Biosignal-based Spoken Communication

Tanja Schultz

Cognitive Systems Lab, Universität Bremen

Cognitive Systems Lab

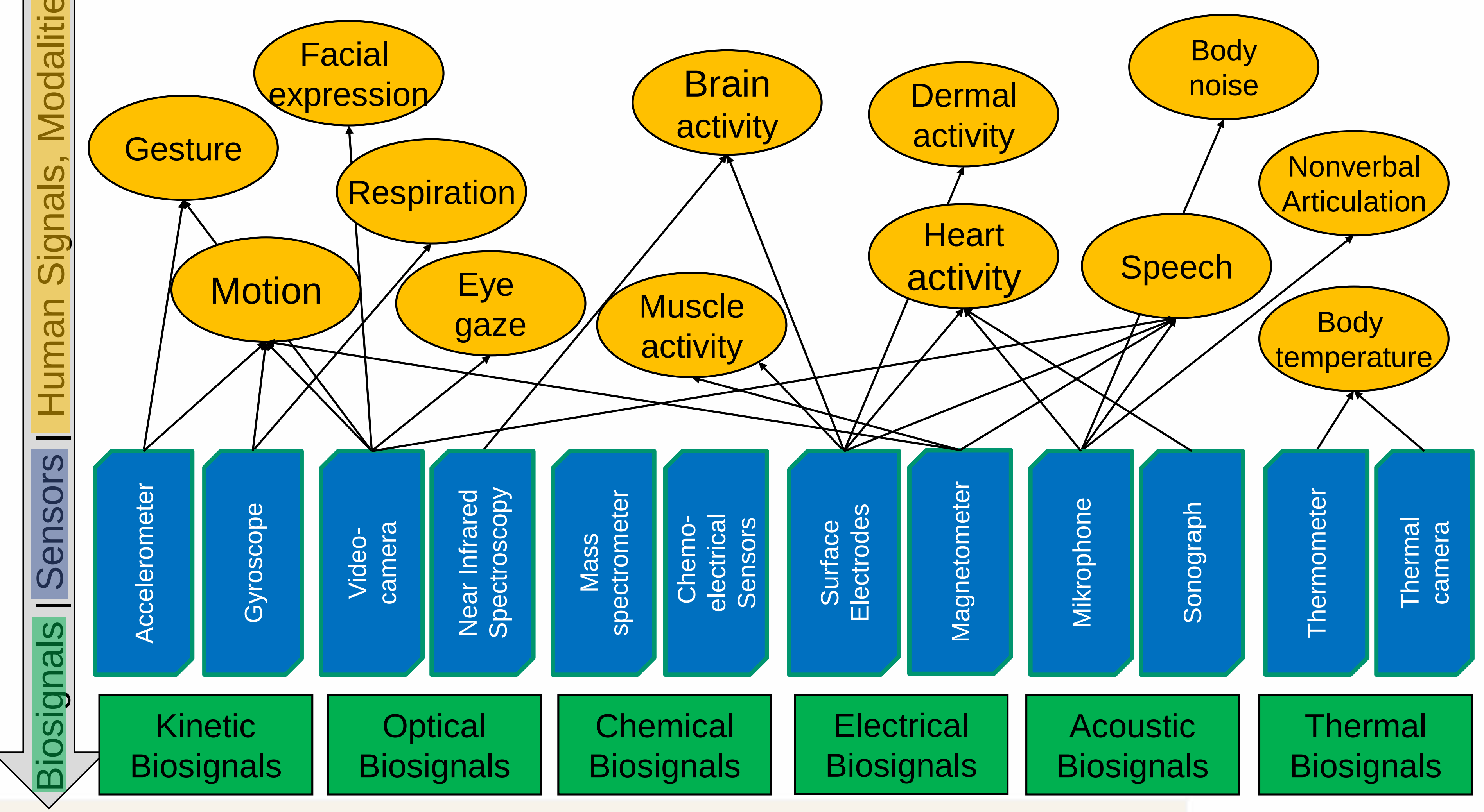
IberSPEECH2018
BARCELONA NOVEMBER 21-23

November 21st 2018

Universität Bremen

Definition Biosignals

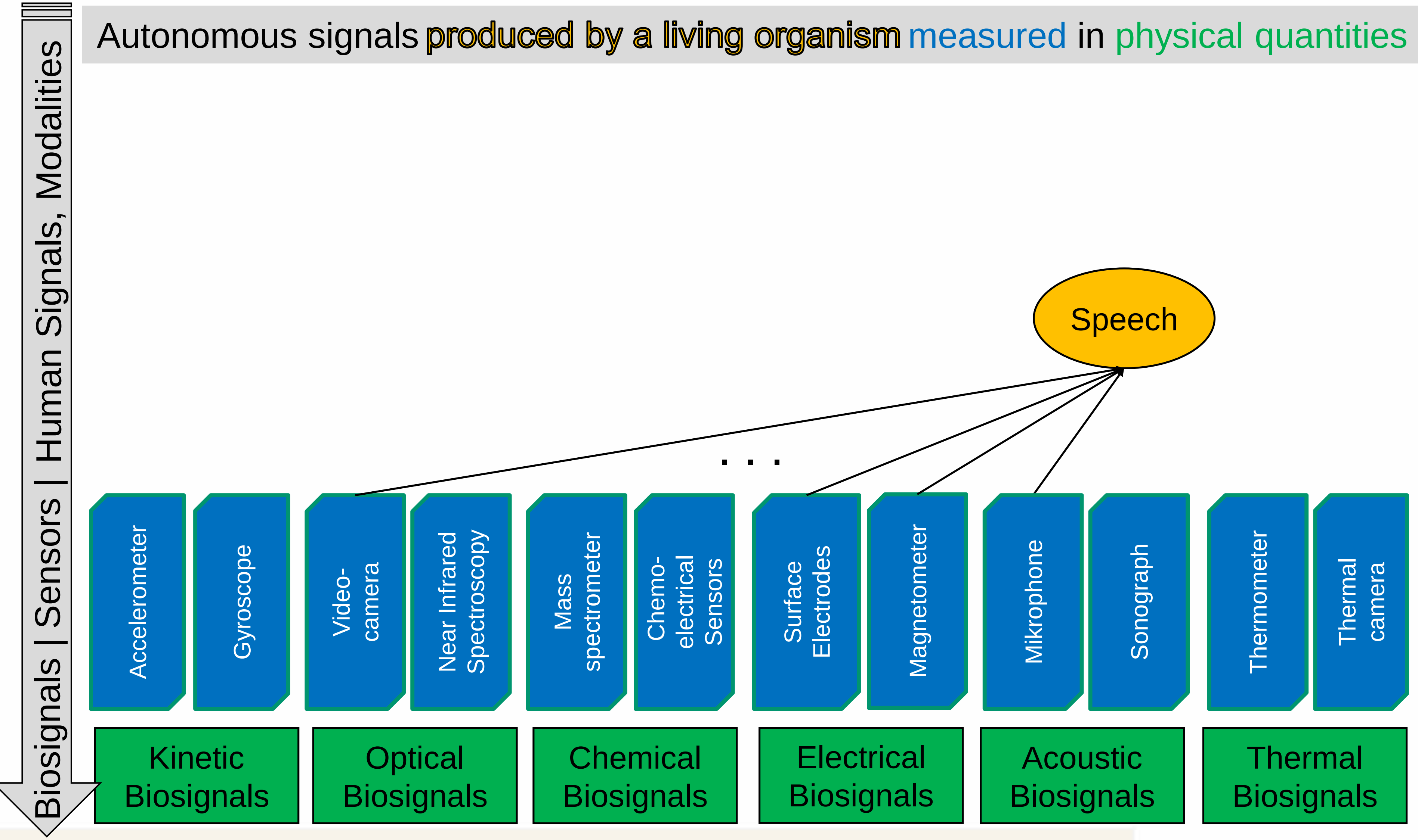
Autonomous signals produced by a living organism measured in physical quantities



T. Schultz, C. Amma, D. Heger, F. Putze, M. Wand Biosignale-basierte Mensch-Maschine-Schnittstellen, In at - Automatisierungstechnik, 2013, volume 61, 2013.

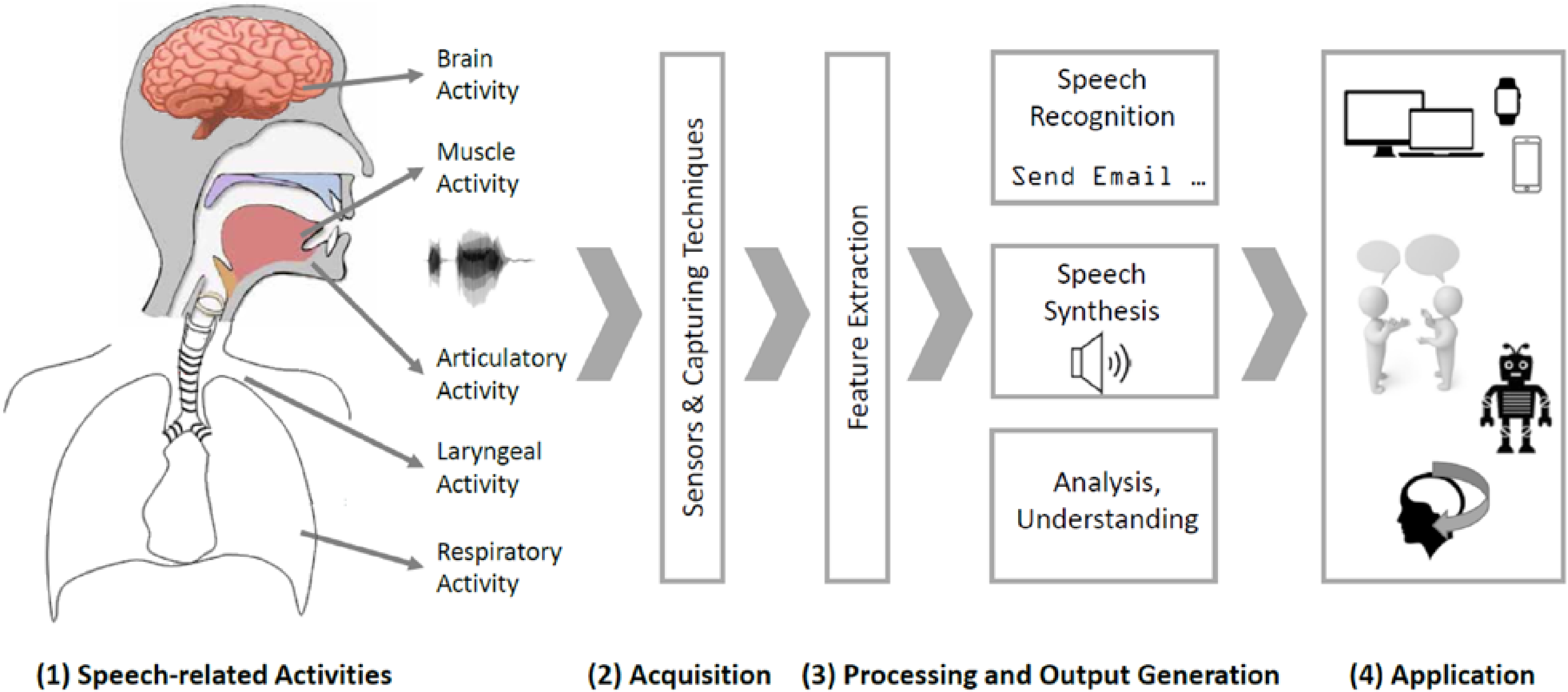
Definition Biosignals

Autonomous signals produced by a living organism measured in physical quantities



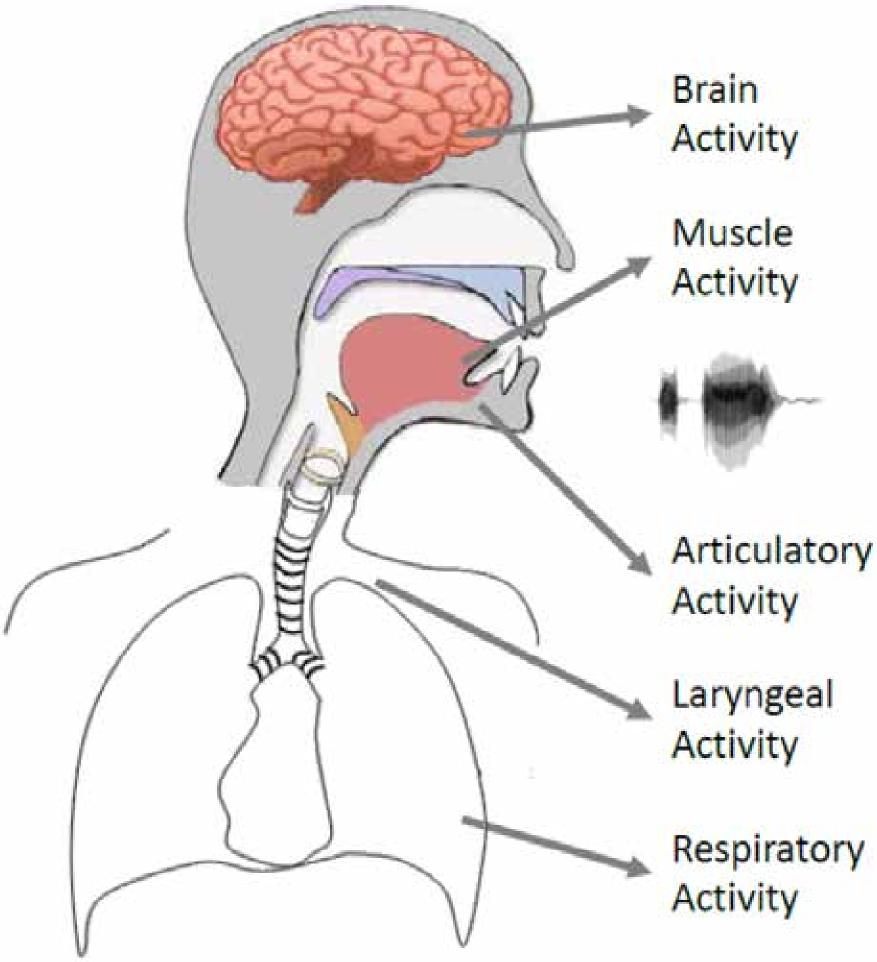
T. Schultz, C. Amma, D. Heger, F. Putze, M. Wand Biosignale-basierte Mensch-Maschine-Schnittstellen, In at - Automatisierungstechnik, 2013, volume 61, 2013.

Biosignal-based Spoken Communication



T. Schultz, M. Wand, T. Hueber, D. Krusienski, C. Herff, J. Brumberg. Biosignal-based Spoken Communication: A Survey. In IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, pp 2257-2271, 2017.

Speech-related Activities and Biosignals



(1) Speech-related Activities

SPEECH production is a complex process resulting from human activities

It is …

- initiated in the brain, …
- leading to muscle activities that produce ...
- respiratory, laryngeal, and articulatory gestures which create acoustic signals

Speech-related activities can be measured at each level of speech processing, including

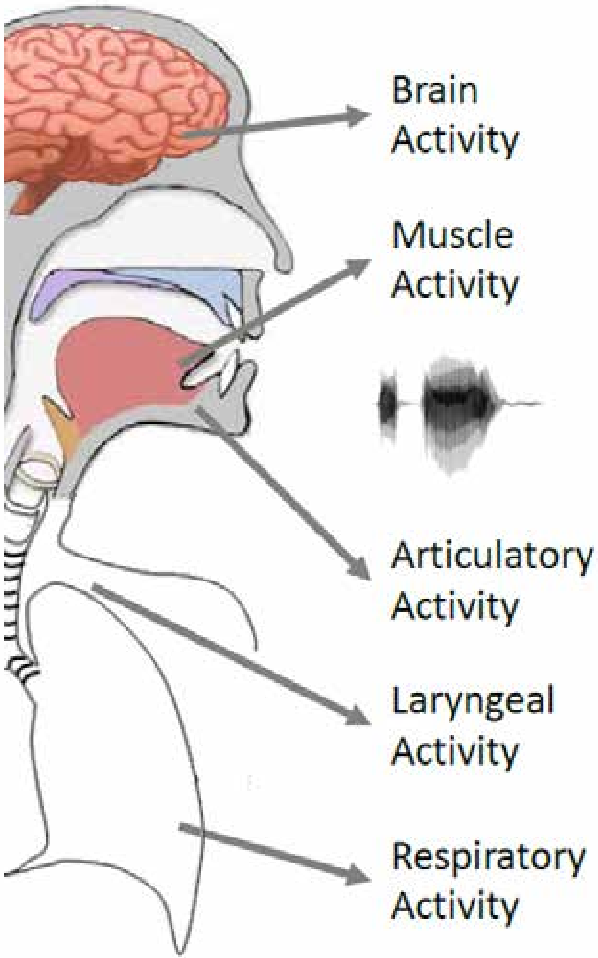
- the central and peripheral nervous system,
- muscular action potentials,
- speech kinematics.

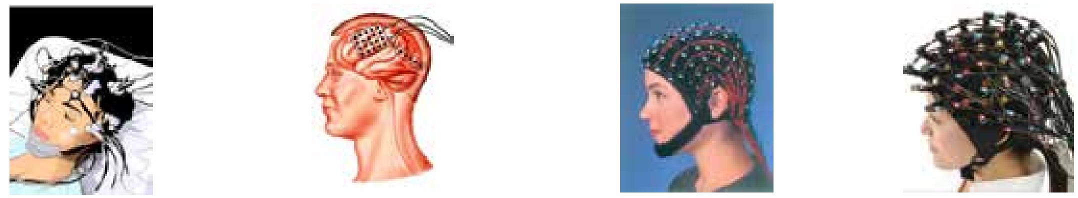
Their measurement,

- recorded with various sensor technologies, results in "**speech-related biosignals**"

Panopoly of Sensor Technologies

Cognitive Systems Lab





ECoG (Schalk, Herff), microelectrodes (Brumberg), EEG (Wester, D'Zmura), fNIRS (Herff/Schultz)

EMG (Jorgensen, Schultz), Lipreading (Petajan, others)

US+video (Hueber), EMA (Schönle), OPG (Birkholz), PMA (Gilbert/Gonzalez, Erro/Hernaez), NAM (Nakajima), intraoral (Bos), Radar, …

Universität Bremen

Why Biosignals for Spoken Comm.

- On the shoulders of giants:
  - Biosignals have been studied for decades to better understand the mechanisms of human speech processing

- Novel Applications, New insights
  - Traditional speech processing focuses mostly on acoustic
  - Alternative biosignals could overcome current limitations of speech processing for humans and machines, e.g.
    - Reduce delay: Capture speech-related activities prior to the airborne acoustic signal
    - Reduce disturbance: Capture speech-related activities even if no acoustic output is suitable/wanted
    - Extend applicability to otherwise mute people (e.g. laryngectomy)

Spectrum of Speaking Modes

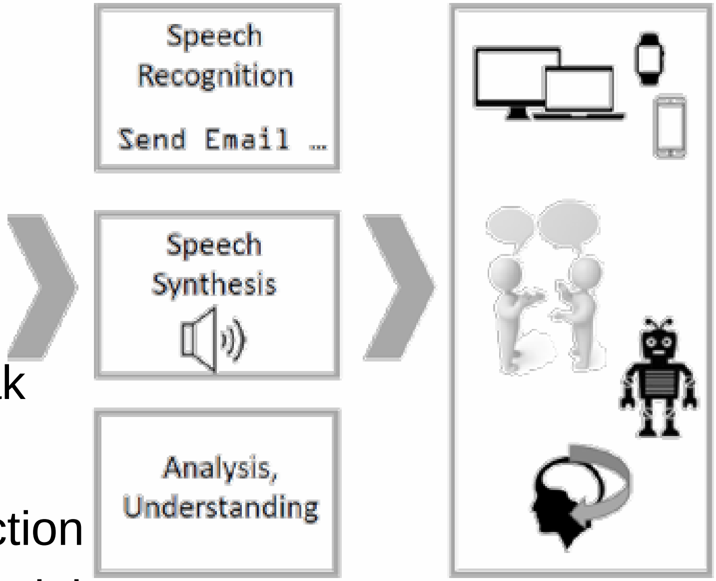
- Speaking modes – acoustic output available:
  - Modal (normal) speech: vocal folds vibrate for voice sounds
  - Whispered speech: turbulent flow through constant aperture between vocal folds
  - Different levels of effort: normal – shouted – murmured

- Speaking modes – no acoustic output available:
  - **Silent speech**: articulators are moving but airstream is suppressed (mouthing speech)
  - **Imagined speech**: like silent speech but no articulation movement (sometimes referred to as “attempted” speech)
  - **Inner speech**: internalized process in which one thinks in pure meaning (no phonological properties, no turn-taking, etc.)

Biosignal-based Spoken Communication

Like acoustics, speech-related Biosignals can be automatically processed:
- Feature extraction followed by speech recognition, speech synthesis, …

Opens up **novel use cases**, coined Biosignal-based Spoken Communication:
- **Robust Spoken Communication**
  - Enhance performance under adverse noise conditions
  - Fuse complementary biosignals
- **Mute-Spoken Communication**
  - Avoid disturbance in quiet environments
  - Secure against eavesdropping in public places
- **Restore Spoken Communication**
  - Voice prostheses for individuals unable to speak
- **Speech Training and Therapy**
  - Deliver articulatory biofeedback of voice production
  - Increase articulatory awareness for therapy & training



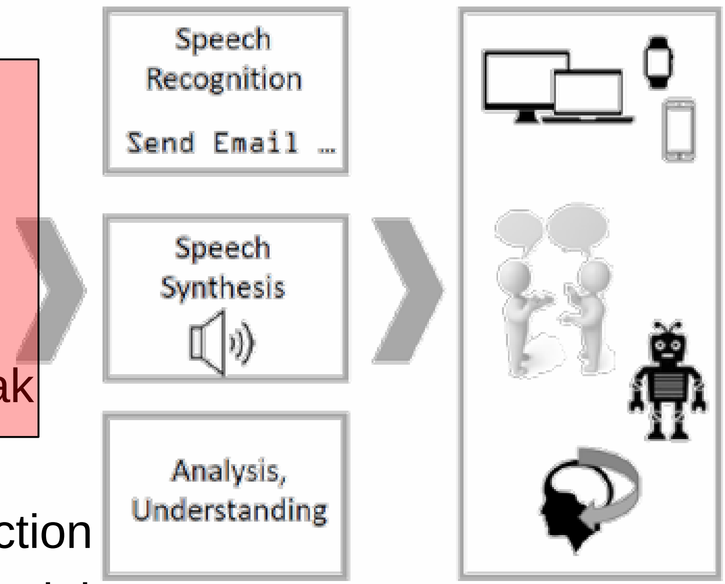
Biosignal-based Spoken Communication

Like acoustics, speech-related Biosignals can be automatically processed:

- Feature extraction followed by speech recognition, speech synthesis, …

Opens up **novel use cases**, coined Biosignal-based Spoken Communication:

- **Robust Spoken Communication**
  - Enhance performance under adverse noise conditions
  - Fuse complementary biosignals
- **Mute-Spoken Communication**
  - Avoid disturbance in quiet environments
  - Secure against eavesdropping in public places
- **Restore Spoken Communication**
  - Voice prostheses for individuals unable to speak
- **Speech Training and Therapy**
  - Deliver articulatory biofeedback of voice production
  - Increase articulatory awareness for therapy & training

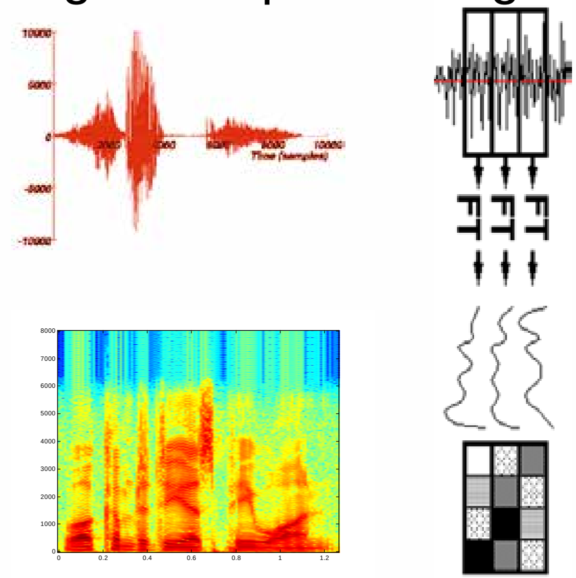


Automatic Speech Recognition (ASR)

Speech Signal Capturing

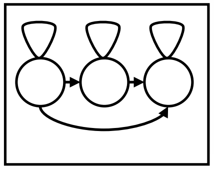
Modality: Speech
Sensor: Microphone
Acoustic Biosignal

Signal Preprocessing



Automatic Speech Recognition (ASR)

Acoustic



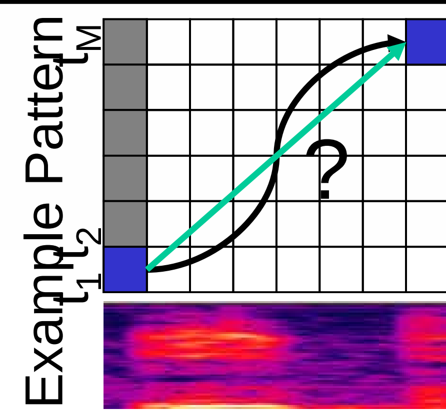
Dictionary

I	/i/
you	/j/ /u/
we	/w/ /e/

Language Model

I am
You are
We are

$$\arg\max_{W} P(W \mid X) = \arg\max_{W} P(W) \times p(X \mid W) \times \frac{1}{P(x)}$$



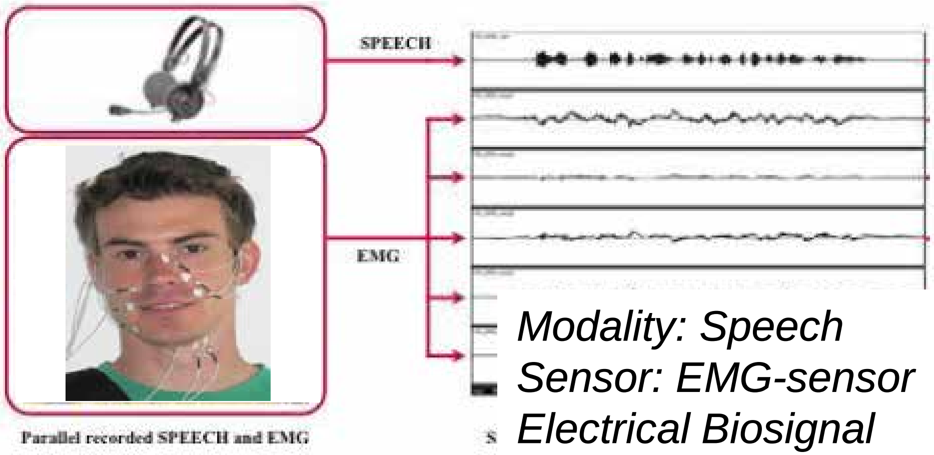
Text Output

"Hello"

Use Muscle Activity instead of Acoustics

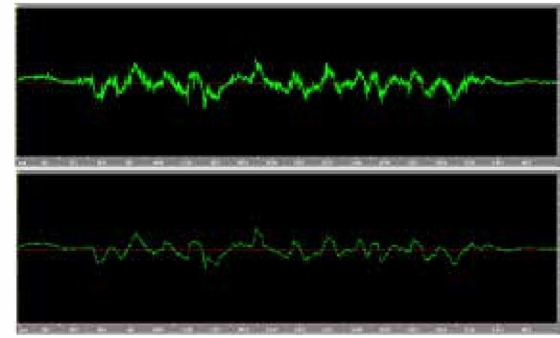
Cognitive Systems Lab

Speech Signal Capturing



Modality: Speech
Sensor: EMG-sensor
Electrical Biosignal

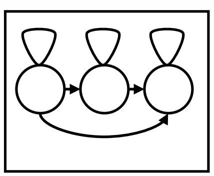
Signal Processing



- Time domain features
- Artefact reduction
- Contextual features
- Compression

Automatic Speech Recognition

Acoustic



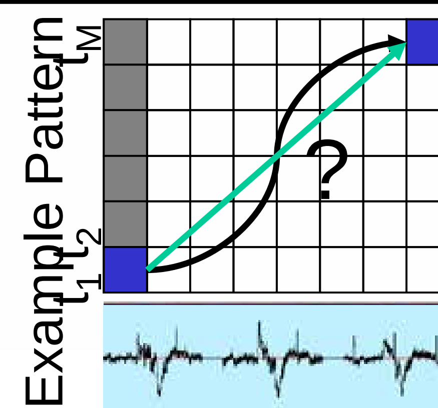
Dictionary

I /i/
you /j/ /u/
we /w/ /e/

Language Model

I am
You are
We are

$$\arg\max_{W} P(W \mid X) = \arg\max_{W} P(W) \times p(X \mid W) \times \frac{1}{P(x)}$$



Text Output

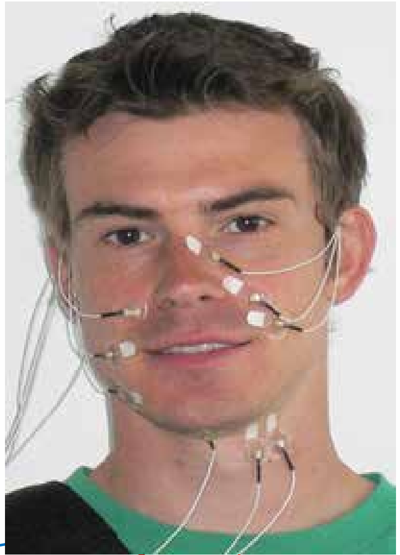
"Hello"

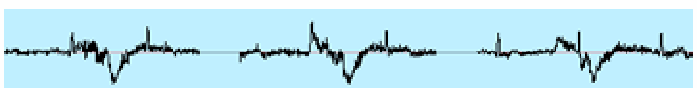
Maier-Hein et al, ASRU 2005, Jou/Schultz 2006-2009, Wand/Schultz 2007-2014, Janke/Schultz 2010-2016, Diener/Schultz 2015-
Wand, Janke, Schultz: Tackling Speaking Mode Varieties in EMG-based ASR, IEEE Biomedical Engineering, Vol 61, 2014.

Silent Speech Interfaces: EMG

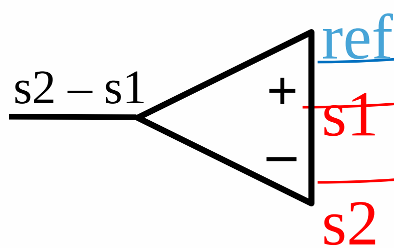
<u>Surface ElectroMyoGraphy (EMG)</u>

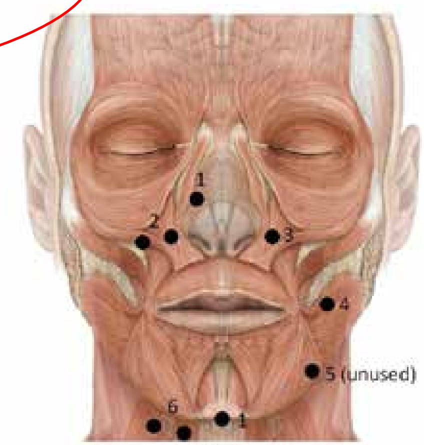
- Surface = No needles
- Electro = electrical activity
- Myo = muscle
- Graphy = recording





EMG-Signal „zero zero zero“





- Speech results from the activity of articulatory muscles
- Electrodes capture the electrical potentials of the muscle activity in the face
- EMG records *Motion*, not the acoustic signal

Denby, Schultz, Honda, Hueber, Gilbert, Brumberg (2010): Silent Speech Interfaces. Speech Communication, Vol 52 (4).

Silent Speech Interfaces: Benefits

EMG records *motion*, not acoustics Þ Silent Speech can be processed

In **Silent speech** the speakers are instructed to move their articulators as if they were producing normal modal speech but to suppress the pulmonary airstream, so that no sound is heard

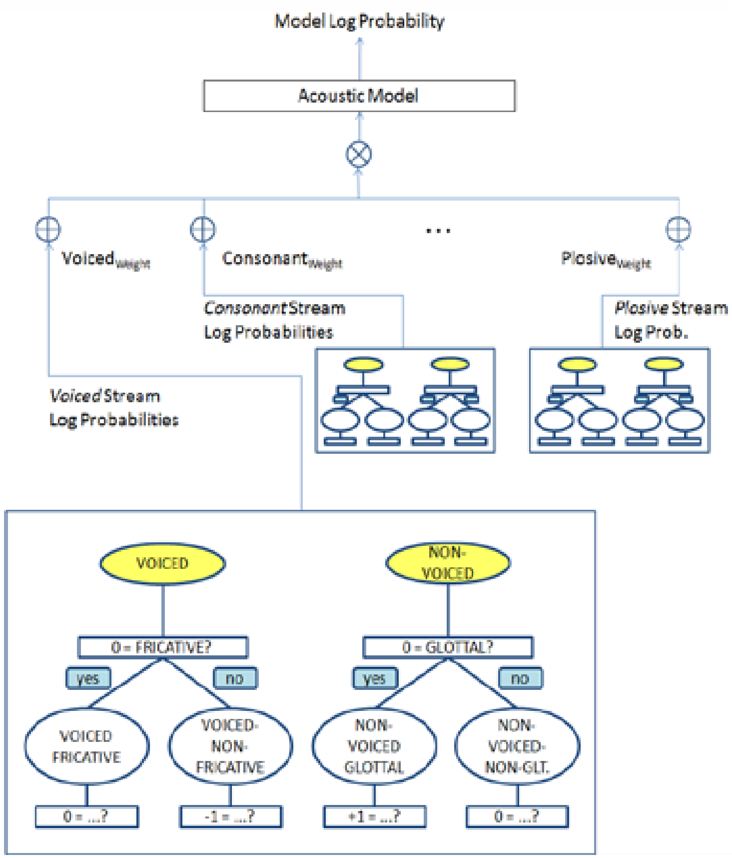
- No Disturbance: Speak silently in quiet environments
- Keep your Privacy: Transmit confidential information
- Noise Robustness: No corruption in noisy environment
- Speech Augmentation: Support speech impaired people

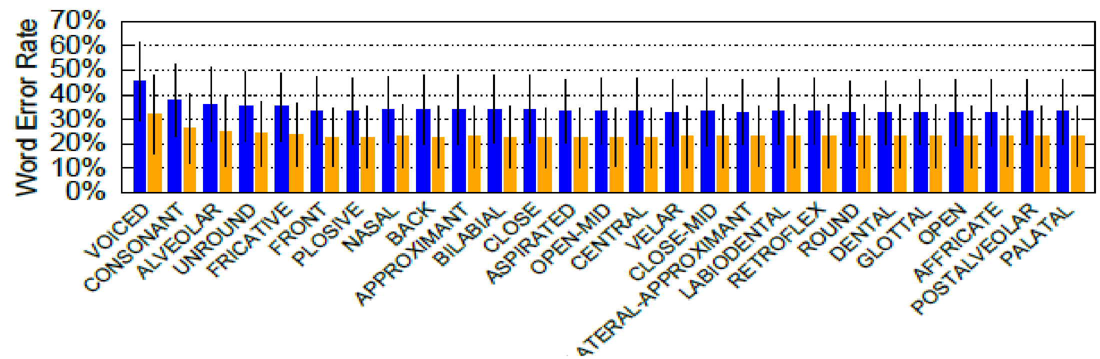
Challenge 1 – Low-ressource ASR

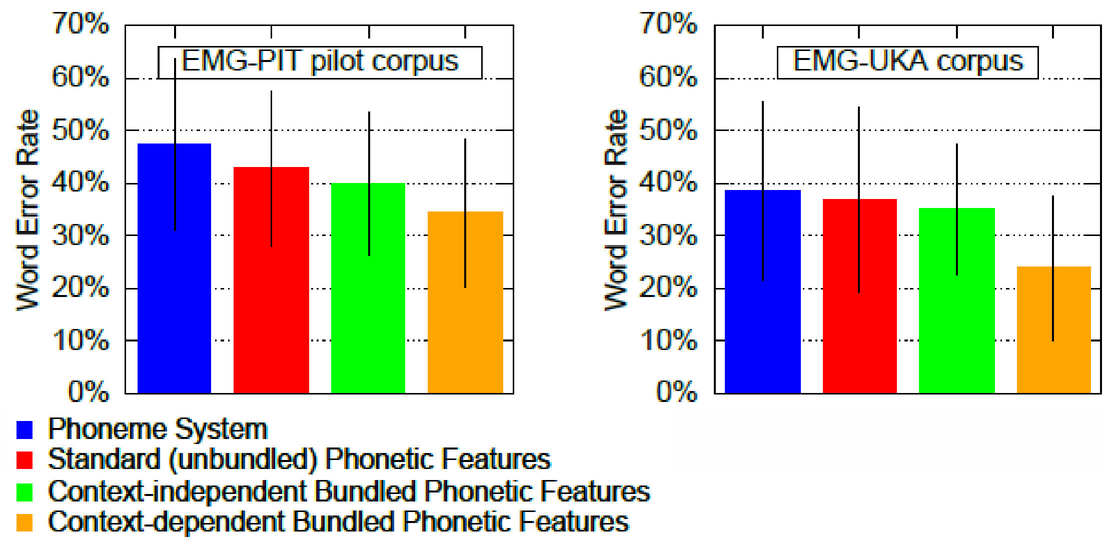
- **Bundles of Phonetic Features** (BDPF) (e.g. voiced fricative, …)
- Context dependent modeling (using decision trees)

– *Multi-Stream* decoding system: nine most frequent PFs

Þ **~ 30% relative WER gain**

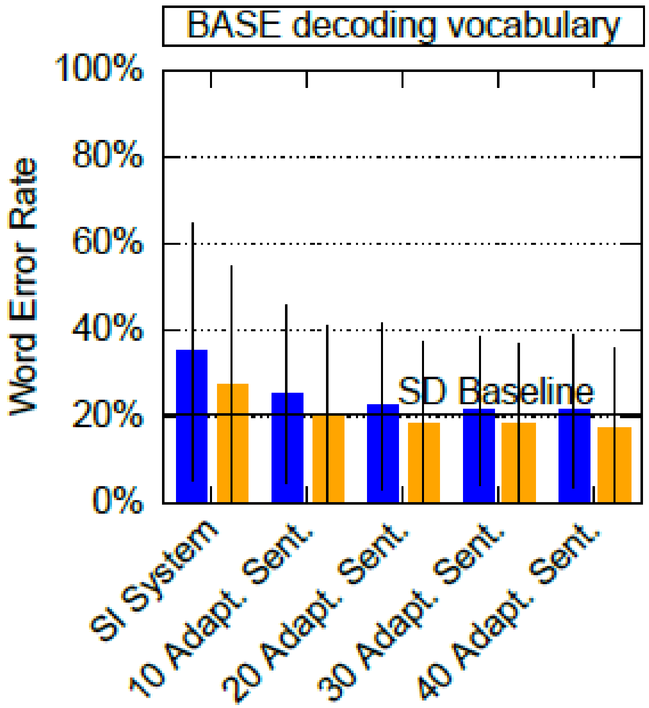


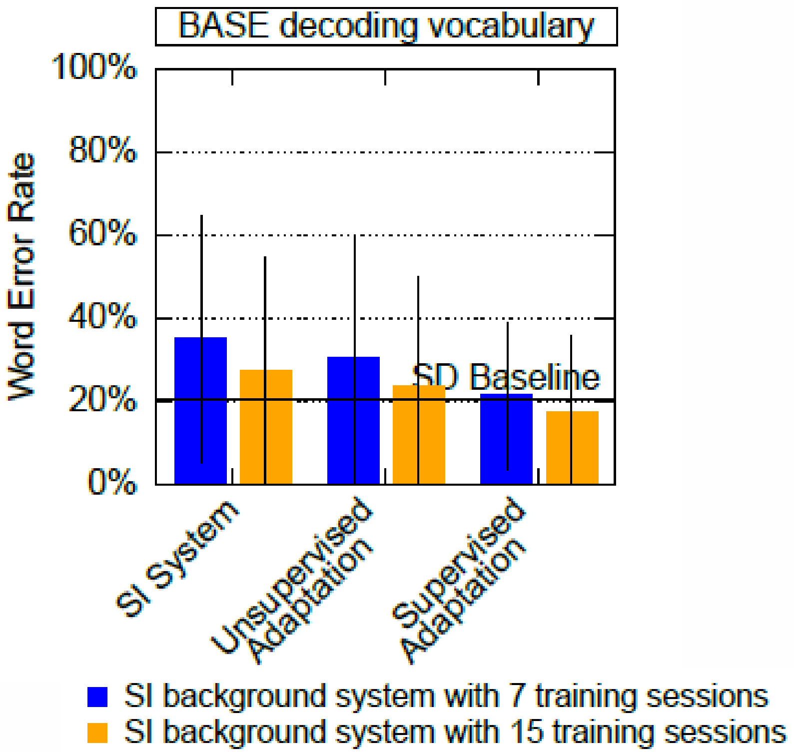




T. Schultz, M. Wand: *Modeling Coarticulation in EMG-based Continuous ASR.* Speech Communication, 52 (4), 2010

Challenge 2 – Session/Spk Dependencies



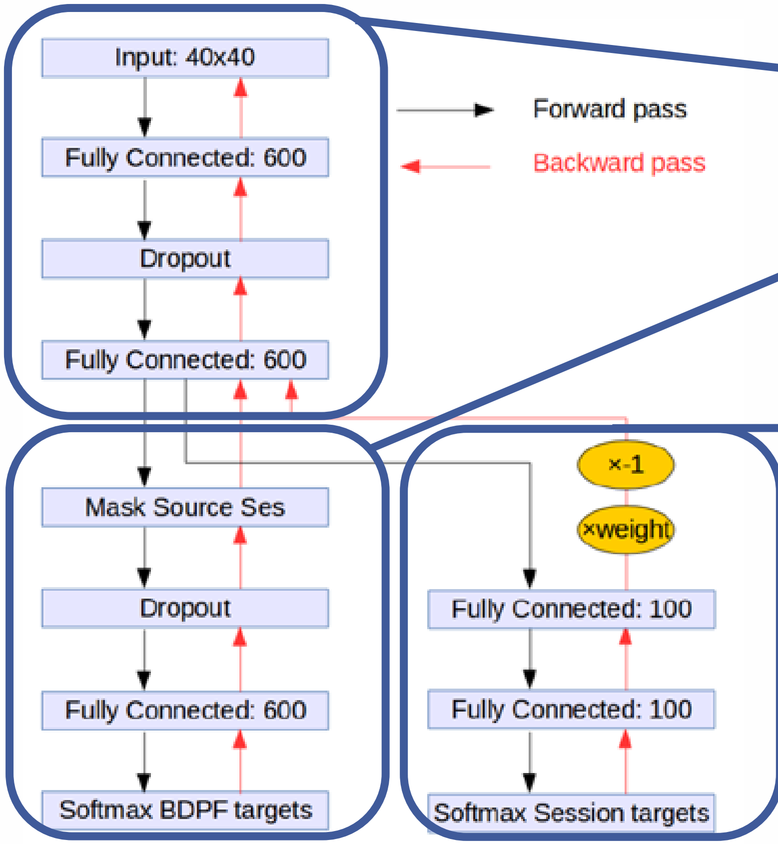


Lessons Learned: What works best

- Train Session-Independent (SI) Systems (the more sessions the better)
- Rapidly adapt SI System to session, MLLR, unsupervised okay
- Training across sessions works well, across speakers not (yet?)

Deep EMG-to-Text (ASR)

Apply Adversarial Training to session-independent EMG-based ASR, i.e. make data of different sessions more confusable, to improve the target classification accuracy



Common part

BDPF state classifier
(only on source sessions)

Session classifier
- inverted gradient (confuse sessions)
- configurable contribution

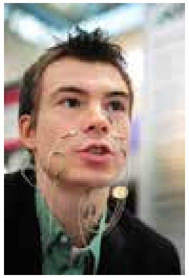
	Dataset	
System	**Development**	**Evaluation**
GMM frontend	30.6%	41.4%
DNN frontend	18.8%	29.9%
DNN + adv. training	15.6%	28.5%

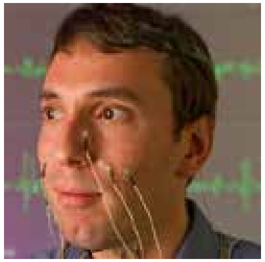
M. Wand, T. Schultz, J. Schmidhuber: *Domain-Adversarial Training for Session Independent EMG-based Speech Recognition, Interspeech 2018*

EMG Multi-speaker, Multi-Session Data

Cognitive Systems Lab

- EMG-PIT corpus: 78 subjects, 18-35 yrs, normal vocal qualities
- About 12 hrs read speech, BN style, large vocabulary
- Audible (normally spoken) and Silent (mouthed)





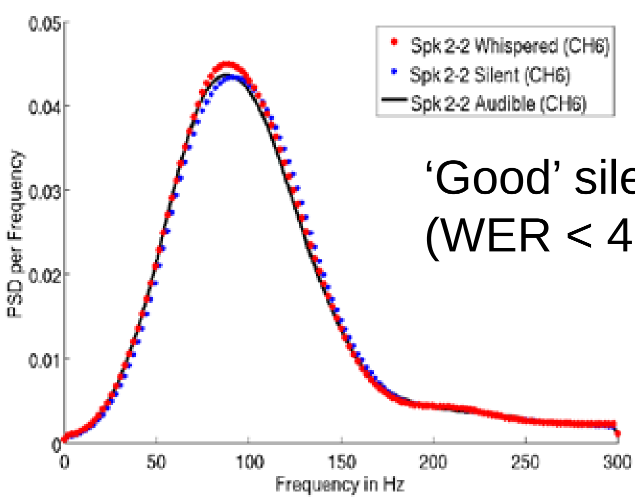
Phase	Speakers	Sessions	Utterances		Duration [min]	
			Audible	Silent	Audible	Silent
Pilot	14	28	1400	1400	108	110
Main	64	64	3200	3200	287	251
Total	78	92	4600	4600	395	361

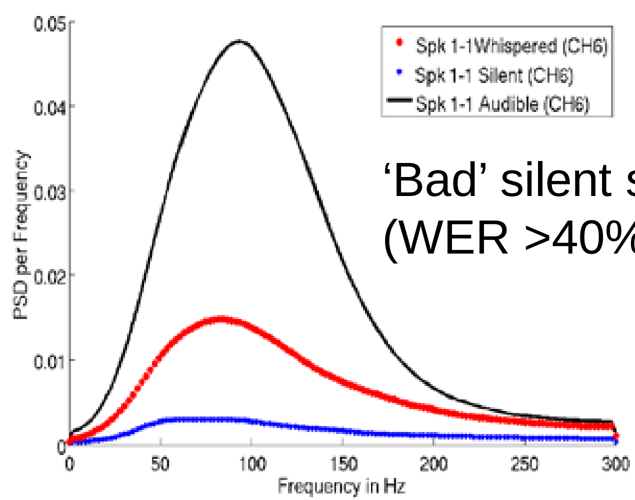
- EMG-UKA corpus: Many sessions of eight subjects, same scenario
  - Audible, Silent & Whispered speaking mode
  - Study Impact of speaking modes
- **Free download of trial corpus** (benchmarks in paper)
- **Full corpus available via ELRA** (research and commercial license)

Wand/Schultz: “The EMG-UKA Corpus for Electromyographic Speech Processing”, Interspeech 2014

Challenge 3 – Lack of Auditory Feedback

- EMG signals of Silent speech are different from those of audible speech
- Effect weaker for experienced speakers; group “good/bad” speakers
- Power Spectral Densities (PSD) of audible, whispered and silent EMG
  Significantly smaller variations for “good” speakers

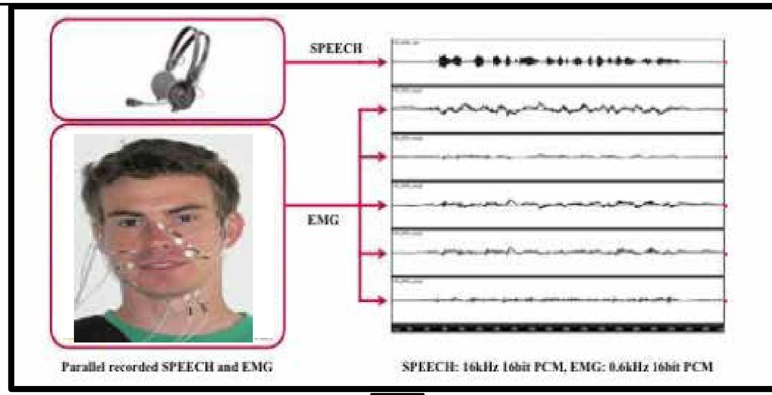




- *Spectral Mapping Algorithm* to compensate for differences: ~12% rel D

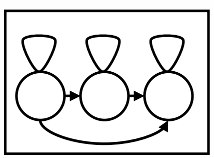
Ultimate Cure: Provide instant auditory feedback ® Direct Synthesis

Two Methods: ASR versus Direct Synthesis



Signal Processing
A/D, Artifacts, Feature Extraction …

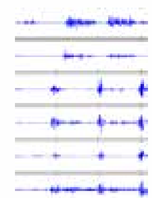
Automatic Speech Recognition (ASR)



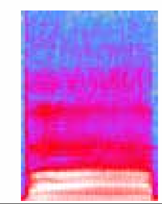
I /i/ you /j/ /u/ we /v/ /e/	I am you are we are

TEXT: Hello World

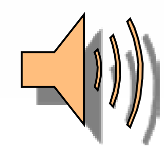
Feature Transform + Vocoding



EMG-2-MFCC



SPEECH:



EMG-to-Text

EMG-to-Speech

Two Methods: ASR versus Direct Synthesis

ASR – PROS

- High output quality
- Text for application backend

ASR – CONS

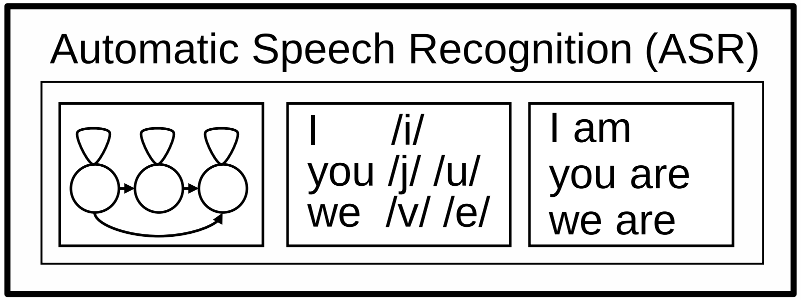
- Limited vocabulary
- Recognition errors
- No emotion, emphasis, …

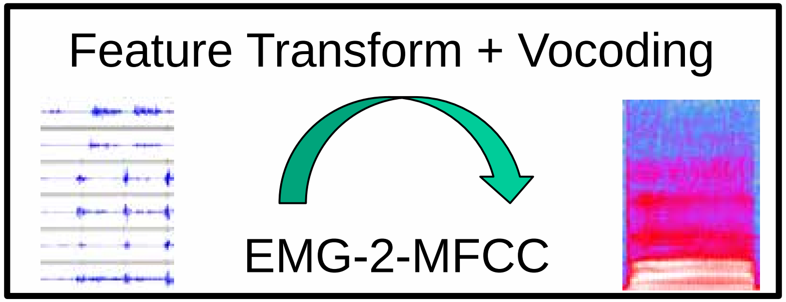
Direct Synthesis – PROS

- No vocabulary restrictions
- Speaker identity, emotion, …
- Minimal delay: user-in-the-loop

Direct Synthesis – CONS

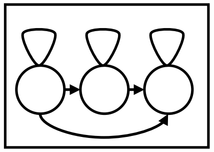
- Output quality (quality vs time)
- No text for application backend





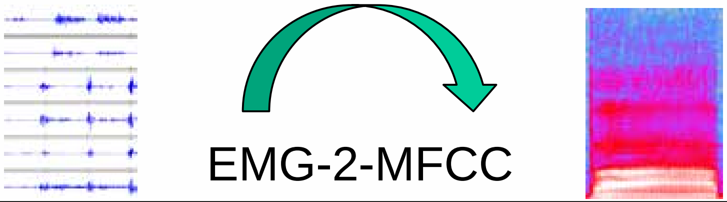
Two Methods: Applications

Automatic Speech Recognition (ASR)



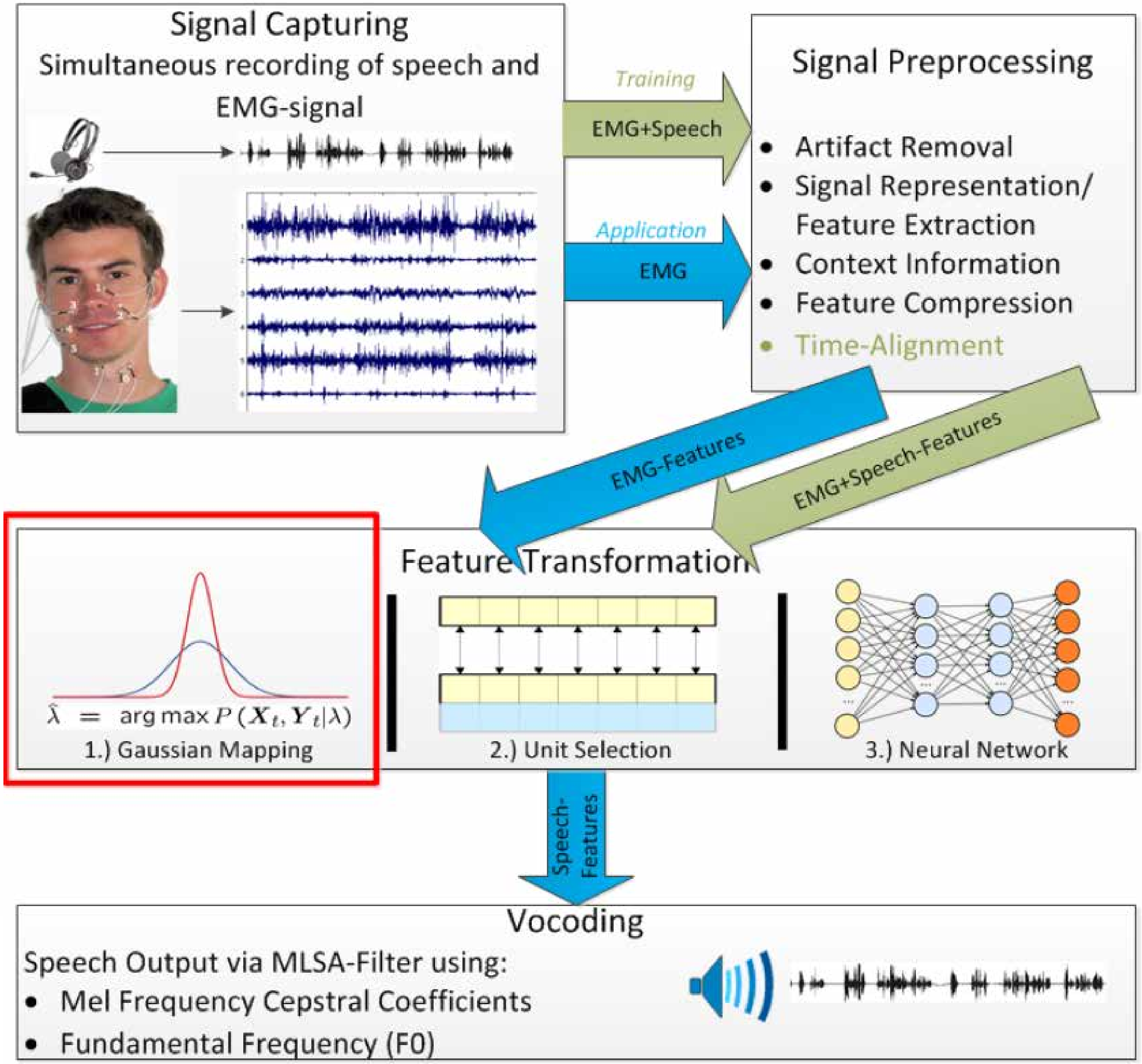
I /i/ you /j/ /u/ we /v/ /e/	I am you are we are

Feature Transform + Vocoding



ASR	Application	Direct Synthesis
YES	Robust Spoken Communication (indirect)	YES
YES	Mute Spoken Communication (indirect)	YES
YES	Silent Command & Control	(no text)
NO	User-in-the-loop, Coadaptation	YES
NO	Biofeedback for Therapy and Training	YES
NO	Voice Prostheses (face-to-face)	YES

EMG-to-Speech (Feature Transform + Vocoding)



Feature Transformation – 3 Approaches

Cognitive Systems Lab

Gaussian Mapping

Source feature vectors

$$x_t = [x_t(1), \ldots, x_t(d_x)]^\top$$

Target feature vectors

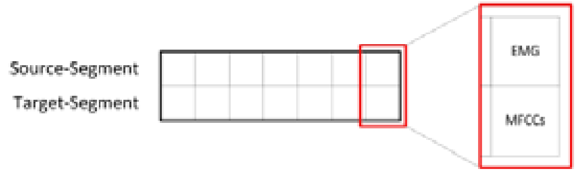
$$y_t = [y_t(1), \ldots, y_t(d_y)]^\top$$

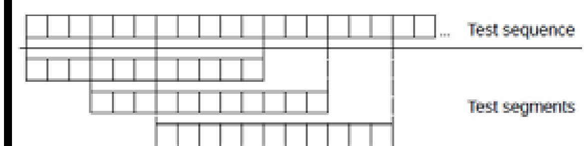
Train GMM to describe the joint probability density of source and target feature vectors:

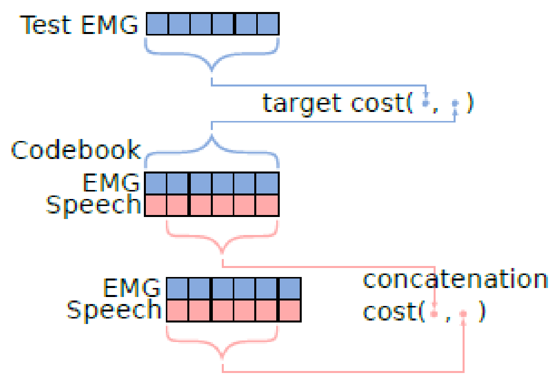
$$P(x_t, y_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}\left([x_t^\top, y_t^\top]^\top; \mu_m^{(x,y)}, \Sigma_m^{(x,y)}\right),$$

$$\mu_m^{(x,y)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \qquad \Sigma_m^{(x,y)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix},$$

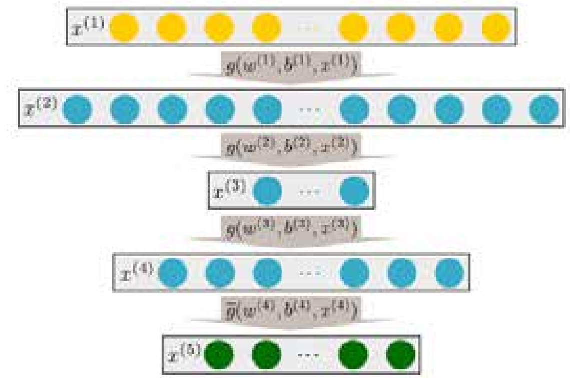
Unit Selection

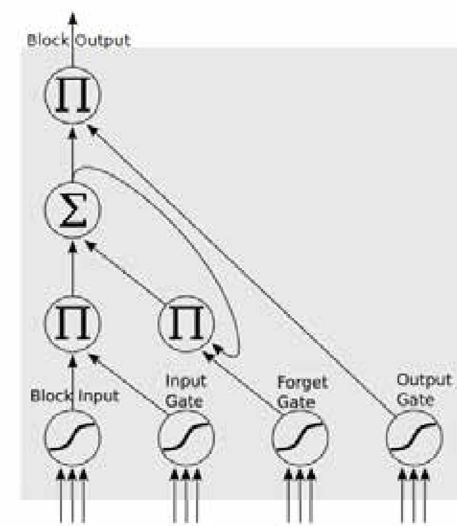




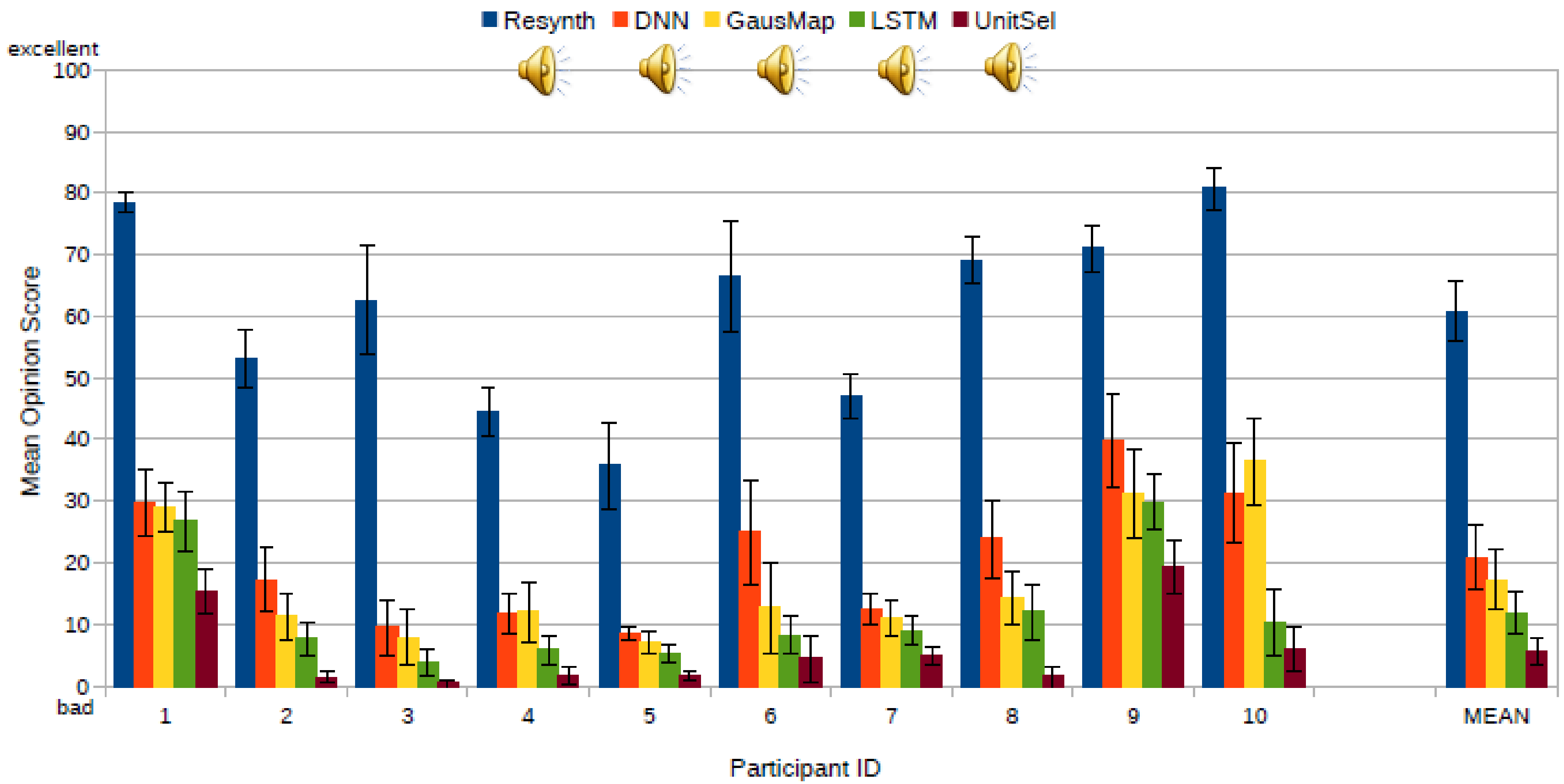


Neural Networks





Approach Comparison (Subjective Eval)



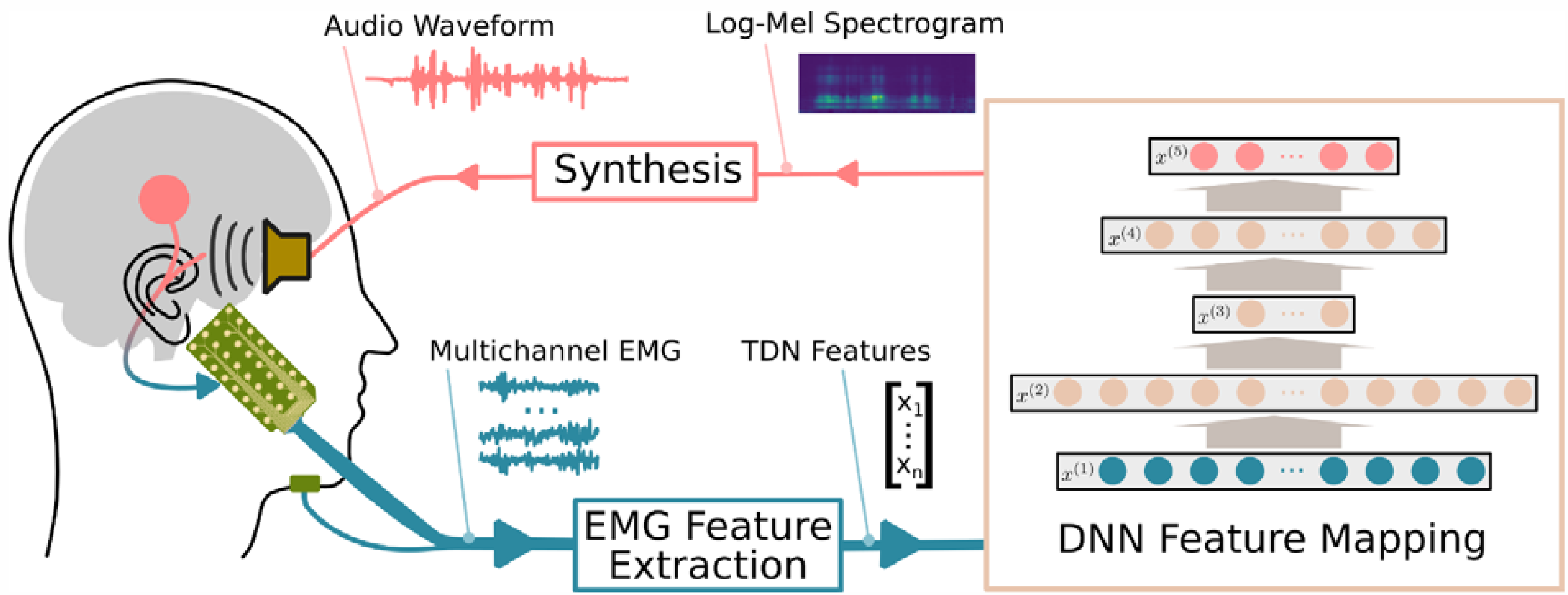
Mean Opinion Scores from Listening test: Resynthesized reference (Resynth), 10 subjects rated the speech quality from 0 (bad) to 100 (excellent); error bars = standard deviation

M. Janke and L. Diener, “EMG-to-speech: Direct generation of speech from facial electromyographic signals,” IEEE/ACM Trans. Audio, Speech, Language Process., Special Issue Biosignal-based Spoken Communication, December 2017.

Low-Latency EMG-to-Speech

Real-time speech output to enable

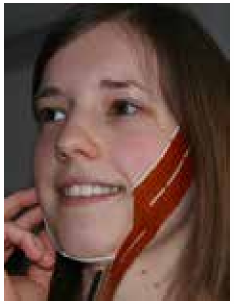
- Natural Conversation (retain paralinguistic information)
- Auditory feedback with acceptable delay



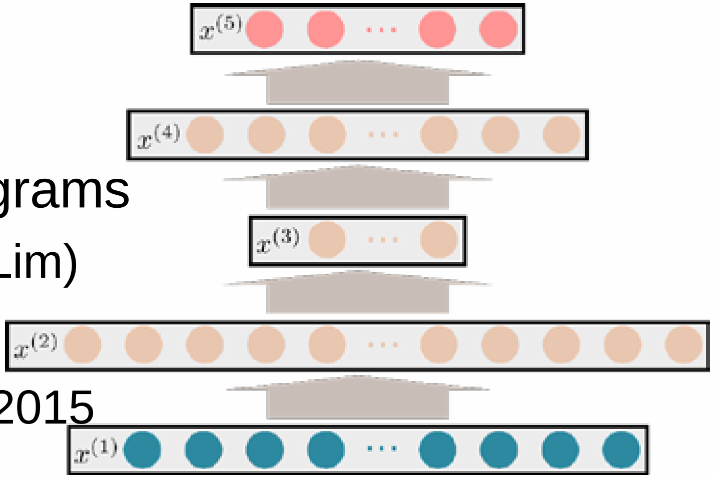
Diener/Schultz: Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion. Interspeech 2018

Pilot Experiments

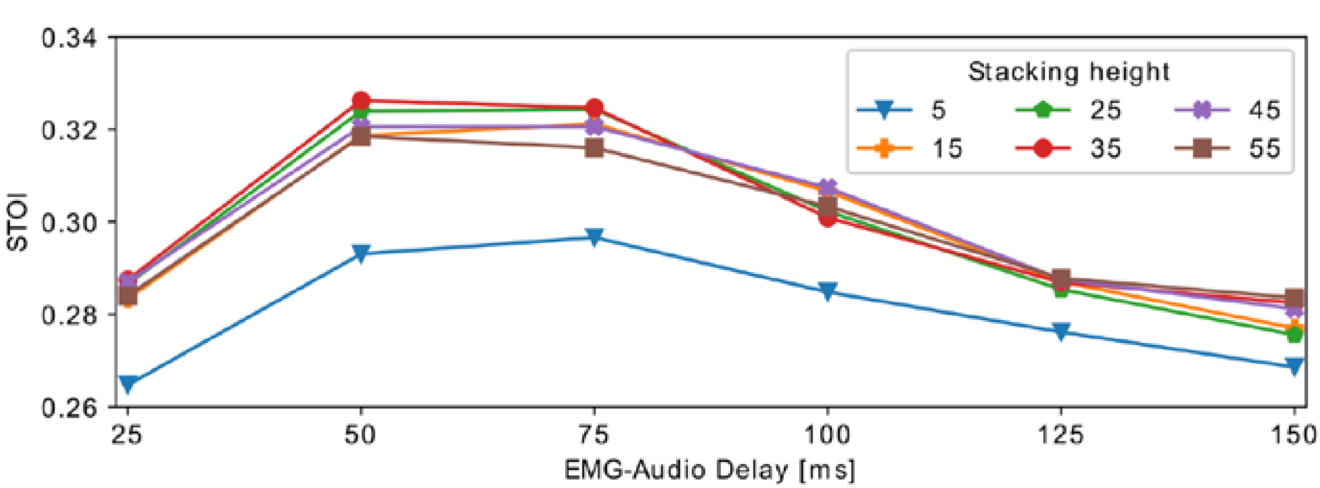
- Challenges: few data (session dependence, time constraints)
- First study on 1 speaker, 2 sessions
  - Array: More sensors, easy-to-use, set up time, …
  - 300 utts = 20min speech / session
  - 135/200 utts training, rest dev and eval,
  - Spectral frame based measure of intelligibility (STOI)

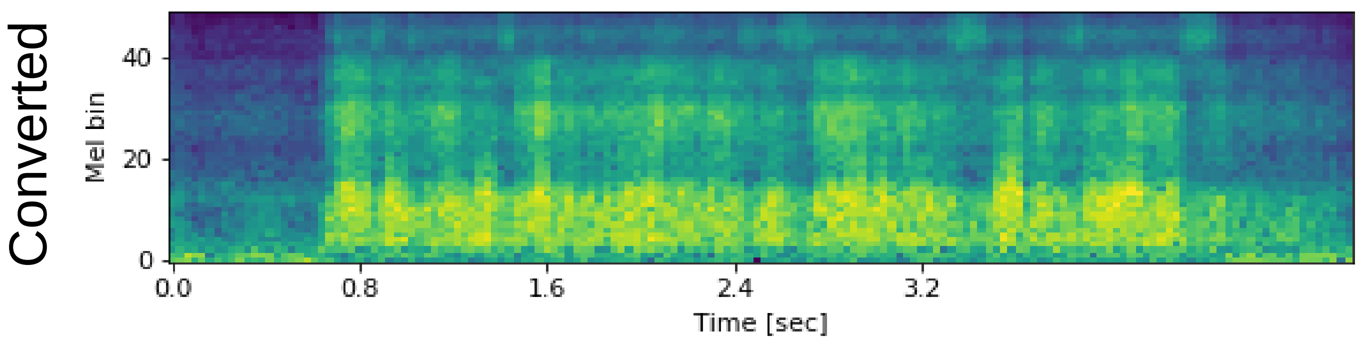


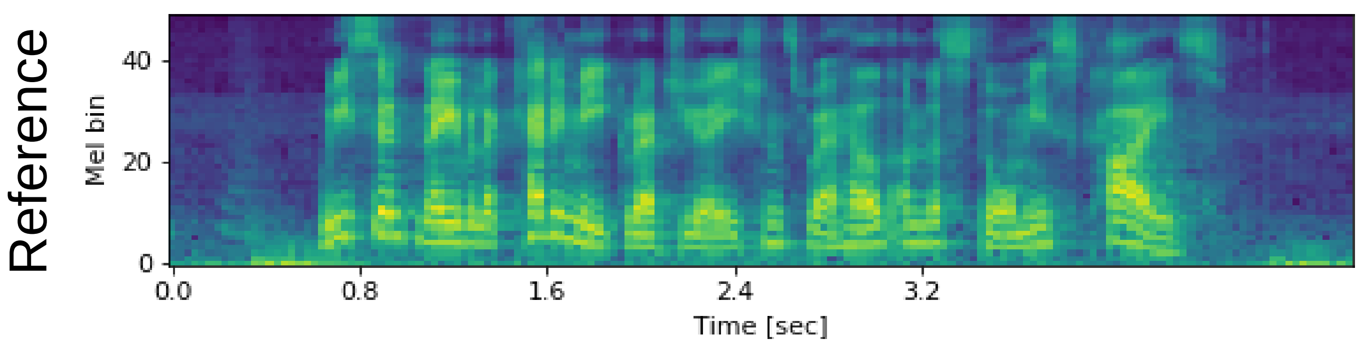
- Features: usual time-domain TDN
  - Stacked into the past only
  - 32ms window size, 5ms shift works best
- Speech output representation: Mel spectrograms
  - Waveforms by phase reconstruction (Griffin-Lim)
- EMG-to-Mel Conversion: Feedforward NN
  - Same shape but different sizes as in Janke, 2015



Results: EMG-Audio Offset (training)







→ 50ms delay, i.e.: (EMG preceeds Audio) confirms Jou, 2006

→ Stacking: more is better but limited by data due to dimensionality

→ Fair ressemblance but still rather noisy

Up to the Brain: Imagined Speech

Acoustic Speech Recognition

– Audible Speech produced by the (excited) human articulatory apparatus

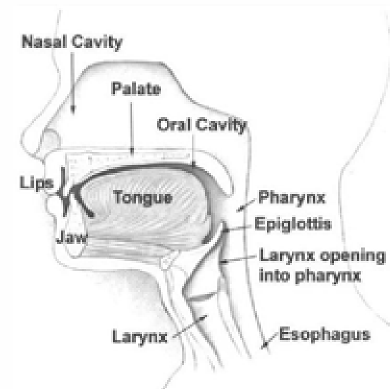
Þ **Traditional Speech-to-Text**

Silent Speech Recognition

– Silent Speech captured by muscle activities which move the articulatory apparatus

– Speech involves innervation of muscles

Þ **EMG-to-Text, EMG-to-Speech**

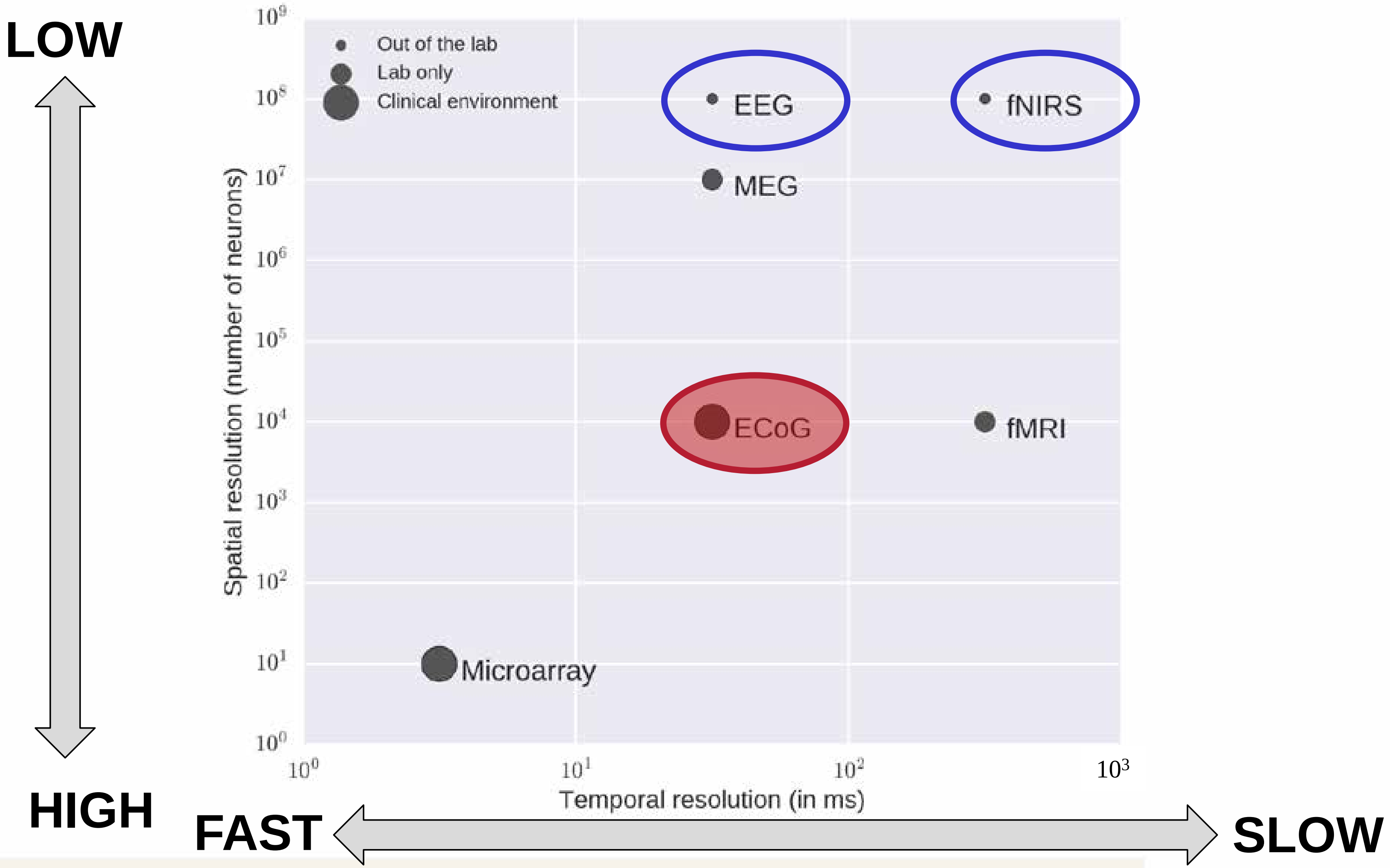


Imagined Speech Recognition

– Thinking about producing speech

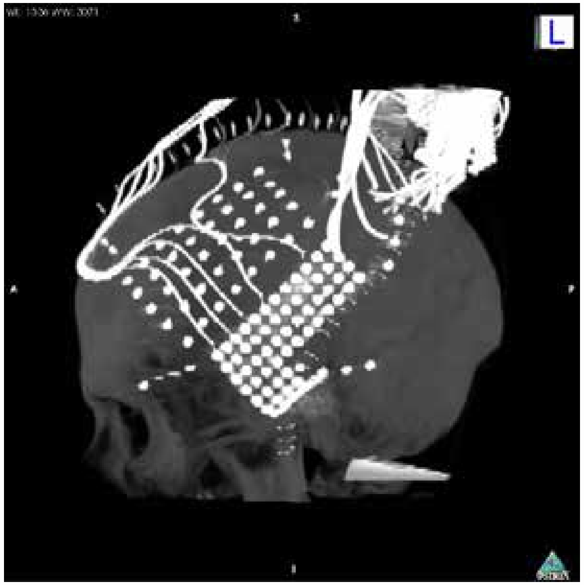
Þ **Brain-to-Text, Brain-to-Speech**

Measuring Brain Activity

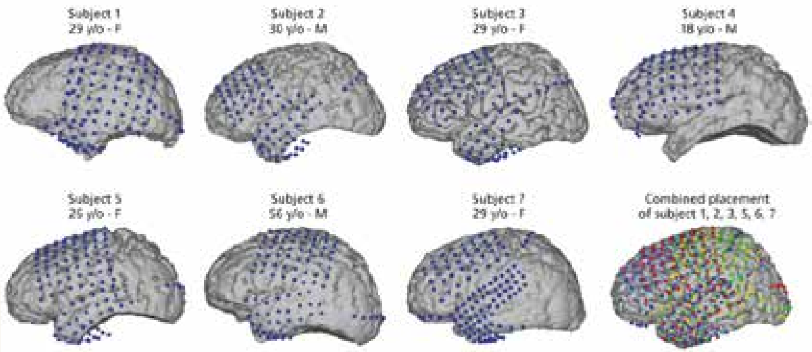


ElectroCorticoGraphy (ECoG)

- ECoG: captures electrical activity of the brain (like EEG) with high temporal resolution (like EEG) but also high spatial resolution (unlike EEG)
  - records directly on the brain surface

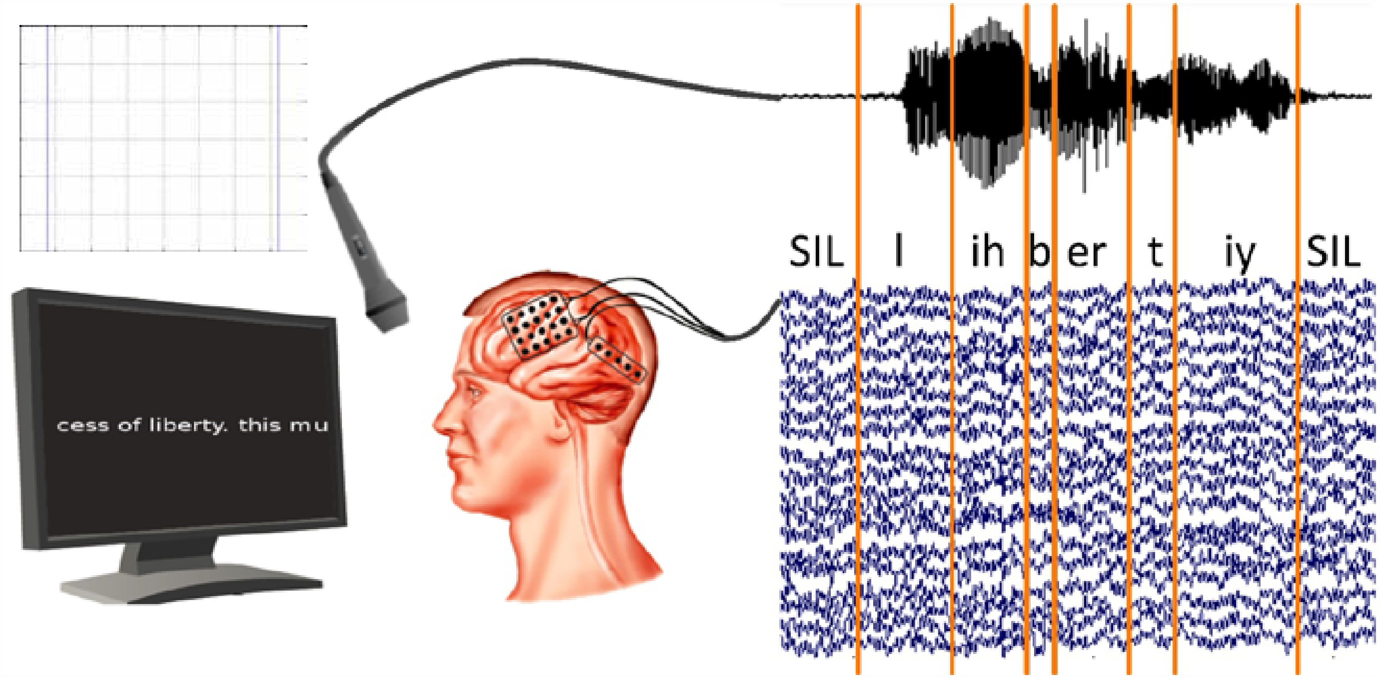


- 7 subjects with intractable epilepsy, Albany Medical Center (NY, US)
- Electrode locations only determined by clinical needs
- 1 – 4 sessions with very little data per session (about 5 minutes)
  - Political speeches, Fan-fiction, Children rhymes
  - Between 20 and 48 phrases per session
- Electrode positions were co-registered in common Talairach space



Experiments with Audible Speech

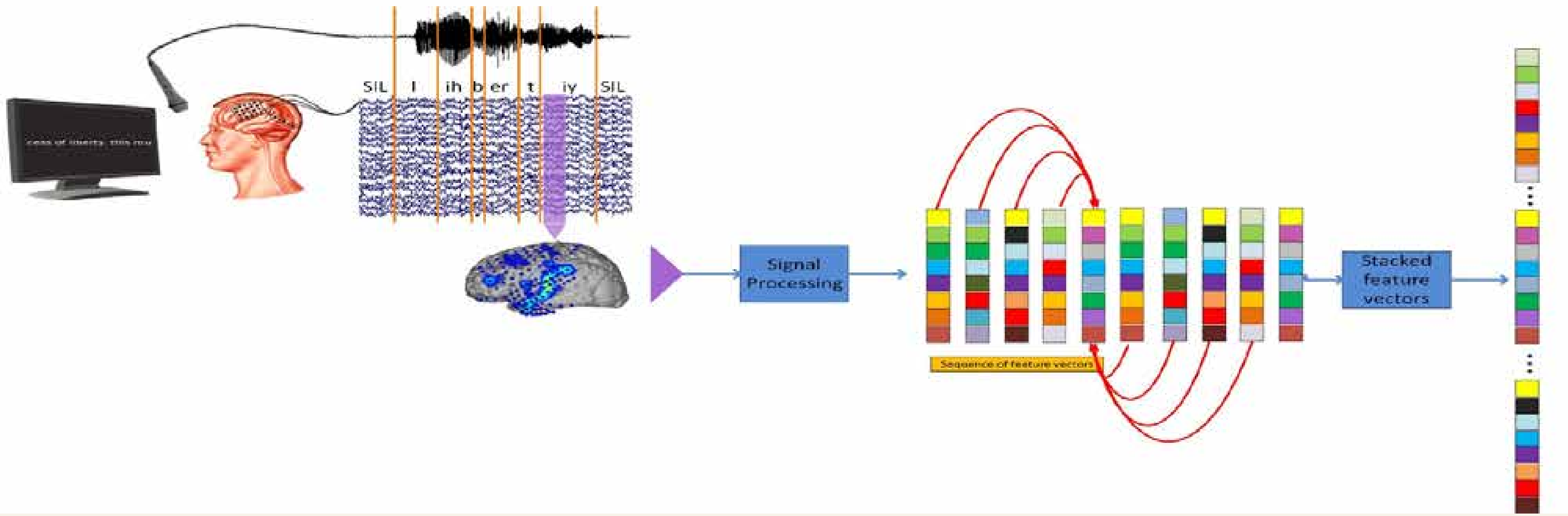
- Participants read aloud scrolling text (Ticker task)
- ECoG & acoustics recorded simultaneously
- Assign phones labels from the acoustic stream (forced alignment, ASR)
- Impose labels on neural data; model phones solely from neural data



C Herff, D Heger, A de Pesters, D Telaar, P Brunner, G Schalk and Tanja Schultz Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci., 12 June 2015 | http://dx.doi.org/10.3389/fnins.2015.00217

Feature Extraction

- 58 – 120 electrodes
- Linear detrending, CAR filtering (re-reference channels to common average)
- Elliptic IIR notch filter (118-122, order 13) attenuates first harmonic of 60 Hz line noise
- For each channel *c* extract logarithmic power of Broadband gamma (70-170 Hz) in 50 ms intervals *i*, 25 ms overlap: $E_{i,c} = \log(\frac{1}{n}\sum_{t=1}^{n} x_{i,c}(t)^2)$
- Assign each interval/frame the corresponding label from the acoustic stream
- **Context:** stack with ± 4 neighboring frames (up to 200 ms prior and after)

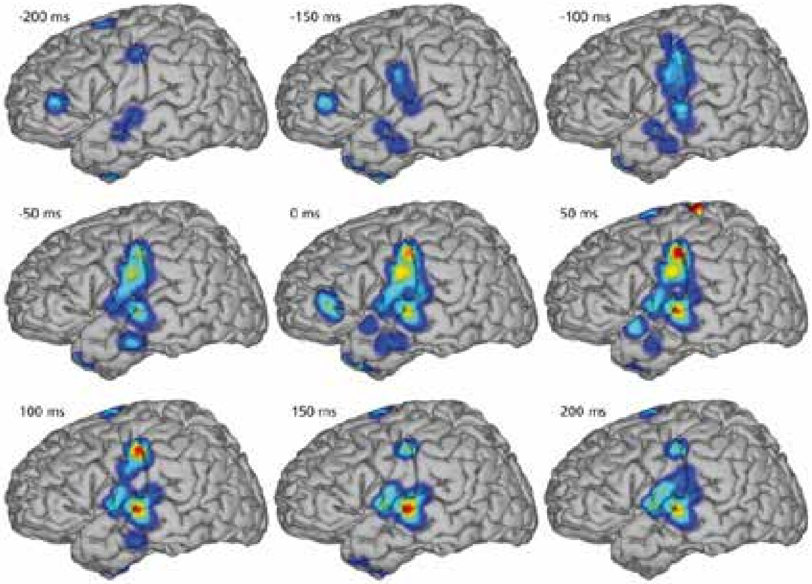


Feature Selection

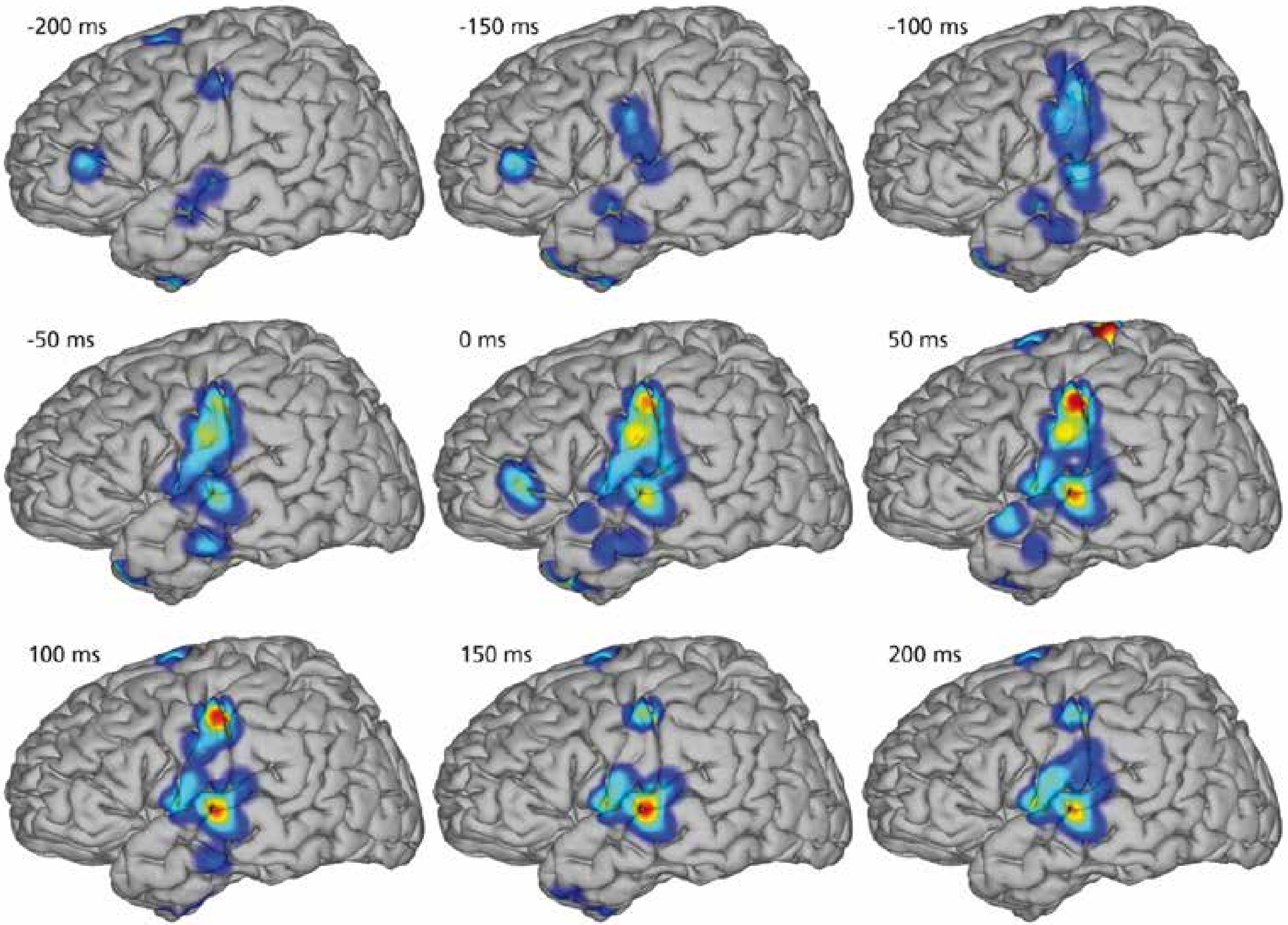
- Large Feature Space: Over 900 dimensions
- Use discriminability as criterion for feature selection
- Mean Kullback-Leibler divergence (KLDiv) for each feature

$$D_{KL}(N_0||N_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - d - log_2(\frac{det(\Sigma_0)}{det(\Sigma_1)})$$

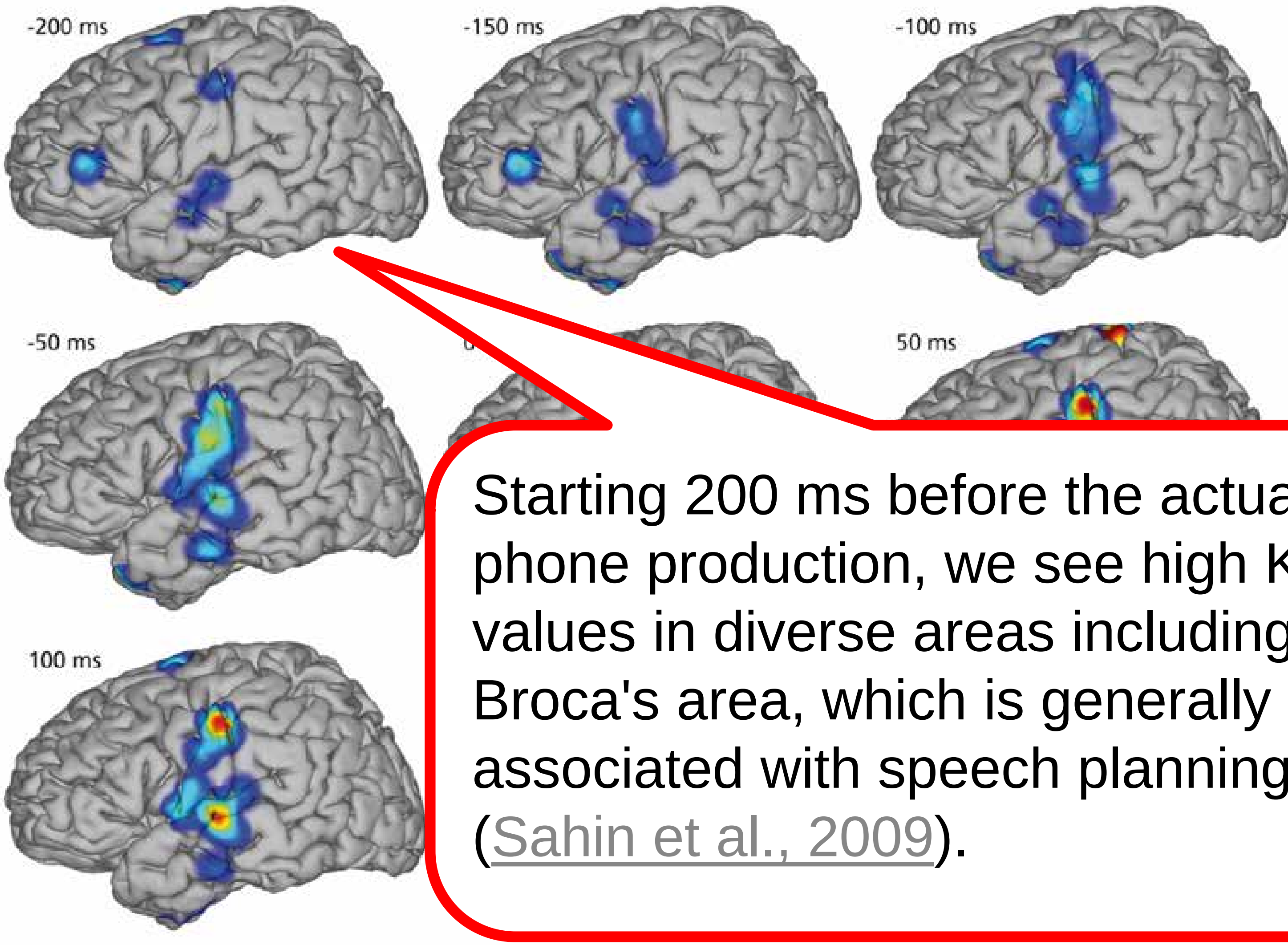
- Calculate KLDiv between each pair of phones à Mean is mean Discriminability for the current phone at location and time offset
- Plotting the mean KLDiv allows interpretation of relevant areas and time offsets



Brain Activity while producing speech

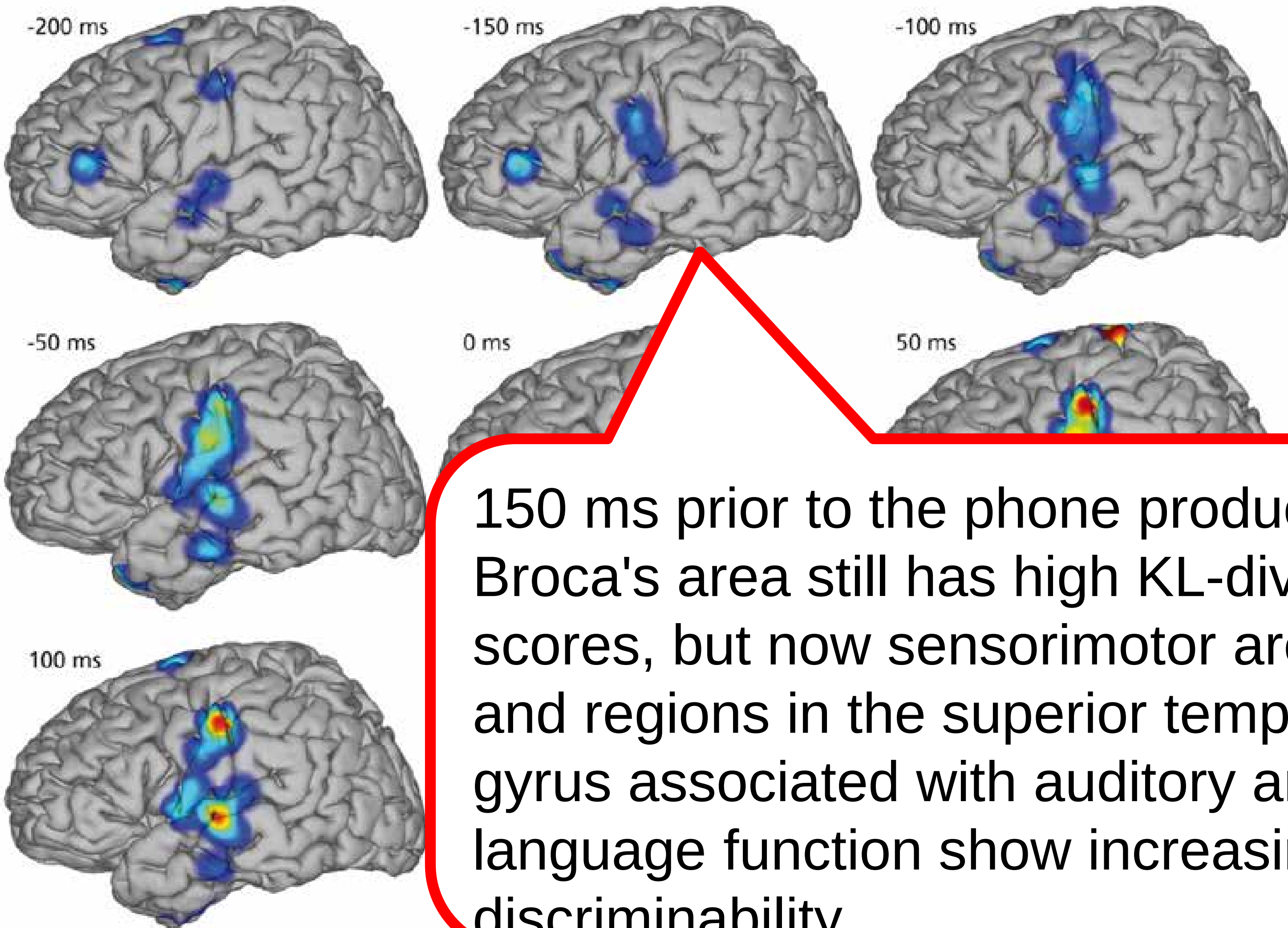


Brain Activity while producing speech



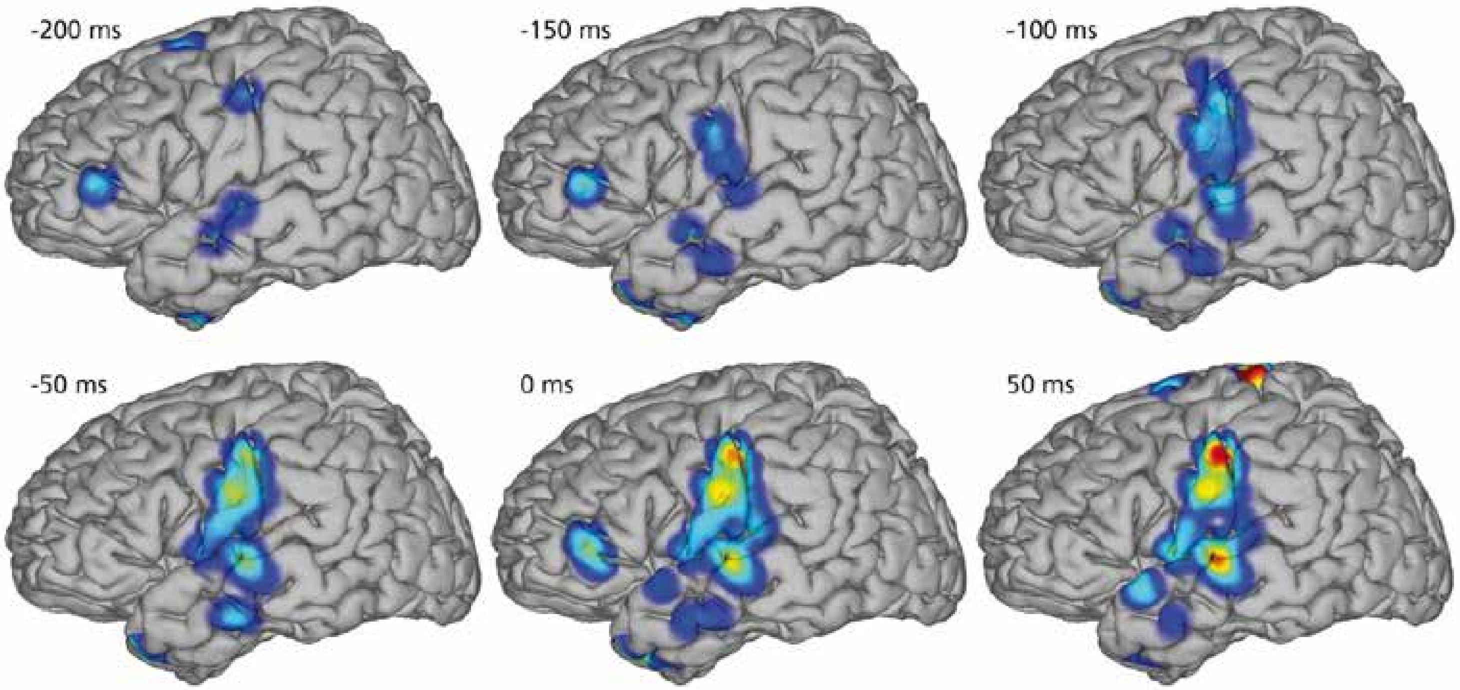
Starting 200 ms before the actual phone production, we see high KL-div values in diverse areas including Broca's area, which is generally associated with speech planning (Sahin et al., 2009).

Brain Activity while producing speech



150 ms prior to the phone production, Broca's area still has high KL-div scores, but now sensorimotor areas and regions in the superior temporal gyrus associated with auditory and language function show increasing discriminability.

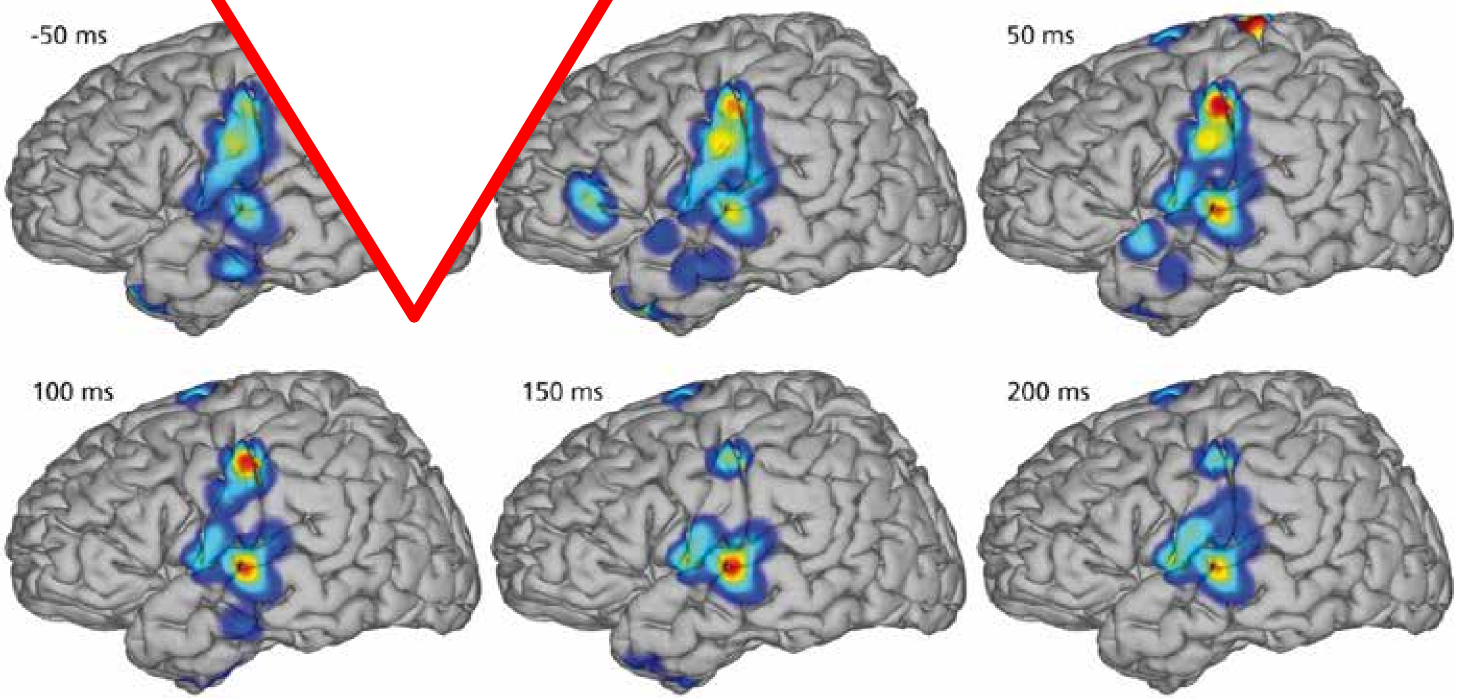
Brain Activity while producing speech



Subsequently, activations in Broca's area vanish and motor area discriminability increases until peaking at the interval between 0 and 50 ms (which corresponds to the average length of phones).

Brain Activity while producing speech

Discriminability increases in auditory regions until approximately 150 ms after phone production.

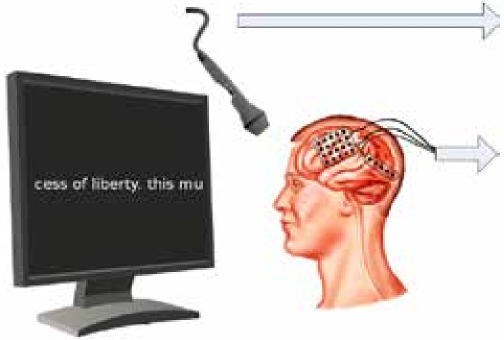


Brain-to-Text (Speech Recognition)

Speech Signal Capturing

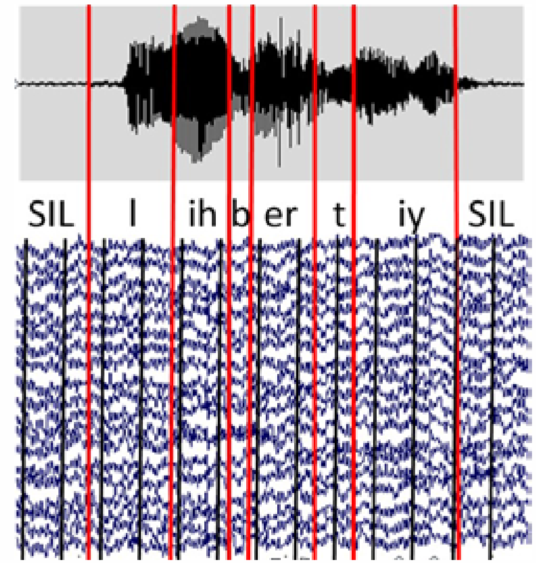
EMG+EEG

fNIRS



Electrocorticography (ECoG)

Signal Processing

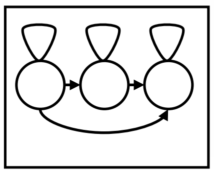


Automatic Speech Recogntion

Text Output

"Hello"

Acoustic

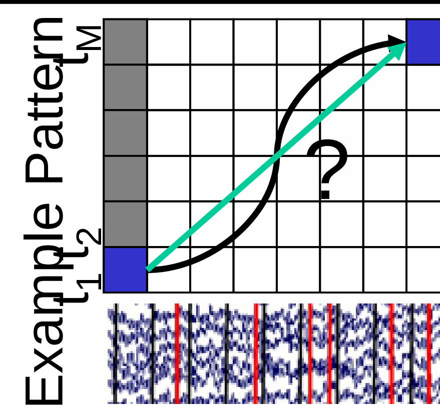


Dictionary

I	/i/
you	/j/ /u/
we	/v/ /e/

Language Model

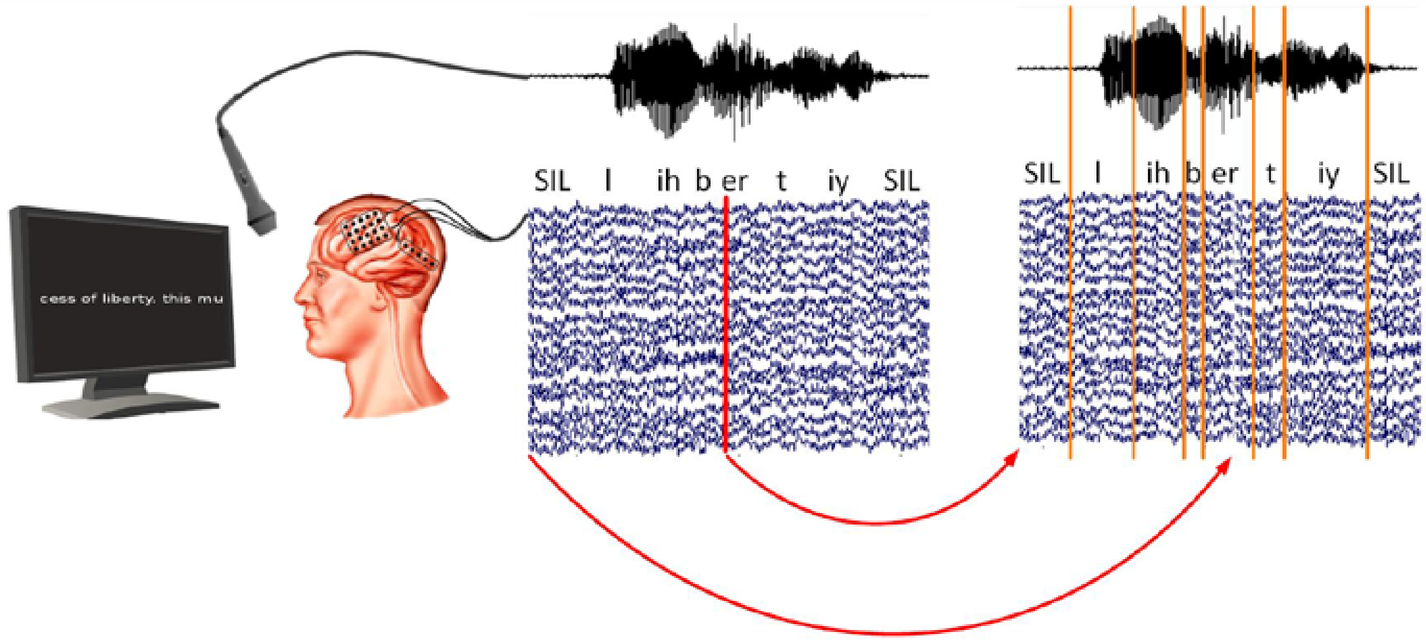
I am
You are
We are



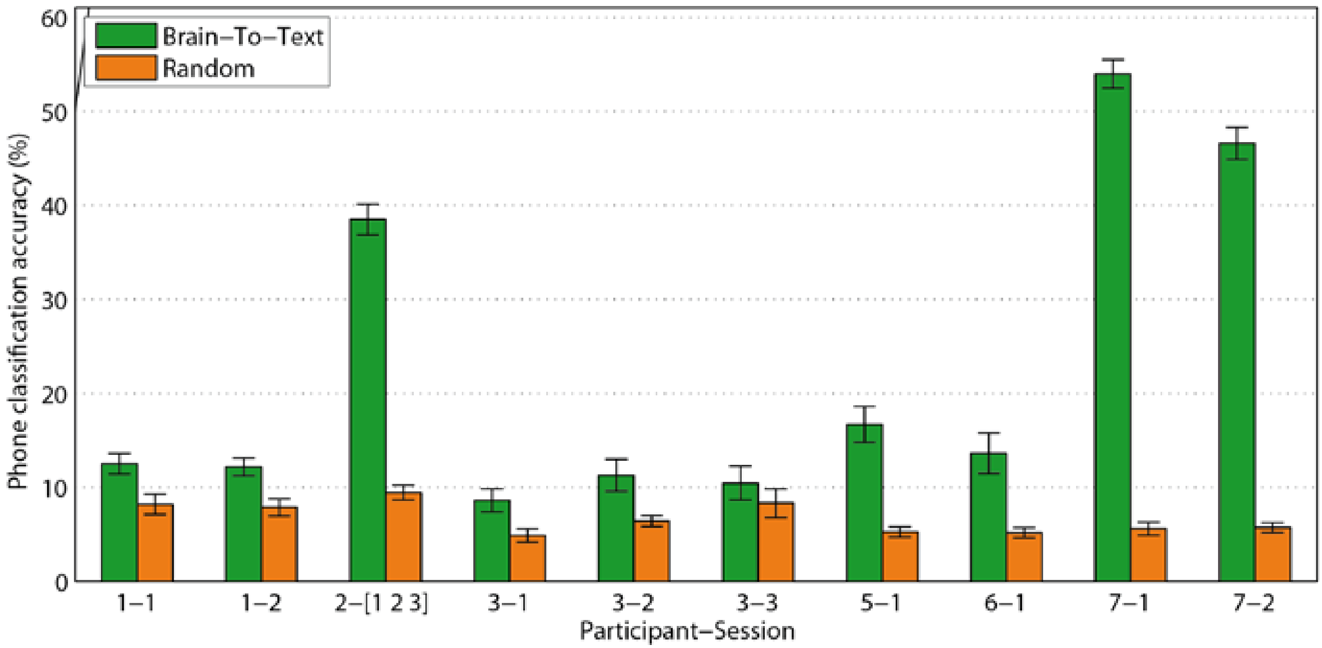
$$\arg\max_{W} P(W \mid X) = \arg\max_{W} P(W) \times p(X \mid W) \times \frac{1}{P(x)}$$

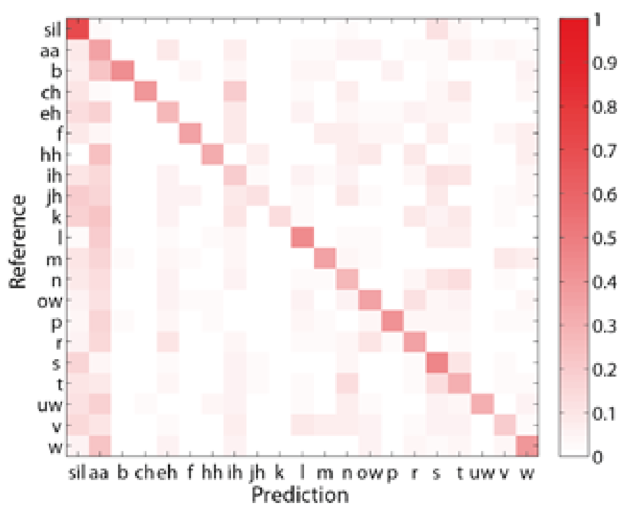
Experimental Results: Randomization Testing

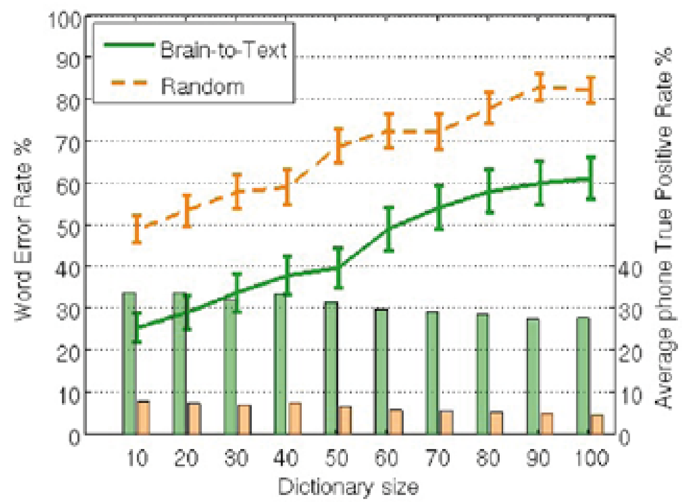
- Shift ECoG data by half of the session, keep labels
- Typical ECoG data, but does not match labels anymore
- Train a full system “Random” for comparison with Brain-to-Text
- All evaluation done in a leave-one-phrase-out cross-validation



Brain-to-Text: Experimental Results

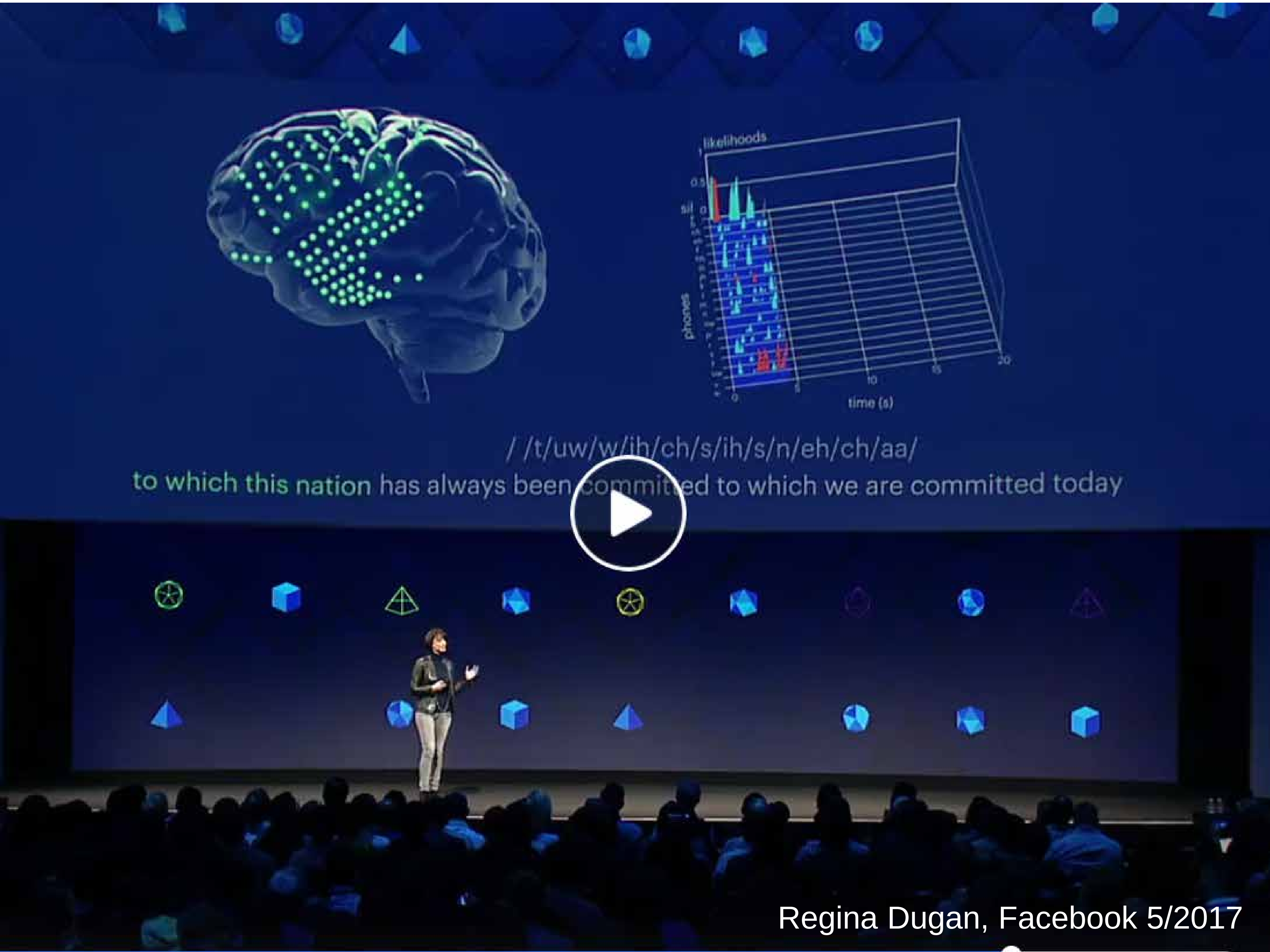






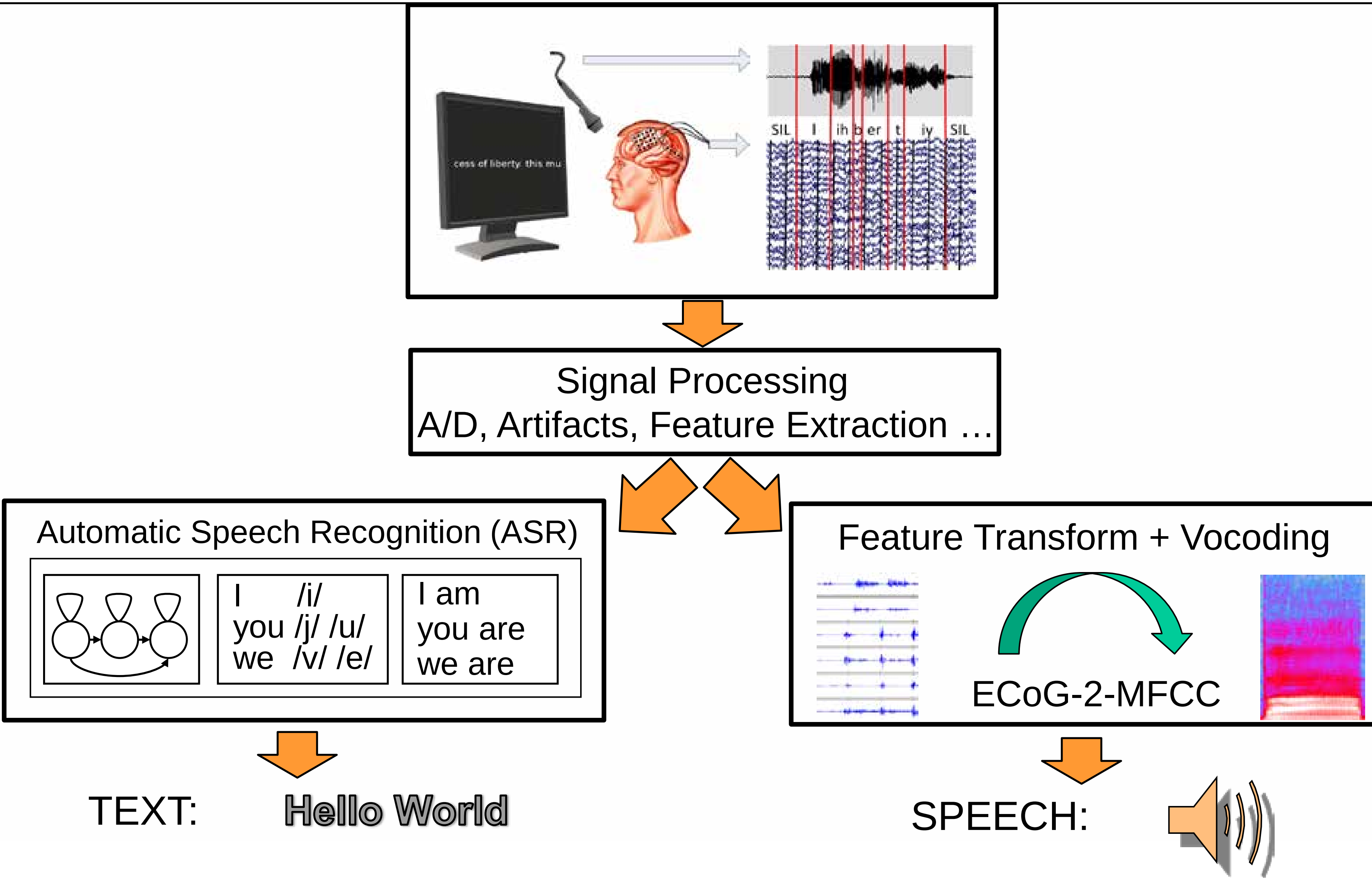
Brain-to-text:
Decoding spoken sentences from phone representations in the brain

Karlsruhe Institute of Technology
Cognitive Systems Lab
SCHALK LAB



Regina Dugan, Facebook 5/2017

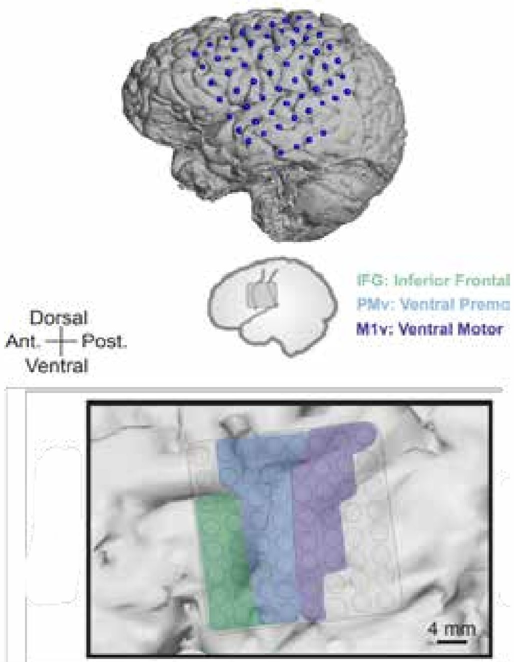
Two Methods: ASR versus Direct Synthesis



Brain-to-Text **Brain-to-Speech**

Towards Brain-to-Speech

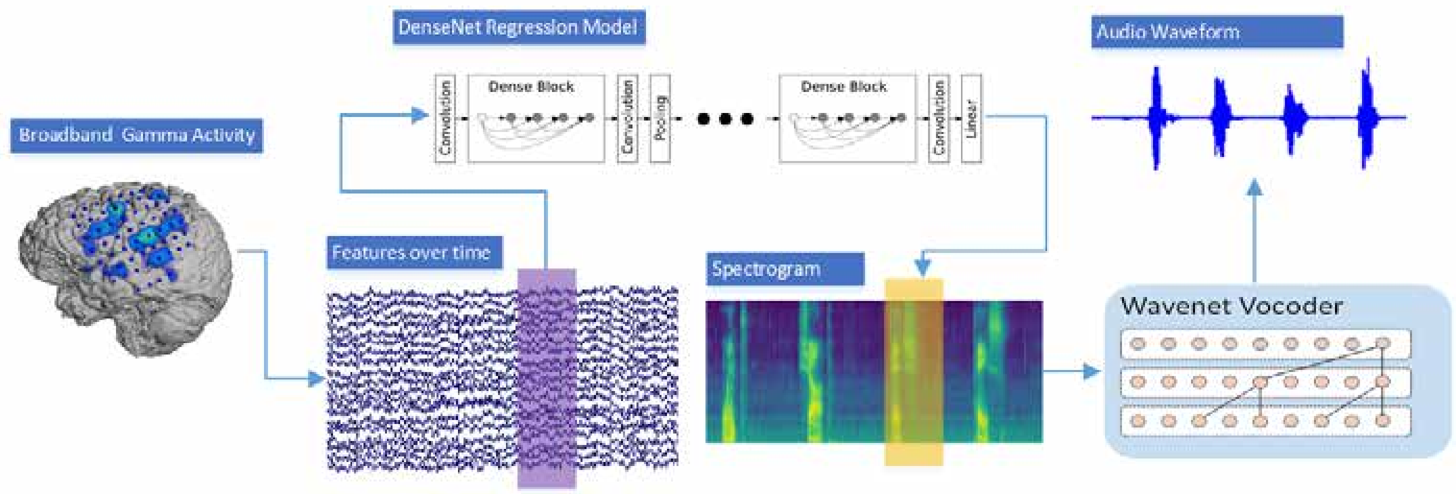
- Direct synthesis of speech from neural activity
  - Current BCI do not convey acoustic cues like **stress, intonation, …**
  - Instant feedback allows for human-in-the-loop concept: Co-adaptation
- RESPONSE Project (with D. Krusienski ODU and J. Shih UCSD)
  - **Re**vealing **SPON**taneous Speech processes in **E**lectrocorticography
  - CR Computational Neuroscience, NSF and BMBF (2017-2020)



- **UCSD** Data (Shih et al.)
  - 6 Epilepsy patients implanted with ECoG grids, strips or depth electrodes for surgical mapping
  - Spontaneous and 50 Harvard Sentences in 3 modes: audible, silent, imagined
- **Northwestern** Data (Slutzky et al.)
  - 6 patients undergoing glioma removal
  - 8x8 electrode high-density ECoG grids placed on IFG, M1v and PMv
  - Audible repetition of >280 words

Two Synthesis Approaches

(1) High-quality Speech Output with Dual Neural Network Approach:
- Densely Connected Convolutional NN maps ECoG to spectral features
- Wavenet transforms spectral features to speech waveform



(2) Fast and straight-forward codebook-based Unit Selection Approach:

Herff, Johnson, Diener, Shih, Krusienski, Schultz: Towards direct speech synthesis from ECoG: A Pilot Study, EMBC 2016
Herff, Diener, Mugler, Slutzky, Krusienski, Schultz: Brain-To-Speech: Direct Synthesis of Speech from Intracranial Brain Activity Associated with Speech Production, BCI 2018

Brain-To-Speech: Direct Synthesis of Speech from Intracranial Brain Activity Associated with Speech Production

Christian Herff, Lorenz Diener, Emily Mugler,
Marc Slutzky, Dean Krusienski, Tanja Schultz

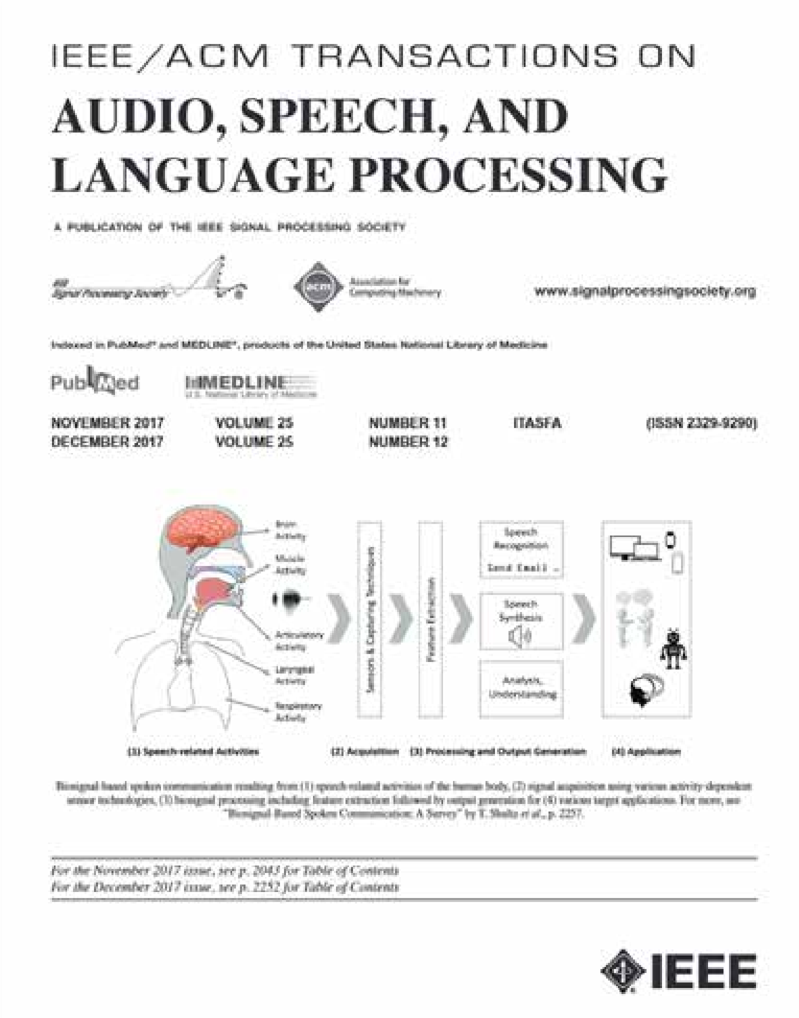
Biosignal-based Spoken Communication

Special Issue **T-ASL, Dec 2017, Vol 25, Number 12**

Editors: Tanja Schultz, Thomas Hueber,
Dean J Krusienski, Jonathan Brumberg

In total 13 papers covering the field, including survey
"Biosignal-based Spoken Communication: A Survey"

Use Cases (Section V)	Speaking Modes (Section II, Table 1-3)				
	modal	murmer	whisper	silent	imagine
(A) Restore SC			*EMG* *PMA* *IMG* *ECoG*	EMG PMA *IMG* ECoG	- - - ECoG
(B) Therapy & Training	EMA EPG IMG *intraoral*	*EMA* *EPG* *IMG* *intraoral*	*EMA* *EPG* *IMG* -	EMA *EPG* *IMG* -	- - - -
(C) Robust SC	EMG *EPG* *PMA* *IMG* intraoral	EMG *EPG* *PMA* *IMG* intraoral	EMG *EPG* *PMA* *IMG* -		
(D) Mute SC		NAM		EMG EMA *PMA* IMG EEG ECoG	- - - - *EEG* ECoG
Insights in SC	All biosignals captured by described technologies including fMRI, fNIRS, MEG, and their combination				



CSL-Team and Collaborators

Thank You!

CSL-Team members involved:
Miguel Angrick, Lorenz Diener, Dominic Heger, Christian Herff, Matthias Janke, Szu-Chen Jou, Udhay Nallasamy, Felix Putze, Kristina Schaaff, Christopher Schulte, Dominic Telaar, Michael Wand, Jochen Weiner, Marek Wester, Marlene Zahner.

CSL-Visitors:
Hugo Gamboa (Nove Lissbon, Plux Inc.), Keigo Nakamura (NAIST), Kishore Prahallad (IIT Hyderabad), Arthur Toth (CMU), S. Umesh (IIT Madras), Micheal Wand (since 2015 with IDSIA)

CSL-Collaborators:
Alan W Black, Carnegie Mellon University, Pittsburgh, PA,
Andrea Bottin, OT Bioelettronica s.n.c.
Jonathan Brumberg, University Kansas,
Hugo Da Silva, Plux Inc.,
Cuntai Guan, Haizhou Li, A*Star, Singapore,
Jose A. Gonzalez, University of Malaga, Spain,
Thomas Hueber, CNRS/GIPSA,
Dean J. Krusienski, ODU, Norfolk, VA,
Gerwin Schalk, P. Brunner, Wadsworth Center, Albany, NY,
Jerry Shih, Epilepsy Center, UC San Diego Health, CA,
Marc W Slutzky, Neurologic Surgery, Northwestern University, Chicago, IL

Biosignals-Lab @ CSL