Transfer Learning for Robust Neural Speech Decoding Across Subjects Using Uncertainty-Aware Representations and Adaptive Calibration

Caragh Morahan
School of Computing, Clemson University, Clemson, SC, USA.
caragh.morahan588@clemson.edu

Eitel Kraushaar
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
eitelkraushaar72@binghamton.edu

Skyla Searle
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
searle122@uc.edu

Hermelinda Archuleta
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
hermelinda.archuleta@colostate.edu

Abstract

Neural speech decoding has advanced rapidly, yet decoders trained on one participant often degrade when transferred to another because electrode geometry, neural dynamics, articulation strategies, and recording stability shift the input distribution and the mapping to speech units. This article develops a transfer-learning framework for cross-subject neural speech decoding that treats uncertainty-aware representation learning and adaptive calibration as coupled design requirements rather than post hoc corrections. The framework specifies a subject-conditioned latent decoder, decomposes predictive uncertainty into aleatoric and epistemic components, aligns second-order latent statistics across subjects with uncertainty-weighted penalties, and calibrates subject-specific output probabilities through constrained affine temperature and bias terms. A risk-controlled decision rule then converts calibrated probabilities and residual uncertainty into selective decoding decisions under an explicit error budget. The central claim is analytical: robust transfer in neural speech decoding requires preserving label-relevant subject variation while suppressing nuisance variation, and calibration must be learned as a transferable component because confidence, rather than accuracy alone, determines whether decoded speech can be acted upon. The framework also identifies failure modes, including negative transfer from over-alignment, unreliable uncertainty estimates under small target samples, and calibration guarantees that weaken when cross-subject exchangeability fails. The resulting design provides a technically grounded basis for cross-subject decoder evaluation and for integrating uncertainty into neural speech interfaces.

Keywords

neural speech decoding; transfer learning; uncertainty quantification; adaptive calibration; cross-subject generalization; selective prediction

1. Introduction

Neural speech decoding aims to convert recorded brain activity into linguistic units or control signals for communication. Brain-computer interfaces have long been motivated by the needs of users with severe motor impairment [1], and recent work has moved from isolated classification tasks toward continuous speech synthesis and neuroprosthetic communication [2], [3]. High-performance demonstrations now include neural-to-text and neural-to-speech systems that exploit cortical population dynamics with substantial temporal structure [4], [5]. These advances establish the feasibility of neural speech decoding, but they also expose a transfer problem: a decoder tuned to one participant's electrode array, cortical coverage, and articulatory strategy is not automatically valid for another participant, another session, or another recording condition.

Cross-subject transfer is difficult because the observed neural representation is not a stable, subject-independent encoding of speech. Electrode placement changes the spatial sampling of neural populations; tissue response and impedance changes alter signal quality; and speakers differ in articulation kinematics, speaking rate, and the neural strategies that support phonation and articulation. Transfer learning is a natural response to this setting [6], and domain adaptation has produced several mature strategies, including adversarial feature invariance [7], second-order covariance alignment [8], and probabilistic treatments of model uncertainty [9]. Yet neural speech decoding imposes constraints that generic adaptation methods do not address well. The label space is structured by phonetics and articulation, the temporal dependence is strong, and the cost of confident errors is high because a decoded command may be used for communication or control.

Structured generalization-error analyses for brain-to-vocalization decoders further suggest that error is not uniform across phonemes, articulatory features, or temporal positions, and that decoder design can be guided by how these error structures transfer across conditions [10]. This observation motivates a framework in which representation alignment is selective rather than global: the goal is not to erase subject identity, but to separate label-relevant variation from nuisance variation. It also motivates uncertainty-aware training. Deep ensembles provide a practical route to predictive uncertainty [11], while modern neural networks are often poorly calibrated even when accurate [12]. Distribution-free uncertainty methods such as conformal prediction offer risk control under explicit assumptions [13], but their guarantees depend on exchangeability, an assumption that cross-subject neural data can violate.

This article proposes a transfer-learning framework for robust neural speech decoding across subjects using uncertainty-aware representations and adaptive calibration. The contribution is a coherent methodological design and analytical account, not a report of a completed empirical benchmark. The framework has four linked components. First, a subject-conditioned latent decoder represents speech-related factors separately from subject-specific observation effects. Second, predictive uncertainty is decomposed into aleatoric and epistemic parts and used to weight representation alignment. Third, adaptive calibration learns subject-specific affine corrections in log-probability space with hierarchical shrinkage toward a pooled calibration function. Fourth, a risk-controlled decision rule uses calibrated probabilities and residual uncertainty to decide when a decoded output should be accepted. The argument is that accuracy alone is an insufficient target for cross-subject neural speech decoding; transferable confidence is a first-class requirement.

2. Problem Formulation and Sources of Cross-Subject Shift

Let x denote a neural feature sequence from a participant s and let y denote a speech target sequence, such as phonemes, articulatory features, or words. A subject-conditioned decoder defines a predictive distribution p(y | x, s). Because speech targets are generated through latent articulatory and linguistic factors that are only partially observed, the decoder can be written with an explicit latent variable z:

$$p(y \mid x, s) = \int p(y \mid z, \theta)\, p(z \mid x, s)\, dz$$

Here z is a latent representation intended to capture speech-relevant structure, θ denotes decoder parameters, and s indexes the subject or recording domain. The marginalization states that decoding is not a direct deterministic map from electrodes to labels; it passes through an intermediate representation whose distribution can depend on the subject. This formulation is useful because it separates two questions: whether z retains speech-relevant information, and whether the mapping from z to y is stable across subjects.

Cross-subject shift can enter through p(z | x, s), through p(y | z, θ), or through both. If electrode placement changes the observation model, the same speech factor may appear with different spatial and spectral signatures, shifting p(z | x, s). If articulation strategies differ, the conditional mapping from a similar latent state to a phonetic target may also shift. A transfer method that aligns only marginal feature distributions can therefore remove discriminative information when subject-specific variation is partly label-relevant. This is a central failure mode in neural speech decoding: over-alignment may make subjects indistinguishable in latent space while degrading the decoder's ability to distinguish phonetic classes.

A second source of difficulty is uncertainty. Neural recordings are noisy, labels may be sparse for a target subject, and some speech units are inherently ambiguous from the available cortical coverage. A decoder that outputs a point prediction hides this ambiguity. For communication interfaces, the relevant object is often a calibrated predictive distribution and a decision rule that can abstain or request confirmation. The framework therefore treats uncertainty as part of the representation and transfer problem, not as a separate visualization after training.

3. Uncertainty-Aware Transfer Framework

3.1 Subject-conditioned latent representation

The encoder maps a neural feature sequence and subject context to a distribution over latent speech factors. A practical parameterization uses a diagonal Gaussian:

$$q\varphi(z \mid x, s) = N(\, \mu\varphi(x, s), diag(\sigma\varphi^2(x, s))\,)$$

The encoder parameters φ produce a latent mean μφ and a variance σφ² for each latent dimension. The subject context s can be supplied as a learned embedding or as structured metadata such as electrode layout and session indicators. The variance term allows the representation to express uncertainty about a latent factor rather than forcing all variation into a deterministic code. This is important when a target subject has few labeled trials: the model can represent ambiguity in latent speech factors instead of overcommitting to a subject-specific mapping.

The decoder p(y | z, θ) then maps latent samples to speech targets. For sequence targets, the decoder can use temporal attention [19] or recurrent memory [20], depending on the duration and sampling characteristics of the neural features. The choice of backbone is

secondary to the design requirement: the latent representation must be evaluated for both label relevance and subject invariance. A representation that is invariant but not discriminative is useless for decoding, while a representation that is discriminative but brittle will not transfer.

3.2 Uncertainty decomposition

Predictive uncertainty can be decomposed into aleatoric and epistemic components. Aleatoric uncertainty reflects irreducible noise or ambiguity in the neural signal and target mapping; epistemic uncertainty reflects uncertainty about model parameters and is especially relevant when target-subject data are scarce. For a latent-variable decoder, the total predictive variance for a scalar output can be expressed as:

$$\mathrm{Var}(y|x,s) = E_q\left[\sigma_y^2(z)\right] + \mathrm{Var}_q\left[\mu_y(z)\right]$$

The first term is the expected conditional variance and represents aleatoric uncertainty. The second term is the variance of the predicted mean across latent states and represents epistemic uncertainty. In practice, the expectation and variance can be approximated by sampling multiple latent states and, when ensembles are used, multiple parameter settings [11]. Monte Carlo dropout provides a lower-cost approximation to Bayesian model uncertainty [18], while variational training supplies a principled latent distribution [17]. These methods are not interchangeable: ensemble disagreement captures parameter uncertainty, whereas latent sampling captures ambiguity in the intermediate speech representation.

The decomposition affects transfer because the two uncertainty types have different implications for alignment. High aleatoric uncertainty in a latent dimension may indicate noisy or uninformative neural features; aligning such dimensions aggressively can propagate noise across subjects. High epistemic uncertainty may indicate that the target subject is outside the source distribution; in that case, adaptation should be conservative and calibration should widen predictive intervals. An uncertainty-aware alignment penalty can therefore down-weight dimensions or samples whose representation is unreliable, reducing negative transfer.

3.3 Uncertainty-weighted cross-subject alignment

Let μ_s and Σ_s denote the mean and covariance of latent representations for subject s, and let μ_0 and Σ_0 denote a reference distribution formed from pooled source subjects or a designated anchor subject. A simple second-order alignment objective is:

$$L_align = \sum_s \left(||\mu_s - \mu_0||^2 + ||\Sigma_s - \Sigma_0||_F^2 \right)$$

Here $||\cdot||^2$ is the squared Euclidean norm and $||\cdot||_F^2$ is the squared Frobenius norm. The first term aligns first-order statistics; the second term aligns second-order statistics. This objective is related in spirit to covariance alignment methods [8] and to moment-matching adaptation, but it is used here with subject-specific weighting and diagnostic constraints. The reference distribution should be chosen to preserve class-conditional structure. When target labels are available, alignment can be performed within predicted or known phonetic classes; when labels are unavailable, pseudo-labels must be filtered by calibrated confidence and uncertainty to avoid reinforcing errors.

Adversarial domain-invariance methods [7], deep adaptation networks [14], adversarial discriminative adaptation [15], and maximum classifier discrepancy [16] offer alternative routes to cross-domain invariance. They can be effective when domains are well matched and labels are abundant, but they also risk removing subject-specific information that is predictive of speech. The proposed alignment is deliberately narrower: it aligns global and class-

conditional latent moments while retaining subject-conditioned variance heads and uncertainty weights. The design assumption is that nuisance variation is more likely to appear in low-confidence or high-variance latent directions, whereas stable speech factors should remain discriminative across subjects.

3.4 Adaptive calibration

Even a well-transferred decoder can produce poorly calibrated probabilities. Neural networks often become overconfident under distribution shift [12], and cross-subject shift is a particularly strong form of shift because both inputs and label conditionals can change. Adaptive calibration introduces subject-specific affine corrections in log-probability space:

$$p_cal(k \mid x, s) = softmax_k(\, a_s \, log \, p(k \mid x, s) + b_s \,)$$

The scalar a_s controls the sharpness of the subject's predictive distribution, and the vector b_s controls class-specific bias. For a target subject with few labels, a_s and b_s should not be estimated freely. A hierarchical prior can shrink them toward pooled values, with the amount of shrinkage determined by the number of labeled target trials and the estimated epistemic uncertainty. This is more flexible than global temperature scaling and more stable than unconstrained per-subject calibration.

Calibration is a transferable component because it corrects the relationship between confidence and correctness rather than the decoder's representational geometry. A subject may have a shifted class prior or a different noise level while retaining a useful latent decoder. Adaptive calibration can correct such shifts without retraining the entire network. However, calibration cannot repair a decoder that has lost label-relevant information. It should therefore be applied after representation transfer and evaluated separately from decoding accuracy.

3.5 Risk-controlled decision rule

For communication interfaces, the decoder should sometimes abstain rather than emit a low-confidence output. Let α be a tolerable error rate among accepted predictions and let u_s be a residual uncertainty score for the subject or trial. A simple selective decision rule is:

$$accept(x, s) = 1[\, max_k \, p_cal(k \mid x, s) \geq \tau(\alpha, u_s) \,]$$

The threshold τ can be estimated by conformal risk control on calibration data [13], with the important caveat that exchangeability may fail across subjects. Group-conditional or subject-adaptive thresholds are therefore preferable to a single global threshold. The indicator function returns one when the calibrated maximum probability exceeds the threshold and zero when the decoder should abstain or request confirmation. The uncertainty score u_s can widen the threshold when the target subject is poorly represented, producing conservative behavior under distribution shift.

The training objective combines prediction, alignment, uncertainty, calibration, and smoothness terms:

$$L = L_pred + \lambda_1 \, L_align + \lambda_2 \, L_unc + \lambda_3 \, L_cal + \lambda_4 \, R_smooth$$

L_pred is the speech-target loss, such as connectionist temporal classification or cross-entropy for frame-level targets. L_align is the moment-alignment term. L_unc penalizes overconfident variance estimates and encourages reliable uncertainty decomposition. L_cal is a calibration surrogate, such as a Brier score or differentiable expected calibration error. R_smooth regularizes temporal predictions to respect the continuity of articulation. The

coefficients λ_1 through λ_4 control trade-offs; they should be selected by held-out subject performance and calibration, not by source accuracy alone.

4. Analytical Assessment and Technical Discussion

The framework's central analytical claim is that cross-subject robustness depends on the interaction between representation invariance and calibrated confidence. This claim has several consequences. First, alignment strength should be conditional on label relevance. If latent moments are aligned without preserving class-conditional structure, the decoder may become subject-invariant but phonetically indiscriminative. The uncertainty-weighted alignment term reduces this risk by down-weighting unreliable dimensions, but it does not eliminate it. A diagnostic test should therefore compare class-conditional latent distances before and after alignment. If phonetic classes collapse across subjects while within-class variance increases, the adaptation is too aggressive.

Second, uncertainty estimates must themselves be validated. Deep ensembles and variational approximations can produce useful uncertainty scores [11], [17], [18], but they are not automatically calibrated under cross-subject shift. A model may be uncertain for the wrong reason, such as electrode noise rather than unfamiliar speech content. The framework's use of uncertainty to weight alignment and set decision thresholds therefore requires an audit of uncertainty-error correlation. If epistemic uncertainty is not predictive of transfer error, the weighting scheme can degrade performance. This is a boundary condition rather than a flaw in the general design: uncertainty-aware transfer is only as reliable as the uncertainty estimator.

Third, adaptive calibration introduces a trade-off between subject specificity and statistical stability. A target subject with many labeled trials can support a flexible calibration function; a subject with few trials requires strong shrinkage toward pooled parameters. The hierarchical formulation makes this trade-off explicit. It also clarifies why calibration should be evaluated with proper scoring rules and reliability diagnostics rather than accuracy alone. A decoder that is slightly less accurate but substantially better calibrated may be preferable for selective communication because it allows the user or system to avoid confident errors.

Fourth, risk-controlled decisions depend on assumptions that cross-subject data can violate. Conformal prediction provides distribution-free guarantees under exchangeability [13], but a new subject is not exchangeable with source subjects in a strict statistical sense. Adaptive or group-conditional conformal methods can mitigate this problem, but they require enough calibration data per group. When such data are unavailable, the threshold should be treated as a conservative heuristic rather than a formal guarantee. This distinction matters for clinical and assistive use: a decision rule that appears to control risk on pooled data may not control risk for a subject with a different electrode configuration or pathology.

Comparisons with alternative transfer strategies clarify the design. Global fine-tuning on target data can be effective when labels are plentiful, but it offers no mechanism for separating nuisance variation from speech-relevant variation and can overfit small target samples. Domain-adversarial training [7] and maximum classifier discrepancy [16] can learn invariant features, but they may remove subject-specific information that is useful for articulation. Covariance alignment [8] and deep adaptation networks [14] are computationally attractive, yet they do not by themselves provide calibrated confidence or selective prediction. The proposed framework combines moment alignment with uncertainty weighting and

calibration because these components address different failure modes: representation shift, unreliable confidence, and unsafe decisions.

Evaluation of such a framework should use leave-one-subject-out protocols and report accuracy, negative log-likelihood, Brier score, expected calibration error, and coverage-risk curves for selective prediction. It should also report uncertainty decomposition diagnostics and class-conditional transfer error, because aggregate accuracy can hide systematic failures on specific phonetic or articulatory categories. The structured error perspective is especially relevant here: if error concentrates in particular phonological classes or temporal positions, alignment and calibration should be analyzed at that level rather than only at the level of whole utterances [10]. These evaluation choices follow from the framework's logic and do not depend on a particular dataset or reported benchmark.

5. Conclusion

This article has developed a transfer-learning framework for robust neural speech decoding across subjects that treats uncertainty-aware representation learning and adaptive calibration as linked requirements. The framework represents speech through a subject-conditioned latent variable, decomposes predictive uncertainty into aleatoric and epistemic components, aligns cross-subject latent moments with uncertainty-sensitive penalties, calibrates subject-specific probabilities through constrained affine corrections, and converts calibrated confidence into risk-controlled decoding decisions. The analytical contribution is a design account of why these components must be coordinated: alignment without discriminative preservation can cause negative transfer, calibration without representation transfer cannot recover lost speech information, and decision thresholds without validated uncertainty cannot support reliable communication interfaces. The framework also identifies the conditions under which its guarantees weaken, particularly when target-subject data are scarce, uncertainty estimates are miscalibrated, or exchangeability across subjects fails. These boundary conditions define the technical space in which cross-subject neural speech decoding can be made more robust and more trustworthy.

References

1. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. Clinical Neurophysiology, 113(6), 767-791.

2. Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568(7753), 493-498.

3. Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3), 217-227.

4. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. Nature, 593(7858), 249-254.

5. Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., ... & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620(7976), 1037-1046.

6. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

7. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, M., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1-35.

8. Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 2058-2065.

9. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584.

10. Huang, J., Tostado-Marcos, P., Narasimha, S. M., Patel, A. N., Arneodo, E. M., Gentner, T. Q., ... & Gilja, V. (2024, July). Guiding brain-to-vocalization decoder design using structured generalization error. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-4). IEEE.

11. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402-6413.

12. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330.

13. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

14. Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on Machine Learning, 37, 97-105.

15. Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7167-7176.

16. Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3723-3732.

17. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

18. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1050-1058.

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

20. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.