UNSUPERVISED DISCOVERY AND ANALYSIS OF THE VOCAL REPERTOIRES AND PATTERNS OF SELECT CORVID SPECIES

Nitin Sudarsanam[⋆] *Sahla Kader*[†] *Isaac Fernandezlopez*[⋆] *Sophie Huang*[†]
Tuan M. Dang[†] *Theron S. Wang*[†] *Hridayesh Lekhak*[†] *Kenny Q. Zhu*[†]

[⋆] Brown University [†] University of Texas at Arlington

ABSTRACT

Corvids are renowned for cognitive and social complexity, yet the structure of their vocal communication remains poorly understood. We analyzed recordings from five *Corvus* species, extracting call units and sequences, and identified call types via unsupervised clustering. Calls were described by 24 acoustic features, including a novel within-call repetition metric, and sequential structure was assessed with 1-4 gram models using perplexity. Features distinguishing species differed from those defining clusters, revealing substantial intra-species vocal diversity. Bigram models best captured vocal sequences across four species, with higher-order models also fitting the American crow. These results show Corvids produce structured, repeated sequences and demonstrate how big data analyses can help interpret songbird vocal structure.

Index Terms— Corvids, Vocalization, Bioacoustics, Clustering, N-gram

1. INTRODUCTION

Crows (Genus *Corvus*) have long been recognized for their cognitive abilities and complex social lives [1, 2]. Complex social environments are accepted to drive species evolution of more intricate communication systems to aid in navigating social relationships [3]. This makes crows ideal for studying more intricate communication and its evolutionary implications, including potential insights into the development of human language.

Previous research has largely focused on the acoustic structure of calls and their potential referential or behavioral meaning [2]. Studies in songbirds, including Japanese tits (*Parus minor*) [4] and southern pied babblers (*Turdoides bicolor*) [5] suggest the presence of compositional syntax, raising the possibility that corvids exhibit compositionality.

Although simplistic, Markov and n-gram models have proven effective in studying animal communication and NLP [6, 7]. Although Kershenbaum et al. (2014) [8] note their limitations for animal vocal sequences, we believe simpler models remain valuable as a part step for understanding call structure.

However, much of this work relies on small datasets, both in terms of audio clips and individual birds, limiting the generalizability of findings. Manual annotation of these data were helpful and necessary in the past, but modern machine learning approaches now allow analysis of much larger datasets, overcoming previous constraints and enabling big data investigation of vocal structure across multiple species.

This study presents three novel contributions: (i) the first cross-species acoustic analysis across five corvid species conducted at this scale (380 hours and 87,000+ calls); (ii) a novel "peak count" repetition metric capturing temporal structure; (iii) an empirical demonstration of shared acoustic structures across corvid species and both substantial inter and intra-species diversity. This represents the largest-scale evidence that corvid vocalizations exhibit structured sequential properties beyond species-specific call descriptions.

2. DATASET

2.1. Audio Denoising

For our work, species were chosen for comparison based on the large amount of data available in these public datasets, as well as their phylogenetic placements in relation to the American crow. We used the eight clade system and genetic phylogeny as presented by Jønsson [1] to gauge genetic relationships between species. Clade IV includes the Hooded Crow (*C. cornix*/HCRW), Carrion Crow (*C. corone*/CACR), and the American Crow (*C. brachyrhynchos*/AMCR), while the Common Raven (*C. corax*/CORA) represents the closer Clade V and the Fish Crow (*C. ossifragus*/FICR) represents the more distant Clade III. All our audio was sourced from Macaulay Library [9], broken down in Table 1.

To address environmental and recording noise present in our data, we applied audio denoising before segmenting samples into call sequences. We evaluated three commonly used methods: AudioSep [10], a foundation model for open-domain source separation, biodenoising [11], designed by the Earth Species Project for denoising animal recordings,

This work is conducted at the National Science Foundation *REU site: Animal Language Processing and Understanding* at University of Texas at Arlington and is supported by NSF Award No. 2349713.

Table 1: Data statistics by species. Note: Avg Length is weighted by # of calls per species.

Species	Raw Audio	# of Calls	Avg Length (s)
AMCR	125:34:50	34,343	2.84
CORA	103:30:48	24,168	1.82
FICR	66:41:03	23,677	1.77
CACR	54:41:27	4,091	2.24
HCRW	30:31:26	1,468	2.11
Total	**380 hrs**	**87,747**	**2.23**

and noisereduce [12], a Python noise reduction algorithm using spectral gating for time series data. AudioSep removed some background noise, but incompletely, and biodenoising introduced sound distortion, while noisereduce was the only model to significantly improve audio (see Section 4.1).

3. METHODS

3.1. Call and Sequence Extraction

After denoising, we segmented audio into call sequences, defined as one or more calls within a single behavioral context. Corvid calls often contain short silence intervals [13]. To avoid splitting within such calls, we defined sequences as bouts of calls separated by $>=$ 10 s of silence (defined as audio below -60 dBFS). To preserve calls, the beginning and end of each sequence was padded with an extra 0.5 s of the original audio, yielding 35,783 sequences.

We defined a call as a continuous vocal utterance delimited by at least 0.5 seconds of silence. We decided on both thresholds through manual review as shorter thresholds fragmented continuous call bouts, while longer thresholds merged distinct sequences. From our dataset of call sequence audio clips, we used the PANNs [14] SED model to identify crow-associated frames above a 0.05 confidence threshold. Sequences with $>=$ 50 consecutive frames (0.5 s) below this threshold were treated as silence, splitting the preceding segment into a call. Each call had to be at least 2 frames long, with 3 frames original audio padded on both sides. We kept track of the source sequence and call order, yielding 87,747 calls.

3.2. Clustering and Features

A set of 24 pre-determined acoustic features (PAFs) was used to both assess cluster quality and to compare vocal qualities across our focal species. Of these characteristics, 23 were drawn from Mates et al. [15], using 2 KHz as the maximum pitch measurement due to evidence that it is the maximum of the optimal vocal range [16]. We further included an additional feature, referred to as "Peak Count", to reflect recent findings that suggest that some crow species can reliably control the number of vocal units produced in a single call and discriminate between numbered stimuli [17, 18]. For each audio signal $y[n]$, the RMS amplitude envelope was computed:

$$e[m] = \sqrt{\frac{1}{L}\sum_{k=0}^{L-1} y[mH+k]^2},$$

then smoothed with a Gaussian filter

$$\tilde{e}[m] = \sum_j e[m-j]\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{j^2}{2\sigma^2}\Big),$$

and normalized as $\hat{e}[m] = \tilde{e}[m]/\max_m \tilde{e}[m]$. Peaks were detected as local maxima exceeding a relative height $h = 0.3$ and separated by at least $d = 5$ envelope frames. This serves as an automated way to estimate the number of vocal units (caws or rattle chirps) within a single call, a task that often requires hours of manual annotation.

To cluster, we selected the GMM clustering method [19]. To select the optimal number of clusters, we used the Bayesian Information Criterion (BIC) to test all cluster numbers between 2 to 100 using the Gaussian Mixture Model clustering method with a diagonal covariance matrix. As the global BIC minimum (at $n = 98$) produced over-fragmentation, we selected $n = 20$ which lies approximately one standard deviation below the global minimum.

3.3. Cross-Species and Cluster Analyses

We performed a comprehensive acoustic cluster analysis on GMM-derived soft-cluster labels, with corresponding PAF vectors. After assigning each call to a GMM cluster, we characterized call counts per cluster and identified primary features that distinguished clusters. We also identified primary features that distinguished our 5 species. Multivariate differences among clusters and species were assessed with MANOVA [20] on z-scored features, followed by univariate ANOVAs [21] per feature. ANOVA p-values were adjusted with Benjamini-Hochberg [22] and accompanied by the η^2 effect. We downsampled all species to 1,468 for the cross-species tests to avoid class imbalance.

3.4. Sequence Analyses

After clustering 80,000+ audio clips with our PAFs, we evaluated sequence structure using 1-4 gram models with 1,000 bootstrap train/test splits. This bootstrap procedure provided robust distributions of perplexity for comparing model orders. Pairwise two-sample t-tests (or Wilcoxon signed-rank tests [23] for paired differences), with p-values adjusted using the Benjamini-Hochberg procedure [22], assessed differences between models and groups, while ANOVA [21] and Kruskal-Wallis [24] handled multi-group comparisons. To identify

Our data and code is available at: `https://github.com/UTA-ACL2/corvids_vocal_repertoire`

over-represented transitions, we applied one-sided binomial tests [25] (null: uniform cluster probability) with FDR correction [22] to account for multiple comparisons. These tests ensured reliable evaluation of model fit, group differences, and transition structure while controlling for false positives.

4. RESULTS AND DISCUSSION

4.1. Denoising Results

To assess intra-rater reliability across our 4 annotators, we calculated the intraclass correlation coefficient (ICC). This gave us an ICC score of 0.848, 95% CI [0.81, 0.88], p = 6.46 x 10-72, suggesting good reliability between scorers [26]. We then calculated the average score for each audio file and used this score for all subsequent analyses on model performance.

The average score for each treatment is as follows: AudioSep = 2.40, biodenoising = 2.43, noisereduce = 2.56, Raw = 2.13. Regardless of treatment, all scores were fairly low. We believe that this is an artifact of strict scoring criteria and does not indicate poor data quality overall. We then ran a one-way ANOVA test in R (Version 4.3.1) which revealed a significant effect of treatment on average audio quality score, F(3, 196) = 2.995, p = 0.032. We then ran a post hoc Tukey's Honest Significant Difference (TukeyHSD) [27] test to identify significant differences in pairwise comparisons. This analysis revealed that only Noisereduce and Raw treatment had a significant difference, p = 0.0206.

4.2. Species Comparisons

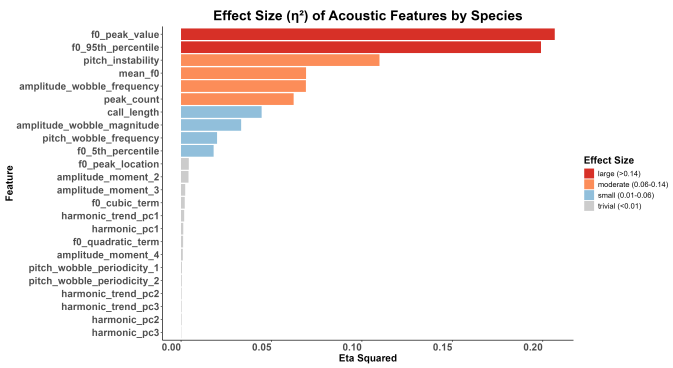

Fig. 1: Effect Size of Acoustic Feature by Species.

Our univariate ANOVA test run on the 5 *Corvus* species found the most important features in distinguishing species to be the fundamental frequency peak value and 95th percentile (see Figure 1). All the 3 clades differ in pitch in the expected order, confirming the biological consensus with big data. Interestingly, the Carrion Crow and Hooded Crow, which both belong to the same clade have similar distributions to each other and are relatively similar to the American Crow. However, the need to downsample prevents firm conclusions.

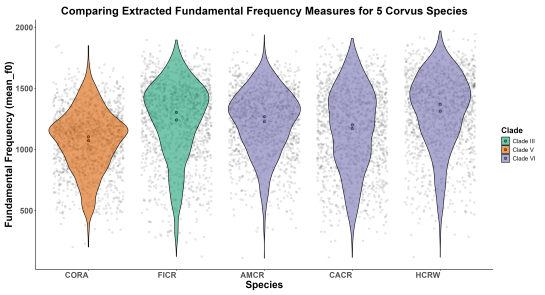

Fig. 2: Comparing Extracted F_0 Measures.

4.3. Clustering Results

To validate cluster distinctiveness, we characterized each cluster using our PAFs and applied the aforementioned statistical tests for discrimination and homogeneity. One-way ANOVAs revealed differences across all acoustic parameters ($p < 0.001$) with large effect sizes [28] ($F > 1{,}900$, $\eta^2 > 0.28$) for 14 out of 24 features. Post-hoc comparisons showed 88% of cluster pairs differed in pitch wobble frequency and amplitude wobble magnitude. Within cluster homogeneity was high (CV < 0.10 for 21/24 features), though cluster 8 was comprised of miscellaneous non-corvid sounds.

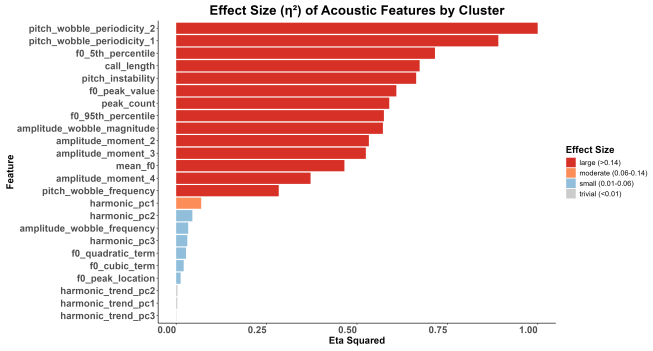

Fig. 3: Effect Size of Acoustic Feature by Cluster.

For supervised discrimination and feature ranking, we trained Random Forests to predict cluster identity. Models used a 70/30 train/test split, class-weighting to offset class imbalance, and out-of-bag error profiles to choose tree counts (final model ntree = 72). Performance was evaluated on the held-out test set with a normalized confusion matrix. As shown in Figure 4, clusters were easily distinguishable by their acoustic features.

Importantly, the features with the highest effect sizes for distinguishing clusters differ from the primary features for species differentiation. This strongly indicates that variance of vocal repertoire is not determined solely, or even primarily, by difference in species. Rather, significantly more features encode variation intra-species than inter-species, suggesting that each *Corvus* species independently possesses substantial vocal variation, which is a foundation for language [29].

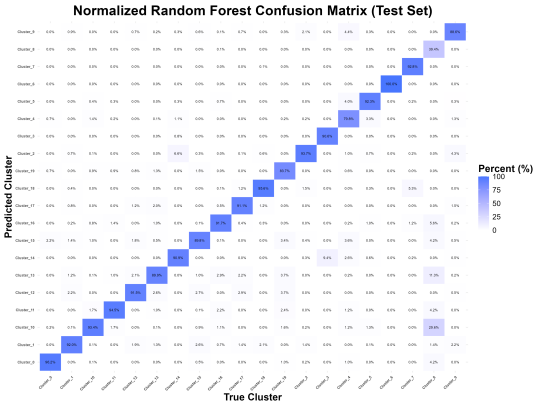

Fig. 4: Normalized Random Forest Confusion Matrix.

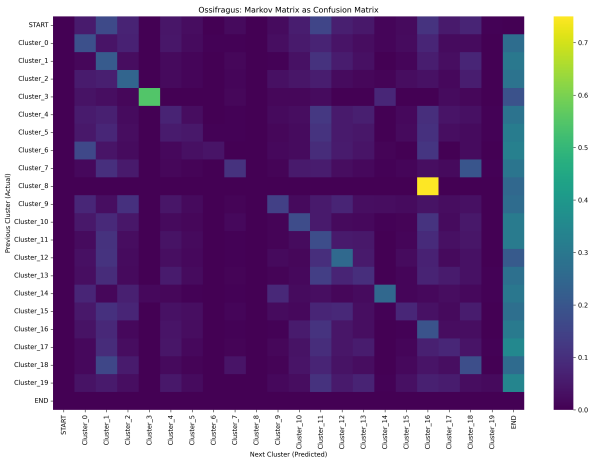

Fig. 5: Visualized Markov Matrix Across All Species.

4.4. Sequence Results

Our cross-species ANOVA test revealed all species' ngram model perplexities differed statistically significantly ($p < 0.001$) across all n from 1-4, with strong effect sizes ($F > 1{,}000$) for all n. This, combined with the previous result showing that clusters and species were distinguished by different features may indicate that variation between species may be less with in fundamental acoustic features, and more in call type usage.

Further, for all *Corvus* species post-hoc pairwise comparisons between N = 1-4 revealed highly significant differences ($p < 0.001$) across all orders and tests, confirming that increasing model order consistently alters the representation of vocal sequence structure.

Table 2: Perplexity values for N-gram orders across species.

Species	1-gram	2-gram	3-gram	4-gram
American Crow	12.79	10.46	10.86	14.13
Common Raven	12.45	10.42	19.78	273.39
Fish Crow	11.57	9.89	17.63	165.52
Hooded Crow	11.04	17.90	385.59	3917.27
Carrion Crow	10.56	10.37	40.72	172.54

Across species, unigram and bigram perplexities are low ($\sim$10-13 for unigrams, $\sim$10 for bigrams), indicating highly stereotyped single and double call sequences. This aligns with corvid communication where basic call types are often observed and easily recognizable across individuals. Trigram perplexities show greater variability, especially in the Hooded and Carrion Crows, which likely reflects limited data. Notably, the American Crow is the only species with trigram and 4-gram perplexities comparable to bigrams, suggesting more fixed phrases reflecting the American crow's high sociality, socially learned vocalizations [2], and greater data.

The first order Markov transition matrix showed a faint diagonal trend, indicating consistent call repetition that corvids also exhibit consistent repetition in line with other species [30]. Species-specific matrices of significant transitions revealed similar diagonals, particularly for the 3 species with the most data. Certain columns in the transition matrices reflect clusters with more data. For most call types, the next call was either itself or the end node, while the start node transitioned into only a few clusters. This suggests any call type can end a sequence, but not necessarily start one, which may partly reflect unbalanced cluster sizes.

5. CONCLUSION

Corvid vocalizations exhibit substantial intra-species variation, with consistent sequence repetition supporting the self-repetition hypothesis and the idea that corvids have intricate communication systems. Bigram structure best captures sequential organization, with some evidence for higher order complexity for the American crow. Future work will test more complex, hierarchical and hidden Markov models. Interpretation remains constrained by noise, unbalanced datasets, and lack of individual crow ID, highlighting the need for more balanced and comprehensive datasets to fully assess communicative complexity in corvids.

6. REFERENCES

[1] K. A. Jønsson, P.-H. Fabre, and M. Irestedt, "Brains, tools, innovation and biogeography in crows and ravens," *BMC Evol. Biol.*, vol. 12, no. 1, pp. 72, 2012.

[2] C. Wascher and S. Reynolds, "Vocal communication in corvids: a systematic review," *Anim. Behav.*, vol. 221, pp. 123073, 2025.

[3] L. Peckre, P. M. Kappeler, and C. Fichtel, "Clarifying and expanding the social complexity hypothesis for

communicative complexity," *Behav. Ecol. Sociobiol.*, vol. 73, no. 11, 2019.

[4] T. N. Suzuki, "Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls," *Anim. Behav.*, vol. 87, pp. 59–65, 2014.

[5] S. Engesser, A. R. Ridley, and S. W. Townsend, "Meaningful call combinations and compositional processing in the southern pied babbler," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 21, pp. 5976–5981, 2016.

[6] K. Katahira, K. Suzuki, K. Okanoya, and M. Okada, "Complex sequencing rules of birdsong can be explained by simple hidden markov processes," *PLOS ONE*, vol. 6, no. 9, pp. 1–9, 09 2011.

[7] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.

[8] A. Kershenbaum et al., "Animal vocal sequences: not the markov chains we thought they were," *Proc. R. Soc. B Biol. Sci.*, vol. 281, no. 1775, pp. 20141370, 2014.

[9] E. Scholes III, "Macaulay library audio and video collection," Cornell Lab of Ornithology. Occurrence dataset, 2015, Accessed via GBIF.org on 2025-09-15.

[10] A. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. Plumbley, and W. Wang, "Separate anything you describe," *IEEE Trans. Audio Speech Lang. Process.*, vol. 33, pp. 458–471, 2025.

[11] M. Miron, S. Keen, J. Liu, B. Hoffman, M. Hagiwara, O. Pietquin, F. Effenberger, and M. Cusimano, "Biodenoising: animal vocalization denoising without access to clean data," 2024.

[12] T. Sainburg, "timsainb/noisereduce: v1.0," June 2019.

[13] C. Heij and W. Verboom, "Bird vocalizations: Common raven (corvus corax)," Juno Bioacoustics, 2022, [Online]. Available: `https://www.researchgate.net/publication/361925309_Bird_vocalizations_Common_raven_Corvus_corax`.

[14] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[15] E. A. Mates, R. R. Tarter, J. C. Ha, A. B. Clark, and K. J. McGowan, "Acoustic profiling in a complexly social species, the american crow: caws encode information on caller sex, identity and behavioural context," *Bioacoustics*, vol. 24, no. 1, pp. 63–80, 2015.

[16] K. K. Jensen, O. N. Larsen, and K. Attenborough, "Measurements and predictions of hooded crow (corvus corone cornix) call propagation over open field habitats," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 507, 2008.

[17] D. A. Liao, K. F. Brecht, L. Veit, and A. Nieder, "Crows "count" the number of self-generated vocalizations," *Science*, vol. 384, no. 6698, pp. 874–877, 2024.

[18] K. F. Brecht, S. R. Hage, N. Gavrilov, and A. Nieder, "Volitional control of vocalizations in corvid songbirds," *PLOS Biol.*, vol. 17, no. 8, pp. e3000375, 2019.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. R. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] H. Hotelling, "The generalization of student's ratio," *Ann. Math. Stat.*, vol. 2, no. 3, pp. 360–378, 1931.

[21] R. A. Fisher, "Statistical methods for research workers," *Oliver and Boyd*, 1925.

[22] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[24] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

[25] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer, 3rd edition, 2005.

[26] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016.

[27] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

[28] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

[29] D. K. Oller and U. Griebel, "Functionally flexible signaling and the origin of language," *Front. Psychol.*, vol. 11, 2021.

[30] R. J. H. Payne and M. Pagel, "Why do animals repeat displays?," *Anim. Behav.*, vol. 54, no. 1, pp. 109–119, 1997.