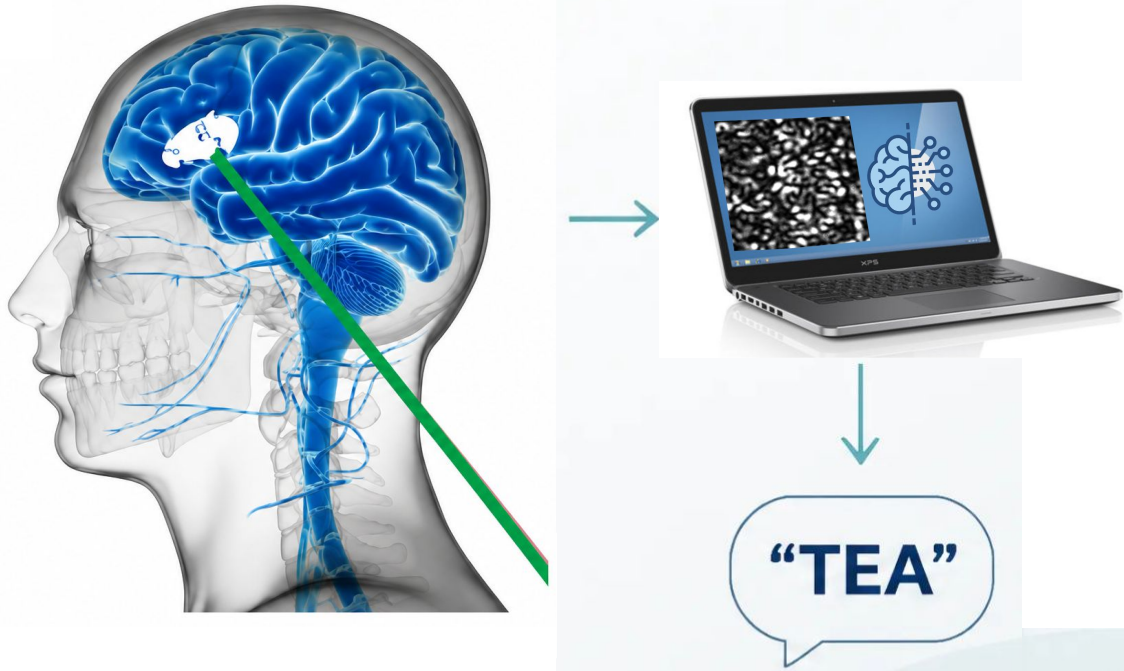
The Leslie and Susan Gonda
Multidisciplinary Brain
Research Center
Bar-Ilan University

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

IMVC 2026

LV-MAE for contactless Inner-speech decoding



Natalya Segal, **Daniel Rubinstein, Moshe Bar, Zeev Zalevsky**

Adapting Long-Video Masked Autoencoders to High-Speed Optical Brain Imaging

Portable, affordable, non-invasive, and contactless. No electrodes or helmets.

Agenda

- Motivation and scope
- Data collection
- Preprocessing, model architecture and post processing
- Results
- Future work

Motivation and scope

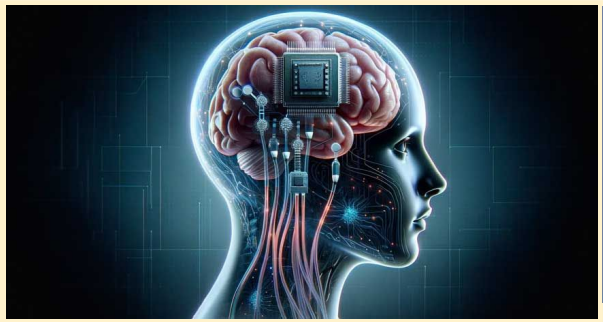
Need contactless, portable, affordable technology

Over 20 million people worldwide lose speech following stroke, ALS or advanced Parkinson's disease

invasive, surgery

NON invasive

*Typical English speech rate ~120 **W**ords **P**er **M**inute*



*Stanford / **Neuralink***

- Implants
- **Clinical trials**, Nov 2025
- **Fast:** ~60-78 WPM
- **125K+ words/unlimited**

Motor cortex → x Parkinson's

Bulky, lab-bound and expensive

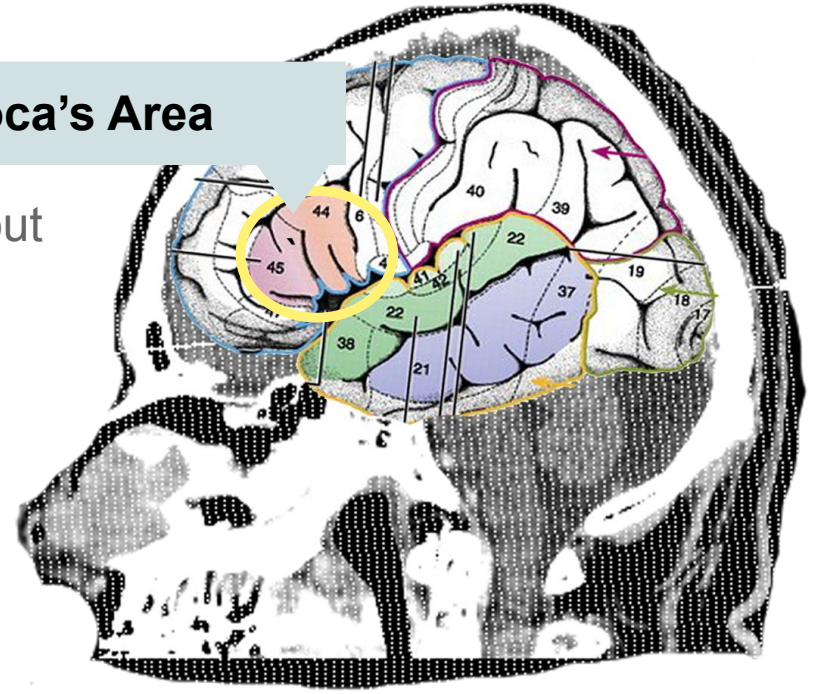
Mature (whole brain) technologies

Require scalp contact

Portable

	fMRI	MEG	EEG	fNIRs
Words **P**er **M**inute	< ~30	10-15	3	~ 1
Vocabulary	unlimited	4	limited	2
Accuracy %	~84, nouns	low	~92, commands	~76%, “yes”/“no”

Neural basis for inner - speech decoding



Inner speech means thinking words silently in the mind without speaking out loud. No physical movement or whispering.

Broca's area vs. **Motor cortex** approach to inner speech BCI

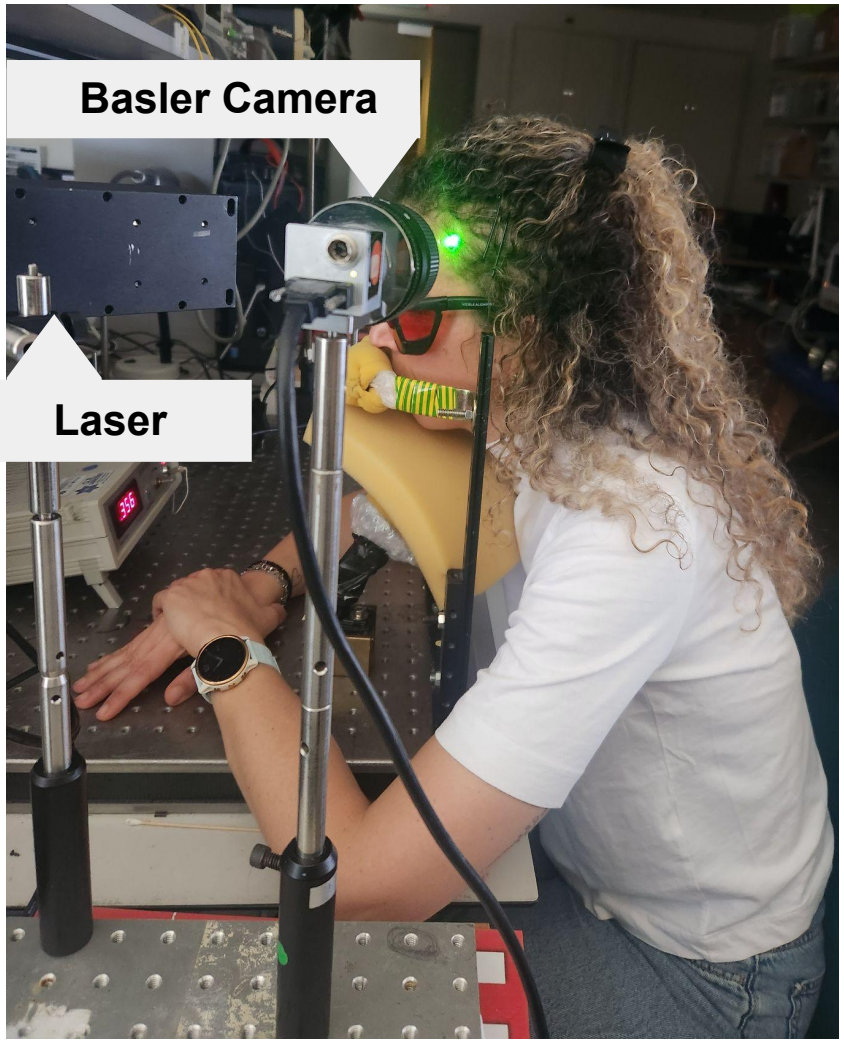
We focus on Broca's area:

- **Language planning**
- Intact in **advanced Parkinson's patients** despite motor cortex impairment.
- Located in the **left hemisphere** in 95% of right-handed and 70% of left-handed individuals

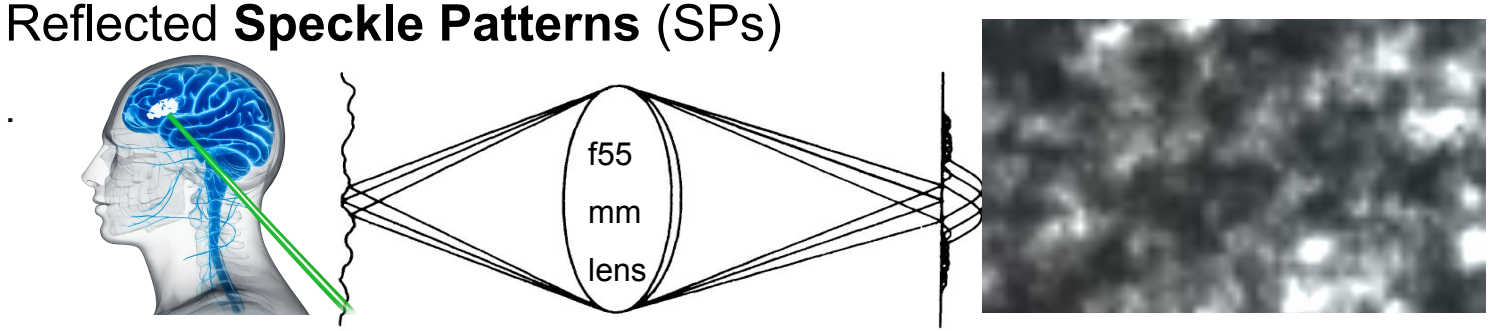
Mechanics:

- Neural **activation** →
- Localized **increase in cerebral blood volume** →
- Elastic micro - deformations of tissues manifesting as **vibrations**.

The Sensing System



Reflected **Speckle Patterns** (SPs)



Formed by interference of **coherent** light scattered from a rough surface:

Bright → **Constructive interference**, high-intensity areas.

Dark → **Destructive interference**, low-intensity areas.

- **Participant silently thinks, no speech, no movement**
- Camera: **1000 fps**, 128×128 px
- Laser: 532 nm green or 850 nm NIR

minimizable

- **Remote**: ~30 cm from the head
- Defocused → Far field: micro-tilts appear as speckle shifts

No electrodes · No gel · No skin contact· Minimizable · Portable

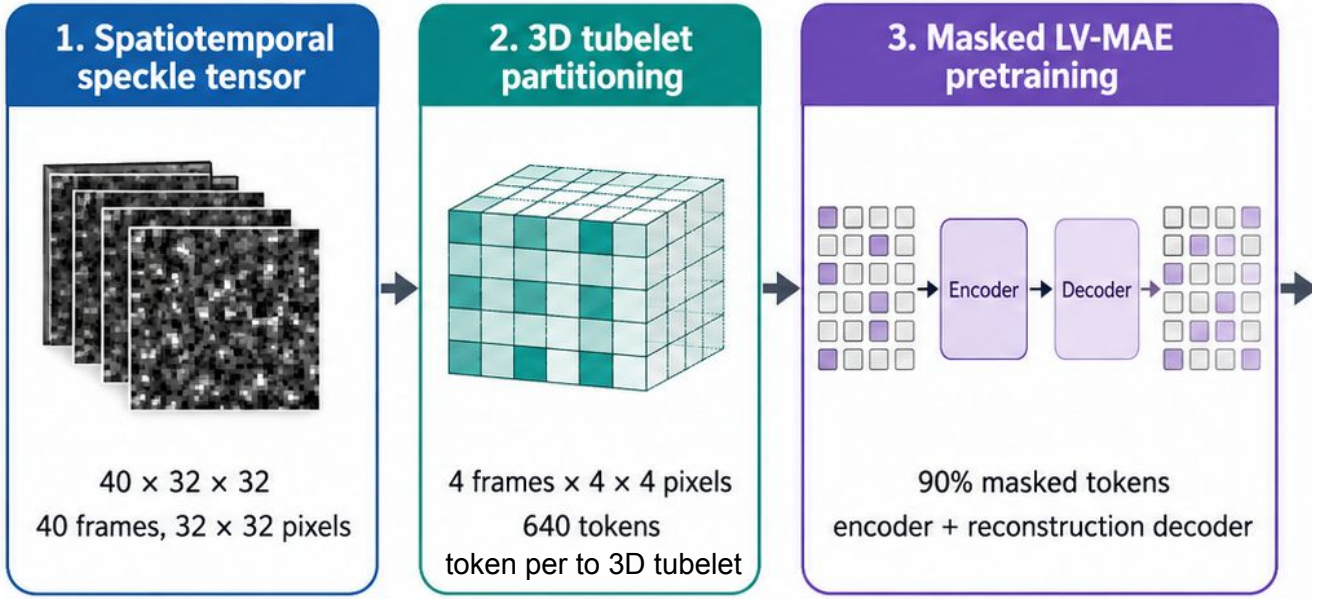
LV-MAE: From Semantic to Non-Semantic Video

Original Long-Video Masked Autoencoder (LV-MAE) uses InternVideo2 as a frozen short-video encoder;

1 — Encode short segments
off-the-shelf encoder per segment
each segment → one embedding

2 — Masked-embedding AE
mask & reconstruct segment embeddings
learns long-range cross-segment context

Our model learns short-segment embeddings directly from speckle dynamics.



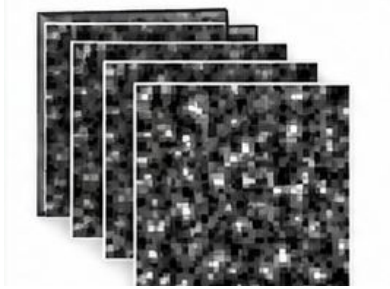
Naiman, I. *et al.*,2025, **ICCV**, LV-MAE: learning long video representations through masked-embedding autoencoders. https://doi.org/10.48550/arXiv.2504.03501,

LV-MAE: From Semantic to Non-Semantic Video

Our model learns short-segment embeddings directly from speckle dynamics.

Objects · Faces · Actions · Scenes

Domain shift



No objects · No actions · Pure dynamics

LIGHTWEIGHT

~2.6 M parameters total ·

Fast training

~30-50 epoches for good results·

- Resolution HD (1920 × 1080 and above).
- Frame rate 24 – 60 fps
- Temporal scale Seconds to minutes
- Short segments **5 s**

- **No recognizable semantic content**
- The signal emerges in how the speckle *shifts and evolves across frames*.
- 28x128 px, 1000 fps,
- milliseconds temporal scale
- Short segments **4 ms**

LV-MAE was designed for long semantic videos; our work shows that masked video representation can be adopted to high-speed, non-semantic biomedical video, enabling contactless decoding of inner-speech content from speckle dynamics.

Datasets, Downstream Tasks, and Experimental Design

Balanced datasets

LV-MAE AUTOENCODERS TRAINING:

- Leave one out on 10 participants,
- Participant silently thinks "yes" / "no"
- No speech, no movement
- 10 models
- 20 s clips, 10 clips per subject per class
- 400K frames per participant
- 3.6M frames per autoencoder

Downstream Tasks: Training/Validation/Test:

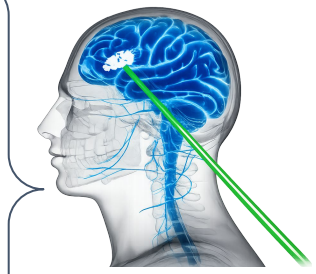
2-words experiment:

Held-out participant · 400K frames · 9-fold CV

4-word experiment:

15 participants · 800K frames each, 12M total,

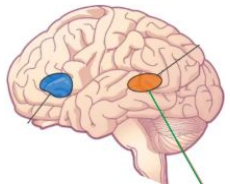
new vocabulary, same area

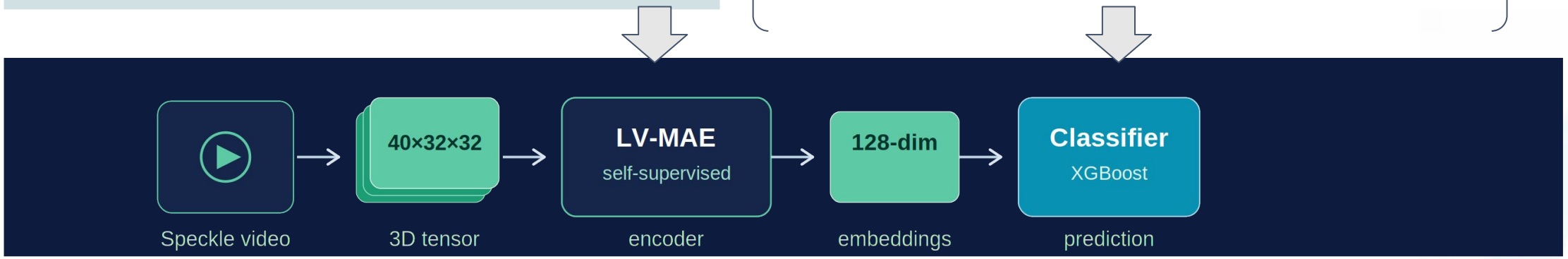


Comprehension experiment: 14 participants ·

new task + different cortical region,

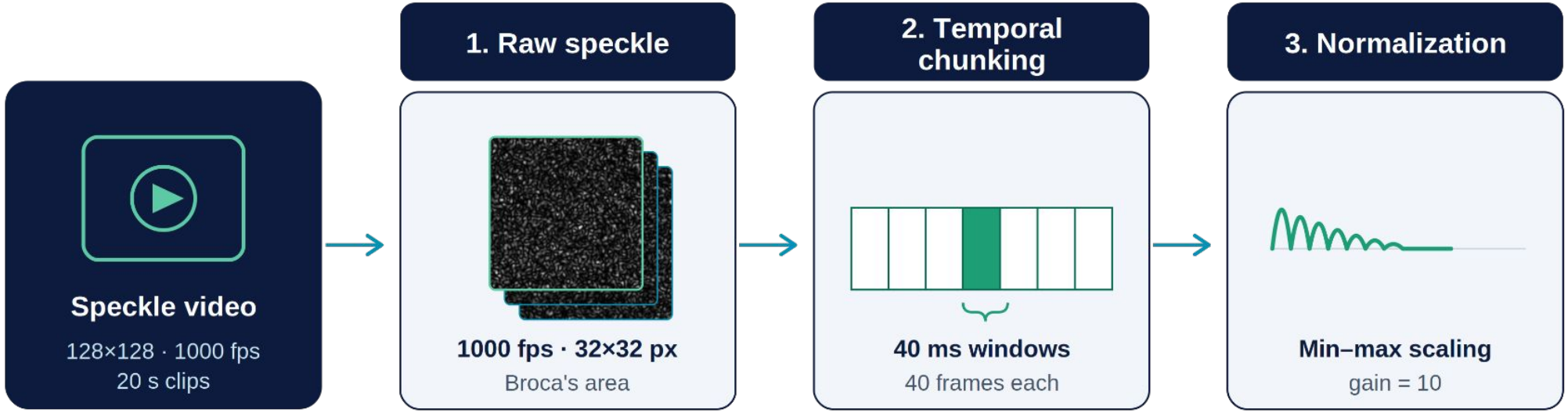
testing framework transferability



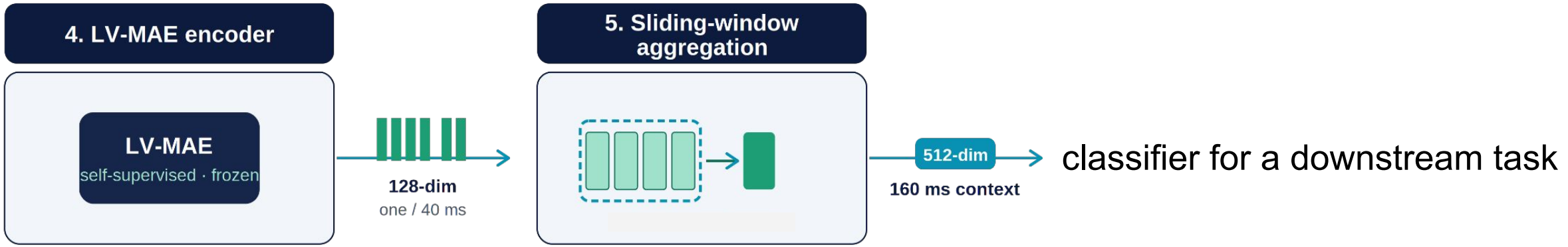


Preprocessing

Common preprocessing before LV-MAE training and inference:



Before classification, LV-MAE embeddings are extracted and aggregated over time:

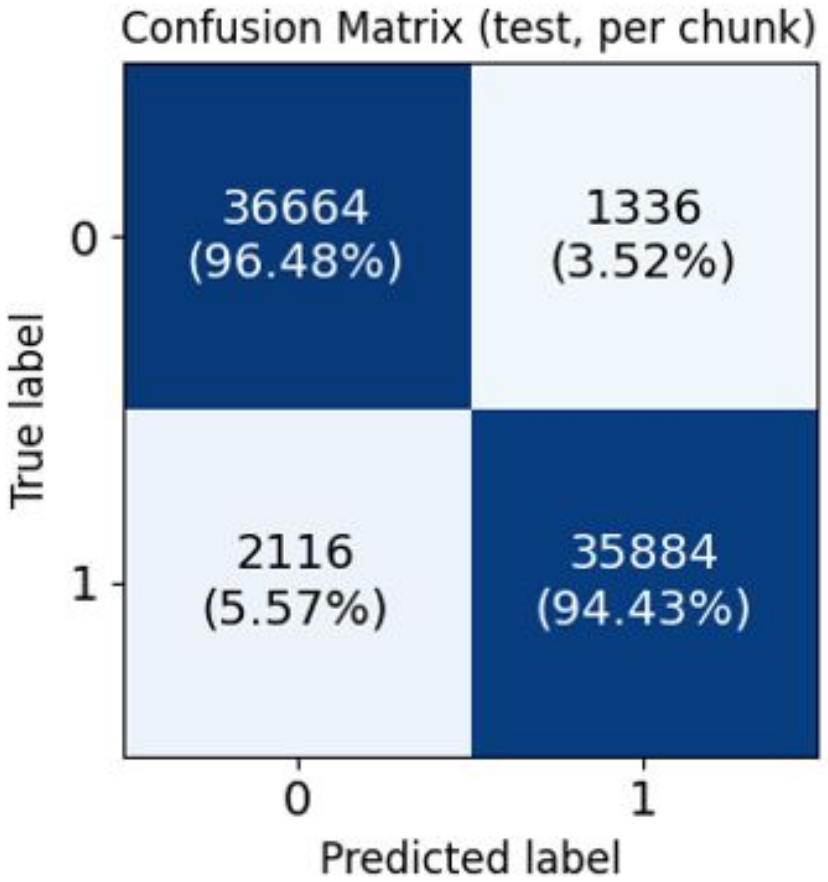


Results - “yes” vs. “no”. Proof-of-concept.

Preprint DOI: 10.21203/rs.3.rs-8703722/v1

Calibration

Classifier trained on embeddings of first 2 clips, validated on the 3rd one and tested on the remaining 7 clips, 40ms

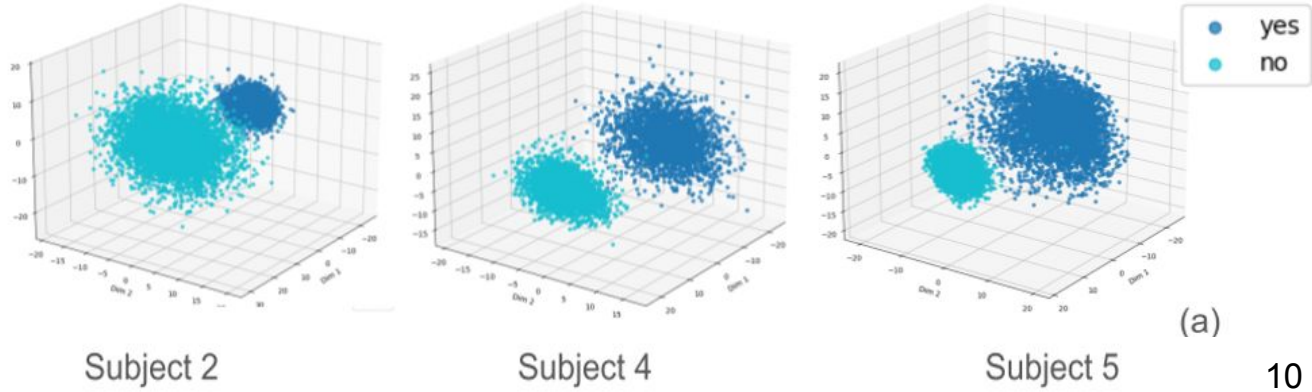


LOW-DATA REGIME

As 60 s of labeled data per class: 40-s for training and subsequent 20-s for validation

Across 10 splits

95.7%



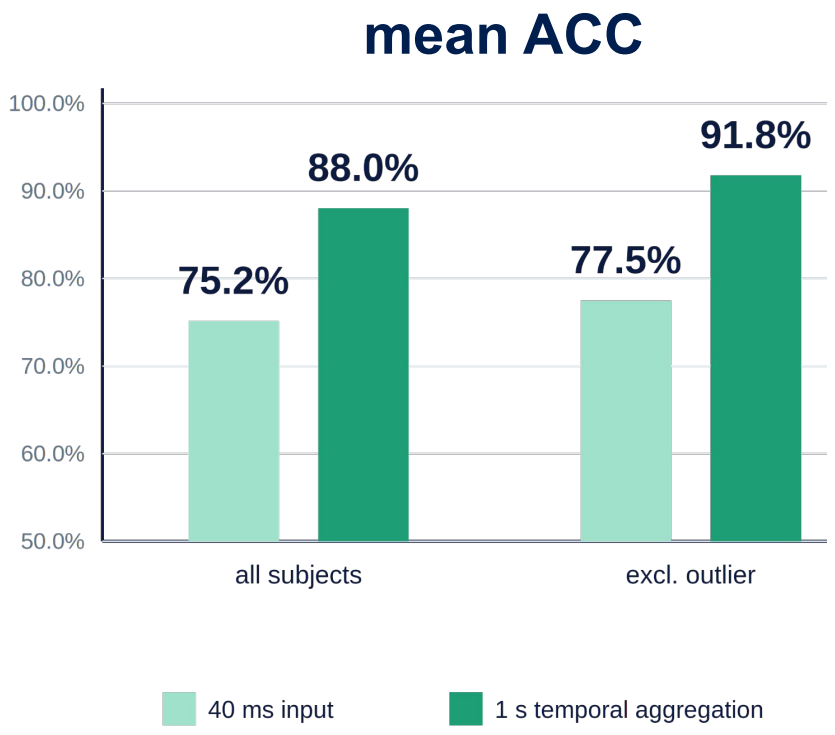
Results - “yes” vs. “no”. Proof-of-concept.

Preprint DOI: 10.21203/rs.3.rs-8703722/v1

Generalization

Classifier trained on embeddings of 9 other subjects

Above-chance decoding in 10/11 splits, outlier: the only participant with right-lateralized Broca’s area.



88.0-91.8% at 60 WPS

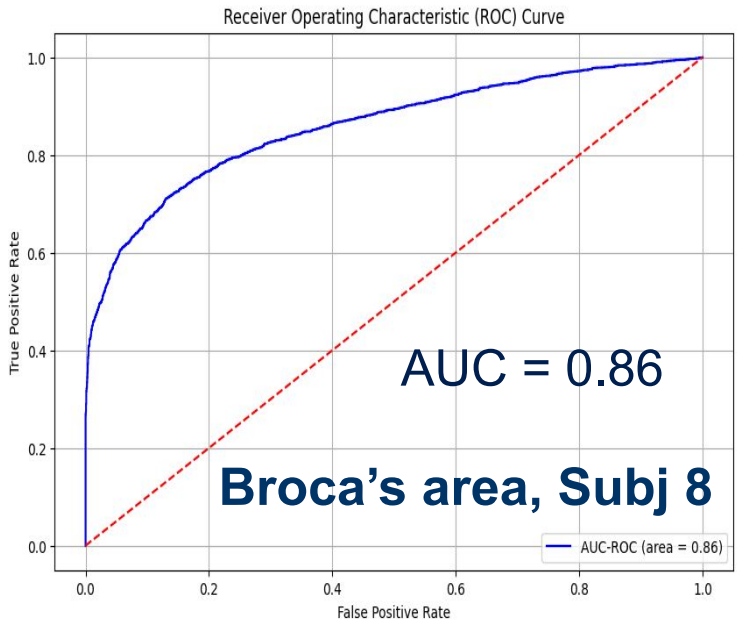
One clip: 20s per class for threshold calibration.

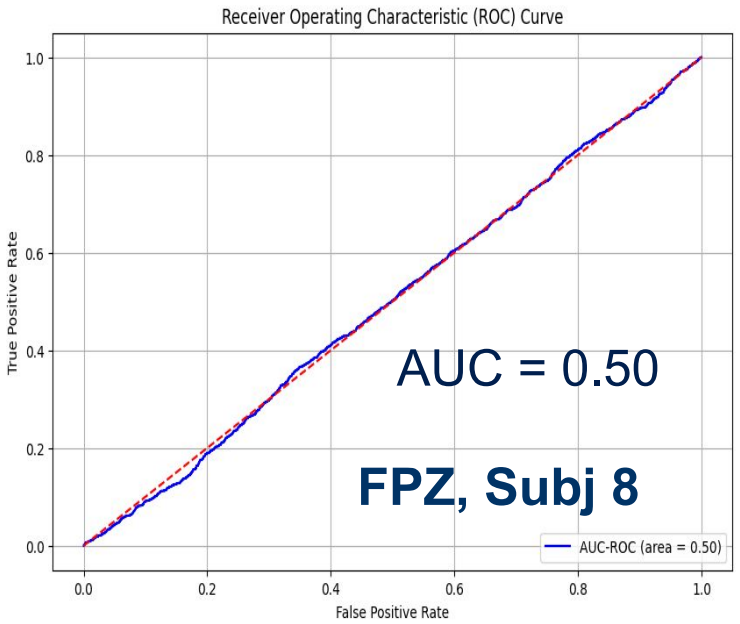
Controls Ruling Out Mumbling and Systemic Movements

Preprint DOI: 10.21203/rs.3.rs-8703722/v1

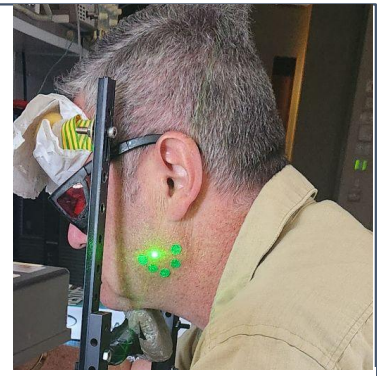
Forehead (FPZ) control

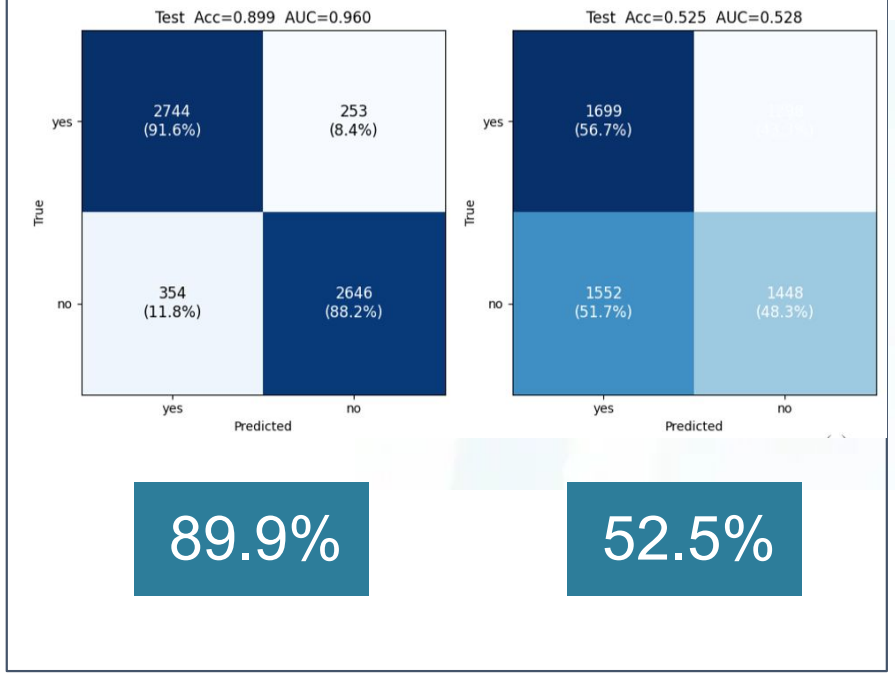
mean ACC on FPZ = 57%, 1 s, 10cv





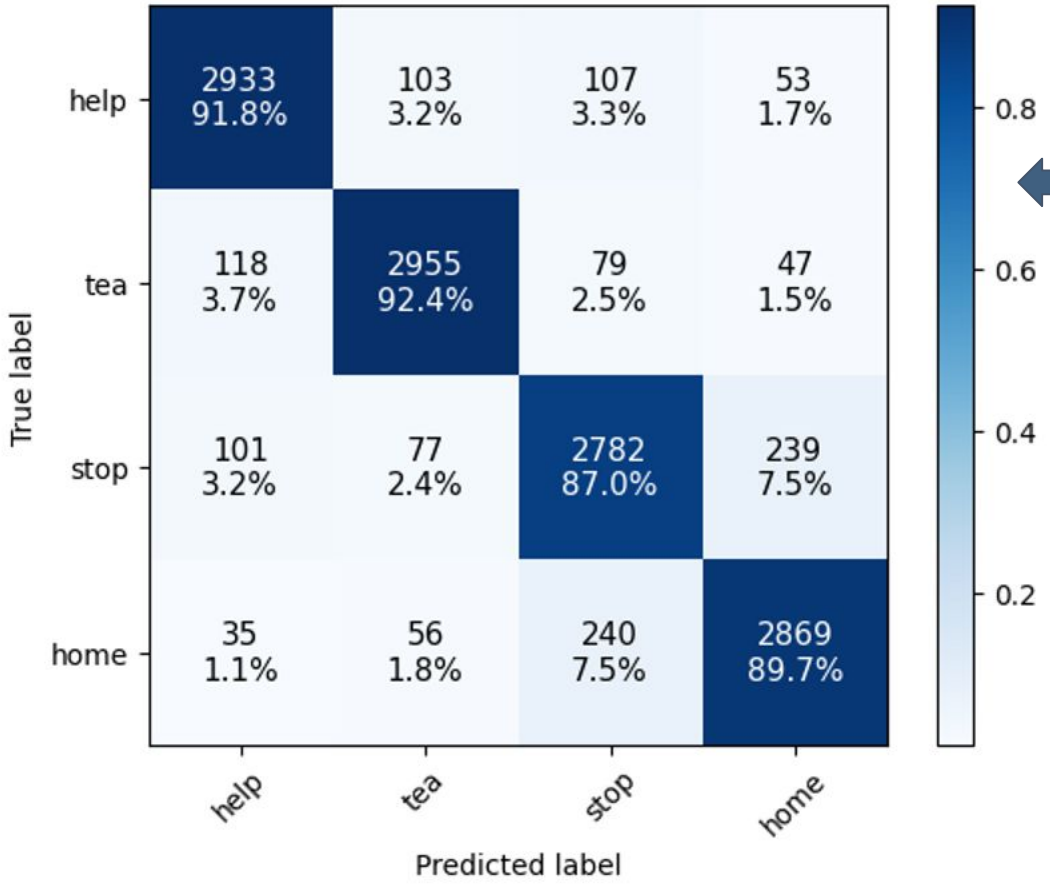
mandibular area control



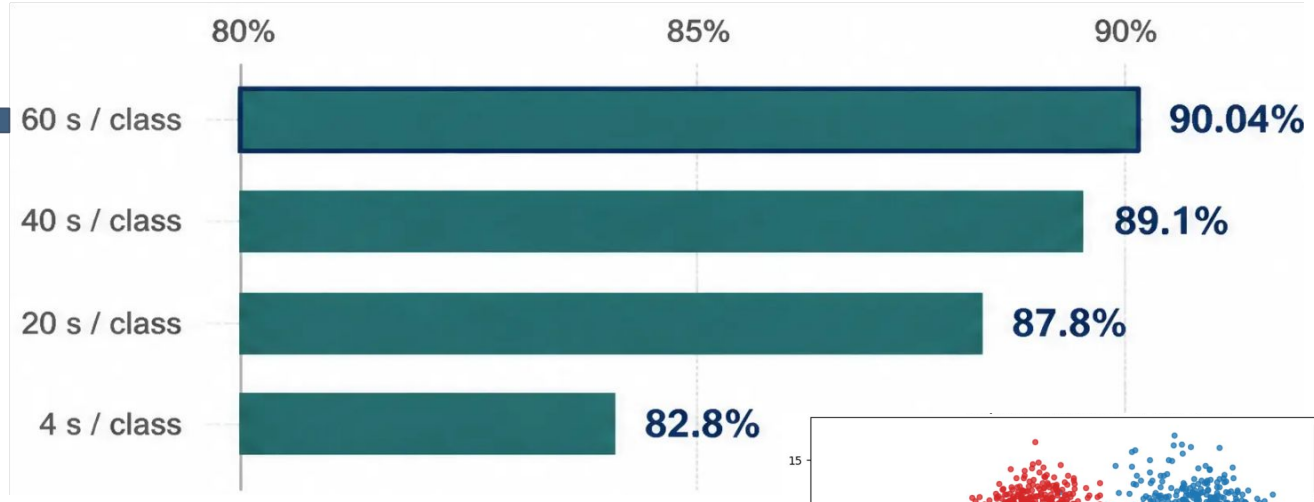


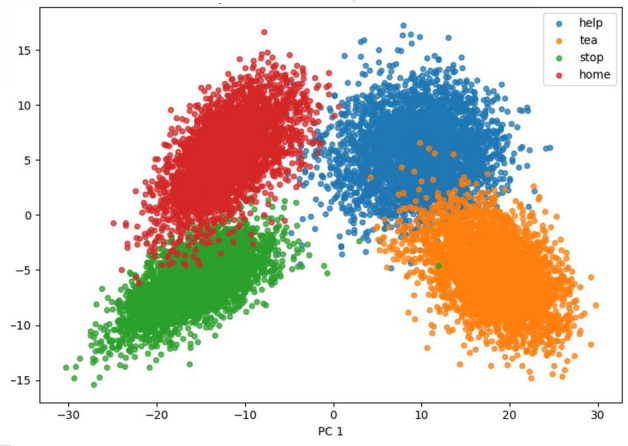
Results: Vocabulary: help · tea · stop · home

Contactless recording over Broca's area. interleaved acquisition | 15 participants | Interleaved



Mean Macro-Accuracy vs. Calibration Budget (per class)





≈ 1 English syllable

160-ms input window

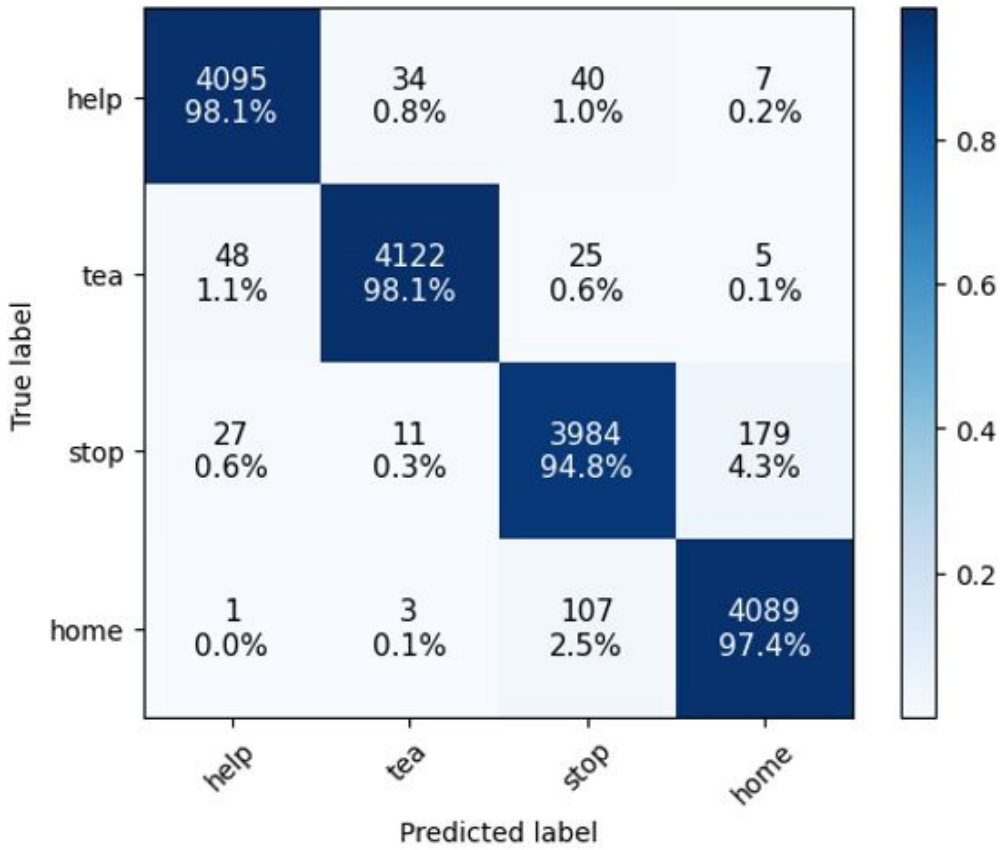
90.04%

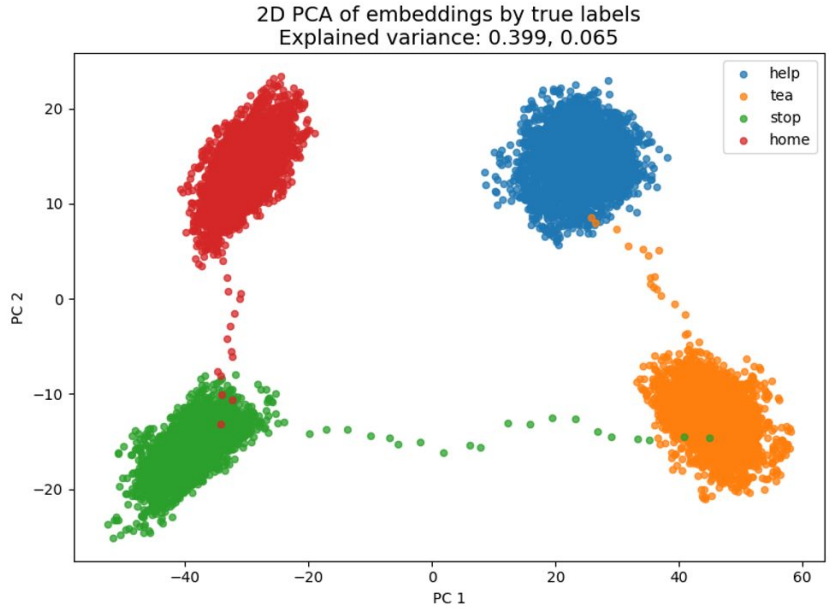
Macro-Accuracy

60 s calibration per class,

4-min total calibration (4 words × 60 s)

Results — 4-Words, word per second





“help” clip · “tea” clip · “stop” clip · “home”clip · pause clip·
…
“help” clip · “tea” clip · “stop” clip · “home”clip · pause clip

10 iterations

Temporal Aggregation Result

1-s input window, 60WPS

97.0%
mean Macro-Accuracy

F1 = 0.97
mean Macro-F1

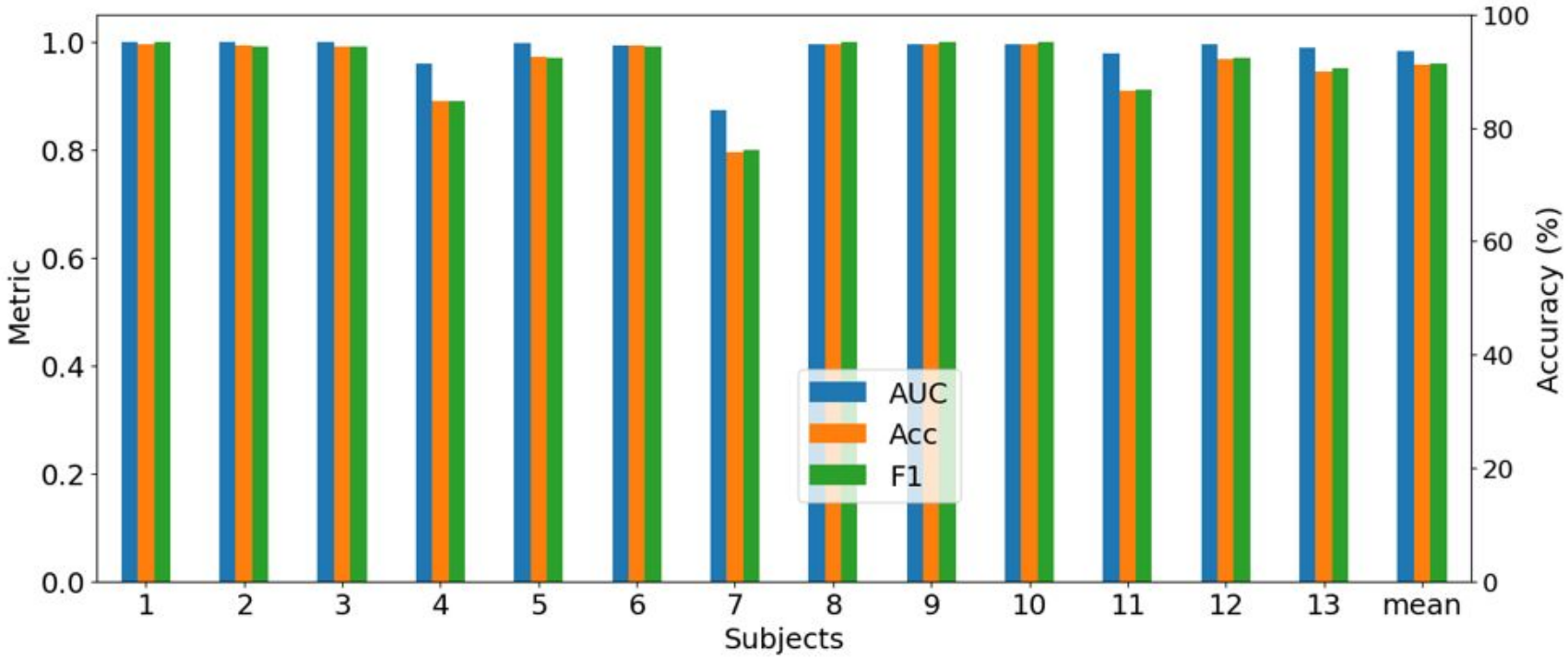
Only **20 s** calibration per class required

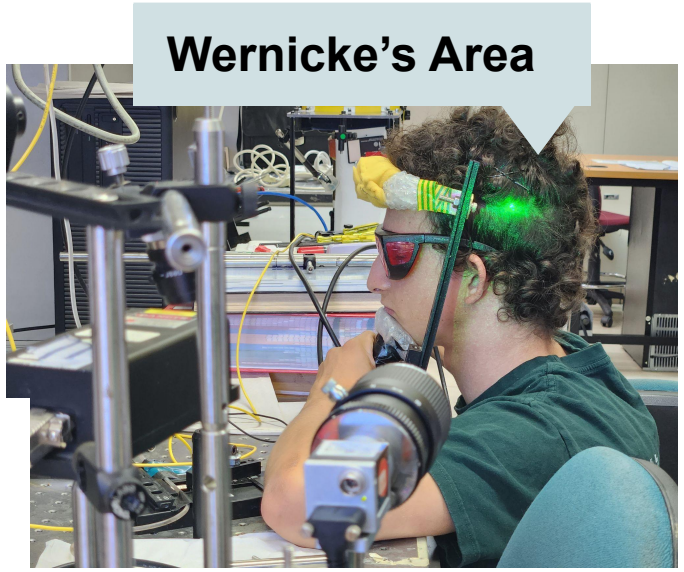
Vocabulary: help · tea · stop · home | 15 participants

Results - transferability to Brain different region

DOI: 10.21203/rs.3.rs-9141990/v1

Comprehension:
English vs. Swedish using LV-MAE pretrained on Broca's area to extract embeddings





96.5%

Mean Accuracy

20-s calibration per class
40-ms inputs

Next Steps

Future directions for contactless inner-speech decoding

Larger Vocabularies

Extend to larger vocabulary , explore multi syllable words and phrase-level decoding.

Clinical Populations

Test with stroke patients, ALS / locked-in syndrome, and advanced Parkinson's disease.

From classification to generative models

Employ predictive approach to word and phrase level decoding

Wearable

Minimize camera and laser. Compress models to enable inference on edge devices

Stability over time

Cross-day calibration transfer and session-independent representations over time.

Architecture improvements

Explore higher levels of representation hierarchy and other architectural adjustments to reduce calibration requirements in larger vocabularies.

The Leslie and Susan Gonda
Multidisciplinary Brain
Research Center
Bar-Ilan University

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

QUESTIONS?

Making speech restoration as natural and unobtrusive as corrective eyewear.

With thanks to our amazing collaborators and colleagues.

Scan the QR code to get the papers and preprints behind this talk.

Code: https://github.com/natalyasegal/SpecklesAI I https://github.com/danielrubinsteinishere/SpeckleAI_Visuals

References

https://www.researchgate.net/profile/Natalya-Segal/research

1. Naiman, I. *et al.* LV-MAE: learning long video representations through masked-embedding autoencoders. *arXiv* 2504.03501 (2025). https://doi.org/10.48550/arXiv.2504.03501

2. Kalyuzhner, Z. *et al.* Remote photonic detection of human senses using secondary speckle patterns. *Sci. Rep.* **12**, 519 (2022). https://doi.org/10.1038/s41598-021-04558-0

3. Segal, N. *et al.* AI-powered remote monitoring of brain responses to clear and incomprehensible speech via speckle pattern analysis. *J. Biomed. Opt.* **30**, 067001 (2025). https://doi.org/10.1117/1.JBO.30.6.067001

4. Segal, N., Kalyuzhner, Z., Agdarov, S., Beiderman, Ya., Beiderman, Y. & Zalevsky, Z. Conference abstract: Remote decoding of inner speech. In *Dynamics and Fluctuations in Biomedical Photonics XXIII*, Proc. SPIE **PC13850**, PC138500D (2026).

5. Natalya Segal, Daniel Rubinstein, Moshe Bar, Zeev Zalevsky et al. Remote Optical Decoding of Inner Speech in Broca's Area via AI-based Speckle Pattern Analysis (July 2026, IOP Physics Photonics), DOI: 10.1088/2515-7647/ae869f.

7. Natalya Segal, Moshe Bar, Daniel Rubinstein et al. Contactless optical decoding of cortical language responses via region-transferable speckle dynamics, 25 March 2026, PREPRINT: https://doi.org/10.21203/rs.3.rs-9141990/v1