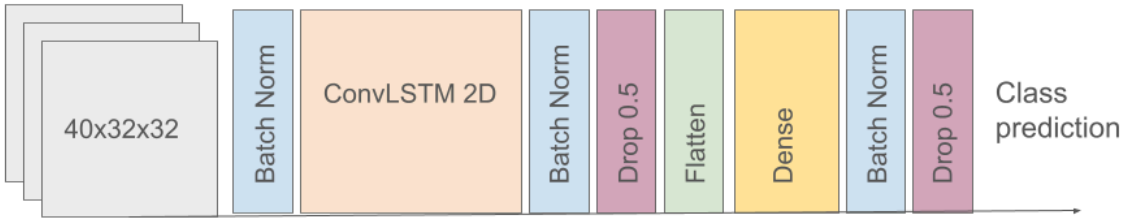
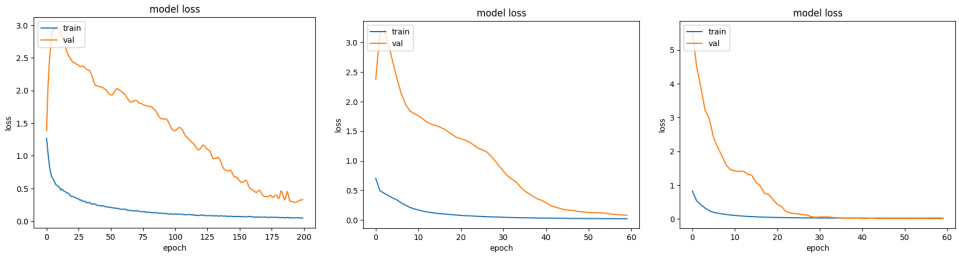
Remote Optical Decoding of Inner Speech in Broca's Area via AI Speckle Pattern Analysis

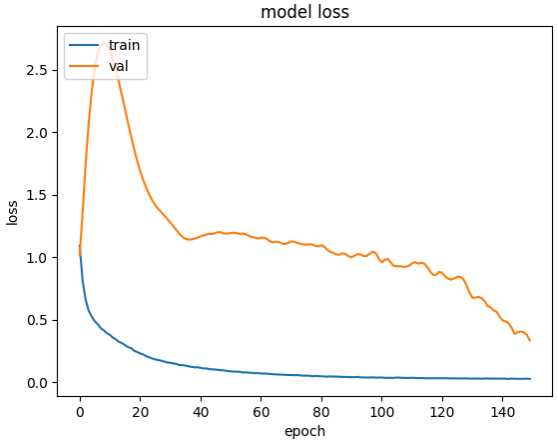
Supplementary Information



Supplementary Fig. 1 (Fig. 2 from ref 21): Layers of the DNN model. The input is a speckle-pattern video tensor of size 40 × 32 × 32, and the output is a classification score.

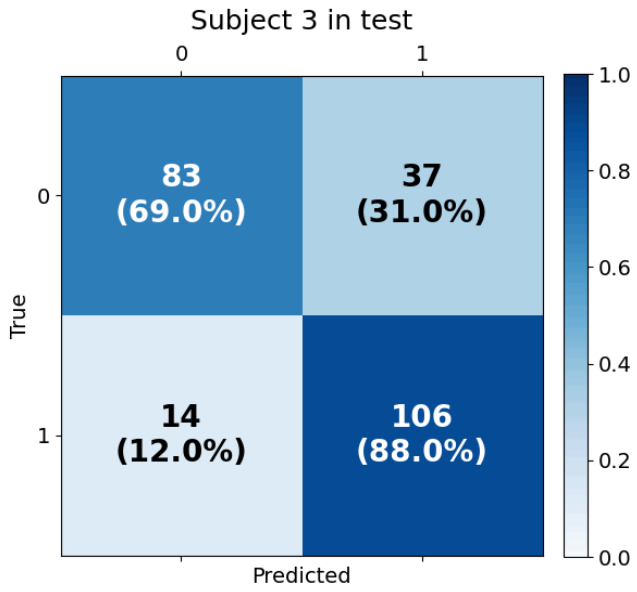


(a)

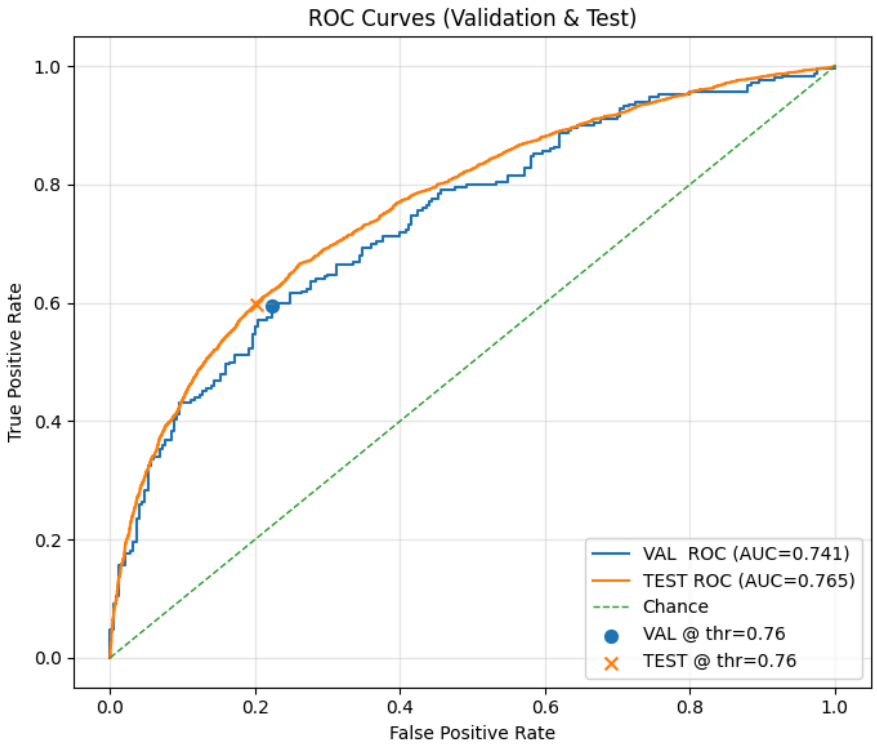


(b)

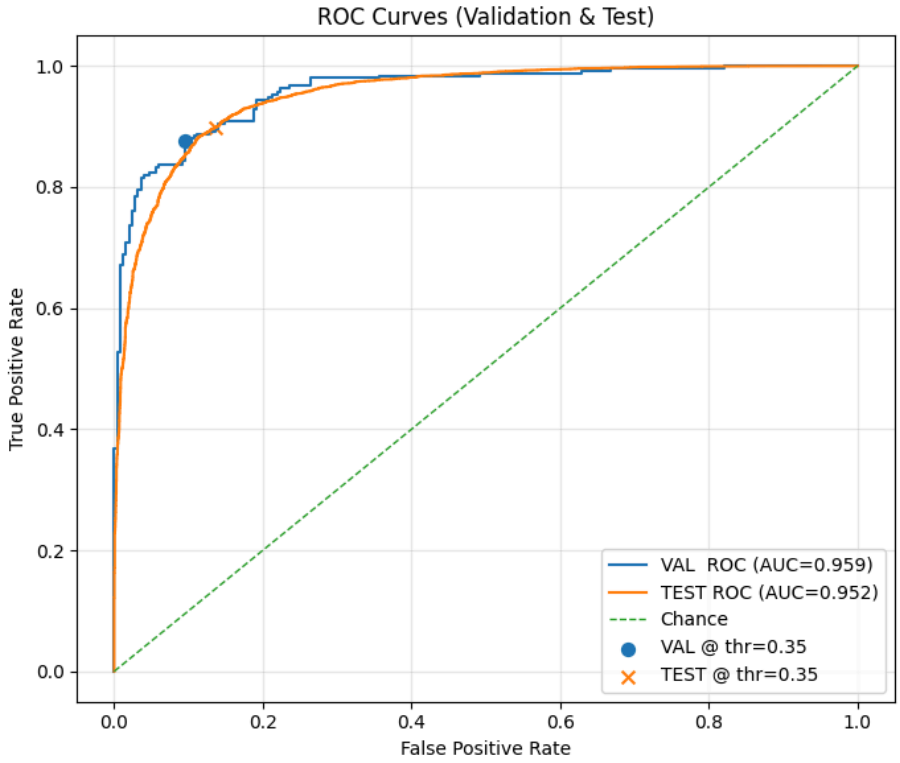
Supplementary Fig. 2: ConvLSTM convergence in per-subject experiments. (a) Representative loss curves over training epochs, with training loss shown in blue and validation loss in orange. The plots correspond to the training of per-subject models for subjects 4, 6, and 8 (left to right). **(b)** Training loss curves for Subject 5 indicate that model performance may benefit from additional training epochs, as convergence has not yet been fully reached.



Supplementary Fig. 3: Confusion matrix of cortical-selective classifier, on the control validation set of Subject 3 (morning data, Table 3), with 1 s temporal aggregation (N = 240). A cortical-selective classifier was trained exclusively on data from subjects who showed no forehead decoding.



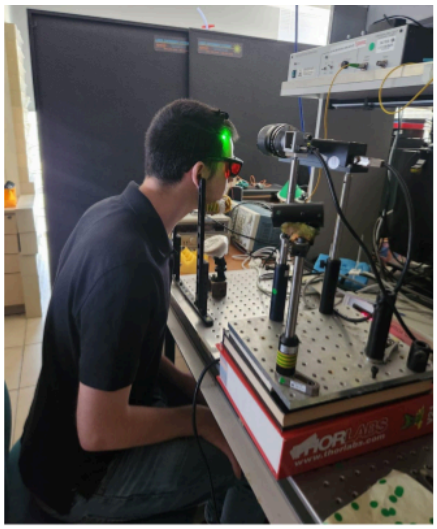
(a)



(b)

Supplementary Fig. 4: Effect of LV-MAE pretraining duration (30 vs. 50 epochs). Classifier ROC curves for Subject 10 (split 1) after LV-MAE pretraining for 30 and 50 epochs. Receiver operating

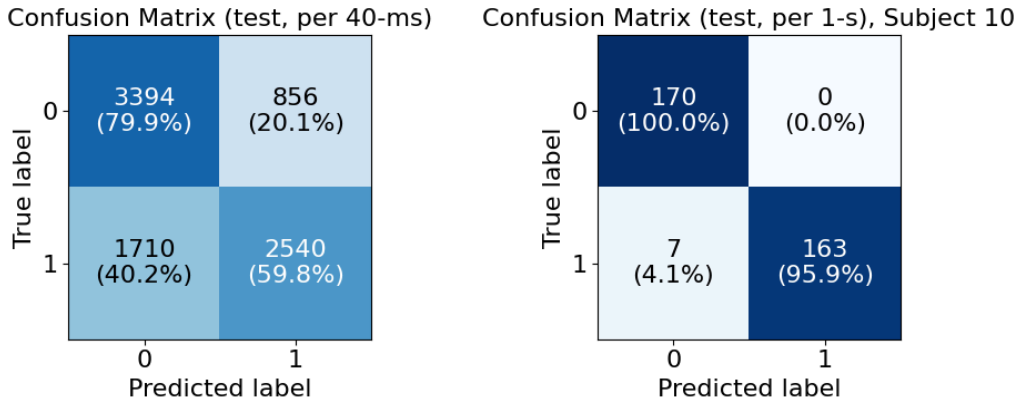
characteristic (ROC) curves for Subject 10 evaluated on 40-ms input segments. The left curve corresponds to the LV-MAE model trained for 30 epochs (a), and the right curve to the model trained for 50 epochs (b). Evaluated on min–max normalized inputs (gain = 10), with normalization applied per 40-ms segment.



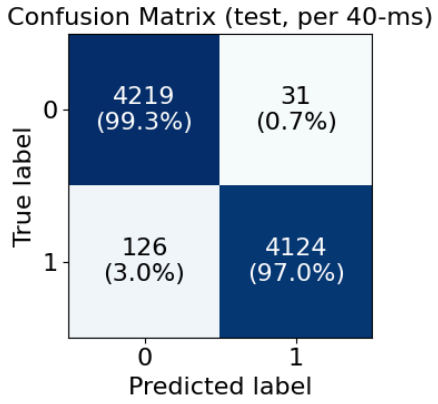
(a)

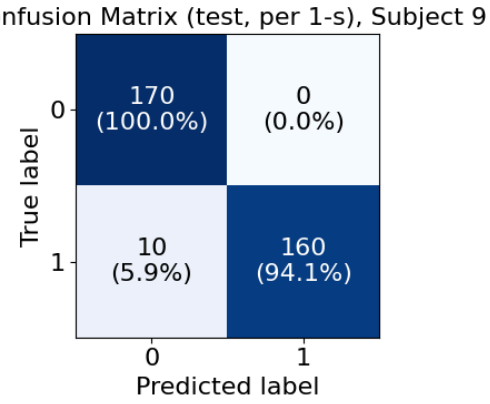
(b)

Supplementary Fig. 5: (a) Photograph showing the experimental setup for subjects with right-hemisphere lateralization of Broca's area. (b) Photograph showing the experimental setup for subjects with left-hemisphere lateralization of Broca's area (the majority of participants). Written informed consent was obtained from the participants for publication.

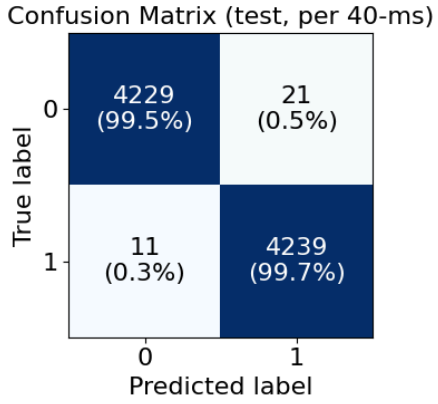


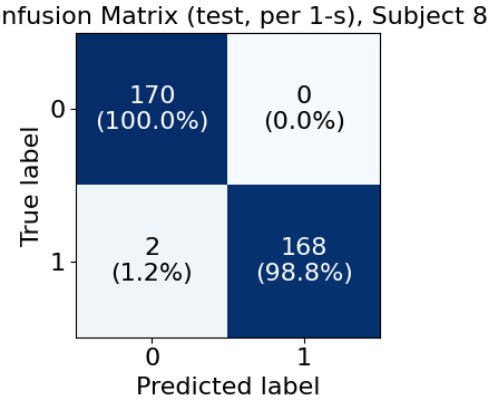
(a)


Confusion Matrix (test, per 40-ms)
4219
(99.3%)
31
(0.7%)
126
(3.0%)
4124
(97.0%)
True label
Predicted label
0
1

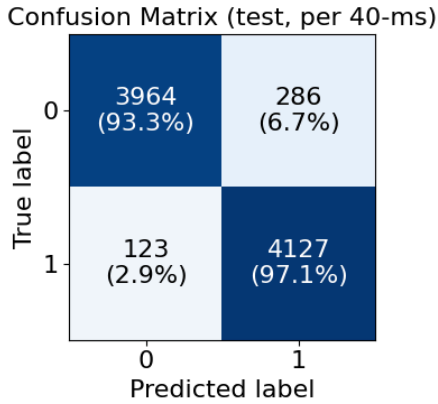

Confusion Matrix (test, per 1-s), Subject 9
170
(100.0%)
0
(0.0%)
10
(5.9%)
160
(94.1%)
True label
Predicted label
0
1

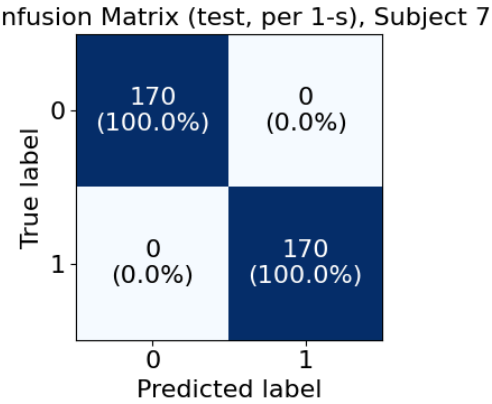
(b)


Confusion Matrix (test, per 40-ms)
4229
(99.5%)
21
(0.5%)
11
(0.3%)
4239
(99.7%)
True label
Predicted label
0
1

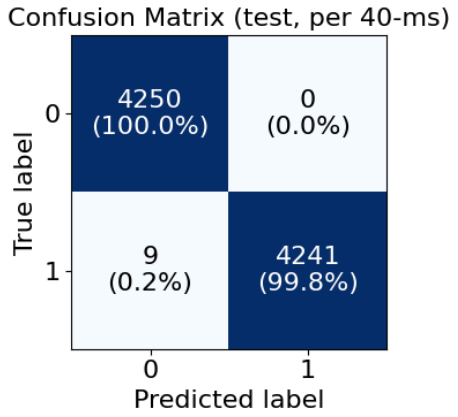

Confusion Matrix (test, per 1-s), Subject 8
170
(100.0%)
0
(0.0%)
2
(1.2%)
168
(98.8%)
True label
Predicted label
0
1

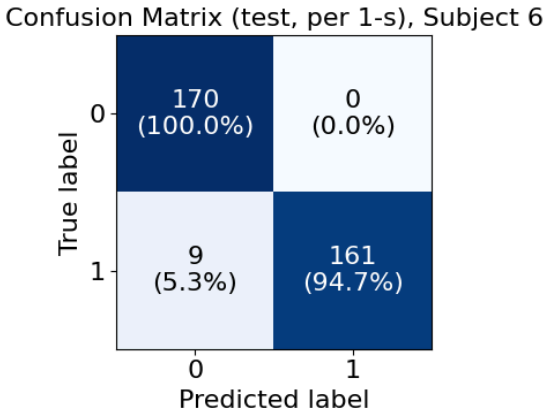
(c)


Confusion Matrix (test, per 40-ms)
3964
(93.3%)
286
(6.7%)
123
(2.9%)
4127
(97.1%)
True label
Predicted label
0
1

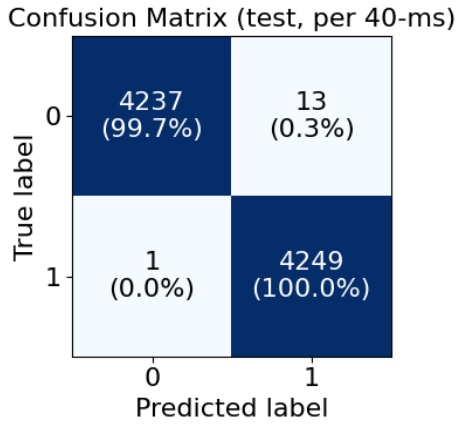

Confusion Matrix (test, per 1-s), Subject 7
170
(100.0%)
0
(0.0%)
0
(0.0%)
170
(100.0%)
True label
Predicted label
0
1

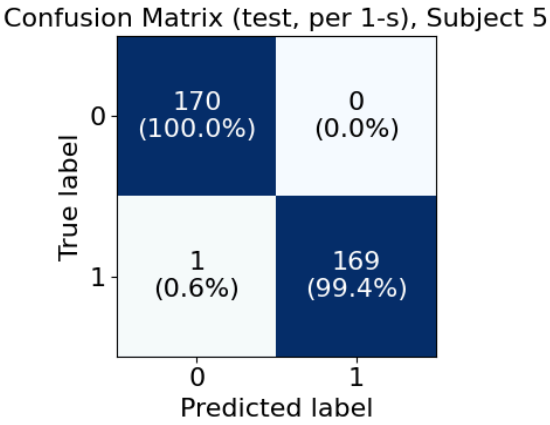
(d)


Confusion Matrix (test, per 40-ms)
4250
(100.0%)
0
(0.0%)
9
(0.2%)
4241
(99.8%)
True label
Predicted label
0
1

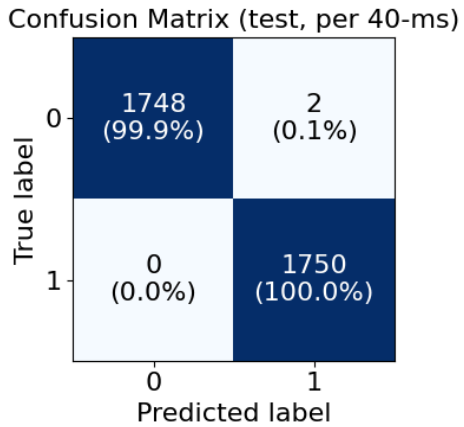

Confusion Matrix (test, per 1-s), Subject 6
170
(100.0%)
0
(0.0%)
9
(5.3%)
161
(94.7%)
True label
Predicted label
0
1

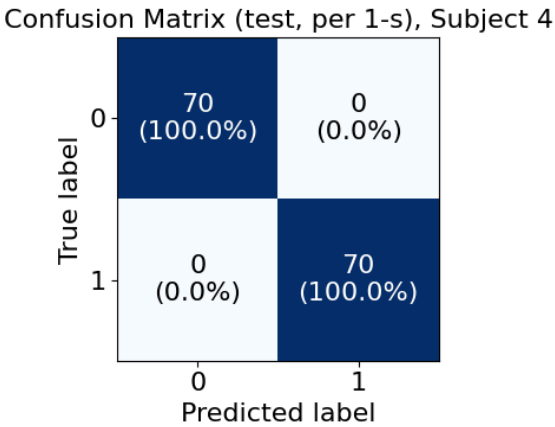
(e)


Confusion Matrix (test, per 40-ms)
4237
(99.7%)
13
(0.3%)
1
(0.0%)
4249
(100.0%)
True label
Predicted label
0
1

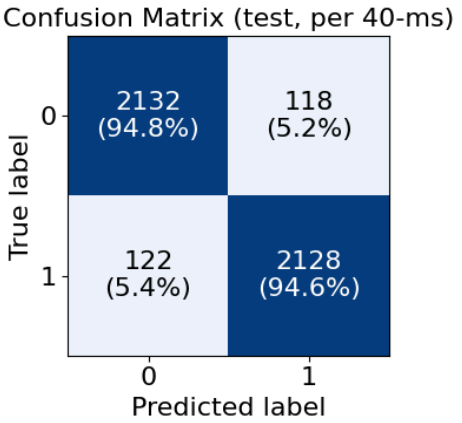

Confusion Matrix (test, per 1-s), Subject 5
170
(100.0%)
0
(0.0%)
1
(0.6%)
169
(99.4%)
True label
Predicted label
0
1

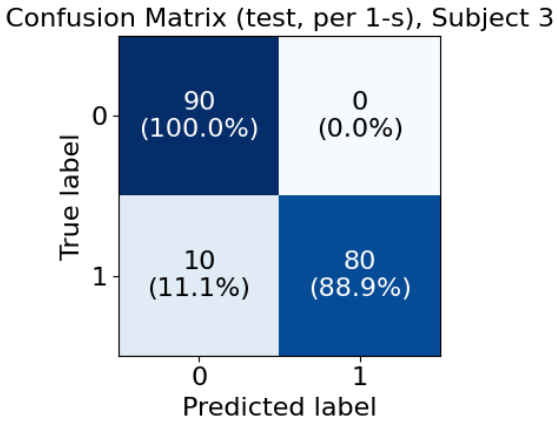
(f)


Confusion Matrix (test, per 40-ms)
1748
(99.9%)
2
(0.1%)
0
(0.0%)
1750
(100.0%)
True label
Predicted label
0
1

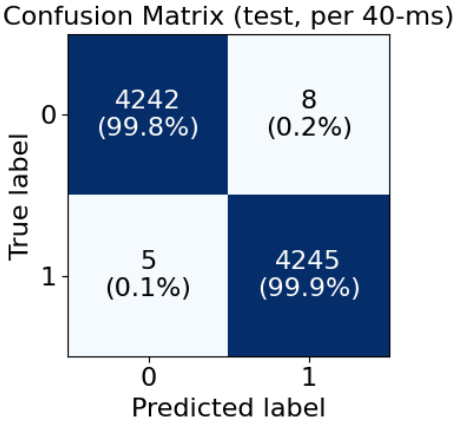

Confusion Matrix (test, per 1-s), Subject 4
70
(100.0%)
0
(0.0%)
0
(0.0%)
70
(100.0%)
True label
Predicted label
0
1

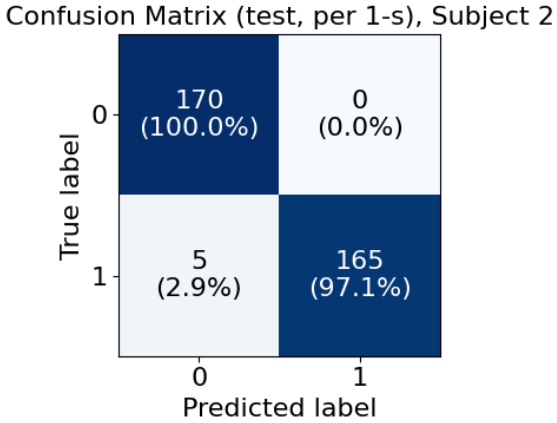
(g)


Confusion Matrix (test, per 40-ms)
2132
(94.8%)
118
(5.2%)
122
(5.4%)
2128
(94.6%)
True label
Predicted label
0
1

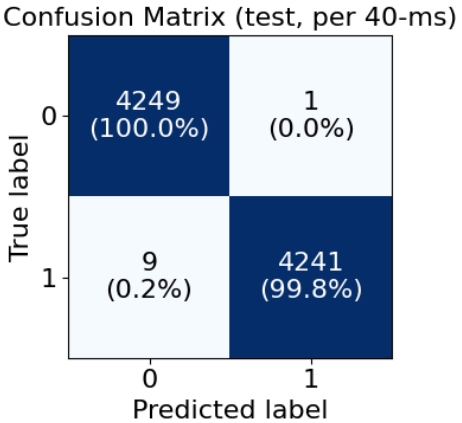

Confusion Matrix (test, per 1-s), Subject 3
90
(100.0%)
0
(0.0%)
10
(11.1%)
80
(88.9%)
True label
Predicted label
0
1

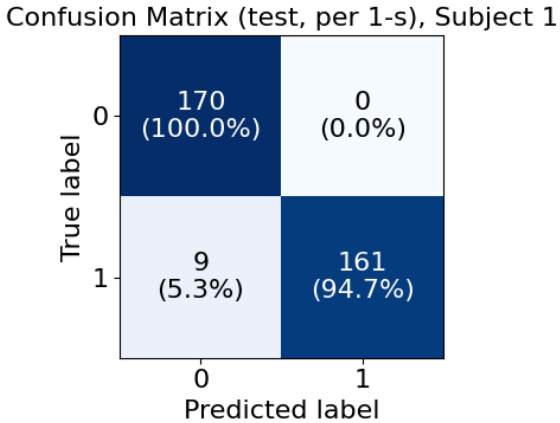
(h)


Confusion Matrix (test, per 40-ms)
4242
(99.8%)
8
(0.2%)
5
(0.1%)
4245
(99.9%)
True label
Predicted label
0
1


Confusion Matrix (test, per 1-s), Subject 2
170
(100.0%)
0
(0.0%)
5
(2.9%)
165
(97.1%)
True label
Predicted label
0
1

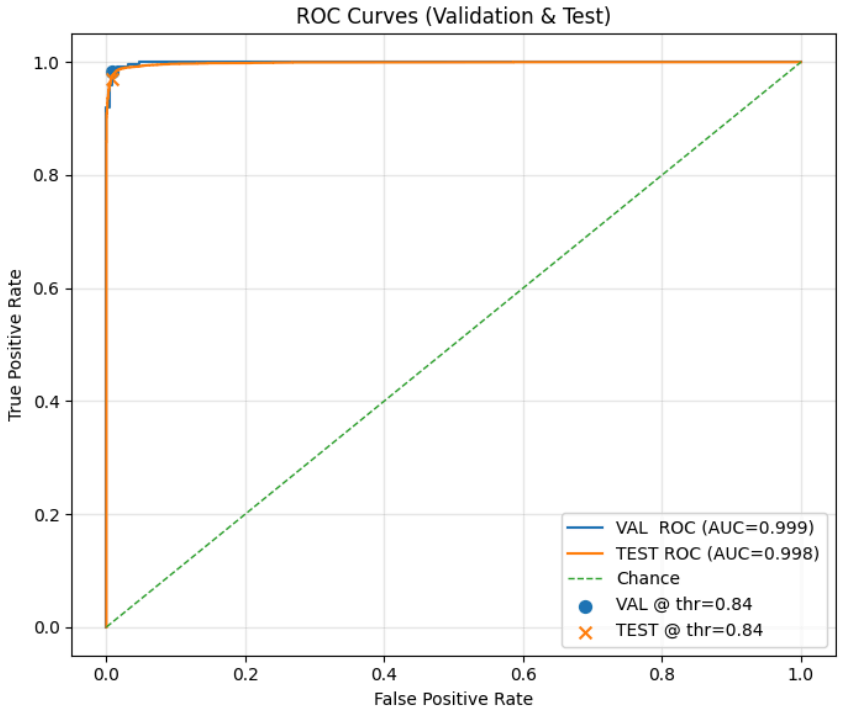
(i)


Confusion Matrix (test, per 40-ms)
4249
(100.0%)
1
(0.0%)
9
(0.2%)
4241
(99.8%)
True label
Predicted label
0
1

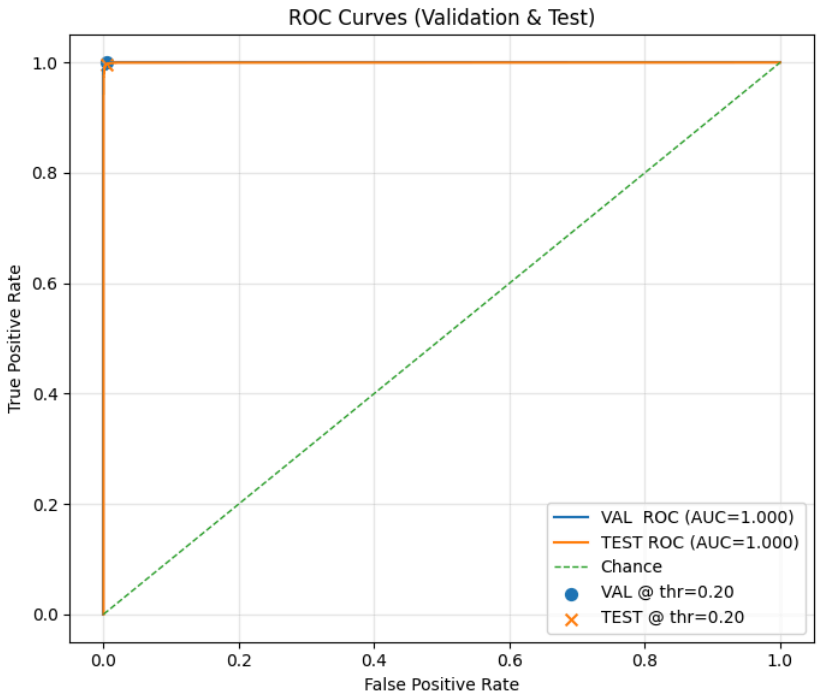

Confusion Matrix (test, per 1-s), Subject 1
170
(100.0%)
0
(0.0%)
9
(5.3%)
161
(94.7%)
True label
Predicted label
0
1

(j)

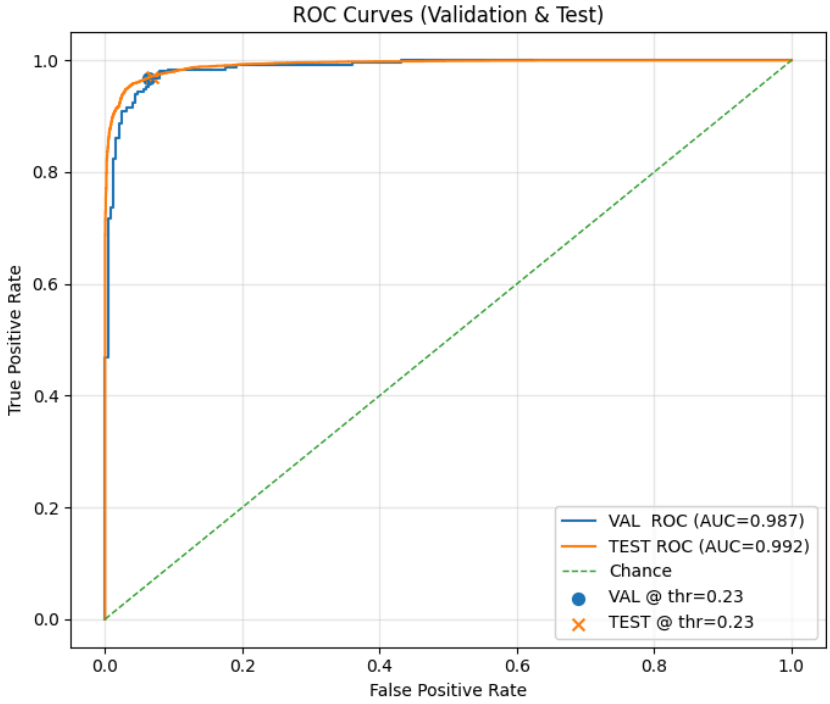
Supplementary Fig. 6: (a–j) Confusion matrices for each split, where the held-out subject is indicated. Calibration experiment using the continuous normalized approach, showing 40-ms and 1-s confusion matrices per split (see Table 2). Thresholds were selected on the validation set using ROC-optimized criteria. Class 0 corresponds to a silent "yes" and class 1 to a silent "no".



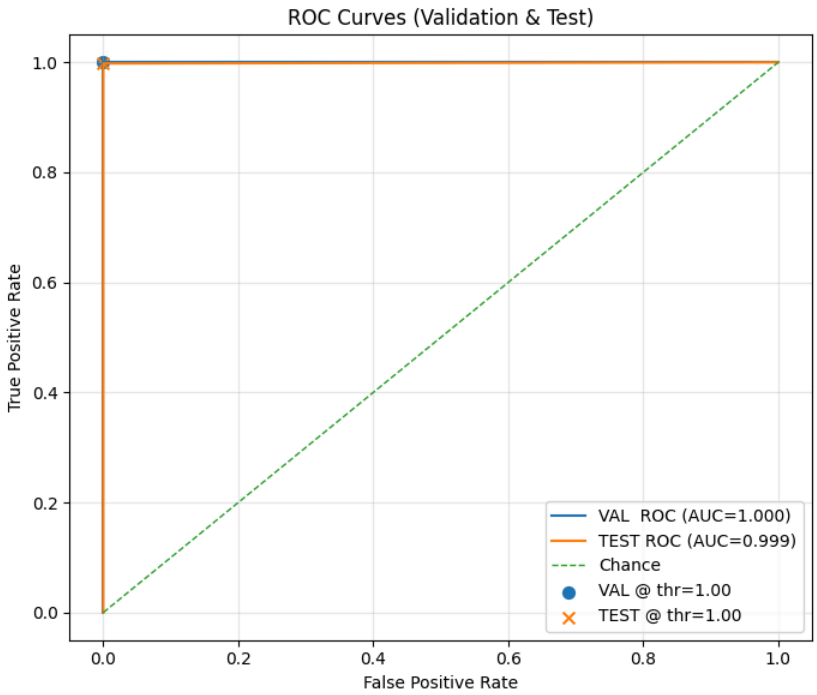
(a) Split 2: Tested on Subject 9



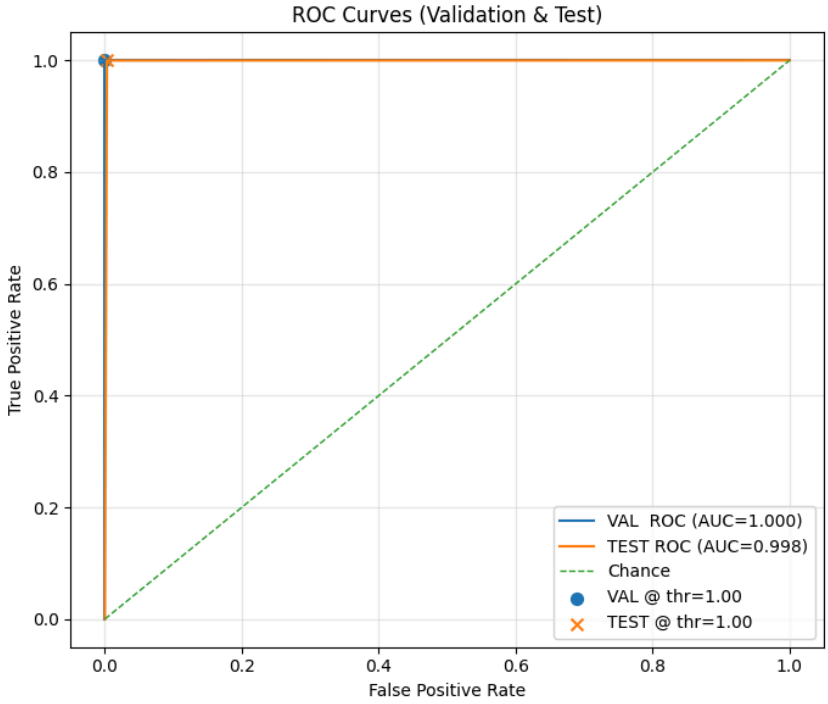
(b) Split 3: Tested on Subject 8



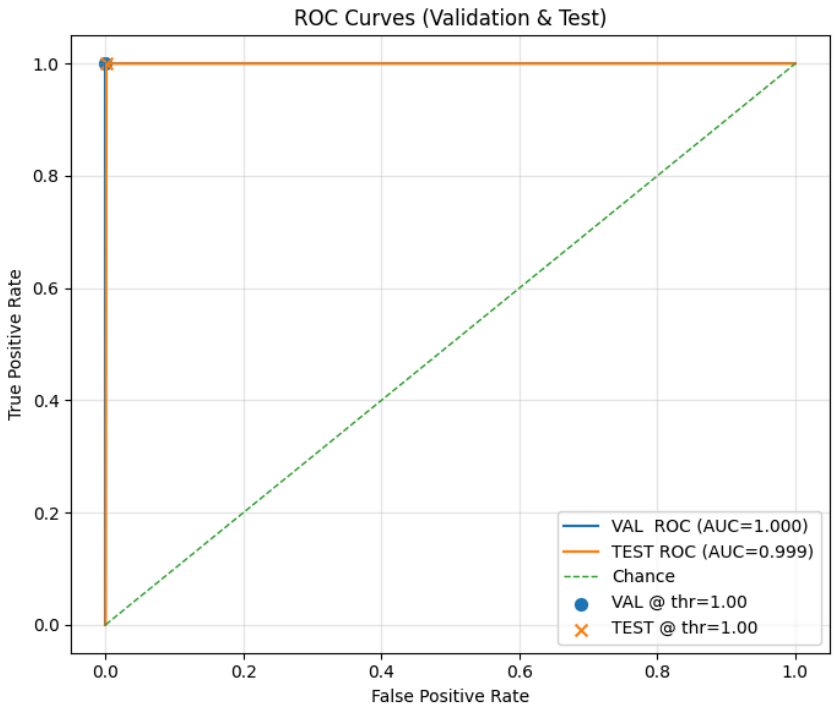
(c) Split 4: Tested on Subject 7



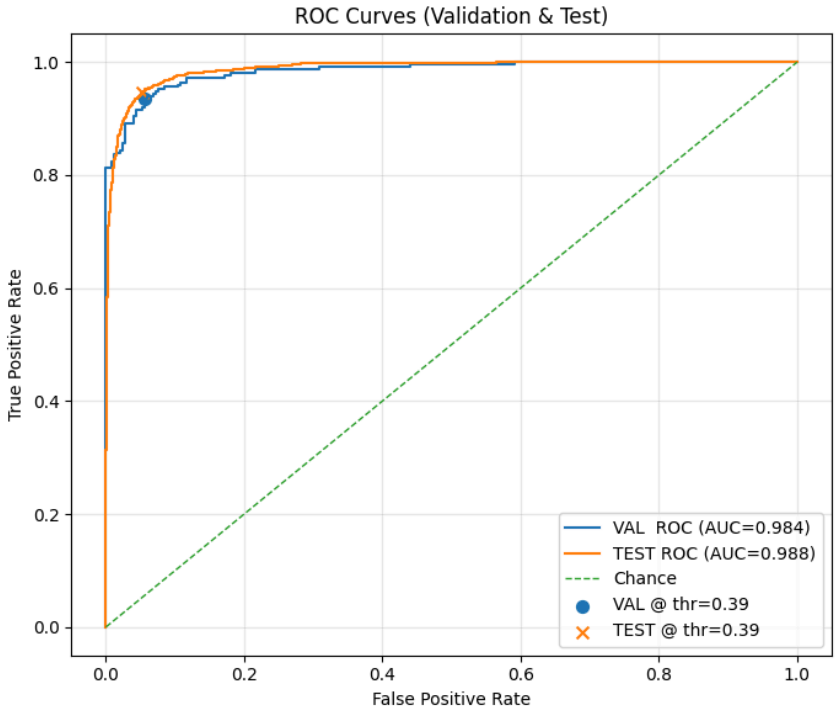
(d) Split 5: Tested on Subject 6



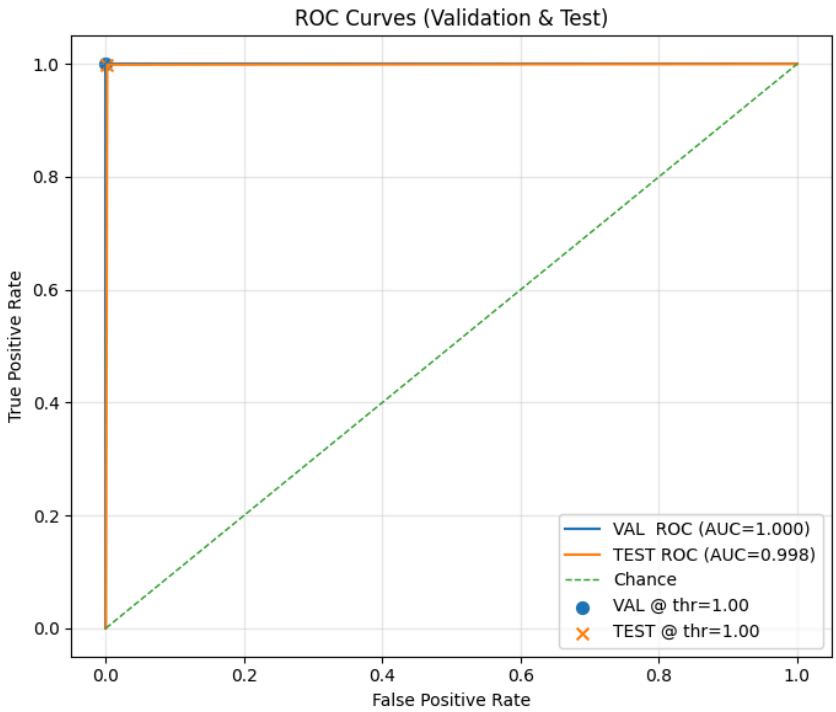
(e) Split 6: Tested on Subject 5



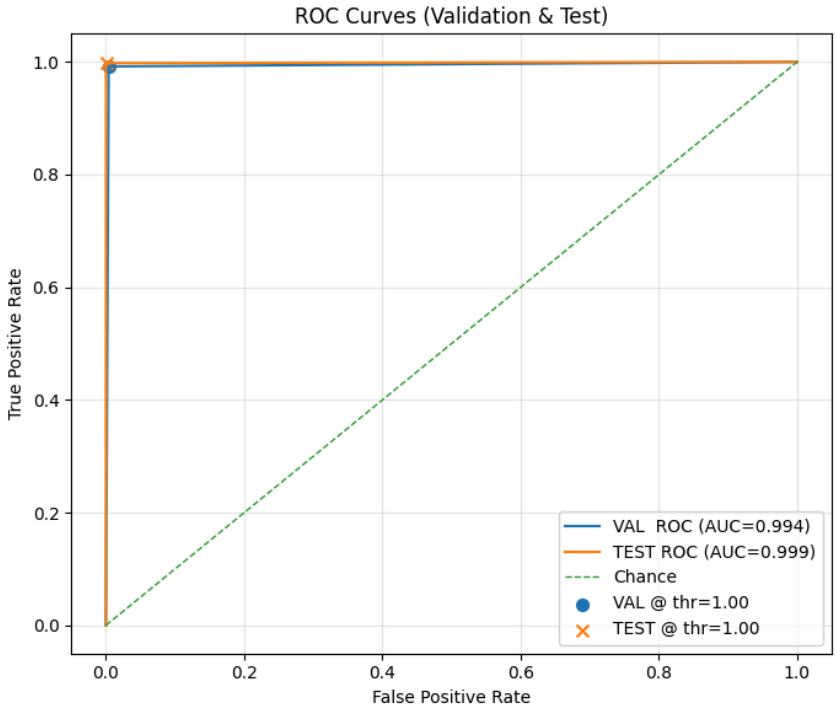
(f) Split 7: Tested on Subject 4



(g) Split 8: Tested on Subject 3

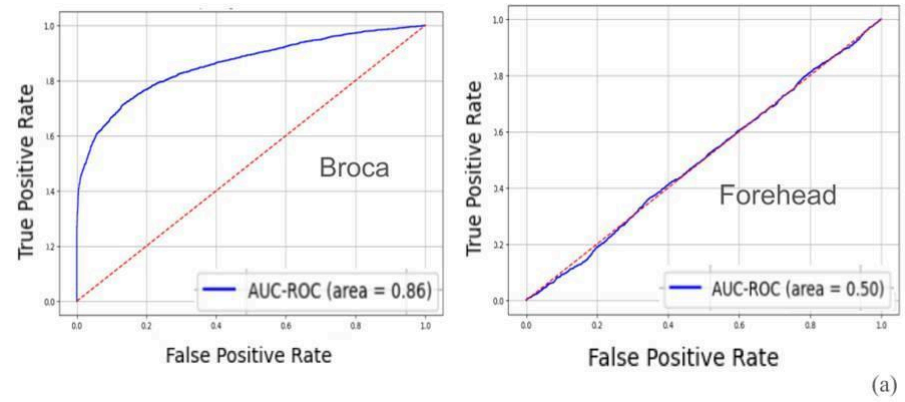


(h) Split 9: Tested on Subject 2



(i) Split 10: Tested on Subject 1

Supplementary Fig. 7: Calibration experiment, AUC-ROC curves for the validation and test sets using the sequential normalized calibration (Table 2), evaluated on 40-ms inputs across splits 2-10, for split 1, see Supplementary Fig. 4 (a).



Supplementary Fig. 8: Location specificity control using forehead recordings. ROC curves for subject 8 comparing decoding performance from Broca's area versus the forehead. (a) For 40-ms chunks, AUC = 0.86 on Broca's area, whereas forehead recordings remain at chance (AUC = 0.50). (b) After temporal aggregation over 1-second, AUC = 1 on Broca's area, while the forehead remains near chance (AUC = 0.47).

Supplementary Tables

Supplementary Table 1:

Results of the classifier on the control validation set on Broca's area, including metrics per subject. Training was performed with an A100 GPU. Metrics per chunk (40-ms input duration) are shown on the left, and metrics aggregated over 1-s on the right. Mean: AUC = 0.87, accuracy = 81.7% for 40 ms of input → AUC = 0.96, accuracy = 96% for 1 s of input.

			Control validation set results on Broca's area								
			40-ms of input					1-s of input			
Subject #	Training set protocol	epochs #	chunks #	AUC	Acc (%)	K	F1	AUC	Acc (%)	K	F1
1	cont	50	3000	1.00	99.9	0.99	0.999	1.000	99.9	0.98	0.990
2	cont	120	1000	0.93	85.7	0.71	0.860	1.000	97.5	0.95	0.970
3	cont	50	2500	0.94	86.4	0.73	0.870	1.000	99.5	0.99	0.995
4	cont	200	2500	0.97	92.6	0.85	0.930	1.000	99.0	0.98	0.990
5	sparse	150	3000	0.53	53.0	0.05	0.510	0.650	65.0	0.30	0.640
5 (5_2)	cont	-	5000	0.67	62.0	-	-	0.990	96.0	-	-
6	cont	60	5000	0.90	84.7	0.69	0.860	1.000	99.5	0.99	0.995
7	cont	70	5000	0.88	79.7	0.60	0.800	1.000	99.5	0.99	0.995
8	cont	60	5000	0.85	79.0	0.58	0.770	1.000	99.5	0.99	0.995
9	cont	40	5000	0.93	85.9	0.72	0.860	1.000	99.5	0.99	0.995
10	cont	20	5000	0.73	69.9	0.40	0.730	0.999	98.0	0.96	0.980

K = Cohen's Kappa, cont = continuous, sparse = sparse cue protocol

Supplementary Table 2:

Performance metrics for two training samples from Subject 5. The first sample ("5_old") was recorded using a sparse-cue protocol, while the second ("5_new" = 5_2 sample, see Table 3) followed the standard continuous-response protocol. The model trained on the standard protocol sample achieved markedly better performance on the test set. Evaluation on a control sample recorded from the forehead (same day, same subject) yielded near-chance results in both cases, suggesting that classification was not driven by muscle activity.

Subject #	Test set results on Broca's area				Forehead (control)	
	40-ms of input		1-s of input		1-s of input	
	Acc (%)	AUC	Acc (%)	AUC	AUC	Acc (%)
5_old	53.0%	0.53	65.0%	0.65	0.48	55.0%
5_new (5_2)	62.0%	0.67	96.0%	0.99	0.37	50.5%

Supplementary Table 3:

Comparison of decoding performance in two recordings from Subject 7. In the initial sample (7_m), the subject self-reported subtle tongue movements during internal speech. To control for this potential artifact, a second sample (7) was collected under stricter immobility conditions. The table shows classification metrics for Broca's area and corresponding forehead-based control recordings. The 7_m sample exhibited comparable performance between Broca and the forehead. In contrast, the new sample (7) demonstrated high Broca-specific decoding with minimal signal detected from the forehead, supporting a cortical origin of the observed effect.

	Results on the control validation sets					
	Broca's area				forehead (control)	
Subject #	Cohen Kappa	F1	Acc (%)	AUC	AUC	Acc (%)
7	0.99	0.995	99.5	1.000	0.000	50.0
7_m	0.91	0.955	95.5	0.994	0.990	95.5

Supplementary Table 4: Generalization performance, demographics test on Subject 10

For generalization without calibration experiment, additional exploratory analysis was performed. Split d (d stands for demographics) includes only female subjects (3, 7) data in the XGBoost training set. The same LV-MAE autoencoder was used for both split 1 and split d. The same validation set was used in both cases.

s #	subj in test	Norm on XGBoost train set	Norm on XGBoost val set	Broca's area								
				40-ms of input				1-s of input (40x25 frames)				
				AUC	Acc %	F1	N in test *	AUC	Acc %	F1	K	N in test
1	10	no	yes	0.57	59.01	0.59	9000	0.77	70.56	0.71	0.41	360
d	10	yes	yes	0.66	61.24	0.61	9000	0.97	92.50	0.93	0.85	360

s# stands for split number.

Supplementary Table 5:

Extended LV-MAE training improves performance for Split 1.

Continuing LV-MAE autoencoder training to 50 epochs yielded higher decoding accuracy and AUC compared with the 30-epoch baseline. The XGBoost classifier was retrained using the same calibration protocol as before, demonstrating that longer representation learning enhances downstream classification performance. Thresholds are selected on validation sets as ROC-optimal. Min-max normalization with a gain of 10 applied to 40-ms chunks prior to XGBoost.

split #	Broca's area				
		40-ms of input, normalized.			
	Test subj	AUC	Accuracy %	F1	LV- MAE epochs
1	10	0.9518	88.13	0.88	50
		0.7647	69.81	0.7	30
2	9	0.9991	98.79	0.99	50
		0.9985	98.15	0.98	30
3	8	0.9999	99.44	1	50
		0.9999	99.62	0.99	30
4	7	0.9990	98.59	0.99	50
		0.9924	95.19	0.95	30

Supplementary Table 6:

Statistics

		20-s				40-ms			
subj	class	Mean	SD	min	max	Mean	SD	min	max
1	0 - yes	2.931	0.086	2.623	3.271	2.931	0.086	2.623	3.271
	1 - no	2.291	0.093	1.951	2.598	2.291	0.093	1.951	2.598
2	0	1.583	0.039	1.442	1.730	1.583	0.039	1.442	1.730
	1	1.982	0.080	1.683	2.252	1.982	0.080	1.683	2.252
3	0	3.419	0.060	3.229	3.643	3.419	0.060	3.229	3.643
	1	3.454	0.056	3.258	3.663	3.454	0.056	3.258	3.663
4	0	1.695	0.103	1.350	2.041	1.695	0.103	1.350	2.041
	1	2.978	0.127	2.613	3.402	2.978	0.127	2.613	3.402
5	0	1.999	0.131	1.588	2.452	1.999	0.131	1.588	2.452
	1	3.253	0.154	2.681	3.814	3.253	0.154	2.681	3.814
6	0	2.172	0.118	1.721	2.609	2.172	0.118	1.721	2.609
	1	2.703	0.059	2.475	2.899	2.703	0.059	2.475	2.899
7	0	3.359	0.121	2.885	3.800	3.359	0.121	2.885	3.800
	1	3.431	0.148	2.924	4.066	3.431	0.148	2.924	4.066
8	0	2.533	0.181	1.925	3.145	2.533	0.181	1.925	3.145
	1	2.070	0.116	1.684	2.541	2.070	0.116	1.684	2.541
9	0	3.134	0.155	2.544	3.712	3.134	0.155	2.544	3.712
	1	3.145	0.170	2.461	3.724	3.145	0.17	2.461	3.724
10	0	2.770	0.228	2.027	3.557	2.770	0.228	2.027	3.557
	1	2.666	0.282	1.775	3.831	2.670	0.282	1.780	3.830

Supplementary Table 7:

Classification performance with calibration on 20-second clips prior to XGBoost. For comparison. A mean AUC of 0.974 ± 0.069 and a mean accuracy of 95.69 ± 8.84% for 40-ms inputs, and a mean AUC of 1 with a mean accuracy of 97.76 ± 1.68% for one-second aggregation (10-fold cross-validation on 3,180 s of balanced recordings). Thresholds calculated in validation sets are ROC-optimal.

Split #	Subject In test	Broca's area, Normalization per 20-s clip								
		40-ms of input				1-s of input (40x25 frames)				
		AUC	Acc %	F1	N in test *	AUC	Acc %	F1	K	N in test
1	10	0.767	69.80	0.70	8500	1.000	96.18	0.96	0.92	340
2	9	0.999	98.15	0.98	8500	1.000	97.06	0.97	0.94	340
3	8	1.000	99.62	1.00	8500	1.000	99.41	0.99	0.99	340
4	7	0.992	95.19	0.95	8500	1.000	100.00	1.00	1.00	340
5	6	0.999	99.89	1.00	8500	1.000	97.35	0.97	0.95	340
6	5	0.999	99.89	1.00	8500	1.000	97.35	0.97	0.95	340
7	4	0.999	99.94	1.00	3500	1.000	100.00	1.00	1.00	140
8	3	0.988	94.67	0.95	4500	1.000	94.40	0.94	0.89	180
9	2	0.999	99.85	1.00	8500	1.000	98.53	0.99	0.97	340
10	1	0.999	99.88	1.00	8500	1.000	97.35	0.97	0.95	340

In the table, "**N in test ***" stands for the number of 40-frame chunks in the test.

Supplementary Results

Exploratory cross-session analysis for temporal stability

As an exploratory analysis, we evaluated cross-session generalization in subject 2 using a second recording (session 2_2; 3,000 balanced chunks) acquired approximately one month later, under slightly altered experimental conditions (minor differences in the relative angles between the camera, laser, and scalp). An XGBoost classifier trained on the earlier session achieved an AUC of 0.92 at the 40-ms chunk level, which increased to 1.0 after 1-s temporal aggregation, indicating that rank ordering was preserved across sessions. At the 40-ms scale, accuracy was 65% without threshold calibration. After 1-s aggregation, accuracy increased to 87.1% following threshold calibration based on a single 1-s chunk per class from the target session.

Supplementary Theory

Metrics:

Below, we formally define accuracy (Equation 1), F1 (Equation 2), and Cohen's kappa (Equations 8-9):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

Where:

Let yi denote the true class label for sample i. We use the standard classification terminology: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$TP = \sum_i (p(xi) == 1) \; and \; (yi == 1) \quad (3)$$

$$TN = \sum_i (p(xi) == 0) \; and \; (yi == 0) \quad (4)$$

$$FP = \sum_i (p(xi) == 1) \; and \; (yi == 0) \quad (5)$$

$$FN = \sum_i (p(xi) == 0) \; and \; (yi == 1) \quad (6)$$

$K = \frac{P_0 - P_e}{1 - P_e}$, P_0 is accuracy, and is an P_e expected agreement by chance. (8)

$$P_e = \sum_i^N p_{true_i} * p_{predicted_i} \quad (9)$$

Cohen's kappa (κ) quantifies the agreement between predicted and true labels while correcting for agreement that would be expected by chance (Pe). Its values range from -1 (systematic disagreement) to 1 (perfect agreement), with 0 indicating chance-level performance. It is computed by comparing the observed agreement (Po) to the expected agreement (Pe), normalized by the maximum possible agreement beyond chance.[30] In Equation 9, N denotes the number of classes, p_{true_i} represents the proportion of class i in the true labels, and $p_{predicted_i}$ denotes the proportion of class i in the predicted labels.