Hao et al. *Zool. Res. Divers. Conserv.* 2025, 2(3): 196−206
https://doi.org/10.24272/j.issn.2097-3772.2025.015
https://cstr.cn/32249.14.zrdc.2025.015

ZRDC | Zoological Research: Diversity and Conservation

Article Open Access

Frequency reflects region in birdsong recognition: Quantified by mutual information and mitigated through adaptive normalization

Zhu-Lin Hao[1], Meng-Kun Zhu[1], Jiang-Jian Xie[1,2,3,*], Chang-Qing Ding[4,*]

[1] *School of Technology*, *Beijing Forestry University*, *Beijing*, 100083, *China*
[2] *State Key Laboratory of Efficient Production of Forest Resources*, *Beijing Forestry University*, *Beijing*, 100083, *China*
[3] *Research Center for Biodiversity Intelligent Monitoring*, *Beijing Forestry University*, *Beijing*, 100083, *China*
[4] *School of Ecology and Nature Conservation*, *Beijing Forestry University*, *Beijing*, 100083, *China*

ABSTRACT

Geographic and environmental heterogeneity generates pronounced variability in bird vocal dialects, complicating reliable species identification across spatially distinct populations. To quantify such dialectal divergence, maximum mean discrepancy (MMD) analysis was applied to multiple regional birdsong datasets, revealing significant distributional differences between recording locations. To overcome these challenges, an adaptive normalization and recognition framework was developed that integrates task-driven feature normalization with a multi-head attention ResNet (MHAResNet) classifier. The normalization module dynamically reweights frequency, time, and channel dimensions according to their contextual importance, enhancing feature alignment across domains. The classifier concurrently distinguishes both species identity and regional provenance. This coupled architecture suppresses domain-induced variability while preserving salient acoustic cues critical for recognition. To dissect the contribution of individual feature dimensions, mutual information neural estimation (MINE) was employed to quantify their relevance to species and region classification. Across three geographically diverse birdsong datasets, the proposed method improved species recognition by an average of 2.9% and region recognition by 3.0% relative to non-normalized baselines. MINE results indicated that frequency features were the most predictive of geographic origin, whereas channel-based features most strongly encoded species discrimination. These results advance understanding of birdsong feature attribution and offer a scalable framework for acoustic biodiversity assessment across biogeographic gradients.

Keywords: Birdsong recognition; Dialectal variation; Mutual information neural estimation; Maximum mean discrepancy analysis; Acoustic feature normalization

INTRODUCTION

Avian species contribute fundamentally to global biodiversity and ecosystem stability through ecological functions spanning pollination, seed dispersal, and trophic regulation (Lu et al., 2023). Vocalizations constitute the primary medium for avian communication, mediating reproductive signaling, territorial defense, and social coordination, and encapsulate ecologically and taxonomically informative acoustic signatures. The deployment of passive acoustic monitoring (PAM) has transformed avian field monitoring by enabling long-term, non-invasive ecological surveillance at spatiotemporal scales unattainable through traditional observation methods (Sedláček et al., 2015). PAM reduces logistical burden while expanding coverage and temporal resolution, presenting a scalable and minimally disruptive tool for biodiversity monitoring (Ma, 2016). However, the sheer volume of audio data generated necessitates high-throughput, automated analytical solutions to conserve resources and improve efficiency (Kasten et al., 2012). In this context, automatic bird vocalization recognition (BVR) has emerged as a key solution to the challenge of processing large-scale ecological audio data (Xie et al., 2025a). Recent advances in deep learning-based approaches have increasingly outperformed traditional rule-based and handcrafted feature methods in avian

Received: 29 September 2025; Accepted: 26 November 2025; Online: 23 December 2025

Foundation items: This work was supported by the National Natural Science Foundation of China (62303063, 32270554), National Key Research and Development Program of China (2024YFF1307202), Beijing Natural Science Foundation (5252014), and Beijing Forestry University Science and Technology Innovation Program (2024XY-G002).

*Corresponding authors, E-mail: shyneforce@bjfu.edu.cn; cqding@bjfu.edu.cn

classification accuracy (Xie et al., 2023b). Nevertheless, a persistent obstacle in real-world ecological applications is cross-domain distributional shifts. Birdsong recordings vary substantially across geographic regions due to dialectal divergence, habitat-specific environmental conditions, and heterogeneity in recording hardware (Xie et al., 2023a, 2025b), causing domain mismatches that reduce recognition performance (Stowell et al., 2019).

Empirical evidence increasingly demonstrates that geographic variability in avian vocalizations significantly impairs recognition accuracy in cross-regional scenarios. Kahl et al. (2021) evaluated convolutional neural network architectures, including BirdNET, across multiple ecological regions and reported marked declines in classification performance when applied to novel habitats, even when leveraging pre-trained feature extractors. Similarly, Lauha et al. (2022) trained models on recordings from Southern Finland and observed strong within-region accuracy but substantial performance deterioration when the same models were deployed in other geographic areas, highlighting the impact of regional dialectal variation on cross-location recognition. In response to these domain discrepancies, a range of adaptation and generalization frameworks have been proposed. Drossos et al. (2019) proposed an unsupervised adversarial strategy that uses Wasserstein distance to align source and target feature distributions, reducing regional bias by emphasizing invariant acoustic characteristics. Xiao et al. (2024) applied MixStyle to the frequency domain of Mel-spectrograms, enabling domain-generalized sound event detection by reducing regional bias across recording conditions. Self-supervised learning paradigms, such as audio-visual alignment (Owens & Efros, 2018), have also yielded domain-agnostic embeddings that enhance robustness under previously unseen environmental settings.

To further address cross-domain variability, adaptive normalization techniques have been proposed to reweight spectral features dynamically. Kim et al. (2021) proposed Residual Normalization (ResNorm), a frequency-wise instance normalization method incorporating residual connections to reduce device-specific artifacts while preserving discriminative features, thereby improving domain generalization in acoustic scene classification. Tang et al. (2022) combined instance normalization with a hybrid augmentation pipeline to suppress device-induced variability in avian audio detection tasks. Kim et al. (2022) adopted a relaxed instance frequency-wise normalization framework that dynamically normalizes spectral features along the frequency axis, improving generalization across multiple recording devices. Xie et al. (2023a) further developed instance frequency normalization to mitigate spectral distortions driven by rapid frequency modulations, enhancing robustness across geographically heterogeneous birdsong datasets. These approaches reduce domain-specific variability and promote domain-robust representations by emphasizing salient frequency components while minimizing spurious variability. However, critical limitations persist. Most frequency-wise normalization schemes apply uniform normalization strength across spectral bins, ignoring differential relevance of acoustic dimensions to specific recognition tasks such as species identification versus regional attribution. Furthermore, conventional instance-based normalizers operate independently across channel, time, and frequency dimensions, lacking coordinated attention to interdependent cues. This decoupled process risks either discarding informative signals or retaining noise. Finally, existing methods typically offer limited interpretability, neither pinpointing which dimensions are affected by domain shifts nor quantifying their contributions to generalization.

This study systematically investigated geographic heterogeneity in avian vocalizations by applying maximum mean discrepancy (MMD) analysis (Gretton et al., 2012) across multiple birdsong datasets, revealing statistically significant distributional divergence among recording regions. Based on this observation, a unified recognition framework was developed that combines adaptive feature normalization with a Multi-Head Attention ResNet (MHAResNet) architecture to mitigate performance degradation caused by geographic variability. In contrast to conventional fixed-frequency or independently applied instance normalization strategies, the normalization module in this framework dynamically adjusts weights across channel, time, and frequency dimensions based on task-specific relevance, effectively suppressing region-specific shifts while retaining biologically meaningful acoustic cues. The attention-enhanced ResNet architecture captures both local and global acoustic patterns, enabling joint recognition of species identity and geographic origin.

To enhance model interpretability, mutual information neural estimation (MINE) (Belghazi et al., 2018) was employed to quantify the mutual information (MI) shared between individual feature dimensions and classification targets, offering insights into the contribution of each dimension to species and region discriminations. Experimental validation on three geographically diverse birdsong datasets demonstrated consistent improvements in recognition accuracy under cross-region evaluation scenarios. By simultaneously enhancing robustness and interpretability, the proposed framework offers a scalable solution for large-scale biodiversity monitoring and ecological assessment.

Quantifying geographic variation in birdsong provides critical ecological insight into population-specific acoustic adaptations, facilitating improved generalization of recognition models and informing conservation efforts. Regionally distinct vocal traits, as revealed through this framework, may serve as acoustic indicators of environmental pressures or incipient evolutionary divergence, supporting biodiversity surveillance and habitat-specific management strategies.

MATERIALS AND METHODS

Datasets

Three geographically distributed birdsong datasets (D3BV, S1S2, and R1R2R3) were used to examine acoustic variation across distinct ecological regions. These datasets capture a range of environmental and dialectal differences, enabling cross-regional species recognition. A unified preprocessing pipeline was applied across all recordings: audio clips were downsampled to 16 kHz, converted to mono, and segmented into 8 s clips with 50% overlap. Files were converted to WAV format when required. To illustrate regional acoustic diversity, representative spectrograms from each dataset were constructed (Supplementary Figures S1–S3) (Jing et al., 2024; Xie et al., 2023a), highlighting frequency-domain differences across geographic regions, supporting further distributional analysis.

D3BV dataset: The Dialect Dominated Dataset of Bird Vocalisation (D3BV) (Jing et al., 2024) contains over 25 h of birdsong recordings from 10 avian species sampled across

three geographic zones within the contiguous United States (CONUS). This dataset emphasizes broad dialectal variation linked to regional biogeography rather than localized site-specific differences. The selected regions include the Western Cordillera (D1), Interior Plains (D2), and Eastern Highlands (D3), each defined by distinct topographical and ecological features. Recordings were obtained from public repositories such as Xeno-Canto and were annotated with species identity and corresponding GPS coordinates. A summary of species distribution across regions is provided in Table 1.

Western Cordillera (D1): Encompasses mountainous terrain, including the Rocky Mountains and Coastal Ranges.

Interior Plains (D2): Includes a flat, expansive region comprising the Great Plains and characterized by open grasslands.

Eastern Highlands (D3): Comprises the Appalachian Mountains and adjacent forested valleys.

S1S2 dataset: The S1S2 dataset (Xie et al., 2023a) comprises annotated birdsong recordings collected from two ecologically distinct sites in upstate New York: Albany (S1) and Lake George (S2). These locations differ markedly in vegetation structure, elevation, and degree of anthropogenic disturbance, providing a natural setting for evaluating site-dependent acoustic variation. The dataset includes bird species shared between both sites, enabling direct assessment of inter-site acoustic variability within identical taxa. Recordings were acquired using handheld digital recorders, with accompanying metadata specifying species identity and recording location. The final dataset includes 10 435 audio clips representing three bird species present at both sites (Table 2).

R1R2R3 dataset: To evaluate geographic variation in birdsong and assess model generalization across spatial domains, a cross-regional dataset (R1R2R3) was constructed, comprising field recordings from three ecologically and geographically distinct locations: Beijing (R1), Zhejiang (R2), and Cambodia (R3). This dataset includes vocalizations from eight bird species with known or inferred dialectal differences: *Acrocephalus bistrigiceps*, *Acrocephalus orientalis*, *Eudynamys scolopaceus*, *Lonchura striata*, *Phylloscopus borealis*, *Phylloscopus borealoides*, *Phylloscopus fuscatus*, and *Spilopelia chinensis*. Recordings were sourced from multiple public repositories, including Xeno-Canto, Macaulay Library of the Cornell Lab of Ornithology, and BirdCLEF challenge datasets (Table 3).

Species selection was informed by prior evidence of vocal divergence across region. For instance, the black-browed reed warbler (*A. bistrigiceps*) and oriental reed warbler (*A. orientalis*) demonstrate marked dialectal divergence, likely shaped by local adaptation and mate selection (Catchpole, 1983). The common cuckoo (*E. scolopaceus*) exhibits geographically structured call variation associated with isolation and genetic differentiation (Wei et al., 2015). Vocal learning and inheritance have been documented bar-winged Prinia (*Prinia familiaris*) and Chinese sparrow (*S. chinensis*) (Ritschard & Brumm, 2011). Additionally, the leaf warblers (*P. borealis*, *P. borealoides*, and *P. fuscatus*) display substantial interregional variation in song architecture, reflecting evolutionary divergence among populations (Tietze et al., 2015). These documented acoustic differences establish a

Table 1 Distribution of bird species in D3BV dataset across different regions

Bird species	Common name	Region D1	Region D2	Region D3
Turdus migratorius	American robin	1 038	187	791
Tringa semipalmata	Willet	138	29	106
Setophaga aestiva	American yellow warbler	730	9	297
Cardinalis	Northern cardinal	778	166	1 299
Agelaius phoeniceus	Red-winged blackbird	1 295	54	839
Setophaga ruticilla	American redstart	199	107	579
Spinus tristis	American goldfinch	221	94	283
Corvus brachyrhynchos	American crow	645	123	435
Certhia americana	Brown creeper	345	12	132
Molothrus ater	Brown-headed	392	50	96

Table 2 Distribution of select bird species in S1S2 dataset across different regions

Bird species	Common name	Region S1	Region S2
Corvus brachyrhynchos	American crow	221	1 477
Turdus migratorius	American robin	1 411	5 554
Cyanocitta cristata	Blue jay	770	1 002

Table 3 Distribution of select bird species in R1R2R3 dataset across different regions

Bird species	Common name	Region R1	RegionR2	Region R3
Acrocephalus bistrigiceps	Black-browed reed warbler	581	816	690
Acrocephalus orientalis	Oriental reed warbler	535	792	624
Eudynamys scolopaceus	Asian koel	253	126	342
Lonchura striata	White-rumped munia	290	300	408
Phylloscopus borealis	Arctic warbler	126	222	276
Phylloscopus borealoides	Kamchatka leaf warbler	534	234	276
Phylloscopus fuscatus	Dusky warbler	599	396	600
Spilopelia chinensis	Spotted dove	384	822	780

strong basis for the inclusion of these taxa in cross-region recognition research.

Proposed method

This study introduces a novel integrated framework designed to characterize and mitigate geographic variability in avian vocalizations. A schematic overview of the proposed approach is illustrated in Figure 1. The framework comprises three principal components: (1) MMD analysis, which quantifies cross-regional distributional shifts in acoustic feature space; (2) an adaptive normalization module that dynamically modulates normalization strength across channel, time, and frequency dimensions, with the normalized representations subsequently classified using the MHAResNet model; and (3) MINE, which evaluates the contribution of each feature dimension in species-level classification.

Recognition model architecture: To ensure experimental consistency, the same input representation and architecture as defined in our previous work (Xie et al., 2025a) was adopted. Each vocalization was encoded as a 3×224×224 tensor composed of power (POW), instantaneous frequency (IF), and group delay (GD) spectrograms. These feature maps were fed into the MHAResNet model, which integrates multi-head attention layers with a ResNet34 backbone to capture discriminative patterns in bird vocalizations. The final representation was aggregated using global average pooling and passed to a softmax classifier to produce posterior probabilities over predefined species and regional classes.

Quantifying distributional divergence with MMD: To assess acoustic variability across geographic domains, MMD was employed as a non-parametric metric for measuring the divergence between probability distributions. This approach is particularly well suited for detecting subtle distributional shifts in high-dimensional spectrogram feature spaces. Pairwise MMD values were computed between regions to quantify the extent of dissimilarity and identify divergence hotspots. MMD was formally defined as the distance between mean embeddings of two distributions within a reproducing kernel Hilbert space (RKHS), estimated via empirical kernel-based methods. The MMD formulation is shown in Equation (1):

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{m(m-1)}\sum_{i\neq i'} k(x_i, x_{i'}) + \frac{1}{n(n-1)}\sum_{j\neq j'} k(y_j, y_{j'}) - \frac{2}{mn}\sum_{i,j} k(x_i, y_j) \quad (1)$$

where x_i and x_i' are samples from region X, with a total of m samples; y_j and $y_{j'}$ are samples from region Y, with a total of n samples; $k(\cdot,\cdot)$ is the kernel function, typically a Gaussian Radial Basis Function (RBF) kernel; and $\widehat{MMD}_u^2$ is the unbiased MMD estimator that quantifies the dissimilarity between the distributions of bird vocalizations from two regions.

Adaptive normalization: To effectively normalize input features while preserving important information, an adaptive normalization method was proposed. This method dynamically adjusts the normalization proportions based on learnable parameters optimized during training. Adaptive normalization follows Equation 2:

$$AN(x) = \gamma_{\text{channel}} \cdot \text{Norm}_c(x) + \gamma_{\text{time}} \cdot \text{Norm}_T(x) + \gamma_{\text{freq}} \cdot \text{Norm}_F(x) \quad (2)$$

where $\text{Norm}_C(x)$, $\text{Norm}_T(x)$, and $\text{Norm}_F(x)$ represent the normalization results along the channel, time, and frequency dimensions, respectively, and γ_{channel}, γ_{time}, and γ_{freq} are the weights corresponding to these dimensions, learned during training to dynamically adjust the normalization proportions. Although these weights change dynamically, the sum remains equal to 1. Each dimension was normalized using the standard normalization formula shown in Equation (3).

$$\text{norm\_x} = \frac{x - \text{mean}(x)}{\text{std}(x) + \varepsilon} \quad (3)$$

where x is the input feature, mean(x) is the global mean of the input feature, std(x) is the global standard deviation, and ϵ is a small value added to prevent division by zero.

During training, the model optimizes three learnable parameters — γ_{channel}, γ_{time}, and γ_{freq}—which dynamically modulate the relative influence of channel, time, and frequency dimensions. These weights enable the model to emphasize the most informative dimensions for cross-regional generalization and species recognition. In the proposed architecture, the adaptive normalization module was

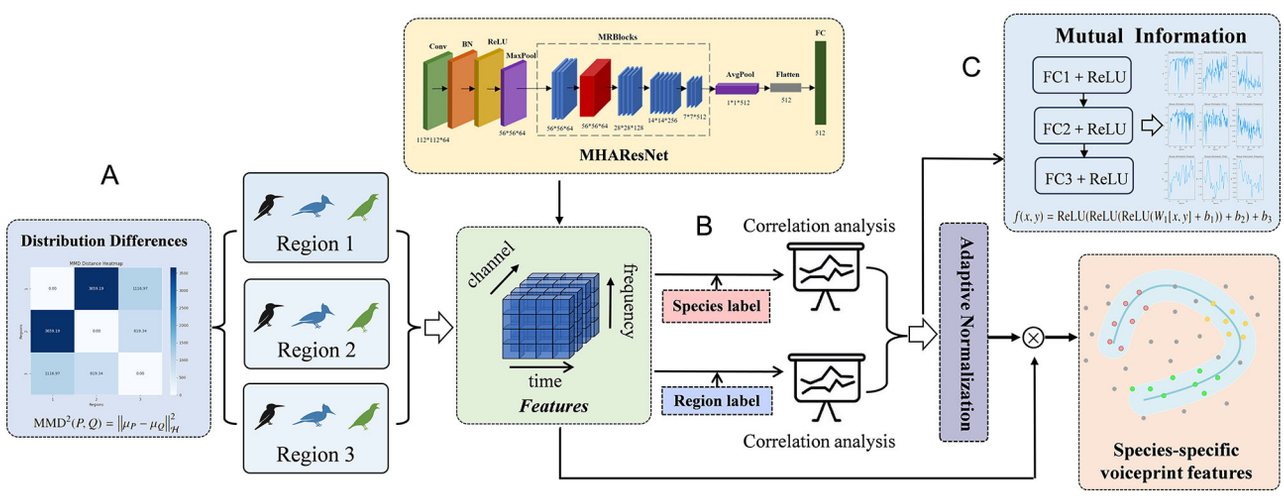


Figure 1 Overview of the proposed framework

A: MMD analysis for quantifying inter-regional divergence in acoustic feature distributions. B: Adaptive normalization module that dynamically reweights channel, time, and frequency dimensions based on task-specific relevance. C: MHAResNet classifier integrated with MINE for interpretability. Arrows indicate flow of information from raw spectrogram features (POW, IF, GD) to normalized representations and final classification outputs.

embedded directly within the neural network. Values for $\gamma_{channel}$, γ_{time}, and γ_{freq} were initialized with equal weights and updated through backpropagation. A softmax function was applied to these parameters to constrain their sum to unity, ensuring a normalized attention distribution over the three dimensions. This adaptive normalization layer precedes the ResNet backbone and operates on the three-channel spectrogram input prior to residual and attention processing. This ensures that subsequent convolutional and attention layers receive balanced and stable inputs.

Mutual information estimation via MINE: MINE was employed to accurately estimate mutual information (MI) between different dimensions of input features and target labels, using Equation 4 as follows:

$$I(X;Y) = E_{p(x,y)\sim p_{XY}}\left[\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right] \quad (4)$$

where $I(X;\ Y)$ is the mutual information between random variables X and Y, p_{XY} $(x,\ y)$ is the joint probability distribution, and $p_X(x)$ and p_Y (y) are the marginal probability distributions of X and Y, respectively.

In this implementation, mutual information is estimated using a neural network composed of three fully connected layers. The network receives concatenated feature-label pairs as input and produces a scalar output representing the estimated dependence. Training optimizes an objective that maximizes the difference between the expected network response under the true joint distribution and the expected response under the product of the corresponding marginal distributions.

The neural network architecture used in MINE was defined based on Equation (5):

$$f(x,y) = \mathrm{ReLU}\left(\mathrm{ReLU}\left(\mathrm{ReLU}\left(W_1[x,y] + b_1\right)\right) + b_2\right) + b_3 \quad (5)$$

where x and y are the input features and labels, respectively; W_1 is the weight matrix; b_1, b_2, and b_3 are the bias terms, and function f $(x,\ y)$ represents the output of the neural network, used to estimate the mutual information between input features and labels.

The MINE model was trained by minimizing the loss function defined in Equation (6):

$$L = -E_{(x,y)\sim p_{XY}}[f(x,y)] + \log E_{(x',y')\sim p_X p_Y}\left[\exp\left(f\left(x',y'\right)\right)\right] \quad (6)$$

where p_{XY} is the joint probability distribution of the input features x and labels y and p_X and p_Y are the marginal probability distributions of x and y, respectively. Terms $(x',\ y')$ represent samples drawn from the product of marginal distributions $p_X(x')$ and $p_Y(y')$.

During training, the model learns to assign higher values to the true joint distribution and lower values to the product distribution, effectively estimating the mutual information. In our experiments, MINE was applied to evaluate the mutual information between different dimensions of the input features (channel, time, and frequency) and target labels.

EXPERIMENTS AND RESULTS

This section presents the experimental evaluation of the proposed framework. First, regional divergence of birdsong distribution was quantified using MMD. Next, cross-regional generalization capability was assessed through recognition experiments involving geographically distinct datasets. Finally, model performance was compared before and after the application of adaptive normalization, and MINE was used to interpret the relevance of individual feature dimensions.

Experimental setup and evaluation metrics

All experiments were conducted using PyTorch v2.0.1 and Python v3.8.0 and executed on a workstation with an Intel i5-9300H CPU and an NVIDIA GTX 1050 GPU. Each model was trained for 100 epochs using the Adam optimizer, with an initial learning rate of 1×10^{-4} and a batch size of 16. To promote convergence and maintain training stability, a learning rate decay factor of 0.8 was applied every three epochs. Early stopping was employed to prevent overfitting, and L2 regularization was applied to all trainable parameters to improve generalization. Total loss function is defined in Equation (7):

$$L'(\theta) = L(\theta) + \lambda\frac{1}{2}||\theta||_2 \quad (7)$$

where $L'(\theta)$ and $L(\theta)$ denote the loss function before and after regularization, respectively, θ denotes all weighting parameters, and λ regulates the influence of the regularization term on loss.

Cross-entropy was used as the classification objective, as it balances class similarity and accelerates convergence in multi-class recognition tasks. The specific formula is shown in Equation (8):

$$L(\theta) = -\frac{1}{n}\sum_x\left[Y_r \ln Y_p + (1-Y_r)\ln\left(1-Y_p\right)\right] \quad (8)$$

where Y_r is the actual output label and Y_p is the corresponding predicted output.

For in-region recognition, each regional dataset was partitioned into training, validation, and test subsets with a 6:2:2 ratio. For cross-region evaluation, the model was trained on all samples from one region (Dn) and tested on a different region (Dm, n≠m), ensuring no data leakage to accurately reflect domain shift across geographic areas.

Model performance was assessed using Accuracy, Precision, Recall, and F1-score. Accuracy was treated as the primary metric and was calculated according to Equation 9. Precision, defined in Equation (10), measures the proportion of correctly classified samples within the predicted class. Recall, defined in Equation (11), measures the proportion of correctly identified samples within the true class. The F1-score, given in Equation (12), provides a balanced metric by combining Precision and Recall.

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$F1 = \frac{2TP}{2TP+FN+FP} \quad (12)$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

Distributional divergence and cross-region recognition

To evaluate the impact of regional variation in birdsong distributions on model generalization, two complementary experiments were conducted: (1) distributional divergence analysis using MMD, and (2) cross-region recognition

involving training on one region and testing on others. These experiments were performed on the D3BV, S1S2, and R1R2R3 datasets to characterize dialectal shifts and domain transfer effects across geographic contexts.

MMD was first applied to compute pairwise distributional distances between regions within each dataset. Heatmaps summarizing these results are presented in Supplementary Figure S4. For the D3BV dataset (Supplementary Figure S4a), the highest divergence was observed between D1 and D2 (3 659.19), followed by a moderate divergence between D1 and D3 (1 116.97), and the lowest between D2 and D3 (819.34). These patterns reflect their underlying ecological contrasts, with D1 comprising mountainous terrain, D2 representing flat grasslands, and D3 occupying an intermediate position in both geography and acoustic similarity. In the S1S2 dataset (Supplementary Figure S4b), a substantial divergence of 1 401.78 was observed between the two sites, attributable to differences in elevation, vegetation density, and anthropogenic disturbance despite their spatial proximity. In the R1R2R3 dataset (Supplementary Figure S4c), divergences were most pronounced between R1–R2 (22 677.81) and R1–R3 (19 157.00), while R2–R3 exhibited the smallest separation (489.53). These values are consistent with regional bioclimatic differences: R1 is geographically and ecologically distant from R2 and R3, while the latter two share more similar habitat types. Overall, MMD values corresponded closely with geographic and ecological distances, confirming that birdsong features are strongly shaped by environmental context and highlighting the importance of addressing domain shifts in cross-region recognition.

Cross-region recognition performance was then evaluated using the same region-to-region combinations. As shown in Table 4, the highest classification accuracy was consistently achieved under in-region evaluation: 89.8% for D1, 82.9% for D2, and 88.9% for D3. However, performance declined substantially when models were tested on unseen regions (e.g., D1→D2 or D2→D1), as shown in Supplementary Figure S5. These performance trends followed a geographic gradient: larger distances between training and test regions were associated with greater recognition degradation. This pattern is consistent with previous findings linking dialectal divergence to geographic separation (Lewis et al., 2021).

Importantly, the relationship between ecological contrast and recognition loss was consistent across datasets. The largest MMD value and performance drop were observed between the Western Cordillera (D1) and Interior Plains (D2), which differ sharply in topography and vegetation. In contrast, D1 and D3 (Eastern Highlands) exhibited smaller divergence values and milder recognition degradation, likely due to shared ecological features such as forested terrain. Collectively, these results demonstrate that regional ecological distance is a key determinant of dialectal variation and cross-region model generalization. The MHAResNet classifier effectively captured these gradients, reinforcing the link between environmental context and acoustic divergence.

Recognition performance using adaptive normalization

The effectiveness of adaptive normalization was evaluated through five experimental configurations on the D3BV dataset: no normalization, comprehensive normalization across all dimensions, and normalization applied individually to each dimension (channel, time, and frequency). Results are presented in Table 5. Compared to the unnormalized baseline, adaptive normalization yielded consistent improvements. For species classification, comprehensive normalization (ALL) increased accuracy from 85.3% to 88.7%. For region recognition, accuracy improved from 80.4% to 82.4%. These findings confirm the utility of adaptive feature scaling in enhancing recognition performance under cross-region conditions. A comparative visualization of these outcomes is provided in Supplementary Figure S6.

The combination of single-dimension normalization experiments and the evolution of adaptive normalization weights identified the functional contributions of each feature axis for species and region recognition. For species recognition, frequency-only normalization yielded the highest accuracy (90.1%), indicating that frequency variability is the dominant source of dialectal interference. In contrast, channel-only normalization slightly reduced performance (84.8%), suggesting that channel-specific information is important for distinguishing species and should be preserved. These results

Table 4 Cross-region recognition performance on the D3BV dataset

Model	Train data	Test data											
		D1				D2				D3			
		ACC	REC	PRE	F1	ACC	REC	PRE	F1	ACC	REC	PRE	F1
MHAResNet	D1	89.8%	89.5%	87.4%	88.6%	78.8%	76.7%	74.1%	69.9%	74.7%	67.1%	70.1%	89.8%
	D2	46.5%	46.8%	45.6%	40.1%	82.9%	82.6%	81.9%	82.4%	63.2%	49.7%	53.4%	46.5%
	D3	69.1%	64.5%	65.0%	62.8%	84.6%	66.9%	74.6%	68.1%	88.9%	87.9%	88.6%	69.1%

Note: Performance was evaluated using cross-corpus training and testing across three distinct regions (D1–D3), with results reported in terms of Accuracy (ACC), Recall (REC), Precision (PRE), and F1-score (F1).

Table 5 Recognition performance on the D3BV dataset with and without normalization

	Species				Region			
	ACC	REC	PRE	F1	ACC	REC	PRE	F1
Unnormalized	85.3%	83.9%	84.8%	83.9%	80.4%	81.1%	81.0%	82.1%
ALL	88.7%	88.5%	88.7%	88.7%	82.4%	81.4%	82.0%	81.7%
Channel	84.8%	84.8%	84.9%	84.2%	82.7%	82.9%	81.6%	81.6%
Time	85.1%	85.1%	85.2%	85.1%	81.4%	81.6%	81.4%	81.3%
Freq	90.1%	89.6%	90.1%	89.8%	72.1%	72.1%	70.9%	70.9%

ALL denotes comprehensive normalization across all dimensions; Channel, Time, and Freq represent normalization applied individually to the channel, time, and frequency dimensions, respectively.

are consistent with the adaptive weight trajectories shown in Figure 2A, where the model assigned the highest normalization weight to frequency and the lowest to channel—effectively suppressing frequency-induced variability while retaining channel cues critical for species discrimination.

For region recognition, the pattern was reversed. Channel-only normalization achieved the highest region-level accuracy (82.7%), whereas frequency-only normalization led to a pronounced decline in performance (72.1%). This suggests that frequency-domain features encode essential geographic signals and should not be aggressively normalized. The corresponding weight dynamics in Figure 2B reinforce this interpretation: the model progressively reduced the normalization weight on the frequency dimension—preserving region-discriminative spectral patterns—while increasing the weights on the channel and time dimensions to suppress task-irrelevant variation. The alignment between classification outcomes and normalization weight trajectories indicates that the adaptive module effectively identified and regulated feature dimensions based on their relevance to the recognition objective.

Mutual information estimation results

To quantify the relative contribution of each input dimension, MI was estimated using the MINE framework for both species and region recognition tasks on the D3BV dataset (Figure 3). For species recognition (Figure 3A), the channel dimension exhibited the highest MI value (~1.8), followed by time (~1.0), while the frequency dimension showed the lowest (~0.3). These results indicate that species-discriminative information is primarily encoded in channel-level features, such as power, instantaneous frequency, and group delay, with limited contribution from frequency patterns.

For region recognition (Figure 3B), the frequency dimension emerged as the most informative (~0.8), highlighting its capacity to capture geographic spectral signatures, including pitch contours and spectral shapes. Time features carried moderate information (~0.55), while channel features were the least informative (~0.22). These MI distributions correspond closely with the behavior of the adaptive normalization module: in each task, the model applied minimal normalization to the most informative dimension—channel for species recognition and frequency for region recognition—thereby preserving task-relevant signal structure.

Collectively, MI-based analysis confirmed that input dimensions contribute unequally to different recognition tasks. Channel-level features dominate species identification, while frequency-domain characteristics play a more prominent role in capturing geographic variation. These findings support the design of task-specific adaptive normalization and provide quantitative insight into the acoustic structure of birdsong across regions.

Validation across diverse datasets

To assess the generalizability of the proposed adaptive normalization framework, additional experiments were conducted on the S1S2 and R1R2R3 datasets, which differ from D3BV in geographic scale, ecological conditions, and species composition. Recognition results with and without adaptive normalization are summarized in Table 6. Across all datasets, adaptive normalization consistently improved species recognition accuracy from 85.3% to 88.7% on D3BV, 95.4% to 99.3% on S1S2, and 92.4% to 93.7% on R1R2R3, with corresponding gains in F1-score. Region recognition showed comparable improvements, with accuracy increasing from 80.4% to 82.4% for D3BV, 95.1% to 97.9% for S1S2, and 86.3% to 90.5% on R1R2R3. Notably, the largest performance gains were observed on the R1R2R3 dataset, highlighting the effectiveness of the model under ecologically heterogeneous, cross-continental settings.

These trends are consistent with prior findings from MI analysis and adaptive weight trajectories, confirming that the normalization framework enhances task-relevant feature representation—amplifying species-discriminative cues while attenuating region-specific variability.

To determine whether the learned normalization behavior generalizes across datasets, the evolution of adaptive normalization weights was tracked during training on D3BV, S1S2, and R1R2R3. As shown in Figure 4, the model consistently adjusted the importance of feature dimensions in a task-specific manner. For region recognition, the frequency dimension received the highest normalization weight in early training—particularly on D3BV—reflecting its strong geographic variability. Over the course of training, this weight gradually decreased, suggesting that the model extracted sufficient region-discriminative cues and prevented over-suppression. In contrast, for species recognition, the channel dimension consistently received the lowest normalization weight across all datasets, indicating that channel-level features encode stable, species-specific information that should remain intact. The frequency dimension, by contrast, was assigned higher normalization weight, consistent with its role in encoding region-dependent variability that may hinder species classification.

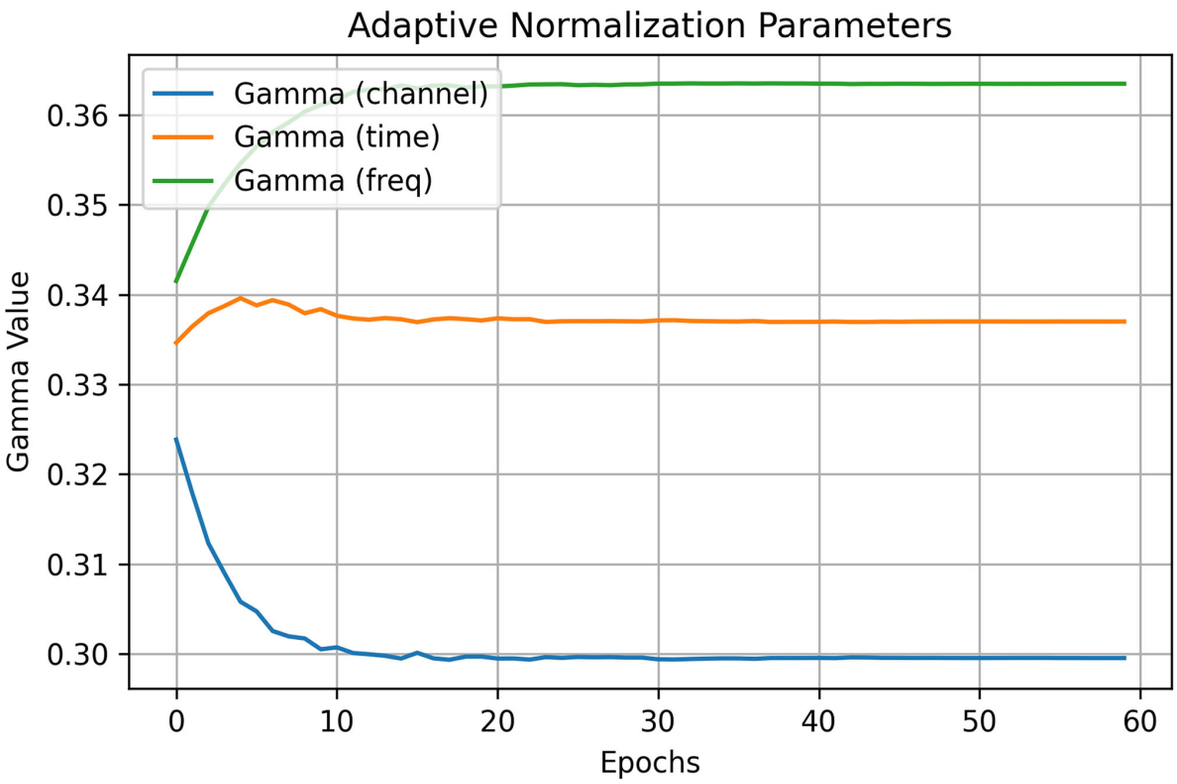
To validate these trends quantitatively, MI between input dimensions and classification targets was estimated using MINE across the D3BV, S1S2, and R1R2R3 datasets. This analysis aimed to identify the dimensions most relevant to each task and determine whether MI patterns align with adaptive normalization dynamics. Detailed MI curves and dataset-specific analyses are provided in Supplementary Figures S7 and S8.

Results revealed consistent task-specific trends across datasets. For species recognition, the channel dimension exhibited the highest MI, followed by time, with frequency contributing the least. For region recognition, the pattern reversed: frequency carried the highest MI, time remained moderate, and channel was the least informative. These findings mirrored the adaptive normalization behavior observed during training—dimensions with higher MI were assigned lower normalization weights to preserve task-relevant content. Overall, MI analyses confirmed that species recognition relies most heavily on channel-derived features, while region recognition depends more strongly on frequency-domain cues (Supplementary Figures S7–S8), underscoring the effectiveness and generalizability of the adaptive normalization strategy.

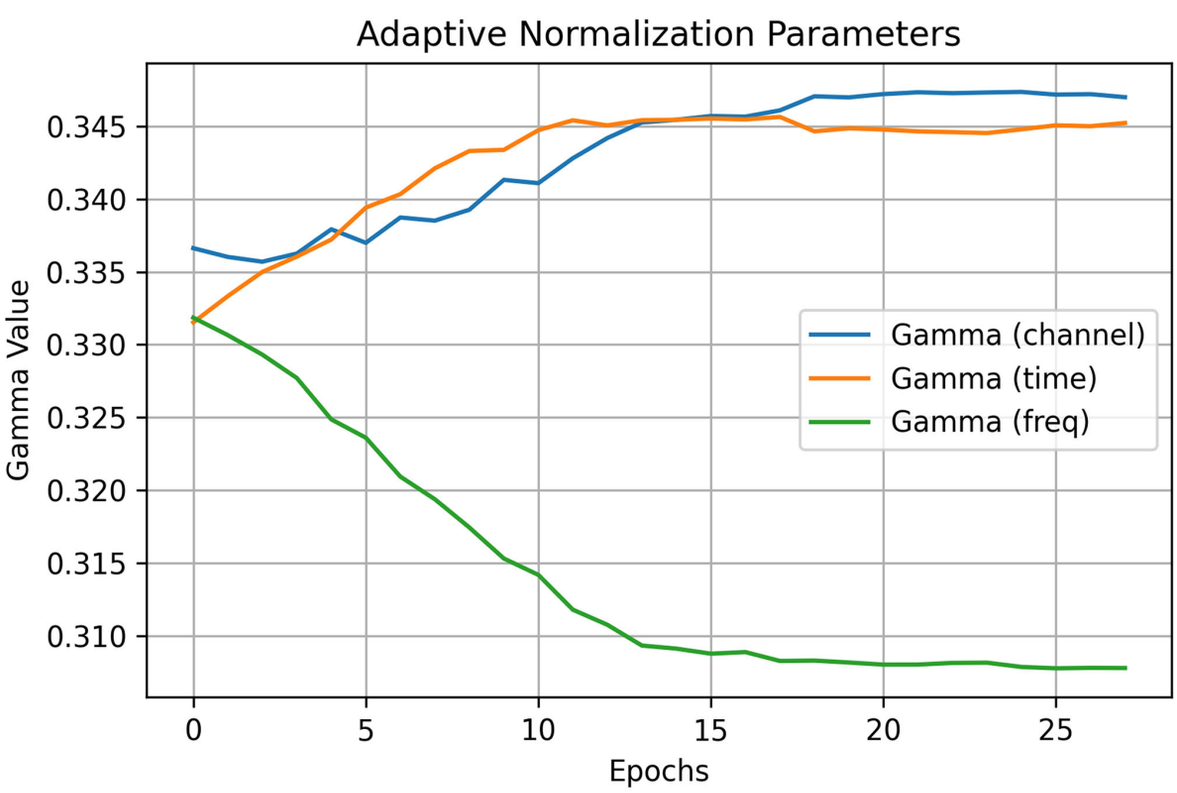
DISCUSSION

Cross-regional variability and the role of adaptive normalization

Spatial heterogeneity in birdsong—driven by dialectal divergence, environmental conditions, and recording devices—induces distributional shifts that pose significant challenges to acoustic recognition systems (Stowell et al., 2019). MMD-based analyses of the D3BV, S1S2, and



(a)



(b)

Figure 2 Evolution of adaptive normalization weights during training

A: Species recognition task. Each curve represents the learned weight assigned to a specific feature dimension (blue: channel, orange: time, green: frequency). B: Region recognition task. Colors are the same as in (A).

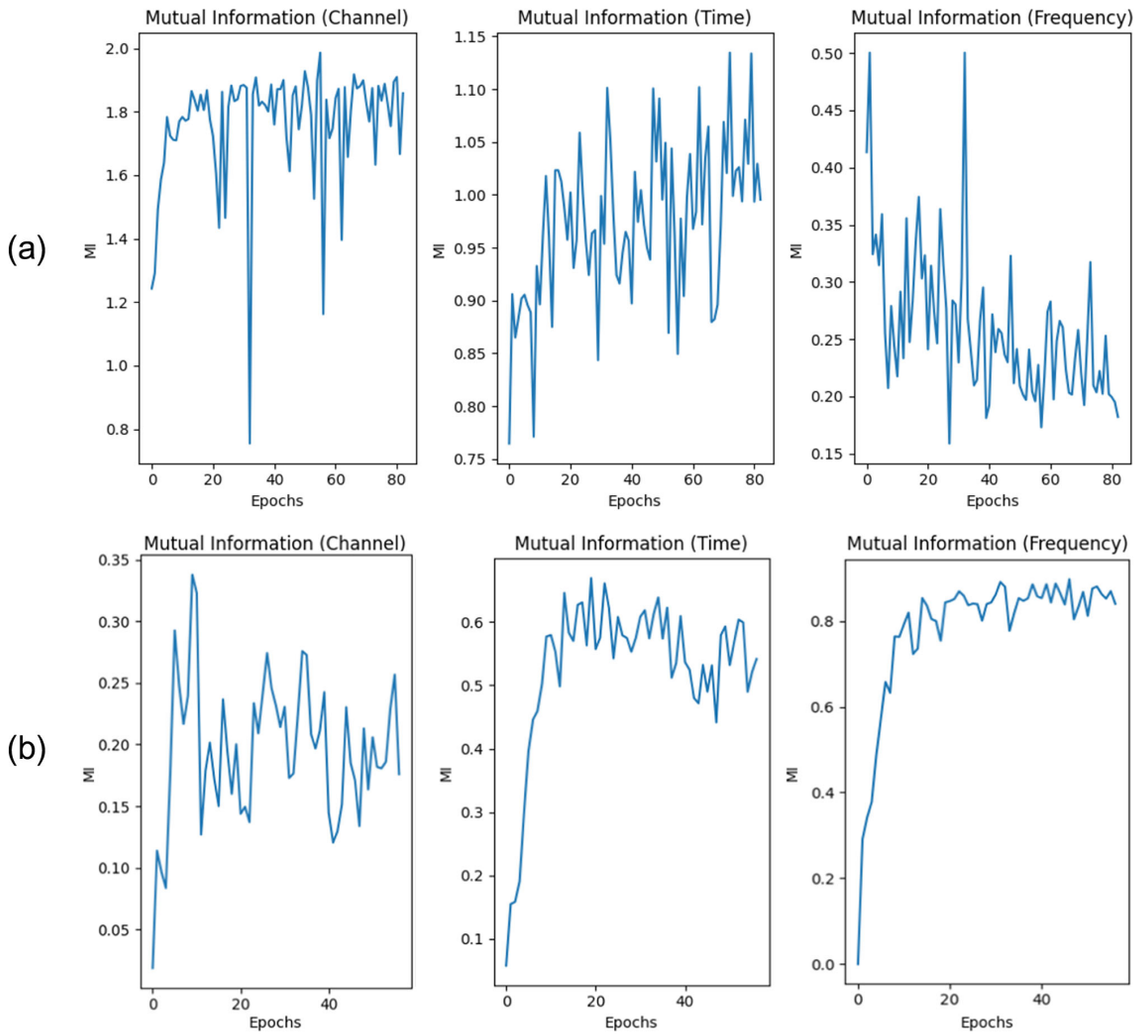


Figure 3 Mutual information between feature dimensions and recognition targets on the D3BV dataset

A: Species recognition; B: Region recognition.

Table 6 Recognition performance with and without adaptive normalization across datasets

Dataset	Species				Region			
	ACC	REC	PRE	F1	ACC	REC	PRE	F1
D3BV	85.3%	83.9%	84.8%	83.9%	80.4%	81.1%	81.0%	82.1%
D3BV-a	**88.7%**	88.5%	88.7%	88.7%	**82.4%**	81.4%	82.0%	81.7%
S1S2	95.4%	91.4%	93.5%	93.6%	95.1%	94.1%	93.7%	94.9%
S1S2-a	**99.3%**	99.3%	99.3%	99.3%	**97.9%**	96.8%	96.7%	96.7%
R1R2R3	92.4%	92.8%	92.5%	92.7%	86.3%	86.4%	87.6%	86.5%
R1R2R3-a	**93.7%**	93.8%	93.8%	93.8%	**90.5%**	90.5%	90.5%	90.5%

R1R2R3 datasets confirmed that the magnitude of such domain shifts varied substantially across regions, reinforcing the need for adaptive mechanisms that can account for dataset-specific variability. Our proposed framework addresses this by dynamically reweighting frequency, time, and channel dimensions according to their task relevance, consistently improving both species and region recognition across ecological contexts. Although earlier approaches have leveraged instance normalization with augmentation (Tang et al., 2022) or frequency-wise normalization (Kim et al., 2022; Xie et al., 2023a) to counteract acoustic shifts, the present findings underscore the necessity of explicitly identifying which dimensions encode region-dependent signals.

Frequency as a marker of geographic identity: Evidence from performance and information metrics

A key finding of this study is the dominant role of frequency-domain features in encoding regional variation in birdsong. This conclusion is supported by two lines of evidence. First, region classification performance deteriorated sharply when normalization was applied exclusively to the frequency dimension, indicating that frequency patterns are essential for distinguishing regional provenance. Second, MI estimates consistently revealed that frequency features carried the highest information content related to regional identity (e.g., 0.85 for frequency vs. 0.22 for channel in the R1R2R3 dataset), corresponding with ecological studies that associate dialectal differentiation with spectral attributes (Slabbekoorn & Smith, 2002; Tietze et al., 2015). Adaptive normalization weight trajectories further validated these findings: during

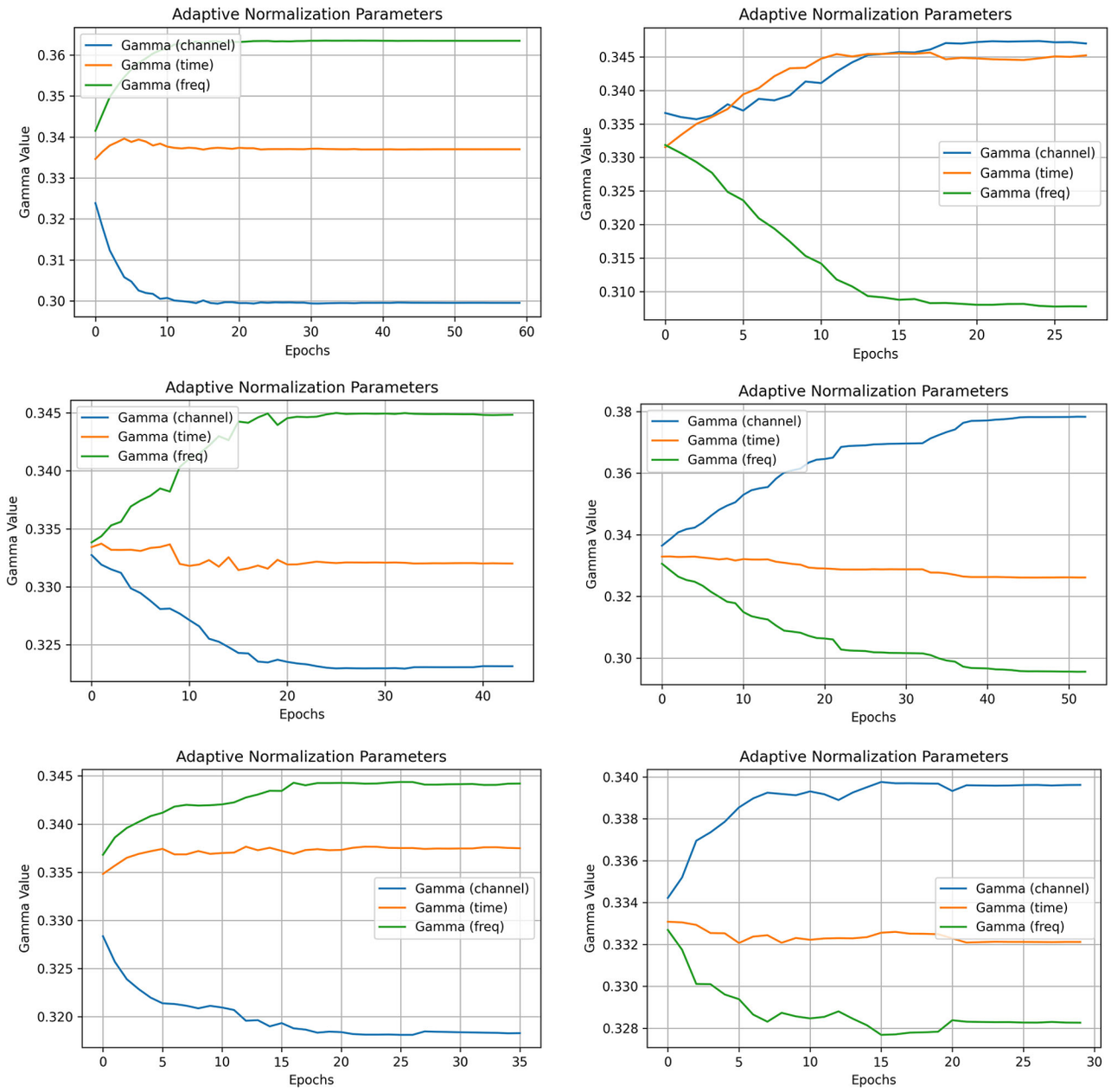


Figure 4 Evolution of adaptive normalization parameters across datasets

A: D3BV; B: S1S2; C: R1R2R3. Left: Species recognition; Right: Region recognition; Curves follow the same color scheme (blue: channel, orange: time, green: frequency).

training, the model progressively reduced normalization pressure on the frequency axis for region recognition tasks, while assigning greater normalization strength to frequency for species classification. This behavior aligns with principles of domain generalization, in which suppressing task-irrelevant variability enhances robustness and preserves discriminative content. These observations collectively establish frequency as a domain-specific axis for regional differentiation, whereas channel-level features provide a more stable basis for species classification.

Enhancing interpretability in data-driven ecological modeling

The MI-based analysis introduced here improved model interpretability by quantifying the associations between acoustic feature dimensions and classification targets, offering a transparent alternative to conventional black-box models. This framework enables: (1) more reliable species distribution tracking under changing climatic conditions by decoupling species-level cues from region-dependent variability (Lewis et al., 2021); (2) enhanced cross-site comparisons of vocal behavior within and between populations; and (3) systematic analysis of how environmental pressures shape vocal learning and dialectal divergence (Ritschard & Brumm, 2011). By isolating task-relevant features from those confounded by geography, the approach supports targeted conservation strategies and deepens mechanistic understanding of avian communication, complementing broader ecoacoustic research linking vocal variation to genetic differentiation and environmental stress gradients (Wei et al., 2015).

Implications for scalable PAM

The proposed framework reduced region-specific bias in large-scale PAM systems, enhancing consistency of species

detection across ecologically diverse habitats and improving the reliability of long-term population assessments. Despite these gains, several challenges persist in field deployments, including background noise contamination, variability in recording equipment, limited spatial coverage, sparse annotations in remote regions, and computational constraints on edge devices. Addressing these barriers will require future exploration of lightweight model variants, semi-supervised training paradigms using weak or noisy labels, and adaptive calibration strategies to maintain performance across heterogeneous monitoring conditions.

CONCLUSION

MI analysis across three geographically distinct datasets revealed that frequency-domain features predominantly encoded regional dialectal variation, while channel-wise spectral features more effectively captured species identity. Guided by this observation, an adaptive normalization framework was developed to dynamically adjust normalization strength across frequency, time, and channel dimensions. This task-aware mechanism suppressed frequency-driven variability during species recognition while preserving frequency content essential for regional classification. The resulting strategy consistently improved performance across both tasks and yielded interpretable normalization weights that clarified how species- and region-related information are separated. By enhancing domain generalization and feature-level interpretability, the proposed approach advances the robustness and transparency of passive acoustic monitoring systems, supporting scalable and accurate biodiversity assessment across heterogeneous acoustic landscapes.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy or ethical restrictions.

SUPPLEMENTARY MATERIALS

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests

REFERENCES

Belghazi MI, Baratin A, Rajeswar S, et al. 2018. MINE: mutual information neural estimation. arXiv preprint arXiv: 1801.04062.

Catchpole CK. 1983. Variation in the song of the great reed warbler *Acrocephalus arundinaceus* in relation to mate attraction and territorial defence. *Animal Behaviour*, **31**(4): 1217–1225.

Drossos K, Magron P, Virtanen T. 2019. Unsupervised adversarial domain adaptation based on the Wasserstein distance for acoustic scene classification. *In*: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz: IEEE, 259–263.

Gretton A, Borgwardt KM, Rasch MJ, et al. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, **13**(25): 723–773.

Jing X, Zhang LY, Xie JJ, et al. 2024. DB3V: a dialect dominated dataset of bird vocalisation for cross-corpus bird species recognition. *In*: Proceedings of the 25th Annual Conference of the International Speech Communication Association. Kos.

Kahl S, Wood CM, Eibl M, et al. 2021. BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics*, **61**: 101236.

Kasten EP, Gage SH, Fox J, et al. 2012. The remote environmental assessment laboratory's acoustic library: an archive for studying soundscape ecology. *Ecological Informatics*, **12**: 50–67.

Kim B, Yang S, Kim J, et al. 2021. Domain generalization on efficient acoustic scene classification using residual normalization. *In*: Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events 2021. 21–25.

Kim B, Yang S, Kim J, et al. 2022. Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification. *In*: Proceedings of the 23rd Annual Conference of the International Speech Communication Association. Incheon, 2393–2397.

Lauha P, Somervuo P, Lehikoinen P, et al. 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution*, **13**(12): 2799–2810.

Lewis RN, Williams LJ, Gilman RT. 2021. The uses and implications of avian vocalizations for conservation planning. *Conservation Biology*, **35**(1): 50–63.

Lu J, Zhang Y, Lv DJ, et al. 2023. Improved broad learning system for birdsong recognition. *Applied Sciences*, **13**(19): 11009.

Ma KP. 2016. Biodiversity monitoring relies on the integration of human observation and automatic collection of data with advanced equipment and facilities. *Biodiversity Science*, **24**(11): 1201–1202. (in Chinese)

Owens A, Efros AA. 2018. Audio-visual scene analysis with self-supervised multisensory features. *In*: Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 639–658.

Ritschard M, Brumm H. 2011. Effects of vocal learning, phonetics and inheritance on song amplitude in zebra finches. *Animal Behaviour*, **82**(6): 1415–1422.

Sedláček O, Vokurková J, Ferenc M, et al. 2015. A comparison of point counts with a new acoustic sampling method: a case study of a bird community from the montane forests of Mount Cameroon. *Ostrich*, **86**(3): 213–220.

Slabbekoorn H, Smith TB. 2002. Bird song, ecology and speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **357**(1420): 493–503.

Stowell D, Wood MD, Pamuła H, et al. 2019. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, **10**(3): 368–380.

Tang TT, Long YH, Li YJ, et al. 2022. Acoustic domain mismatch compensation in bird audio detection. *International Journal of Speech Technology*, **25**(1): 251–260.

Tietze DT, Martens J, Fischer BS, et al. 2015. Evolution of leaf warbler songs (Aves: Phylloscopidae). *Ecology and Evolution*, **5**(3): 781–798.

Wei CT, Jia CX, Dong L, et al. 2015. Geographic variation in the calls of the common cuckoo (*Cuculus canorus*): isolation by distance and divergence among subspecies. *Journal of Ornithology*, **156**(2): 533–542.

Xiao Y, Yin H, Bai JS, et al. 2024. Mixstyle based domain generalization for sound event detection with heterogeneous training data. arXiv preprint arXiv: 2407.03654.

Xie JJ, Hao ZL, Hu CH, et al. 2025a. Beyond amplitude: phase integration in bird vocalization recognition with MHAResNet. *Avian Research*, **16**(1): 100229.

Xie JJ, Wang YQ, Qian XY, et al. 2025b. Improving bird vocalization recognition in open-set cross-corpus scenarios with semantic feature reconstruction and dual strategy scoring. *IEEE Signal Processing Letters*, **32**: 1515–1519.

Xie JJ, Zhang LY, Zhang JG, et al. 2023a. Cross-corpus open set bird species recognition by vocalization. *Ecological Indicators*, **154**: 110826.

Xie JJ, Zhong YJ, Zhang JG, et al. 2023b. A review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecological Informatics*, **73**: 101927.