Evaluating Vocal Variety in a Suburban Resident Population of *Melospiza melodia*: Machine Learning as a Means to Survey Songbird Populations

Vidhur Prabhu

Word Count: 2562

Figure 1: Juvenile *Melospiza melodia*

1 Introduction

Avian vocal communication is a complex phenomenon which represents complex communication and cultural adaptation. Songbirds learn from a process known as song tutoring, or mimicry from older birds, similar to language development in humans (ten Cate, 2018). New World sparrow species (Passerellidae) are illustrators of this concept, observing both continental and microgeographical dialects in their vocalizations (Hensel et al., 2022)(Otter et al., 2020).

Bird vocalizations consist of both songs and calls. Calls, typically innate, tend to be uniform throughout populations and are not developed through song tutoring. Songs are complex and carry encoded semantic meaning, often sang by male individuals and learned through song tutoring (Arcese et al., 1988). Additionally, song must function over long distances due to cross-territory communication (Nelson et al., 2016).

Dialects and song variations play an important role in songbird culture and population genetics, as homogenized dialects result in reduced gene flow due to territory formation and mate choice (Lewis et al., 2021). High quality songs can also be preserved for many generations (Pichkar et al., 2024).

The song sparrow (*Melospiza melodia*) is among the most prevalent avian species in the contiguous United States. Its song is highly studied, colloquially known to resemble the opening of Ludwig van Beethoven's Symphony No. 5. Western subspecies of *M. melodia* exhibit a higher level of song sharing than their eastern counterparts (Foote and Barber, 2007). However, *M. melodia* across the continent are consistently familiar with neighboring songs (DuBois et al., 2016). Unlike other species of New World sparrows, female *M. melodia* show a preference for songs from neighboring territories (Hernandez et al., 2009)(Williams et al., 2024)(Aubin et al., 2024). Preferential tutoring has also been demonstrated in juvenile sparrows, where song sharing is increased by both tutor-neighbor songs which survive the winter and songs which are shared by several tutors (Beecher, 2017).

Low-frequency anthropogenic noise has been explained as an effector for *M. melodia* living in urban environments (Wood and Yezerinac, 2006). Thus, noisier locations resulted in higher frequency low notes and relatively less energy in the low frequency range of songs (Foote and Barber, 2007). In other species, similar characteristics have been observed, through with a constrained minimum (Phillips et al., 2020).

Generally, it is possible to understand which aspects of a song indicate semantic or geographical features through analysis of songs among various populations. Both the terminal strophe in Puget Sound white-throated sparrow *Zonotrichia albicollis* and the middle segment of savannah sparrow (*Passerculus sandwichensis*) song have been identified as population-specific indicators (Ramsay and Otter, 2015)(Otter et al., 2020)(Hensel et al., 2022). While there is research in analyzing the presence of dialects, there exists a barrier in that monitoring of vocalizations remains a specialized and resource-intensive task. Machine learning increases accessibility in understanding these dialects.

Passive-acoustic monitoring (PAM) is a noninvasive technique used to capture acoustic signals from field environments through the use of autonomous recording units (ARUs) (Kahl et al., 2021). PAM is increasingly cost-effective and widely used, while also acting as an permanent archive that can be verified independently (Kahl et al., 2021). However, analysis of PAM datasets remains technically challenging, requiring either intensive manual extraction or, more recently, efficient algorithmic analysis through machine learning (Kahl et al., 2021).

Various machine learning methods have been used in bioacoustics, though supervised deep learning models, namely Convolutional Neural Networks (CNNs) remain incredibly popular (Kahl et al., 2021). However, unsupervised learning algorithms, such as k-Means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), have also been incredibly powerful in analyzing unlabeled datasets, or when paired with supervised learning. As machine learning problems become focused towards intradataset classification and structure discovery, unsupervised clustering methods become more prevalent.

Dialect identification is a strong indicator of population fitness and health, especially in threatened or endangered species. Machine learning techniques have also been shown to aid in novel identification of new dialect features, which may explain cultural trends unrecognizable by human investigators (Wang et al., 2022). It can also provide insight into translocation of juveniles and migration patterns.

Despite increased use of machine learning methods in bioacoustics, unsupervised clustering remains underutilized in dialect identification among avian species (Story et al., 2024). While supervised algorithms have distinguished between species and dialects, unsupervised machine learning has not yet been assessed as a viable candidate in environments where labeled training data is not feasible to produce. I set out to survey and understand the vocal variety in a population of *M. melodia*, assessing the quality and feasibility of a unsupervised machine learning algorithm as a high-quality tool in bioacoustics and

vocal dialect research.

2 Methods

2.1 Data Collection

The primary dataset consisted of eight weeks of recordings from January through March 2026 using a PAM device from the Cornell University Lab of Ornithology K. Lisa Yang Center for Conservation Bioacoustics, the SwiftOne terrestrial autonomous recording unit (ARU). The device was mounted to a tree nearing a grass field in a suburban park 2. Recording was scheduled from 06:00 each morning and ended at 19:00 each evening local time. This ensured that the *M. melodia* vocalization activity window was captured during the recording period. Recordings were captured on alternating days during the recording period, yielding a dataset of 23520 minutes. Recording settings remained consistent with Cornell Lab guidance (factory default).

Figure 2: SwiftOne PAM Device

Figure 3: Captured Recording Area

2.2 Data Preparation

The full recorded archive was processed using an offline version of the BirdNET algorithm to isolate regions where *M. melodia* vocalizations occurred. BirdNET utilies a CNN trained on bird vocalizations from citizen science recordings to assign confidence levels for the prevalence of specific species throughout three-second audio clips (Kahl et al., 2021). A confidence threshold of 70% was required to use any given sample, minimizing false positives while preserving a sufficient final dataset of 376 clips.

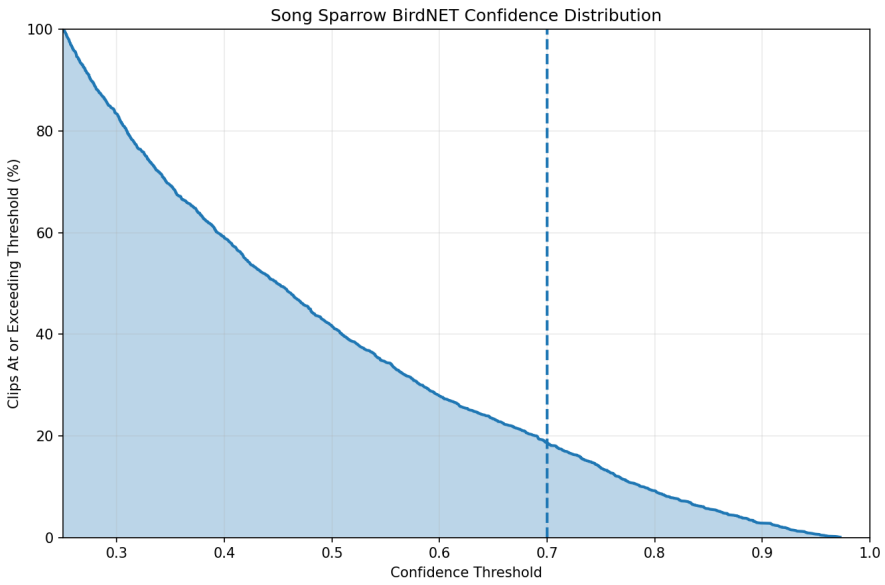

Figure 4: BirdNET Confidence Distribution

2.3 Preprocessing

Each retained clip was resampled to 22050Hz. Leading and trailing silence was trimmed. A mel-frequency cepstrum (MFC) time-frequency-amplitude representation of each clip was computed using a short-time Fourier Transform (STFT), which created a logarithmic spectrogram graph of the present *M. melodia* vocalization. See 5 for a representative spectrogram. The STFT is a linear transformation used to extract frequency content from a waveform audio while retaining temporal context. This becomes especially applicable to machine learning models, serving as a semantically meaningful representation. Mel-spectrograms are computed using a logarithmic mel scale, which more closely align with meaningful sound-patterns by emphasizing perceptually relevant frequency differences. This is particularly relevant to vocalization analysis as temporal patterns remain consistent while harmonic and frequency structures are clearly defined, improving the discriminability of vocalization types in machine learning algorithms.

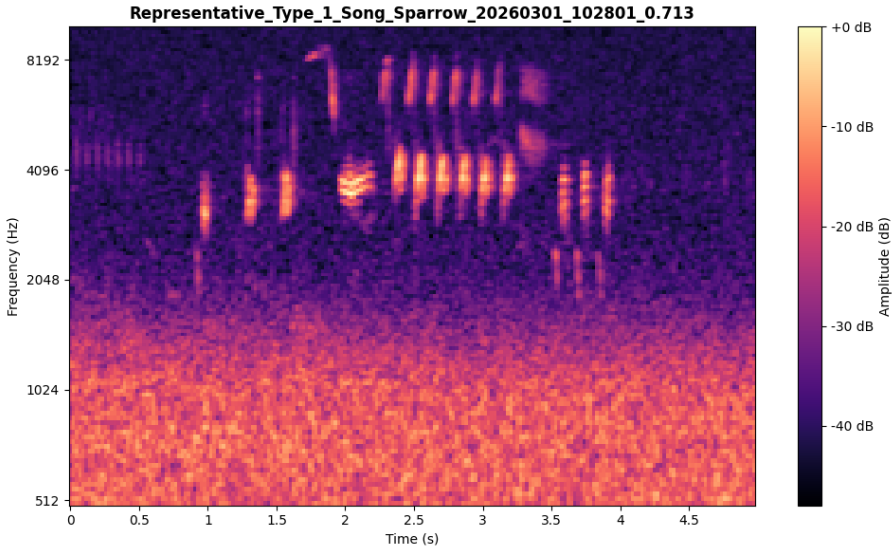

Figure 5: Example Spectrogram

The spectrogram was generated using a log-mel scale between 1-8 kHz, which encompasses the range of *M. melodia*. Each mel-band was independently normalized for loudness, then time-warped to 128 frames so that all spectrogram inputs would be temporally equivalent. The final feature vector was

formed through L2 normalization of the flattened one-dimensional vector. Each clip was presented to the model as a $8,192$ dimension vector. No data augmentation methods were applied.

2.4 Clustering

An unsupervised machine learning network was used to classify spectrograms by vocalization. As the final number of classes was not known, an unsupervised HDBSCAN algorithm pipeline was utilized.

Principal Component Analysis (PCA) was first applied to the $8,192$ dimension feature matrix to reduce it down to 50 components. PCA, in practice, reduces low-variance dimensions that are unlikely to carry meaningful acoustic information (i.e., similar across all clips or noisy).

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) was then applied for further dimensionality reduction from 50 components to a 2-dimension projection (McInnes et al., 2020). UMAP builds a weighted k-nearest-neighbor high-dimensional map from the PCA and finds the closest 2D projection that preserves topological structure. Compared to alternatives such as t-SNE (t-distributed Stochastic Neighbor Embedding), UMAP preserves structure more accurately and scales favorable to larger datasets, improving computational efficiency.

Clustering on the final UMAP embedding was performed using a Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2013). The number of clusters in HDBSCAN are unassumed, which is important in the application of categorizing vocalizations in *M. melodia*, as repetoire size was not known *a priori*. Insufficient density points were also labeled as as noise (unclassified). Due to ambient acoustic interference and overlapping vocalizations, some number of unclassified clips are expected.

2.5 Model Evaluation

Manual verification of clustering was conducted on a subset of the complete dataset. Each clip was reviewed by listening to the audio and inspecting the spectrogram, then assigned a binary label, song or call. Verification was calculated with a boolean accuracy as the proportion of clips within each cluster that matched its majority label. Unclassified clips were labeled as calls.

3 Results

3.1 Unsupervised

All 376 feature vectors were applied to the unsupervised algorithm clustering pipeline. The HDBSCAN model identified 27 clusters and one additional unclassified category. 22 clusters, and the unclassified category, were manually verified as call vocalizations. 46 of 376 clips (12.2%) were unclassified. The remainder of the clips were distributed with various membership sizes, with cluster 3 as a size outlier. Cluster distribution is shown in 6. All clips showed relatively stable prevalence throughout the recording period.

Of the five song vocalization clusters, two were shown to have highly similar acoustic structures upon visual inspection of spectrograms. These two clusters likely represent the same vocalization type with a minor acoustic difference that caused HDBSCAN to treat them as distinct dense regions in the UMAP embedding.

Song clusters were grouped similarly, generally separated from calls apart from cluster 24. The noise category was scattered throughout regions and particularly within the various call types, expected for limited-context clips that share partial features with multiple features. HDBSCAN clusters are shown in 7.

3.2 Manual Evaluation

Manual evaluation of clips drawn randomly showed that the model achieved a 95.29% accuracy when categorizing between call and song vocalizations, indicating strong agreement between the cluster assignments and the manually assigned vocalization labels. Cluster-level reports are shown in 1. Vocalization cluster 24 had a particularly high error rate, likely due to its proximity to many call clusters unlike other song clusters.

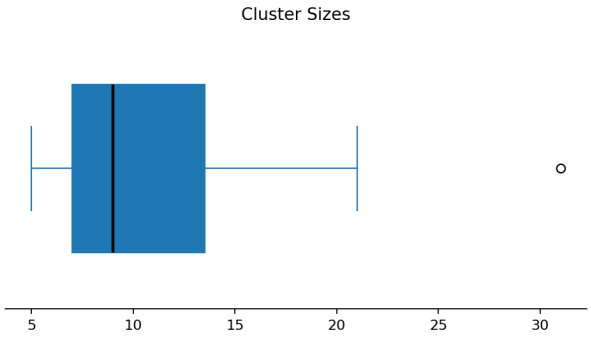

Figure 6: Cluster Size Distribution Box Plot

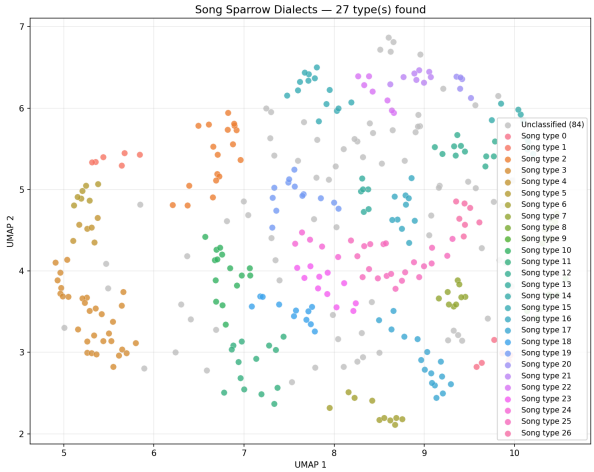

Figure 7: Categorized HDBSCAN Plot

3.3 Seasonal Variation

Vocalization detections and cluster composition varied across the eight-week recording period. A decline in call prevalence over the recording period was observed, suggesting a shift in *M. melodia* behavior from late Winter to the early Spring breeding season. However, no trend in song detections occurred, which indicates that the two behavioral functions are distinct and variable. Vocalization detections by recording week are shown in 8.

3.4 Temporal Variation

Vocalization detections varied substantially throughout the day. Song vocalizations peaked at 10:00 and the majority of all song detections occurred before 13:00. However, total detections peaked at 17:00, indicating greater call activity in the late afternoon. Vocalization detections by hour are shown in 9.

Vocalization	Correct/Total	% Correct
Vocalization 24	2/3	66.7%
Vocalization 3	10/13	76.9%
Vocalization 2	8/10	80.0%
Vocalization 11	7/8	87.5%
Unclassified	44/46	95.7%
Other		100.0%

Table 1: Manual Verification Results

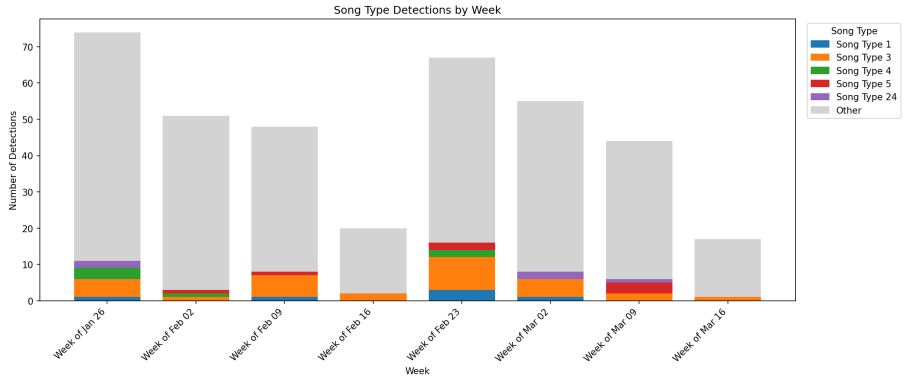

Figure 8: Vocalization Type Detections by Week

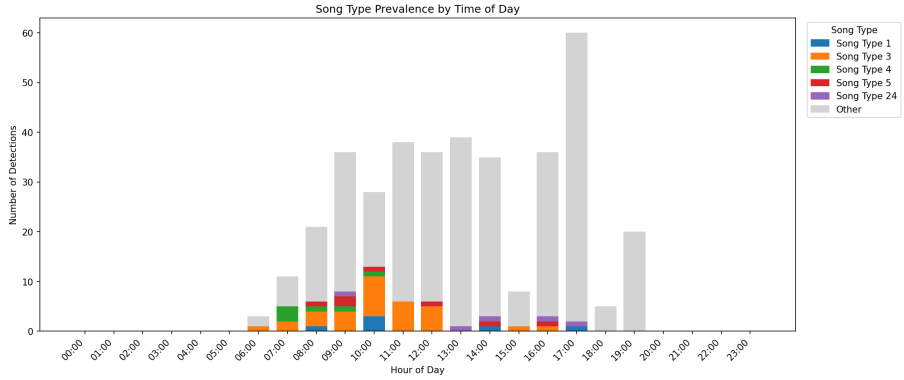

Figure 9: Vocalization Type Detections by Hour

4 Discussion

4.1 Population-Level Vocal Fingerprint

The unsupervised clustering pipeline identified four song vocalization clusters across 376 retained clips spanning eight weeks of passive-acoustic monitoring at a suburban park. The small song repertoire size suggests that this resident population of *M. melodia* presents as homogeneous in song vocalization during late winter months. This is consistent with prior observations of *M. melodia*, which exhibit elevated song sharing relative to eastern counterparts and that wintering resident populations stabilize, especially in wintering months (Foote and Barber, 2007). However, more vocalizations may emerge during spring breeding, indicating hibernation of song dialects.

The large number of calls indicates significant variance in frequency and tempo. However, it is unclear if there is a preference to any particular call type.

Little evidence in song change has been observed in Golden-crowned sparrows over longitudinal studies (Shizuka et al., 2016). In savannah sparrows, the buzz segment is a stable population marker, while the introductory segment has been known to change over the course of decades (Williams, 2021)(Williams et al., 2019). In rufous collared sparrows, geographical boundaries of dialects remained consistent, while certain acoustic variables changed over a 24 year period (García et al., 2015). Vocal variation can also ben an indicator of environment stress, observed through reduced song learning and complexity. The late afternoon surge in calls while songs remain low potentially represent the anthropogenic impact of the suburban setting, where increased noise may suppress song behavior and increase contact and alarm calls.

4.2 Machine Learning Effectiveness

The binary accuracy achieved in manual verification demonstrates that this pipeline is capable of analyzing culturally meaningful categorical structure from raw spectrogram features without labeled training data. Without prior knowledge of categorical data, the algorithm was able to demonstrate two a distinc-

tion between song and call types. This suggests that acoustic differences sufficiently evident and that HDBSCAN is a particularly well-suited algorithm to this classification task.

The identification of two similar song clusters as distinct represents both the sensitivity and a key limitation of HDBSCAN. Minor spectral divergence, potentially through recording angle, individual variation in a bird, or slight background noise, can produce distinct clusters. This was additionally evident through the numerous call clusters, which likely did not differ apart from syllable spacing and slight frequency variation. However, HDBSCAN also preserves this variation for manual investigation, which may prove useful to future work in understanding specific vocalization types.

4.3 Limitations

However, several limitations limit conclusions. Since recording was conducted during a period lacking cultural transmission and song tutoring, a full understanding of how machine learning interacts with larger categories and quantities of vocalizations is unclear. Applying the same pipeline to a breeding season dataset would substantially increase acoustic complexity, with increased song variants and novel vocalizations not present in the winter population.

Furthermore, as manual verification was conducted by a single reviewer, song clustering differences were not assessed. Future validation efforts should compare spectrograms between clusters and be cross verified by multiple reviewers.

The use of a single ARU also limits inference about spatial structure and population-level repertoires. The single device captures vocalizations from an unknown number of individuals within the detection radius. Cluster membership cannot be attributed to individual birds and observed song cluster structure could represent a single male's repertoire, a shared dialect among several neighboring males, or simply the range of the recording device.

4.4 Future Research and Conservation

The immediate extension of this work is replication with multiple ARUs deployed across a study area. Within a single park, each device would capture distinct local profiles, reflecting individuals within the recording area. Comparing cluster compositions across devices would also reveal dialect distribution, song tutoring, and individual movement. This could potentially reduce the need for banding, radio, or physical capture. Song migration across seasons could also reveal dispersal patterns and novel song variant formation through the population.

At a broader scale, vocalizations and preference for specific songs can be tracked must more accessibility and continuously using machine learning. Standardized feature extraction and clustering across multiple parks or habitats would allow quantitative comparison of diversity, repertoire stability, and species-level trends. This could indicate anthropogenic impacts and threats to a species.

Longitudinally, PAM devices across multiple breeding seasons would allow tracking of dialect stability and generational vocal culture. This dramatically lowers the barrier to long-term population surveillance, especially for threatened or geographically isolated species where population declines may precede detectable changes. Vocal diversity could also serve as early indicators of cultural erosion and genetic reduction.

Song repertoire size and complexity have also been associated with HVC volume in male songbirds (Pfaff et al., 2007). Machine learning-based repertoire profiling could, in future studies, serve as a non-invasive proxy for population-level health without requiring surgical examination of individuals.

Finally, multi-species acoustic monitoring also becomes available, enabling community-level vocal profiling to understand interspecies vocal responses to noise, seasonal shifts, and temporal activity windows. Ultimately, low-cost PAM harware and unsupervised machine learning positions bioacoustic monitoring as an efficient and accessible tool for population ecology monitoring for researchers and conservation organizations.

5 Conclusion

This study demonstrates that HDBSCAN is capable of identifying meaningful vocal structure within the songs of a *M. melodia* population without labeled training data. From 8 weeks of training data from a non-breeding season, the algorithm distinguished song from call vocalizations at 95.29% accuracy. These results establish unsupvised clustering as an informative bioacoustic tool capable of efficiently discovering trends within field datasets.

Acknowledgments

Thank you to the Eastside Audubon Society, Marymoor Park Management, and King County Department of Natural Resources and Parks for providing access to the recording space. Special thanks to the Cornell Lab of Ornithology K. Lisa Yang Center for Conservation Bioacoustics for use of the SwiftOne Terrestrial Autonomous Recording Unit.

References

Arcese, P., Stoddard, P. K. and Hiebert, S. M. (1988), 'The form and function of song in female song sparrows', *The Condor: Ornithological Applications* **90**(1), 44–50.
URL: *https://doi.org/10.2307/1368431*

Aubin, J. A., Dobney, S. L., Foreman, S. A., Doucet, S. M., Norris, D. R., Williams, H. and Mennill, D. J. (2024), 'Birds respond more strongly to locally common versus locally rare songs: a playback experiment with savannah sparrows', *Animal Behaviour* **212**, 127–135.
URL: *https://www.sciencedirect.com/science/article/pii/S000334722400099X*

Beecher, M. D. (2017), 'Birdsong learning as a social process', *Animal Behaviour* **124**, 233–246.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347216302020*

Campello, R. J. G. B., Moulavi, D. and Sander, J. (2013), Density-based clustering based on hierarchical density estimates, *in* J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu, eds, 'Advances in Knowledge Discovery and Data Mining', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172.

DuBois, A. L., Nowicki, S. and Searcy, W. A. (2016), 'A test for repertoire matching in eastern song sparrows', *Journal of Avian Biology* **47**(2), 146–152.
URL: *https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/jav.00811*

Foote, J. R. and Barber, C. A. (2007), 'High level of song sharing in an eastern population of song sparrow (melospiza melodia)', *The Auk* **124**(1), 53–62.
URL: *https://doi.org/10.1093/auk/124.1.53*

García, N. C., Arrieta, R. S., Kopuchian, C. and Tubaro, P. L. (2015), 'Stability and change through time in the dialects of a neotropical songbird, the rufous-collared sparrow', *Emu - Austral Ornithology* **115**(4), 309–316.
URL: *https://doi.org/10.1071/MU14099*

Hensel, A. L., Dobney, S. L., Doucet, S. M., Ryan Norris, D., Newman, A. E., Williams, H. and Mennill, D. J. (2022), 'Microgeographical variation in birdsong: Savannah sparrows exhibit microdialects in an island population', *Animal Behaviour* **188**, 119–131.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347222001026*

Hernandez, A. M., Pfaff, J. A., MacDougall-Shackleton, E. A. and MacDougall-Shackleton, S. A. (2009), 'The development of geographic song preferences in female song sparrows melospiza melodia', *Ethology* **115**(6), 513–521.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0310.2009.01642.x*

Kahl, S., Wood, C. M., Eibl, M. and Klinck, H. (2021), 'Birdnet: A deep learning solution for avian diversity monitoring', *Ecological Informatics* **61**, 101236.
URL: *https://www.sciencedirect.com/science/article/pii/S1574954121000273*

Lewis, R. N., Williams, L. J. and Gilman, R. T. (2021), 'The uses and implications of avian vocalizations for conservation planning', *Conservation Biology* **35**(1), 50–63.
URL: *https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13465*

McInnes, L., Healy, J. and Melville, J. (2020), 'Umap: Uniform manifold approximation and projection for dimension reduction'.
URL: *https://arxiv.org/abs/1802.03426*

Nelson, D. A., Szeyller, E. and Poesel, A. (2016), 'Alerting and message components of white-crowned sparrow song differ in structure and environmental transmission', *Behaviour* **153**(3), 263 – 285.
URL: *https://brill.com/view/journals/beh/153/3/article-p263$_2$.xml*

Otter, K. A., Mckenna, A., LaZerte, S. E. and Ramsay, S. M. (2020), 'Continent-wide shifts in song dialects of white-throated sparrows', *Current Biology* **30**(16), 3231–3235.e3.
URL: *https://doi.org/10.1016/j.cub.2020.05.084*

Pfaff, J. A., Zanette, L., MacDougall-Shackleton, S. A. and MacDougall-Shackleton, E. A. (2007), 'Song repertoire size varies with hvc volume and is indicative of male quality in song sparrows (melospiza melodia)', *Proceedings of the Royal Society B: Biological Sciences* **274**(1621), 2035–2040.
URL: *https://doi.org/10.1098/rspb.2007.0170*

Phillips, J. N., Rochefort, C., Lipshutz, S., Derryberry, G. E., Luther, D. and Derryberry, E. P. (2020), 'Increased attenuation and reverberation are associated with lower maximum frequencies and narrow bandwidth of bird songs in cities.', *Journal of Ornithology* p. 593–608.

Pichkar, Y., Searfoss, A. M. and Creanza, N. (2024), 'Detecting cultural evolution in a songbird species using community science data and computational modelling', *Animal Behaviour* **210**, 331–345.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347224000204*

Ramsay, S. M. and Otter, K. A. (2015), 'Geographic variation in white-throated sparrow song may arise through cultural drift', *Journal of Ornithology* **156**(3), 763–773.
URL: *https://doi.org/10.1007/s10336-015-1183-8*

Shizuka, D., Lein, M. R. and Chilton, G. (2016), 'Range-wide patterns of geographic variation in songs of golden-crowned sparrows (zonotrichia atricapilla)', *The Auk* **133**(3), 520–529.
URL: *https://doi.org/10.1642/AUK-16-27.1*

Story, B., Gillespie, P., Derryberry, G., Derryberry, E., Fefferman, N. and Maroulas, V. (2024), 'Dialectdecoder: Human/machine teaming for bird song classification and anomaly detection', *Ecological Informatics* **82**, 102657.
URL: *https://www.sciencedirect.com/science/article/pii/S1574954124001997*

ten Cate, C. (2018), 'The comparative study of grammar learning mechanisms: birds as models', *Current Opinion in Behavioral Sciences* **21**, 13–18. The Evolution of Language.
URL: *https://www.sciencedirect.com/science/article/pii/S2352154617301651*

Wang, D., Forstmeier, W., Farine, D. R., Maldonado-Chaparro, A. A., Martin, K., Pei, Y., Alarcón-Nieto, G., Klarevas-Irby, J. A., Ma, S., Aplin, L. M. and Kempenaers, B. (2022), 'Machine learning reveals cryptic dialects that explain mate choice in a songbird', *Nature Communications* **13**(1), 1630.
URL: *https://doi.org/10.1038/s41467-022-28881-w*

Williams, H. (2021), 'Mechanisms of cultural evolution in the songs of wild bird populations', *Frontiers in Psychology* **Volume 12 - 2021**.
URL: *https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.643343*

Williams, H., Dobney, S. L., Robins, C. W., Norris, D. R., Doucet, S. M. and Mennill, D. J. (2024), 'Familiarity and homogeneity affect the discrimination of a song dialect', *Animal Behaviour* **209**, 9–20.
URL: *https://www.sciencedirect.com/science/article/pii/S000334722300310X*

Williams, H., Robins, C. W., Norris, D. R., Newman, A. E. M., Freeman-Gallant, C. R., Wheelwright, N. T. and Mennill, D. J. (2019), 'The buzz segment of savannah sparrow song is a population marker.', *Journal of Ornithology* p. 217–227.

Wood, W. E. and Yezerinac, S. M. (2006), 'Song sparrow (melospiza melodia) song varies with urban noise', *The Auk* **123**(3), 650–659.
URL: *https://doi.org/10.1093/auk/123.3.650*