Voice print analysis model for PAM without deep learning

Mitsuo Iwayanagi[1], Akiko Urago[2]

Abstract

The effectiveness of voice monitoring using a combination of recording devices such as IC recorders and automatic detection of target voice by AI has been recognized in habitat condition surveys of raptors and other rare bird species. Currently, most of the knowledge on automatic detection of target sounds by AI in Japan is based on machine learning. However, there have been reported issues in versatility, such as the difficulty of preparing teacher data due to the lack of large-scale open data on birds of prey and other species in Japan, and the fact that some teacher data may cause the correctness rate of models to drop in different recording environments. Therefore, it was considered a hurdle for small-scale organizations to introduce speech detection models using machine learning. Therefore, a speech detection model that could be created with a small sample was developed on a trial basis. In addition, focusing on the percentage of correct responses, we made efforts to avoid undetected speech as much as possible, while allowing for false positives that can be scrutinized by screening. The main method was to extract one characteristic contour of the voiceprint from a typical call of the target species, compare the center of gravity position, area, aspect ratio, and shape of the contour of each voiceprint in the recorded voice data and perform feature point matching, and based on these results, the target voiceprint was ranked into three levels and detected.

1 Introduction

In recent years, Passive Acoustic Monitoring (PAM) technology, which identifies species based on their vocalizations, has been advancing. Research on sound detection using machine learning and deep learning is also progressing, but there are several issues that prevent PAM from being used effectively in various field settings. One issue is that there is insufficient open data on bird sounds in Japan to build machine learning (including deep learning) models, necessitating the collection and preparation of audio data in-house, which is a time-consuming and labor-intensive process in practice. Another issue arises when machine learning models are applied in locations different from their training environments, as variations in recording conditions and background noise can reduce identification accuracy, increasing the likelihood of missed detections. (Sato et al., 2023; Yamakawa et al., 2021; Maegawa et al., 2022; Maegawa et al., 2022). Therefore, in this study, we developed an automatic voice detection method using voiceprint image processing.

2 Features of Rule-Based AI

There are multiple methods for identifying species names from voice data, one of which is machine learning-based, and the other is rule-based. The machine learning-based method involves training AI using multiple voice data samples of a single species, which we refer to as machine learning AI in this paper. The other method involves codifying the conditions for extracting audio data for each species into rules and having the AI search for items that match the rules. This method is referred to as rule-based AI in this paper.

2.1 Mechanism of Species Identification Using Rule-Based AI

Rule-based AI for species identification involves converting the recorded sound of the target organism into a voiceprint image, extracting shape and acoustic features, and then using predefined rules (threshold values) to determine the species name. The image on the left in Figure 2-1 is a voiceprint image of a Japanese bush warbler, and the image on the right is a voiceprint image of its song, with the white U-shaped portion representing the song. It is also possible for humans to identify species by looking at voiceprint images of sounds sampled by humans and searching for this shape. Rule-based AI essentially automates and accelerates the human process of identifying

[1] Raven Ltd.

[2] Raven Ltd.

species from these images.

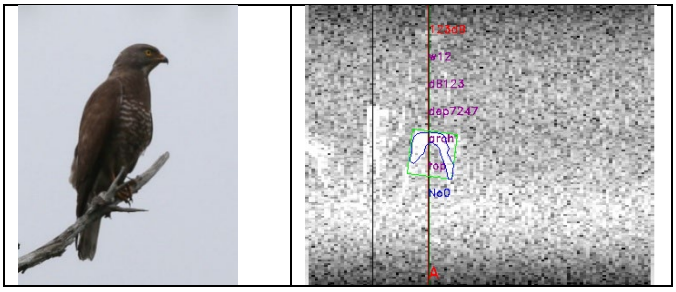

Figure 2-1: Image of voice detection

2.2 Differences between machine learning AI and rule-based AI

Machine learning AI enables the system to directly distinguish voiceprints through learned patterns. Machine learning AI can learn various types of audio data, enabling it to correctly distinguish not only bird songs but also ground noises. However, machine learning AI has several drawbacks, including the need for a large amount of sample data for construction, reduced recognition rates when background noise is present, and the fact that the reasons for misclassification are often incomprehensible to humans (see Table 2-1). An existing system that uses machine learning AI is Kaleidoscope. Kaleidoscope excels at detecting specific types of clear bird calls and offers high operability as a desktop application (Wildlife Acoustics, 2022).

Rule-based AI is a mechanism in which humans first specify the conditions for the contours of voiceprints from images of characteristic voiceprints, and the AI extracts voiceprints that match the conditions. Rule-based AI can be configured with a small number of sample data and is less affected by background noise in the case of bird songs. Additionally, it can extract sounds even with poor recording quality or changes in sound quality. However, as of now, it cannot distinguish between ground calls or synchronized calls by a flock (see Table 2-1).

Table 2-1: Differences between machine learning and rule-based

Comparison items	Machine learning AI	Rule-based AI
Data for AI	Voiceprint data	Spectrogram (voiceprint image)
Classification method	Machine learning, templates	Rule-based
Number of samples for model construction	Several hundred to several thousand or more per type	1 to several dozen per type
Use of open source	Basically not possible	Possible
Target sounds (types)	All sounds	Only distinctive sounds such as chirping
Target sounds (intensity)	All sounds	Only clear sounds
Influence of background noise	Depends on the training data	Not easily affected
Revision of judgment criteria	Difficult	Easy
PC performance	Medium to high	Low to medium
Operational style	Desktop application	Script-based (Python)
Reusability of models in other regions	A bit difficult	Easy
Advantages	・Can distinguish even non-distinctive bird calls ・High accuracy (ability to identify correctly)	・Can be used with even a single sample of distinctive bird calls ・Few omissions ・No restrictions on recording quality
Disadvantages	• Requires a large amount of training data • No publicly available general-purpose database • Human inability to understand classification rules • Increased risk of omissions in surveys of areas with no existing survey data	• Inability to classify species with non-distinctive calls or ground calls • Human-based supplementary surveys are required for bird species inventory

2.3 Strengths of Rule-Based AI

Rule-based AI has the following strengths.

No false negatives, and accuracy can be improved as the investigation progresses

One of the characteristics of rule-based AI is that it has few false negatives. In designing rule-based AI, we adopted a design policy of “avoiding false negatives and correcting false positives.” In the early stages of investigation, we allow the AI to detect audio on the safe side when determining species, so there are more misclassifications than with machine learning AI. However, rule-based AI can detect audio that machine learning AI misses. Additionally,

by having humans check the classification results and fine-tune the rules, the system gradually improves to better suit the specific context. The ability to improve accuracy as the investigation progresses is a key strength of rule-based AI.

Clear classification criteria and human-adjustable classification conditions

Another strength of rule-based AI is that judgment conditions can be adjusted intuitively through scoring. While machine learning AI also adjusts subtle judgment criteria through learning, the adjustment points become a black box, making human intervention impossible. However, rule-based AI allows users to freely manipulate the criteria for measurement, making it accessible even to humans with limited expertise in biological surveys. Additionally, the ability to adjust the system as surveys progress for species whose calls vary by area is another strength of rule-based AI.

Can be implemented with a small number of samples (as few as one)

Rule-based AI is also advantageous because it can be implemented even with a very small number of samples. Generally, machine learning AI requires a large amount of training data, specialized knowledge, computational resources, and repeated model tuning, but practical implementation is difficult due to the limited availability of publicly available sample data. Therefore, the ability to operate with as few as one sample is a key strength of rule-based AI.

3 How to build a rule-based AI model

Rule-based AI requires the construction of a model for each species to be classified. The steps for constructing rule-based AI are (1) preparation of audio samples, (2) creation of spectrograms, (3) voiceprint extraction, (4) consideration of measurement items, and (5) setting of classification criteria. By repeating this process for each species to be investigated, rule-based AI can be constructed for multiple species. This process takes approximately one to two hours per species. The details of each step are described below.

3.1 Preparation of audio samples

The first step is to prepare audio samples that are characteristic of the target species. The audio source can be open-source audio samples or audio samples recorded by the user. When using open-source audio sources, prioritize “whether the voiceprint is clearly visible to the human eye” over recording quality. In this case, we prepared the song of the Japanese bush warbler as an audio sample. Although it is possible to build a rule-based model with as few as one audio sample, the more samples you have, the easier it is to evaluate the stability and versatility of the recognition accuracy. First, collect audio samples that are suitable (i.e., with clear voiceprints and little distortion) and, if possible, also collect unsuitable samples for comparison. In actual field tests, sufficient detection accuracy was achieved even with about 10 audio samples containing a mix of clear and unclear samples.

3.2 Creating spectrograms

Once the audio samples have been collected, the next step is to convert them into images. This involves applying a short-time Fourier transform (STFT) to the audio samples to generate spectrograms (visualizations of time, frequency, and sound pressure).

3.3 Voiceprint extraction

After the spectrogram is generated, the voiceprint is extracted. Since the spectrogram contains various audio information, it is converted into a black-and-white image with clear contours to make it easier for AI to recognize, and the voiceprint is extracted. Figure 3-1 shows the images before and after voiceprint extraction, with the left image being the extracted spectrogram and the right image being the extracted voiceprint image.

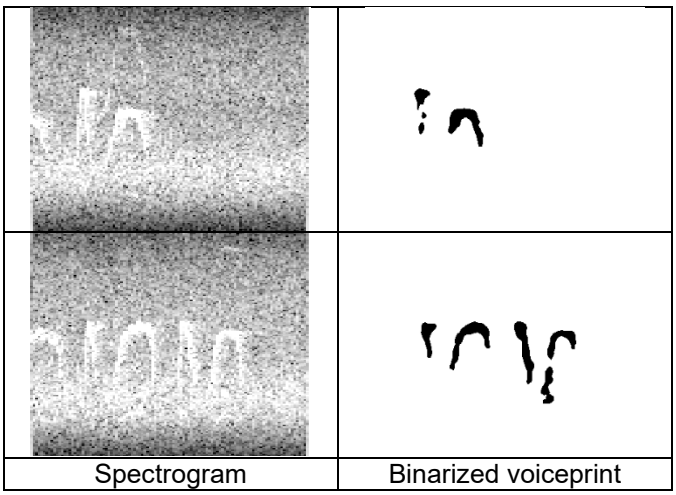

Figure 3-1: Spectrogram and binarized voiceprint of a Japanese bush warbler

3.4 Consideration of measurement items

Consider measurement items for the digitized voiceprints. The items adopted for measurement are sound pressure prominence, sound pressure peak occurrence position, maximum frequency (main component), frequency change rate (before and after), convexity defect depth and angle, skeleton branch number, and bottleneck structure. The purpose of extracting each item is shown in Table 2-1. However, these measurement items were selected based on the voiceprints of the Japanese bush warbler, so if they are applied to other species or voiceprints with different shapes, the rules or measurement items may need to be revised.

When selecting measurement items, other candidates included Spectral Centroid (center of gravity), Spectral Bandwidth (bandwidth), Spectral Flatness (flatness), and Inharmonicity Score (dissonance score). However, these were excluded due to significant variability in background noise (noise), recording equipment, and recording distance. The time difference between the sound pressure peak within the vocalization and the peaks before and after it, known as pitch (periodicity), was also considered as a measurement item. However, when insect sounds overlapped in the background, even though the Sashi-ba vocalization itself could be extracted, the pitch component was dragged by the insect sounds, resulting in the detection of incorrect periods. Therefore, pitch was not adopted as a measurement item in this study.

Table 3-2: Binary vocalization measurement items and their purposes

Measurement items	Purpose
Sound pressure prominence	Confirmation of clarity of voice
Sound pressure peak occurrence position	Stability of sound source
Maximum frequency (main component)	Identification of specific frequency bands
Frequency change rate (before and after)	Tendency of sound pattern changes
Depth and angle of convexity defects	Evaluation of uneven structure
Number of skeleton branches	Voiceprint structure classification (arch or trident or more)
Bottleneck structure	Voice distortion detection and quality evaluation
Voiceprint height and width	Exclusion of noise and abnormal shapes

3.5 Setting the evaluation criteria

After measuring the sample voiceprints, we set the evaluation criteria. First, we created scatter plots of the measurement items and analyzed the patterns observed in clear voiceprints. Next, we designed evaluation criteria for each measurement item, scored them, and classified the results into detection ranks of “high,” “medium,” and “low (requires verification)” based on the total score and the results of some important items.

The evaluation scores for each measurement item were set as follows:

- “Maximum frequency (main component)”, “Depth and angle of convexity defects”, “Number of skeleton branches”, “Bottleneck structure”
  → If any of these is rated -1, it is classified as “Low”
- Other features (sound pressure prominence, frequency change rate, voiceprint height and width, etc.)
  → These are scored as 0 or +1 and added to the total score, with the total value determining whether it is classified as “Medium” or “High”

Table 3-3: Detection rank

Detection rank	Note
High	High similarity to typical voiceprints of the target
Medium	Similarity to the target voiceprint exists
Low	Little similarity to the target voiceprint (treated as undetected)
Low (requires verification)	Little similarity to the target voiceprint, but strong voice and possible distortion, so verification is required

This weighting design takes into account that bird vocalizations are easily affected by movement and posture changes during recording, and that the main components of the voiceprint may not be extracted completely due to binarization processing. Even with clear audio, there may be discrepancies in the judgment of fine-tuning items, so only the four elements that must be passed are emphasized, and the rest are treated as supplementary elements.

Furthermore, among the vocalizations classified as “low,” those with high sound pressure prominence and clear vocalization were classified as “low (requires verification)” in a separate category, considering the possibility of misclassification due to

distortion of the main components. Since “low” has a high possibility of misdetection and a low priority as a screening target, it was treated as undetected in the evaluation.

4 How to operate rule-based AI

When operating the constructed rules, the following steps are taken: (1) collection of audio data, (2) audio processing, (3) voiceprint extraction, and (4) voiceprint identification. The process from audio processing to voiceprint identification is automated by a computer and takes about one second for one hour of recorded data.

4.1 Collection of audio data

The first task is to collect audio data by operating recording devices in the target area. Specialized devices designed for biological sounds such as bird calls or IC recorders are preferable, but smartphones equipped with external microphones (including low-cost models) can also be used as an alternative.

The recording format should be MP3 or WAV. WAV format is superior in terms of voiceprint reproduction, but tends to be more expensive in terms of recording equipment. On the other hand, MP3 (44.1 kHz / 128 kbps) is theoretically capable of extracting the main components of voiceprints for all bird species, making it a practical choice depending on the budget.

For audio data, clear audio with minimal background noise is preferred. However, “clear” does not necessarily mean “loud.” If the sound pressure is too high or if the recording device is positioned above the recording area and the bird is circling while vocalizing, the voiceprint may be distorted, so caution is required.

Regarding background noise, uniform sounds such as wind, water, or rustling leaves have relatively minor effects. However, sounds with similar frequencies (pitch) from other bird species or insects, or sounds with a wide pitch range such as metal impacts, may reduce detection accuracy.

Table 4-1: Sample suitability for audio data

	Suitable	Possible	Not suitable
Sound source quality	WAV	MP3 44.1kHz/128kbps or higher	MP3 44.1kHz/128kbps or lower
Recording equipment	Dedicated device, IC recorder	Smartphone + external microphone	Smartphone only
Position of sound source and recording device	Stationary	Moves but distance and direction remain constant	Moves around recording device
Volume	Medium	High	Low, maximum
Background noise	Low	Medium	Strong
Birds, insects, etc. with the same pitch	Weak	Medium	Strong

4.2 Audio processing

The collected audio data is first divided into 3-second segments. A bandpass filter with a bandwidth of 2000–5000 Hz is applied to the prepared audio data, and then the audio data is divided into 3-second segments with 1-second overlap. This setting is intended to capture the typical vocalization time of the Japanese bush warbler and limit the frequency range (pitch) to suppress the influence of noise. Note that for species whose vocalizations exceed 3 seconds, we plan to introduce an evaluation system that combines multiple adjacent segments.

The segmented audio data is subjected to short-time Fourier transform (STFT) to generate spectrograms (visualization of time, frequency, and sound pressure).

Noise gates for noise removal are not applied uniformly across the entire 3-second interval but are used for localized and limited processing, such as calculating the frequency change rate of individual vocal patterns.

4.3 Phoneme extraction

The audio data divided into 3-second segments was binarized by performing a stepwise threshold search process (hereinafter referred to as threshold scanning) to extract phonemes. Specifically, this involves repeatedly performing binarization and shape extraction on each image while gradually lowering the threshold for the brightness of the spectrogram. This structure detects high-intensity areas (strong sound pressure areas) first, and when the extracted shape roughly matches the typical voiceprint area set in advance, the result is adopted as the “optimal extraction result,” and subsequent scanning is terminated.

This method tends to produce variations in the size

of extracted shapes depending on sound pressure and voiceprint size, with larger voiceprints being slightly underestimated and smaller voiceprints being slightly overestimated. While such variations have a certain impact on detection accuracy, they are corrected in the subsequent scoring process by comparing shape features and frequency structures, resulting in stable final judgment accuracy.

Note that in this method, the voiceprint with the strongest sound pressure is prioritized, so in cases where multiple voiceprints exist within the same image, the voice of the Japanese bush warbler may be buried by strong other sound sources (such as insect sounds or wind) and not detected.

4.4 Voiceprint Identification

The binary voiceprint data is measured against seven predefined criteria, and the species name is identified according to the classification criteria.

5 Field Verification Results

As a result of analyzing 41 target sounds contained in the sample data (total recording time: 286 hours) using rule-based AI, 33 were detected. The identification accuracy (precision) was 2.81%, and the recall rate was 80.49%.

Although the identification accuracy was low, the recall rate was high, and detection with few missed target sounds was achieved. False detections were limited to 1,140 cases, and compared to the approximately 17,000 cases (286 hours divided into 1-minute intervals) that would have been detected by workers alone, the number of cases to be identified was significantly reduced, suggesting a significant reduction in the workload.

Note that while the AI performs detection in 3-second intervals, manually screening all detection intervals every 3 seconds for a large dataset is impractical. Therefore, evaluations were aggregated at 1-minute intervals. If multiple voiceprints were detected within a 1-minute interval, the voiceprint with the highest evaluation rank was selected as the representative, and the detection determination for that interval was based on that voiceprint.

Table 4-1: Detection Results Summary

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

6 Potential for expansion and practical application

Rule-based AI is more suitable for surveys aimed at identifying the habitats of specific species rather than for biodiversity surveys, due to limitations such as the time-consuming nature of rule construction and the current inability to respond to bird chirping or the collective calls of a flock. Additionally, since rules can be created using as little as one sample, this approach is effective for studying species with limited vocalization data. By fine-tuning a single rule, it can be applied across multiple locations, making it versatile for surveys conducted in various regions.

Furthermore, by utilizing cloud platforms such as the paid version of Google Colab, it is possible to analyze large amounts of data, making it usable even when high-performance PCs are not available locally.

Note that this method currently consists of a script-based configuration that runs on a Python environment and is not yet compatible with application formats equipped with a GUI. However, the processing flow has a simple and lightweight structure, making it easy to develop into an application or integrate into other systems in the future.

7 Considerations

7.1 Support for multiple species and improvement of accuracy

At this stage, the trial is limited to one species of Japanese bush warbler with a stable arch-shaped voiceprint, but it is considered possible to support other species with distinctive songs or calls in the future.

Furthermore, many cases of false detection have clear causes, and further accuracy improvements are considered possible by introducing supplementary rules or processing branches tailored to these causes.

7.2 Detection Issues and Future Improvement Possibilities

The following issues have been identified:

- Difficulty in extraction due to overlapping sound sources: In environments where multiple sound sources overlap, the target voiceprint may be buried by other strong sounds, making extraction difficult.

- Subjectivity of threshold and condition settings: Currently, thresholds and rules are set subjectively, resulting in issues with reproducibility.
- Unquantified relationship between detection and sound pressure: No clear criteria have been established regarding the relationship between sound pressure prominence and detection success rates.
- Unsupported calls with no distinctive features: No established method exists for distinguishing voiceprints with few distinctive features from noise.

When multiple types of sounds are mixed in a speech segment, or when the same type of sound is emitted by a group, the main component of the sound is often unclear, making it difficult to determine using the current detection rules. Although sound source separation technology has advanced in recent years, it is difficult to achieve sufficient separation performance in natural environments where unknown species' calls overlap, and there are limitations to its implementation in the field.

Therefore, introducing a mechanism to automatically identify sections with high acoustic complexity and divert them to separate processing systems (e.g., analysis using a different model or prioritization for human verification) may reduce overall false detection rates and improve processing efficiency.

In addition, we are considering optimizing the thresholds and scores for each measurement item through statistical methods to improve the reproducibility and objectivity of rule construction. On the other hand, since this method emphasizes a structure that reflects the knowledge of observers, the optimization results will be used only as reference values for design, and the final rules will be determined through human interpretation and verification.

Furthermore, if quantitative optimization is advanced through large-scale detection experiments and threshold-based evaluations, it may be possible to determine the detection limits that depend on the detectable vocalization distance and recording environment based on the relationship between sound pressure prominence and detection success rate.

For sounds such as the chirping of small birds, which have poor voiceprint dispersion and are very faint, it is difficult to quantitatively compare shapes and distinguish them from other similar sounds. Therefore, it is currently difficult to automatically detect small birds that migrate or overwinter using this method, and it is considered essential to supplement it with conventional methods such as visual surveys by humans in the field in order to grasp the occurrence status of such species.

8 Conclusion

In this study, we proposed a rule-based AI method for automatically detecting the vocalizations of gray-faced buzzard. We demonstrated that high-accuracy identification is possible under specific conditions through flexible binarization using threshold scanning, a combination of shape and acoustic features, and simple scoring. However, the evaluation in this study was conducted based on limited test data collected at the present time, and although the accuracy rate was high, the number of correct audio samples used for verification was not sufficient. We are currently collecting additional audio data and plan to aim for formal publication based on more comprehensive verification results using these data in the future.

9 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