Parlinguistic speech processing – Practical aspects

Dr. Mathew Magimai Doss

Voice anonymization

Voice privacy: Why?

Voice is a personal identifier

"*Recent studies indicate that nearly all children aged 3 to 17 in the UK engage with online content, with smartphones (72%) and tablets (69%) being the most commonly used devices*."

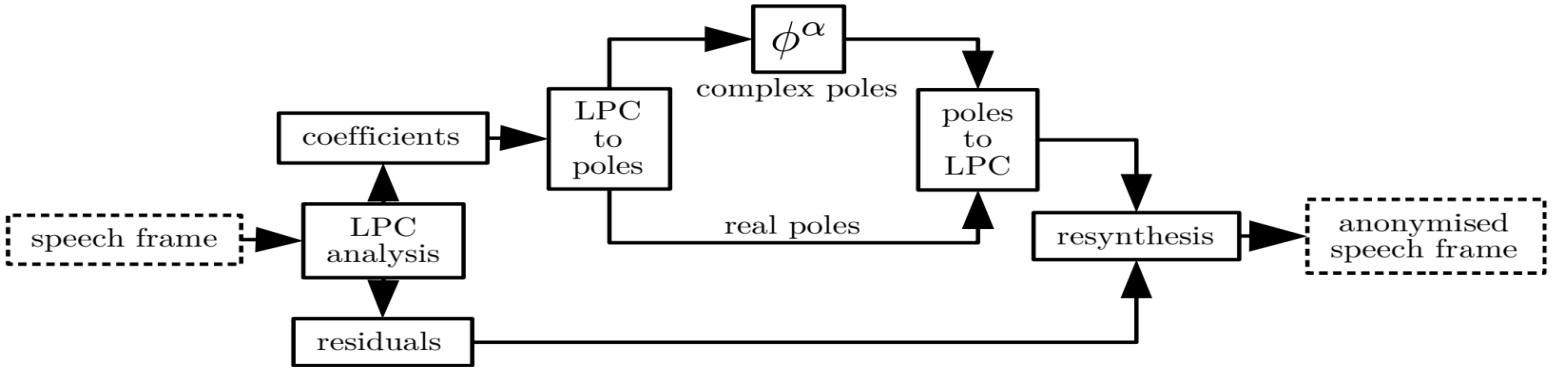
https://www.ofcom.org.uk/media-use-and-attitudes/media-habits-children

"*It took only 20 minutes and 1 US dollar to generate the fake audio of President Biden, discouraging voters to cast their ballots.*"

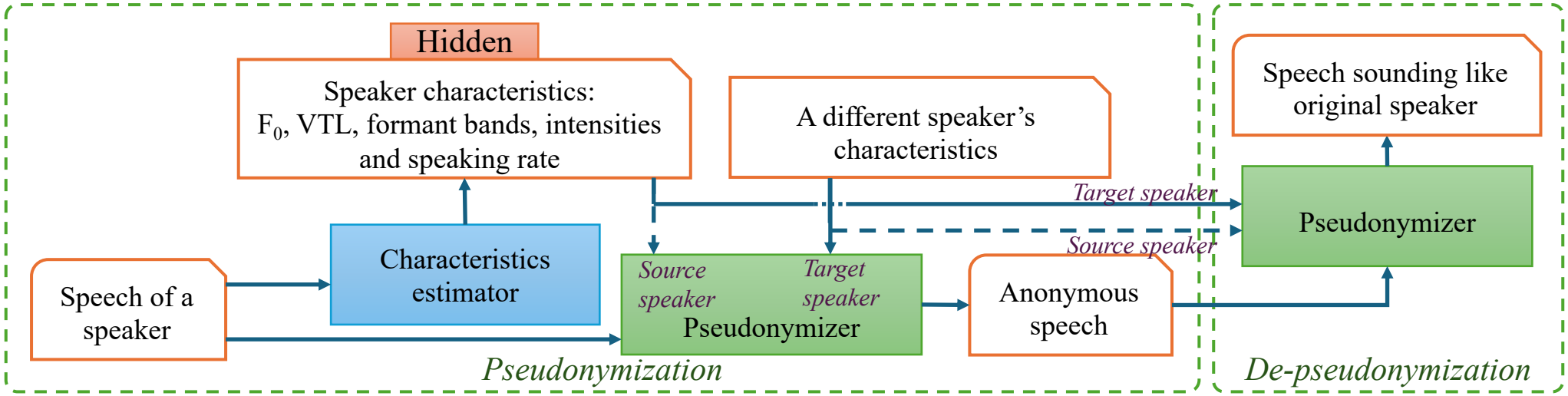
- *Protection from Exploitation and Abuse*
- *Avoiding Behavioral Targeting*
- *Encouraging Safe Exploration*
- *Allowing unbiased, fair use of speech processing applications*

Voice Privacy Challenge (https://www.voiceprivacychallenge.org/)

Signal-processing based approaches

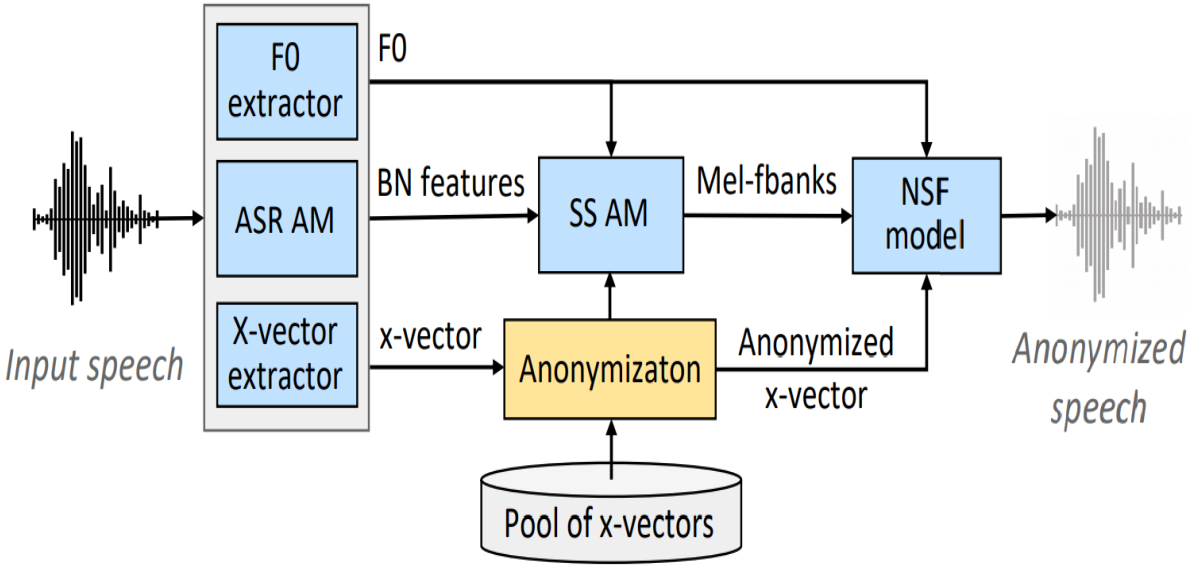


Patino, et al. (2020). Speaker anonymisation using the McAdams coefficient. arXiv preprint arXiv:2011.01130.

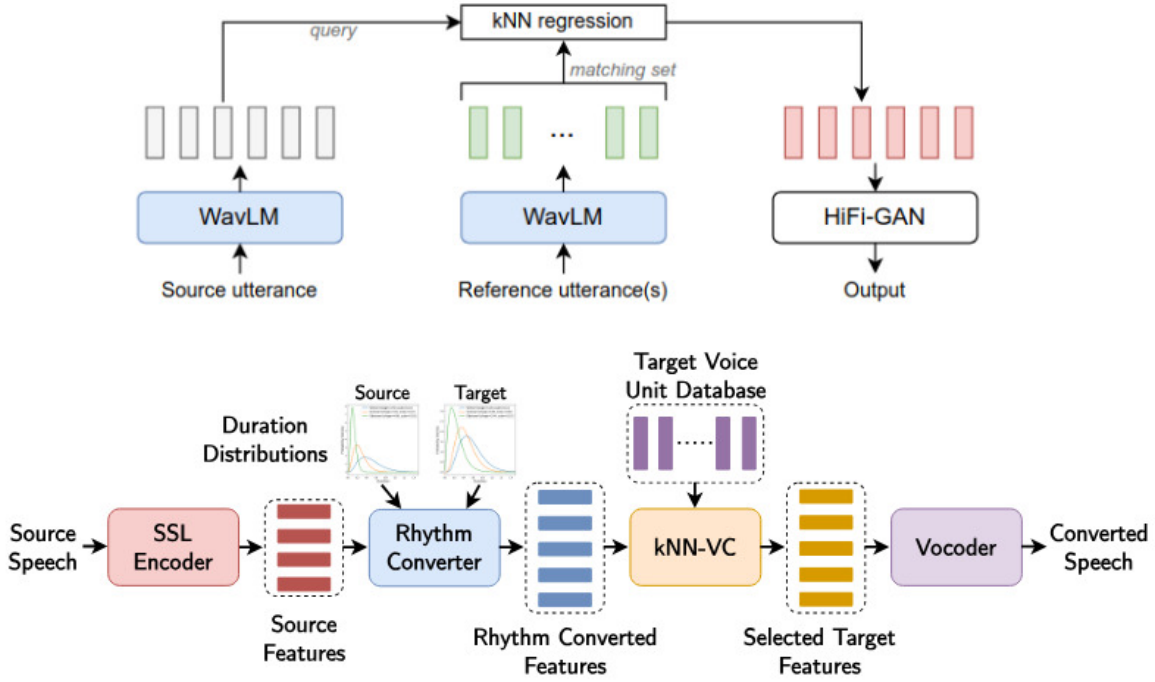


Adjustable Deterministic Pseudonymization of Speech, S. Pavankumar Dubagunta, Rob J. J. H. van Son and Mathew Magimai-Doss, in: Computer, Speech & Language, 72, 2022

Neural-based approaches

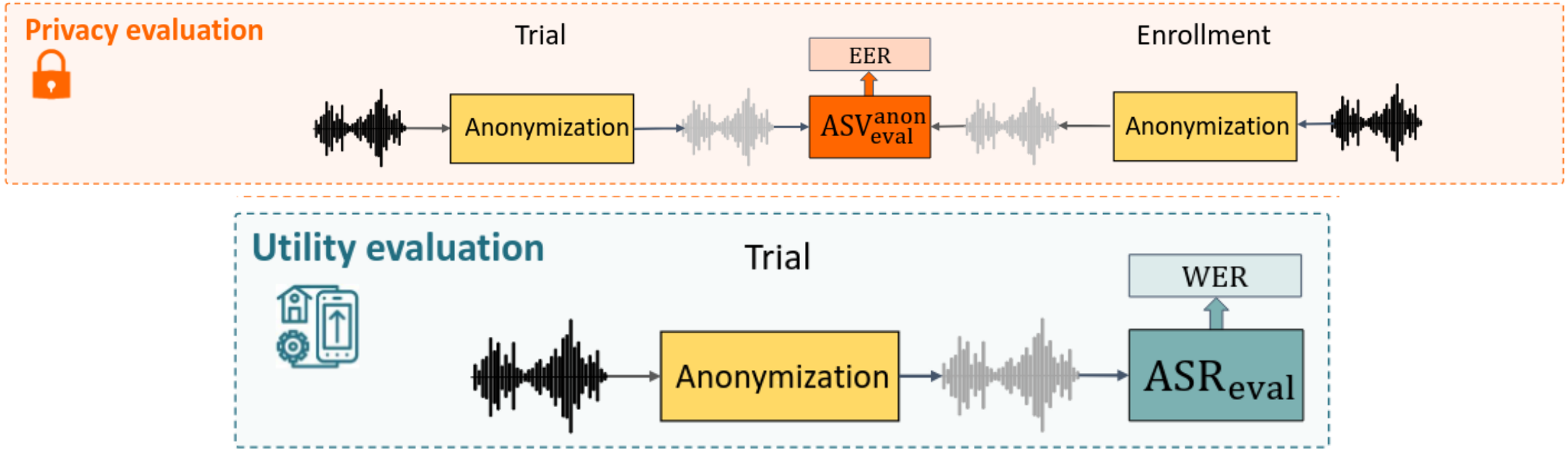


Tomashenko, *et al.* (2020). Introducing the VoicePrivacy initiative. Proc. Interspeech.



kNN-VC

Evaluation



Utility beyond intelligibility?

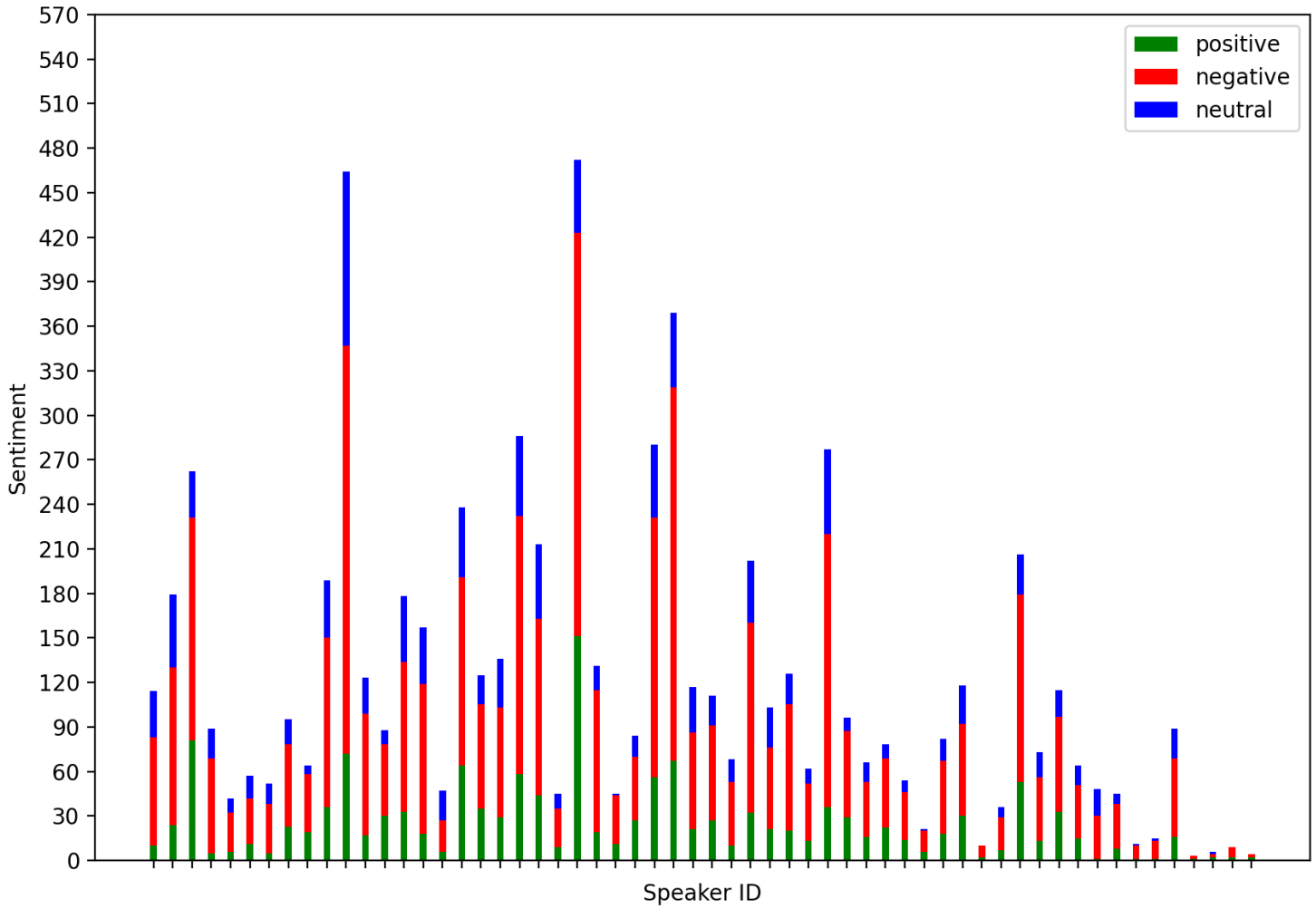
Emotion Modeling in Remote Learning

Dimensional Emotion Modelling of Student Speech in Remote Setting, Sargam Vyas, Bogdan Vlasenko, André Mayoraz, Egon Welen, Per Bergamin, and Mathew Magimai-Doss. (Manuscript under preparation)

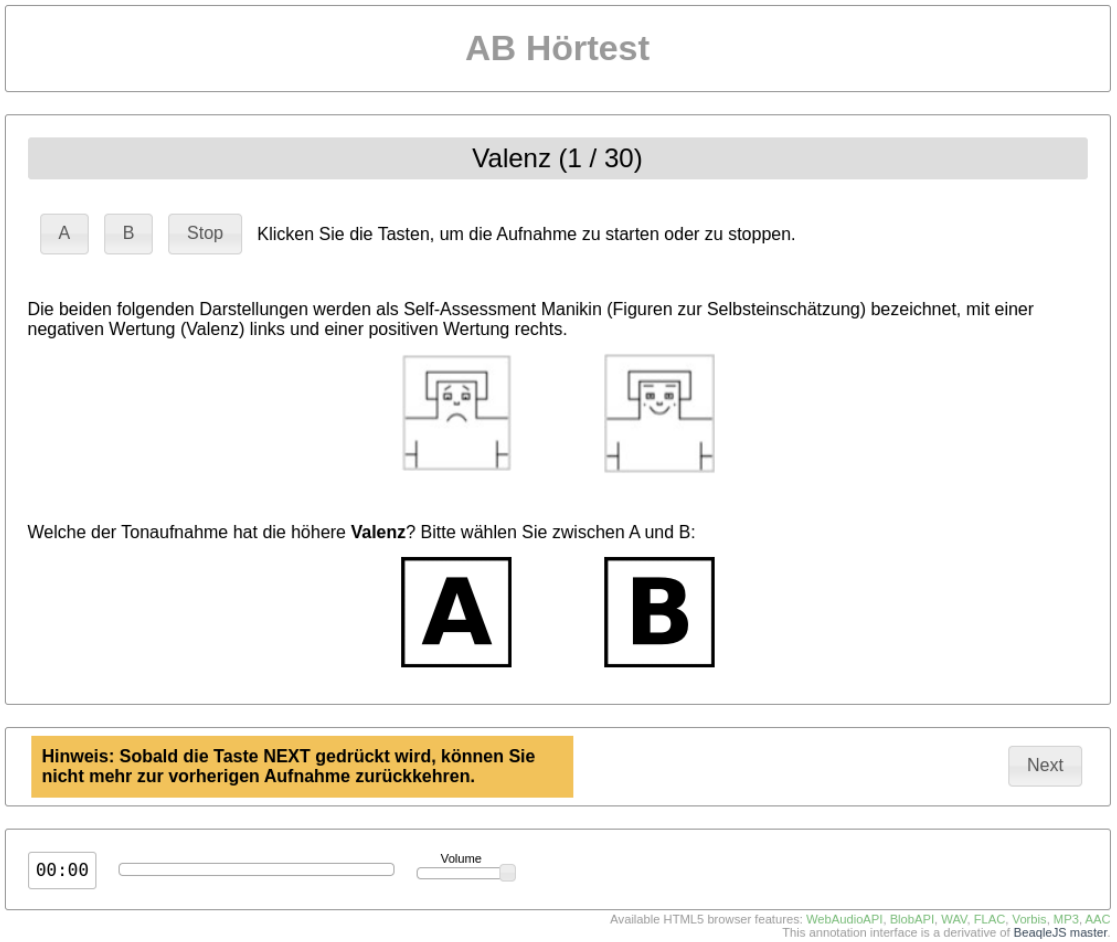
A collaboration between Idiap and FFHS, Brig

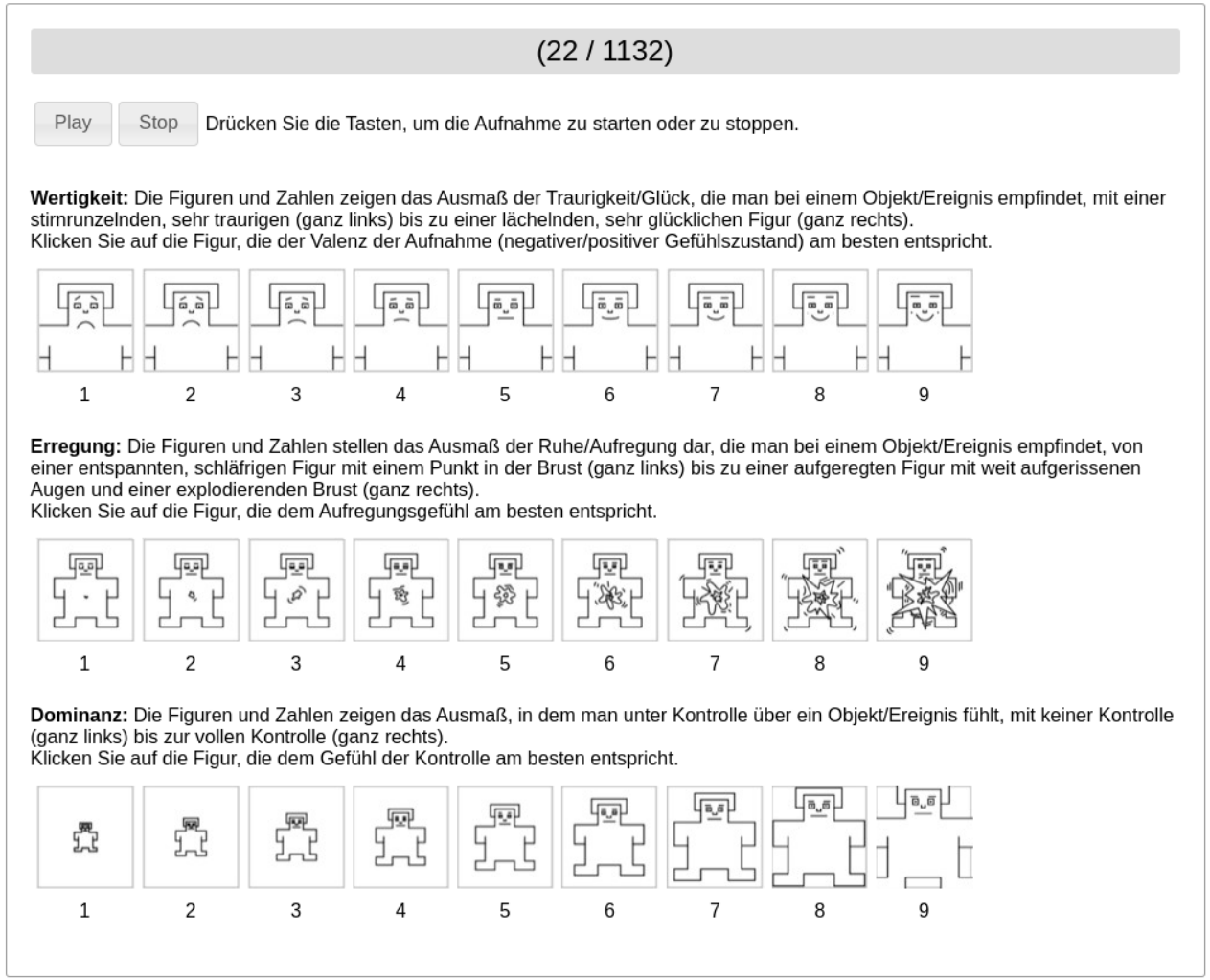
How to elicit emotion?

- Self-control tasks with sequence of steps
  - Information about the task
  - Open question
  - Open answer
  - Level of difficulty assessment
  - Self-evaluation w.r.t sample answer
  - Self-reflection w.r.t sample answer
- 5 out of 11 tasks offered on Moodle with speech input
  - 56 students from “Introduction to Project Management” course
  - Open response and self-reflection recorded
  - Transcribed by off-the-shelf German speech recognition system
  - Students corrected the transcription
- Recording’s chunked into “semantically completed” chunks and automatic sentiment analysis using off-the-shelf German BERT-based system.



Whether emotional variation are perceived? (1)



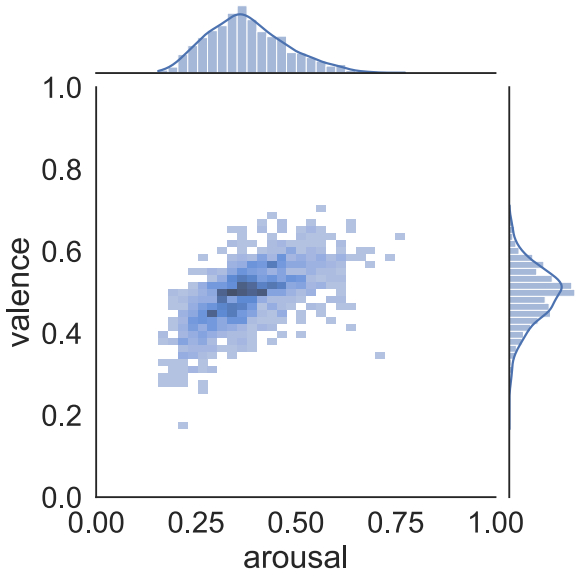


Emotional speech sample pairs from Vera-am-Mittag database were used to train raters and get familiar with SAM.

Whether emotion changes are perceived? (2)

Six raters

ID	VALENCE [%]	AROUSAL [%]	DOMINANCE [%]
1	100	92	67
2	100	100	100
3	42	67	83
4	50	83	100
5	9	100	100
6	100	100	100



Inter-rater agreement

CORPORA	VALENCE	AROUSAL	DOMINANCE
	μ OF CORRELATION COEFFICIENT r		
VAM I [60]	0.49	0.78	0.68
VAM II [60]	0.48	0.66	0.54
SPOT-ED	0.65	0.60	0.67
	$\overline{\sigma}$ OF CORRELATION COEFFICIENT r		
VAM I [60]	0.30	0.38	0.33
VAM II [60]	0.28	0.30	0.29
SPOT-ED	0.12	0.18	0.19

SPOT-ED: developed corpus

Whether emotions can be predicted?

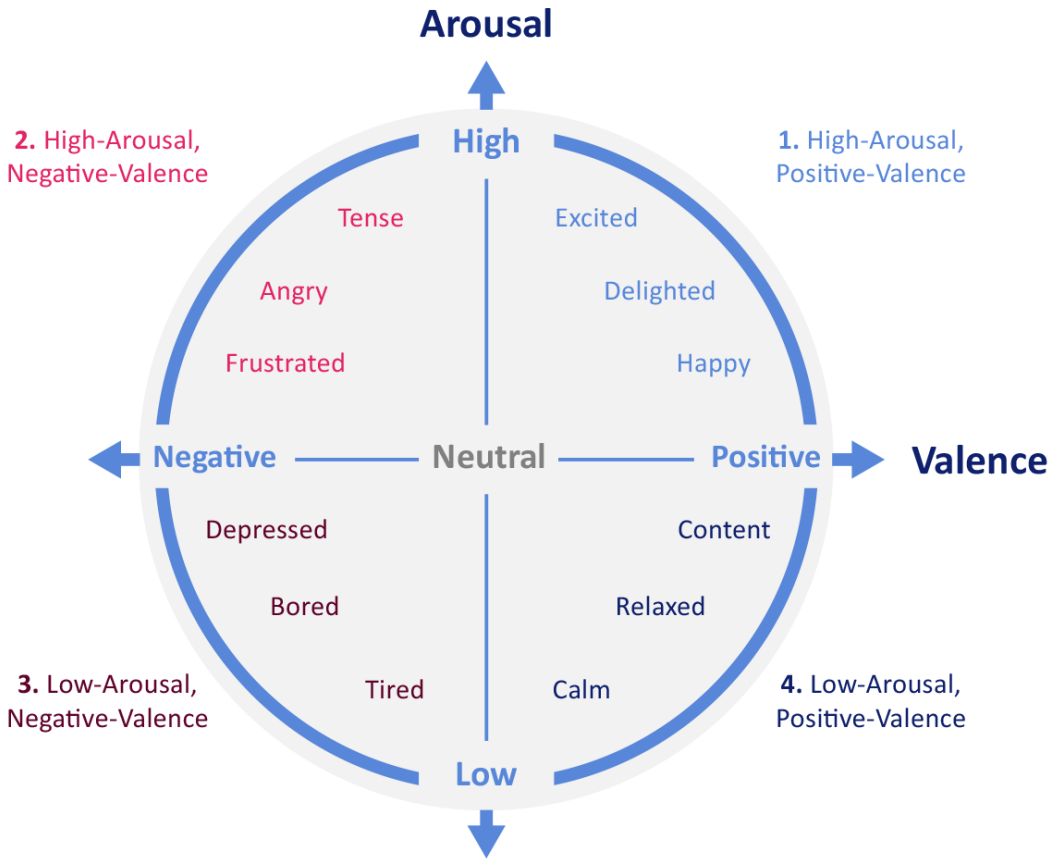
		SPEAKER-INDEPENDENT			SPEAKER-DEPENDENT		
FR	**Dimension**	**Corr$_{spea}$↑**	**Corr$_{pear}$↑**	**RMSE ↓**	**Corr$_{spea}$↑**	**Corr$_{pear}$↑**	**RMSE ↓**
SINGLE FEATURE REPRESENTATIONS							
1. COMPARE	VALENCE	0.429	0.439	0.064	0.489	0.511	0.061
	AROUSAL	0.509	0.531	0.084	0.605	0.630	0.078
	DOMINANCE	0.638	0.640	0.086	0.688	0.687	0.082
2. W2V2-MSP	VALENCE	0.525	0.551	**0.060**	0.555	0.584	0.058
	AROUSAL	0.587	0.611	0.080	0.621	0.657	0.076
	DOMINANCE	0.622	0.640	0.088	0.664	0.682	0.084
3. HUBERT	VALENCE	0.428	0.437	0.062	0.493	0.528	0.060
	AROUSAL	0.503	0.532	0.085	0.607	0.642	0.079
	DOMINANCE	0.598	0.615	0.090	0.669	0.680	0.085
4. WAVLM	VALENCE	0.449	0.447	0.062	0.514	0.535	0.060
	AROUSAL	0.510	0.552	0.084	0.599	0.642	0.078
	DOMINANCE	0.623	0.635	0.088	0.706	0.710	0.082
EARLY FUSION: COMBINED FEATURE REPRESENTATIONS							
1.+2.	VALENCE	**0.536**	**0.562**	**0.060**	**0.585**	**0.613**	**0.057**
COMPARE +	AROUSAL	**0.630**	**0.651**	**0.076**	**0.686**	**0.713**	**0.072**
W2V2-MSP	DOMINANCE	**0.737**	**0.744**	**0.078**	**0.766**	**0.767**	**0.074**
1.+3.	VALENCE	0.476	0.481	0.062	0.537	0.556	0.059
COMPARE +	AROUSAL	0.554	0.578	0.082	0.643	0.667	0.076
HUBERT	DOMINANCE	0.675	0.677	0.084	0.723	0.721	0.079
1.+4.	VALENCE	0.473	0.480	0.062	0.528	0.549	0.060
COMPARE +	AROUSAL	0.548	0.573	0.082	0.64	0.664	0.076
WAVLM	DOMINANCE	0.678	0.679	0.084	0.726	0.723	0.079

Characterization of Embarrassment

Multidisciplinary characterization of embarrassment through behavioral and acoustic modeling, Dajana Šipka, Bogdan Vlasenko, Maria Stein, Thomas Dierks, Mathew Magimai-Doss and Yosuke Morishima, in: Scientific reports, 2025

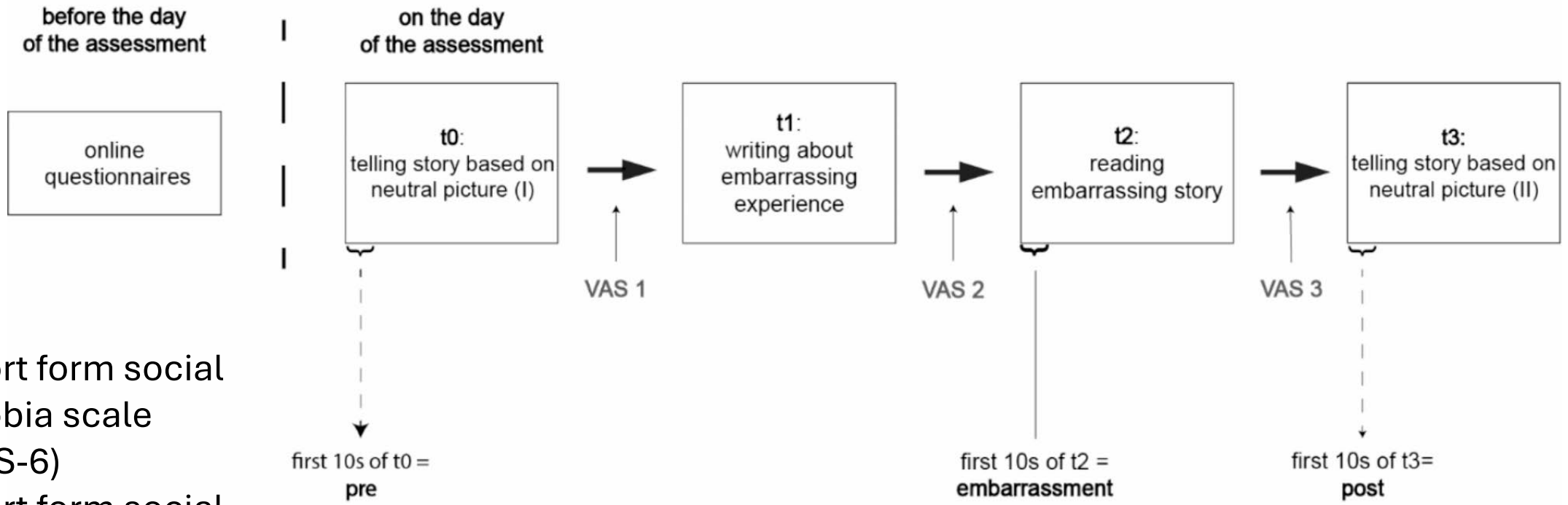
Embarrassment

- Social emotion
  - Self-concious + other-concious emotion
  - Can be experienced publicly (more likely around strangers) as well as in private
  - Personal vs. empathic
  - Previously considered to be part of shame, not anymore
- Typical physiological and behavioral changes
  - Blushing
  - Changes in voice
  - Changes in non-verbal behavior e.g., no eye contact or lowering of head
- Aspects of embarrassment
  - Fear of negative evaluation
  - Fear of rejection
  - Heightened self-conciousness
- Shares many characteristics with social anxiety



<u>Voice-based assessment can be less-biased to subjects' responses and more objective than self-questionnaires.</u>

Study design



- Short form social phobia scale (SPS-6)
- Short form social interaction anxiety scale (SIAS-6)

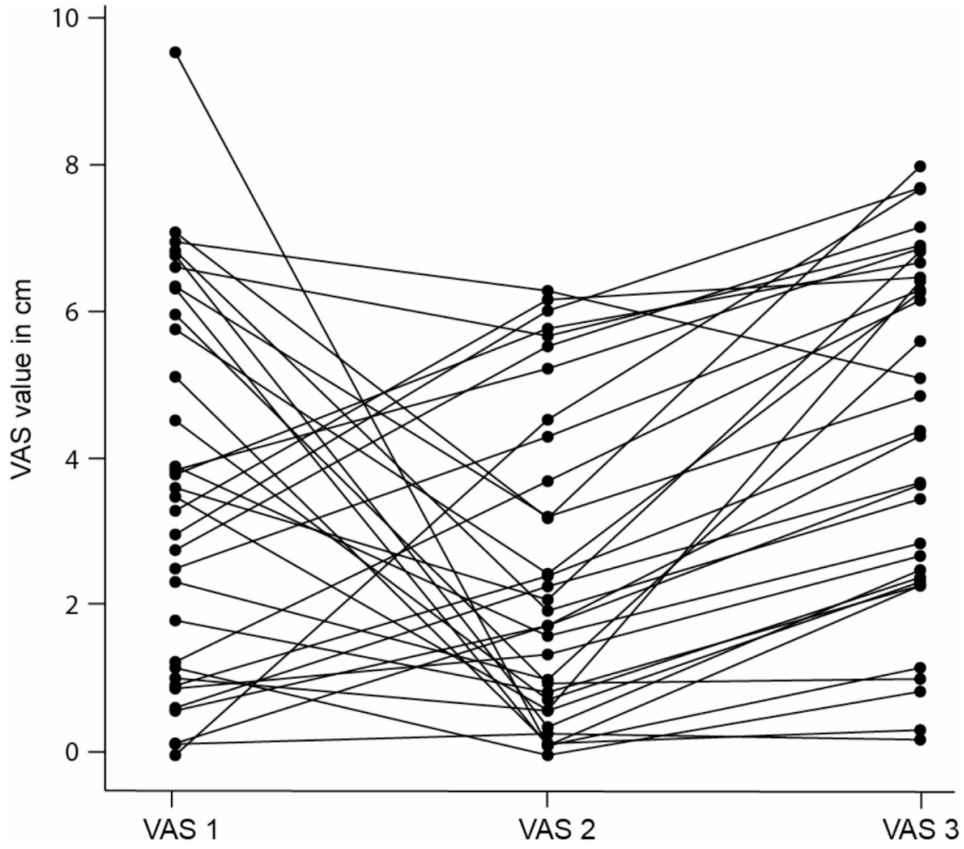
VAS: Visual anxiety scale
Participants mark embarrassment on 10 cm long line (0: not embarrased at all and 10: extremely embarrased)

Thirty three subjects (undergraduate psychology students from Univ. of Bern) selected after applying exclusion criteria

Subjective analysis

Questionnaires	*N*	*M*	*SD*	*Mdn*
SPS-6	33	2.52	3.01	2.00
SIAS-6	33	7.00	4.02	6.00
S (SPS–6 & SIAS–6)	33	4.76	4.18	4.00

16 participants had SIAS-6 score above cut-off 7
17 participants has SPS-6 score above cut-off 2
9 participants (above both SIAS-6 and SPS-6 cut-off)



Significant difference between VAS-2 and VAS-3

31 out of 33 significantly embarrassed after reading the embarrassing story

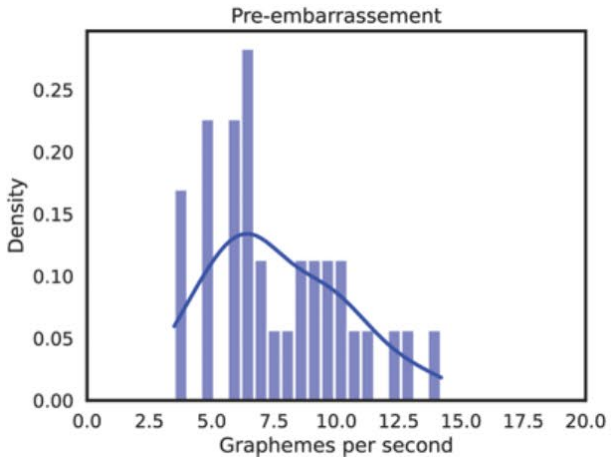
Statistically significant Negative correlation between SPS-6+SIAS-6 score and (VAS-2 –VAS-3)

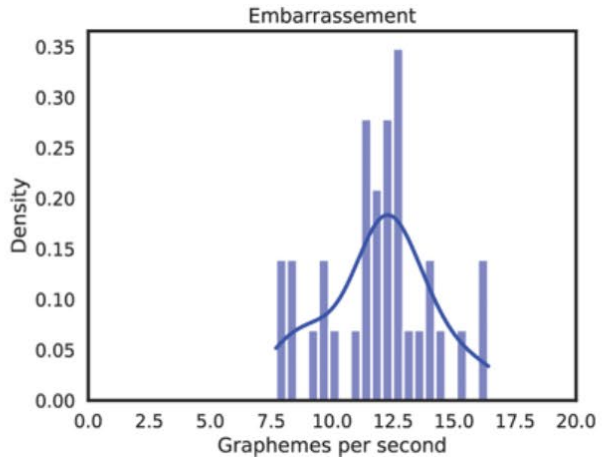
Acoustic classification-based analysis

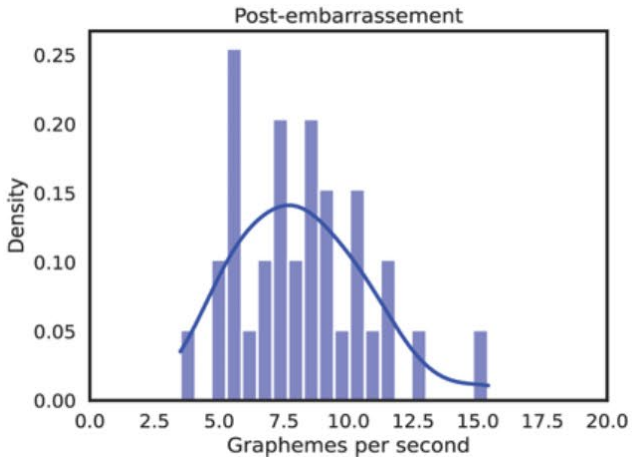
6373 ComParE features

	Pre vs. emb	Post vs.emb	Pre vs. post	(Pre + post) vs. emb	Pre vs. emb vs. post
RF	**0.848**	**0.833**	0.621	**0.803**	0.636
SVM	**0.864**	**0.818**	0.591	**0.818**	0.596

Feature ranking analysis: spectral modulation features (e.g., delta, delta-delta features)





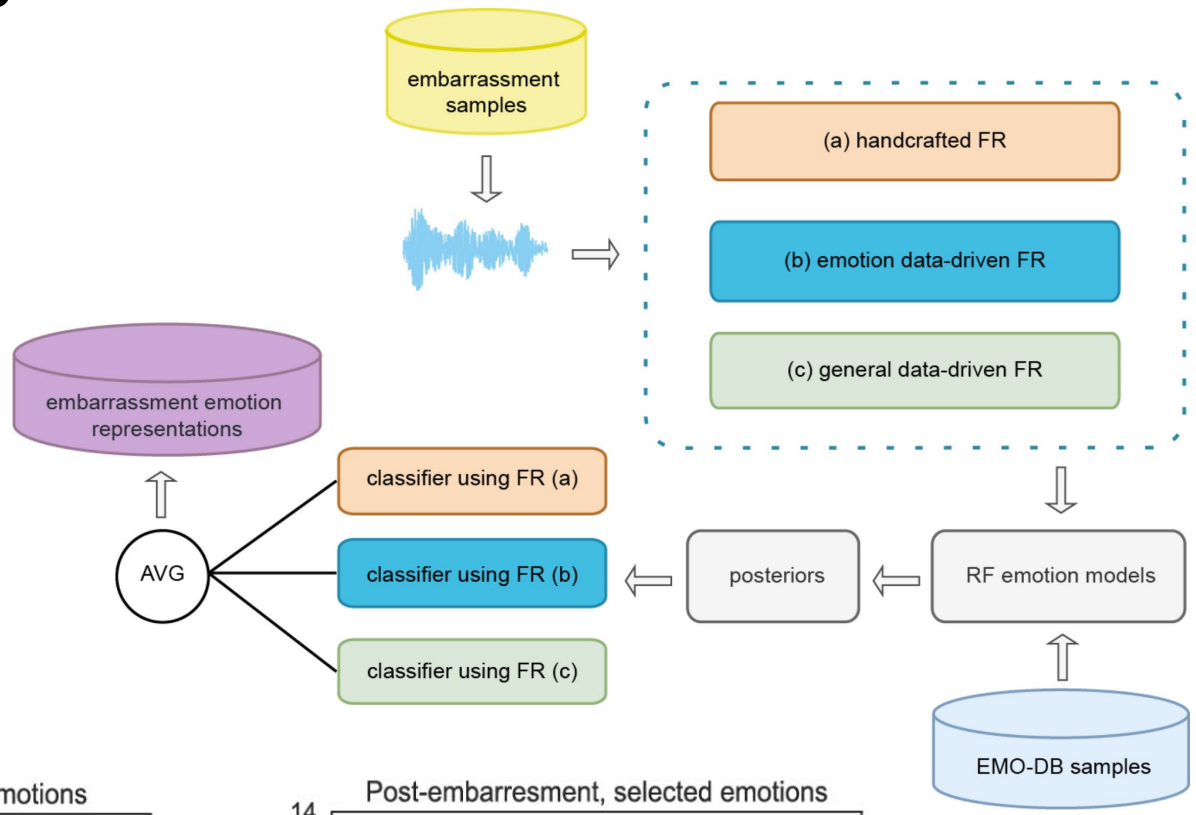


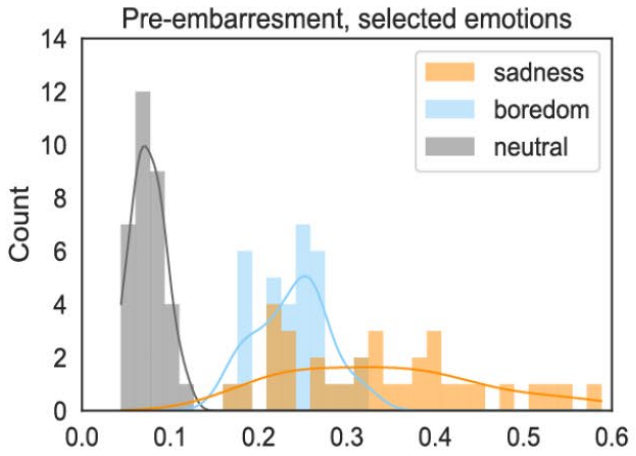
Whisper speech recognition system

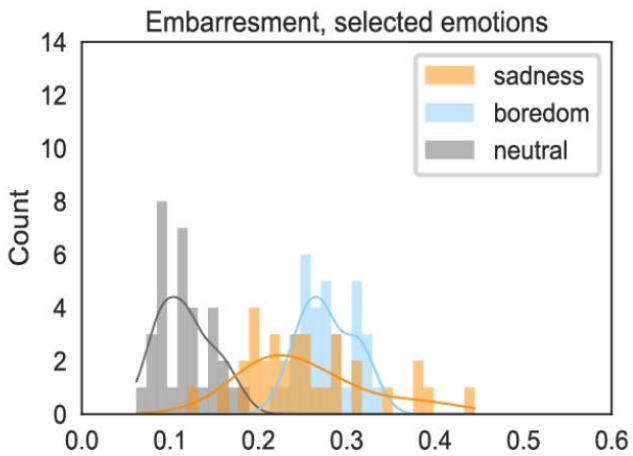
Emb high speaking rate higher than Pre and Post

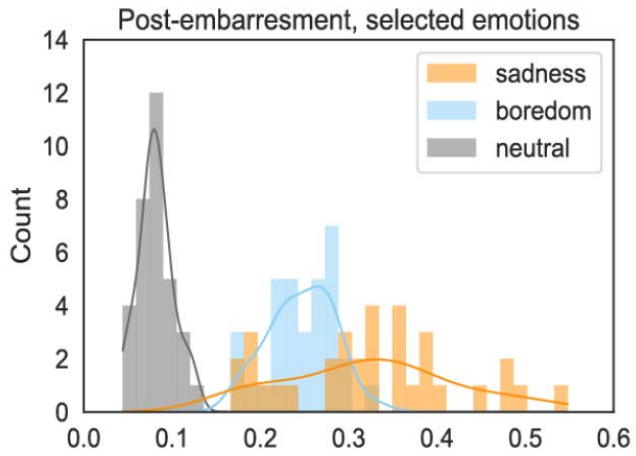
Categorical emotion analysis

Emotion	Mean			*p*-value		
	Pre	Emb.	Post	Pre vs. emb.	post vs. emb.	Pre vs. post
Happiness	0.06	0.06	0.06	0.503	0.652	0.818
Sadness	0.34	0.25	0.33	< 0.001	0.001	0.506
Disgust	0.15	0.15	0.15	0.699	0.397	0.697
Fear	0.08	0.09	0.09	0.079	0.146	0.685
Boredom	0.24	0.28	0.25	< 0.001	< 0.001	0.402
Anger	0.05	0.05	0.05	0.574	0.582	0.317
Neutral	0.08	0.12	0.08	< 0.001	< 0.001	0.200







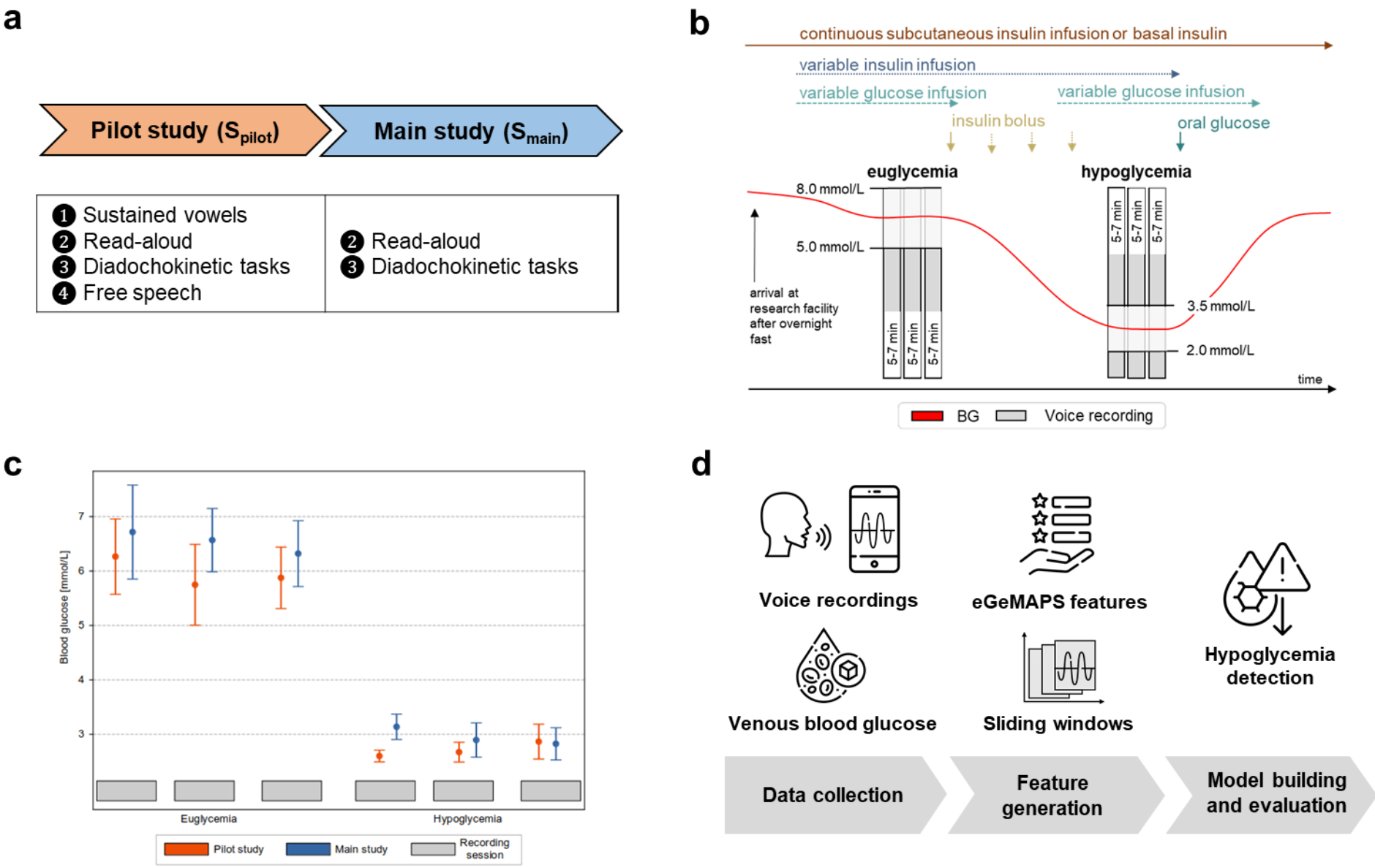


Voice-based hypoglycemia detection

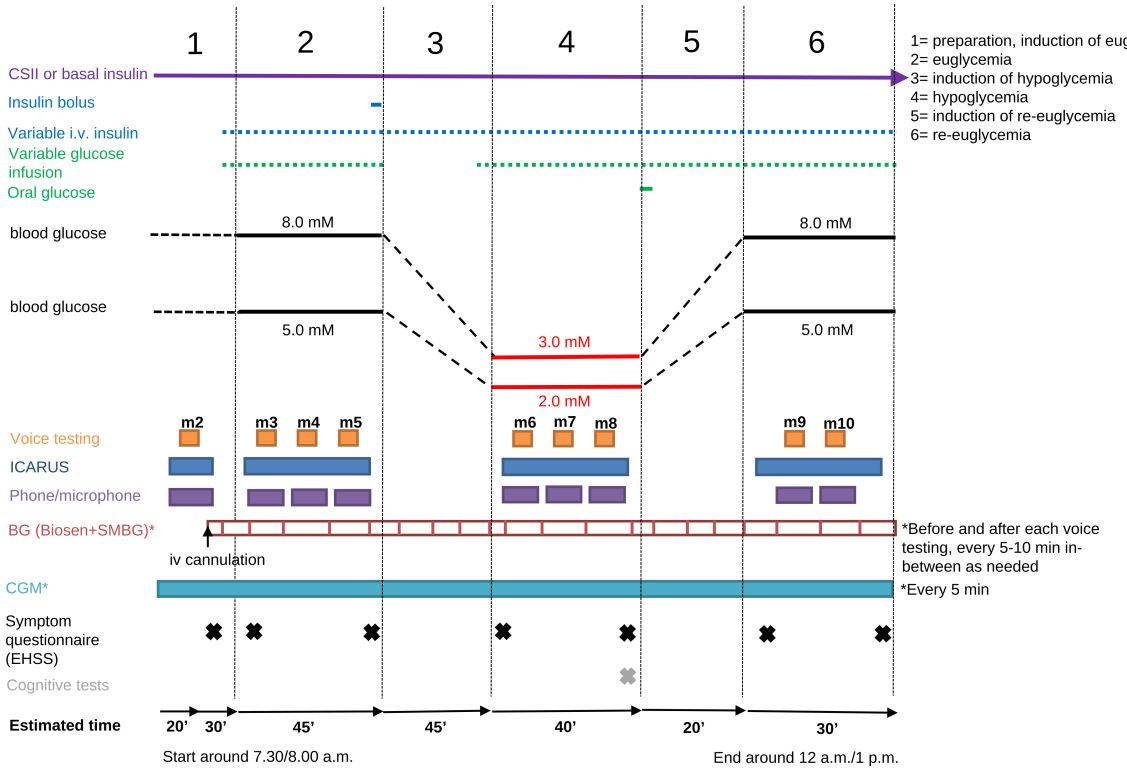
- Hypoglycemia is a hazardous diabetes-related emergncy
- Detected by blood glucose level measurement
- Can voice serve as a tool to detect hypoglycemia in a non-intrusive manner?

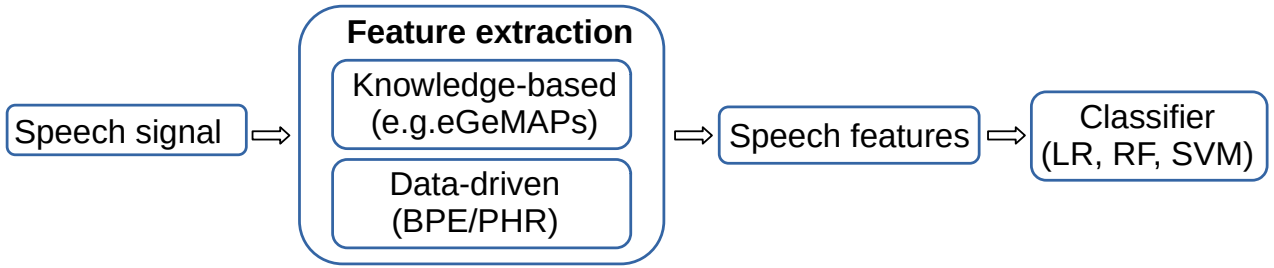
Listening to Hypoglycemia: Voice as a Biomarker for Detection of a Medical Emergency Using Machine Learning, Vera Lehmann, Martin Hilpert, Zohreh Mostaani, Sevada Hovsepyan, Esmé Wallace, Colombine Verzat, Stefan Feuerriegel, Mathias Kraus, James Rosenthal, Gürkan Yilmaz, Mathew Magimai-Doss and Christoph Stettler, in: Diabetes Care, 2025

Study summary

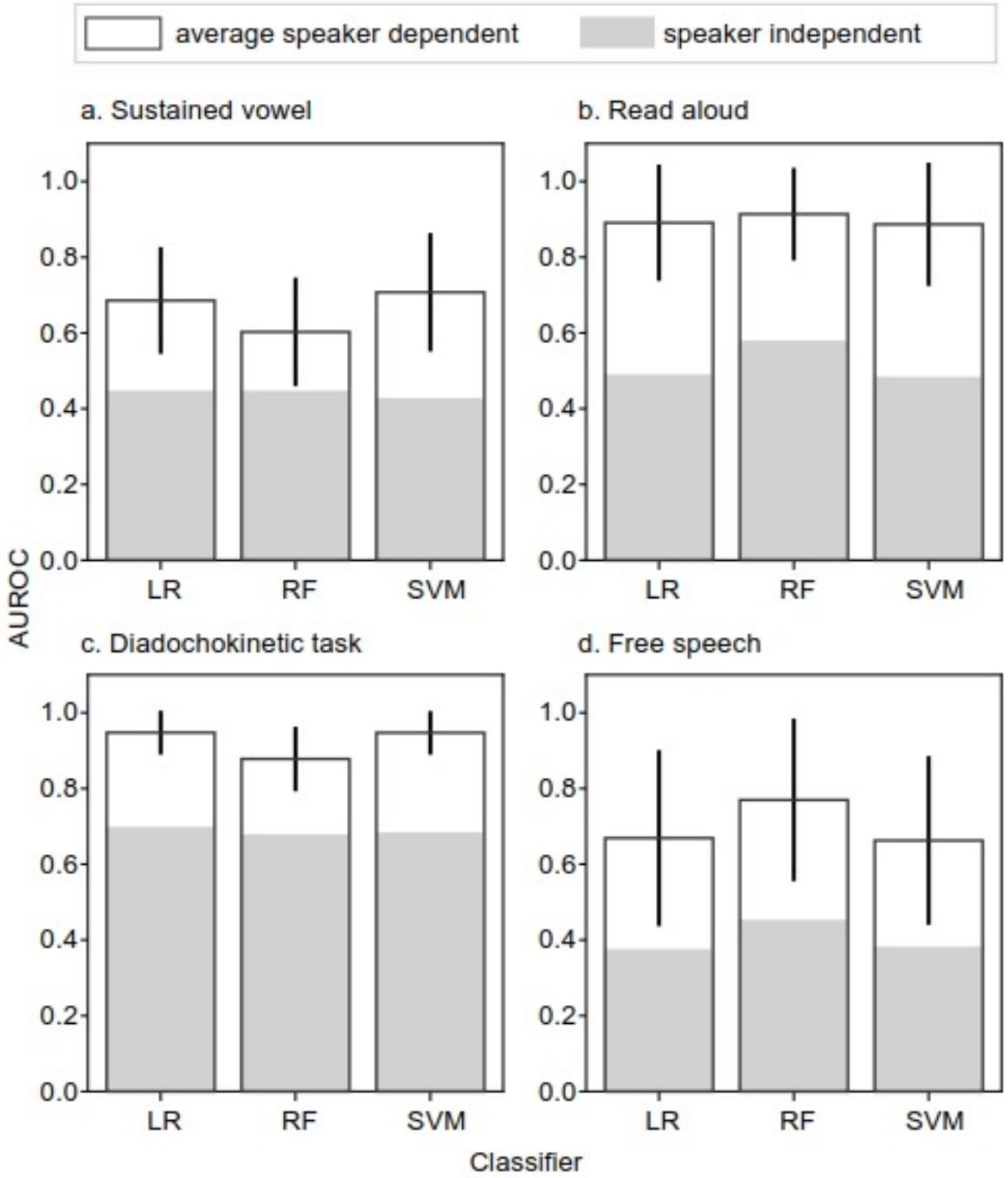


Pilot study (6 subjects)

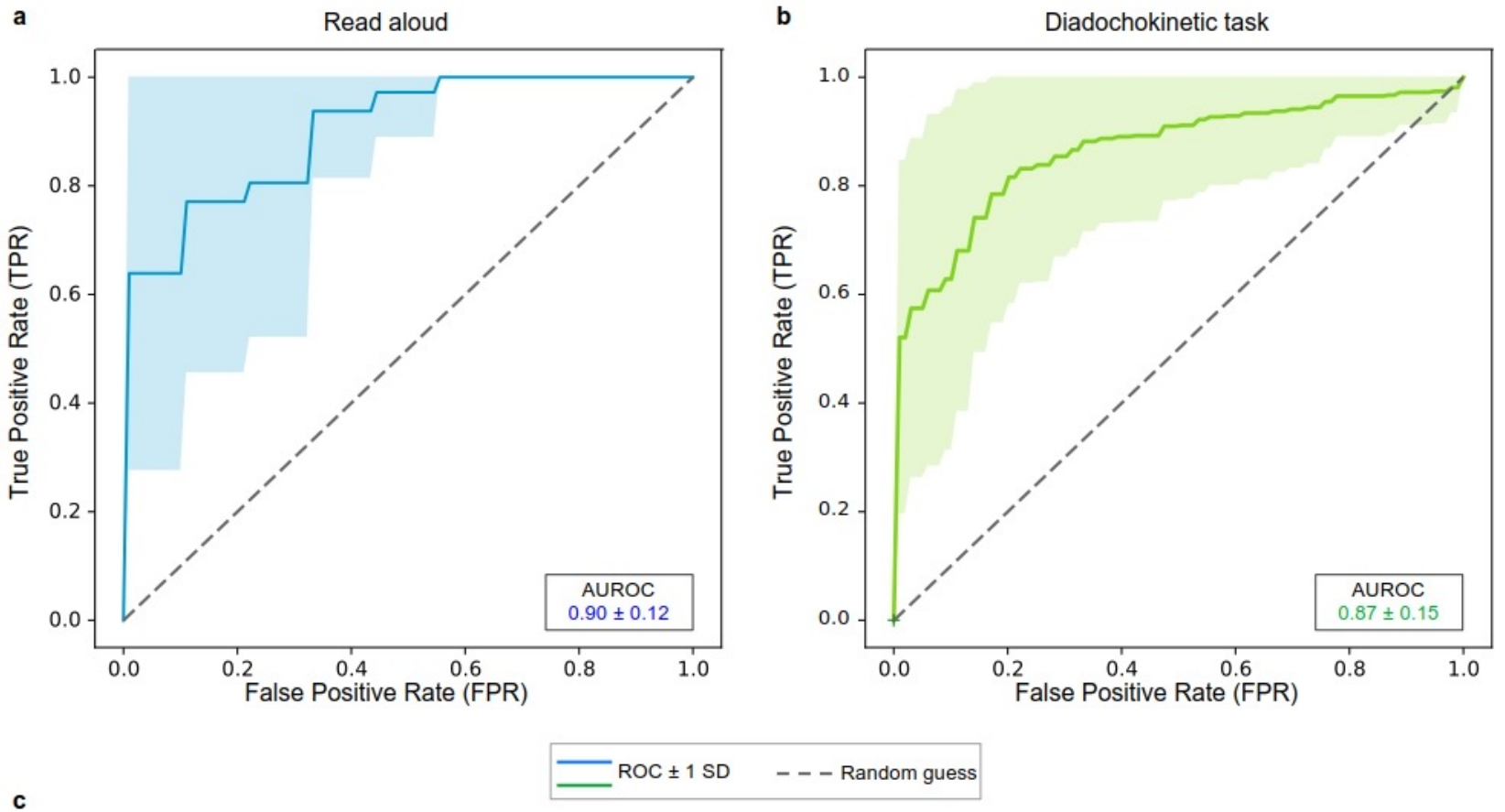




eGeMAPs feature-based system



Main study (16 subjects)

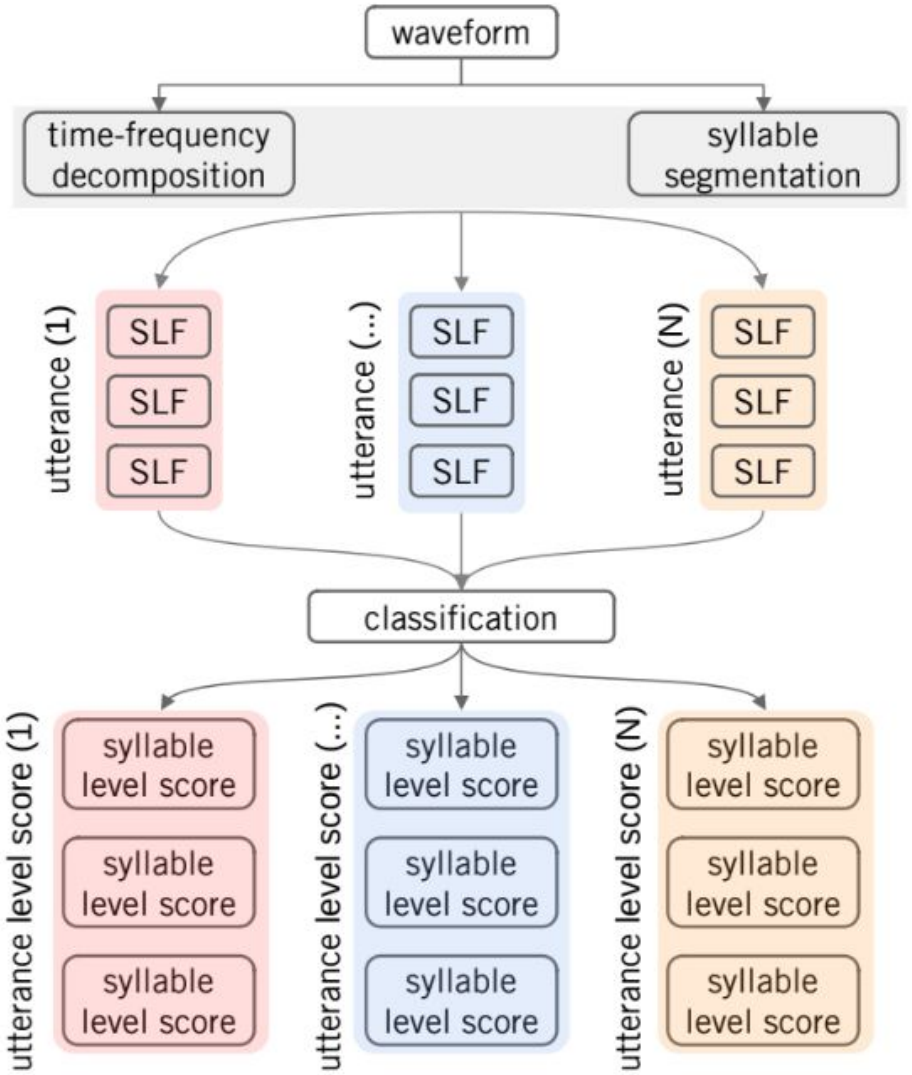


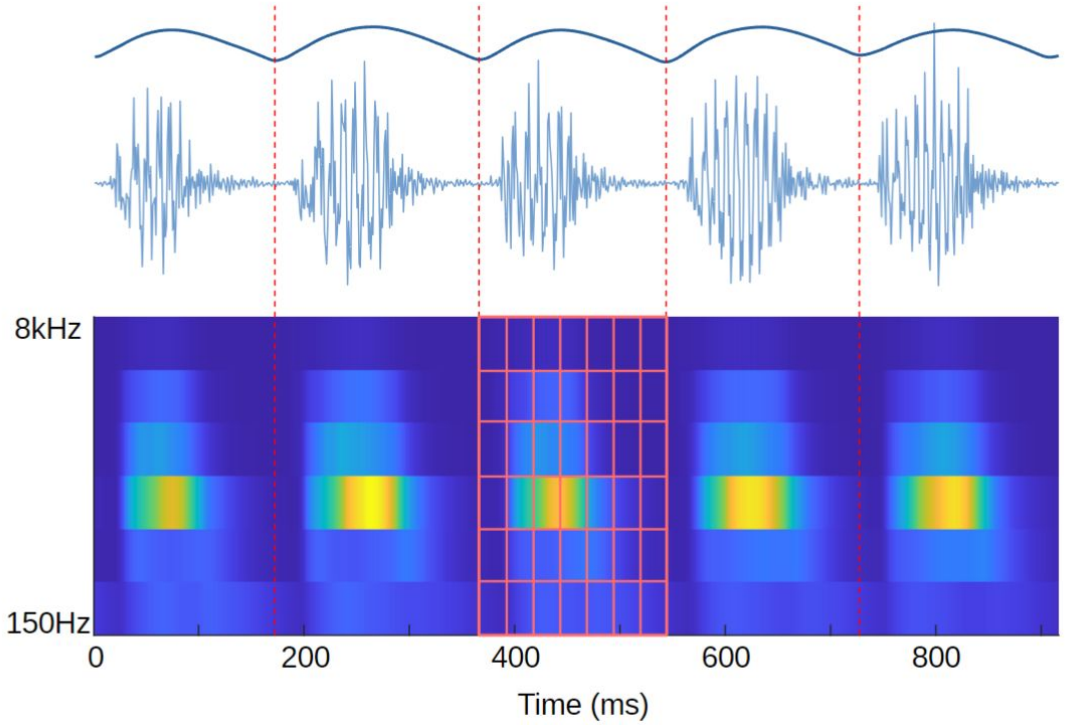
c

task	AUROC	AUPRC	BACC	MCC	F1	Sensitivity	Specificity
Read aloud	0.90 ± 0.12	0.90 ± 0.13	0.87 ± 0.15	0.75 ± 0.30	0.87 ± 0.15	0.88 ± 0.16	0.87 ± 0.15
Diadochokinetic task	0.87 ± 0.15	0.87 ± 0.14	0.82 ± 0.13	0.64 ± 0.26	0.82 ± 0.13	0.82 ± 0.14	0.82 ± 0.13

AUROC: Area under ROC curve
AUPRC: Area under precision-recall curve
BACC: Balanced accuracy
MCC: Matthew's correlation coefficient
F1: F1 score

Segmental-approach validation (1)

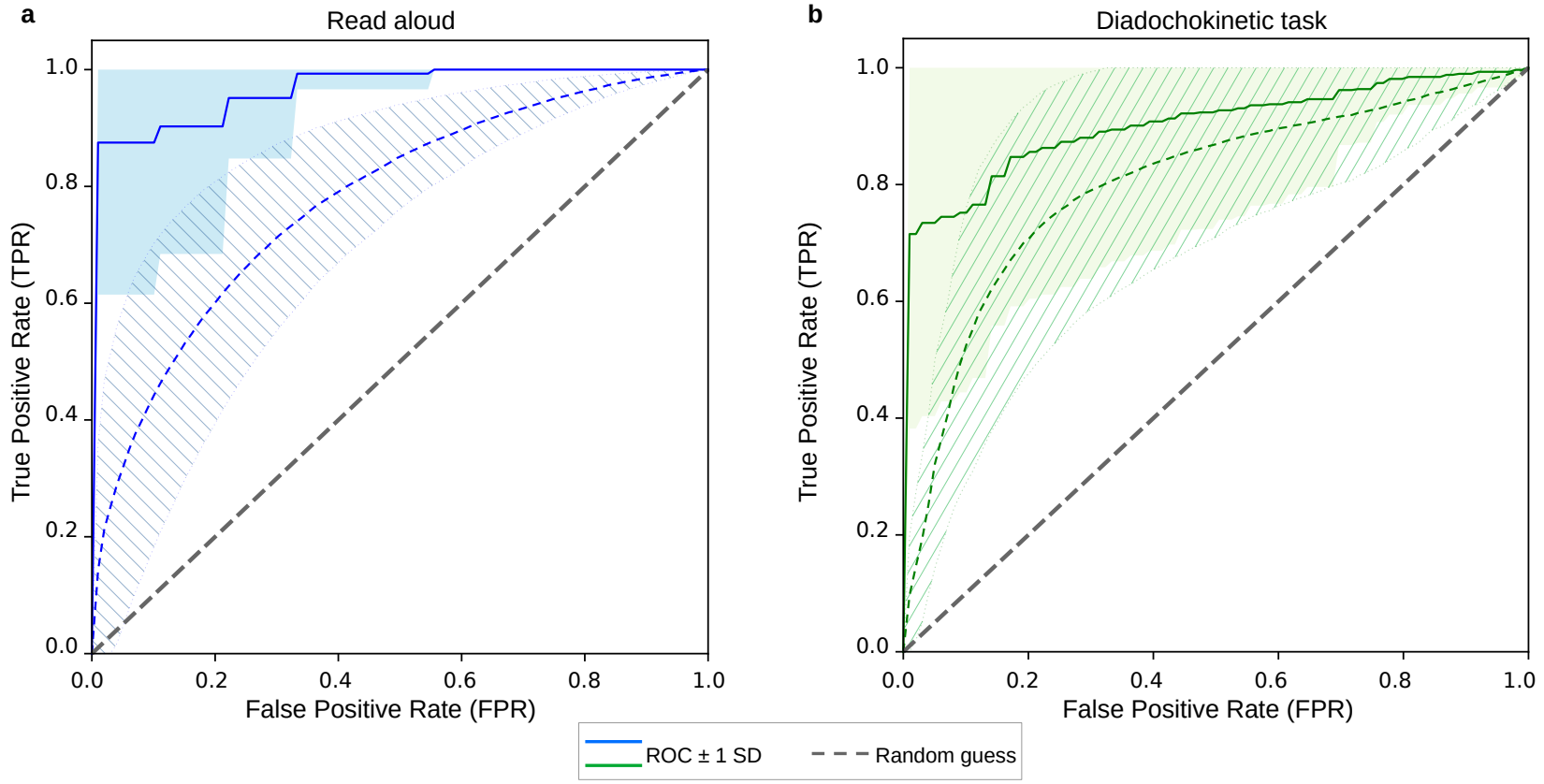




- Multi-point evidence
  - Features are extracted acoustically/linguistically defined segments at multiple points of the speech utterance
-

SYLLABLE LEVEL FEATURES FOR PARKINSON'S DISEASE DETECTION FROM SPEECH, Sevada Hovsepyan and Mathew Magimai-Doss, in Proc. of ICASSP, 2024

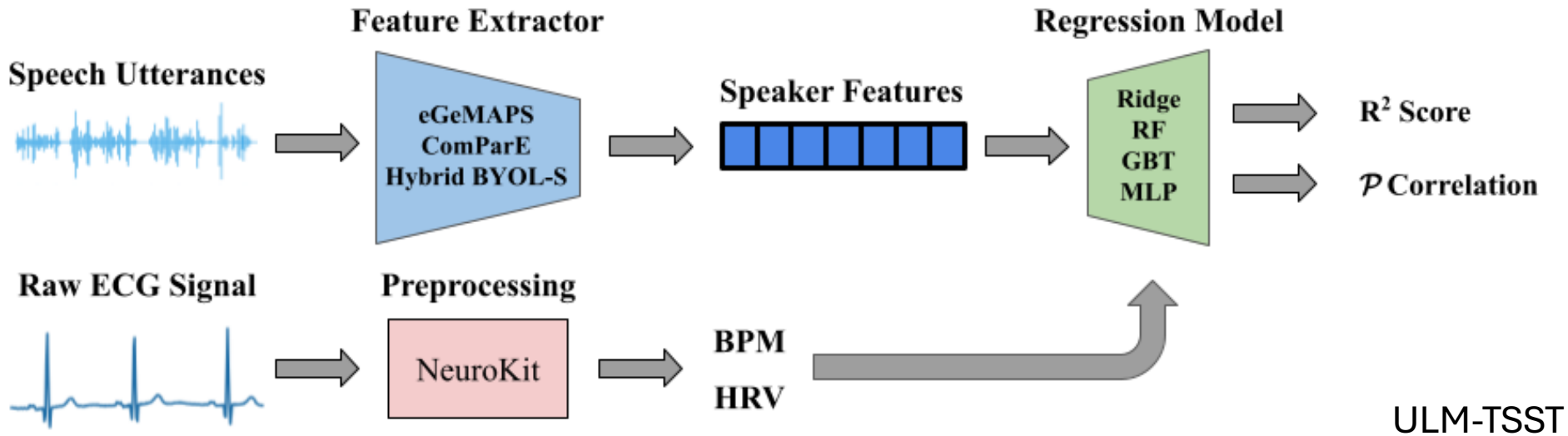
Segmental-approach validation (2)



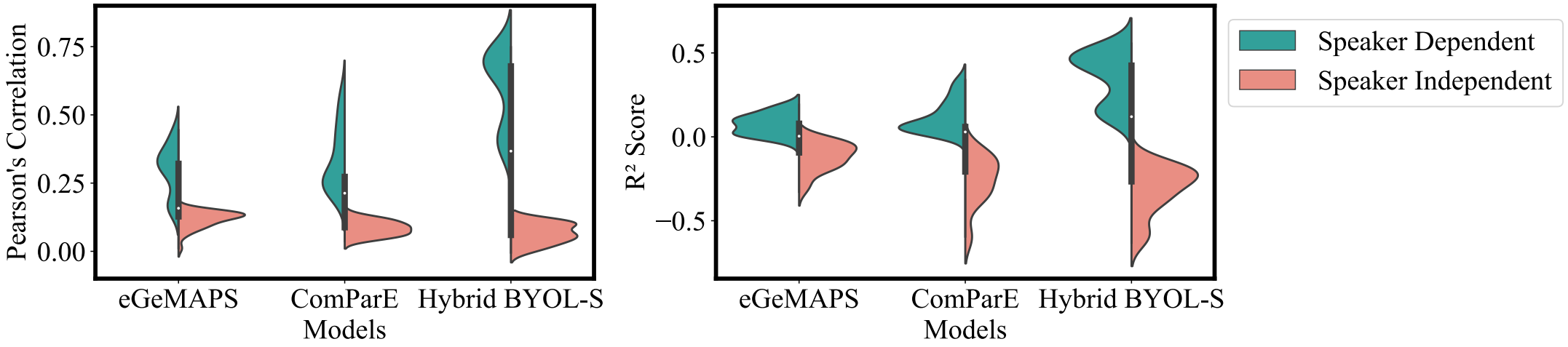
c

task	AUROC	AUPRC	BACC	MCC	F1	Sensitivity	Specificity
Read aloud (utterance based)	0.97 ± 0.06	0.96 ± 0.08	0.96 ± 0.06	0.93 ± 0.12	0.96 ± 0.06	0.96 ± 0.08	0.97 ± 0.09
Read aloud (syllable based)	0.78 ± 0.1	0.77 ± 0.11	0.73 ± 0.11	0.45 ± 0.21	0.74 ± 0.1	0.74 ± 0.1	0.72 ± 0.11
DDK (utterance based)	0.9 ± 0.17	0.91 ± 0.16	0.89 ± 0.14	0.79 ± 0.27	0.89 ± 0.13	0.88 ± 0.15	0.9 ± 0.17
DDK (syllable based)	0.8 ± 0.14	0.78 ± 0.12	0.76 ± 0.12	0.53 ± 0.23	0.76 ± 0.12	0.78 ± 0.13	0.76 ± 0.1

Speech-based heart rate estimation (1)

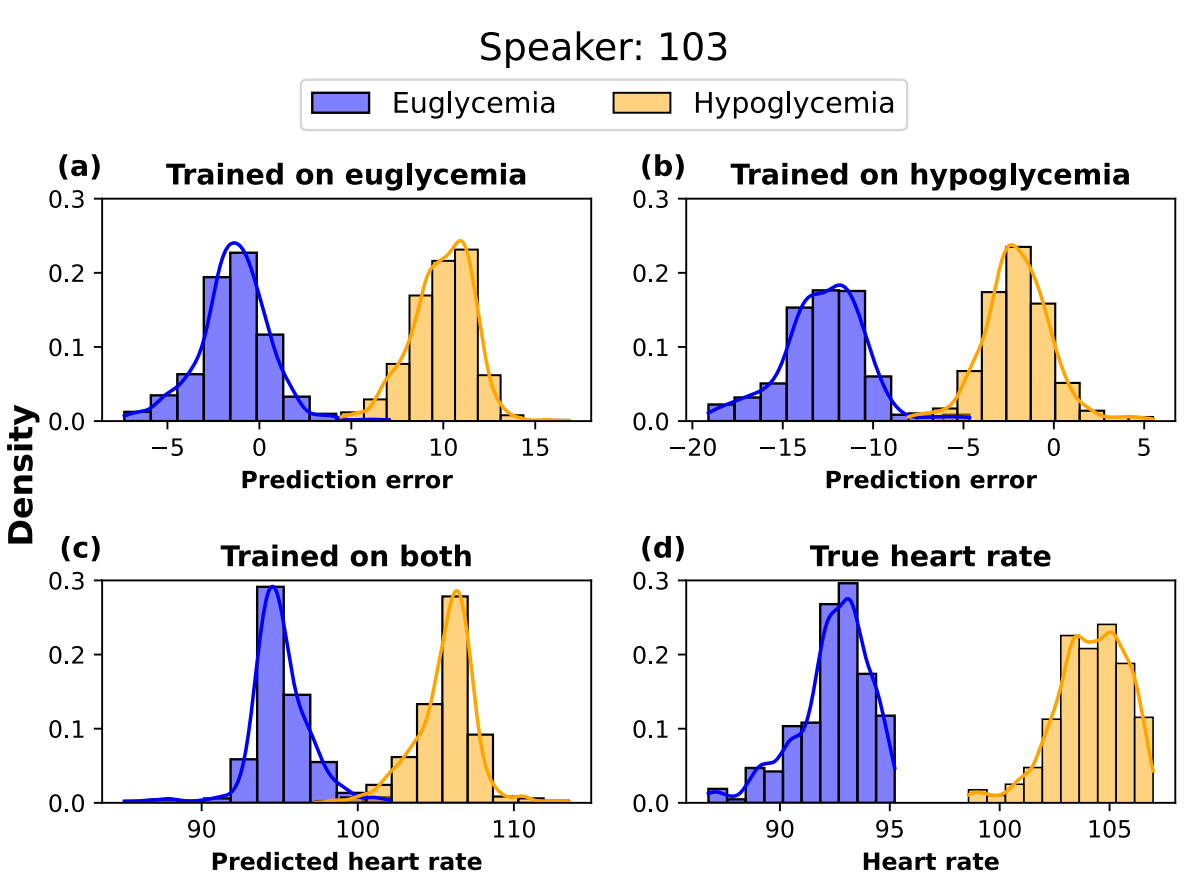


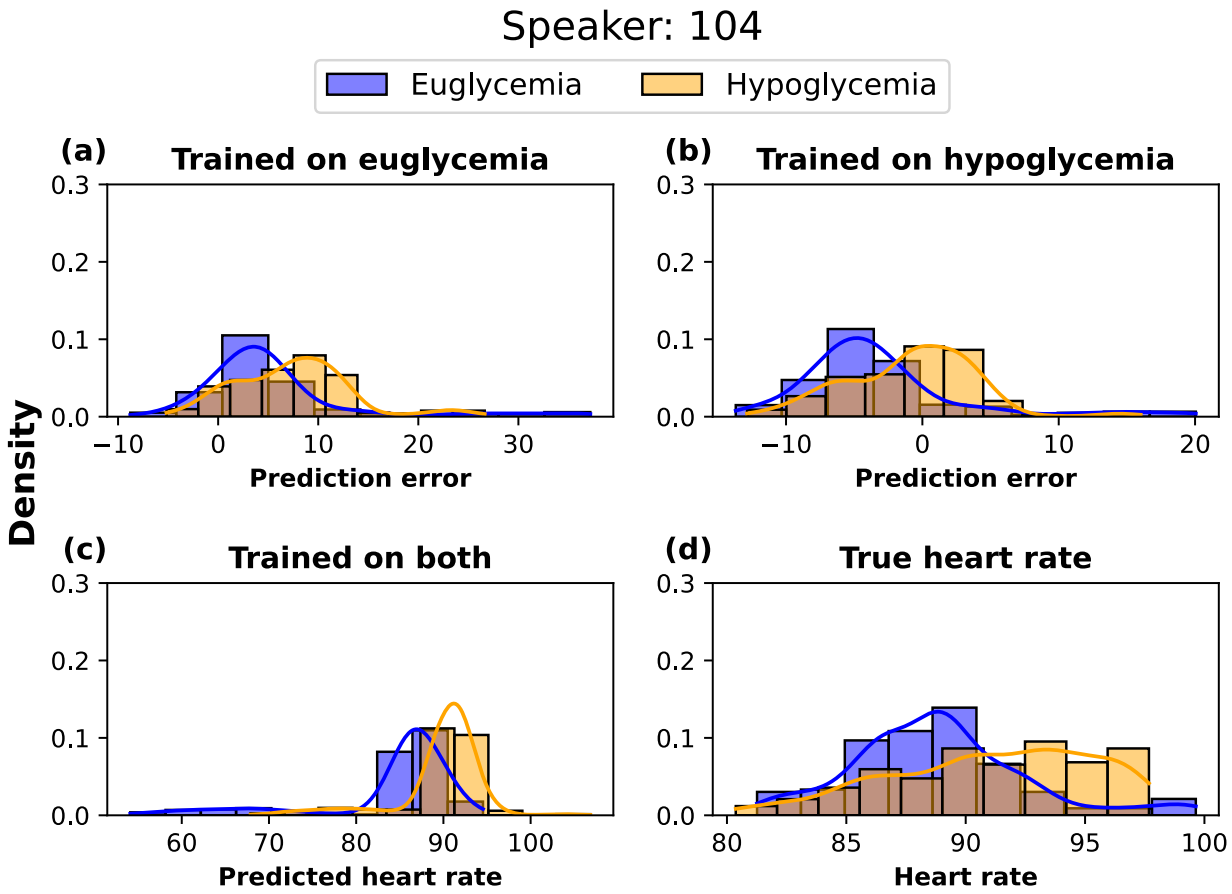
ULM-TSST dataset



Predicting Heart Activity from Speech using Data-driven and Knowledge-based features, Gasser Elbanna, Zohreh Mostaani and Mathew Magimai-Doss, in: Proc. of Interspeech, 2024

Speech-based heart rate estimation (2)





Speakers from pilot study

Towards Integrated Processing of Physiological Signals and Speech, Zohreh Mostaani, PhD Thesis 10679, EPFL, 2025