PRETRAINING DOMAIN SHAPES AUDITORY NEURAL PREDICTIVITY DIFFERENTLY ACROSS SPECIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have successfully captured aspects of neural representation in both humans and nonhuman primates, particularly in vision. However, whether such generalization extends to species-specific domains, such as speech, remains unclear. We tested self-supervised and supervised models spanning training corpora from human speech to animal vocalizations on both animal and human neural responses prediction tasks. Across all animal neural-prediction tasks, models achieved nearly identical best-layer encoding accuracy, however, on human conditions, diverged across domains. To examine whether this convergence versus divergence pattern is already present at the level of model representations, we applied unbiased Centered Kernel Alignment (uCKA) to the activations and observed a pretraining-corpora-dependent input geometry across all datasets. To investigate why this geometry is not fully reflected in encoding accuracy, we introduced a sensitive subspace analysis to the frozen predictor and found pretraining-corpora-dependence in speech models disappeared on animal datasets but preserved on human datasets. Our findings demonstrate a species-dependent effect of pretraining domain and, using speech as a case study, explain how brain filters out domain-specific model representations absent from that species.

1 INTRODUCTION

Deep neural networks (DNNs) have become powerful computational models of sensory cortex. In vision, object-recognition networks capture aspects of neural response profiles and hierarchical organization in the human ventral stream (Güçlü & Van Gerven, 2015; Eickenberg et al., 2017; Ratan Murty et al., 2021; Cao et al., 2026), and also predict responses in nonhuman visual cortex, particularly in primates with homologous visual systems shaped by broadly similar sensory environments (Yamins et al., 2014; Nayebi et al., 2018; Xiao et al., 2024; Ramos-Traslosheros & Ponce, 2026). Similar progress has emerged in audition, where speech-optimized DNNs can predict responses in the human auditory cortex (Kell et al., 2018; Vaidya et al., 2022; Millet et al., 2022; Li et al., 2023). Speech, however, is a human-specific communication domain and is absent from the natural auditory experience of nonhuman animals. It therefore remains unclear whether neural predictivity depends on pretraining in a species-matched auditory domain, or whether both representations learned from speech and animal vocalizations generalize similarly across human and nonhuman brains (Fig. 1a).

There are facts supporting that the cross-species could happen. Human speech and animal vocalizations share low-order spectrotemporal statistics (Singh & Theunissen, 2003), suggesting that speech-pretrained DNNs may generalize to the auditory cortex of nonhuman animals. Also, some nonhuman species possess the abilities of vocal learning and reference (Pepperberg, 2002; Oren et al., 2024). There are also reasons to expect this generalization to be difficult. Human speech and language depend on specialized areas, which do not have functionally equivalent networks discovered in animal brains (Friederici et al., 2017), opposing this cross-species generalization and leaving the question open: Does auditory neural predictivity require pretraining on domain-matched corpora, or can models trained on diverse auditory domains learn representations that remain predictive across species? Answering this question requires model–brain comparisons in animals that possess complex vocal behavior.

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.

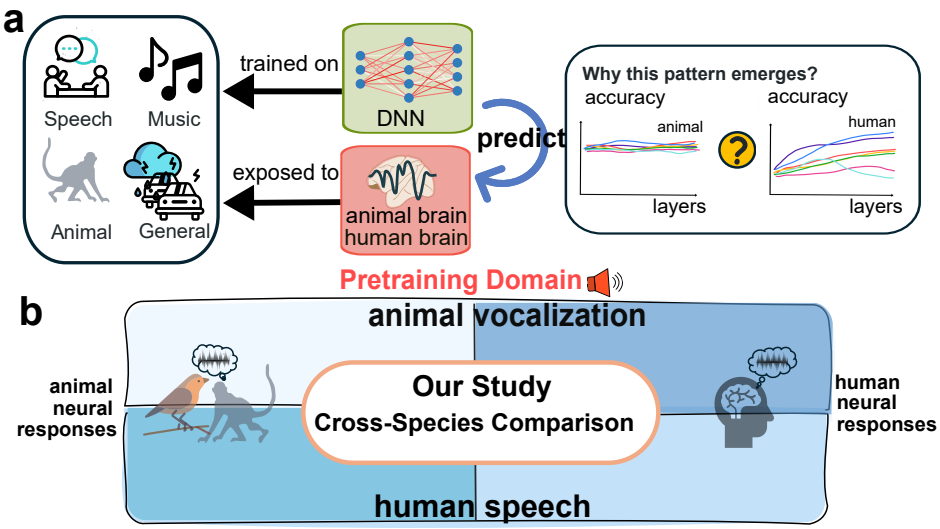

Figure 1: **(a)** We used deep neural networks (DNNs) pretrained on different auditory domains to predict neural responses in humans and non-human animals, and systematically compared prediction accuracy across model layers. **(b)** The study space was defined along two dimensions, the DNN pretraining domain and the source of neural responses, resulting in four paradigms. Our study systematically compared all four combinations of pretraining domains and neural response sources.

We addressed this question using two representative nonhuman species: zebra finches, an evolutionary distant vocal-learning songbird with rich vocal song types, and squirrel monkeys, a nonhuman primate with a complex species-specific call system. These species provide distinct evolutionary tests of vocal-sound processing without treating animal communication as equivalent to human language.

We compared nine self-supervised audio models and six from the same architecture spanning human speech, general sounds, and animal vocalizations, including matched-objective subsets that isolate the effects of pretraining corpus. Using a common neural-encoding framework, we evaluated these models on zebra finches responses to conspecific vocalizations, squirrel monkey responses to both conspecific calls and human speech, and human responses to speech and diverse natural sounds (Fig. 1b). This design tests whether neural predictivity broadly transfers across pretraining domains and species, while revealing task-specific advantages of matched pretraining.

Our main contributions are: First, across animal neural-prediction tasks, models pretrained on markedly different auditory domains achieved nearly identical best-layer encoding accuracy, whereas their predictivity diverged substantially on human data (Fig. 3). Second, unbiased Centered Kernel Alignment (uCKA) revealed clear pretraining-corpus-dependent geometry in model activations across both animal and human datasets, showing that the convergence in animal neural predictivity does not arise from convergence of the underlying model representations. Third, we introduce an input-sensitivity subspace analysis of the complete fitted neural predictor. This analysis shows that speech-corpus-specific structure present in model activations is selectively lost in animal neural-predictive subspaces while remaining preserved in human neural-predictive subspaces.

2 Related Works

Auditory DNNs as models of neural responses. Early studies found that DNNs optimized for auditory tasks can effectively predict auditory cortical responses (Kell et al., 2018).This predictive ability does not depend on a particular DNN architecture (Tuckute et al., 2023). Further studies showed that DNN-based neural predictivity cannot be fully explained by traditional acoustic representations (Keshishian et al., 2020; Giordano et al., 2023; Vaidya et al., 2022).

Effects of Factors on Neural Predictivity. The training paradigm can influence neural predictivity. self-supervised models often achieve stronger neural correspondence than supervised model, particularly at intermediate layers (Vaidya et al., 2022). Pretraining data can also substantially shape the auditory representations learned by DNNs and their correspondence with neural responses. Training on more naturalistic acoustic data has been associated with higher auditory cortical predictivity (Tuckute et al., 2023),while differences in the domain and linguistic composition of training data can systematically influence model–brain correspondence (Millet et al., 2022).

Cross-Species and Cross-Domain Neural Prediction. Prior studies have shown that neural predictivity can persist even when the model's pretraining domain does not match the neural stimulus domain (Millet et al., 2022), and that speech models can retain neural predictive ability when the language used for model training differs from that of the experimental stimuli (Li et al., 2023). Moreover, speech-trained models have been shown to predict auditory neural responses in non-human species (Ahmed et al., 2025). In parallel, work on bioacoustic transfer has demonstrated that representations learned from human speech can generalize across species and acoustic domains in downstream detection and classification tasks (Sarkar & Doss, 2025). However, existing studies have examined a relatively limited range of training paradigms and pretraining domains. It remains unclear whether the effects of different training paradigms on neural predictivity are consistent across species, and whether domain-matched models provide advantages for cross-species neural prediction.

3 METHODS

3.1 MODELS

HuBERT is a widely used BERT-inspired self-supervised speech model that has demonstrated strong performance in automatic speech recognition (Hsu et al., 2021), as well as keyword spotting and emotion recognition (Yang et al., 2021). Its broad adoption has also produced numerous variants pretrained on different corpora. To our knowledge, however, no publicly available HuBERT checkpoint has been pretrained exclusively on animal vocalizations.

We therefore selected `AVES-bio` as the closest available bioacoustic counterpart. It was pretrained on 153 hours of the core corpus and an additional 207 hours animal recordings. Because its pretraining corpus was not exclusively bioacoustic, we also included `AVES-core`, `AVES-nonbio`, and `AVES-all` to examine the effect of corpus composition. All four AVES models used the same learning rate, number of pretraining updates, and pretraining procedure (Miron et al., 2026). For speech controls sharing the HuBERT objective, we selected `chinese-hubert-base`and `hubert-base-ls960`. These models were pretrained on diverse Chinese speech from WenetSpeech and curated English read speech from LibriSpeech, respectively. We additionally included `MERT-v0-public` (Li et al., 2024) , a HuBERT-style model pretrained on music, as a music-domain control. Additionally, to test whether our conclusions generalize across self-supervised objectives, we included two wav2vec 2.0 models pretrained on the corresponding Chinese and English corpora (Baevski et al., 2020). All evaluated models used base-sized backbones comprising a seven-layer convolutional feature encoder followed by 12 Transformer blocks, thereby minimizing architectural variation while allowing the pretraining corpus and objective to vary.

To further examine whether neural predictivity is specific to self-supervised representation learning, we included six task-optimized speech recognition models as additional controls. These included three fully supervised models trained from scratch (wav2letter: Collobert et al. (2016), Speech2Text: Wang et al. (2020), and DeepSpeech2: Amodei et al. (2016)), two weakly supervised sequence-to-sequence models (Whisper-tiny and Whisper-base: Radford et al. (2023)), and one self-supervised model followed by supervised ASR fine-tuning (Wav2Vec2-FT: Baevski et al. (2020)). Together, these models provide complementary comparisons across different training paradigms.

3.2 DATASETS

To examine how pretraining domain influences neural predictivity across species and auditory stimuli, we evaluated the same models on three animal and two human neural datasets. Together, these datasets span three species and five evaluation conditions, covering responses to conspecific vocalizations, human speech, and general sounds (Fig. 2).

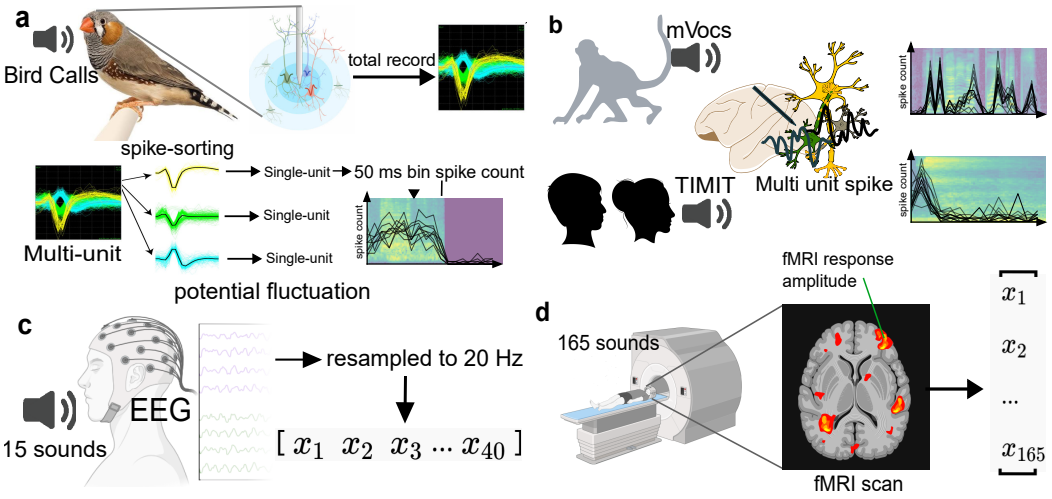

Figure 2: **Auditory datasets and neural response preparation.** **(a)** Squirrel monkey auditory cortical spiking responses to conspecific vocalizations (mVocs) and human speech (TIMIT). **(b)** Zebra finch AA-4 recordings during conspecific calls. **(c)** Human EEG responses to 15 natural sounds. **(d)** Human NH2015 fMRI responses to 165 natural sounds. These processed responses provide the targets for linear neural encoding models.

We first examined responses to conspecific vocalizations in zebra finches, a songbird with a well-characterized vocal repertoire. We used the CRCNS Avian Auditory 4 (AA-4) dataset, which contains extracellular auditory recordings from six adult zebra finches (Elie & Theunissen, 2015). Spikes were counted in non-overlapping 50-ms bins from stimulus onset to 400 ms after stimulus offset.

We next extended the analysis to squirrel monkeys, allowing us to examine a primate auditory system responding to both conspecific vocalizations and human speech.

We used auditory-cortex recordings released by Ahmed et al. (2025). Recording channels were transformed into multi-unit spike count of 50 ms bins, following the original study. The human EEG datasets complemented these animal conditions by testing whether models pretrained on bioacoustic recordings also predict human auditory responses to Speech and broader auditory content. We restrict our primary evaluation to Experiment 2 of Zuk et al. (2020) comprising speech, music and impact sounds. The original response epochs were resampled separately to 50 ms per sample point.

To extend the human evaluation beyond speech, we used the NH2015 natural-sound fMRI dataset (Norman-Haignere et al., 2015). For each sound and voxel, response estimates from three scanning sessions were averaged to obtain a single neural target.

3.3 Liner ridge fitting, scoring and unbiased CKA

To compare the neural information accessible from different pretrained representations, we fitted a separate L2-regularized linear readout for each candidate model layer and neural target (unit, electrode, or voxel) within each subject or recording session. For the time-resolved datasets, model activations were aligned to 50-ms neural-response bins, and features from causal lags were concatenated to predict each target bin. For NH2015 fMRI, model activations were averaged across time and fitted to each voxel. Within each dataset, all candidate models were evaluated using the same fitting and scoring procedure.

We compared model representations using unbiased linear uCKA. For each dataset and model, we extracted activations from the best layer selected by held-out neural prediction performance and concatenated them by time. We formed linear Gram matrices, applied unbiased U-centering, and normalized their cross-HSIC by the geometric mean of the two self-HSIC values.

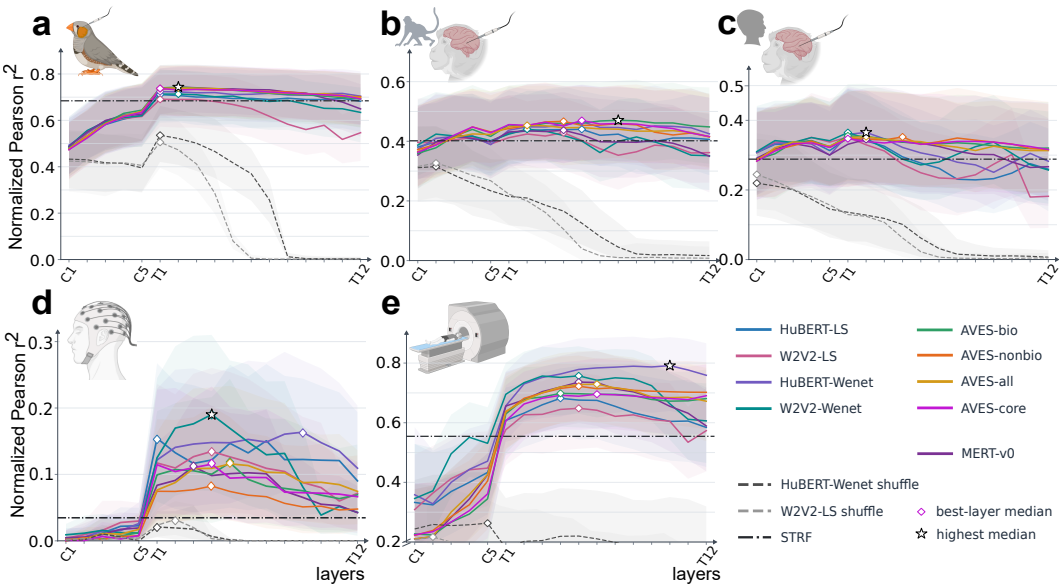

Figure 3: **Layerwise neural prediction across species and stimulus domains.** Panels show zebra-finch **(AA-4; a)**, squirrel-monkey **(mVocs; b)** and human speech **(TIMIT; c)**, human scalp EEG **(d)**, and human fMRI **(NH2015; e)**. Lines and bands show the median and inter-quartile range across neural targets after target-level reliability normalization.

3.4 Input subspace estimation

uCKA only specifies how stimuli are arranged relative to one another while leaving the input directions along which that geometry is realized undetermined. To anchor this relational geometry in concrete stimulus directions, we turn to the subspace defined by the prediction encoder itself by propagating gradients through the same fitted prediction encoding function.

Our analysis was inspired by Wingert et al. (2026), who showed that the gradients of a CNN's predicted neural responses with respect to the input spectrogram could be compressed into a low-dimensional tuning subspace, suggesting that gradients across stimuli provide a compact description of the stimulus temporal sensitive structure used by a nonlinear encoding function. Here we identified the input waveform perturbation subspaces to which each frozen encoding model was most sensitive. We applied this principle to the complete frozen encoding pipeline and summarized sensitivity jointly across neural outputs, sounds, and individuals. Specific steps are as follows.

Backpropagation. For model m, individual r, sound s, and response bin t, we computed the Jacobian of the predicted neural response with respect to a local waveform perturbation $\boldsymbol{\delta}$ spanning the same lag window used for neural encoding:

$$\boldsymbol{J}_{mrst} = \left.\frac{\partial \hat{\boldsymbol{y}}_{mrst}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right|_{\boldsymbol{\delta}=0}. \tag{1}$$

The matrix $\boldsymbol{J}_{mrst}^{\top}\boldsymbol{J}_{mrst}$ governs the leading-order mean squared change at response bin. Aggregating these matrices across held-out sounds and valid response bins yields directions of greatest average sensitivity, following the vector-valued formulation of active subspaces by Zahm et al. (2020). NH2015 fMRI has one fitted response per sound and voxel.

To avoid explicitly constructing the Jacobian, we used Rademacher output projections (Clarkson & Woodruff, 2009). For each valid response bin,

$$\boldsymbol{g}_{mrst} = \boldsymbol{J}_{mrst}^{\top}\boldsymbol{z}_{rst}, \qquad z_{rst,i} \in \{-1, +1\}, \tag{2}$$

where the same projection was shared across models for a fixed individual, sound, and response bin. We aggregated projected gradients within each individual, normalized their second moment to unit

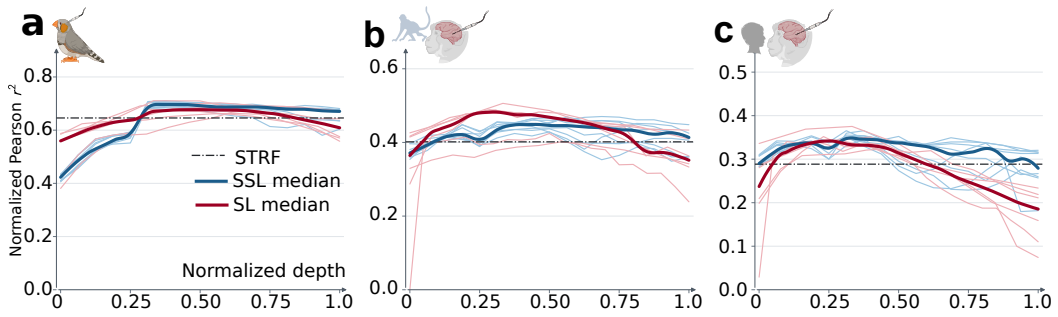

Figure 4: **Depth-dependent neural predictivity across training paradigms.** Layer-wise neural predictivity is shown across different neural datasets with layer positions normalized to a shared relative-depth scale. **a-c** is an comparison between 9 SSL and 6 SL models. Solid line represents the median among models.

trace, and then averaged equally across individuals:

$$\widehat{C}_m = \frac{1}{R}\sum_{r=1}^{R} \frac{\sum_s \sum_t \boldsymbol{g}_{mrst}\boldsymbol{g}_{mrst}^\top}{\operatorname{tr}\left(\sum_s \sum_t \boldsymbol{g}_{mrst}\boldsymbol{g}_{mrst}^\top\right)}. \tag{3}$$

The leading eigenvectors of $\widehat{C}_m$ define the model's input sensitivity subspace. We used rank 64 for the primary analysis and estimated the leading directions with randomized SVD (Halko et al., 2011).

Finally, we quantified the overlap between the rank-k subspaces of models a and b using the RMS cosine of their principal angles:

$$O_{ab}^{(k)} = \frac{1}{\sqrt{k}} \left\| \boldsymbol{U}_{a,k}^\top \boldsymbol{U}_{b,k} \right\|_F. \tag{4}$$

$O_{ab}^{(k)}$ ranges from 0 for orthogonal subspaces to 1 for identical subspaces. For NH2015, we applied the same analysis at the stimulus level, and additionally repeated it using only nonspeech stimuli.

4 Results

4.1 Model accuracy diverges primarily on human datasets

Table 2 shows that the contrast between animal and human neural prediction persists within HuBERT family. Across the animal datasets, best-layer scores differed by at most 7.09% of the highest-performing model. By contrast, the human datasets showed both larger improvements over the STRF baseline and substantially greater variation across models, with relative score ranges of 35.71% in EEG and 13.82% in fMRI. Despite the relatively small margins over the traditional STRF baseline in the animal datasets, all models significantly outperformed STRF except W2V2-LS on AA-4 ($q < 0.01$; one-sided paired Wilcoxon signed-rank tests; Fig. 3).

4.2 Pretraining-domain structure in model rankings

We next examined model-ranking patterns in datasets showing relatively large between-model divergence. **Effects of pretraining corpus.** In both human datasets, WenetSpeech-trained models outperformed their LibriSpeech-trained counterparts within the corresponding architectures. Speech-trained models occupied the leading ranks in EEG, with W2V2-Wenet achieving the highest score (0.190). **Similarity between AVES-bio and AVES-core.** Within the AVES family, AVES-bio and AVES-core achieved nearly identical best-layer scores in both human datasets, whereas AVES-bio scored above AVES-nonbio in EEG and below AVES-nonbio and AVES-all in fMRI (Fig. 3d,e).

4.3 CROSS-SPECIES CONSISTENCY OF TRAINING-PARADIGM EFFECTS

To test whether the effect of training paradigm on neural predictivity generalizes across species, we additionally evaluated six task-optimized speech-recognition models alongside the nine self-supervised learning (SSL) models introduced above (Fig. 4).

The two model groups achieved comparable performance at their best encoding layers.Across Bird AA-4, Monkey mVocs, and Monkey TIMIT, the differences in median peak predictivity between SSL and task-optimized models were +0.013, -0.031, and +0.005, respectively, indicating no consistent best-layer advantage for either training paradigm. At the final layers, however, SSL models showed higher median predictivity across all three datasets, with differences of +0.061, +0.062, and +0.097 for Bird AA-4, Monkey mVocs, and Monkey TIMIT, respectively. We also engaged a fine-tuned control of wav2vec 2.0 (Fig. 9), which yields the same conclusion. We therefore focused the subsequent representational analyses on the SSL models.

4.4 PRETRAINING-CORPUS-RELATED GEOMETRY IN MODEL ACTIVATIONS

We next asked whether the corpus-dependent effects observed in neural predictivity were already present in the geometry of the model activations. **AVES models form a consistent representational cluster.** Across all datasets, the AVES models formed a robust cluster. Mean within-AVES uCKA ranged from 0.816 to 0.943 and exceeded mean Speech–AVES similarity by 0.059–0.213. Within the AVES family, AVES-core and AVES-all were comparatively less similar (mean uCKA = 0.865, versus 0.908 across the remaining AVES pairs)(Fig. 5).

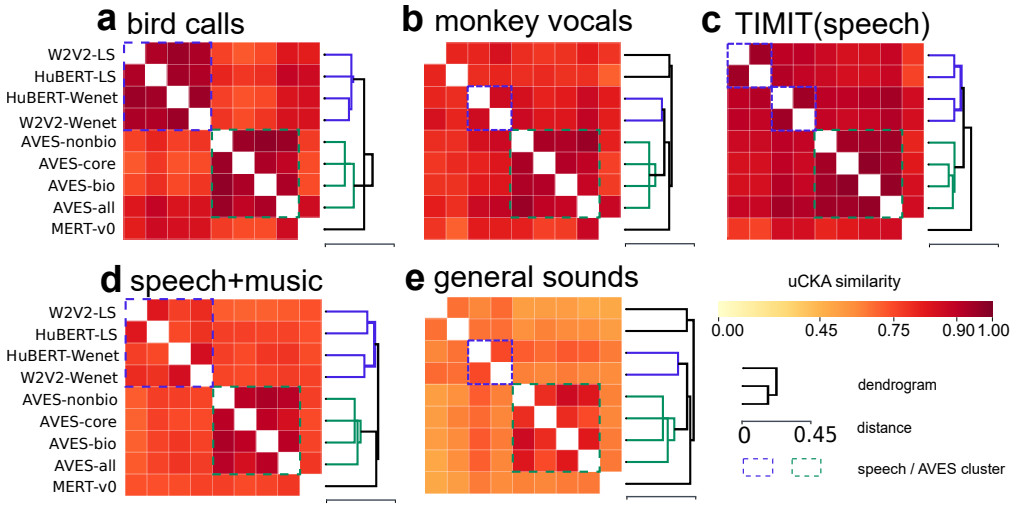

Figure 5: **Unbiased CKA (uCKA) of model features.** Each complete 9 × 9 matrix reports the raw pairwise signed uCKA values for the same nine pretrained models. **We use only activations of the sound stimuli, while the brain responses are not engaged.** Frames remark clusters formed by speech models and AVES models.

Speech models retain pretraining-corpus structure. In AA-4 and human EEG, the four speech models formed a clear cluster. Their mean pairwise distances ($1 - \text{uCKA}$) were 0.070 and 0.199, respectively, compared with 0.201 and 0.256 between speech models and the remaining pretrained models. In mVocs, the wav2vec 2.0 and HuBERT models pretrained on WenetSpeech formed a compact pair, with a pairwise distance of 0.129 compared with a mean cross-corpus speech-model distance of 0.178. Monkey TIMIT showed approximately symmetric corpus-specific pairing: both the LibriSpeech and WenetSpeech pairs had a distance of 0.062, compared with 0.099 across corpora. The fMRI result was more asymmetric, with the WenetSpeech pair showing substantially lower distance than the cross-corpus speech-model pairs (0.252 versus 0.356). MERT showed no stable family affiliation across datasets. In AA-4, it was slightly more similar to speech models than

to AVES models (mean uCKA = 0.853 versus 0.779), whereas in the remaining datasets it did not consistently cluster with either family.

Thus, pretraining domain leaves a detectable structure in model activations across both animal and human datasets, despite the much weaker domain dependence observed in animal neural predict accuracy. In summary, in uCKA, we find a relatively high similarity of activation representations within the AVES model family, possibly due to shared core corpus. Clusters within same pretraining corpora shows that models of different pretraining corpora already diverge at the level of representation.

4.5 Speech-specific subspace emerges only in human datasets

This raises the question of whether such geometry survives when the activations are projected onto the neural-predictive subspace. **Speech-model organization.** At rank 32, 64, or 128, the four speech-pretrained models formed a complete dendrogram clade in human EEG and in both fMRI conditions, but not in any animal dataset (Fig. 6 7 8).

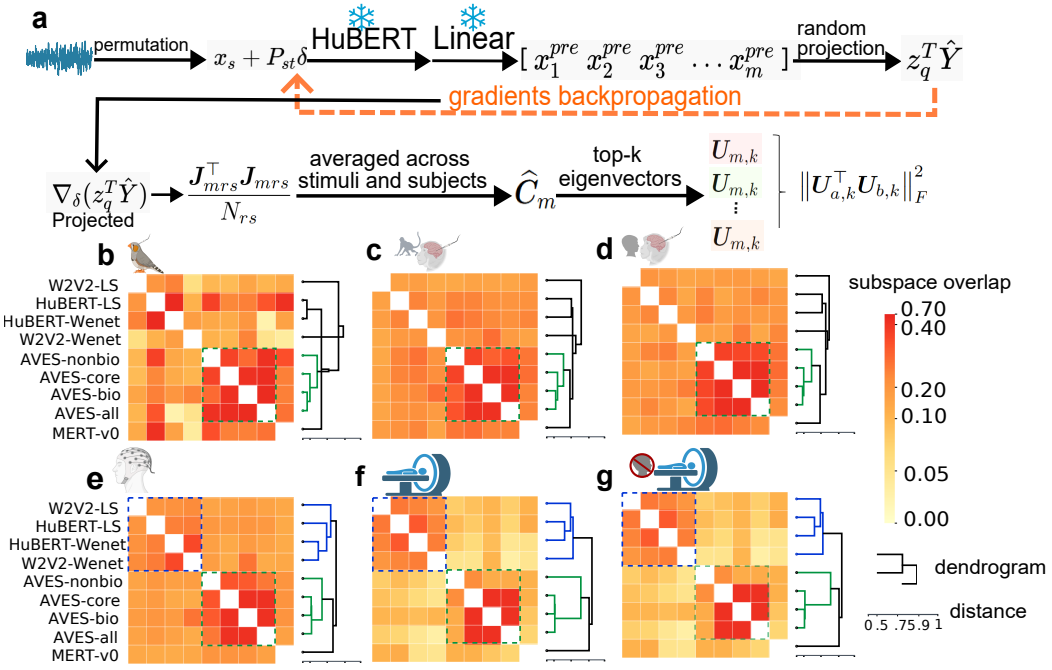

Figure 6: **Backpropagation steps and subspace overlaps. (a)** Overview of the response-aligned gradient analysis. Gradients are aggregated across valid bins, stimuli, and individuals to estimate the most sensitive lag-domain directions. The two fMRI conditions retain the stimulus-level coordinate because their fitted responses have no time-bin axis. **(b–g)** Overlap matrices and the corresponding dendrograms for each dataset. Matrix values are rank-64 RMS principal cosines, and pairwise dendrogram distances are defined as $1 - \text{overlap}$.

Moreover, neither the LibriSpeech-matched pair nor the WenetSpeech-matched pair formed a two-model dendrogram clade in AA-4, mVocs, or monkey TIMIT at ranks 32, 64, or 128, indicating that corpus-specific organization was not preserved in the animal neural-predictive subspaces (Appendix I).

The effect follows the listener rather than the stimulus. The preservation of speech-model structure was better explained by the species providing the neural responses than by the presence of speech in the stimulus set. Monkey TIMIT and human EEG both contained human-speech stimuli, yet their neural-predictive subspace geometries differed markedly: in monkey TIMIT, mean within-speech overlap was slightly lower than speech-to-other-model overlap (0.187 versus 0.193), and the four speech models did not form a clade, whereas in human EEG, within-speech overlap was substantially higher (0.252 versus 0.175), with all four speech models forming a complete clade.

Conversely, within human fMRI, removing all speech stimuli from gradient aggregation left the complete speech-model clade unchanged. Together, these results indicate that speech-specific subspace organization tracks the neural response source rather than the acoustic content of the stimuli.

4.6 AVES-NONBIO CLUSTER WEAKER IN AVES FAMILY

We next examined the internal structure of the AVES family. AVES-core, AVES-bio, and AVES-all constituted the most consistent recurring clade, forming an exact dendrogram clade in every dataset and fMRI condition. Their mean within-trio overlap was 0.528, 0.440, 0.512, 0.470, and 0.483 in AA-4, mVocs, monkey TIMIT, EEG, and full fMRI, respectively. By comparison, the mean overlap of AVES-nonbio with this trio was only 0.371, 0.338, 0.363, 0.307, and 0.161. The same organization persisted in the non-speech fMRI analysis, with a within-trio mean of 0.479, an AVES-nonbio-to-trio mean of 0.162. Despite this weaker alignment, AVES-nonbio still joined the three models within a higher-level four-model AVES clade in every condition.

Within the AVES family, we found that bio corpora counts for sensitive subspaces on all datasets. We found that subspaces of AVES-nonbio, with the lowest proportion of animal vocalizations, is relatively distant from the rest AVES models across all datasets, which may indicate that the bio component is important for shaping the sensitive subspaces on both human and nonhuman neural representations, in contrast to speech component.

5 DISCUSSION

Although achieve comparable peak neural predictivity, models trained with different paradigms diverge on layerwise accuracy. While task-optimized models tend to lose predictive power in intermediate and deeper layers, SSL models maintain or even enhance their neural predictivity with increasing depth. our work extends previous studies (Li et al., 2023; Vaidya et al., 2022) to cross species neural prediction. We also found that models pretrained on real-World datasets performs better. Across all datasets, speech models trained on WenetSpeech consistently outperform those trained on LibriSpeech with the corresponding objective (Fig. 3), which verifies the conclusion of Chen et al. (2026) across species. Within the AVES model family ranking is not stable (Fig. 3), yet AVES-all consistently remains the best, suggesting that a broader span of pretraining corpora may be more favorable for brain-response prediction. Meanwhile, AVES-bio and AVES-core perform comparably, which demonstrates that animal vocalizations do not confer much additional advantage for human neural response prediction.

Models generalized well to the brain-response prediction task across all neural datasets, even when the pretraining did not fully match the stimulus domain. However, a cost of linguistic abstraction in animal datasets was observed: accuracy of AVES models in deeper layers was generally higher than that of speech models. We attribute this to the fact that the deep representations of speech models encode more language-related information than AVES models, whereas animal brain responses do not contain. On the TIMIT dataset, which consists exclusively of speech stimuli, the relative decline of speech models in deeper layers is even more obvious, corroborating our proposed explanation.

We find that subspace overlaps differ considerably between the animal and human datasets, which corresponds precisely to the difference in accuracy between animals and humans in Fig. 3. In the human data, speech models all form clear clusters which indicates that language pretraining exerts a consistent influence on the sensitive directions of the corresponding human-brain prediction encoders. In animals, however, the sensitive subspaces of these models fail to form a cluster within WenetSpeech, LibriSpeech or all four speech models. The subspace finding partly explain results from prediction accuracy. That is, in animal-brain prediction encoders, language pretraining or certain speech corpora does not shape the sensitive subspace along a unified direction. This corresponds precisely to the absence of speech in animals.

We consider two pretraining objectives and a single model parameter scale. A comparison with models pretrained exclusively on animal vocalizations, and more specifically, on the vocalizations of a single species, is absent from our analysis. Owing to the limited availability of open-source datasets, we discuss only three animal datasets, from two nonhuman species, and two human datasets.

6 AI USE STATEMENT

Generative AI tools were used only for LaTeX formatting, language polishing and code reviewing. All AI-assisted content was carefully verified by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

7 ETHICS STATEMENT

This study did not recruit new participants or collect new human or animal data. We conducted secondary analyses of existing, publicly available datasets. The zebra finch data were originally collected under procedures approved by the Animal Care and Use Committee of the University of California, Berkeley. The squirrel monkey data were collected under protocols approved by the Institutional Animal Care and Use Committee of the University of California, San Francisco . The EEG dataset was obtained from Zuk et al. (2020), whose study was conducted at Trinity College Dublin; the original publication did not specify the approving ethics committee. The original fMRI study was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), and all participants provided informed consent. The present study involved only secondary analyses of these existing datasets.

8 REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the computational models, datasets, feature extraction procedures, representational analyses, cross-validation schemes, statistical inference procedures, and evaluation metrics in the Methods and Appendix. Model identities and pretrained checkpoints are specified in the Appendix. The datasets used in this study are publicly available from the sources cited in the paper. Code and additional materials required to reproduce the reported analyses and figures are provided as supplementary materials.

REFERENCES

Bilal Ahmed, Joshua D Downer, Brian J Malone, and Joseph G Makin. Deep neural networks explain spiking activity in auditory cortex. *PLOS Computational Biology*, 21(8):e1013334, 2025.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Runnan Cao, Jie Zhang, Jie Zheng, Yue Wang, Peter Brunner, Jon T Willie, and Shuo Wang. Computational single-neuron mechanisms of visual object coding in the human temporal lobe. *Nature communications*, 17(1):2234, 2026.

Chenggang Chen, Zhiyu Yang, and Xiaoqin Wang. Real-world unsupervised models generalize to predict brain responses to out-of-distribution stimuli. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=H1HHss5Zj4.

Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 205–214, 2009.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

Julie E Elie and Frédéric E Theunissen. Meaning in the avian auditory cortex: neural representation of communication calls. *European Journal of Neuroscience*, 41(5):546–567, 2015.

Julie E Elie and Frédéric E Theunissen. Invariant neural responses for sensory categories revealed by the time-varying information for communication calls. *PLoS computational biology*, 15(9): e1006698, 2019.

Angela D Friederici, Noam Chomsky, Robert C Berwick, Andrea Moro, and Johan J Bolhuis. Language, mind and brain. *Nature human behaviour*, 1(10):713–722, 2017.

Bruno L Giordano, Michele Esposito, Giancarlo Valente, and Elia Formisano. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4):664–672, 2023.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of neuroscience*, 35(27):10005–10014, 2015.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Menoua Keshishian, Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *Elife*, 9:e53445, 2020.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, volume 2024, pp. 12181–12204, 2024.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H Carney, Junfeng Lu, Jinsong Wu, and Edward F Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35: 33428–33443, 2022.

Marius Miron, David Robinson, Milad Alizadeh, Ellen Gilsenan-McMahon, Gagan Narula, Emmanuel Chemla, Maddie Cusimano, Felix Effenberger, Masato Hagiwara, Benjamin Hoffman, et al. Avex: What matters for animal vocalization encoding. In *International Conference on Learning Representations*, volume 2026, pp. 113288–113313, 2026.

Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel Yamins. Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, 31, 2018.

Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, 88(6):1281–1296, 2015.

Guy Oren, Aner Shapira, Reuven Lifshitz, Ehud Vinepinsky, Roni Cohen, Tomer Fried, Guy P Hadad, and David Omer. Vocal labeling of others by nonhuman primates. *Science*, 385(6712): 996–1003, 2024.

Irene M Pepperberg. In search of king solomon's ring: cognitive and communicative studies of grey parrots (psittacus erithacus). *Brain, Behavior and Evolution*, 59(1-2):54–67, 2002.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Giordano Ramos-Traslosheros and Carlos Ponce. Feature segregation by signed weights in artificial vision systems and biological models. In *International Conference on Learning Representations*, volume 2026, pp. 41740–41766, 2026.

N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

Eklavya Sarkar and Mathew Magimai Doss. Comparing self-supervised learning models pre-trained on human speech and animal vocalizations for bioacoustics processing. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Nandini C Singh and Frédéric E Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394–3411, 2003.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39, 2020.

Jereme C Wingert, Satyabrata Parida, Sam V Norman-Haignere, and Stephen V David. Convolutional neural network models describe the encoding subspace of local circuits in auditory cortex. *Nature Neuroscience*, 29(4):876–887, 2026.

Will Xiao, Saloni Sharma, Gabriel Kreiman, and Margaret S Livingstone. Feature-selective responses in macaque visual cortex follow eye movements during natural vision. *Nature neuroscience*, 27(6):1157–1166, 2024.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

Olivier Zahm, Paul G Constantine, Clémentine Prieur, and Youssef M Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

Nathaniel J Zuk, Emily S Teoh, and Edmund C Lalor. Eeg-based classification of natural sounds reveals specialized responses to speech and music. *NeuroImage*, 210:116558, 2020.

A MODEL SOURCES

- AVES-bio, AVES-core, AVES-nonbio, and AVES-all: AVES pretrained models.
- chinese-hubert-base: Hugging Face repository.
- hubert-base-ls960: Hugging Face repository.
- MERT-v0-public: Hugging Face repository.
- chinese-wav2vec2-base: Hugging Face repository.
- wav2vec2-base-960h (SSL and fine-tuned): fairseq pretrained models.
- Speech2Text: Hugging Face repository.
- Whisper-base: Hugging Face repository.
- DeepSpeech: pretrained checkpoint.
- Whisper-tiny: Hugging Face repository.
- wav2letter: pretrained weights.

B DATA SOURCES

- Zebra finches: CRCNS AA-4 dataset.
- Squirrel monkeys: Zenodo repository.
- Human EEG: OSF repository.
- Human fMRI (NH2015): GitHub repository.

C MODEL DETAILS

Table 1: Pretrained models evaluated in this study. AS: AudioSet; VS: VGGSound. Model sources are listed in Appendix A.

Model	**SSL objective**	**Pretraining corpus**
AVES-bio	masked reconstruction	Core + AS/VS (animal), 360 h
AVES-core	masked reconstruction	FSD50K + AS core, 153 h
AVES-nonbio	masked reconstruction	Core + AS/VS (non-animal), 360 h
AVES-all	masked reconstruction	Core + AS/VS (all), 5,054 h
chinese-hubert-base	masked reconstruction	WenetSpeech (Chinese)
hubert-base-ls960	masked reconstruction	LibriSpeech (English)
MERT-v0-public	masked reconstruction	Music4All + FMA subset, 900 h
chinese-wav2vec-base	contrastive prediction	WenetSpeech (Chinese)
wav2vec2-base-960h	contrastive prediction	LibriSpeech (English)

D SUPPLEMENTARY FIGURES

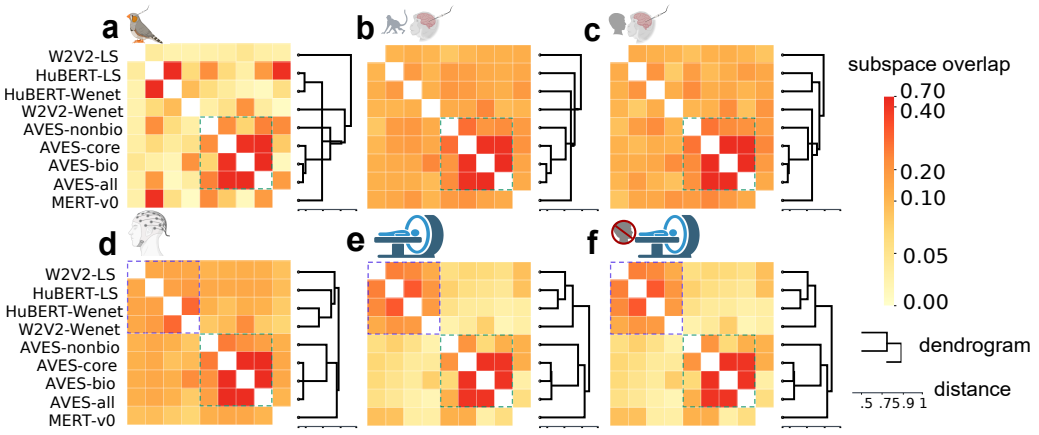

Figure 7: subspace overlaps on rank 32.

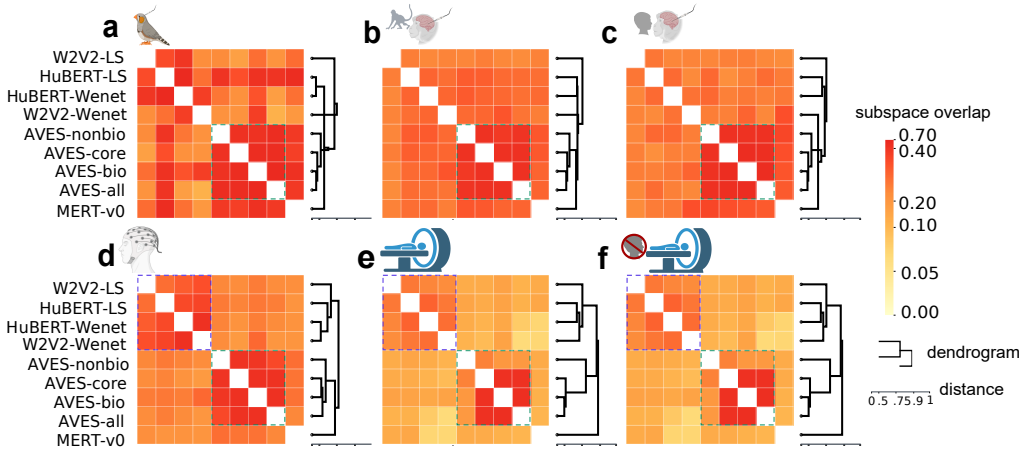

Figure 8: subspace overlaps on rank 128.

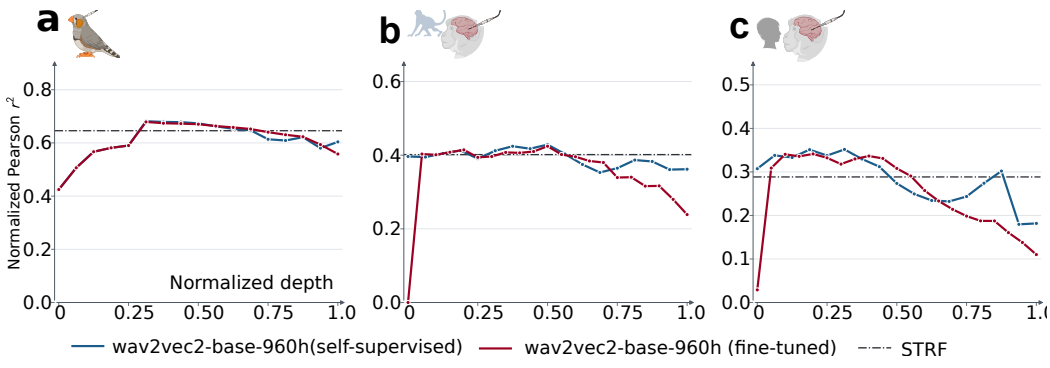

Figure 9: Comparison of W2V2-Base-960 trained under two different paradigms

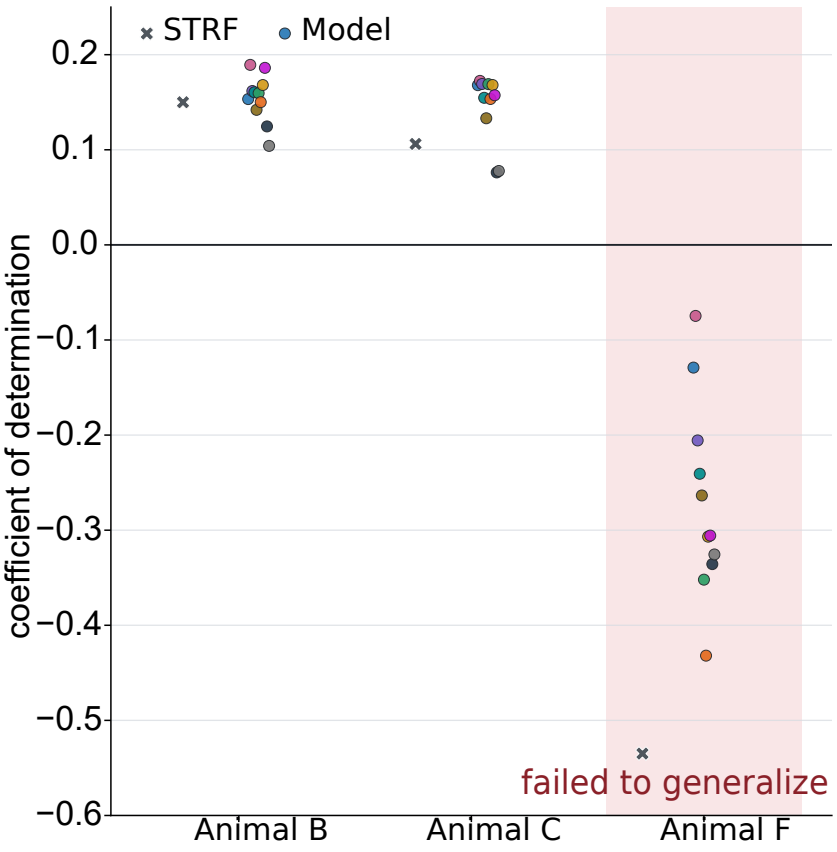

Figure 10: model-wise coefficient of mVocs. STRF and pretrained-model squirrel-monkey mVocs. Crosses show STRF; circles show model-only encoders. Whereas every displayed B/C value was positive, animal F yielded a negative coefficient of determination for both STRF (-0.535) and models (range -0.432 to -0.075).

E DATASETS

CRCNS Avian Auditory 4 (AA-4) dataset contains extracellular auditory recordings from six adult zebra finches(Elie & Theunissen, 2015). The vocalization stimuli cover nine behaviorally meaningful categories, with multiple callers and renditions represented within categories, providing a broad sample of the species' communication repertoire Elie & Theunissen (2019). Most stimuli were presented approximately 10 times. We retained spike-sorted auditory units with a Spearman–Brown-corrected split-half reliability of at least 0.5, yielding 644 units across all six birds. Spikes were counted in non-overlapping 50-ms bins from stimulus onset to 400 ms after stimulus offset, and responses were averaged across repetitions.

The TIMIT condition comprised 499 American English sentences from speakers spanning multiple dialect regions: 489 sentences were presented once, and 10 were presented 11 times each. The conspecific-vocalization condition (mVocs) comprised 303 squirrel-monkey calls: 292 were presented once, and 11 were presented 15 times each. These two conditions enabled us to compare model predictions for human speech and animal vocalizations within the same nonhuman species.

Following the authors' released reliability procedure, we retained channels with positive repeat-response reliability that exceeded the mean of a circular-shift null distribution by at least 0.5 null standard deviations. The main analysis included 404 multi-units for TIMIT and 438 for mVocs. In both conditions, stimuli presented once are used to fit the linear readout and select its regularization hyperparameter; repeated stimuli formed a fixed held-out test set. For each test repeat, predictions were correlated with the corresponding neural response time series across test stimuli. The resulting Pearson correlations were averaged across repeats and subsequently corrected for the noise ceiling using repeat-response reliability. Subject F consistently yielded negative coefficient of determination for both STRF and models in mVocs. Therefore, f was excluded from all mVocs analyses in our work (Fig. 10).

The original human EEG measured non-invasive scalp EEG during presentation of thirty two-second sounds: fifteen natural sounds, comprising five speech, five music, and five impact sounds, together with fifteen acoustically synthetic counterparts. For ridge fitting and inner-loop model selection, EEG responses were averaged across all available valid non-target presentations of each sound. For held-out evaluation, we used only natural sound and engaged 74 valid trial responses per sound. Thirteen participants have completed the test and contribute 128 channels each.

NH2015 fMRI dataset (Norman-Haignere et al., 2015) has 165 recognizable natural sounds span speech in multiple languages, instrumental music and song, human and animal vocalizations, and a wide range of nonvocal biological, mechanical, and environmental sounds. This stimulus diversity allowed us to assess model predictivity across a broader range of auditory content. We used the released version analyzed by Tuckute et al. (2023), comprising eight participants and 7,694 reliability-screened voxels. For each sound and voxel, response estimates from three scanning sessions were averaged to obtain a single neural target for encoding-model fitting and held-out evaluation.

F RIDGE FITTING DETAILS

For the time-resolved datasets, model activations were aligned to 50-ms bins and expanded over causal lags of 0–400 ms for bird AA-4, 0–200 ms for monkeys and human EEG. This expansion allowed each neural response bin to depend on both current and preceding model activations, accommodating response delays and temporal integration. For NH2015, predictions were correlated with three-scan-averaged voxel responses across the test sounds. Following Tuckute et al. (2023), we corrected for both measured-response reliability and prediction reliability, accounting for measurement noise and its propagation into the fitted readout. Measured-response reliability was estimated from correlations between scans; prediction reliability was estimated from predictions generated by separately fitted per-scan readouts. Each estimate used the median pairwise correlation with Spearman–Brown correction for three scans. We report corrected Pearson r^2 and summarize each voxel by its median score across the ten splits.

For AA-4, mVocs, and TIMIT, temporal Pearson correlations of predicted and collected responses were computed within repeats and then averaged across repeats. To account for trial-to-trial variabil-

Table 2: Best-layer neural prediction and across-model variability within the HuBERT family. The seven models comprise HuBERT-LS, HuBERT-Wenet, four AVES variants, and MERT-v0. Scores are median reliability-normalized squared Pearson correlations across neural targets. Range/best denotes the difference between the highest and lowest model medians divided by the highest median. Population SD/best uses the same normalization, with the standard deviation computed using denominator $N = 7$. Both dispersion measures are reported as percentages.

Dataset	Best-layer median range	STRF median	Range/best (%)	Population SD/best (%)
AA-4	0.711–0.744	0.684	4.39	1.49
Monkey mVocs	0.437–0.470	0.402	7.09	2.80
Monkey TIMIT	0.345–0.365	0.288	5.54	1.68
Human EEG	0.105–0.164	0.035	35.71	12.95
NH2015 fMRI	0.682–0.791	0.554	13.82	4.25

ity not predictable from the stimulus, we applied a reliability normalization to all correlation values using the random-pair response-reliability procedure by Ahmed et al. (2025).

AA-4 and Human EEG. To evaluate generalization to unseen sounds, we used five outer whole-stimulus folds and four inner folds, keeping all observations of a stimulus in the same folds. Within each outer-training set, regularization was selected independently for each neural target by maximizing the coefficient of determination (R^2) over inner out-of-fold predictions. The readout was then refitted on the complete outer-training set and evaluated on the held-out stimuli, separating regularization selection from final evaluation.

mVocs and TIMIT. Following the training–test design of Ahmed et al. (2025), readouts were fitted using non-repeated mapping stimuli, while repeated held-out stimuli were reserved for evaluation and reliability estimation. Regularization was selected by three-fold cross-validation within the mapping set. The final readout was then refitted on all mapping stimuli, without using the repeated test stimuli for fitting or regularization selection.

NH2015. To match the stimulus-level fMRI responses, activations were averaged over time and retained at their native dimensionality. We followed Tuckute et al. (2023), using the same ten overlapping 83-training/82-test splits. Voxel-specific regularization was selected by efficient leave-one-out cross-validation within the 83 training sounds, after which the readout was refitted on all training sounds and applied to the 82 test sounds.

G PCA USED IN AA-4 AND HUMAN EEG FOR MODEL ACTIVATIONS

Considering the relevant insufficient neural data for fitting the high-dimensional activations and the risk of overfitting, each DNN layer was reduced to 80 principal components before lag expansion in AA-4 and human EEG. The principal component analysis was strictly fold-local: the 80 basis functions were fitted only to the current inner-training stimuli, and for the final outer-fold model, it was refitted using the complete outer-training set.

H ENCODING ACCURACY SUMMARY

I RANK SENSITIVITY

The qualitative speech-clade result was rank-robust. From rank 32 to 64 to 128, median pairwise overlap increased from 0.060 to 0.219 to 0.361 in AA-4, from 0.142 to 0.227 to 0.297 in mVocs, from 0.141 to 0.202 to 0.270 in monkey TIMIT, from 0.139 to 0.193 to 0.261 in human EEG, and from 0.069 to 0.103 to 0.141 in human fMRI. No animal dataset produced a complete four-model speech clade, or a corpus-matched LibriSpeech or WenetSpeech pair, whereas human EEG and human fMRI retained the complete speech-model clade at all three ranks.