UNDERSTANDING AND IMPROVING CROSS-SESSION GENERALIZATION IN SURFACE-EMG SILENT SPEECH RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Silent-speech decoding from surface electromyography (sEMG) offers a non-invasive communication interface, but performance can deteriorate across recording sessions. We use Mindscape-137, to our knowledge the largest open silent speech sEMG corpus by utterance count, with 25,957 silent utterances across 25 sessions from one speaker spanning 21,632 unique sentences and 2,951 word types, to examine signal variation and evaluate interventions for cross-session recognition. Repeated sentences and electrode reapplication between sessions support content-matched comparisons. Channel amplitudes and relationships vary across sessions, while temporal amplitude-envelope features remain informative about repeated-prompt identity. Imposed gain and mixing perturbations increase decoding errors, so the decoder is sensitive to these transformations. Out-of-span augmentation reduces word error rate from 72.15% to 65.77%. In a session-scaling study with three fixed held-out sessions, increasing training coverage from 5 to 22 sessions reduces fresh-sentence word error rate from 67.22% to 28.30%. Checkpoint averaging and latent covariance alignment provide further improvements in the scaling study, whereas the tested rate correction and labeled calibration do not consistently help.

1 INTRODUCTION

Invasive brain–computer interfaces (BCIs) have demonstrated accurate attempted-speech communication in people with severe motor impairments, but require surgically implanted electrodes (Willett et al., 2023; Card et al., 2024). Surface electromyography (sEMG) offers a non-invasive, wearable route by recording speech-related muscle activity through skin electrodes. For users who can silently articulate words, these signals support speech synthesis and text decoding without an audible voice (Gaddy & Klein, 2020; Benster et al., 2024). Recent EMG-to-text work includes phonemic sequence models without time-aligned audio targets (Gowda & Miller, 2026) and language model adapters for closed-vocabulary decoding (Mohapatra et al., 2025).

Reliable use requires decoding to remain accurate across recording sessions. Removing and reapplying sEMG electrodes can change the recorded signals, motivating session normalization and adaptation (Maier-Hein et al., 2005; Wand & Schultz, 2014). Prior EMG speech studies have investigated multi-session training (Wand & Schultz, 2011), compensation for recording-position shifts (Wand et al., 2014b), and adaptation using unlabeled target-session recordings, including domain-adversarial training on audibly produced EMG (Wand et al., 2018). Wearable command-recognition systems also evaluate multi-day sensor repositioning and incremental fine-tuning (Spacone et al., 2026). Invasive speech BCIs also accommodate recording changes: Willett et al. (2023) use day-specific input layers for across-day variation and rolling normalization for within-day variation. Long-term decoder adaptation across days has also enabled neural-map stabilization and plug-and-play motor control with electrocorticography (ECoG) (Silversmith et al., 2021).

Training across speakers and recording configurations offers another route. Pretraining on heterogeneous electroencephalography (EEG) and EMG recordings supports word-classification calibration across users and recording days (Inoue et al., 2025), and Le et al. (2026) examine multisubject pretraining and short calibration in a closed 50-sentence sEMG corpus. These evaluations are limited

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by the recorded material: existing multi-session silent-speech corpora hold 1.8 hours of silent speech across eight speakers in EMG-UKA and 3.6 hours from one speaker in Gaddy and Klein's corpus (Wand et al., 2014a; Gaddy & Klein, 2020). Larger longitudinal recordings enable more extensive investigation of recording variability and its relationship to decoding.

We therefore run this study on Mindscape-137, the largest open multi-session silent speech sEMG corpus by utterance count that we are aware of: 25,957 silent productions and 22.1 hours from one speaker across 25 sessions over 14 days, spanning 21,632 unique sentences and 2,951 word types, with electrode reapplication between sessions and a voiced twin for nearly every production. Each session repeats the same 100 anchor and 40 calibration sentences alongside new ones, which supports content-matched signal comparisons and separates transfer of familiar sentences from recognition of new sentences. Using a MONA-derived reference encoder (Benster et al., 2024) decoded by lexicon-constrained CTC beam search with an n-gram language model (Kahn et al., 2022; Heafield, 2011), we evaluate interventions spanning training-session coverage, signal augmentation, checkpoint averaging and target-session adaptation. The reference benchmark uses a chronological split; complementary studies examine additional training sessions and alternative pipelines under their own matched controls.

The study has two halves. We first characterize cross-session variability and its cost to decoding. Channel amplitudes and relationships vary across sessions, with covariance-based mixing measures changing more relative to their own within-session reference than gain measures. Temporal amplitude-envelope features remain useful for identifying repeated sentences, whereas covariance and correlation features identify little sentence content under the tested classifiers. Controlled gain and mixing perturbations increase decoding errors, and phoneme probes characterize the speech content accessible to the tested readouts.

We then ask which interventions improve generalization, and report those that failed alongside those that worked. Signal augmentation helps without additional recordings, and more training sessions improve decoding. After training, checkpoint averaging helps using no target-session data at all, and latent covariance alignment (in the style of CORAL, Sun et al., 2016) adds gains from unlabeled target recordings. The tested rate correction instead raised error relative to checkpoint averaging in every evaluation, and labeled calibration gave no sustained benefit at ten or more training sessions. Supporting prompt and phoneme analyses are reported in Appendix A.

Our contributions are:

- **A longitudinal corpus and cross-session benchmark** for studying signal variation and transfer to familiar and unfamiliar sentences.
- **A characterization of cross-session variability and its decoding cost**, through within- and between-session measurements, perturbations, and sentence and phoneme probes.
- **An evaluation of interventions**, spanning training coverage, augmentation, checkpoint averaging, covariance alignment, rate correction and labeled calibration, under each study's matched controls.

2 METHODS

2.1 DATASET AND ACQUISITION

We use Mindscape-137, a longitudinal corpus of silent and voiced speech from one adult participant. The corpus contains 25 recording sessions collected over 14 days, with electrodes removed and reapplied between sessions. Each session includes 100 anchor sentences and 40 calibration sentences shared across sessions, together with approximately 860 fresh sentences. Prompts are English sentences of 4–20 words. The participant produces each prompt silently and then aloud, providing silent sEMG, voiced sEMG, and audio of the voiced production. The corpus contains 25,957 silent productions totaling 22.1 hours and 25,445 voiced productions, covering 21,632 unique sentences and 2,951 word types. Repeated prompts allow comparisons across recording sessions with sentence content held fixed, while fresh prompts support evaluation on unfamiliar sentences.

Signals are recorded from eight bipolar sEMG channels over facial and neck sites involved in articulation and stored at 1 kHz; acquisition and segmentation details are given in Appendix B.2.

2.2 EVALUATION DESIGN

We use two primary evaluation designs: a chronological split for the reference benchmark and a fixed held-out-session split for the session-scaling study.

Session and sentence terminology. A *seen session* contributes recordings to initial training and an *unseen session* does not. In-session evaluation uses held-out recordings from seen sessions. A *familiar sentence* has the same prompt text as one in the sEMG training data and an *unfamiliar sentence* does not, so an unseen session can contain both. Familiarity is relative to the sEMG training transcripts, not the external language model. All definitions refer to initial training, before target-session adaptation.

The reference benchmark uses a chronological split: sessions 1–10 for training, sessions 11–13 for development, and sessions 14–15 for testing. Fresh-sentence results are reported separately, and a familiar-versus-unfamiliar sentence comparison on the test sessions is given in Appendix A.5.

The session-scaling study uses a separate design. Recordings were collected across multiple dates, with some sessions recorded on the same day. Sessions 6, 13 and 19 were selected to sample the early, middle and late parts of the recording period, respectively, and are excluded from training in every scaling experiment. Models are trained on 5, 10, 15 or 22 of the remaining sessions; the exact training-session lists are given in Table 3 in Appendix B. For each experiment, 10% of the data from each included training session is held out for development. Evaluation uses the same bank of 2,680 fresh recordings from those three sessions, with plain models evaluated at epoch 200. Training can include recordings collected after an evaluation session, so this is a held-out-session design rather than a chronological future-session test.

The studies do not constitute one uniform ablation grid: effects are interpreted relative to each study's matched control and retain their original evaluation population, decoding procedure, and aggregation rule. The reference test sessions were evaluated repeatedly, so they are treated as a transfer check, with checkpoint and method-selection procedures documented per experiment.

2.3 REFERENCE DECODING SYSTEM

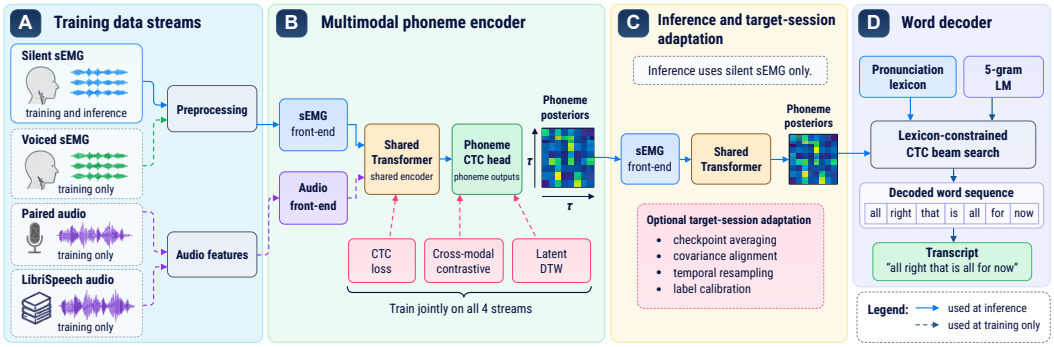

Figure 1: Multimodal training and silent-speech decoding. **(A)** Silent sEMG, voiced sEMG, paired speech audio and LibriSpeech audio provide four training streams. **(B)** Modality-specific front ends feed a shared Transformer and phoneme CTC head, trained with CTC, cross-modal contrastive and latent dynamic-time-warping (DTW) losses. **(C)** Inference needs only silent sEMG, and the listed post-training interventions are evaluated separately or in specified combinations. **(D)** Lexicon-constrained CTC beam search combines phoneme posteriors with a pronunciation lexicon and a 5-gram language model. Solid and dashed arrows mark inference and training-only paths.

Signal preprocessing. Inputs to the reference encoder are high-pass filtered and cleaned of mains-interference harmonics detected per session and channel. Amplitude is scaled by one root-mean-square value estimated over training utterances, with no per-channel gain equalization, per-utterance normalization, rectification or envelope extraction (Appendix B.2).

Model and training. The reference system (Figure 1) is adapted from MONA (Benster et al., 2024). Its connectionist temporal classification (CTC) head (Graves et al., 2006) has 41 units: blank,

39 stress-stripped ARPAbet phonemes and a word boundary. Models train for 200 epochs, with LibriSpeech audio (Panayotov et al., 2015) filling half of each batch and no signal augmentation. The reference checkpoint is selected by greedy WER on held-out training-session utterances, whereas intervention comparisons use stated epochs or checkpoint windows. Table 4 in Appendix B.3 lists all other settings.

Decoding. The reference benchmark reports two decoders separately: greedy phoneme output mapped to the nearest word in the training-text lexicon, and lexicon-constrained CTC beam search with a 5-gram language model built from training sentences (settings in Appendix C). The scaling study uses a 123,432-word lexicon derived from the CMU Pronouncing Dictionary (Weide, 1998) and an external 5-gram language model trained on OpenWebText2 (Gao et al., 2020), with operating points selected per experiment on its own development data (Appendix B.4).

2.4 Interventions for cross-session generalization

Signal augmentation. Training-time augmentations apply channel transformations derived from training-session variation while keeping the transcript target. Out-of-span augmentation includes transformations outside the linear span of between-session variation estimated from the training sessions. A separate drift-augmentation condition, `xs_aug`, uses transformations fitted on training sessions 1–10 (Appendix B.4).

Training-data coverage. Adding training sessions supplies more utterances and recording conditions and, under the 200-epoch schedule, more optimizer updates, so the scaling design does not separate session diversity from data volume and training exposure.

Adaptation and combined systems. Post-training interventions include latent covariance alignment, temporal resampling, checkpoint averaging and labeled calibration. Checkpoint averaging needs no target recordings, whereas the other corrections estimate statistics from unlabeled target-session recordings, which is transductive when the whole session is used. Labeled calibration uses separate support and query recordings. In the historical benchmark, averaging, covariance alignment and temporal resampling form a three-component stack (Appendix C).

In the scaling study, checkpoint averaging takes the arithmetic mean of 25 sets of model weights from epochs 80 through 200 at five-epoch intervals, following weight averaging along a training trajectory (Izmailov et al., 2018). Latent covariance alignment applies a whitening–recoloring transformation related to CORAL (Sun et al., 2016) to the encoder output, mapping the target-session covariance, estimated from unlabeled target-session recordings, toward the equally weighted mean of the training sessions' within-session covariances. Rate correction linearly resamples each recording over its valid length by the ratio of training to target mean valid latent-sequence lengths (Appendix B.4).

Scaling-study calibration evaluates labeled fine-tuning with checkpoint averaging, covariance alignment and rate correction. A previously development-selected configuration is kept fixed while support budgets vary from 0 to 140 labeled utterances per target session in steps of 10. Support comes first from calibration recordings and then from anchors; evaluation uses separate fresh recordings. Zero labels denotes the baseline before labeled fine-tuning, and decoder settings remain fixed at their zero-label values throughout each curve. Support-selection details and additional calibration comparisons are given in Appendices B.4 and B.6.

2.5 Signal and representation analyses

Between-session variation. We compare channel statistics across sessions using repeated anchor sentences to control sentence content. In the reference covariance analysis, a whitening–recoloring map between session covariances is fitted on 25 anchor sentences and evaluated on 25 others:

$$A_{i \to j} = C_j^{1/2} C_i^{-1/2}.$$

Channel gain is summarized by the diagonal of this map, and mixing by the root-mean-square off-diagonal magnitude after gain normalization. Each component's between-session variation is expressed relative to its own within-session variation. The covariance map fixes the otherwise un-

determined rotation and therefore describes a particular statistical transport, rather than a uniquely identified physical electrode transformation.

Prompt identification. The complementary signal analysis uses the first silent production of each of 100 anchor prompts in ten sessions, yielding 1,000 trials. We compare onset-aligned low-pass amplitude envelopes with band-power, covariance, and correlation features using shrinkage linear discriminant analysis for 100-way prompt identification. The held-out evaluations use sessions 1, 5, and 6 in turn, with preprocessing statistics and shrinkage fitted using only the training sessions (Appendix A.3). These probes measure how readily prompt identity can be recovered using each specified feature and readout.

Phoneme probes. To examine phonetic content within the decoder, we fit multinomial logistic-regression probes to frozen activations and evaluate balanced accuracy. We compare probes fitted within a session with probes fitted on training sessions and transferred to unseen sessions. Phone labels for voiced recordings come from Montreal Forced Aligner alignment of paired audio (McAuliffe et al., 2017), and silent labels are transferred from the voiced twin through the model's DTW alignment (Appendix A.4). Probe accuracy describes accessibility to the specified readout, not a fraction of all speech information retained.

Controlled perturbations. Gain and mixing transformations derived from measured recording statistics are imposed, singly and together, on held-out recordings, and a fixed decoder's WER is compared with its WER on the unmodified recordings. This measures decoder sensitivity, and because the effects interact and are imperfectly isolated, their sum is not a causal partition of drift.

2.6 Outcomes and reporting

We report word error rate, computed as substitutions, deletions and insertions divided by the number of reference words. The scaling study, its correction and calibration experiments, and the controlled-perturbation experiment pool word edit counts and reference-word counts across evaluated recordings. For the historical reference/correction benchmark, aggregate WER is the recording-count-weighted mean of session WERs (903 and 894 fresh recordings in sessions 14 and 15), rather than a pooled word-count estimate. Improvements are reported in percentage points as baseline minus method error, so positive values indicate better performance. Independent training runs are distinguished from repeated checkpoint evaluations within a run (Appendix B.7).

3 Experiments

We examine how sEMG recordings vary across sessions and test whether generalization improves by broadening the signals encountered during training, combining model weights across training checkpoints, and adapting to the target session.

3.1 Characterizing signal variation across sessions

Repeated sentences allow us to compare recordings while holding sentence content fixed. The same sentence showed changes in both channel amplitudes and relationships across sessions. In the example from sessions 3 and 8, two channels changed amplitude in opposite directions, while their correlation fell from 0.552 to 0.092 (Figure 2). Across content-matched comparisons, gain and channel-mixing variation were greater between sessions than within sessions (Appendix A).

To test decoder sensitivity, we imposed gain and mixing transformations derived from measured recording statistics on 86 held-out fresh-sentence recordings from each of training sessions 1–10, 860 recordings in total, using the same fixed checkpoint and greedy decoder for the original and transformed recordings. Gain, mixing, and their combination increased WER by 3.09, 4.36, and 6.70 percentage points, respectively (Table 1). These controlled tests establish sensitivity to the imposed transformations; they do not assign causal shares of natural cross-session errors. Paired-bootstrap intervals are in Appendix A.

These observations motivate exposing the decoder to representative signal changes during training.

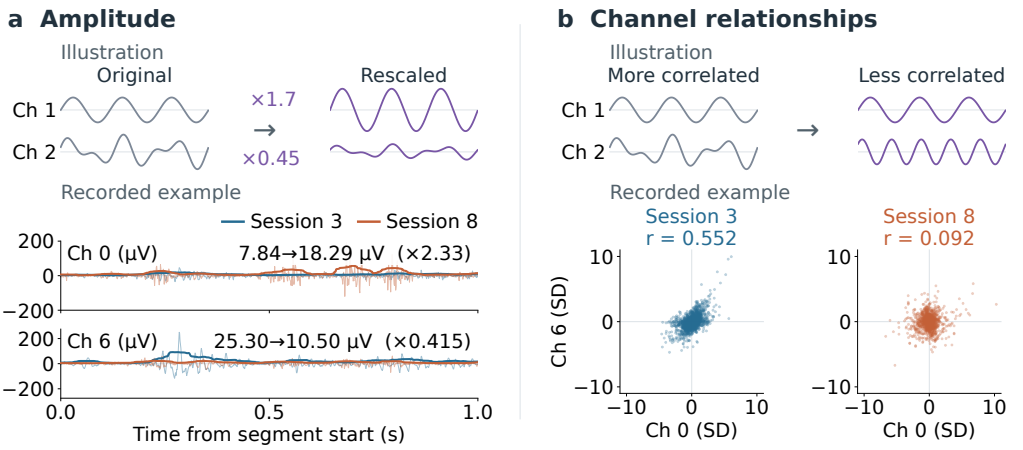

Figure 2: Amplitude and channel relationships. Conceptual illustrations (top) accompany recorded examples of the same sentence in sessions 3 and 8 (bottom). **(a)** Voltage waveforms (thin) and 51-sample RMS envelopes (thick) show amplitude changes; 0.3–1.3 s after GO is relabeled 0–1 s. **(b)** Within-trial standardization removes scale differences while preserving correlation. RMS and correlation annotations use full trials. Illustrations are synthetic, not fitted transformations. Measurement details are in Appendix A.

Table 1: Controlled perturbations of a fixed decoder.

Imposed transformation	**Greedy WER (%)**	**Increase over unmodified input (percentage points)**
None	34.04	—
Channel gain	37.13	3.09
Channel mixing	38.40	4.36
Gain and mixing	40.74	6.70

3.2 Augmenting signals using measured session variation

We use training-session statistics to guide channel transformations while retaining the original sentence labels. Drift augmentation applies transformations estimated from training sessions, while out-of-span augmentation includes transformations outside the linear span of the estimated between-session variation. This broadens the signal variation encountered during training without collecting additional recordings.

In the sessions-14–15 experiment, epoch-200 WER was 72.15% without augmentation, 68.40% with drift augmentation, and 65.77% with out-of-span augmentation (Figure 3a). Each condition used one training run, with the same evaluation recordings and fixed greedy decoder. Out-of-span augmentation also yielded lower WER at all 19 paired checkpoints evaluated between epochs 110 and 200 in both sessions; checkpoint-window statistics are in Appendix B.4.

3.3 Expanding training across recording sessions

Additional sessions provide new utterances and naturally occurring recording variation. We trained models using 5, 10, 15, or 22 sessions and evaluated them on the same fresh recordings from held-out sessions 6, 13, and 19.

Averaged over the two runs (Figure 3b), WER fell from 67.22% with 5 training sessions to 41.18%, 32.93%, and 28.30% with 10, 15, and 22 sessions, respectively. The two runs differed by no more than 0.78 percentage points at any training size, and per-run values are given in Table 5. Decoder settings are selected separately for each model. Adding sessions also shortens the recording interval between training and evaluation: the training lists in Table 3 bracket session 6 with immediate

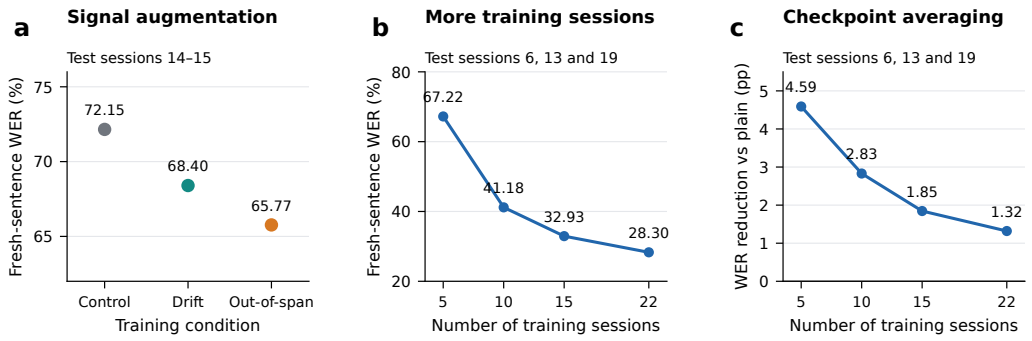

Figure 3: Augmentation, training-session coverage and checkpoint averaging. **(a)** Epoch-200 augmentation comparison on sessions 14–15 with a fixed greedy decoder. **(b)** Two-run mean fresh-sentence WER on sessions 6, 13 and 19 for epoch-200 models. Per-run values are in Table 5. **(c)** Two-run mean WER reduction from checkpoint averaging relative to the corresponding plain checkpoints. Panel a weights session WER by recording counts; b and c pool word counts within each evaluation. Augmentation and scaling use separate protocols (Section 2.2).

neighbours from ten sessions onward, session 13 from 15 onward, and session 19 only at 22, so temporal proximity to the held-out sessions increases along with coverage.

Adding sessions broadens the recording conditions encountered during training, while checkpoint averaging combines different states reached during that training. A single checkpoint may reflect fluctuations specific to that stage of optimization; averaging weights across checkpoints may smooth these fluctuations and reduce dependence on a particular training state (Izmailov et al., 2018).

3.4 Improving generalization through checkpoint averaging

We averaged the weights of 25 checkpoints saved every five epochs from epoch 80 to 200. Relative to the corresponding plain epoch-200 checkpoints, averaging reduced WER by 4.59, 2.83, 1.85, and 1.32 percentage points at 5, 10, 15, and 22 training sessions, respectively (Figure 3c). These are mean gains across two independent training runs; averaging improved WER in both runs at every training size, by 1.09 to 5.07 percentage points (Table 6). The absolute gain shrank as WER fell, but stayed at 4.7–6.9% of the plain-checkpoint WER.

Checkpoint averaging needs no additional recordings, target-session data or training, so its gains add to those from more training sessions at no data cost. It is also the starting point for the target-session corrections in Section 3.5, and Figure 4a plots its gains alongside theirs.

3.5 Target-session corrections and labeled calibration

Checkpoint averaging uses no target-session data. We next tested corrections that do: latent covariance alignment and rate correction estimate statistics from unlabeled target recordings, and calibration fine-tunes on labeled ones (Section 2.4).

Covariance alignment added gains. Rate correction did not. Applied on top of checkpoint averaging, covariance alignment further reduced WER in all eight evaluations, both training runs at all four training sizes, by 0.34 to 3.33 percentage points. The two-run mean reduction relative to the plain checkpoint grew from 4.59, 2.83, 1.85 and 1.32 percentage points with averaging alone to 7.36, 4.61, 2.66 and 1.86 with alignment, at 5, 10, 15 and 22 training sessions (Figure 4a). The increment shrank with more training sessions, and at 22 sessions it falls inside the run-to-run spread. Adding rate correction to averaging instead increased WER by 0.17–0.70 percentage points in all eight evaluations, and adding it to averaging plus covariance alignment increased WER in seven of the eight (Table 6).

Labeled calibration gave no sustained gain after ten or more training sessions. We fine-tuned the averaging/covariance/rate system on 0 to 140 labeled utterances per target session, with decoder

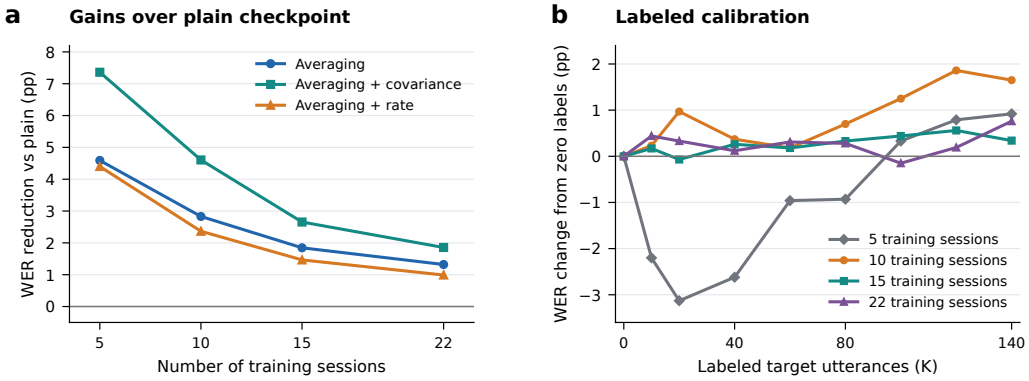

Figure 4: Target-session correction and calibration. **(a)** Two-run mean WER reductions relative to each run's plain checkpoint; covariance and rate conditions include checkpoint averaging and use unlabeled target recordings. **(b)** Change from zero-label WER for the averaging/covariance/rate system, with one training run per size and decoder settings fixed within each curve. All evaluations pool word counts across sessions 6, 13 and 19. Positive values indicate improvement in a and deterioration in b; changes are in percentage points.

settings fixed at their zero-label values (Figure 4b). The five-session model improved from 60.15% without labels to 57.02% at 20 labeled utterances per target session, a reduction of 3.13 percentage points, then ended at 61.07% by 140. The ten-, 15- and 22-session models showed no sustained decrease at any budget, moving from 37.24% to 38.89%, 30.37% to 30.71%, and 26.93% to 27.69%. Each curve uses one training run and a fixed fine-tuning configuration; per-budget WERs and a comparison with fresh-text support are in Appendix B.6.

4 Discussion and Conclusion

Session differences were evident in channel amplitudes and relationships, and controlled gain or mixing perturbations degraded a fixed decoder. Yet amplitude-envelope features and phoneme probes still recovered speech-related information in unseen sessions. Together, these results characterize cross-session variability as substantial changes in the recorded signal that affect decoding without eliminating the underlying speech information. This helps separate changes in signal statistics from a loss of speech content, and suggests that a major challenge is learning representations and decoders that remain stable across recording conditions.

Broader exposure to this variability substantially improved generalization. Expanding training-session coverage improved unseen-session recognition in both training runs; averaged over the two runs, fresh-sentence WER fell from 67.22% with 5 sessions to 28.30% with 22. Out-of-span augmentation also improved a separate historical benchmark without additional recordings. Together with prior work on across-day sampling, cross-speaker training, and pooled-user training (Natraj et al., 2025; Scheck et al., 2024; Fogg et al., 2026), these results support broader training exposure as a way to improve robustness to session-to-session variation.

Post-training methods provided further gains. Checkpoint averaging improved decoding without target-session data, while covariance alignment further improved performance using unlabeled target recordings, consistent with related alignment approaches for intracortical decoding (Degenhart et al., 2020; Zhang et al., 2026). The additional benefit of these corrections became smaller as training-session coverage increased, suggesting that broader training can absorb part of the mismatch that would otherwise remain at test time. Labeled calibration marks where this trend ends: it helped the five-session model, but improved no model trained on ten or more sessions by more than 0.15 percentage points.

Overall, the results connect the structure of cross-session variability to practical strategies for improving generalization. Training across a broader range of sessions provides the largest improve-

ment, and checkpoint averaging and covariance alignment reduce the remaining mismatch without target-session labels. Once training covers ten or more sessions, labeled calibration has nothing left to fix. These findings point toward silent-speech decoders that stay accurate across recording sessions without session-specific labels.

5 Limitations and Future Work

The study uses one speaker, one electrode montage, and 14 recording days, with electrode replacement between sessions. Expanding to more speakers, recording setups, and longer time spans would help distinguish session effects from electrode effects and test generalization. The session-scaling split uses interleaved held-out sessions rather than a prospective future-day evaluation, and historical sessions were reused across analyses. Prospective held-out sessions and additional training runs would provide a clearer estimate of uncertainty. Finally, our experiments were offline and used a constrained lexicon, motivating future work on online adaptation, unrestricted-vocabulary decoding, and calibration protocols beyond those tested here.

REPRODUCIBILITY STATEMENT

The Mindscape-137 dataset and the code for every experiment reported here will be publicly available by the date of publication, through a public dataset repository and a public code release that will be linked in the camera-ready version. Section 2 specifies the corpus, the evaluation designs, the model and the training and decoding procedures, Appendix B gives the training-session lists for each scaling experiment and the preprocessing, architecture, optimization and intervention settings, and the remaining appendices give the per-session results and the historical model and decoder settings.

ETHICS STATEMENT

The corpus comes from a single adult participant who is one of the authors and who recorded their own data.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, literature retrieval and discovery, writing and debugging code, preparing plots and figures, and result interpretation and consistency checks. The authors take responsibility for the final manuscript, scientific claims, references, code, and figures.

REFERENCES

Tyler Benster, Guy Wilson, Reshef Elisha, Francis R. Willett, and Shaul Druckmann. A cross-modal approach to silent speech with LLM-enhanced recognition. arXiv preprint arXiv:2403.05583, 2024. URL https://arxiv.org/abs/2403.05583.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132.

Alan D. Degenhart, William E. Bishop, Emily R. Oby, Elizabeth C. Tyler-Kabara, Steven M. Chase, Aaron P. Batista, and Byron M. Yu. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature Biomedical Engineering*, 4(7):672–685, 2020. doi: 10.1038/s41551-020-0542-9.

Zachery M. Fogg, Nicholas S. Card, Maitreyee Wairagkar, Aparna Srinivasan, Tyler Singer-Clark, Xianda Hou, Elizaveta Okorokova, Hamza Peracha, Carrina Iacobacci, Tiffany Brailow, Justin Jude, Hadar Levi-Aharoni, Trung Le, Domenick Mifsud, Pranav I. Deevi, Samuel R. Nason-Tomaszewski, Anna L. Pritchard, Yunnuo Zhang, Brice Richards, Payton Bechefsky, Leigh R. Hochberg, Ziv Williams, Kiarash Shahlaie, Nicholas AuYong, Daniel B. Rubin, Chethan Pandarinath, David M. Brandman, and Sergey D. Stavisky. A generalizable speech neuroprosthesis. bioRxiv preprint, 2026. URL https://www.biorxiv.org/content/10.64898/2026.07.23.739430v1.

David Gaddy and Dan Klein. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5521–5530. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.445.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. doi: 10.48550/arXiv.2101.00027. URL https://arxiv.org/abs/2101.00027.

Harshavardhana T. Gowda and Lee M. Miller. Non-invasive electromyographic speech neuroprosthesis: A geometric perspective. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11636–11650. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.564. URL https://aclanthology.org/2026.findings-acl.564/.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 369–376, 2006. doi: 10.1145/1143844.1143891. URL https://www.cs.toronto.edu/~graves/icml_2006.pdf.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*, pp. 187–197, Edinburgh, Scotland, 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2123/.

Masakazu Inoue, Motoshige Sato, Kenichi Tomeoka, Nathania Nah, Eri Hatakeyama, Kai Arulkumaran, Ilya Horiguchi, and Shuntaro Sasai. A silent speech decoding system from EEG and EMG with heterogenous electrode configurations. In *Proceedings of Interspeech 2025*, pp. 5603–5607, 2025. doi: 10.21437/Interspeech.2025-1183.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 876–885. AUAI Press, 2018. URL https://auai.org/uai2018/proceedings/papers/313.pdf.

Jacob D. Kahn, Vineel Pratap, Tatiana Likhomanenko, Qiantong Xu, Awni Hannun, Jeff Cai, Paden Tomasello, Ann Lee, Edouard Grave, Gilad Avidov, Benoit Steiner, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. Flashlight: Enabling innovation in tools for machine learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10557–10574. PMLR, 2022. URL https://proceedings.mlr.press/v162/kahn22a.html.

Chenqian Le, Beatrice Fumagalli, Yasamin Esmaeili, Xupeng Chen, Tianyu He, Nikasadat Emami, Adeen Flinker, and Yao Wang. Multi-subject pretraining enables short-calibration personalization for closed-corpus surface EMG speech decoding. arXiv preprint arXiv:2609.21288, 2026. URL https://arxiv.org/abs/2609.21288.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. Session independent non-audible speech recognition using surface electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 331–336, 2005. doi: 10.1109/ASRU.2005.1566521.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pp. 498–

502, 2017. doi: 10.21437/Interspeech.2017-1386.

Payal Mohapatra, Akash Pandey, Xiaoyuan Zhang, and Qi Zhu. Can LLMs understand unvoiced speech? exploring EMG-to-text conversion with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 703–712, 2025. doi: 10.18653/v1/2025.acl-short.56.

Nikhilesh Natraj, Sarah Seko, Reza Abiri, Runfeng Miao, Hongyi Yan, Yasmin Graham, Adelyn Tu-Chan, Edward F. Chang, and Karunesh Ganguly. Sampling representational plasticity of simple imagined movements across days enables long-term neuroprosthetic control. *Cell*, 188(5):1208–1225.e32, 2025. doi: 10.1016/j.cell.2025.02.001.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://www.danielpovey.com/files/2015_icassp_librispeech.pdf.

Kevin Scheck, Zhao Ren, Tom Dombeck, Jenny Sonnert, Stefano van Gogh, Qinhan Hou, Michael Wand, and Tanja Schultz. Cross-speaker training and adaptation for electromyography-to-speech conversion. In *Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, 2024. doi: 10.1109/EMBC53108.2024.10781707.

Daniel B. Silversmith, Reza Abiri, Nicholas F. Hardy, Nikhilesh Natraj, Adelyn Tu-Chan, Edward F. Chang, and Karunesh Ganguly. Plug-and-play control of a brain-computer interface through neural map stabilization. *Nature Biotechnology*, 39(3):326–335, 2021. doi: 10.1038/s41587-020-0662-5.

Giusy Spacone, Sebastian Frey, Giovanni Pollo, Alessio Burrello, Daniele Jahier Pagliari, Victor Kartsch, Andrea Cossettini, and Luca Benini. SilentWear: An ultra-low power wearable system for EMG-based silent speech recognition. arXiv preprint arXiv:2603.02847, 2026. URL https://arxiv.org/abs/2603.02847.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2058–2065, 2016. doi: 10.1609/aaai.v30i1.10306. URL https://ojs.aaai.org/index.php/AAAI/article/view/10306.

Michael Wand and Tanja Schultz. Session-independent EMG-based speech recognition. In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*, pp. 295–300. SciTePress, 2011. doi: 10.5220/0003169702950300.

Michael Wand and Tanja Schultz. Towards real-life application of EMG-based speech recognition by using unsupervised adaptation. In *Proceedings of Interspeech 2014*, pp. 1189–1193, 2014. doi: 10.21437/Interspeech.2014-301.

Michael Wand, Matthias Janke, and Tanja Schultz. The EMG-UKA corpus for electromyographic speech processing. In *Proceedings of Interspeech 2014*, pp. 1593–1597, 2014a. doi: 10.21437/Interspeech.2014-379.

Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz. Compensation of recording position shifts for a myoelectric silent speech recognizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2094–2098, 2014b. doi: 10.1109/ICASSP.2014.6853968.

Michael Wand, Tanja Schultz, and Jürgen Schmidhuber. Domain-adversarial training for session independent EMG-based speech recognition. In *Proceedings of Interspeech 2018*, pp. 3167–3171, 2018. doi: 10.21437/Interspeech.2018-2318.

Robert L. Weide. The Carnegie Mellon pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Zhanqi Zhang, Shun Li, Bernardo L. Sabatini, Mikio Christian Aoi, and Gal Mishne. ALIGN: Adversarial learning for generalizable speech neuroprosthesis. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 8006–8025. PMLR, 2026. URL https://proceedings.mlr.press/v337/zhang26a.html.

A SUPPORTING SIGNAL ANALYSES

A.1 RECORDED-EXAMPLE MEASUREMENTS

Figure 2 uses session 3, trial 70, take 0, and session 8, trial 45, take 0. Waveforms are shown without additional filtering or amplitude normalization, with centered 51-sample moving RMS envelopes computed before cropping. Color identifies session. Full GO-to-DONE windows are 2.233 s and 1.871 s, and GO is a cue rather than a detected articulation onset. RMS and correlation annotations use these windows without time warping. In panel (b), each channel is centered and divided by its within-trial standard deviation, and all channel-0/channel-6 sample pairs use identical axis limits and equal aspect. The illustrations are conceptual, not fitted mixing models or augmentation outputs.

A.2 BETWEEN- AND WITHIN-SESSION VARIATION

Covariance maps fitted and evaluated on disjoint anchor sets (Section 2.5) showed between-session re-mixing at 7.45 times its within-session reference scale and gain at 5.07 times its own reference. All 45 between-session re-mixing values exceeded the largest within-session reference value. Early-to-late comparisons within sessions were 0.99 times the reference for re-mixing and 1.64 times for gain: channel relationships were relatively stable, while amplitude could drift within a session. These statistical differences do not identify a unique physiological cause; electrode reapplication was not varied independently of session.

A.3 PROMPT-IDENTIFICATION FEATURES

The prompt-identification analysis uses the first silent production of each of 100 anchor prompts in ten sessions, yielding 1,000 trials, with voiced productions analyzed separately. Features are extracted from the active interval after excluding the two 0.5-second margins. The headline envelope representation uses a 5 Hz low-pass envelope sampled at 100 Hz, aligned to the first and last crossings of a threshold 30% of the way from the 10th percentile to the maximum of the channel-standardized mean envelope, then resampled to 64 time points. For each held-out session, the shrinkage of the linear discriminant classifier is selected from 0.1, 0.4, and 0.8 by inner leave-one-session-out validation on the nine training sessions.

Accessible sentence and phonetic information. In a complementary 100-way identification task, onset-aligned amplitude envelopes identified known anchor prompts across held-out sessions with mean accuracy 81.67% (64%, 91% and 90% for held-out sessions 1, 5 and 6). Band-power, covariance and correlation features reached 17.00%, 3.33% and 0.67%, respectively, against a 1% chance level. These comparisons concern repeated prompts and specified readouts. They do not measure recognition of new sentences or the fraction of speech information preserved, and the envelope features contain amplitude as well as temporal structure.

A.4 PHONEME-PROBE PROTOCOL

Utterances, rather than individual frames, are split between fitting and evaluation; feature scaling and 128-dimensional PCA are fitted only on the fitting partition. Voiced probes support comparisons across network depths; the silent probe is evaluated at one depth to avoid changing its alignment across layers.

Frozen-activation probes likewise distinguished information accessible to a target-fitted readout from information accessible to a transferred readout. For silent speech at the encoder latent, mean balanced accuracy was 15.72% when a probe was fitted and evaluated on separate utterances within unseen sessions 11–13, versus 11.38% when the probe was fitted on training sessions and transferred, against a 2.55% chance level and 20.52% for a probe fitted and evaluated within the training sessions. The transferred readout therefore recovers 67% of the target-fitted probe's above-chance accuracy. In voiced recordings, the analogous values at the final transformer layer were 37.95% and 32.12%. These results are separate from sequence decoding: target-fitted probes use target-session labels, silent labels depend on alignment to voiced twins, and the voiced depth comparison cannot be taken as a silent-speech depth result.

Table 2: Sentence familiarity and WER on unseen sessions.

Sentence category	Session 14 WER (%)	Session 15 WER (%)
Repeated anchor sentences (familiar text)	49.37	36.68
New sentences (unfamiliar text)	67.93	60.83

A.5 Sentence familiarity on unseen sessions

Sentence familiarity. On unseen sessions 14–15, the historical reference model trained on sessions 1–10 (epoch 187) decoded repeated training prompts more accurately than new sentences (Table 2). All recordings came from unseen sessions and used the same greedy training-lexicon word decoder without a language model.

Anchor counts were 102/101 recordings for sessions 14/15, the separate 40-sentence calibration block is not shown, versus 903/894 new-sentence recordings; each WER pools word errors and reference words within its cell. Repeated-prompt evaluation therefore understates new-content difficulty here, although sentence content also differs between categories. Without matched within-session WER, this comparison does not quantify the cross-session penalty for either sentence category.

A.6 Uncertainty in the word-level perturbation experiment

For the Section 3.1 WER increases, 95% paired-bootstrap intervals were [2.23, 3.97] percentage points for gain, [3.49, 5.22] for mixing, and [5.68, 7.74] for both. Intervals use 10,000 paired utterance resamples and are conditional on the evaluated checkpoint, sessions, and perturbation assignment; they do not quantify variability across future sessions or training runs.

B Training, evaluation and implementation details

B.1 Study designs

Table 3: Study-specific training and evaluation designs.

Study	Training and evaluation design	Role in the comparison
Historical reference benchmark	Train s1–10; dev s11–13; test s14–15. The full training configuration includes external audio.	Chronological cross-session decoding and the main signal/representation analyses.
Session-scaling study	5 sessions: s1–5. 10 sessions: s1–5, s7–11. 15 sessions: s1–5, s7–12, s14–17. 22 sessions: s1–5, s7–12, s14–18, s20–25. Hold out 10% within each training session for development; evaluate on s6/13/19 at epoch 200.	Training coverage, checkpoint averaging, unlabeled corrections and labeled calibration with pooled-word WER.
Complementary signal study	Analyze s1–10; prompt identification holds out s1, s5 or s6. Word perturbations use held-out recordings from trained sessions s1–10.	Signal characterization and fixed-decoder sensitivity.

B.2 Acquisition and preprocessing

Signals are recorded from eight bipolar sEMG channels over facial and neck sites involved in articulation, using 16 Ag/AgCl electrodes and a separate ground. Recordings are acquired at 30 kHz with an Intan RHD headstage and stored at 1 kHz. A foot pedal marks each production, and utterance segments include 0.5 seconds of margin before and after the marked interval. These margins are part of the recorded segment and are not assumed to be physiologically silent reference periods.

For the reference encoder, mains interference is removed by spectral interpolation at harmonics detected above a 1.32 dB excess threshold for each session and channel, followed by a zero-phase,

third-order 2 Hz high-pass filter. Amplitude is scaled by a single root-mean-square value estimated over training utterances, with a tanh soft clip at 50 times that scale. Because mains detection uses session-specific measurements, this step is distinguished from the training-fitted global amplitude scale.

B.3 Model and optimization

Table 4 lists the architecture and optimization settings of the reference system.

Table 4: Reference architecture and optimization settings.

Setting	Value
Shared encoder	Six-layer transformer, hidden dimension 768
sEMG front end	Four residual blocks; downsampling by a factor of 16 to a 62.5 Hz latent sequence; final convolution with dilation 3
Audio front end	80-bin log-mel input; downsampling by a factor of two
CTC output	41 units: blank (index 0), 39 stress-stripped ARPAbet phonemes (indices 1–39), word boundary (index 40)
Parameters	67.96 million
Losses	CTC on each stream, cross-modal contrastive, latent DTW alignment; the inherited supervised temporal contrastive term is disabled
Silent-stream CTC weight	Increases from 0.2 to 1.0 over the first 100 of 200 epochs
Optimizer	AdamW (Loshchilov & Hutter, 2019); peak learning rate 3×10^{-4}; weight decay 0.1
Schedule	500 warm-up steps, then cosine learning-rate decay to 1% of the peak; 200 epochs
Regularization	Dropout 0.2; gradient-norm clipping at 1.0
Input normalization	None per channel or per utterance
Augmentation	None in the reference configuration

B.4 Interventions and decoding

Signal augmentation. Out-of-span augmentation corresponds to experiment identifier A1. The drift-augmentation condition `xs_aug` uses configuration values `drift_span=1.072` and `redraw=8000`; it is distinct from out-of-span augmentation.

Checkpoint-window augmentation results. In the sessions-14–15 experiment, out-of-span augmentation reduced average fresh-sentence WER from 72.93% to 67.19%, an improvement of 5.74 percentage points. This average covers 19 checkpoints from epochs 110–200, with lower WER at every checkpoint in both sessions. The checkpoints come from one training run per condition.

The epoch-200 augmentation conditions use the same fixed greedy decoder without a language model. Their training setup differs from the historical reference model in Appendix C.

Checkpoint averaging and latent corrections. Checkpoint averaging uses checkpoints from the cosine-decay schedule in Table 4. Floating-point tensors are accumulated in 64-bit precision and cast back to the checkpoint's storage type. For covariance alignment, each covariance is shrunk toward an isotropic matrix with the same trace, with coefficient 0.10, and the transformation preserves the target latent mean. Each component is compared within a training run against that run's own plain checkpoint.

Decoding. Both the historical and the scaling-study beam-search evaluations use Flashlight (Kahn et al., 2022) with KenLM (Heafield, 2011). Scaling-study decoder operating points are selected using up to 430 development utterances per experiment, with a 48-point coarse grid followed by 24 sequential refinement trials. These decoded-development evaluations differ from the greedy metric logged during training.

Calibration support sets. Support budgets are swept from 0 to 140 labeled utterances per target session in steps of 10, and Figure 4b plots 0, 10, 20, 40, 60, 80, 100, 120 and 140. The repeated-text

support pool comprises 40 calibration and 100 anchor utterances per session, the latter after retakes are removed. Budget and block composition move together in this pool: budgets up to 40 are drawn from the calibration block alone, and larger budgets add anchors, so a dose effect and a block effect are not separated here. All four calibration curves use one training run at each training size and pool fresh recordings from sessions 6, 13 and 19. The additional support-text comparison is described in Appendix B.6.

B.5 Scaling replication and component comparisons

Two models were independently trained at each of the four training-session counts, under the same configuration, and runs 1 and 2 identify the two series. Table 5 gives the pooled and session-specific WERs of the plain epoch-200 models. Pooled WER weights reference words rather than averaging the three session WERs.

Table 5: Fresh-sentence WER (%) for two training runs at each training-session count.

Sessions	Run	Pooled	s6	s13	s19
5	1	67.47	53.56	83.15	65.43
5	2	66.97	54.00	81.74	64.94
10	1	41.42	24.94	62.19	36.78
10	2	40.94	24.48	61.62	36.36
15	1	32.82	21.09	47.76	29.36
15	2	33.05	21.69	48.18	29.02
22	1	28.69	19.50	42.44	23.88
22	2	27.91	18.22	41.54	23.72

Table 6 gives one evaluation per trained model, that is two runs at each of the four training sizes. Every component comparison within a row uses that row's own plain checkpoint, and mean gains average the within-row differences at each training size. Figure 3c shows checkpoint averaging, while Figure 4a also shows the target-session corrections. The plain baselines in these correction evaluations differ slightly from the scaling-series values in Table 5 because the two tables report decodes of the same weights in different computing environments, and each correction is compared only against its matching baseline.

Table 6: Pooled fresh-sentence WER (%) for checkpoint averaging and latent corrections. The covariance (+Cov), rate (+Rate) and combined (+Both) columns all include checkpoint averaging.

Sessions	Run	Plain	Average	+Cov	+Rate	+Both
5	1	67.39	63.28	59.95	63.49	60.15
5	2	66.97	61.90	59.69	62.07	59.52
10	1	41.61	38.67	36.87	38.90	37.24
10	2	40.94	38.22	36.47	38.92	36.71
15	1	32.56	30.77	30.14	31.14	30.37
15	2	33.05	31.15	30.16	31.54	30.50
22	1	28.68	27.59	26.86	28.01	26.93
22	2	27.91	26.36	26.02	26.60	26.13

B.6 Target-session corrections and calibration: details

Section 3.5 summarizes these experiments. Every correction comparison is made within a run against that run's own plain checkpoint before gains are averaged across runs (Table 6). Across the eight evaluations, averaging alone reduced WER by 1.09 to 5.07 percentage points. Averaging plus rate correction still lowered WER relative to the plain checkpoint, by 4.40, 2.37, 1.47 and 0.99 percentage points at 5, 10, 15 and 22 training sessions, but less than averaging alone. The tested

temporal resampling procedure therefore provided no consistent benefit, which does not establish that temporal adaptation is ineffective in general.

Table 7 lists calibration WER at every plotted budget. Each curve pools fresh recordings from sessions 6, 13 and 19, and each model is fine-tuned on top of checkpoint averaging, covariance alignment and rate correction. The 22-session curve has a shallow minimum of 26.78% at 100 labels; this and the other intermediate minima are descriptive rather than development-selected label budgets.

Table 7: Calibration WER (%) on fresh recordings from sessions 6, 13 and 19. Column headings indicate the number of training sessions; each column uses one training run. Support consists of repeated calibration and anchor recordings.

Labels per target session	**5**	**10**	**15**	**22**
0	60.15	37.24	30.37	26.93
10	57.95	37.47	30.54	27.37
20	57.02	38.21	30.30	27.26
40	57.53	37.61	30.63	27.05
60	59.19	37.42	30.55	27.24
80	59.22	37.94	30.70	27.21
100	60.48	38.49	30.81	26.78
120	60.94	39.10	30.93	27.12
140	61.07	38.89	30.71	27.69

Support-text comparison for the 15-session model. This separate experiment compares repeated calibration/anchor support with fresh-text support on identical evaluation recordings. We reserve 140 fresh support utterances per target session by recording group, keeping retakes together, and exclude the same reserved groups from evaluation in both conditions at every label budget. Both curves therefore start at 30.80% WER on this reduced query set, rather than the 30.37% baseline in Table 7. At 140 labels, WER reaches 31.10% with repeated-text support and 34.42% with fresh-text support. Fresh-text support does not produce a sustained improvement under this configuration. Repeated support comes from approximately the first 14% of the session, whereas fresh support is distributed through the session, so the comparison changes both text familiarity and temporal sampling. It does not isolate either factor or establish an overfitting mechanism.

B.7 Reporting conventions

Partial evaluations retain their actual session subset and are not substituted for the complete three-session aggregate. An improvement of $100 \times (\mathrm{error}_{\mathrm{baseline}} - \mathrm{error}_{\mathrm{method}})$ is reported in percentage points, and absolute baseline and method error rates accompany the improvement. Scaling-study correction gains and calibration changes are calculated from the reported WER values; means and differences are then rounded to two decimal places. In the historical training pipeline, the seed flag also changes the sentence split, so comparisons across seeds must account for both initialization and split variation. Epoch-specific effects, checkpoint-window means, and evaluations of averaged model weights are different quantities and are labeled separately. Paired utterance-bootstrap intervals for the fixed-checkpoint perturbation experiment describe uncertainty over the evaluated utterances, conditional on those sessions and models; they do not quantify variability across new training runs or future recording sessions.

C Supporting correction and calibration comparisons

These comparisons retain their original protocols and controls and are not pooled with the scaling series in Figures 3b–c and 4.

Historical reference model. The reference model used for the sentence-familiarity and historical correction/calibration analyses was trained from scratch. Valid sequence lengths are propagated through the network, and normalization statistics exclude padded positions using a two-pass variance calculation in 32-bit precision.

Historical decoding settings. The historical reference/correction evaluations use beam size 1,000, a base LM weight of approximately 4.8663, and a base word score of -1.7951. A development-selected scale multiplies both values; the scale used for each system is reported with its results. Lexicon resources and selection policies therefore remain part of each comparison's protocol.

Correcting internal representations and using labeled calibration. Latent covariance alignment adjusts the relationships among learned features, while temporal resampling changes the number of latent frames supplied to the sequence decoder. We combined these operations with averaging of late training checkpoints to form a three-component correction stack. Checkpoint averaging requires no target-session recordings; covariance alignment and temporal resampling estimate statistics from the whole target session before decoding. The stack thus uses unlabeled target data in an offline, transductive setting. Its duration-based resampling estimate also assumes that the target and training sessions have comparable sentence-length distributions.

Stack and calibration settings. The stack averages 25 checkpoints from epochs 80–200. The covariance reference is the mean of the ten source sessions' within-session frame covariances; target covariance and duration-based resampling statistics are estimated from the whole unlabeled target session. Historical labeled calibration fine-tunes all model parameters at learning rate 0.00001 for 300 steps, with 40% source-session replay. Support contains 40 calibration examples per target session; fresh query recordings remain separate. These optimization settings describe this historical study.

For the word-level transfer evaluation, the reference model was trained on sessions 1–10, with methods and hyperparameters selected on development sessions 11–13, and evaluated on fresh sentences from sessions 14 and 15. All systems use the same reference model and the 5-gram decoder. The reference model and unlabeled stack share LM scale 1.00; both 40-label systems use 1.25. Aggregate WER weights the two session WERs by fresh-recording counts (Section 2.6); improvements are calculated before rounding. Alongside the unlabeled stack, we evaluated fine-tuning with 40 labeled calibration sentences per target session, scored on the separate fresh-sentence set. Table 8 summarizes these comparisons.

Table 8: Historical target-session correction and labeled calibration.

System	Session 14 WER (%)	Session 15 WER (%)	Aggregate WER (%)	Improvement over reference (percentage points)
Reference model	59.68	52.85	56.28	—
+ Unlabeled correction stack	55.44	49.82	52.65	3.64
+ 40 labeled calibration sentences	54.56	51.04	52.81	3.48
+ Stack and 40 labeled calibration sentences	52.66	48.45	50.57	5.71

The unlabeled stack and labeled calibration each improved WER on both sessions. Combining the stack with 40 labeled calibration sentences reduced aggregate WER by 5.71 percentage points relative to the reference model. The unlabeled-stack improvement holds the decoder settings fixed. The labeled-calibration comparisons also change LM scale relative to the reference, so they do not isolate the effect of calibration. The full-stack comparison does not isolate each component's word-level contribution. Sessions 14 and 15 were evaluated repeatedly across the study, so these historical results are transfer checks rather than a single untouched final evaluation.