan (2012). “Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition“, in Proc. Interspeech.

Humans vs. machine listening: Simple tasks: Noisy digits



Meyer (INTERSPEECH 2013)

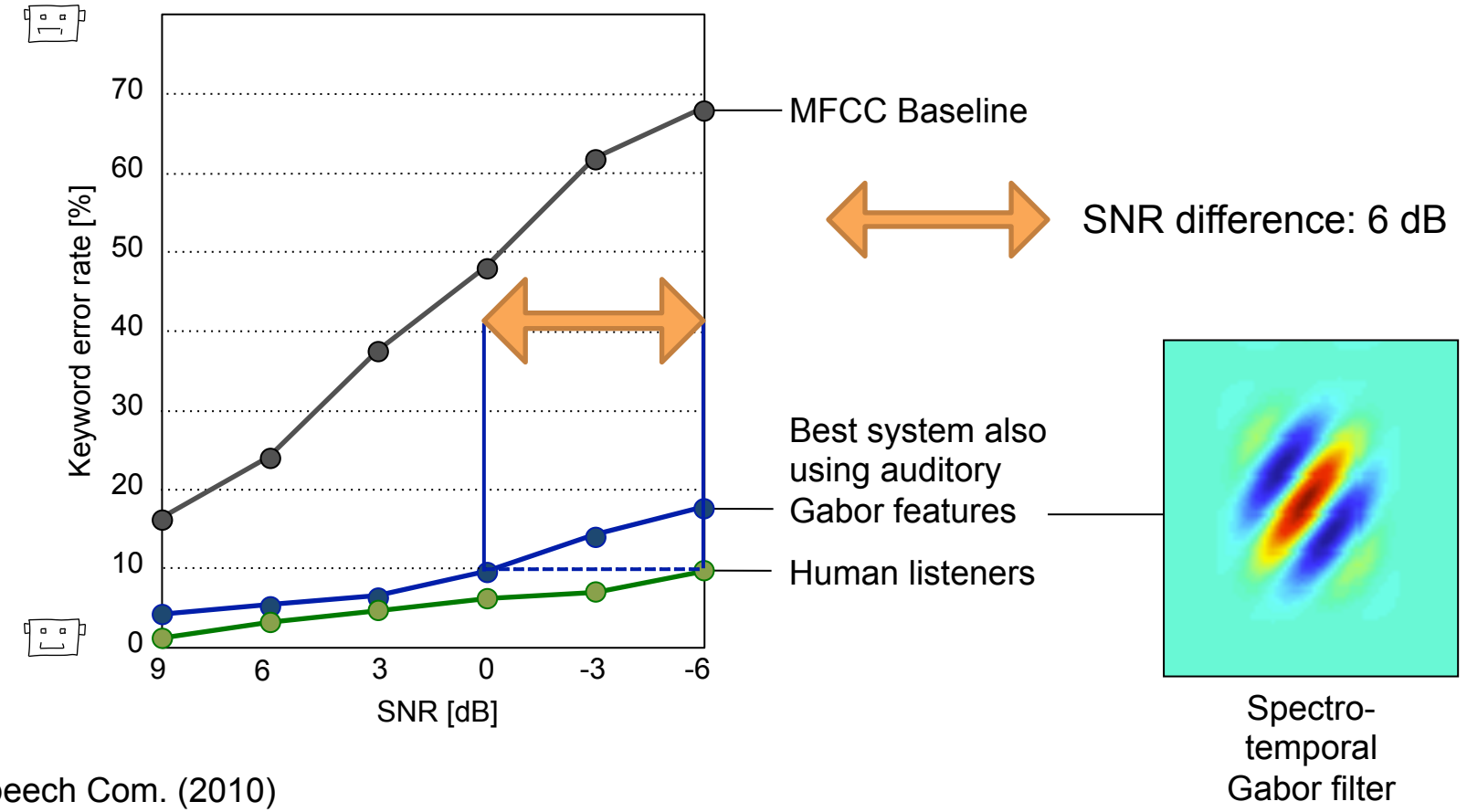
Humans vs. machine listening: Realistic, moderately difficult



Spectro-temporal Gabor filter

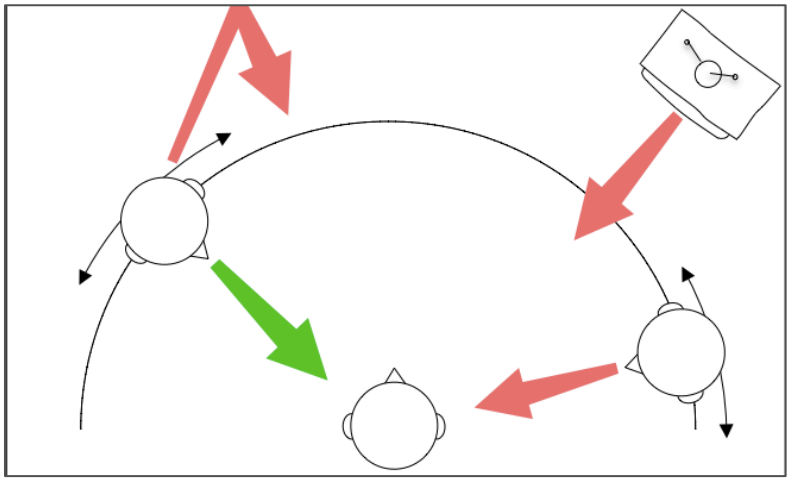
Meyer et al., Speech Com. (2010)
Schädler et al., JASA (2012)
Moritz et al., CHiME 2013 Workshop (2013)

Humans vs. machine listening: Realistic, moderately difficult

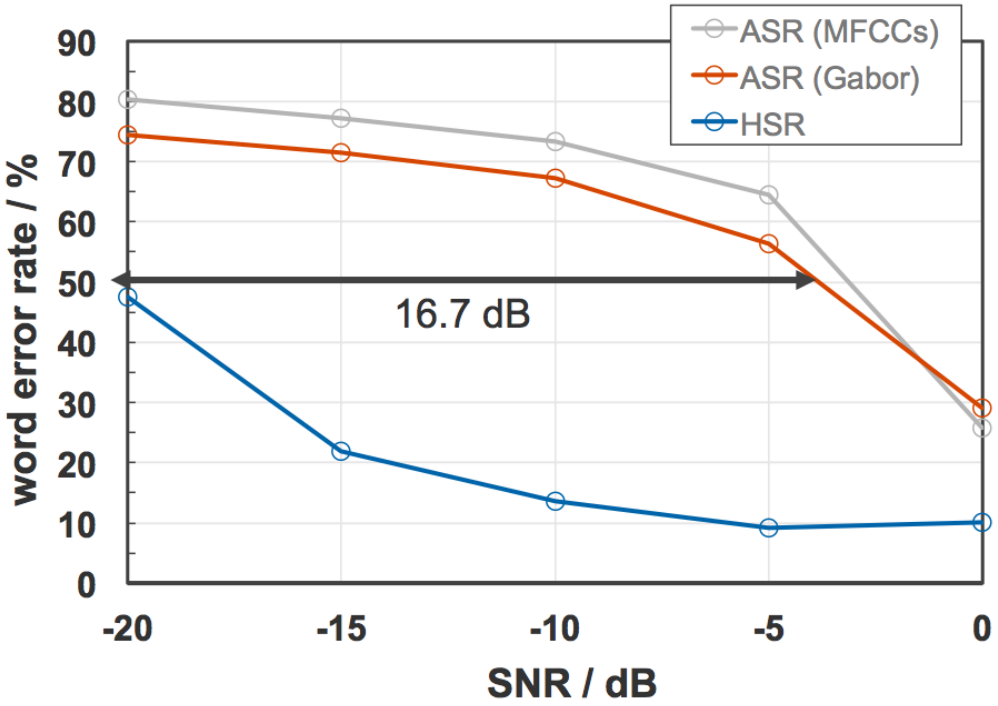


Meyer et al., Speech Com. (2010)
Schädler et al., JASA (2012)
Moritz et al., CHiME 2013 Workshop (2013)

Humans vs. machine listening: Realistic, very difficult



Fast moving speakers at 0 dB + additive noise + reverberation



Spille and Meyer (INTERSPEECH 2014)

Summary

1 The auditory approach to speech processing often improves robustness of ASR systems

...but there is still quite a gap between HSR and ASR. Can ASR still be useful for models of human speech perception?

Overview

1 Improving automatic speech processing based on „auditory inspiration“

2 Models of speech intelligibility

3 Models of speech perception and cortical correlates

Modelling human speech perception

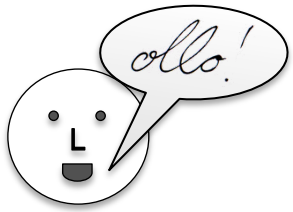
A good model of human speech perception…

- …could be used to evaluate speech compression algorithms (e.g., “Does the proposed algorithm decrease speech intelligibility”)
- …could predict performance of new hearing aid algorithms
- …without the need of performing (expensive) listening experiments.

Modelling human speech perception

Test utterances

- Oldenburg logatome corpus (3,600 non-words)
- Added stationary, speech-shaped noise

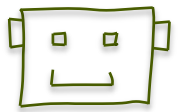


Human speech recognition

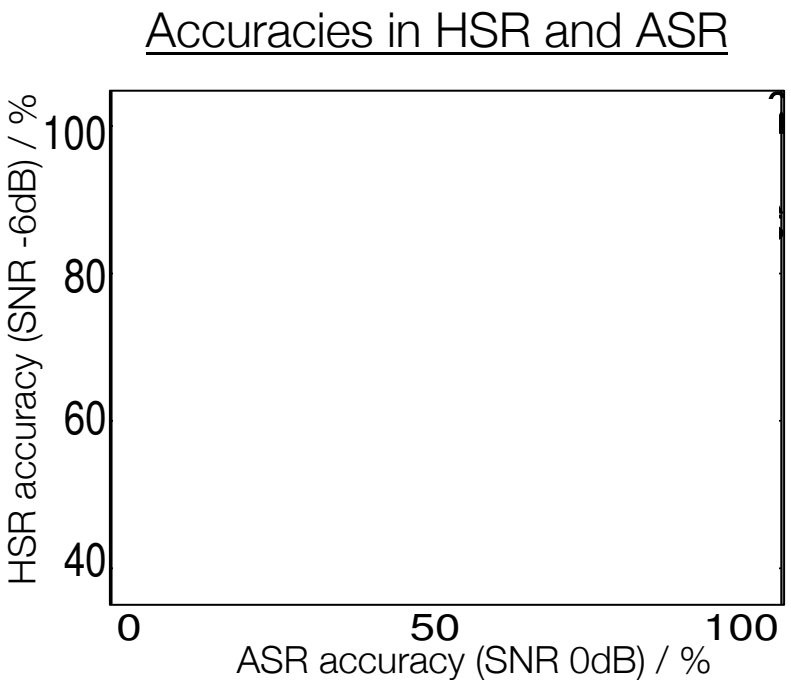
- 6 normal-hearing listeners
- Signal-to-noise ratio: -6 dB
- ~21k responses

Automatic speech recognition

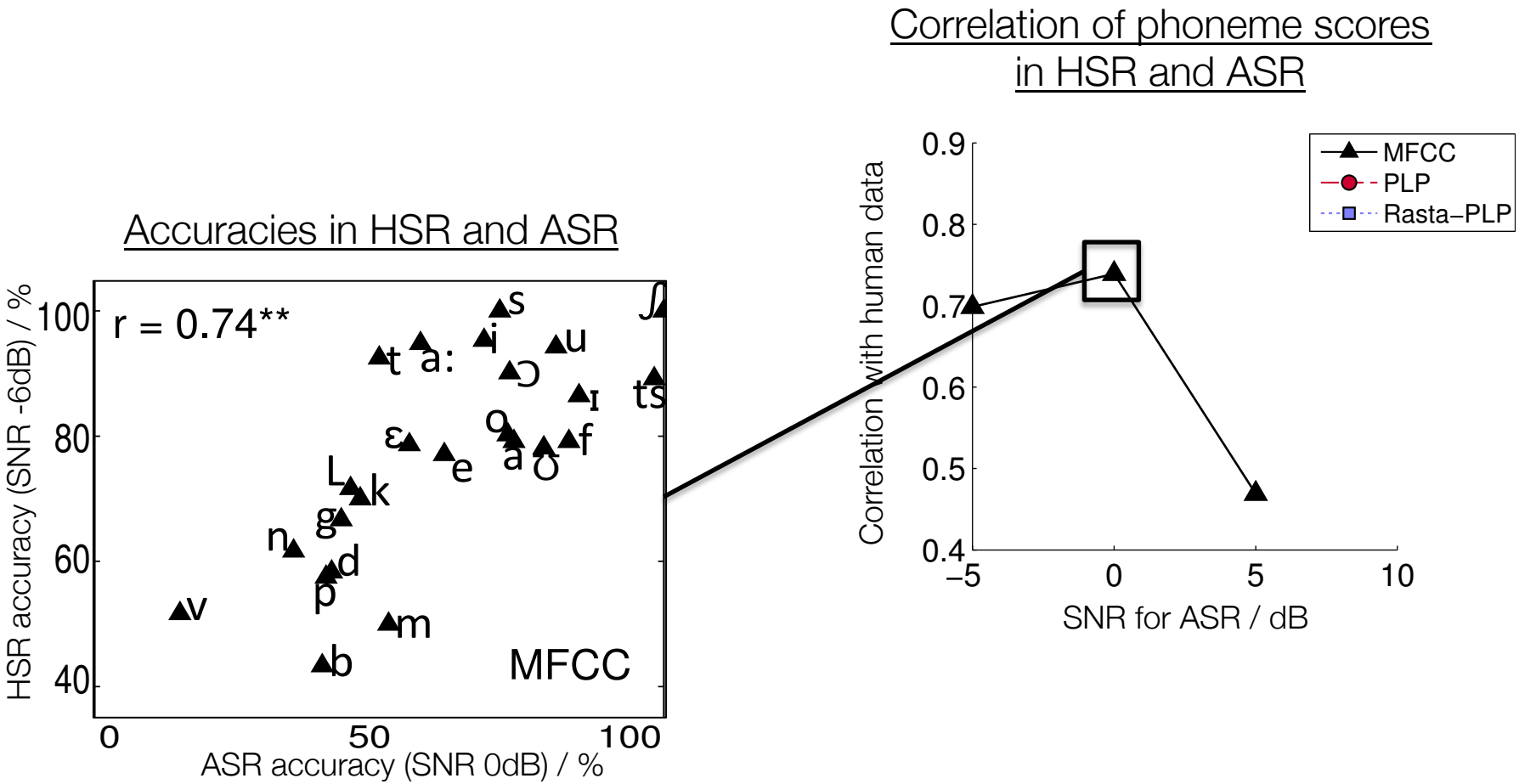
- Three different feature types
- Wide range of SNRs
- Classifier: Hidden Markov Model (HMM)



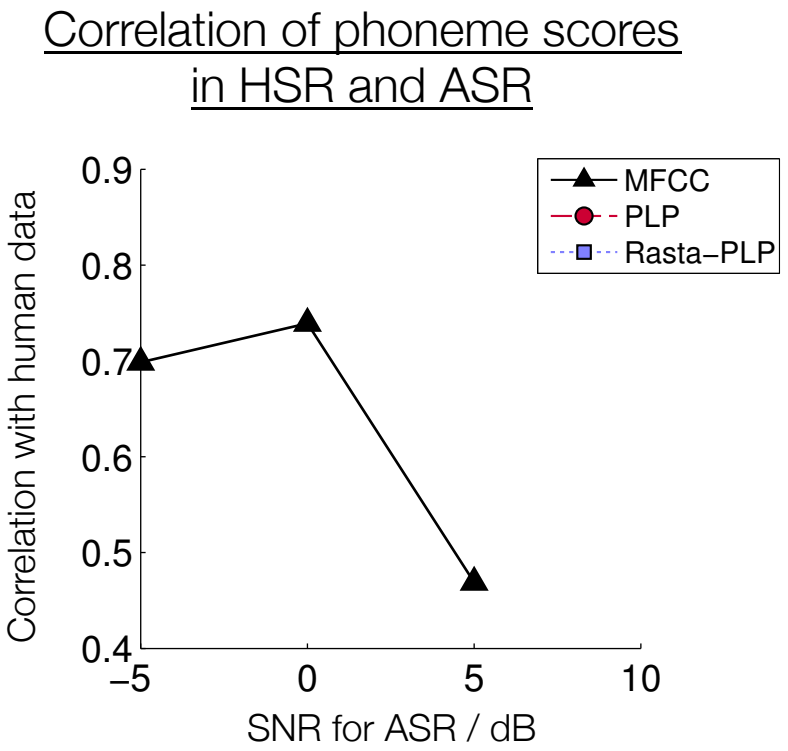
Modelling human speech perception



Modelling human speech perception

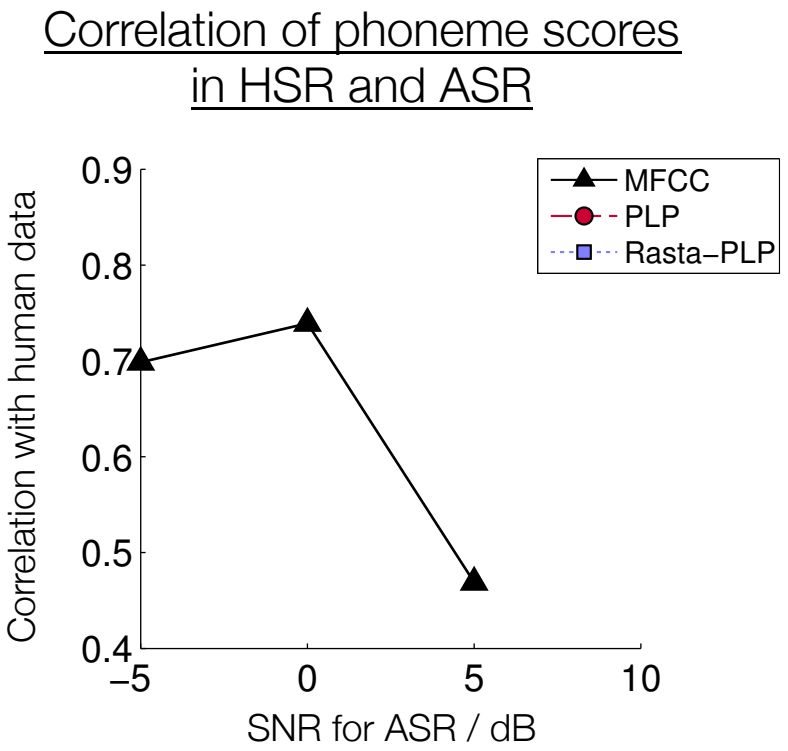


Modelling human speech perception



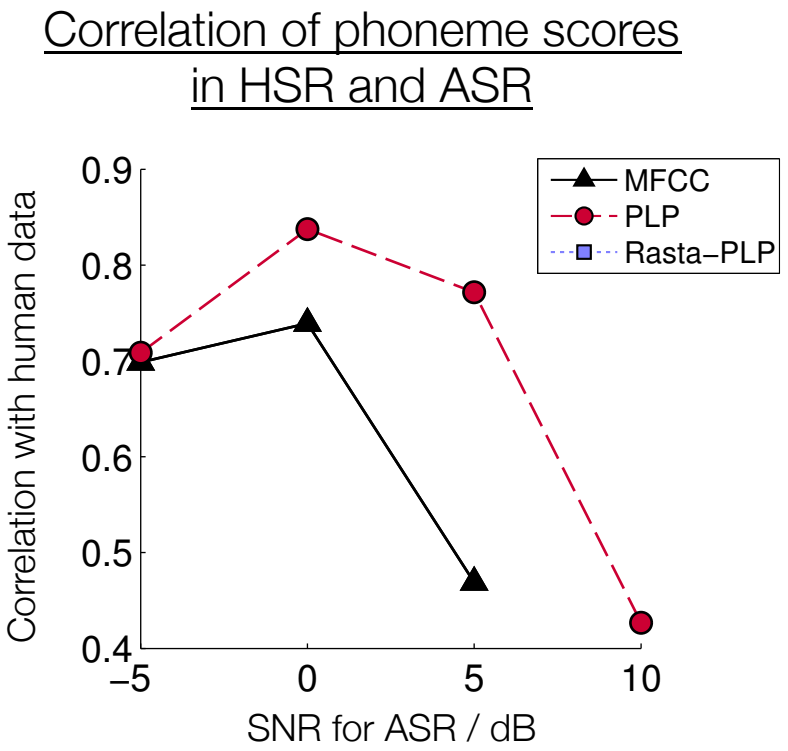
Modelling human speech perception

- Feature types
  - Mel-frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction coefficients (PLP)
  - Rasta-PLP (Relative spectra PLP)



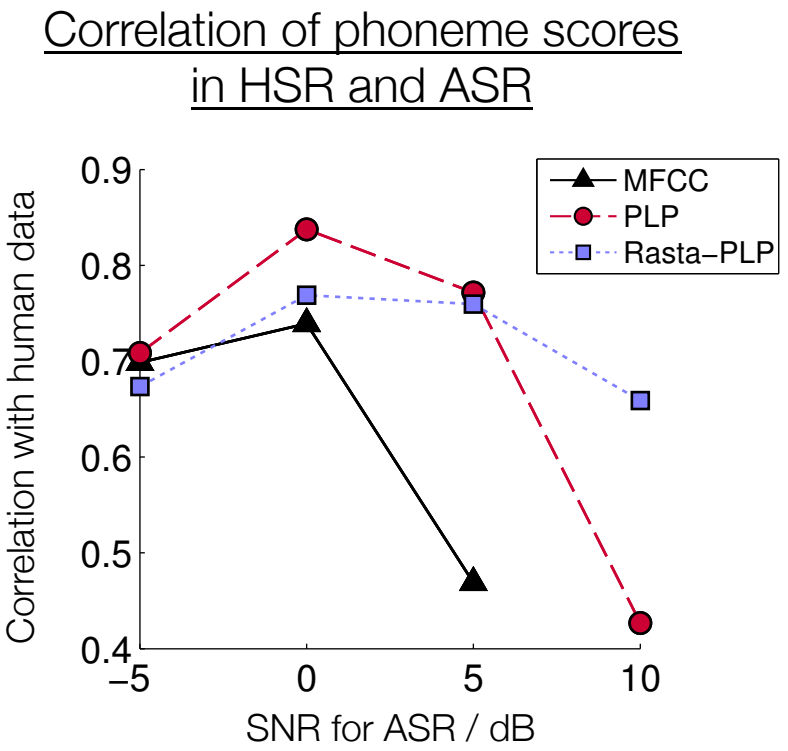
Modelling human speech perception

- Feature types
  - Mel-frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction coefficients (PLP)
  - Rasta-PLP (Relative spectra PLP)



Modelling human speech perception

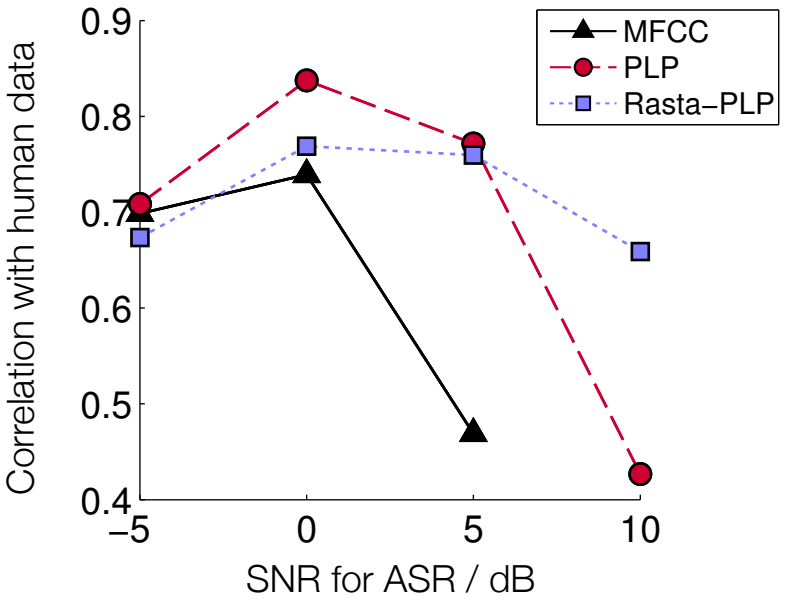
- Feature types
  - Mel-frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction coefficients (PLP)
  - Rasta-PLP (Relative spectra PLP)



Modelling human speech perception

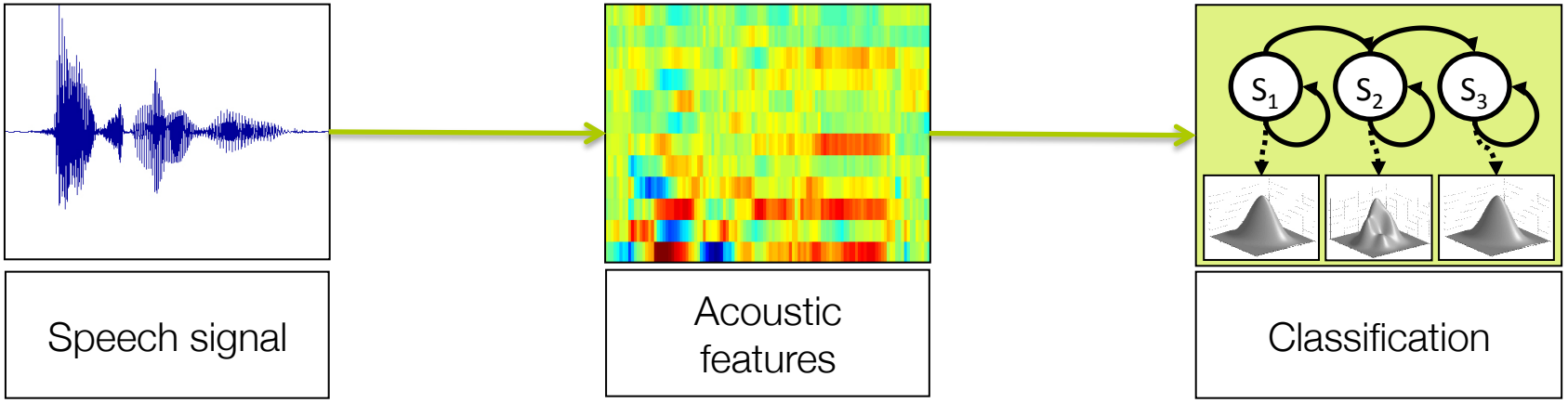
- Feature types
  - Mel-frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction coefficients (PLP)
  - Rasta-PLP (Relative spectra PLP)
- Features with higher “auditory influence”: better predictions
- ASR is in principle suitable to model phoneme confusions

Correlation of phoneme scores in HSR and ASR



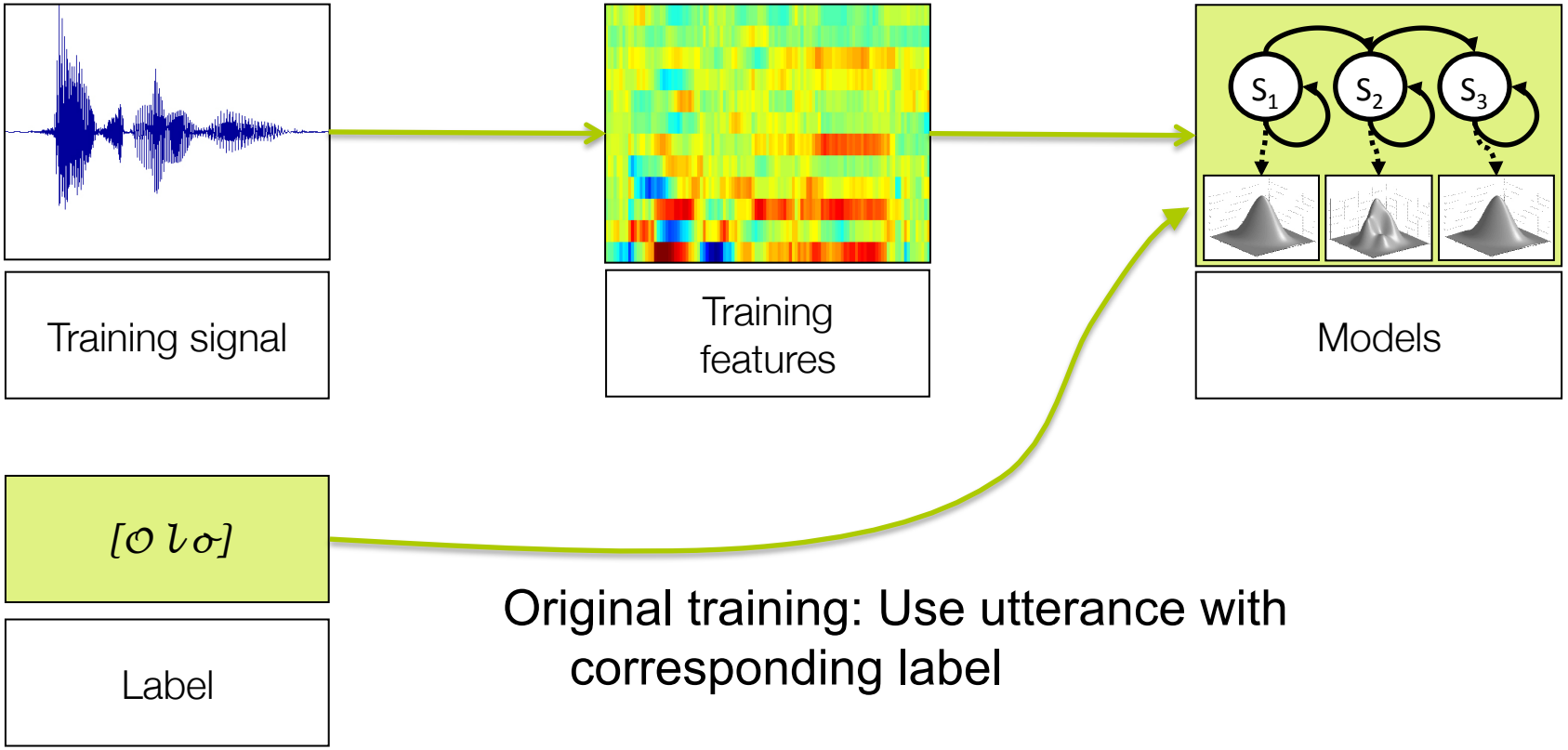
Modelling human speech perception

What can we do to improve the predictions of speech intelligibility?



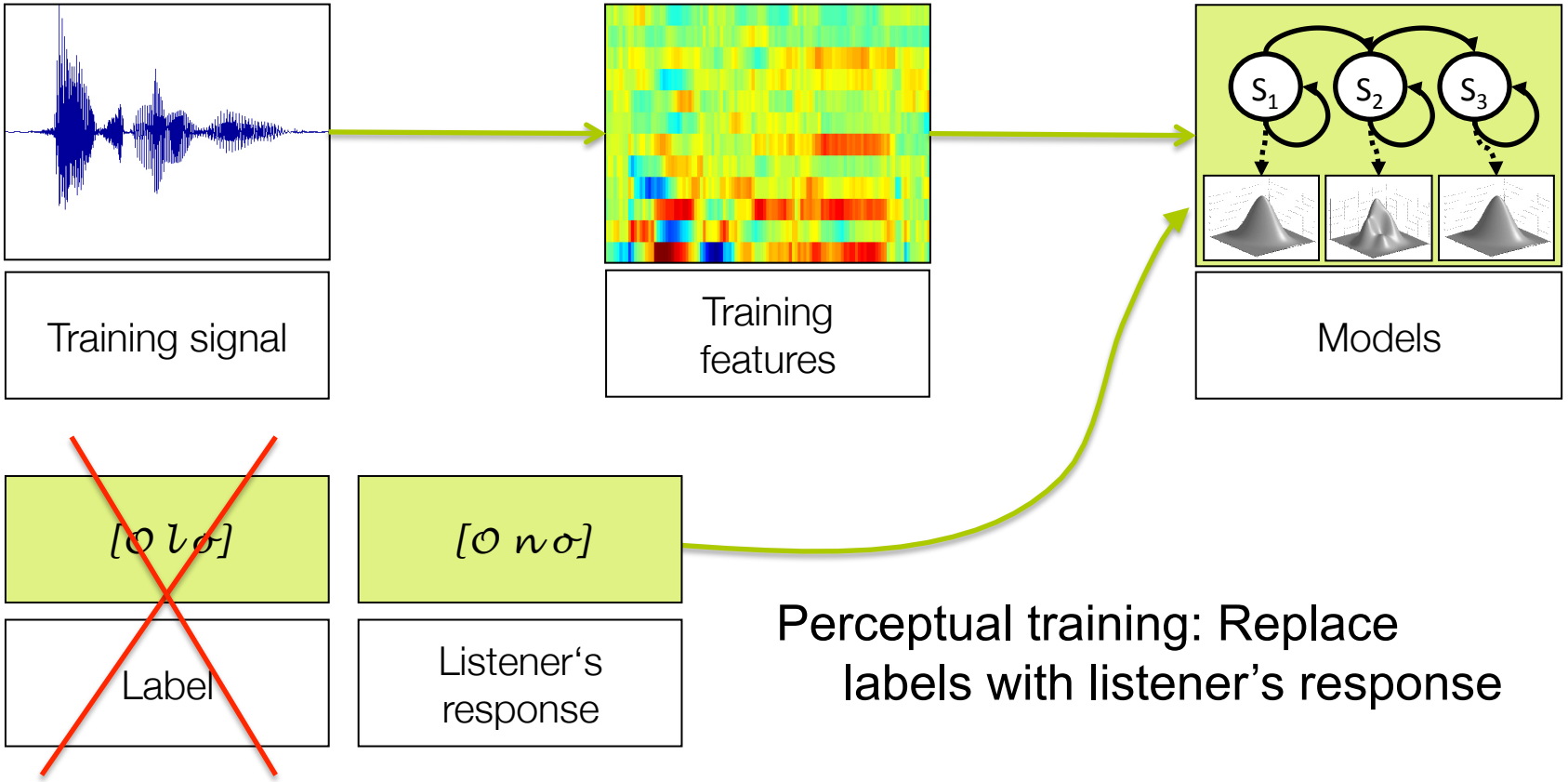
Modelling human speech perception

What can we do to improve the predictions of speech intelligibility?

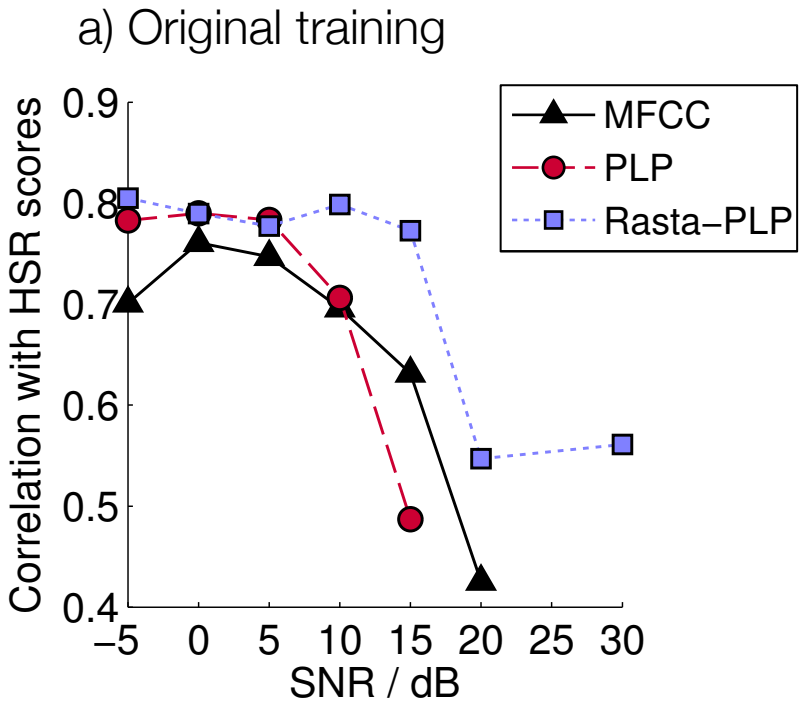


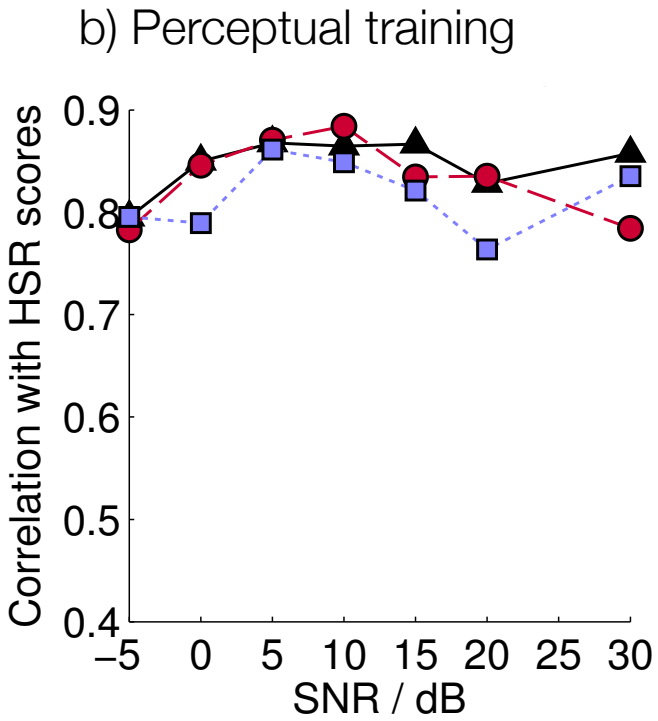
Modelling human speech perception

What can we do to improve the predictions of speech intelligibility?



Modelling human speech perception





Comparison of original and perceptual training:

Best correlation increased from 0.80 to 0.89 ($p < 0.01$)

Summary

2

Models of speech intelligibility can profit
from methods used in speech research.

...but we need lots of listeners responses
to continue with perceptual training for large scale models

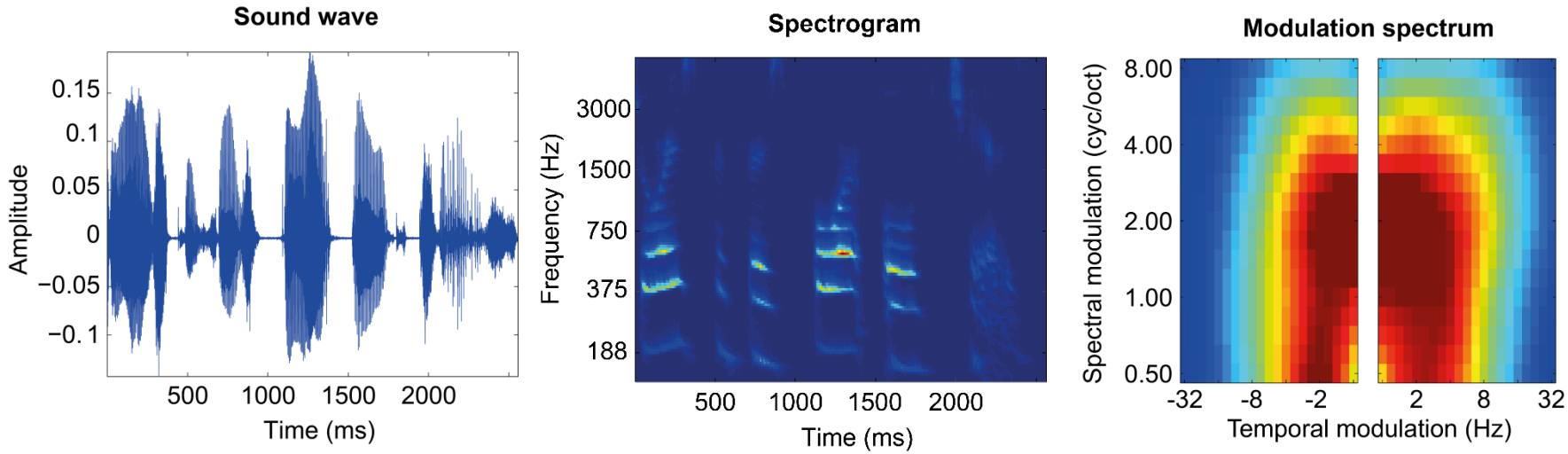
Overview

1 Improving automatic speech processing based on „auditory inspiration“

2 Models of speech intelligibility

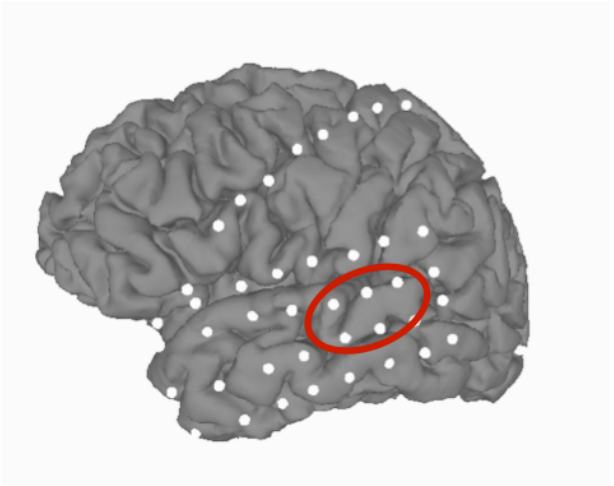
3 Models of speech perception and cortical correlates

Models of speech perception and cortical correlates



- Cooperation with the Neuropsychology Lab (Jochem Rieger)
- What speech features are represented in human cortex?
- How are these speech features represented?

Electrocorticography recordings



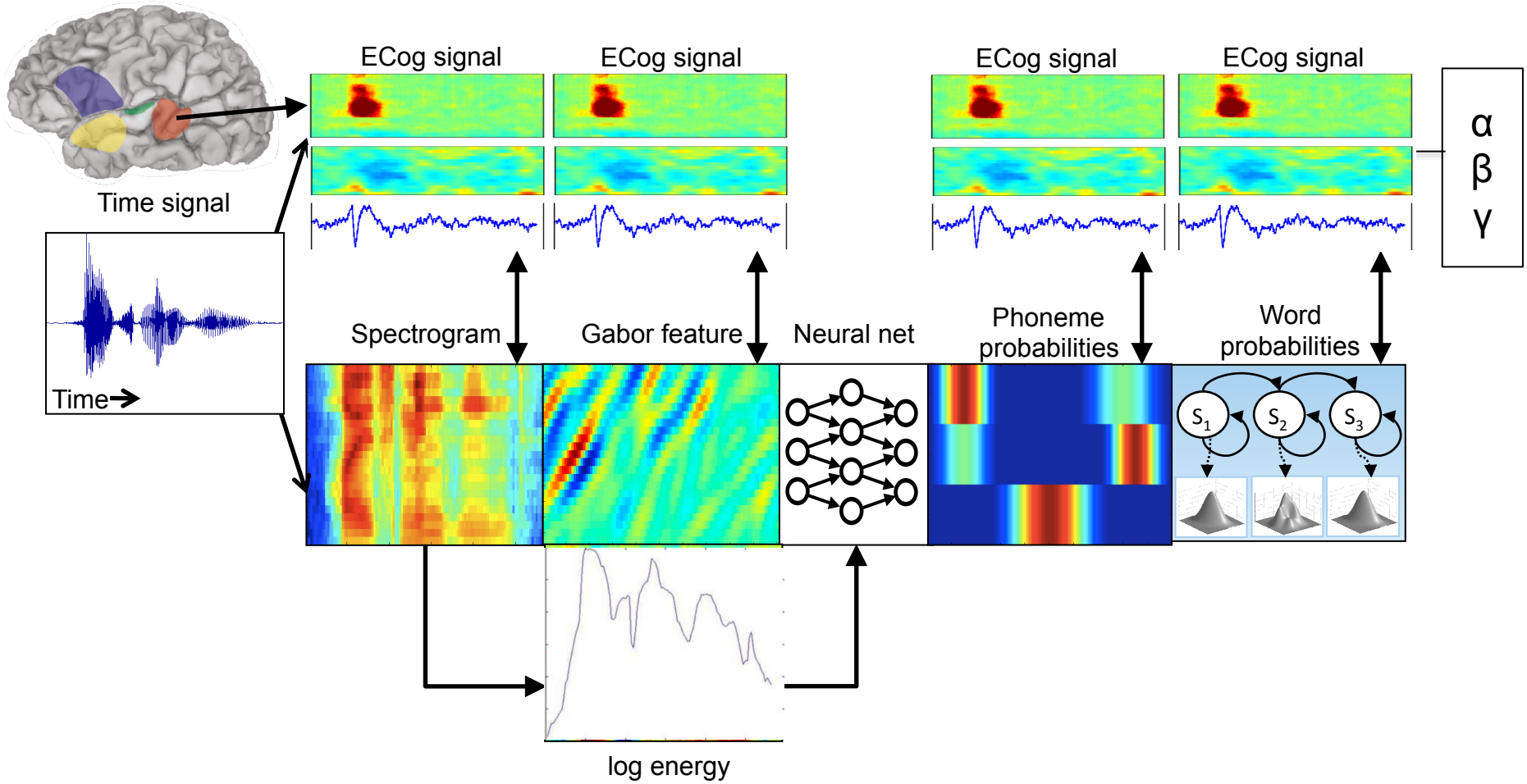
- Direct subdural recordings from patients with intractable epilepsy (we aim for N >= 5)
- Focus on posterior superior temporal gyrus (pSTG) electrodes
- 2 recording sites (Berkeley and Houston)

Speech material and task

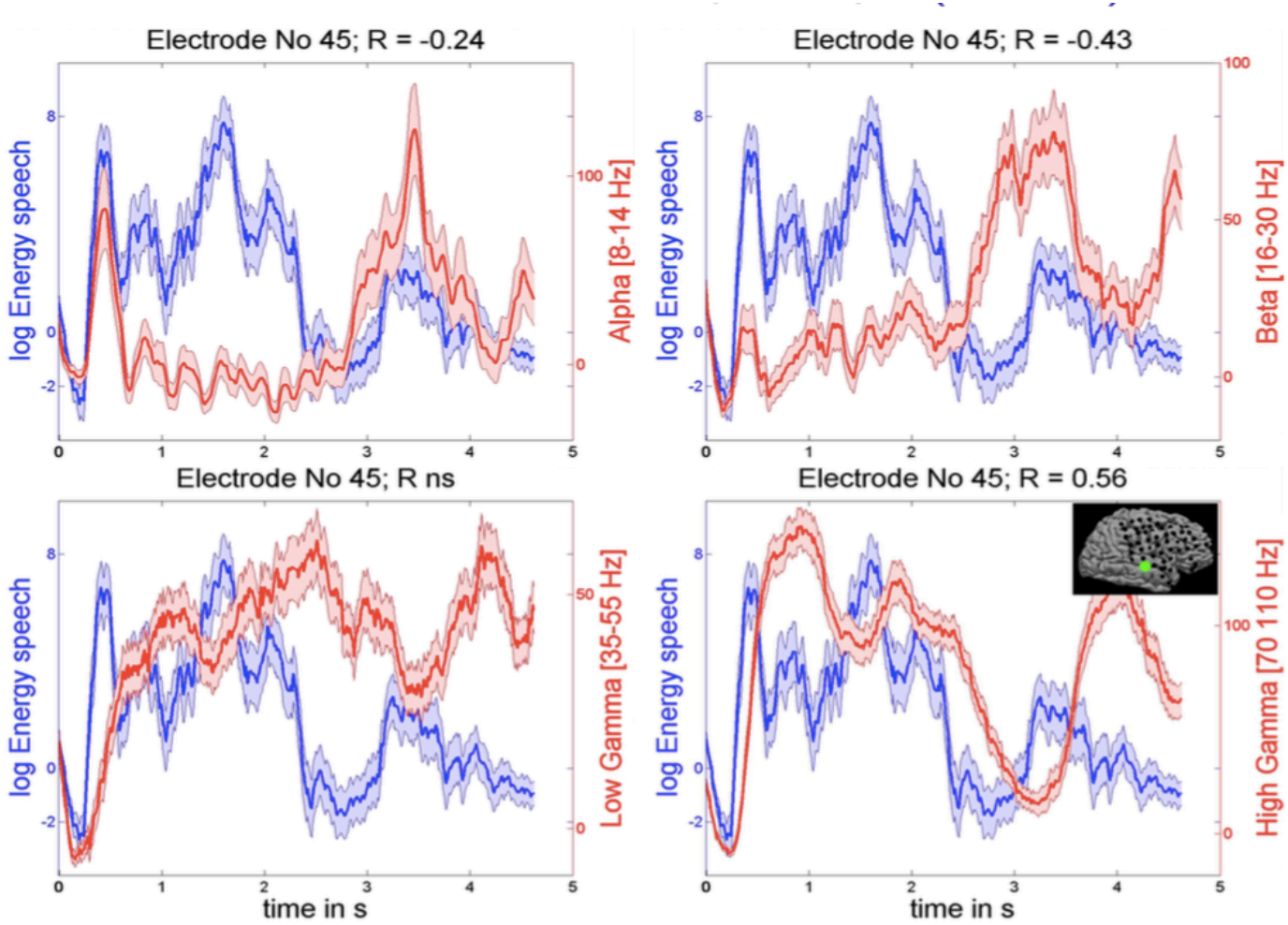
- 210 sentences (105 per experimental condition)
- Re-recording of English matrix test
- Task: Did the target word occur in the sentence?

a	b	c	d	e
Peter	got	three	large	desks
Kathy	sees	nine	small	chairs
Lucy	bought	seven	old	tables
Alan	gives	eight	dark	toys
Rachel	sold	four	heavy	spoons
William	prefers	nineteen	green	windows
Steven	has	two	cheap	sofas
Thomas	kept	fifteen	pretty	rings
Doris	ordered	twelve	red	flowers
Nina	wants	sixty	white	houses

Models of speech perception and cortical correlates

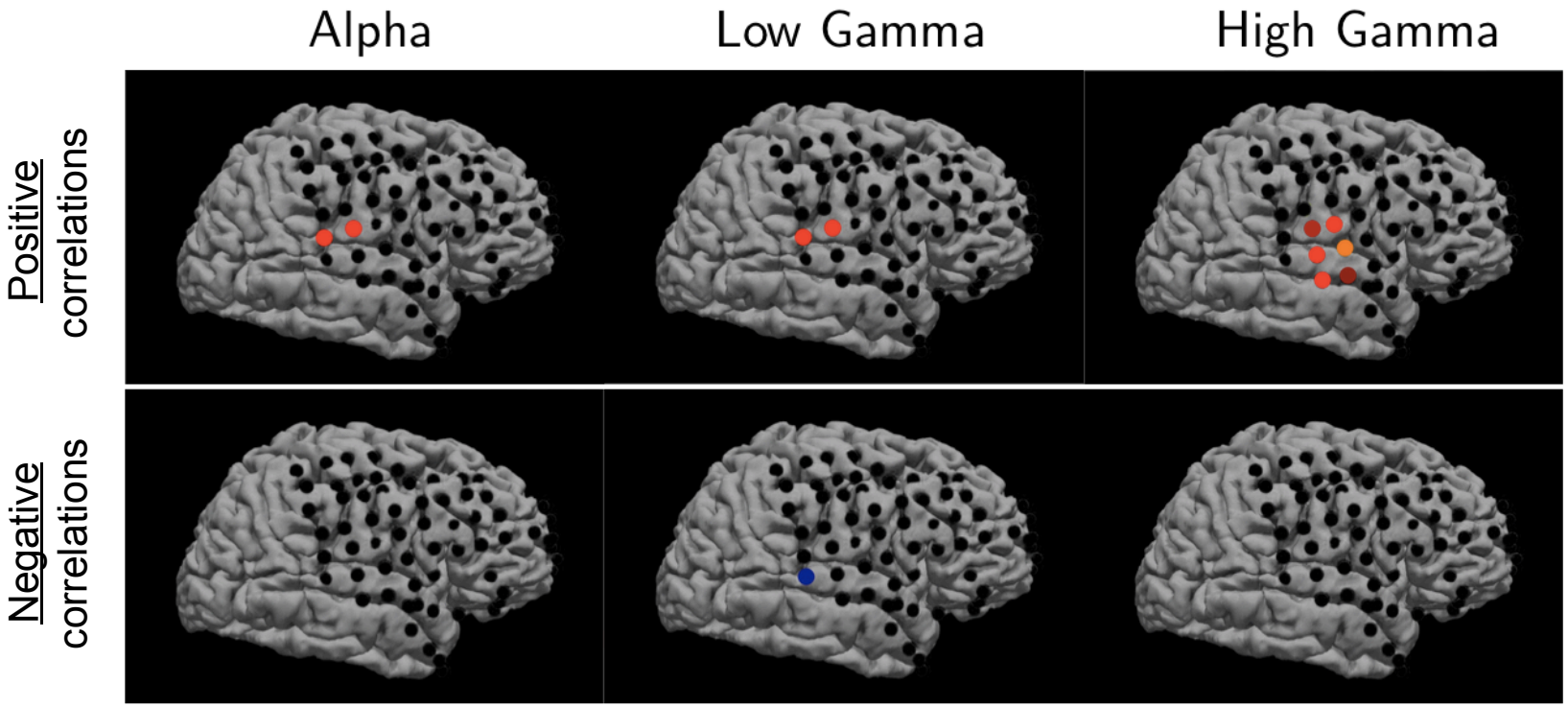


Comparison of frequency bands

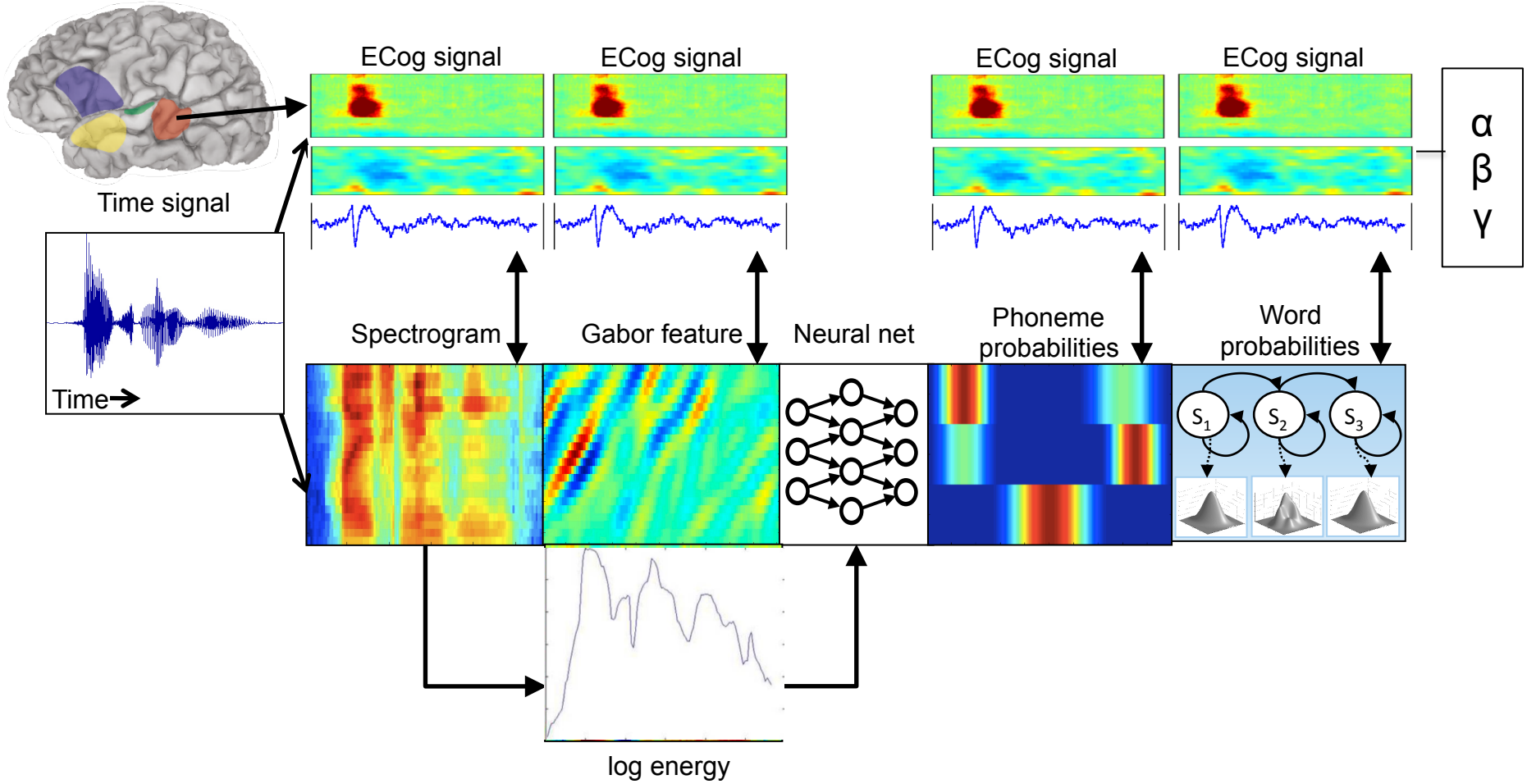


Models of speech perception and cortical correlates

Most positive significant correlations are found in posterior superior temporal gyrus in high gamma band (70-110 Hz)



Models of speech perception and cortical correlates



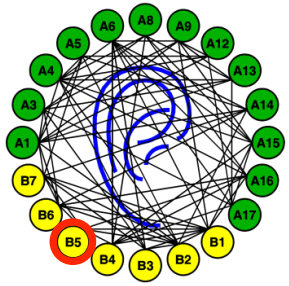
Summary

1. The auditory approach to speech processing often improves robustness of ASR systems

2. Models of speech intelligibility can profit from methods used in speech research.

3. Log energy is a decent 0th feature for analyzing activity data obtained with electrocortography.

Thanks




Constantin Spille


Jasper Ooster


Marina Frye


Angel Castro


Feifei Xiong


Franz Kunze

- Inga Schepers, Cristiano Micheli, Jochem Rieger from the Neupsychology department

Our research is funded by

- CRC (SFB/TRR) 31 "The active auditory system"
- Cluster of Excellence Hearing4all

Thank you for your attention!