Exploratory analysis of yellow mongoose vocalization: detection from in-the-wild recordings and call classification

Sevada Hovsepyan [1,*,**], *Imen Ben Mahmoud* [1,*], *Vanessa Rüegg* [2,*], *Marta Manser* [2], *Mathew Magimai.-Doss* [1]

[1] Idiap Research Institute, Switzerland
[2] University of Zurich, Switzerland

sevada.hovsepyan@idiap.ch, imen.benmahmoud@idiap.ch, vanessa.rueegg3@uzh.ch, marta.manser@ieu.uzh.ch, mathew@idiap.ch

Abstract

Understanding the vocal repertoire of social animals, such as the yellow mongoose (YM), is crucial for deciphering their social structure and evolutionary history. In this study, we present an exploratory analysis of YM vocalization detection and classification using both signal processing and machine learning approaches. The analysis uses two distinct datasets: (1) expert-annotated, clean pup vocalizations (n=940), and (2) noisy field recordings (n=29) captured with directional microphones. Our results indicate that handcrafted features, originally developed for human speech modeling to represent syllables, are effective for animal vocalization classification, suggesting potential shared acoustic structures. Although detecting YM vocalizations in wild recordings proved challenging, our objective is to aid annotators by flagging potential vocalization segments. The results presented in this work provide a foundation for scalable, semi-automated analysis of animal vocal repertoires.

Index Terms: animal vocalization, yellow mongoose, language evolution, handcrafted features, animal VAD

1. Introduction

Social animals rely on vocal communication to navigate complex social structures, but most research has focused on highly social or obligate social species [1, 2]. The yellow mongoose (Cynictis penicillata), as a facultatively social carnivore, provides an important model for exploring communication in species with flexible social systems [2, 3]. The Social Complexity Hypothesis suggests that vocal complexity increases with social complexity [4, 5], yet communication in less social or facultatively social species remains understudied. Investigating the vocal repertoire and ontogeny of such species, including how vocalizations develop from pups to adults, can offer insights into the evolutionary drivers of communication and whether vocalizations are innate or learned [2, 6, 7].

Analyzing YM vocalizations is challenging due to their dependence on social context, which necessitates naturalistic recordings in group settings [8]. Even after data collection, the process remains labor-intensive, requiring repeated listening, annotation, cleaning, and categorization of vocalizations into distinct types and contexts. These challenges underscore the need for scalable, semi-automated tools to assist researchers in analyzing animal vocal repertoires.

While machine learning approaches have been developed to address these challenges [9, 10, 11, 12], tools tailored to YM vocalizations are lacking. This manuscript explores how current methods [13, 14, 15] can be adapted for YM vocalization analysis, focusing on two key research questions: (1) the identification and classification of vocalization/call types, and (2) the detection of vocalizations in noisy field recordings.

For the first research question, while some call types are well-defined (e.g. "krr voc", "begging call", "chime", etc), others remain ambiguous: either because they are difficult to assign to known categories or because they may represent previously undiscovered call types. The challenge here is twofold: first, to systematically classify these ambiguous vocalizations, and second, to determine whether they belong to known categories or constitute novel types.

The second research question focuses on detecting vocalizations in real-world recordings using directional microphones. Despite the focused capture, recordings are often noisy, obscuring YM vocalizations due to background interference or calls from other species. Our goal is to assist researchers by flagging potential vocalization segments, reducing the annotation burden (e.g. accept high false positives, if accuracy is also high). This approach lays the foundation for scalable, semi-automated analysis of animal vocal repertoires.

This exploratory analysis evaluates current methods for addressing these challenges, laying the groundwork for future end-to-end solutions. Such solutions would enable vocalizations to be detected in wild recordings and classified into distinct groups, ultimately advancing our understanding of YM communication. The remainder of the manuscript is organized as follows: 2. Methods describes the dataset, handcrafted feature sets, and the signal processing and machine learning tools used. 3. Results presents the vocalization detection and classification results, accompanied by analysis. Finally, 4. Discussion interprets the findings and concludes the study.

2. Methods

2.1. Datasets

The manuscript utilizes YM recordings collected in the Kalahari, South Africa, using a Sennheiser ME66 directional microphone and Marantz PMD660 recorder (48 kHz, 24-bit).

2.1.1. Yellow Mongoose Pup Vocalizations

We analyzed a dataset of expert-annotated YM pup vocalizations, consisting of 940 high-quality recordings labeled into 9 call types, 8 of which were well-defined, while the 9th category,

[*] These authors contributed equally.
[**] indicates the corresponding author.

"undefined pup voc," includes vocalizations that could not be confidently assigned to the known types, potentially representing ambiguous or novel call types (Figure 1). The dataset is highly imbalanced, with "krr voc" being the most frequent and "undefined pup voc" the second largest group. This imbalance was considered during analysis and interpretation of results.

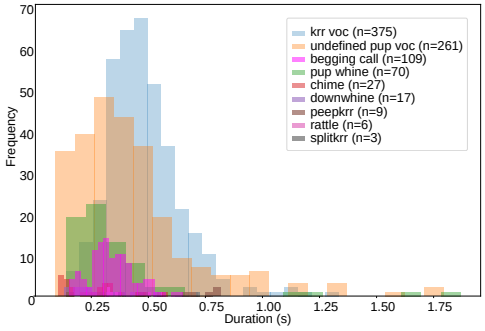

Figure 1: *Vocalization duration distribution for yellow mongoose pup vocalizations.*

2.1.2. Field Recordings

The second dataset comprises 29 field recordings captured using directional microphones. These recordings include a mix of pup and adult yellow mongoose vocalizations, as well as background noise and calls from other species. Recordings were expert annotated, with timestamps marking the start and end of YM vocalizations (overall, 1948). The total duration of the field recordings spans several hours.

2.2. Feature Extraction and Classification

2.2.1. Handcrafted Features

For the pup vocalizations, we employed a signal processing-based approach using handcrafted features. These features parametrize spectrotemporal patterns of vocalizations and are rooted in neurocomputational models of speech perception [16, 17, 18]. They have been previously used in paralinguistic challenges, such as Parkinson's disease detection from speech [19]. The pipeline involved calculating the short-term Fourier transform (STFT) for each vocalization (downsampled to 20.5kHz) using the librosa library [20], collapsing the resulting power spectrogram into F=24 frequency channels by averaging the spectral energy in adjacent channels, and binning it into T=16 temporal segments. The standardized spectrogram (16x24) was then converted to a dB scale and flattened to create a feature vector for each vocalization (1x(16x24) = 1x384).

2.2.2. Deep Learning Approach

To classify vocalizations directly from raw waveforms (downsampled to 16kHz), we employed a convolutional neural network (CNN). This approach originated from speech processing [21, 22], and later was extended to animal vocalization [10, 23]. The architecture, detailed in Table 1, consists of four convolutional blocks. Each block includes a convolutional layer, batch normalization, max pooling, and Rectified Linear Unit (ReLU) activation. The network concludes with an adaptive average pooling layer (target size set to 1), allowing it to handle variable-length waveform inputs. The final layers are fully connected, producing class probabilities for each vocalization type. The model was implemented using PyTorch [24] and trained using the cross-entropy loss function to optimize classification performance.

Table 1: *CNN architecture for CNN-crafted call classifier.* n_f *denotes the number of filters. HU denotes the number of hidden units.*

Block	Operation	Kernel	Stride	Padding	n_f/HU
1	Convolution	5	2	0	20
	Batch Normalization	-	-	-	20
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
2	Convolution	5	2	0	40
	Batch Normalization	-	-	-	40
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
3	Convolution	3	1	0	40
	Batch Normalization	-	-	-	40
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
4	Convolution	3	1	0	40
	Batch Normalization	-	-	-	40
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
5	Adaptive Avg Pooling	-	-	-	-
	Flatten	-	-	-	-
	Fully Connected Layer	-	-	-	10
	ReLU Activation	-	-	-	-
	Fully Connected Layer	-	-	-	9

2.2.3. Classification Protocol

We evaluated both approaches using 5-fold stratified cross-validation to ensure equal representation of each vocalization type. For the handcrafted features, we used a random forest classifier (number of estimators=100, random state=42) implementation in Scikit-learn [25]. For the CNN, raw waveforms were used as input. To analyze feature importance and construct the confusion matrix, we split the dataset into 30 % test and 70 % train sets. Feature importance was assessed using permutation tests [25], which allowed us to identify the spectrotemporal patterns most critical for classifying each call type. This approach provided insights into which frequency and temporal bins carried the most discriminative information for each vocalization category.

2.3. Vocalization Detection in Field Recordings

For detecting YM vocalizations in field recordings, we employed two state-of-the-art models: rVAD [13] and WhisperSeg (wSeg) [14]. rVAD is an unsupervised model designed for robust voice activity detection, making it suitable for noisy environments without requiring labeled training data [13]. WhisperSeg, in contrast, is a supervised model pretrained on both human speech and a diverse set of animal vocalizations, including those from non-human primates and other mammalian species [14]. We evaluated the performance of both models in detecting YM vocalizations, focusing on overlap ratio (ratio of detected vocalizations to actual vocalizations), false positive rate, and interval duration between detected segments. Additionally, we

analyzed which call types were more easily detectable, given their distinct frequency and temporal profiles.

3. Results

3.1. Yellow Mongoose Pup Call Classification Performance

We first compared the performance of the CNN and random RF approaches using 5-fold cross-validation. As shown in Table 2, the RF method with handcrafted features consistently outperformed the CNN, achieving a mean accuracy of 0.684 (±0.016) compared to 0.61 (±0.029) for the CNN. Given its superior performance, we selected the RF approach for further analysis.

Table 2: *Cross-validation accuracy per fold and mean across folds*

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
CNN	0.63	0.61	0.59	0.56	0.64	0.61 ± 0.029
RF	0.67	0.699	0.674	0.669	0.709	0.684 ± 0.016

Using a 30-70 test-train split, we trained the RF classifier on the handcrafted features and evaluated its performance on the test set.

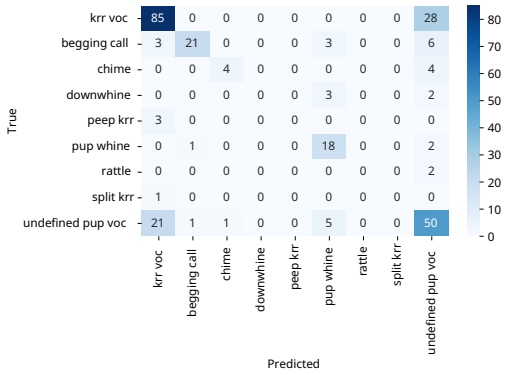

Figure 2: *Confusion matrix with handcrafted feature sets and RF classifier.*

Table 3: *Classification report on the Test set with RF classifier*

	precision	recall	f1-score	support
krr voc	0.75	0.75	0.75	113
begging call	0.91	0.64	0.75	33
chime	0.80	0.50	0.62	8
downwhine	0.00	0.00	0.00	5
peepkrr	0.00	0.00	0.00	3
pup whine	0.62	0.86	0.72	21
rattle	0.00	0.00	0.00	2
splitkrr	0.00	0.00	0.00	1
undefined pup voc	0.53	0.64	0.58	78
accuracy			0.67	264
macro avg	0.40	0.38	0.38	264
weighted avg	0.67	0.67	0.66	264

The resulting confusion matrix (Figure 2) and classification report (Table 3) reveal that the RF classifier successfully identified several call types with high accuracy. However, the dataset's imbalance, particularly the predominance of "krr" vocalizations, likely contributed to the high precision for this class. Notably, a portion of the "undefined pup voc" samples were classified as known call types, while a distinct subgroup remained classified as "undefined." This pattern suggests that the handcrafted features capture discriminative spectrotemporal patterns, aiding classification even in ambiguous cases. Furthermore, the persistence of a subgroup of "undefined pup voc" samples may indicate the presence of a novel call type not yet characterized in the dataset.

To further investigate the acoustic characteristics of the "undefined pup voc" class, we conducted permutation tests to identify the most important features for each call type.

3.2. Feature Importance Analysis

To elucidate the acoustic basis for the classification of "undefined pup voc" and other call types, we performed permutation tests to identify the most important spectrotemporal features for each class. The feature importance maps for the most prominent call types (based on the number of samples in the dataset) are presented in Figure 3. The color-map highlights the frequency and temporal bins that contribute most to the classification of each vocalization type.

3.3. Vocalization Detection in Field Recordings

We evaluated the ability of rVAD and wSeg to detect yellow mongoose vocalizations in noisy field recordings, which included both pup and adult call types. Figure 4 presents the overlap ratio (detected vs actual vocalizations) for each audio file, as well as the detection ratio for each call type. rVAD consistently outperformed wSeg, demonstrating higher sensitivity across call types and recording conditions.

Both rVAD and wSeg produced a high number of false positives, but analysis revealed that most false detections were brief and temporally clustered. Applying post-processing (merging predictions closer than a set threshold (0.1, 0.5, or 1 second)) reduced false positives for both models without affecting true positives (Table 4). This demonstrates that simple temporal smoothing can improve precision while maintaining sensitivity, making these tools effective for flagging vocalization segments in real-world recordings.

Table 4: *Detection results for rVAD and wSeg after merging the predictions if the distance is less than a certain threshold (in seconds).*

Method	threshold	sensitivity	precision
rVAD	unmerged	0.81 ± 0.27	0.19 ± 0.18
	0.1 s	0.81 ± 0.27	0.22 ± 0.21
	0.5 s	0.81 ± 0.27	0.30 ± 0.25
	1 s	0.81 ± 0.27	0.37 ± 0.27
wSeg	unmerged	0.16 ± 0.21	0.09 ± 0.14
	0.1 s	0.16 ± 0.21	0.10 ± 0.15
	0.5 s	0.16 ± 0.21	0.15 ± 0.18
	1 s	0.17 ± 0.21	0.18 ± 0.20

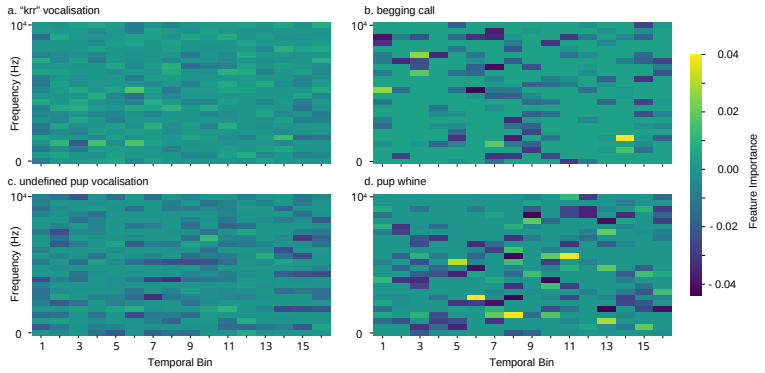

Figure 3: *Feature importance map with RF permutation test.*

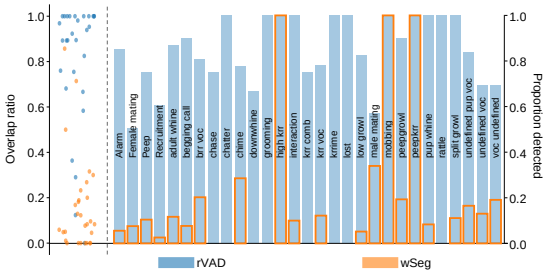

Figure 4: *Performance for rVAD and wSeg across audio files and call types.*

4. Discussion

4.1. Summary of Findings

This manuscript presents an exploratory analysis of yellow mongoose (YM) vocalization detection and classification, addressing two key research questions using distinct datasets. While the long-term goal is to develop an end-to-end framework for detection, classification, and grouping of animal vocalizations, this study focuses on evaluating current methods for each task separately. Our results demonstrate the effectiveness of state-of-the-art detection algorithms and signal processing techniques for YM vocalization analysis, providing a foundation for scalable, semi-automated analysis of animal (YM) vocal repertoires.

4.2. Classification Performance and Evolutionary Insights

Our classification results, with an F1 score of approximately 70%, indicate that handcrafted features, originally developed for human speech analysis [17, 26, 18], are highly effective for YM vocalization classification. Cross-validation revealed that the RF approach using these features consistently outperformed CNN-based approach. These features, rooted in neurocomputational models of speech perception [17, 26, 18], parametrize syllables as fundamental units of vocal communication [27, 28, 29]. Their success in classifying YM vocalizations suggests a striking parallel: like human speech, YM communication is structured around syllable-like bouts. Furthermore, the confusion matrix, show that some "undefined" vocalizations were confidently assigned to known call types, while others formed a distinct subgroup. While this does not constitute definitive proof of novel call types, permutation tests and feature importance analyses provide objective criteria for re-evaluating ambiguous vocalizations. The feature importance maps (Figure 3) reveal the specific temporal and frequency ranges that distinguish each call type: providing researchers with a data-driven approach to resolve annotation ambiguities and potentially discover new vocalization categories.

4.3. Detection Performance and Practical Utility

The detection analysis focused on real-world recordings containing a mix of pup and adult vocalizations, which introduced additional variability compared to the controlled pup vocalization dataset used for classification. rVAD [13] consistently outperformed wSeg [14], achieving high sensitivity across call types. However, its performance came with a trade-off: a higher false positive rate, primarily due to brief, clustered segments. This pattern suggests that simple post-processing, such as merging nearby segments, significantly improves precision without sacrificing sensitivity (Table 4). The primary goal of our detection analysis is to assist annotators by flagging potential vocalization segments, rather than providing fully automated, high-precision segmentation. By focusing only on preselected segments, researchers can drastically reduce the time required for manual annotation. For example, instead of reviewing hours of audio, annotators can prioritize segments flagged by rVAD, confirming or correcting detections as needed. This semi-automated approach is particularly valuable in real-world settings, where background noise and overlapping calls complicate manual annotation. Future refinements, such as integrating context-aware filtering or hybrid detection models, could further enhance the tool's utility for field researchers.

4.4. Limitations and Future Directions

While this study advances our understanding of YM vocalizations, several limitations should be acknowledged. First, the dataset imbalance, particularly the predominance of "krr voc", may influence classification performance. Future work should aim to collect more balanced datasets, including underrepresented call types. Second, while rVAD demonstrated robust detection performance, its false positive rate highlights the need for improved noise suppression or complementary detection methods. Finally, extending this framework to other species could reveal broader patterns in animal communication and further elucidate the evolutionary origins of vocal complexity.

5. Acknowledgments

We are grateful to the Kalahari Research Trust and Northern Cape Department of Environment and Nature Conservation for research permission at the Kalahari Research Centre, as well as for the logistic and financial support of the Universities of Zurich, Cambridge and Pretoria, MAVA Foundation, Zoo Zurich, Exekias and IMS Foundation on the maintenance of the field site. This work was partially funded by the NCCR Evolving Language (phase II), Swiss National Science Foundation Agreement #51NF40_225146.

6. Generative AI Use Disclosure

A Generative AI (LLMs) were exclusively used for checking the grammar and spelling, editing and polishing parts of manuscript.

7. References

[1] T. C. Schneider and P. M. Kappeler, "Social systems and life-history characteristics of mongooses," *Biological Reviews*, vol. 89, no. 1, pp. 173–198, 2014.

[2] M. B. Manser, D. A. Jansen, B. Graw, L. I. Hollén, C. A. Bousquet, R. D. Furrer, and A. le Roux, "Chapter six - vocal complexity in meerkats and other mongoose species," ser. Advances in the Study of Behavior, M. Naguib, L. Barrett, H. J. Brockmann, S. Healy, J. C. Mitani, T. J. Roper, and L. W. Simmons, Eds. Academic Press, 2014, vol. 46, pp. 281–310.

[3] G. Veron, M.-L. Patou, and A. P. Jennings, "Systematics and Evolution of the Mongooses (Herpestidae, Carnivora)," in *Small Carnivores*. John Wiley & Sons, Ltd, 2022, ch. 3, pp. 61–78.

[4] T. M. Freeberg, "Social Complexity Can Drive Vocal Complexity: Group Size Influences Vocal Information in Carolina Chickadees," *Psychological Science*, vol. 17, no. 7, pp. 557–561, Jul. 2006.

[5] T. M. Freeberg, R. I. M. Dunbar, and T. J. Ord, "Social complexity as a proximate and ultimate factor in communicative complexity," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1597, pp. 1785–1801, Jul. 2012.

[6] A. E. Ames and V. Vergara, "Trajectories of Vocal Repertoire Development in Beluga (Delphinapterus leucas) Calves: Insights from Studies a Decade Apart," *Aquatic Mammals*, vol. 46, no. 4, pp. 344–366, Jul. 2020.

[7] L. I. Hollén and M. B. Manser, "Ontogeny of alarm call responses in meerkats, Suricata suricatta: The roles of age, sex and nearby conspecifics," *Animal Behaviour*, vol. 72, no. 6, pp. 1345–1353, 2006.

[8] A. le Roux, M. I. Cherry, and M. B. Manser, "The audience effect in a facultatively social mammal, the yellow mongoose, *Cynictis penicillata*," *Animal Behaviour*, vol. 75, no. 3, pp. 943–949, Mar. 2008.

[9] R. D. Daneshyari, "Deep learning approaches for acoustic animal classification," *International Journal of Computing and Artificial Intelligence*, vol. 5, no. 2, pp. 199–204, Jul. 2024.

[10] E. Sarkar, K. Wierucka, A. B. Bosshard, J. Burkart, and M. Magimai Doss, "On feature representations for marmoset vocal communication analysis," *Bioacoustics*, vol. 34, no. 3, pp. 355–369, May 2025.

[11] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar. 2022.

[12] E. Sarkar and M. Magimai.-Doss, "Towards leveraging sequential structure in animal vocalizations," in *NIPS Workshop: AI for non-human animal communication*, 2025. [Online]. Available: https://openreview.net/forum?id=a8sQqweMIM

[13] Z.-H. Tan, A. kr. Sarkar, and N. Dehak, "rvad: An unsupervised segment-based robust voice activity detection method," *Computer Speech Language*, vol. 59, pp. 1–21, 2020.

[14] N. Gu, K. Lee, M. Basha, S. K. Ram, G. You, and R. H. Hahnloser, "Positive transfer of the whisper speech transformer to human and animal voice activity detection," in *Proc. ICASSP*. IEEE, 2024, pp. 7505–7509.

[15] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[16] I. B. Yildiz, K. von Kriegstein, and S. J. Kiebel, "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," *PLoS Computational Biology*, vol. 9, no. 9, pp. e1 003 219–e1 003 219, Sep. 2013.

[17] S. Hovsepyan, I. Olasagasti, and A.-L. Giraud, "Combining predictive coding and neural oscillations enables online syllable recognition in natural speech," *Nature Communications*, vol. 11, no. 1, p. 3117, Jun. 2020.

[18] M. Nabé, J.-L. Schwartz, and J. Diard, "Combining top-down syllabic duration prediction with bottom-up envelope processing for syllabic segmentation in speech perception: A computational Modeling study with the COSMO-Onset model," *Language, Cognition and Neuroscience*, vol. 40, no. 8, pp. 1085–1110, 2025.

[19] S. Hovsepyan and M. Magimai.-Doss, "Syllable Level Features for Parkinson's Disease Detection from Speech," in *Proc. ICASSP*. IEEE, Apr. 2024, pp. 11 416–11 420.

[20] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and Music Signal Analysis in Python," pp. 18–24. [Online]. Available: https://conference.scipy.org/proceedings/scipy2015/brian_mcfee.html

[21] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. Interspeech 2013*, 2013, pp. 1766–1770.

[22] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15–32, Apr. 2019.

[23] I. B. Mahmoud, E. Sarkar, M. Manser, and M. Magimai.-Doss, "Feature Representations for Automatic Meerkat Vocalization Classification," in *4th Intl. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)*, 2024.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," arXiv:1912.01703, Dec. 2019.

[25] F. Pedregosa, 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, and A. L. Giraud, "Speech encoding by coupled cortical theta and gamma oscillations," *eLife*, vol. 4, no. MAY, pp. 1–45, May 2015.

[27] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Science Advances*, vol. 5, no. 11, p. eaay6279, Nov. 2019.

[28] A. L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neuroscience*, vol. 15, no. 4, pp. 511–517, Mar. 2012.

[29] O. Ghitza, "The theta-syllable: A unit of speech information defined by cortical function," *Frontiers in Psychology*, vol. 4, no. MAR, 2013.