original article to Primates

Analysis of individual identification and age-class classification of wild macaque vocalizations without pitch- and formant-based acoustic parameter measurements

Rentaro Kimpara[1*], Fumiya Kakuta[2], Hiroki Koda[3,4], Ikki Matsuda[1,5,6], and Goro Hanya[2]

[1] Wildlife Research Center, Kyoto University, Kyoto, Japan
[2] Center for Ecological Research, Kyoto University, Inuyama, Japan
[3] Graduate School of Arts and Sciences, The University of Tokyo, Japan
[4] Center of Evolutionary Cognitive Science, The University of Tokyo, Japan
[5] Chubu Institute for Advanced Studies, Chubu University, Kasugai, Japan
[6] Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Correspondence to RK (kimpara.rentaro.75c@st.kyoto-u.ac.jp)

Key words: wild macaques, vocalizations, vocal identification, age classification, mel spectrogram

Abstract

In recent years, deep learning has achieved high performance in bioacoustic classification tasks by leveraging automatically processed acoustic features for large datasets. However, few performance evaluations of automatically processed acoustic features have been conducted on small-scale data because deep learning requires large datasets. To test whether mel spectrograms (an automatically processed acoustic features) are effective for classifying relatively small acoustic data, we evaluated the performance of two classification machines (random forest and support vector machine) using mel-spectrograms of 651 coo calls of six wild female Japanese macaques on two tasks: 1) individual identification and 2) age-class classification between younger (<10 yrs) and the older animals (>20 yrs). For the individual identification task, the mean balanced accuracy was 0.81 for random forest and 0.82 for support vector machine. For the age-class classification task, the mean balanced accuracy was 0.91 for random forest and 0.93 for support vector machine. Considering that of all the calls were recorded in the wild, methods using automatically processed acoustic features, such as mel spectrogram, are effective in classifying small acoustic data for the individual identification task. The high performance in the age-class classification task might be due to the ability of mel spectrograms to capture the characteristics (e.g. harshness) of older individuals.

Introduction

To elucidate the vocal communication of animals, bioacoustic studies have implemented tasks to classify calls from acoustic features for such purposes as identifying individuals making vocalizations and classifying vocal repertoires. This work has been pursued for more than half a century (Arnaud et al., 2023; Fukushima et al., 2015; Marler, 1956; Marler and Hobbett, 1975; Mitani, 1986; Rendall et al., 1998; Winter et al., 1966). Historically, conventional acoustic analysis has measured numerous acoustic features like pitch and formant and has examined classification performance using dimensionality reduction and multivariate analysis (Ceugniet and Izumi, 2004; Oyakawa et al., 2007; Rendall et al., 1998). Although these conventional approaches are still in use, they have suffered from problems such as the arbitrary nature of selecting acoustic features and the lack of versatility due to the definition of species-specific acoustic features.

Recent machine learning, especially deep learning, can automatically process acoustic signals for classification tasks without explicitly measuring specific features, which solves the problems associated with the conventional method (Arnaud et al., 2023; Lakdari et al., 2024). Nevertheless, current deep-learning techniques have a methodological shortcoming in that they require a large amount of supervised data. This is an obstacle to building models because preparing a large amount of supervised data requires a huge amount of time or is nearly impossible due to difficulties such as collecting sufficient data in the field. Therefore, developing classification models for relatively small amounts of data without measuring specific acoustic features could alleviate the difficulty of model building and facilitate bioacoustic studies of many species. However, no such models have so far been built.

Acoustic features are relatively unstable due to age-related changes. Aging in animals is

associated with physical and cognitive changes (Hindle et al., 2009; Manrique and Call, 2015). In humans, aging affects acoustic features such as fundamental frequency and signal-to-noise ratio (Ferrand, 2002; Ryan and Burk, 1974; Stathopoulos et al., 2011), suggesting that the age of a human could be identified from the voice (Harnsberger et al., 2008; Huntley et al., 1987). Interestingly, some studies suggest that human behavior is influenced by age-related acoustic features such as pitch (Feinberg et al., 2005; Klofstad et al., 2015; Tigue et al., 2012). While fewer studies have explored age-related acoustic changes in non-human animals, some research suggests that such changes influence receiver behavior in a way similar to humans (Lemasson et al., 2010; Zipple et al., 2020). However, bioacoustics research has largely focused on early life stages, with limited attention to senescence.

Individual identification of vocalizations or classification of vocal repertoires, notably in group-living primates, has been attempted repeatedly over the past half-century (Arnaud et al., 2023; Fukushima et al., 2015; Marler and Hobbett, 1975; Mitani, 1986; Rendall et al., 1998). However, few studies have attempted to identify vocalizers or classify age classes in wild primates without measuring specific acoustic features such as pitch or resonance frequency. In this study, therefore, we address the tasks of vocalizer identification and age-class classification (younger: <10 yrs or older: 20 > yrs) in wild Japanese macaques (*Macaca fuscata yakui*) without measuring specific acoustic features.

Materials and methods

Recordings of vocalization

The Japanese macaques (*Macaca fuscata yakui*) of Yakushima Island in southern Japan (30.41N, 130.41E), have been studied intensively since the 1970s, and

multiple groups have been identified (Yamagiwa and Hill, 1998). We investigated one of the identified groups, named the Petit group, from October to December 2023. At that time, the Petit Group was composed of about 33 individuals, including 10 adult females (>=6 years), 5 or more adult males (>=6 years), 8 juvenile females (1–5 years), 7 juvenile males (1–5 years), and 3 infants. We selected 10 adult females as focal animals and collected vocalization data using the focal animal sampling method, where RK and FK followed the focal animal. One observation session lasted two hours, and we changed focal individuals during the day. We recorded vocalizations with annotations of the vocalizer using an audio recorder with two channels and 32-bit-float/48 kHz sampling rate formats (48 kHz), which was connected to a directional microphone (Sennheiser MKE600) for recording the animal's vocalizations and a lavalier microphone (COMICA CVM-V02O) for recording the observer's annotations, such as recording vocalizers. The total observation time was 112 hours over 20 days. We created classifiers for six adult females (Kapa, 9 yrs; Rine, 8 yrs; Sasa, 7 yrs; Sazae, estimated 21-22 yrs; Taiko, estimated 23-25 yrs, Taiwu, 9 yrs), for which we obtained sufficient high-quality vocal data.

Analysis

Data selection and preprocessing

We used the acoustic analysis software Praat (Boersma, 2024) for annotation to select data in the recordings for which we could reliably identify the vocalizer. After the annotation, we selected the calls that met all five of the following criteria for further analysis: 1) Little background noise, 2) coo call (see Nakagawa et al., 2010 for its common definition and function), 3) no overlap with other individuals, and 4) certainty about the vocalizer, and 5) less than 1 second. From 651 audio recordings, we finally

selected 116 calls for Kapa, 89 for Rine, 138 for Sasa, 122 for Sazae, 84 for Taiko, and 102 for Taiwu, and we used them in subsequent models for classification tasks.

Generation of Mel Spectrograms

First, we generated mel spectrograms of the vocal region with Python's audio processing module *librosa* (McFee et al., 2015). A mel spectrogram is a set of acoustic features derived through logarithmic scale filters designed using the frequency-sensitivity characteristics of human auditory perception, (Stevens et al., 1937) and it is one of the most widely used values in machine learning (Baowaly et al., 2024; Dossou and Gbenou, 2021; Walsh et al., 2023; Xie et al., 2024). The original code for the generation of mel spectrograms was based on a previous animal bioacoustics study (Thomas et al., 2022), and we set the key parameters of generating mel spectrograms generations as follows: N_MELS=40 (number of Mel bands), FFT_WIN=0.03 (30-ms window size for STFT), FFT_HOP=0.00375 (hop size, set to 1/8th of the window size), WINDOW="hann" (Hann window function), FMIN=0, and FMAX=24000 (frequency range in Hz). To handle varying call durations, zero-padding was applied to match the length of the longest call, ensuring consistent input dimensions for the classifiers. Each spectrogram was also z-transformed to normalize intensity values across calls, minimizing the effects of amplitude variation. These parameter settings, along with zero-padding and z-transformation, were all based on the suggestions of previous study (Thomas et al., 2022). This process resulted in 10,360 dimensions for each call.

Data structure analysis with UMAP

To assess the separability of individual classes and age classes, we performed dimensionality reduction using Uniform Manifold Approximation and Projection

(UMAP) from *umap-learn* (McInnes et al., 2018). We repeated supervised UMAP and unsupervised UMAP 100 times each using all calls (651 calls). Supervised UMAP aims to create a latent space where labels (e.g., individual identity) help guide the projection, making it easier to see how well the classes separate. In contrast, unsupervised UMAP does not use labels and reveals the natural structure of the data. In this study, we used two types of labels: individual and age-class (older (over 20 years) and younger (less than 10 years)). Both methods used the same parameters as those in the previous research (Thomas et al., 2022) except for the n_components used for comparison in the previous research: n_components=2, metric='euclidean'; min_dist=0. We calculated silhouette scores and distances among calls using each UMAP projection to evaluate cohesion within the same class and separation between different classes. Silhouette scores range from -1 to 1, indicating how tightly data points are grouped within a cluster and how well clusters are separated here; higher positive values mean better separation, while higher negative values indicate overlap between clusters.

Train-test split

Given the unbalanced sample sizes among individuals, we randomly selected 80 coo calls per individual in each classifier learning iteration. As with the 80 calls in each individual, we allocated 64 calls to the training data and the remainder (16 calls) to the test data. This process was repeated 1000 times, resulting in 1000 pairs of training and test datasets to account for variability and provide a robust estimate of classifier performance.

Classifier learning

In this study, we tackled two tasks: One was to identify individuals from the calls of six adult females, and the other was to classify age classes (older = over 20 years old and younger = less than 10 years old). For individual and age-class classification, we trained random forests (RF) and support vector machines (SVM) using training datasets and then predicted test datasets 1000 times using *scikit-learn,* a commonly used Python module (Pedregosa et al., 2011). In each iteration, to estimate the performance of classifiers, we calculated balanced accuracy, which is defined as the average of recall in each class and reduced the biases due to the differences in data size between classes compared to the standard accuracy. By using the prediction based on the test data, we calculated the accuracy per call defined as follows: the number of times the predicted label in the test data matched the actual label divided by the number of times the call was used as the test data.

Hyperparameter tuning

Since it is time-consuming to tune the hyperparameters of the classifiers in each iteration, we explored the best parameters using 5-repeated 5-fold cross-validation in the grid search method. In other words, we repeated the random selection of 80 calls from each individual and 5-fold cross-validation 5 times, and then we searched for the combinations of hyperparameters with the highest average balanced accuracy, commonly known as the grid search method. We searched for the best combination of hyperparameters with high average balanced accuracy for RF and SVM in the age-class classification and individual identification tasks, respectively. The grids explored by RF and SVM are as follows: for RF i) max_depth (10, 20, 50, None), ii) max_features (sqrt, log2, None), iii) criterion (gini, entropy, log_loss), iv) min_samples_split (2, 4, 6, 8), v) min_samples_leaf (1, 2, 5); for SVM i) the kernel type (linear, poly, rbf, sigmoid), ii) C,

which is the regularization parameter (0.1, 1, 10, 100), iii) the kernel coefficient in poly, rbf and sigmoid (scale, 0.001, 0.01, 0.1, 1), iv) decision_function_shape (ovo, ovr), v) for poly or sigmoid, coef0 (0.0, 0.1, 0.5, 1), and vi) for poly, degree (1, 2, 3, 4). To reduce the computational costs, we set n_estimator =100 in the RF. The best hyperparameters were as follows: In the individual identification task, for RF best hyperparameter combination was n_estimators=100, max_depth=20, max_features=sqrt, criterion=entropy, min_samples_split=2, and min_samples_leaf=1; for SVM C=10, kernel=rbf, gamma=scale, and decision_function_shape=ovo. In the age-class classification task, for RF the best hyperparameter combination was n_estimators=100, max_depth=20, max_features=sqrt, criterion=gini, min_samples_split=2, and min_samples_leaf=2; for SVM C=0.1, kernel=poly, degree=3, coef0=0.5, gamma=0.001, and decision_function_shape=ovo.

Results

Individual identification

Coo calls were generally clustered by individual in the supervised UMAP, while the unsupervised UMAP did not produce clear clusters (Figure 1). This was also reflected in the results of the silhouette score. The mean silhouette score for all six individuals of the supervised UMAP (0.60) was much higher compared to the unsupervised UMAP (0.05) (Figure 2). The mean silhouette score in each individual was positive (0.21-0.90) in the supervised UMAP; however, the silhouette scores of Kapa and Rine were negative in the unsupervised UMAP. The all-average distances within individuals were smaller than between individuals in the supervised UMAP; on the other hand, the unsupervised UMAP did not show much difference in the within-individual and between-individual distances (Figure 2).

In the individual identification task, the mean balanced accuracy of RF was 81% and that of SVM was 82% (Table 1). The mean accuracy varied across individuals, with values ranging from 0.68 to 0.94 in RF and from 0.72 to 0.94 in SVM. Both RF and SVM had the highest accuracy in Sasa and the lowest in Kapa. The main factor lowering the accuracy for both RF and SVM was the misidentification of Kapa, Rine, and Taiwu among them.

The distribution of the accuracies of all coo calsl was divided into high-accuracy and low accuracy calls (Figure 4). When we defined the calls with accuracy below 1/6 (lower than expected) as low-accuracy calls and over 0.9 as high-accuracy calls, 55 coo calls were classified as low accuracy (Table 2) and 447 as high accuracy (Table 3) in both RF and SVM. The proportion of low-accuracy calls in each individual varied, with the highest for Kapa at 0.15 and the lowest for Sasa at 0.02. On the other hand, the proportion of high-accuracy calls ranged from 0.54 to 0.89.

Age-class classification

In the supervised UMAP, coo calls were clustered tightly by age-class, while the unsupervised UMAP showed a looser separation between coo calls of younger and older individuals (Figure 5). As with the supervised UMAP, the averages of silhouette scores were 0.86 for the younger and 0.80 for the older, respectively, which were much higher than the unsupervised UMAP with 0.18 for the younger and 0.01 for the older (Figure 6). Nevertheless, only some calls of the older individuals had negative silhouette scores in the supervised UMAP. In terms of distances, the mean pairwise distance within age classes (3.05 for the younger and 3.76 for the older) was smaller than between age classes (22.66) in the supervised UMAP, whereas no differences were observed within and between age classes in the unsupervised UMAP (Figure 6).

In the age classification task, the balanced accuracy for RF was 91% and that for SVM was 93% (Table 4), thus exceeding 90%. The accuracy for young individuals exceeded 98% in both RF and SVM, while that for older individuals was less than 87%. Therefore, misclassification was caused by classifying the older calls as younger ones.

In the age classification task, the accuracy scores of each call were sharply divided between high and low values (Figure 8). When we defined the calls with accuracy below 0.5 (lower than expected) as low accuracy calls, the common low-accuracy calls between RF and SVM were 23, comprising of only Sazae and Taiko (Table 5). Here 12.3% of the Sazae calls and 9.5% of Taiko calls were low accuracy.

Discussion

Individual identification

In this study, RF and SVM achieved 81% and 82%, respectively, in identifying the individual from coo calls. The previous research on individual identification from contact calls in macaques achieved variable accuracy (0.32-0.92), although the datasets were obtained in laboratory settings (Ceugniet and Izumi, 2004; Fukushima et al., 2015; Trawicki, 2024). Considering the difficulty of collecting a balanced set of vocalizations of sufficient and well-controlled sound quality from a social group of wild macaques, the performance of our approach was favorable in comparison to classifiers that relied solely on highly controlled calls in a laboratory setting. In previous studies, the performance of classifiers varied greatly depending on the acoustic features used (Fukushima et al., 2015; Rendall et al., 1998). This means that the arbitrary way of selecting acoustic features may distort the performance evaluation. Our approach using mel spectrograms would be more versatile for eliminating arbitrariness by not extracting specific acoustic features.

The method using mel spectrograms for individual identification has the potential to facilitate network analysis of vocal communications and density estimation in the wild. Recently, vocal individual identification has received attention for its use in mark-and-recapture methods due to its non-invasive nature and ability to collect data even at night (Longden et al., 2020). While the average accuracy of individual identification in this study was 80%, more than half of calls achieved over 90% accuracy for any individual with either the RF or SVM models. Highly accurate discrimination is possible by collecting enough data for each individual, identifying calls that are easier to discriminate in advance, and focusing on those calls. Additionally, employing supervised UMAP to pre-check the data structure, as recommended in a previous study (Arnaud et al., 2023), would be beneficial, since supervised UMAP can capture structures that unsupervised UMAP cannot clarify, such as higher silhouette scores and longer inter-class distances (Figures 1, 2, 5, and 6) . Although this study demonstrates the effectiveness of mel spectrograms in small groups, further research is needed to verify whether similar accuracy can be achieved with larger numbers of individuals.

In the individual identification task, the accuracy of each call was sharply divided and skewed into high and the low scores, and low accuracy calls were made by all individuals. These results might be caused by mislabeling or data sampling bias, such as the collection of only a small number of specific types of coo calls in an individual. Otherwise, the results suggest that some of the individuals' calls might be distinctive. Such differences in individuality might differ depending on the context. For example, in some primates, the call tends to be less individualistic when the distance from other individuals is sufficiently close (Leliveld et al., 2011; Mitani et al., 1996). In addition, the coo call in response tends to resemble the call of the individual whose call

was responded to (Sugiura, 1998). These factors may affect the individuality of each coo call. Further study of the relationship between individuality and the context of the calls would help to elucidate which factor affects the difference in individuality.

Age-class classification

This study achieved 91-93% accuracy in the age-class classification task. This high accuracy might be attributed to the use of the entire spectrogram data rather than specific acoustic features. Inoue suggested that older Japanese macaques emit harsher calls (Inoue, 1988), and mel spectrograms have the potential to capture features representing the overall shape of spectrograms, such as harshness, which are difficult to capture from specific acoustic features like pitch or resonance frequencies. Although the accuracy for the older class was relatively low (82.3-86.7%), the precision for the older class (i.e., the probability that a call predicted to be the older class is indeed from the older class) was 97-98%. In other words, not all older individual calls can be classified, but the prediction of older calls is quite accurate. Furthermore, there were no low-accuracy calls common between SVM and RF in younger individuals. Such accurate recognition of older calls may contribute to maintaining spatial cohesion within a lineage, as previous studies have suggested that older individuals serve as hubs of the vocal network in the lineage (Mitani, 1986). However, since the classification mechanism of machine learning is probably different from the recognition mechanism of Japanese macaques, it is necessary to test experimentally whether age classes can be classified from calls by Japanese macaques.

Acknowledgements

We appreciate the residents and researchers on Yakushima Island for supporting

our field study, as well as the Yakushima Forest Ecosystem Conservation Center and the Kagoshima Prefectural Government for permission to perform the research. We also Drs. Hideki Sugiura and Mariko Suzuki for giving thoughtful comments and Sarah Katherine Carnes for preprocessing the vocal data. The study was supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (No. 23H02563 to GH, No. 24H00774 to IM, and No. 23K25168 and No. 23H05428 to HK) and the Collaborative Research Program of Wildlife Research Center, Kyoto University.

Author contributions

Convinced procedures: RK(lead), FK(supporting), HK(equally), IM (supporting), GH (supporting); Data collection: RK (lead), FK(equally); Software: RK (lead), HK(supporting); Grant obtained: HK(supporting); IM (supporting); GH (lead); Supervision; HK(supporting), IM (supporting), GH (lead); Formal analysis: RK(lead), HK(supporting); Manuscript writing: RK(equally), IM(supporting), HK(lead), GH (supporting)

Conflict of interests

The authors declare no conflicts of interest.

Figures

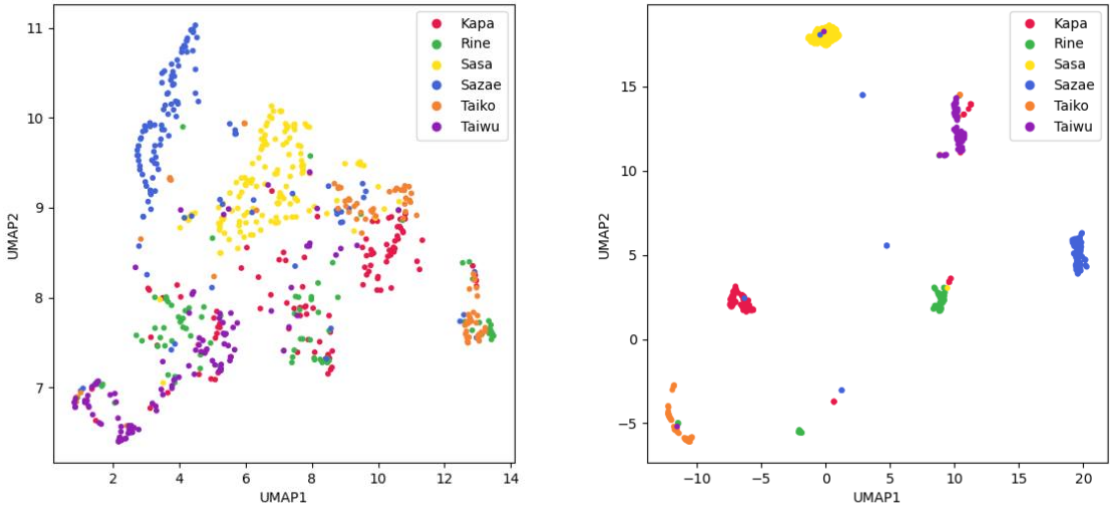

Figure 1 Projections of all coo calls into two-dimensional space through UMAP using individual labels (651 calls; each dot = 1 call; different colors indicate different individuals). Left: unsupervised UMAP. Right: supervised UMAP.

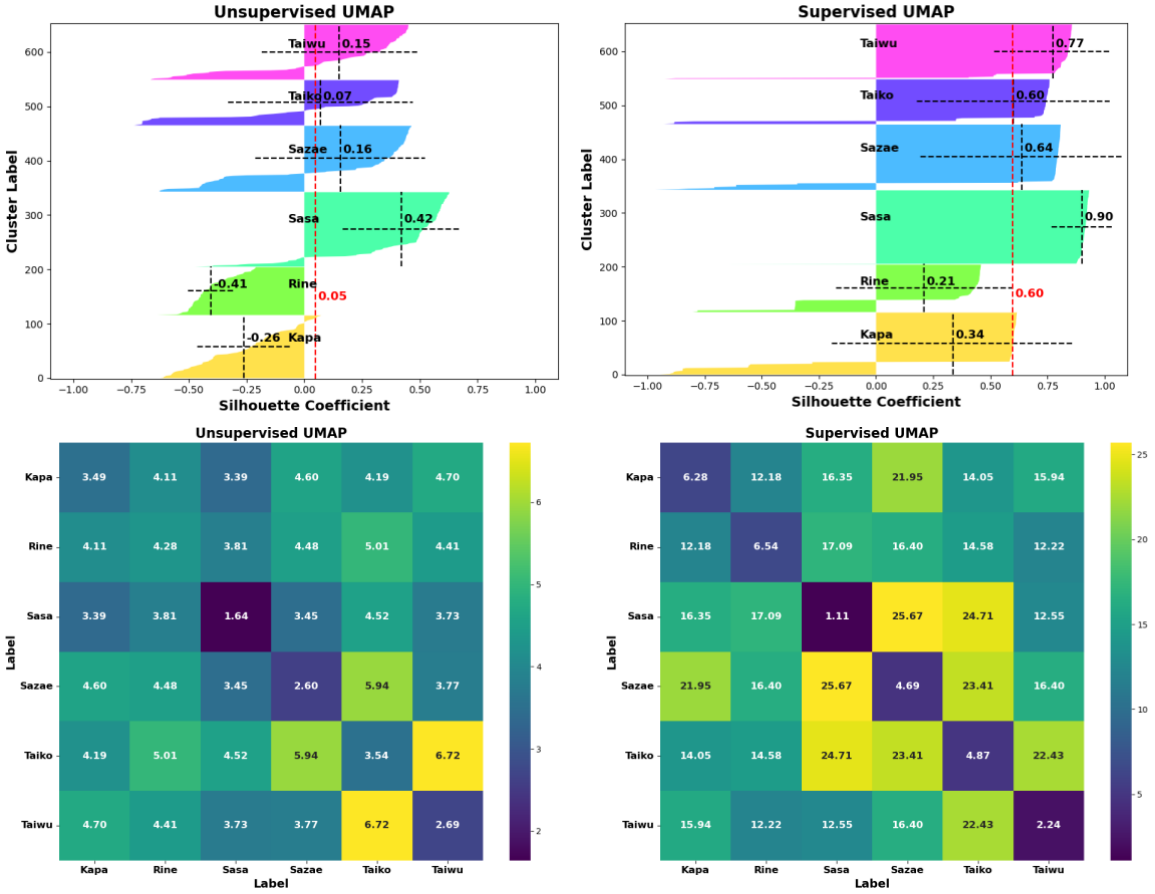

Figure 2 Silhouette plots and distances between individuals using projections of UMAPs and individual labels. Top-left: silhouette plots of unsupervised UMAP with averages and standard deviations per individual (black line) and in all individuals (red line). Bottom-left: distance matrix of unsupervised UMAP. Top-right: silhouette plots of supervised UMAP. Bottom-right: distance matrix of supervised UMAP with averages and standard deviations per individual (black line) and in all individuals (red line).

Table 1 Balanced accuracy of six individual identifications

Classifier	Mean	std	min	25%	50%	75%	max
RF	**0.81**	**0.04**	**0.69**	**0.78**	**0.81**	**0.83**	**0.92**
SVM	**0.82**	**0.04**	**0.71**	**0.79**	**0.82**	**0.84**	**0.93**

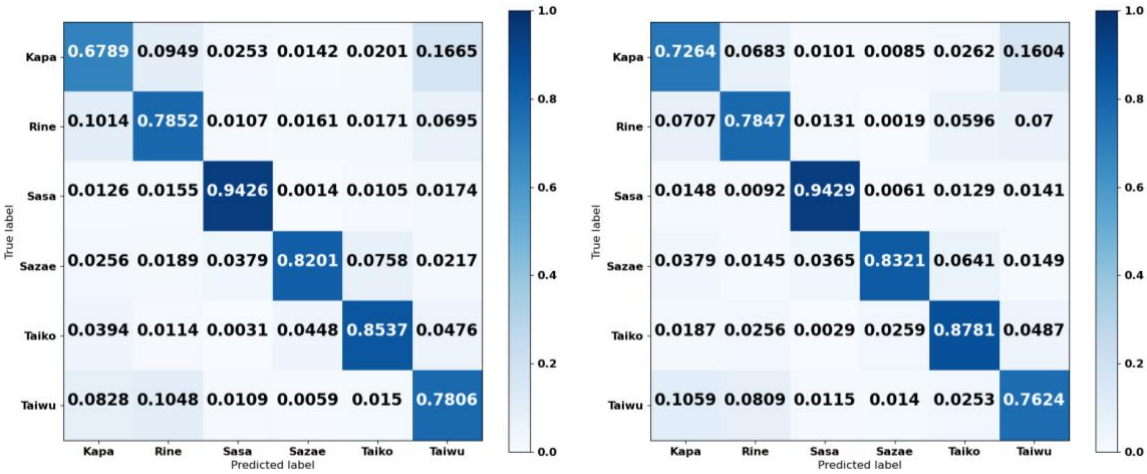

Figure 3 Average confusion matrix for 1000 iterations in identifying six females. Left: RF. Right: SVM.

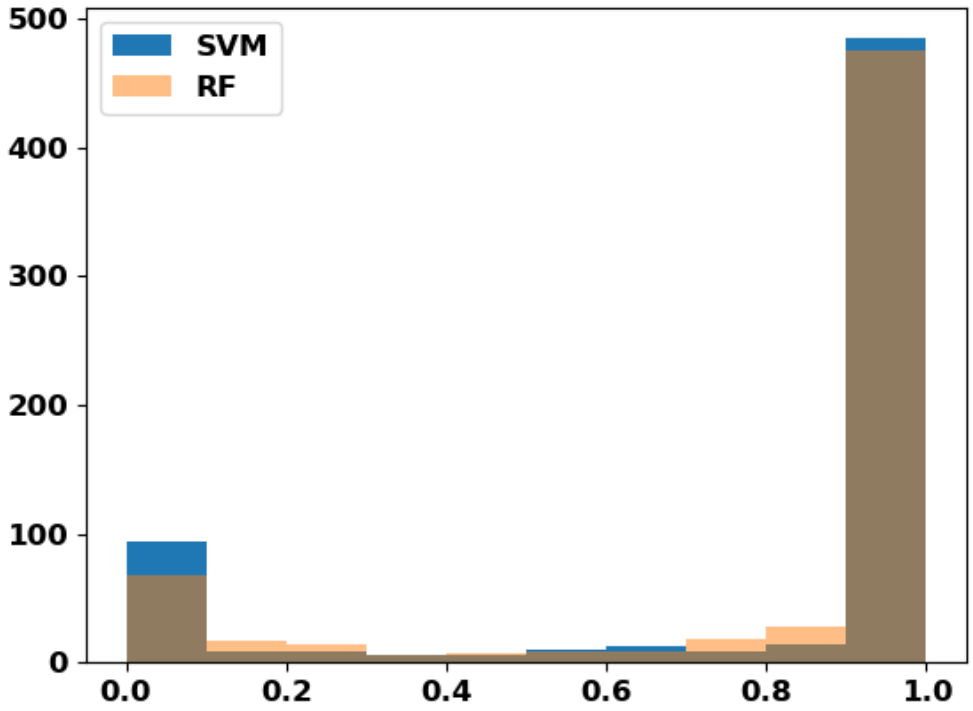

Figure 4 Frequency of accuracy for coo calls in individual identification. Blue bar and orange bar represent SVM results and RF results, respectively.

363 **Table 2 Low-and-high accuracy calls in individual identification**

Individual	Number of low-accuracy calls in RF	Number of low-accuracy calls in SVM	Number of low-accuracy calls in common between RF and SVM	Number of calls
Kapa	**26**	**20**	**17**	**116**
Rine	**12**	**10**	**9**	**89**
Sasa	**6**	**6**	**3**	**138**
Sazae	**18**	**16**	**13**	**122**
Taiko	**7**	**7**	**4**	**84**
Taiwu	**12**	**9**	**9**	**102**

364

365

366 **Table 3 Low-and-high accuracy calls in individual identification**

Individual	Number of high accuracy calls in RF	Number of high accuracy calls in SVM	Number of high-accuracy calls in common between RF and SVM	Number of calls
Kapa	**67**	**74**	**63**	**116**
Rine	**58**	**54**	**49**	**89**
Sasa	**126**	**127**	**124**	**138**
Sazae	**94**	**94**	**92**	**122**
Taiko	**67**	**64**	**60**	**84**
Taiwu	**63**	**72**	**59**	**102**

367

368

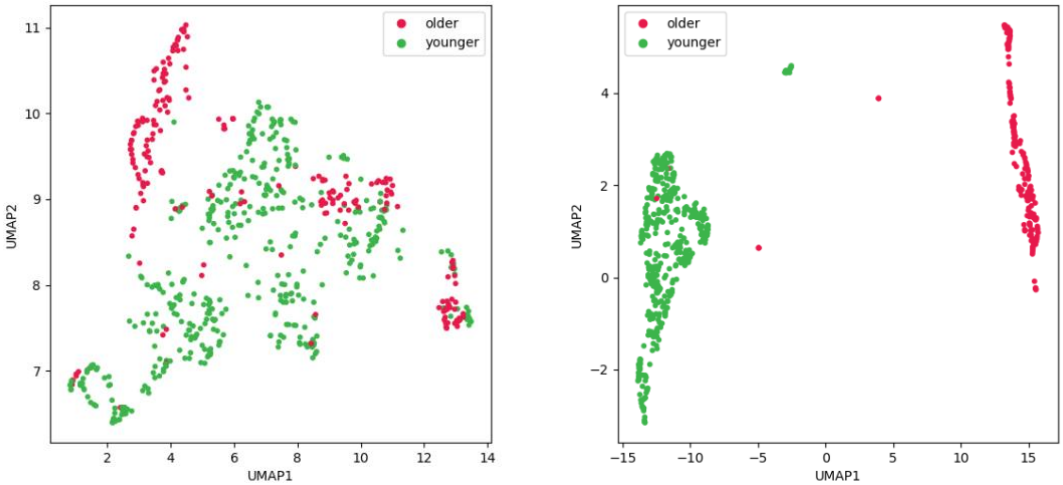

Figure 5 Projections of all coo calls into two-dimensional space through UMAP using age labels (651 calls; each dot = 1 call; different colors indicate different age classes). Left: Unsupervised UMAP. Right: Supervised UMAP

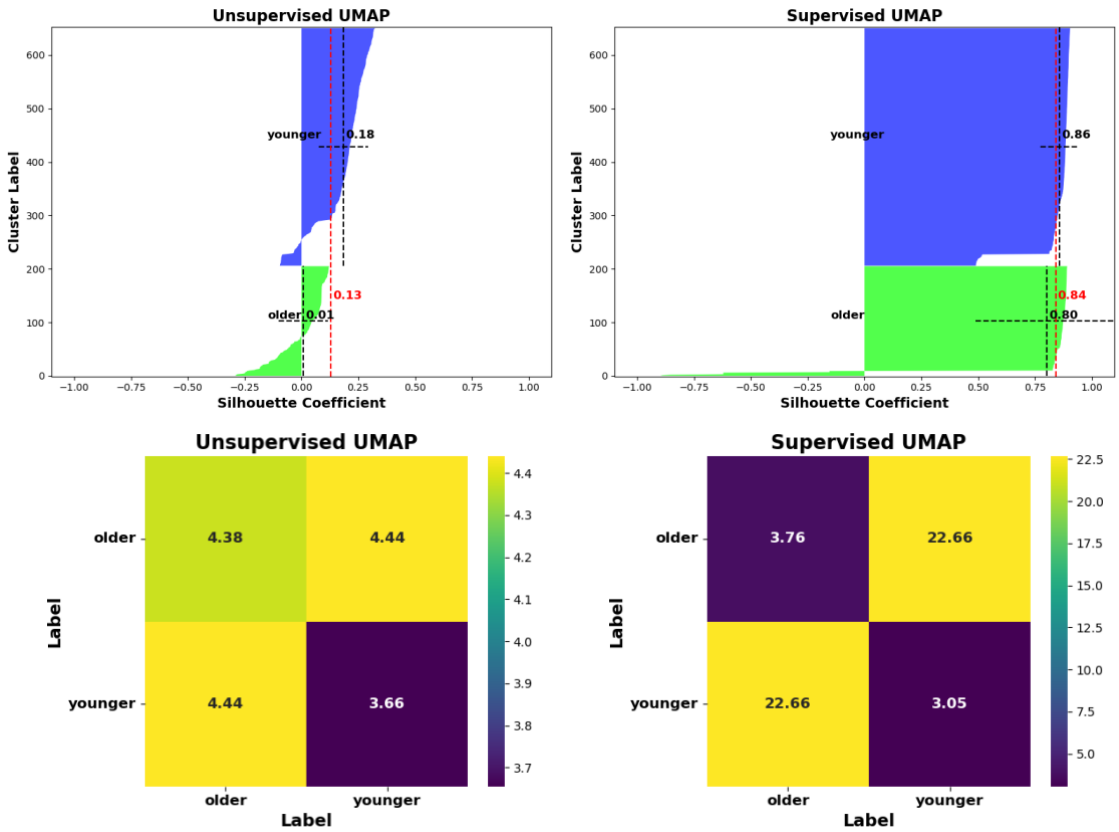

Figure 6 Silhouette plots and distances between individuals using projections of UMAPs and age class labels. Top-left: silhouette plots of unsupervised UMAP with averages and standard deviations per age class (black line) and in all classes (red line). Bottom-left: distance matrix of unsupervised UMAP. Top-right: silhouette plots of supervised UMAP with averages and standard deviations per age class (black line) and in all classes (red line). Bottom-right: distance matrix of supervised UMAP.

Table 4 Balanced accuracy of age classification

Classifier	Mean	std	min	25%	50%	75%	max
RF	**0.91**	**0.03**	**0.78**	**0.88**	**0.91**	**0.93**	**0.98**
SVM	**0.93**	**0.03**	**0.81**	**0.91**	**0.93**	**0.95**	**1.00**

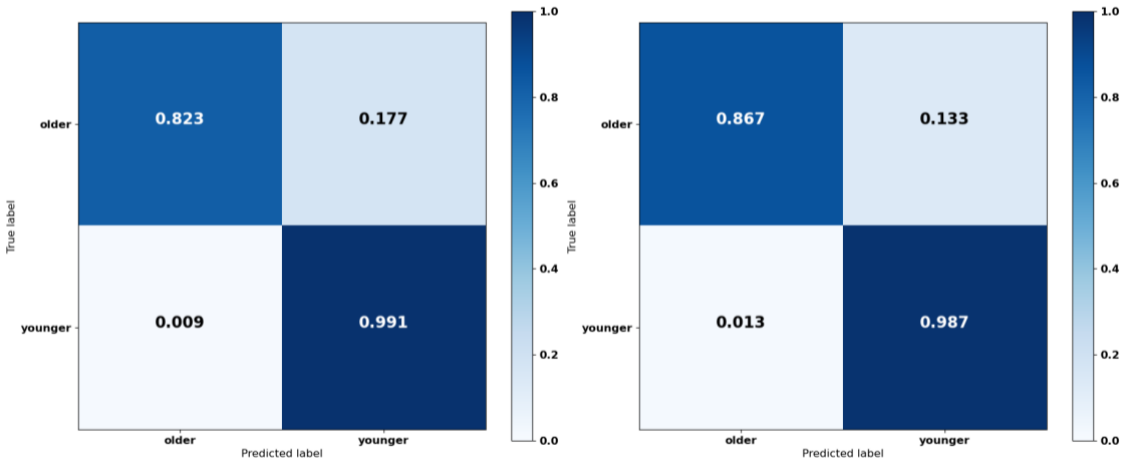

Figure 7 The Confusion matrix for classifying age-class between six females. Left: RF. Right: SVM.

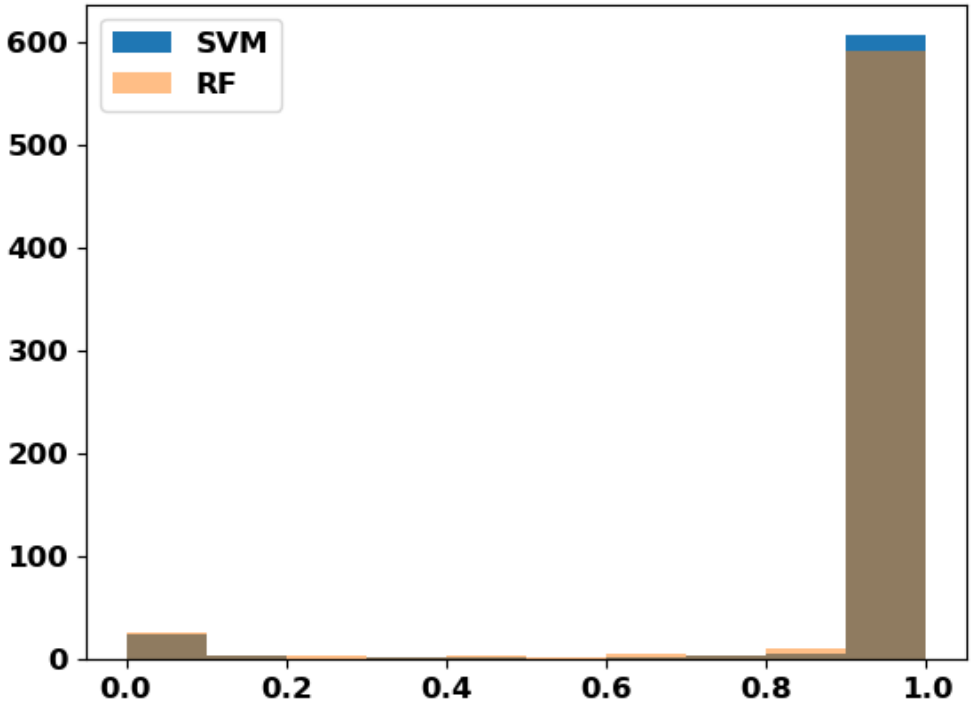

Figure 8 Frequency of accuracy for coo calls in age-class classification. Blue bar and orange bar represent SVM results and RF results, respectively

Table 5 Low-accuracy calls in age classification

Individual	Number of low-accuracy calls in RF	Number of low-accuracy calls in SVM	Number of low-accuracy calls in common between RF and SVM	Number of calls
Kapa	**0**	**2**	**0**	**116**
Rine	**0**	**0**	**0**	**89**
Sasa	**1**	**0**	**0**	**138**
Sazae	**23**	**16**	**15**	**122**
Taiko	**14**	**10**	**8**	**84**
Taiwu	**0**	**1**	**0**	**102**

Reference

Arnaud, V., F. Pellegrino, S. Keenan, X. St-Gelais, N. Mathevon, F. Levréro, and C. Coupé, 2023, Improving the workflow to crack Small, Unbalanced, Noisy, but Genuine (SUNG) datasets in bioacoustics: The case of bonobo calls: PLOS Computational Biology, v. 19, p. e1010325.

Baowaly, M. K., B. C. Sarkar, M. A. A. Walid, M. M. Ahamad, B. C. Singh, E. S. Alvarado, I. Ashraf, and M. A. Samad, 2024, Deep transfer learning-based bird species classification using mel spectrogram images: PLOS ONE, v. 19.

Boersma, P. W., David., 2024, Praat: doing phonetics by computer [Computer program]. Version 6.4.12, retrieved 8 May 2024 from http://www.praat.org/.

Ceugniet, M., and A. Izumi, 2004, Individual vocal differences of the coo call in Japanese monkeys: Comptes Rendus. Biologies, v. 327.

Dossou, B. F., and Y. K. Gbenou, 2021, FSER: Deep convolutional neural networks for speech emotion recognition: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 3533-3538.

Feinberg, D. R., B. C. Jones, A. C. Little, D. M. Burt, and D. I. Perrett, 2005, Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices: Animal Behaviour, v. 69.

Ferrand, C. T., 2002, Harmonics-to-Noise Ratio: An Index of Vocal Aging: Journal of Voice, v. 16.

Fukushima, M., A. M. Doyle, M. P. Mullarkey, M. Mishkin, and B. B. Averbeck, 2015, Distributed acoustic cues for caller identity in macaque vocalization: Royal Society Open Science, v. 2, p. 150432.

Harnsberger, J. D., R. Shrivastav, W. S. Brown, H. Rothman, and H. Hollien, 2008, Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age:

Journal of Voice, v. 22.

Hindle, A. G., M. Horning, J.-A. E. Mellish, and J. M. Lawler, 2009, Diving into old age: muscular senescence in a large-bodied, long-lived mammal, the Weddell seal (*Leptonychotes weddellii*): Journal of Experimental Biology, v. 212.

Huntley, R., H. Hollien, and T. Shipp, 1987, Influences of listener characteristics on perceived age estimations: Journal of Voice, v. 1.

Inoue, M., 1988, Age Gradations in Vocalization and Body Weight in Japanese Monkeys (*Macaca fuscata*): Folia Primatologica, v. 51.

Klofstad, C. A., R. C. Anderson, and S. Nowicki, 2015, Perceptions of Competence, Strength, and Age Influence Voters to Select Leaders with Lower-Pitched Voices: PLOS ONE, v. 10.

Lakdari, M. W., A. H. Ahmad, S. Sethi, G. A. Bohn, and D. J. Clink, 2024, Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons: Ecological Informatics, v. 80.

Leliveld, L. M. C., M. Scheumann, and E. Zimmermann, 2011, Acoustic correlates of individuality in the vocal repertoire of a nocturnal primate (*Microcebus murinus*): The Journal of the Acoustical Society of America, v. 129.

Lemasson, A., E. Gandon, and M. Hausberger, 2010, Attention to elders' voice in non-human primates: Biology Letters, v. 6.

Longden, E. G., S. H. Elwen, B. McGovern, B. S. James, C. B. Embling, and T. Gridley, 2020, Mark–recapture of individually distinctive calls—a case study with signature whistles of bottlenose dolphins (*Tursiops truncatus*): Journal of Mammalogy, v. 101.

Manrique, H. M., and J. Call, 2015, Age-dependent cognitive inflexibility in great apes:

Animal Behaviour, v. 102.

Marler, P., 1956, The voice of the chaffinch and its function as a language: Ibis.

Marler, P., and L. Hobbett, 1975, Individuality in a Long-Range Vocalization of Wild Chimpanzees: Zeitschrift für Tierpsychologie, v. 38.

McFee, B., C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, 2015, librosa: Audio and Music Signal Analysis in Python: SciPy.

McInnes, L., J. Healy, and J. Melville, 2018, Umap: Uniform manifold approximation and projection for dimension reduction: arXiv preprint arXiv:1802.03426.

Mitani, J. C., J. Gros-Louis, and J. M. Macedonia, 1996, Selection for acoustic individuality within the vocal repertoire of wild chimpanzees: International Journal of Primatology, v. 17, p. 569-583.

Mitani, M., 1986, Voiceprint identification and its application to sociological studies of wild Japanese monkeys (*Macaca fuscata yakui*): Primates, v. 27, p. 397-412.

Nakagawa, N., M. Nakamichi, and H. Sugiura, 2010, The Japanese Macaques: Primatology Monographs: Tokyo, Springer.

Oyakawa, C., H. Koda, and H. Sugiura, 2007, Acoustic features contributing to the individuality of wild agile gibbon (*Hylobates agilis agilis*) songs: American Journal of Primatology, v. 69.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011, Scikit-learn: Machine Learning in Python: JOURNAL OF MACHINE LEARNING RESEARCH, v. 12, p. 2825-2830.

Rendall, D., M. J. Owren, and P. S. Rodman, 1998, The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations: The Journal

of the Acoustical Society of America, v. 103.

Ryan, W. J., and K. W. Burk, 1974, Perceptual and acoustic correlates of aging in the speech of males: Journal of Communication Disorders, v. 7.

Stathopoulos, E. T., J. E. Huber, and J. E. Sussman, 2011, Changes in Acoustic Characteristics of the Voice Across the Life Span: Measures From Individuals 4–93 Years of Age: Journal of Speech, Language, and Hearing Research, v. 54.

Stevens, S. S., J. Volkmann, and E. B. Newman, 1937, A Scale for the Measurement of the Psychological Magnitude Pitch: The Journal of the Acoustical Society of America, v. 8.

Sugiura, H., 1998, Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques: Animal Behaviour, v. 55, p. 673-687.

Thomas, M., F. H. Jensen, B. Averly, V. Demartsev, M. B. Manser, T. Sainburg, M. A. Roch, and A. Strandburg-Peshkin, 2022, A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations: Journal of Animal Ecology, v. 91, p. 1567-1581.

Tigue, C. C., D. J. Borak, J. J. M. O'Connor, C. Schandl, and D. R. Feinberg, 2012, Voice pitch influences voting behavior: Evolution and Human Behavior, v. 33.

Trawicki, M. B., 2024, Automatic gender recognition and speaker identification of Rhesus Macaques (*Macaca mulatta*) using hidden Markov models (HMMs): International Journal of Speech Technology, v. 27, p. 179-186.

Walsh, S. L., S. Engesser, S. W. Townsend, and A. R. Ridley, 2023, Multi-level combinatoriality in magpie non-song vocalizations: Journal of the Royal Society Interface, v. 20, p. 20220679.

Winter, P., D. Ploog, and J. Latta, 1966, Vocal repertoire of the squirrel monkey (*Saimiri sciureus*), its analysis and significance: Experimental Brain Research, v. 1, p.

359-384.

Xie, B., V. Daunay, T. C. Petersen, and E. F. Briefer, 2024, Vocal repertoire and individuality in the plains zebra (*Equus quagga*): Royal Society Open Science, v. 11.

Yamagiwa, J., and D. A. Hill, 1998, Intraspecific variation in the social organization of Japanese macaques: Past and present scope of field studies in natural habitats: Primates, v. 39, p. 257-273.

Zipple, M. N., S. Peters, W. A. Searcy, and S. Nowicki, 2020, Sounds of senescence: male swamp sparrows respond less aggressively to the songs of older individuals: Behavioral Ecology, v. 31.