

# Hold Out the Vocal Unit: Choosing Neural Decoders for the Variation They Must Survive

Brandon Fisher, Wesley Harper, Blake Harrison

## Abstract

A neural vocalization decoder can perform well on randomly held-out observations while depending on regularities that disappear when an entire vocal unit is absent from training. This article develops an evidence framework around structured generalization error rather than a new decoder experiment. A study of zebra finch vocalizations provides the motivating comparison: the apparent ranking of neural representations depends on how the evaluation withholds vocal material. We distinguish interpolation, unseen-unit transfer, session transfer, and translation to human communication as different empirical claims. The proposed evaluation record joins each split to a defined nuisance variation, fits every learned representation within the training boundary, and reports uncertainty at the level of genuinely independent recordings or subjects. Efficient adaptation may make these comparisons affordable, but cannot establish transfer by itself. A useful decoder benchmark should therefore identify the variation that a deployed system must survive before choosing the split on which it will be scored.

Keywords: neural decoding; vocalization; structured generalization; latent dynamics; evaluation design; brain-computer interfaces.

## 1. Similar examples do not answer every deployment question

Neural decoding research often begins with a sensible engineering objective: recover a behavioral signal from recorded neural activity as accurately as possible. The evaluation becomes more difficult when the desired behavior differs systematically from the examples used for training. Guiding Brain-to-Vocalization Decoder Design Using Structured Generalization Error makes this problem concrete through a zebra finch setting (J. Huang et al., 2024). Its comparison of neural representations motivates a broader methodological question: what type of future observation is a reported error intended to represent? A numerical score has little transportable meaning until that observation is specified.

The distinction is especially important when repeated behavior provides many superficially separate examples. A recording can contain multiple renditions of similar vocal units, adjacent windows from one rendition, and sessions sharing the same animal and recording apparatus. Randomly assigning windows across training and testing may create a difficult signal reconstruction problem, yet still permit the model to recognize regularities that are already represented in training. Holding out an entire vocal unit changes the information available for reconstruction. Neither split is universally correct. They answer different questions about interpolation and transfer and should be described in those terms.

This article is a methodological synthesis. It does not download bird recordings, fit a latent model, retrain a decoder, or report new performance measurements. The motivating study is used to organize evaluation choices, not to claim that one representation is optimal for every species or recording system. In the reported comparison, latent-factor and firing-rate representations have advantages over spikes under a structured held-out-unit task, while the random-holdout comparison is less discriminating. The relevant lesson is that the test construction can change what an architectural comparison reveals. Broader deployment claims require additional evidence beyond that particular comparison.

## **2. Define the future observation before the split**

An interpolation task asks whether a decoder can reconstruct further examples drawn from behavior already represented in its training experience. An unseen-unit task asks whether it can recover a vocal element whose examples were withheld. A session-transfer task adds changes in recording conditions, neural sampling, and behavior. A subject-transfer task changes the biological system itself. These targets can be arranged as a sequence of increasingly different information environments, but their difficulty is not guaranteed to follow a strict numerical ordering. A new session can be easier than an unusual example from the original session, depending on the signal and the decoder.

A useful evaluation plan states which sources of variation are allowed to remain shared. For a within-session held-out-unit comparison, it may be legitimate to share recording hardware, preprocessing conventions, and the set of recorded channels. It may not be legitimate to learn a behavior-specific representation from the withheld unit and then call the resulting decoder an unseen-unit model. The plan should also specify whether temporal context immediately surrounding the withheld unit is available. That context can contain information about its identity or timing. Context is not inherently an error; it becomes a problem when the reported claim suggests that the model succeeded without information it actually used.

Structured cross-validation research in other domains emphasizes that dependency structure should inform the withholding scheme (Roberts et al., 2017). The same principle applies here, although neural recordings require their own units and scientific interpretation. A block should correspond to the nuisance variation or novelty that the study wishes to test. Grouping examples only because the software supports a convenient identifier is insufficient. The investigator should be able to explain why the held-out group resembles the observation expected during later use and why the remaining training data would legitimately be available at that moment.

## **3. Representations must remain inside the training boundary**

Latent dynamical representations can combine information across neurons and time in ways that expose useful population structure. LFADS provides an influential example of learning latent factors from neural population activity (Pandarinath et al., 2018). Such representations can improve a decoder's access to shared dynamics, but their training process is itself part of the prediction system. If the latent model sees the complete recording before the supervised decoder is split, the evaluation no longer tests the entire pipeline on unseen material. An unsupervised learning objective does not make information exposure irrelevant to a generalization claim.

The appropriate boundary depends on the scientific task. A transductive task may explicitly allow unlabeled neural activity from the future test session when learning a representation. An inductive task may forbid it. Both can be studied, but they should be reported separately because they offer the decoder different resources. The same applies to normalization, channel selection, alignment, dimensionality selection, and temporal smoothing. A transformation estimated from all observations can carry information about the test distribution even when no behavioral labels were provided. Repeating only the final regression within each training fold does not repair a representation learned outside that fold.

Hyperparameter selection needs a second boundary. Choosing a latent dimension, decoder depth, or smoothing window after comparing results on the final held-out unit uses that unit as development data. Nested evaluation separates tuning from the final error estimate, a distinction illustrated by work on bias in cross-validation after model selection (Varma and Simon, 2006). The outer split should encode the intended generalization problem, while inner splits choose hyperparameters using only the outer training

data. Small datasets may make this procedure expensive or unstable. That limitation should be reported as uncertainty about model selection rather than resolved by quietly reusing the test material.

An informative comparison would preserve the same outer split across spikes, rates, and latent factors. It would record the number of training examples, temporal support of each feature, and the computational resources needed to generate that feature. A smoothed rate representation can use a wider temporal neighborhood than a spike-based feature, and a latent model can encode additional context through its state. If these differences are intentional, they belong in the comparison. If they are accidental, a representation ranking may partly reflect unequal information access. Matching nominal input dimensions alone does not equalize the temporal information supplied to the decoder.

#### **4. Error should identify a failed capability**

A reconstruction error averaged over all samples can obscure whether a decoder preserves the distinctions needed for communication. Timing, acoustic envelope, spectral structure, and unit identity are related but separable properties. A system might reproduce an average pattern while losing the contrast between two vocal elements. Alternatively, it might identify the intended unit while introducing timing jitter. A benchmark should state which of these properties its primary outcome measures and which require secondary outcomes. Adding many scores without a designated primary question can turn a comparison into an opportunity to select whichever measure favors a preferred model.

Per-unit reporting is particularly useful when units differ in frequency and complexity. A pooled error weights frequent units more heavily unless the analysis explicitly reweights them. Equal unit weighting changes the estimand to performance on an average unit rather than an average recorded sample. Neither quantity is intrinsically preferable, but they need different interpretations. A deployment context with rare but important vocal distinctions may justify reporting both. Uncertainty intervals should respect the sampling units that generated the data. Thousands of adjacent time bins do not create thousands of independent animals or independently repeated recording conditions.

Error calibration is also different from average accuracy. If a decoder reports confidence in a unit label or a downstream recognizer's output, that confidence should be evaluated on the same structured holdout that defines the deployment target. Calibration methods developed for modern neural networks supply useful tools, but their effectiveness must be checked under the relevant shift (Guo et al., 2017). A model calibrated on familiar units may still be confidently wrong on an unseen one. An abstention or confirmation mechanism would require its own evaluation of coverage, error, and the consequences of delayed output. Calibration cannot be inferred from a visually plausible spectrogram.

The choice of temporal alignment deserves an explicit record. Optimizing alignment separately for every test output can remove errors that a real-time user would experience as latency or incorrect timing. Some analyses legitimately compare shapes after alignment; others need the original time axis. The distinction changes what a low error means. Reporting both aligned shape quality and unaligned timing quality would expose the tradeoff when it matters. The same principle applies to preprocessing that removes silence or difficult transitions: the removed material should remain visible in the coverage description rather than disappearing from the definition of success.

## 5. Biological transfer is a separate evidentiary step

Birdsong offers a valuable setting for studying patterned neural activity and structured vocal behavior. It does not directly establish performance for human speech restoration. Human speech neuroprosthesis studies demonstrate that neural signals can support communication systems under their own experimental conditions (Willett et al., 2023). Other work combines speech output and avatar control in a participant-specific system (Metzger et al., 2023). These are relevant examples of the broader goal, but their outcomes cannot be substituted for measurements in the bird study, nor can a bird decoder's generalization ranking be assumed to persist across human cortical recordings and language tasks.

The required transfer argument should identify which mechanism is expected to carry over. A population representation may suppress channel-specific noise, exploit repeated motor structure, or provide a smoother state for a downstream decoder. These explanations have different observable implications. A study that withholds vocal units primarily tests one aspect of behavioral novelty; it does not automatically test electrode drift, altered speaking strategies, or changes in the user's condition. Speech synthesis from decoded neural activity illustrates how a complete communication pipeline combines several modeled components (Anumanchipalli et al., 2019). Its existence does not establish that another decoder is stable across recording changes. Such stability needs direct measurement.

Clinical relevance also changes the loss function. A reconstruction score may not capture the cost of an incorrect word, a delayed response, or an unintended activation. User adaptation and decoder adaptation can interact, making an offline test an incomplete description of long-term use. A prospective system evaluation would therefore need to specify assistance, correction opportunities, feedback, and the time allowed for learning. None of these requirements invalidates an offline structured-holdout experiment. They define the additional steps between evidence about a decoding component and evidence about a usable communication system.

## 6. Efficient adaptation under a fixed evaluation budget

Repeatedly fitting representations inside nested structured splits can be computationally demanding. Low-rank adaptation is relevant when a large pretrained component forms part of the decoding or acoustic modeling pipeline (Hu et al., 2022). The Hadamard Tensor Ring approach motivates investigating whether more economical fine-tuning can support such repeated evaluations (H. Tong et al., 2026). That is a proposed use rather than a demonstrated result for neural vocalization data. An adaptation method's reported efficiency in its original setting does not establish the same efficiency or accuracy tradeoff for a new neural input representation.

A fair study would specify which weights are adapted, how many trainable parameters are introduced, and what preprocessing remains fixed. It would compare adaptation methods under matched outer splits and equivalent access to training material. Wall-clock time, peak memory, and the cost of selecting adapter settings should be included. A method that trains quickly after extensive tuning can still require a large total research budget. Equally, an efficient method may allow more complete evaluation across sessions even if its best single-split score is unchanged. That benefit should be measured as evaluation coverage rather than presented as improved biological generalization.

The adaptation boundary must remain auditable. A pretrained acoustic component can encode language or sound regularities useful for reconstruction, while the neural adapter learns participant-specific mappings. Such prior knowledge is a legitimate resource when disclosed. It becomes difficult to interpret when different models receive different pretrained corpora or when unit-specific material enters pretraining without documentation. The record should identify these resources and distinguish reconstruction

supported by the neural signal from reconstruction supported by a strong prior. Ablation tests can help, but their conclusions depend on retaining the same evaluation target and scoring conventions.

## 7. A minimum record for an interpretable benchmark

The proposed reporting record begins with a sentence defining the future observation: another rendition, a new vocal unit, another session, or a new subject. It then identifies the grouping variables, permissible shared information, and all learned preprocessing. A split manifest should preserve group membership and the order in which transformations were fitted. The record should include the tuning procedure, failed runs, exclusion rules, and the temporal support of predictions. These details allow a reader to understand why a score estimates the intended capability. They are more informative than a random seed listed without the construction of the corresponding split.

The benchmark should finish with a failure description linked to that capability. If a model reconstructs familiar units well but fails on a withheld one, the result can guide representation design without implying that the model is generally unusable. If structured and random splits produce similar results, the study should examine whether the withheld structure actually introduces the intended novelty. The motivating brain-to-vocalization work is valuable because it places that relationship between split and representation at the center of the design question. Choosing a neural decoder begins with specifying what must remain reliable when the data change, and then collecting evidence for precisely that change.

## References

- Huang, Jingya, et al. "Guiding Brain-to-Vocalization Decoder Design Using Structured Generalization Error." *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1-4. <https://doi.org/10.1109/embc53108.2024.10782761>.
- Tong, H., Xu, G., & Chen, Y. (2026, May). Hadamard Tensor Ring for Efficient Low-Rank Fine-Tuning. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. <https://doi.org/10.1111/ecog.02881>.
- Pandarinath, Chethan, et al. "Inferring single-trial neural population dynamics using sequential auto-encoders." *Nature Methods*, vol. 15, no. 10, 2018, pp. 805-815. <https://doi.org/10.1038/s41592-018-0109-9>.
- Varma, Sudhir and Richard Simon. "Bias in error estimation when using cross-validation for model selection." *BMC Bioinformatics*, vol. 7, no. 1, 2006. <https://doi.org/10.1186/1471-2105-7-91>.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>.
- Willet, Francis R., et al. "A high-performance speech neuroprosthesis." *Nature*, vol. 620, no. 7976, 2023, pp. 1031-1036. <https://doi.org/10.1038/s41586-023-06377-x>.
- Metzger, Sean L., et al. "A high-performance neuroprosthesis for speech decoding and avatar control." *Nature*, vol. 620, no. 7976, 2023, pp. 1037-1046. <https://doi.org/10.1038/s41586-023-06443-4>.
- Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang. "Speech synthesis from neural decoding of spoken sentences." *Nature*, vol. 568, no. 7753, 2019, pp. 493-498. <https://doi.org/10.1038/s41586-019-1119-1>.
- Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*, 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.