EEG-sAAD Grand Challenge Onboarding guide to the training data

What is in the five training datasets, what our two exploratory analyses found, which traps to avoid, and how to start building a decoder.

Who this is for

New team members joining our ICASSP 2027 EEG-sAAD (selective auditory attention decoding) challenge work. You need basic machine-learning and signal-processing background; EEG-specific terms are explained in the glossary (Appendix A).

What this document combines

1. Corpus summary (reports/eda_raw): inventory and integrity check of all 1,979 trials. **2. Example-trial EDA** ('EEG-sAAD Data Story'): a deep look at ONE trial per dataset. It produced twelve hypotheses (H1-H12) and a method roadmap. **3. Full-corpus EDA** (reports/eda_full): every trial of every subject, used to confirm or refute H1-H12 with subject-level statistics. Where the two EDAs disagree, **the full-corpus result wins**; example-trial numbers appear only as 'first estimate'.

How to use it

Read **Part 1 (Start here)** first: it takes ten minutes and lists what everyone on the team must know. Parts 2-6 are reference material. Part 7 tells you where the files are and how to run things. Every number comes from code that was run on the data; the source tables are named so you can check them.

Built 2026-09-25 by reports/onboarding/build_guide.py from reports/eda_full/tables, reports/eda_raw and 'EEG-sAAD Data Story.html'.

Contents

Part	Topic
1	Start here: twelve things to know
2	The challenge in two pages: task, rules, metric, baseline
3	The data: datasets, files, labels and conventions
4	How we built our understanding: from example-trial hypotheses to full-corpus verdicts
5	What the data show: dialects, labels, the attention signal, shortcuts, switch detection
6	What to do about it: recommended pipeline, validation checklist, roadmap, open questions
7	Getting started: environment, directory map, commands
A	Glossary
B	References (literature cited by the example-trial EDA)

1. Start here: twelve things to know

- **The task.** Two people talk at once; the listener attends to one. From the listener's EEG (plus the two speech envelopes) we must output, for every 128 Hz sample, which talker (or which side) is attended. Attention may switch within a trial. The test subjects and recording labs are unseen.
- **The hard rule.** A label at time t may use data only up to t + 1 s, counting every filter, lag, window and smoother. Zero-phase filtering (filtfilt) breaks this rule. Subject, trial and dataset IDs must never be model inputs.
- **The score.** H = harmonic mean of sample accuracy (Acc) and a switch-latency score L = median of exp(-SDT/60). The published baseline scores Acc 68.4%, SDT 18.53 s, H 0.7002 on the KUL-AV-GC hold-out.
- **The data.** Five datasets: 95 subjects, 1,979 trials, 77 h of EEG at 128 Hz (Part 3). They come from different labs, amplifiers, languages and montages; only **57 channel names** are common to all five.
- **'Raw' means five different things.** The amplitude scale differs 16.5x between datasets; DTU, KUL-UHD and USTC carry DC offsets of thousands of units; every KUL-UHD and USTC trial starts and ends with a large jump. Normalise causally before pooling (Part 5.1).
- **There is no speaker switch in the training data.** The only genuine switches are the 99 KUL-AV-GC side switches at exactly t = 300 s, where the SAME talker moves to the other side. Speaker-switch detection can only be trained and tested with synthetic switches (Part 5.2).
- **Never use the trial-level label for direction.** It is the side at trial onset, so it is wrong for 50% of KUL-AV-GC samples. Use the per-sample CND labels (attendedStream, attendedAzimuth).
- **The attention signal is a whisper.** Across datasets, 1-s decisions are right about 53-56% of the time. Accuracy grows predictably with window length: accuracy is about Phi(0.9 x SNR1 x sqrt(w)), where SNR1 is roughly 0.07-0.13 across datasets and 0.14-0.20 within a subject (Part 5.3).
- **Shortcuts are everywhere.** Labels follow trial parity (KUL-UHD, USTC); the attended envelope is louder in KUL2016 (a loudness-only classifier gets 82.5%); covariances identify the dataset at 99.7%; spatial decoders memorise trials. Validate so that none of these can help (Part 5.4).
- **Spatial (direction-only) decoders do not transfer.** Across datasets they are at chance (47-54%); on the 99 genuine KUL-AV-GC switches the best one reaches 51.7% sample accuracy.
- **The scorer has quirks.** A PERFECT decoder is charged a missed switch when the next switch reverts within 50 s, and a decoder with zero information still earns L of about 0.45-0.73. Always report L next to its null (Part 5.5).
- **Causal preprocessing costs nothing.** Causal filtering is never worse than zero-phase filtering or the organisers' preprocessed 32 Hz data. Start from the pipeline in Part 6.

2. The challenge in two pages

2.1 Task and test sets

Selective auditory attention decoding (sAAD): a listener hears two simultaneous talkers, one on each side, and attends to one of them. The EEG follows the slow loudness contour (envelope) of the attended talker slightly more closely than that of the ignored one. A decoder exploits this difference, or other spatial cues, to say who (or which side) is attended, and must keep doing so when attention switches.

The blind test sets are **KUL-DLAAD** (BioSemi, Dutch) and **NJU-SUSTech-sAAD-GC** (NeuroScan, Mandarin). Both have unseen subjects, attention switches at random times within trials, and eye-gaze control, so the eyes cannot give the attended side away. The example-trial EDA quotes the challenge page: KUL-DLAAD has 19 subjects with 8 trials and about 50 min of EEG each; NJU-SUSTech has 12 subjects with about 60 min each. The switch rate and cueing are not published.

2.2 Rules that shape every design decision

Table 2.1. Challenge rules (from the challenge prompt and the organisers' baseline README).

Rule	What it means
Lookahead	A label at time t may use data only up to t + 1 s. This budget covers every filter, decoder lag, window and smoother (Part 6 shows how our pipeline spends it).
Output	One label per 128 Hz sample: attended stream (0 = first, 1 = second) or attended side (0 = left, 1 = right).
Identity	Dataset, subject and trial IDs must never be predictive inputs; no statistics pooled across a subject's test trials.
Adaptation	Test-time updates are allowed only within a trial and only causally.

2.3 The metric

For each ground-truth switch at t0 the scorer searches [t0, t0 + 100 s] for the split point tau* that best explains the predictions as 'old label, then new label'. The detection counts only if the 0.5 s after tau* all carry the new label; otherwise the search moves on, and a switch never confirmed scores 0. SDT = tau* - t0 and the per-switch score is exp(-SDT/60). L is the median over a subject's switches, averaged over subjects; Acc is the fraction of correct samples, averaged over subjects; **H = 2 Acc L / (Acc + L)**.

Table 2.2. The latency score is lenient: even 18.5 s keeps 73% of the credit.

SDT	exp(-SDT/60)
5 s	0.92
10 s	0.85
18.5 s (baseline)	0.73
60 s	0.37
missed	0

2.4 The official baseline

baselines/matlab_linear trains one backward (stimulus-reconstruction) decoder on the pooled training datasets, decodes 1-s windows on a held-out dataset and smooths the decisions with a causal hidden Markov model (HMM). On the KUL-AV-GC hold-out it reaches 53.94% at 1 s and 73.20% at 50 s without smoothing, and **Acc 68.4%, SDT 18.53 s, H 0.7002** with the HMM. The baseline also contains generateSyntheticSwitches.m, which imposes switches on fixed-attention trials by swapping the two streams' roles. The official scorer is baselines/matlab_scoring. MATLAB is not installed here, so we use a verified Python port (reports/eda_full/scripts/scorer.py; 34/34 hand-made tests plus 3,000 random cases).

3. The data

3.1 The five training datasets

*Table 3.1. Training corpus (counts from reports/eda_full/tables/corpus_inventory.csv). *The challenge page lists KUL-UHD as BioSemi/255 channels; the CND files say SynAmps RT; 253 channels are stored after dropping two extra electrodes.*

dataset	subj.	trials	hours	trial (s)	ch.	amplifier	lang.	azimuth	notes
KUL-AV-GC	13	99	16.5	600	64	BioSemi ActiveTwo	Dutch	±90°	audiovisual, gaze-controlled; side switch at 300 s
DTU	18	1080	15.0	50	64	BioSemi	Danish	±60°	reverberation conditions
KUL2016	16	320	20.5	124	64	BioSemi	Dutch	±90°	dry vs HRTF rendering
KUL-UHD	30	120	13.0	391	253	SynAmps RT (per CND file)*	Dutch	±90°	ultra-high-density montage
USTC	18	360	12.0	120	64	NeuroScan SynAmps	Mandarin	±90°	

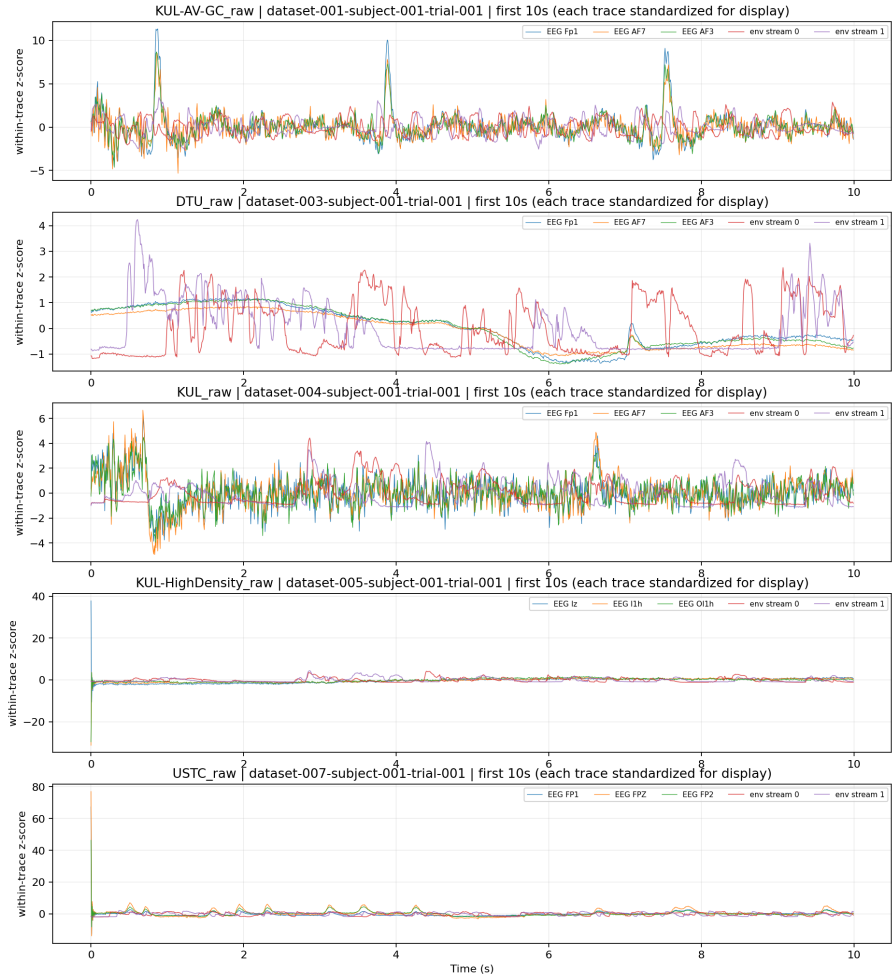


Figure 3.1 (from the corpus summary, reports/eda_raw). The first 10 s of subject 1, trial 1 of each dataset: three EEG channels and the two envelopes, each standardised for display. Note the eye blinks (KUL-AV-GC), the slow drift (DTU) and the huge jump at t = 0 in KUL-UHD and USTC.

3.2 Where the data live and what the files contain

Table 3.2. Directory map of the workspace data.

path	contents
archives/datasets/	original ZIP archives + SHA256.txt (never modify)
data/cnd/<DATASET>/dataCND/dataSub<N>.mat	EEG ('neural'): per trial EEG [time x channels], chanlocs, fs, and PER-SAMPLE ground truth: attendedStream (1-based stream index) and attendedAzimuth (degrees). Some datasets add extChan (EOG), origTrialPosition, deviceName.
data/cnd/<DATASET>/dataCND/dataStim<N>.mat	'stim': envelope_stream1/2 per trial, streamWav (story file names), condIdxs/condNames. KUL-UHD uses one shared dataStim.mat.
data/cnd/.../pre_*.mat	the organisers' preprocessed variant: zero-phase 1-9 Hz band-pass, 32 Hz, plus bad-channel or bad-segment repair
data/superhuge/raw/	the same data converted for the Python DNN baselines: eeg/dataset-DDD-subject-SSS-trial-TTT.npy (time x channels, 128 Hz), stimuli/env/..._env.npy (time x 1 x 2 streams), meta/metadata.pkl
data/superhuge/preprocessed/	same layout for the pre_* (32 Hz) files
baselines/	matlab_linear (decoder + HMM), matlab_scoring (official scorer), python/SuperHugeAAD (DNNs)
tools/convert_cnd_to_superhuge.py	the CND -> SuperHuge converter; its conventions are listed below

3.3 Conventions that trip people up

- **SuperHuge column 0 = the stream attended for most of the trial.** This is a training convention; it will not hold at test time, and it silently leaks the answer if a model can see column order. Columns were checked against the CND files for all 1,979 trials (EEG identical, envelope r = 1.000).
- **The metadata 'label' is trial level:** the attended side at trial onset ('left'/'right' or 'L'/'R'). It is not a per-sample ground truth. Per-sample labels exist only in the CND files.
- **Dataset IDs differ between systems:** the converter numbers datasets 1 (KUL-AV-GC), 3 (DTU), 4 (KUL2016), 5 (KUL-UHD) and 7 (USTC). Subjects are renumbered 1..N in file order, except USTC; KUL-UHD files start at dataSub0.
- **Channels must be matched by upper-cased name, never by index.** The BioSemi sets spell 'Afz'; USTC uses upper case (FP1, FPZ) and adds M1, M2, CB1, CB2, PO5, PO6; KUL-UHD uses 10-5 names. The common montages are 57 channels (all five datasets), 58 (BioSemi-64 and USTC, identical to the baseline's list) and 62 (BioSemi-64 and KUL-UHD).
- **Story names are unreliable.** In KUL-AV-GC, 29 of 30 distinct audios carry more than one streamWav name. All KUL-UHD audio also occurs in KUL2016. Group folds by envelope identity (tables/stim_identity_clusters.csv), not by file name.

4. How we built our understanding

Three analyses were run, each on a larger evidence base than the last. The example-trial EDA was fast and suggestive; the full-corpus EDA was designed to test its claims at scale, with statistics over subjects.

Table 4.1. The three analyses behind this guide.

analysis	where	evidence base	main output
Corpus summary	reports/eda_raw	all 1,979 trials, summary statistics	inventory, integrity (0 missing files, 0 non-finite values), durations, label counts, channel statistics
Example-trial EDA	EEG-sAAD Data Story.html	one trial per dataset (subject 1, trial 1) + corpus statistics	12 hypotheses (H1-H12), a literature-backed method roadmap, a re-implemented scorer, synthetic-switch simulations
Full-corpus EDA	reports/eda_full	every trial and subject; subject-level medians with 95% bootstrap CIs, Wilcoxon tests, FDR	verdicts on H1-H12, LOTO/LOSO/LODO decoders, SNR law, shortcut audits, official-scorer port, integrator comparison, lookahead unit test

4.1 The twelve hypotheses and their verdicts

Table 4.2. Hypotheses H1-H12: example-trial estimate vs full-corpus result (details and CIs: reports/eda_full/tables/hypotheses_verdicts.csv).

ID	claim from the example trial	first estimate (n = 1 trial)	full corpus	verdict
H1	Amplitude scales differ ~16x; DTU, KUL-UHD, USTC carry large DC offsets	SD 8.3/39.2/9.6/132.1/61.5; RMS/SD 1/210/1/58/94	Yes: 16.5x; offsets of thousands of units in DTU, KUL-UHD, USTC.	confirmed
H2	Different pre-filtering; strong 50 Hz in KUL-UHD and USTC	line ratio 10.8 (UHD), 9.3 (USTC)	KUL-AV-GC is band-limited 1-41 Hz and notched; the 50 Hz line in KUL-UHD/USTC is real but 3-4x weaker than the example suggested.	partly
H3	Edge transients in USTC and KUL-UHD; bad channels exist	jumps 9,249 / 13,011 units at t=0; UHD P7 SD 2,381 vs Cz 71	Yes: 100% of USTC and KUL-UHD trials vs 3-5% elsewhere; 5.1% of KUL-UHD channels bad.	confirmed
H4	Accuracy = Phi(SNR1 sqrt(w)); SNR1 ~0.16 within subject, ~0.10 for the baseline	54.5% @1 s, 66.1% @10 s, 88.9% @60 s (n=9 windows)	Yes: k = 0.74-0.98; KUL-AV-GC SNR1 0.185 within subject, 0.107 across datasets (baseline implies 0.099).	confirmed
H5	Causal filtering is no worse than zero-phase	within-trial 50-58% @1 s either way	Yes: never worse; zero-phase is slightly worse where significant.	confirmed: causal is never worse; zero-phase is slightly worse where significant
H6	The KUL-AV-GC 5-min switch changes side, not talker	cross-half: switch hyp. 0.0% @60 s; no-switch hyp. 88.9%	Yes: 0 speaker switches, 99 side switches (all at 300 s); cross-half test 84% vs 20% at 60 s.	confirmed
H7	The attended envelope is louder in KUL2016 and KUL-AV-GC	65% of 320 (p=2.3e-9); 59% of 99 (p=0.0095)	KUL2016 yes (65% of trials; loudness-only classifier 82.5%); KUL-AV-GC only weakly.	confirmed (KUL2016); partly (KUL-AV-GC)
H8	Label follows trial parity in KUL-UHD and USTC	60/60 and 180/180	Yes: 100% of trials, every subject.	confirmed
H9	USTC stream columns may be swapped	r = 0.083 vs 0.026	No: the held-out decoder prefers column 0 in 75% of USTC trials.	refuted
H10	Covariance fingerprints the dataset; alignment removes only part	99.7% -> 76.1% (chance 39%)	Yes: 99.7% dataset ID; still 84% after EA and 87% after causal running EA (chance 31%).	confirmed
H11	Spatial decoders exploit trial fingerprints	not tested (literature: CSP trial-ID ~100%, LOTO 41-45% on KUL-AV-GC)	Yes: trial ID 63-89%; -22 pp when trials are held out; chance across datasets.	confirmed
H12	No alpha or EOG marker at the KUL-AV-GC switch	ALI 0.040 -> 0.045 (p=0.77); hEOG p=0.98	Yes (all p >= 0.08), but KUL-AV-GC is high-passed ~1 Hz, so weak evidence of absence.	confirmed

4.2 What changed when we went from one trial to all trials

- **Confirmed and strengthened:** the direction-only KUL-AV-GC switch (now in all 99 trials, from the CND labels themselves), the parity lock, the dataset fingerprint, the SNR law and the harmlessness of causal filtering.
- **Refuted:** the suspected USTC column swap. USTC is simply the hardest dataset (weakest SNR1, edge transients, a different envelope extraction), not a mislabelled one.
- **Weaker than first thought:** the 50 Hz mains line (3-4x smaller), the 'louder attended column' in KUL-AV-GC (not usable by a classifier), and the gain from integrator tuning. On synthetic switches with within-subject evidence, the example trial gave H of about 0.80 for a 5-10 s moving average. With honest cross-dataset evidence the best integrator reaches H 0.63-0.75, only 0.02-0.07 above plain argmax.
- **Worse than first thought:** Euclidean alignment removes less of the fingerprint (84-87% remain vs 76% in the example), and the chance floor of L depends on the dataset (0.45-0.73 rather than 0.49-0.57), because it depends on trial length and switch spacing.
- **New:** the scorer's 50-s property (a perfect decoder can miss a switch), unreliable story names, the fact that all KUL-UHD audio is reused from KUL2016, and that pooling strategy (pooled covariance vs averaged decoders) makes no difference across datasets.

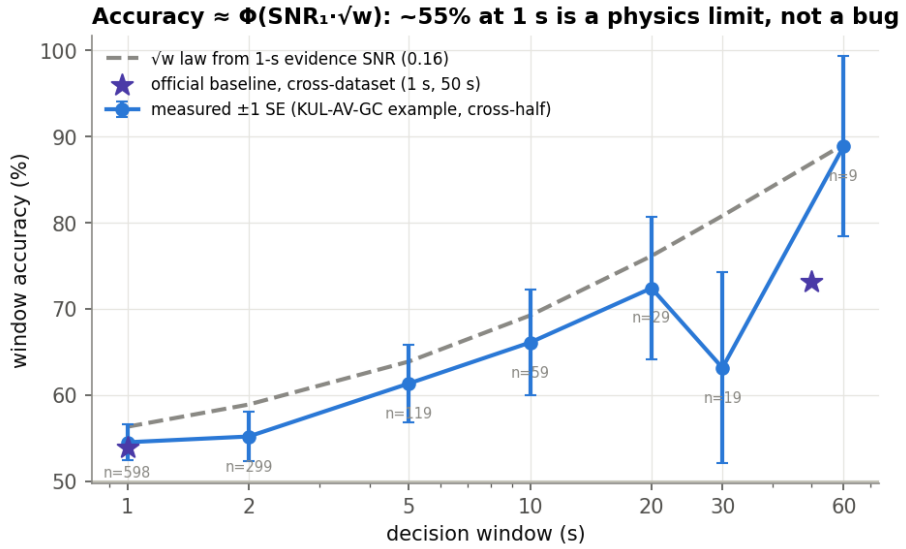


Figure 4.1 (example-trial EDA, n = 1 KUL-AV-GC trial). The SNR law as first seen: window accuracy tracks Phi(0.16 sqrt(w)). Compare with Figure 5.5, the same law confirmed on all 95 subjects. (Colours of example-trial figures follow that document, not the dataset colours used elsewhere.)

5. What the data show

Dataset colours used in all full-corpus figures: **KUL-AV-GC red, DTU blue, KUL2016 green, KUL-UHD purple, USTC orange**; grey marks a baseline or null. Accuracies are medians over subjects with 95% bootstrap CIs.

5.1 Five labs, five dialects of 'raw'

Table 5.1. Recording 'dialects' (subject-level medians over 1,979 trials; stored units). RMS/SD = 1 means no DC offset; 50 Hz ratio > 1 means mains noise, < 1 a notch.

dataset	median ch. SD	RMS / SD	|DC|	high-pass	low-pass	50 Hz ratio	edge-jump trials	bad ch.
KUL-AV-GC	7.96	1	0.001	HP ~1 Hz or higher	LP, roll-off ~40.8 Hz	0.069	3%	0.1%
DTU	39.1	211	8,490	no HP (DC offsets present)	LP, roll-off ~64.0 Hz	1.4	3%	2.3%
KUL2016	9.64	1	0.008	HP < 1 Hz (or none, DC removed)	LP, roll-off ~64.0 Hz	1.1	5%	0.0%
KUL-UHD	131	51.4	7,874	no HP (DC offsets present)	LP, roll-off ~62.8 Hz	3.8	100%	5.1%
USTC	63.6	97.6	5,198	no HP (DC offsets present)	LP, roll-off ~64.0 Hz	2.2	100%	1.4%

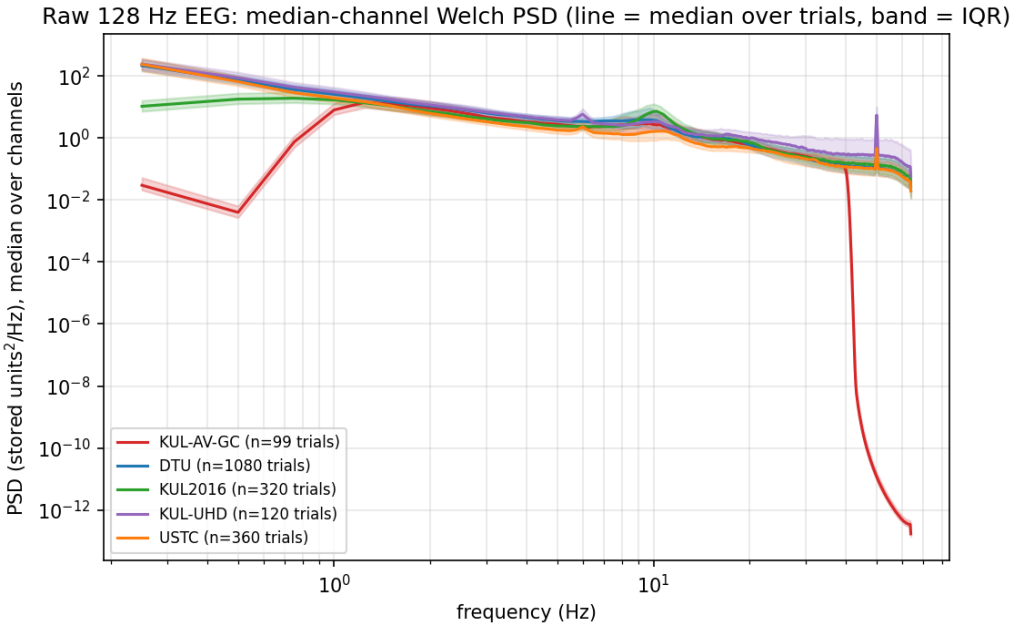


Figure 5.1. Median-channel power spectra (median over trials, IQR band). KUL-AV-GC (red) was band-limited to about 1-41 Hz before release; KUL-UHD and USTC show a 50 Hz line; DTU, KUL-UHD and USTC keep large power below 1 Hz because they were never high-passed.

How to read this

x = frequency, y = power, both logarithmic. Curves that differ in shape mean the datasets were processed differently before release, which a model can use to tell datasets apart.

Consequence: the front end must be causal and per-trial: match channels by name, high-pass with the filter state initialised to the first sample (otherwise the DC offsets ring for seconds), re-reference, band-pass, and scale with a running robust z-score. Trim or warm up the first second, because every KUL-UHD and USTC trial starts with a jump. None of the datasets is stored average-referenced; common-average referencing measurably helps (Part 5.3).

5.2 Labels: there is no speaker switch

Table 5.2. Ground truth from the per-sample CND labels (reports/eda_full/tables/label_audit_summary.csv).

dataset	trials	within-trial speaker switches	within-trial side switches	side samples mislabelled by the trial label
KUL-AV-GC	99	0	99	50%
DTU	1080	0	0	0%
KUL2016	320	0	0	0%
KUL-UHD	120	0	0	0%
USTC	360	0	0	0%

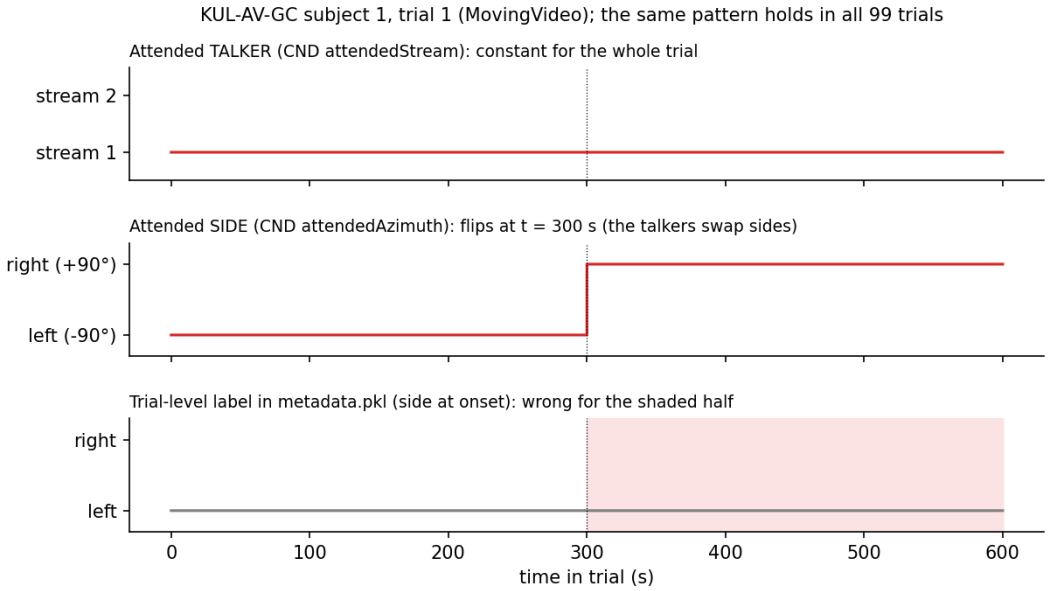


Figure 5.2. What happens in a KUL-AV-GC trial: the attended talker never changes; at t = 300 s the two talkers swap sides, so the attended SIDE flips. The trial-level label is wrong for the second half.

Table 5.3. Cross-half decoding test on all 99 KUL-AV-GC trials (13 subjects): a decoder trained on one half and tested on the other. Accuracy (%) expected under each explanation.

window	if the talker switched at 5 min	if the same talker is attended all trial
1 s	45.4 [44.3, 46.0]	56.6 [55.8, 57.8]
10 s	36.2 [31.9, 39.0]	67.9 [65.9, 71.8]
60 s	20.3 [12.5, 28.1]	84.4 [79.7, 89.1]

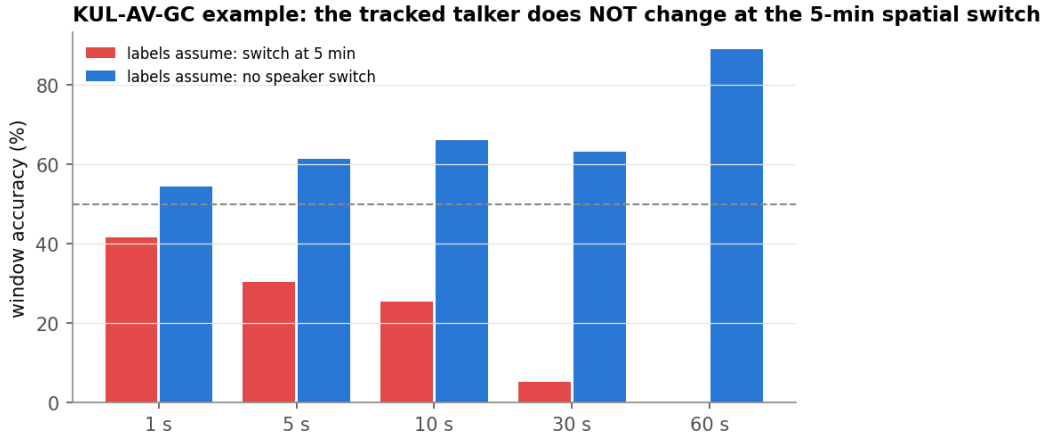


Figure 5.3 (example-trial EDA, n = 1 trial). The same test on one trial first raised the hypothesis; Table 5.3 confirms it on all trials (the full-corpus 60-s values are 84% vs 20%, less extreme than 89% vs 0%).

What this means for you

For a stimulus-reconstruction (speaker) decoder there is no genuine switch anywhere in the training data, so speaker-switch latency can only be measured with SYNTHETIC switches. For a direction decoder, take labels from the per-sample azimuth. The KUL-AV-GC side switches become 'free' for a speaker decoder only if the test data tell us where each stream is; ask the organisers (Part 6.4).

5.3 The attention signal: small, consistent, predictable

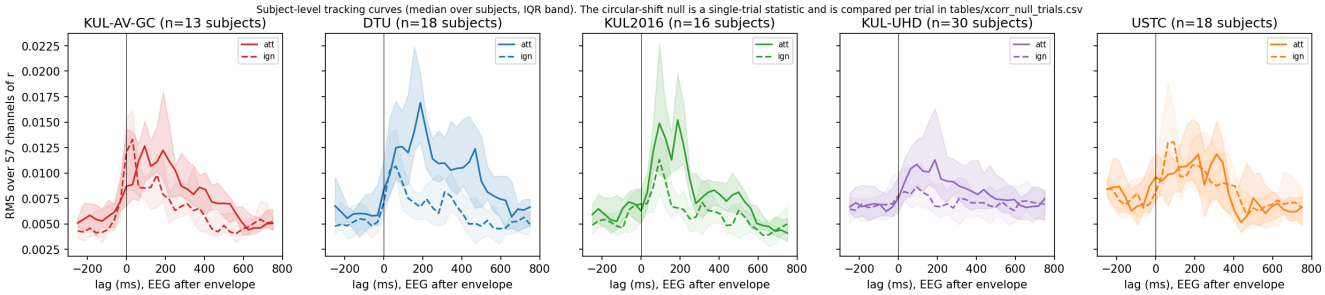


Figure 5.4. Envelope-EEG cross-correlation (RMS over 57 channels, median over subjects). Solid = attended, dashed = ignored talker. The attended talker is tracked more strongly in every dataset (peak 94-188 ms; Wilcoxon p <= 0.01), but correlations are only about 0.01-0.02.

Table 5.4a. Backward-decoder accuracy (%) by validation scheme and window (causal pipeline, 57 channels, lags 0-500 ms). DTU trials are too short for 60-s windows.

validation	window	KUL-AV-GC	DTU	KUL2016	KUL-UHD	USTC
within subject (LOTO)	1 s	57.3 [56.3, 58.2]	57.6 [56.7, 59.8]	58.0 [56.9, 58.9]	55.9 [54.4, 57.0]	56.1 [55.0, 58.3]
	10 s	70.0 [68.9, 71.0]	71.9 [67.9, 74.4]	72.9 [71.3, 75.6]	65.9 [61.8, 69.7]	66.8 [63.4, 72.3]
	60 s	88.9 [84.7, 91.7]	n/a	91.0 [86.1, 94.4]	81.2 [72.9, 87.5]	77.5 [75.0, 85.0]
new subject (LOSO)	1 s	53.8 [53.0, 54.9]	55.3 [53.6, 56.9]	54.7 [54.3, 56.8]	54.6 [53.6, 55.2]	54.1 [52.1, 55.2]
	10 s	61.9 [60.8, 64.8]	66.0 [60.4, 68.3]	66.1 [61.3, 68.9]	63.5 [60.5, 64.9]	60.2 [55.2, 64.8]
	60 s	72.2 [69.4, 81.5]	n/a	84.0 [73.6, 87.5]	77.1 [72.9, 83.3]	67.5 [57.5, 80.0]
new dataset (LODO)	1 s	54.4 [54.0, 55.6]	55.6 [54.6, 56.8]	54.5 [52.9, 55.3]	53.9 [53.0, 54.8]	52.8 [50.4, 54.1]
	10 s	61.9 [59.3, 65.7]	66.5 [62.9, 69.2]	64.1 [57.8, 66.9]	62.8 [56.9, 64.5]	56.4 [52.7, 61.1]
	60 s	77.8 [75.0, 79.2]	n/a	79.2 [66.7, 83.3]	75.0 [70.8, 79.2]	62.5 [50.0, 75.0]

Table 5.4b. 1-s evidence SNR1 = mean(d)/SD(d), d = r(attended) - r(ignored). The official baseline's 53.94% at 1 s implies 0.099. Fitted k in Phi(k SNR1 sqrt(w)): 0.74-0.98.

SNR1	KUL-AV-GC	DTU	KUL2016	KUL-UHD	USTC
within subject	0.185 [0.162, 0.205]	0.177 [0.166, 0.230]	0.197 [0.180, 0.231]	0.142 [0.110, 0.157]	0.157 [0.126, 0.213]
new subject	0.087 [0.078, 0.110]	0.142 [0.091, 0.166]	0.133 [0.105, 0.173]	0.120 [0.090, 0.130]	0.102 [0.047, 0.130]
new dataset	0.107 [0.094, 0.132]	0.131 [0.119, 0.166]	0.120 [0.067, 0.129]	0.097 [0.070, 0.122]	0.069 [0.014, 0.108]

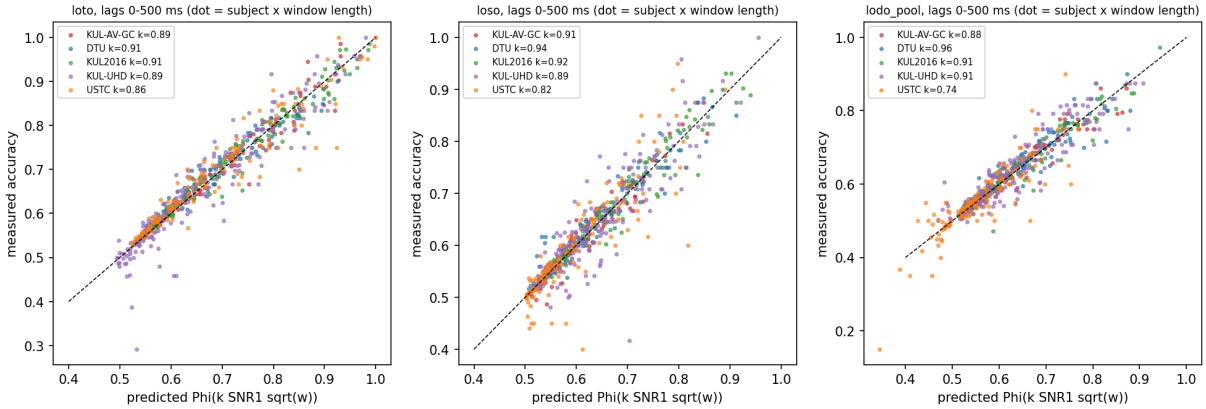


Figure 5.5. The SNR law on all 95 subjects: accuracy predicted from each subject's 1-s SNR (x) vs measured accuracy (y), one dot per subject and window length.

How to read this

Dots on the diagonal mean one number (SNR1) predicts accuracy at every window length. k < 1 means consecutive seconds are correlated, so longer windows help slightly less than sqrt(w). The practical reading: a better decoder (higher SNR1) improves every window length at once.

Pipeline choices (LODO, 95 subjects, paired tests): zero-phase filtering is never better (up to 1.3 pp worse); the organisers' preprocessed 32 Hz data are not better; dropping the common average reference costs up to 1.4 pp; 57, 58 or 62 channels and 0-250 vs 0-500 ms lags make no reliable difference; a causal 0.5-30 Hz variant was +1.3 pp at 10 s (p = 0.04, uncorrected), which is worth a follow-up. How the training data are pooled across datasets (dataset-equal pooled covariance, sample-weighted pooling, or averaging per-subject decoders) changes accuracy by 0.3 pp at most.

5.4 Shortcuts and traps

Table 5.5. Shortcuts found in the training data (full-corpus numbers).

trap	where	evidence	why it hurts	defence
Label = trial parity	KUL-UHD, USTC	100% of trials: odd trials one side, even the other	Anything slow (drift, fatigue, electrode settling) or order-related predicts the label	Fold by subject; never feed trial order or slow drift features
Louder attended envelope	KUL2016	65% of trials; a loudness-only classifier picks the attended stream at 82.5% for unseen subjects	A model can 'decode' attention without looking at the EEG	z-score each stream per window; swap stream columns at random in training
Column 0 = attended	all (training files)	converter convention	Any access to column order leaks the answer	shuffle column order; never rely on it
Reused stimuli, bad names	KUL2016/KUL-UHD, KUL-AV-GC	all KUL-UHD audio occurs in KUL2016; KUL2016 has 8 audios; KUL-AV-GC names do not match audio	Story leakage between folds	group folds by envelope identity
Dataset fingerprint	all	covariance identifies the dataset at 99.7% (84.3% after EA, 87.1% after causal running EA; chance 31%); raw amplitude alone 98.9%	Models learn the lab instead of attention	causal normalisation; LODO validation
Trial fingerprint	spatial decoders	trial ID predicted at 63-89% from 5-s covariances	Random-window splits inflate direction accuracy by a median 22 pp (up to 49 pp)	never score windows from training trials
Gaze / EOG	KUL2016, USTC	USTC's recorded HEO separates left/right trials (d = 0.43); test sets are gaze-controlled	Eye cues will be absent or reversed at test time	exclude EOG and frontal-proxy features from direction models

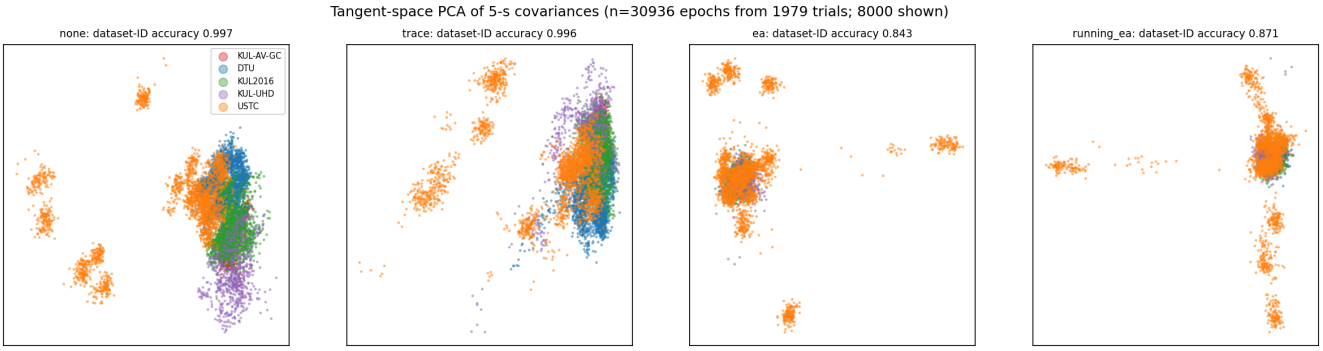


Figure 5.6. Each dot is a 5-s EEG segment placed by its spatial covariance; colour = dataset. Separate islands = a dataset fingerprint. Alignment mixes the colours only partly.

Spatial decoders in numbers (KUL-AV-GC, FBCSP, 5-s windows): within-trial split 57.3%, held-out trial 44.8%, new subject 50.0%, new dataset 53.1%. In KUL-UHD the same decoder drops from 92.7% to 44.3% when whole trials are held out.

5.5 Switch detection and the metric

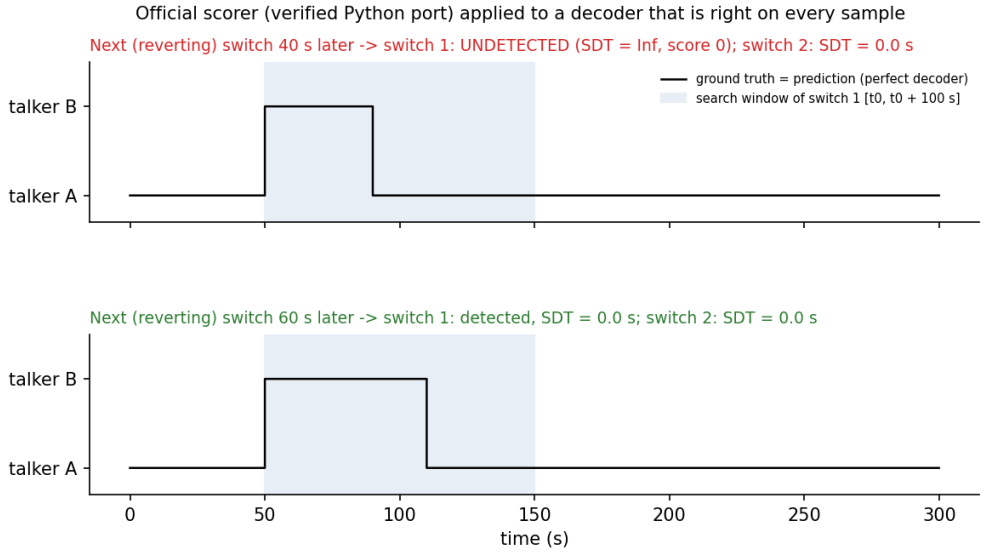


Figure 5.7. The official scorer (Python port) on a decoder that is correct on every sample. When the next switch reverts within 50 s (top), 'old label until the end of the 100-s window' wins the split search, fails the 0.5-s sustain check, and the first switch is scored as missed. With 60 s spacing (bottom) it is detected instantly.

Under the default synthetic schedule (gaps drawn from 30-90 s) this happens to 414 of 3,113 switches for a perfect decoder. It is invisible in the perfect decoder's L (still 1.00, because L is a median) but it costs real decoders. A second property: the 0.5-s sustain check can read the next trial's labels for a switch at the very end of a trial.

Table 5.6. Switch detection on SYNTHETIC switches with cross-dataset (LODO) evidence; best integrator on the median held-out dataset = 10-s causal moving average. 'null L' = the same pipeline on sign-randomised evidence.

held-out dataset	SYNTHETIC switches	Acc %	SDT s	L	H	H argmax	null L	oracle misses
KUL-AV-GC	890	61.2	19.7	0.721	0.662	0.602	0.480	142/890
DTU	220	64.1	6.4	0.904	0.750	0.694	0.730	0/220
KUL2016	934	61.3	15.5	0.730	0.666	0.623	0.549	138/934
KUL-UHD	661	59.9	19.6	0.696	0.644	0.599	0.528	113/661
USTC	408	55.3	19.5	0.746	0.635	0.622	0.649	21/408

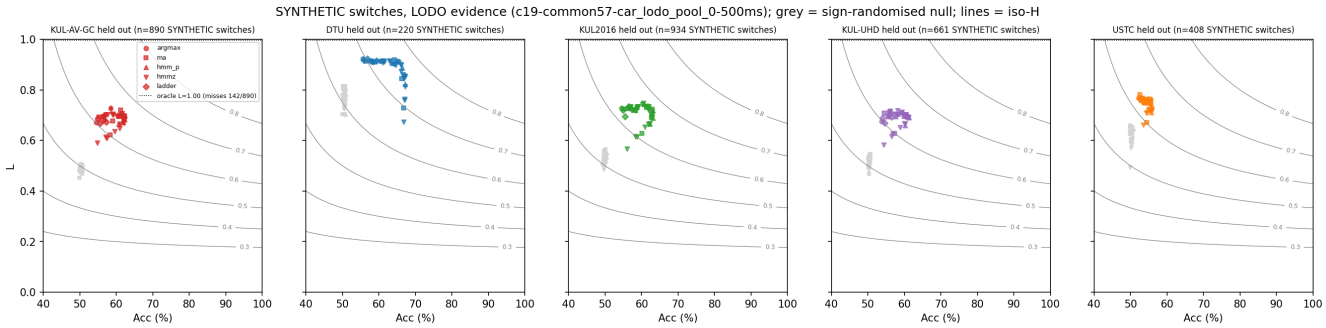


Figure 5.8. Accuracy vs latency score for all integrator settings (colour) and their nulls (grey), with iso-H lines. Everything sits in a narrow band well above the null and far below the perfect decoder.

Genuine switches. On the 99 real KUL-AV-GC side switches, direction evidence from spatial decoders trained on the other four datasets gives Acc 51.7%, SDT 47.4 s, L 0.604, H 0.557 (best setting; null L 0.451). Direction is essentially not decodable across datasets with spatial features. Mapping a speaker decoder through known stream positions gives a high-looking H of about 0.91, but that number is invalid for our purposes: column 0 is the attended talker for the whole trial by construction, and it assumes stream positions are known at test time.

Lookahead unit test. Perturbing every input after t + 1 s leaves all labels up to t unchanged (21/21 checks). Zero-phase filtering and a 2-s budget are both caught as violations; leaks shorter than about 0.5 s may escape the test.

6. What to do about it

6.1 Recommended starting pipeline

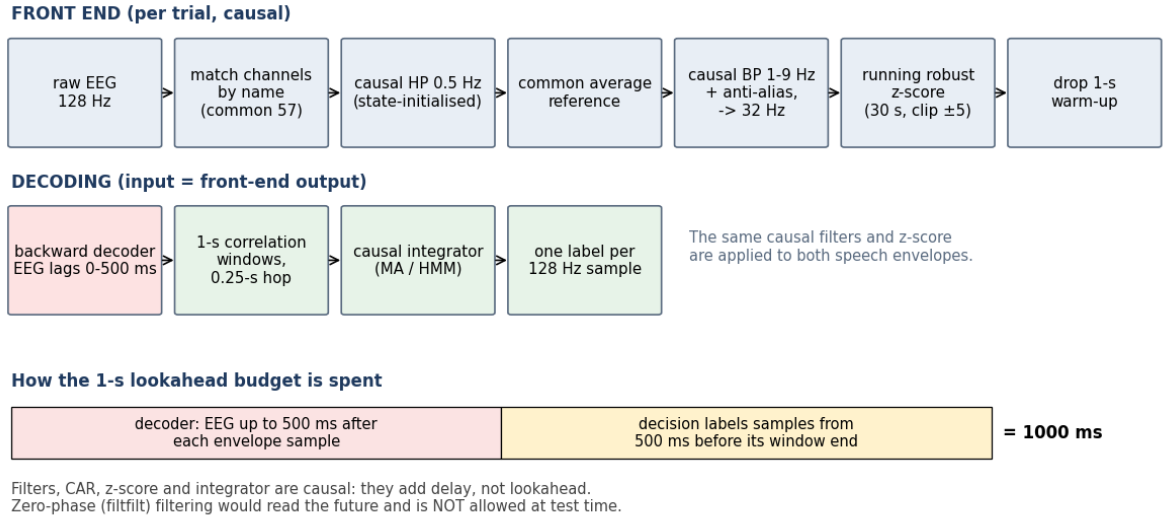


Figure 6.1. The causal pipeline used for every full-corpus decoding result, and how it spends the 1-s lookahead budget exactly.

6.2 Validation checklist

- Outer loop: **leave-one-dataset-out**, with KUL-AV-GC as the main proxy fold; hyperparameters (including the integrator) in an inner loop over the training datasets only. Choose settings on the median held-out dataset, not the mean.
- Aggregate per subject first, then across subjects; report medians with bootstrap CIs.
- Never score windows drawn from a trial that contributed training data. For classification with balanced labels, plain leave-one-trial-out is biased below 50%; use balanced leave-pair-out.
- Group folds by **envelope identity**, and keep KUL2016 and KUL-UHD on the same side of any split.
- Report Acc, the SDT distribution, the miss rate, L and H separately, with the **null L** (sign-randomised evidence) next to every L. Label SYNTHETIC results as such.
- Check the synthetic switch spacing against the scorer's 50-s property; closely spaced switches penalise even a perfect decoder.
- Run the lookahead unit test (reports/eda_full/scripts/e3_lookahead_test.py) on the full submission pipeline.
- Confirm that no step reads trial or subject IDs, trial order, column order, or statistics pooled across test trials.

6.3 Roadmap: the example-trial proposals, updated with full-corpus evidence

Table 6.1. Method roadmap. The literature behind each proposal is listed in Appendix B.

priority	proposal (example-trial EDA)	full-corpus evidence	status
P0	Remove the lab dialects causally (name-matched montage, state-initialised HP, CAR, running robust z)	Confirmed necessary: amplitude alone identifies the dataset at 98.9%; causal filtering costs nothing; CAR helps	keep
P0	Make envelopes scale-symmetric (per-window z-score, random column swap)	Confirmed: loudness-only classifier 82.5% in KUL2016	keep
P0	Retune the integrator against the official scorer	Gain is real but small with cross-dataset evidence (+0.02-0.07 H); a 10-s moving average or variance-normalised HMM is near the frontier	keep, but do not expect big wins
P1	Raise SNR1: pooled-covariance ('avgcorr') decoders, CCA	Pooling strategy made no difference under LODO (<= 0.3 pp); CCA not yet tested	test CCA next
P1	Causal running Euclidean alignment	The fingerprint survives (87%); effect on decoding untested	test, low priority
P2	Within-trial unsupervised adaptation of the decoder	Untested here; allowed by the rules if causal	promising for long trials

priority	proposal (example-trial EDA)	full-corpus evidence	status
P3	Causal DNN envelope decoders with shortcut defences	Untested here; the lookahead rule caps the receptive field	only with LODO validation
avoid	Spatial / direction-only decoders	Confirmed: chance across datasets; memorise trials	auxiliary at most

6.4 Open questions to raise with the organisers

- Are the stream positions (azimuths) available at test time? If so, a speaker decoder also yields direction labels.
- Which label is scored on each blind set (attended stream or attended side), and how are the streams ordered?
- What is the switch spacing in the blind sets, and is the scorer's behaviour for switches closer than 50 s intended?
- How were the envelopes computed? USTC envelopes differ in shape (syllable-rate bump, platykurtic), and KUL-AV-GC envelopes are high-passed at about 1 Hz.
- Which amplifier recorded KUL-UHD (the challenge page and the CND files disagree), and which two channels were dropped?

6.5 Five experiments this analysis justifies

- **Integrator tuning** on LODO evidence against switch schedules that match the test protocol.
- **Raise SNR1** before touching the integrator: CCA with an envelope filter bank, causal within-trial adaptation, or a small DNN trained LODO, all compared on the SNR1 yardstick.
- **Causal front end** as in Figure 6.1, with explicit edge handling for KUL-UHD and USTC.
- **Keep spatial decoders out** unless they survive LOTO and LODO; clarify stream-position availability first.
- **Story-disjoint, order-aware folds** in every model-selection loop.

7. Getting started

7.1 Environment

cd /root/disk_3tb/EEG-sAAD-Grand-Challenge-Training-Dataset && source .venv/bin/activate. The environment has Python 3.12, PyTorch with CUDA (a Tesla V100 is available), the SuperHugeAAD package, numpy/scipy/scikit-learn/pyriemann and reportlab. MATLAB, Octave and MNE are not installed. The machine is shared (8 CPUs, 23 GB RAM): use at most about 4 worker processes.

7.2 Where to find things

Table 7.1. Where the analysis outputs live.

path	what it is
reports/onboarding/	this guide (build_guide.py, make_figures.py, figures/)
reports/eda_raw/	corpus summary: README.md, trial_inventory.csv, channel statistics
EEG-sAAD Data Story.html	example-trial EDA (1.6 MB with embedded images; open it in a browser)
reports/eda_full/README.md, EEG-sAAD_full_corpus_EDA_report.pdf	full-corpus EDA report (all tables, figures and 'how to read' notes)
reports/eda_full/tables/	68 result tables; start with hypotheses_verdicts.csv, per_subject_metrics.csv, label_audit.csv (per trial: labels, switches, column check, story names)
reports/eda_full/scripts/	resumable analysis scripts; run_all.sh drives them; scorer.py is the verified scorer port
reports/eda_full/cache/cnd/<key>.npz	per-trial ground truth extracted from CND: att_stream, att_az, att_col, env_cnd, EOG, story names, condition
reports/eda_full/config.yaml	every analysis parameter (seed 20270101)
reports/eda_full/NEXT_STEPS.md	the five experiments above, with their evidence

7.3 Useful commands

- Re-run the full-corpus EDA (resumable): cd reports/eda_full && bash scripts/run_all.sh
- Score a prediction directory with the Python scorer: use score_subject() / aggregate() in reports/eda_full/scripts/scorer.py on the same CSV format as baselines/matlab_scoring (time_index, label, trial; 0-based)
- Check the scorer port: .venv/bin/python reports/eda_full/scripts/e0_scorer_tests.py (must print PASSED 34/34)
- Load one trial in the causal pipeline: load_pp(row, 'c19', 'common57') in reports/eda_full/scripts/common.py
- Rebuild this guide: cd reports/onboarding && ../../.venv/bin/python make_figures.py && ../../.venv/bin/python build_guide.py

> **Good first tasks for a new team member**
>
> 1. Read reports/eda_full/tables/label_audit.csv for one dataset and find the KUL-AV-GC switch in the CND labels. 2. Reproduce one LODO accuracy number from Table 5.4a with the scripts. 3. Score the sign-randomised null for your own integrator and compare it with Table 5.6.

Appendix A. Glossary

term	meaning
sAAD / AAD	(selective) auditory attention decoding: inferring which of several talkers a listener attends to from EEG
Trial	one continuous recording (50 s to 10 min) with two simultaneous talkers
Envelope	the slow loudness contour of a talker's speech (roughly 1-9 Hz); the thing the brain tracks
Backward decoder	a linear filter mapping EEG (here 0-500 ms after each moment) onto the speech envelope; the talker whose envelope correlates better with the reconstruction is predicted as attended
Forward model / TRF	the reverse direction: how EEG responds over time to the envelope
r(att), r(ign), d	correlation of the reconstruction with the attended / ignored envelope in a window; d = r(att) - r(ign)
SNR1	mean(d)/SD(d) over 1-s windows; accuracy is about Phi(k SNR1 sqrt(w)) (the SNR law)
LOTO / LOSO / LODO	leave-one-trial-out within a subject / leave-one-subject-out within a dataset / leave-one-dataset-out (closest proxy for the blind test sets)
Within-trial split	train and test windows from the same trial; leaky, only used to measure memorisation
Acc, SDT, L, H	official metrics: sample accuracy; switch detection time; median exp(-SDT/60); harmonic mean of Acc and L
SYNTHETIC switch	a switch imposed on a fixed-attention trial by swapping which stream counts as attended (generateSyntheticSwitches.m); measures reaction speed where no real switch happened
Integrator	causal post-processor turning noisy 1-s evidence into stable labels (argmax, moving average, HMM, ladder)
Null (sign-randomised)	the same pipeline fed evidence with random sign: what zero information scores
CAR	common average reference: subtract the mean of all channels at each sample
EA / running EA	Euclidean alignment: whiten each recording by its mean covariance; the running version uses only the past
Causal / zero-phase	causal filters use only past samples (required); zero-phase filters also use future samples
CND	the 'continuous-event neural data' MATLAB format the organisers distribute (neural + stim structs)
SuperHuge	the npy/pickle layout used by the Python DNN baselines, produced by tools/convert_cnd_to_superhuge.py
Montage 57 / 58 / 62	channel sets matched by name: common to all five datasets / BioSemi-64 and USTC / BioSemi-64 and KUL-UHD
ALI, hEOG	alpha lateralisation index (right minus left parieto-occipital 8-13 Hz power, normalised); horizontal eye-movement proxy from frontal electrodes

Appendix B. References

Sources cited by the example-trial EDA ('EEG-sAAD Data Story'). They were **not re-checked** for this guide; items that document marked as unverified (mostly 2026 preprints) remain unverified. Challenge and dataset pages come first.

#	source	link
1	Challenge page (rules, metric, baseline numbers)	https://www.codabench.org/competitions/17856/
2	KUL2016 dataset (Zenodo 4004271)	https://zenodo.org/records/4004271
3	KUL-AV-GC dataset (Zenodo 11058711)	https://zenodo.org/records/11058711
4	KUL-UHD dataset (Zenodo 4518754)	https://zenodo.org/records/4518754
5	USTC dataset (Zenodo 18366436)	https://zenodo.org/records/18366436
6	Heintz et al. 2025, HMM post-processing of AAD decisions (arXiv 2506.24024)	https://arxiv.org/abs/2506.24024
7	Geirnaert et al. 2021, Electroencephalography-based AAD: an overview (SPM; arXiv 2008.04569)	https://arxiv.org/abs/2008.04569
8	Geirnaert et al. 2020, MESD metric for AAD (TNSRE)	https://www.biorxiv.org/content/10.1101/745695v2.full
9	Geirnaert et al. 2021, Unsupervised self-adaptive AAD (JBHI)	https://simongeirnaert.com/publication/unsupervised2021geirnaert/unsupervised2021geirnaert.pdf
10	Time-adaptive unsupervised AAD (bioRxiv 2022.01.07.475386)	https://www.biorxiv.org/content/10.1101/2022.01.07.475386v3.full
11	Initialisation-bias fixes for unsupervised AAD (arXiv 2509.14764)	https://arxiv.org/abs/2509.14764
12	Wong et al. 2018, comparison of regularisation methods for AAD (Front. Neurosci.)	https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00531/full
13	Miran et al. 2018, real-time tracking of selective attention (Front. Neurosci.)	https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00262/full
14	de Cheveigné et al. 2018, decoding the auditory brain with CCA (NeuroImage)	https://www.sciencedirect.com/science/article/pii/S1053811918300338
15	de Cheveigné et al., stepwise match-mismatch study (bioRxiv 2020.11.05.370072)	https://www.biorxiv.org/content/10.1101/2020.11.05.370072v1.full
16	Accou et al. 2023, VLAAI decoder (Sci. Rep.)	https://www.nature.com/articles/s41598-022-27332-2
17	Puffay et al. 2023, relating EEG to continuous speech: review of pitfalls (arXiv 2302.01736)	https://arxiv.org/abs/2302.01736
18	Rotaru et al. 2024, gaze-controlled AV AAD dataset and spatial decoders (bioRxiv 2023.07.13.548824)	https://www.biorxiv.org/content/10.1101/2023.07.13.548824v2
19	Rotaru et al. 2024, spatial AAD under gaze control (arXiv 2412.01401)	https://arxiv.org/abs/2412.01401
20	Yan et al. 2025, block-design shortcuts in DNN AAD (Interspeech)	https://www.isca-archive.org/interspeech_2025/yan25_interspeech.html
21	ListenNet (IJCAI 2025)	https://www.ijcai.org/proceedings/2025/0461.pdf
22	He & Wu, Euclidean alignment (arXiv 1808.05464)	https://arxiv.org/abs/1808.05464
23	Revisiting Euclidean alignment (arXiv 2502.09203)	https://arxiv.org/abs/2502.09203
24	Wilroth et al. 2023, Riemannian transfer for locus-of-attention decoding	https://pubmed.ncbi.nlm.nih.gov/37988748/
25	Adams & MacKay 2007, Bayesian online change-point detection (arXiv 0710.3742)	https://arxiv.org/abs/0710.3742
26	Closed-loop CCA AAD system (bioRxiv 2024.03.01.582668)	https://www.biorxiv.org/content/10.1101/2024.03.01.582668v1
27	Contrastive transformer ensembles for match-mismatch (Sci. Rep. 2025)	https://www.nature.com/articles/s41598-025-13646-4
28	HappyQuokka system, ICASSP 2023 challenge (arXiv 2305.06806)	https://arxiv.org/abs/2305.06806
29	Wang et al., speech features for match-mismatch (arXiv 2401.04964)	https://arxiv.org/abs/2401.04964
30	Deng et al. 2020, alpha lateralisation in spatial attention (NeuroImage)	https://www.cmu.edu/dietrich/psychology/shinn/publications/pdfs/2019/2019neuroimage_deng.pdf
31	Alpha + tracking fusion (JSLHR 2021)	https://pubs.asha.org/doi/10.1044/2021_JSLHR-20-00608
32	Speech onsets and EEG (J. Neural Eng. 2024)	https://iopscience.iop.org/article/10.1088/1741-2552/ad1784
33	Nested cross-validation for EEG (arXiv 2510.19174)	https://arxiv.org/abs/2510.19174
34	Ivucic et al. 2024, causal time-series splits (EMBC)	https://doi.org/10.1109/embc53108.2024.10782636
35	EEG-AAD 2026 locus-of-attention challenge	https://fchest.github.io/icassp-aad/
36	Unverified 2026 preprints cited: channel adaptation (arXiv 2604.23091), unbalanced datasets (arXiv 2605.25605), learned switch gate (arXiv 2608.01618), Markov-switching model (arXiv 2602.13447); EEG foundation-model study (OpenReview 5Xwm8e6vbh)	https://openreview.net/forum?id=5Xwm8e6vbh