Voice print analysis model for PAM without deep learning

Mitsuo Iwayanagi[1], Akiko Urago[1]

Abstract

We developed a rule-based voice print analysis AI model (here after Rule-based model) for PAM that defines explicit identification species rules based on visually recognizable shapes and structures of voiceprints. Different from Machine-leaning AI model, Rule-based model does not require a large amount of training data, specialized expertise, and substantial computational resources. Using a small number of publicly available audio samples, Rule-based model was constructed and tested on the vocalizations of the Grey-faced Buzzard, achieving a recall rate of 80%. Even if the accuracy is lower than Machine-learning model, Rule-based model may serve as a viable alternative in environments where the use of machine learning AI model is not feasible due to budgetary or technical constraints.

1 Introduction

In recent years, many AI-based classification models for Passive Acoustic Monitoring (PAM) have been made (Appendix A). But these classification models cannot be applied easily in the field. One of the reasons is cost and time for development models. Developing machine learning models requires a large amount of training data for each species. It imposes a significant burden of cost, labor, and expertise. It is a major barrier to developing AI-based classification model. Another reason is low transfer ability. Most classification models are evaluated under the conditions similar to where the models were built. Although some general-purpose models such as BirdNET (Kahl et al., 2021) and the CNN model by Marchal et al. (2021) have demonstrated high accuracy across regions, these models require thousands of labeled samples, advanced deep learning techniques, and close collaboration with expert researchers. Then it is not easy to create an AI-based classification model under the limited budget and data.

In light of these challenges, we explored alternatives to machine learning approaches. In the beginning, we examined some methods such as acoustic pattern matching, acoustic indices, and time-series or signal analysis. But we encountered, these approaches tend to be vulnerable to changes in recording equipment and background noise, and may not be reliable for stable detection under varied field conditions. As a potentially more robust option, we focused on the visual patterns of spectrograms that are recognizable to humans and defined explicit detection rules based on the shapes and structures of voiceprints. As a first trial, the Grey-faced Buzzard (*Butastur indicus*) was selected as the target species due to the visual clarity of its arch-shaped voiceprint structure.

2 Basic Structure of the Rule-based model

The Rule-based model is designed to reflect how humans visually identify voiceprint patterns. Instead of analyzing the raw audio waveform directly, the Rule-based model first converts the voiceprint into a spectrogram(Figure 2-1), extracts the visual patterns(Figure 2-2), measures the visual patterns, evaluates measured figures, and identifies the target species.

Identification is done by evaluating the measured figures referring the evaluation criteria. The evaluation criteria are set manually for each species in advance. Rather than generic optimization algorithms, Rule-based model relies on a human observer's visual identification and uses those rules as evaluation criteria. Because of that Rule-based model enables working with a small number of samples.

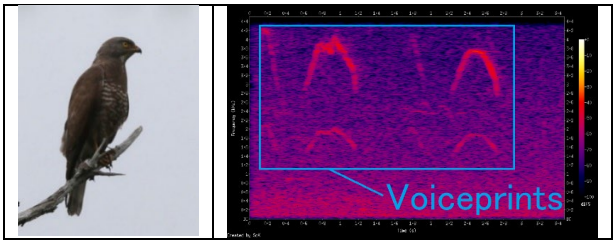

Figure 2-1: Converted spectrogram of Grey-faced Buzzard's voiceprint

[1] Raven Ltd.

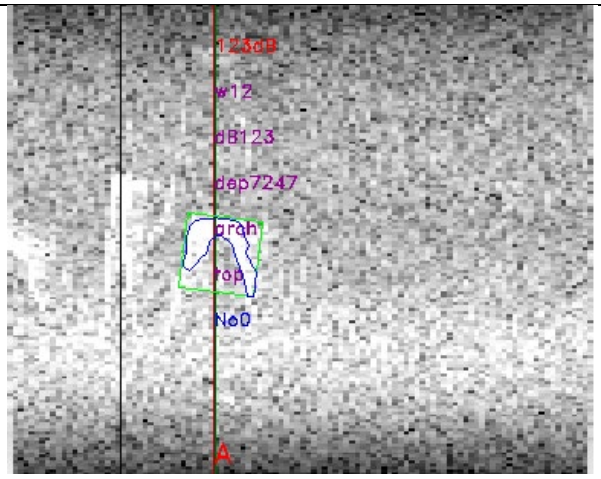

Figure 2-2: Extracted pattern of the spectrogram

3 Mechanism of Rule-based model

3.1 Features used for evaluation criteria

We selected five core features(Table 3-1) and one supplement feature for evaluation criteria focusing on stability and visual clarity. Five core features are selected based on geometric analysis techniques that could explicitly quantify such features, including skeleton analysis and convexity defect detection. As a supplement detection, sound pressure prominence is used. Evaluation criteria were designed to support structural classification (e.g., arch or trident forms), noise filtering (via height and width), and distortion evaluation (via bottleneck structure).

Acoustic indices are excluded from the core features. Because most acoustic indices were highly sensitive to recording devices and background noise. For example calculating the spectral centroid tended to fluctuate due to the influence of subtle noise.

Shape-matching techniques such as ORB matching and template matching were also excluded. Because they failed to capture fine-grained differences, such as taper angle, indentation depth, or asymmetry.

Table 3-1: Core features for evaluation criteria

Measurement item	Purpose
Maximum frequency (main component)	Identification of specific frequency bands
Depth and angle of convexity defects	Evaluation of uneven structure
Number of skeleton branches	Voiceprint structure classification (arch/trident/other)
Bottleneck structure	Voice distortion detection and quality evaluation
Voiceprint height and width	Exclusion of noise and abnormal shapes

3.2 Visual pattern extraction method

We adopted a fixed-length segments method to cut voiceprint. In terms of Grey-faced Buzzard the segment was set to 3 seconds in length with a 1-second overlap, following Maegawa et al. (2022), to cover a typical call in a single segment.

We used Python for audio processing by combining modules listed in Appendix B. First, a bandpass filter was applied to isolate the frequency range relevant to the target species, followed by Short-Time Fourier Transform (STFT) to generate spectrograms. The resulting spectrograms were binarized (Figure 3-1), and the main outlines of visual pattern were extracted.

From these extracted visual patterns, five core features and one supplement feature were measured for evaluation.

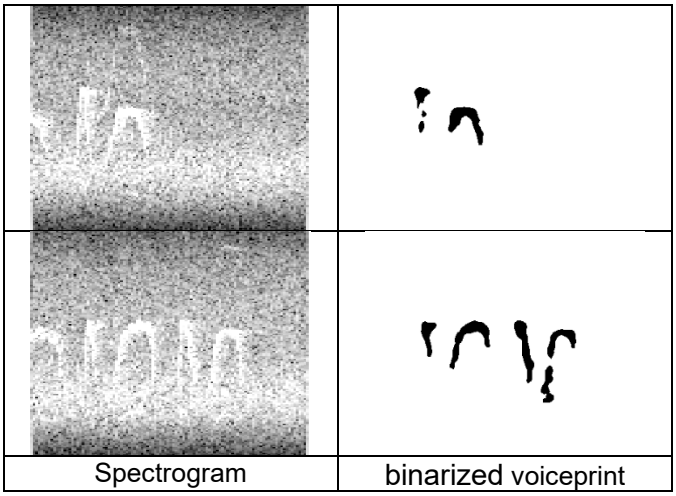

Figure 3-1: Extracted visual pattern

3.3 Evaluation Method

Rule-based approach is used for evaluation the measured figures of extracted visual pattern. Measured figures are evaluated based on evaluation criteria. These criteria are designed manually to fit the species, reducing false detections from similar sounds or noise. To manage variability caused by environmental noise or individual differences, the criteria are defined with a certain margin of tolerance. The criteria were initially derived from sample data and were refined through repeated visual inspection and manual adjustment. Each measured figure is evaluated as a “hit” only if all the figures are satisfied the respective evaluation criteria; if any criteria was not met, the measured figure was evaluated as a “non-hit.” By manual

adjustment of the criteria for each species, this approach enables stable performance even when only a small number of samples are available.

4. Accuracy Verification

4.1 Data Used for setting evaluation criteria

The data for setting the evaluation criteria of Grey-faced Buzzard covers multiple countries, various devices, and many surrounding environments. The number of data is 16. 13 of which were obtained from the open-access bird sound database Xeno-Canto, and the remaining 3 were independently recorded in Japan by the author. The Xeno-Canto data are recorded in China, Japan, Malaysia, Russian Federation, Taiwan, and Thailand.

4.2 Data Used for model Verification

For model verification, continuous audio recordings were collected using IC recorders installed at eight locations in Japan between May and June 2022. A total of 286 hours of audio data was used for the analysis. These recordings included a wide variety of natural and artificial background noise, such as wind, rivers, vehicles, and insects. As such, the verification data were intended to reflect real-world field conditions and were used to verify the model's noise tolerance and species detection performance.

The model performed detection using a sliding window of 3 seconds per segment. Then the evaluation results were aggregated into 1-minute units and identified species based on the number of hits and recall accuracy.

4.3 Verification Results

Identification results showed that the calls of the target species, the Grey-faced Buzzard, appeared for a total of 41 minutes within the 286 hours of recorded audio. The model successfully detected 33 of these minutes, resulting in a recall rate of 80% (Table 4-2). Some false detections were observed, caused by calls from other species or strong background noise. Given the sporadic nature of target calls amid diverse background sounds, the model demonstrates practical usability. These findings indicate that the model is capable of detecting the species' calls from field recordings, even under varying environmental conditions and with a limited number of reference samples.

Table 4-2: Detection Results Summary

Total (Correct answers)	Detected (TP)	Not detected (FN)	False positive (FP)	Precision (%)	Recall (%)
41	33	8	1140	2.81	80.49

5 Considerations

The Grey-faced Buzzard, selected as the target species in this study, exhibits a distinctive arch-shaped voiceprint that is visually recognizable and relatively easy for humans to identify, making it well-suited for rule-based modeling. In contrast, small passerines with short and ambiguous calls, or species that share similar acoustic patterns, may be more difficult to distinguish due to the lack of clear structural features. Additionally, the method is not suitable for situations involving overlapping calls from multiple individuals (chorusing), where voiceprints may become severely distorted.

Most of the sample data used in this study were recorded outside Japan, using diverse equipment and under varying environmental conditions. Nevertheless, the model demonstrated high detection accuracy when applied to domestic recordings, suggesting that for specific vocal phrases produced by a given species, it may be possible to apply a shared detection rule across different regions and recording environments.

It should be noted that the evaluation in this study was based on a limited set of test data available at the time of analysis. Future work will involve expanding the target species and conducting more comprehensive performance assessments. Currently, new audio data are being collected, and improvements to the model are underway to support formal public release.

6 Conclusion

This study developed a Rule-based model designed to detect bird vocalizations based on the distinctive voiceprint structure of a target species, and evaluated its effectiveness. Because the model operates using explicitly defined rules based on visual voiceprint patterns, it can be constructed with limited data and showed stable performance. In particular, this approach may serve as an effective alternative in contexts where it is difficult to secure the extensive sample datasets or high-level computational resources required for deep learning models.

Although most of the sample data used to design the model were recorded outside Japan, the model maintained high detection accuracy when applied to domestic recordings, suggesting that for specific vocal phrases, detection rules may be transferable across regions and recording conditions. These findings indicate that Rule-based model can function effectively even under constraints of limited data and budget, and suggest its potential as a supportive tool for bird community surveys using Passive Acoustic Monitoring (PAM).

References

1. Takumi Sato, Yuko Maegawa, Tomohiro Haga, & Akira Sasaki (2023). Development of a bird species identification system using deep learning and bird calls: Future prospects. Bird Research, 19, A41–A50. https://doi.org/10.11252/birdresearch.19.A41 [in Japanese].
2. Masatake Yamakawa, Fumiaki Takeuchi, & Kazuyoshi Yoshii (2021). Improving the Accuracy of Automatic Species Identification Using Calls for the Advancement and Efficiency of Raptor Surveys: Enhancing Accuracy Through Neural Networks and Noise Reduction. Journal of the Japan Society of Civil Engineers, Series G (Environmental Research), 77(6), II_73–II_79. https://doi.org/10.2208/jscejer.77.6_II_73, [in Japanese].
3. Yusuke Ueno & Masao Kurihara (2016). Experimental simple judgment between existence and non-existence and evaluation of breeding stage of goshawk using sound analysis. Journal of Japan Society of Civil Engineers, Ser. G (Environmental Research), 72(6), II_341–II_349. [in Japanese].
4. Yuko Maegawa, Shun Takagi, Kazuki Komori, Yuki Aoki, & Masahiko Nakamura (2022). A study on bird call monitoring methods using AI technology: A case study of the Japanese Sparrowhawk. Bird Research, 18, A71–A86. https://doi.org/10.11252/birdresearch.18.A71, [in Japanese].
5. Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega-Bermudez, P., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. PeerJ, 1, e103. https://doi.org/10.7717/peerj.103
6. Kahl, S., Stöter, F. R., Klinck, H., et al. (2021). BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236
7. Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M., & Bayne, E. M.（2022）. OpenSoundscape: An open-source Python package for bioacoustic analysis. Methods in Ecology and Evolution, 13(3), 634–640. https://doi.org/10.1111/2041-210X.13772
8. Marchal, J., & Fabianek, F. (2021). A comparison of four software programs for detecting Bicknell's and Gray-cheeked Thrush vocalizations. Avian Conservation and Ecology, 16(1), 13. https://doi.org/10.5751/ACE-01908-160113
9. Fujitsu (2023). QSAS-Bird: AI system for identifying species based on the calls of Blakiston's Fish Owl. Retrieved from https://www.fujitsu.com/jp/solutions/business-technology/tc/special/qsas-bird/ [in Japanese].

Appendix A. Overview of Prior Studies Using AI for Bird Vocalization Analysis

System	Region	Target species	AI method	Training samples	Cross-regional performance	Citation
BirdNET	Germany, USA	Multiple	Machine learning (CNN)	22,960	Validated across multiple regions	Kahl et al. 2021
ARBIMON	Puerto Rico	Multiple	Acoustic detector / Clustering	No training	localized system	Aide et al. 2013
OpenSoundscape	USA	Multiple	Machine learning (Conventional)	2,318	Limited, no transferability tested	Knight et al. 2022
QSAS-Bird	Japan	Single	Machine learning (CNN)	private	Unknown	Fujitsu Web
Goshawk model	Japan	Single	Machine learning (Conventional)	1,500	Performance declined	Ueno & Kurihara 2016
Sashiba CNN	Japan	Single	Machine learning (CNN)	100	Performance declined	Maegawa et al. 2022
CallSeeker	Canada	Multiple	Acoustic template matching / Clustering	No training	Unknown	Marchal et al. 2021
Song Scope	Canada	Multiple	Machine learning (classifier)	6,755	Unknown	Marchal et al. 2021
Kaleidoscope Pro	Canada	Multiple	Machine learning (Clustering + classifier)	6,755	Unknown	Marchal et al. 2021
CNN (Marchal)	Canada	Multiple	Machine learning (CNN)	6,755	Validated across multiple regions	Marchal et al. 2021

Appendix B. Supplementary Acoustic Features

Library Name	Version	Primary Purpose	License
librosa	0.9.2	Extraction of audio features	BSD
opencv-python	4.5.4	Image processing and contour extraction	Apache 2.0
pydub	0.25.1	Audio format conversion (e.g., MP3 to WAV)	MIT
soundfile	0.12.1	Reading and writing audio files	BSD
matplotlib	3.4.3	Visualization (e.g., spectrograms, plots)	PSF
scipy	1.7.1	Numerical processing, filtering, peak analysis	BSD