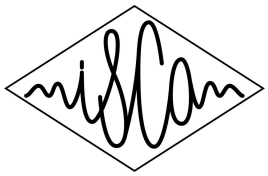

Proceedings of the 4th International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots

VIHAR 2024

Kos, Greece, 6 (onsite) and 9 (online) September 2024

Published by:
Ricard Marxer

ISBN: 978-2-9562029-3-6

Credits:
Editors: Marius Miron, Ricard Marxer
Cover photo: Arne Müseler / www.arne-mueseler.com, via Wikimedia Commons, Recolored and cropped by Ricard Marxer, CC BY-SA 3.0 DE

Workshop took place in Kos, France — September 6 (onsite) and 9 (online), 2024

Published online at http://vihar-2024.vihar.org/ — October 7, 2024

Workshop Organisation

Organising Committee

Marius Miron Earth Species Project

Yossi Yovel Tel Aviv University

Sara Keen Earth Species Project

Eliya Nachmani Google

Paola R. Peña University College Dublin

Björn Schuller Technical University of Munich

Olivier Pietquin Earth Species Project

Steering Committee

Roger K. Moore Sheffield University

Dan Stowell Tilburg University

Angela Dassow Carthage College

Elodie F. Briefer University of Copenhagen

Ricard Marxer University of Toulon

Scientific Committee

Anna Zamansky
Ilyena Hirskyj-Douglas
Sara Keen
Vincent Lostanlen
Leah Govia
Logan James
Maddie Cusimano
Julie Patris
Benjamin Hoffmann
Jennifer Cunha
Ricard Marxer
David Robinson
Ines Nolasco
Masato Hagiwara
Ondřej Cífka
Yossi Yovel
Pedro Ramoneda
Alinta Krauth
Jen-Yu Liu
Luis Joglar-Ongay
Paul Best
Eliya Nachmani
Lonce Wyse
Angelica Lim
Xavier Favory
Hassan Zaal
Marius Miron

Workshop supported by

Conference Program

Keynotes

6 September 2024

9 September 2024

The ape and the first word: Advances in deciphering language evolution

Adriano Lameira, ApeTank, Department of Psychology, University of Warwick, UK

Biography

Dr. Adriano Lameira is an Associate Professor and UK Research & Innovation Future Leaders Fellow at the Department of Psychology, University of Warwick, UK, where he leads the ApeTank, a research lab dedicated to the study of the origins of human behaviour and mind, with a focus on shedding light on language origins, dance and music evolution, and the precursors of imagination. He obtained his PhD at the University of Utrecht (Netherlands), followed by a Junior Research Fellowship at Durham University (UK) and a Marie Skłodowska-Curie Fellowship at the University of St Andrews (UK) before settling in Warwick, a lush campus-based university set in the English Shakespearean countryside and a spin-off of the neighbouring University of Oxford. Adriano and his team study great ape communication, cognition and cultures both in the wild (the peat-swamps of Borneo and low mountain rainforests of Sumatra that harbour the last remaining populations of wild orangutans) and in captivity across Europe and American zoo, complemented with comparative research with children at the department's babylab. Beyond tackling fundamental questions about the nature and evolution of human mind's building blocks, the ApeTank is committed to using new research methods and evidence for advanced cogno-communicative capacities in great apes to (i) improve primate welfare & husbandry in captivity, (ii) advocate primate conservation & protection in the wild, (iii) inform superior bio-inspired computer models and AI applications and (iv) advise stakeholders and law-makers.

Abstract

Why, within a natural world teeming with examples of remarkable communication, has language only evolved in one lineage? Language fundamentally transformed how our species transmits information and knowledge, changing the face of the planet. Yet its origins remain obscure. Language doesn't fossilize, and thus, won't ever be unearthed from an archeological dig. To decipher the puzzle of language evolution, one is bestowed with the task of first deciphering the vocal communication systems of our closest living relatives – nonhuman great apes – the best living models to study and understand the form and function of the precursor system used directly by our first verbal ancestors. Here, I will present some of the most recent strides and findings on the structure and combinatorics of the vocal system of wild orangutans, the only great ape besides humans to combine consonant-like and vowel-like calls into word-sized combination, which in turn combine into sentence-long strings. These new strands of evidence unveil behavioral feedstock for the emergence of several features and capacities classically considered uniquely human or linguistic, challenging notions of a recent all-or-nothing quasi-mystical event at the origin point of language. Instead, each new discovery supports an ape-human vocal-verbal continuum deeply rooted in the communicative, cognitive and cultural capacities of ancient hominids, as predicted by evolution by means of natural selection. Shared ancestry between the communication systems of great apes and humans highlights the difficulty of decoding one system without identifying links and parallels to another closely related and well-understood system. In this sense, language serves as a Rosetta Stone of sorts for deciphering great ape communication. This raises the possibility that some animal communication systems may forever remain undecipherable due to a long, independent evolutionary path that has made them too dissimilar to any system humans might eventually understand. Similar to how the Rongorongo script of Easter Island and the Linear A script of Crete remain unsolved, some animal communication systems may forever elude human understanding. Evolution should, thus, inform and guide all efforts to decipher animal communication.

Decoding animal acoustic communication based on human language and music

Marisa Hoeschele, Austrian Academy of Sciences (OeAW) | ÖAW · Acoustics Research Institute

Biography

Marisa Hoeschele completed a PhD in Psychology with a specialization in Comparative Cognition and Behaviour at the University of Alberta in Edmonton, Alberta, Canada. She now leads the Biology Cluster at the Acoustics Research Institute of the Austrian Academy of Sciences, an interdisciplinary research-only institution with researchers from all different disciplines (e.g., mathematics, phonetics, physics, psychology, machine learning, biology) to address fundamental questions in acoustics. Her own research group, the Musicality and Bioacoustics group, studies parallels between humans and other animals in terms of their acoustic perception and production.

Abstract

Language and music are both critical aspects of human life and seem to emerge spontaneously in all human societies. This spontaneous emergence of language and music suggests that both have roots in human biology. Because we share biological origins and behavioral similarities with other species, we can use cross-species studies to learn more about both the roots of human language and music, and the purpose, depth, and meaning of non-human species-specific sounds. However, a lot of what we know about humans is based on anecdotal and subjective knowledge of our own minds, behaviours, and practices. If we want to make comparisons between humans and other animals, it is critical that we study humans the same way we study any other species. By doing so, we gain powerful insight into 1) how we might advance the study of the behaviours of other species, and 2) what combination of traits is required to make a species human-like. I will present two examples of how we have learned more about humans and made steps towards decoding animal communication by considering humans an acoustic animal. First, I will show how typical bioacoustic methods would fail if we apply them to human vocalizations, and how taking this into account has led us to identify units that are similar to consonants and vowels in budgerigar (Melopsittacus undulatus; a small parrot species) vocalizations. Second, I will show how comparative research supports the idea that cross-culturally shared properties of musical scales likely stem from the physics of vocal sounds.

Jungle chatter:
Eavesdropping on prey alarm calls to detect the presence of predators

Angela Dassow[1], *Arik Kershenbaum*[2], *Bethany Smith*[3], *Andrew Markham*[4], *Casey Anderson*[5], *Riley McClaughry*[5], *Ramjan Chaudhary*[6], *Holly Root-Gutteridge*[7]

[1]Department of Biology, Carthage College, WI, USA
[2]Girton College and Department of Zoology, University of Cambridge, England
[3]Institute of Zoology, Zoological Society of London, England
[4]Department of Computer Science, University of Oxford, England
[5]VisionHawk Films, MT, USA
[6]The Den Adventurer's, Nepal
[7]School of Life Science, University of Lincoln, England

adassow@carthage.edu

Abstract

Encounters between humans and large predators result in several human deaths each year, which erodes local support for large predator conservation, despite many large predator species being endangered or critically endangered. We describe how eavesdropping on the communication of different species can help to detect tigers and alert people to their presence. While tigers sometimes produce loud and distinctive roars, they do not produce sufficient vocal events to be tracked acoustically. However, tigers pose a danger to other animals and these animals reliably produce alarm calls in their presence, and forest rangers commonly use these alarm calls to locate tigers in the field. We tested the responses of prey species in the Terai region of Nepal to an artificial tiger model and used automated detection of chital deer (*Axis axis*) alarm calls to generate a heatmap of tiger presence, which can be used to alert villagers of areas of increased risk of tiger encounters.

Index Terms: alarm calls, automatic detection, human-wildlife conflict, interspecific eavesdropping, tigers

1. Introduction

Tigers (*Panthera tigris*) are a keystone species that are fundamental to maintaining balance and supporting biodiversity in ecosystems but are listed as Endangered throughout their range [1]. Although tiger populations have been increasing in Nepal in recent years, this recovery is threatened by escalating conflicts with humans over attacks on humans and livestock, leading to 32 human fatalities between 2007-2014 [2]. Local populations that use the forest as a subsistence resource suffer the bulk of tiger attacks but remain generally supportive of tiger conservation. Knowing precisely where tigers are present could help prevent some of these conflicts and in turn help to facilitate coexistence between humans and tigers. However, due to their solitary nature and large home ranges, tigers are notoriously difficult to locate and track. Existing options for monitoring tiger movements include GPS tracking and the use of camera traps, but GPS collars are invasive and costly, requiring capturing and sedating individual tigers, and camera traps can only survey a small area in front of individual cameras so can easily miss detecting tigers when they are present. As such, non-invasive methods that can monitor tiger presence over large spatial scales and alert in near real-time, would be invaluable to help those living alongside tigers to make informed decisions about when it is safest to enter the forest or when they need to move or guard their livestock.

The forest is, however, home to many prey species that have evolved natural vigilance behaviors to protect against tiger predation. In particular, certain species of deer and monkeys, particularly chital deer (*Axis axis*), gray langurs (*Semnopithecus schistaceus*), and rhesus macaques (*Macaca mulatta*), issue loud alarm calls when a predator is spotted [3], [4], and this complex interspecific assemblage of vocalizations can be used to assess the risk of tiger presence. In fact, nature guides and forest rangers routinely listen for prey alarm calls to alert them to the presence of large predators. Our ongoing project aims to combine this local knowledge and evolutionary behavior of animals with advancements in technology to create risk maps of tiger presence in an area. By eavesdropping on the alarm calls of prey species and using their naturally evolved response to predators, automating and computerizing detections, and translating this into a central digitized interface where tiger risk can be visualized, we can convey this information to at-risk populations such as local villagers foraging in the forests.

Passive acoustic monitoring (PAM) is a growing field in wildlife research, providing the ability to gather large amounts of data on animal behavior and distribution in a non-invasive way [5]. Autonomous recording devices are deployed across a landscape and record audio continuously if necessary, or at scheduled times of day. Previous studies have demonstrated

the effectiveness of PAM in monitoring landscape use, interspecific interactions, and conservation priorities for a wide range of terrestrial species [6], [7], [8]. However, acoustic monitoring is only effective where a species vocalizes regularly, and at a volume that makes detection on a grid of monitoring devices realistic. Although social predator species such as wolves vocalize to maintain social links between individuals [9], many predator species remain largely silent in an attempt to avoid detection by prey. Tigers, in particular, vocalize loudly, but only in territorial contexts, and only rarely [10]. Therefore, tracking tigers using PAM must rely on the vocal responses of other species to tiger presence.

Here we present the results of a development project in which we deployed PAM devices with onboard automatic detection of chital deer alarm calls in forests in the Terai region of southern Nepal, with a high risk of tiger-human conflict. Each device, based on the open-source CARACAL hardware [11], also uses a sub-Gigahertz radio to communicate with a base station, which gathers alarm vocalization events and generates a heat map indicating the risk of tiger presence, based on the frequency and intensity of alarm calls.

To determine whether prey species alarm calls are reliably generated in response to tiger presence, we presented an artificial tiger model to chital deer, grey langurs, and rhesus macaques, and recorded their vocal responses. We also monitored the vocal activity of these species in the absence of tiger presence, and our findings strongly suggest that prey alarm calls can be used as a reliable indicator of the prey species' perception of predator risk.

2. Methods

The study was carried out in the Dalla (28.40421° N, 81.22958° E) and Khata (28.36813° N, 81.21630° E) community forests around Bardia National Park in southern Nepal (Figure 1). Community forests provide the opportunity for villagers to perform traditional foraging tasks such as collecting firewood and grazing livestock, which maintains a productive balance in the natural ecosystem [12], but exposes them to injury from wildlife. Management of the forests is performed by local trained rangers employed by the national Forestry Department. Rangers are also tasked with monitoring for the presence of tigers and other potentially dangerous species such as leopards (*Panthera pardus*), rhinos (*Rhinocerous unicornis*) and elephants (*Elephas maximus*). The forests are dominated by sal (*Shorea robusta*), and kamala (*Mallotus philippensis*) [13], with various grass species (e.g. *Tripidium bengalense*) collected by villagers for traditional uses [14].

Figure 1. *Map of the study area, showing (A) Dalla Community Forest, and (B) Khata Community Forest.*

We deployed 10 CARACAL acoustic recording devices in each of the forests during December 2023 and March 2024. The CARACAL devices are equipped with four MEMS microphones for beamforming to determine direction of arrival of sound signals, and integrated GPS clock synchronization for localization of sound sources using multilateration. For this project, we added an 868 MHz LoRa radio transmitter (iLabs Challenger RP2040, Invector Labs, Tomelilla, Sweden) that transmitted information on alarm detections every 30 seconds (Figure 2).

Figure 2. *A CARACAL acoustic recording device deployed in the Dalla community forest.*

All three focal prey species produce predator alarm calls. The chital deer alarm call [3] is loudest and most characteristic (Figure 3), being a strongly modulated narrowband chirp between 0.75 and 1.25 kHz.

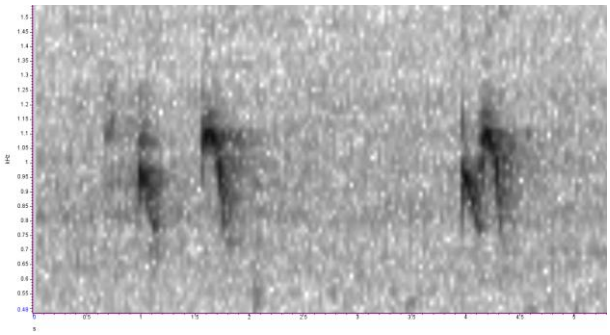

Figure 3. *Spectrographic representation of chital deer alarm calls, showing their characteristic 0.75 – 1.25 kHz chirp.*

Rhesus macaque alarm calls [4] (Figure 4) are noisy, broadband sequences lasting 1-5 seconds and given repeatedly. Each call is a series of short pulses (about 200ms).

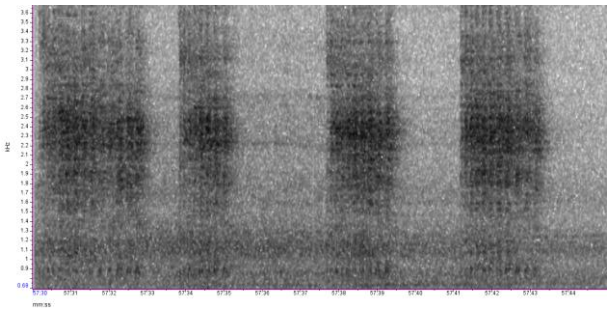

Figure 4. *Rhesus macaque alarm calls, being a series of short pulses, repeating in sets of 1-5 seconds length.*

Grey langurs produce shorter, single alarm calls [3] (Figure 5), each about 200ms in length, very broadband (with significant energy well beyond the 8 kHz Nyquist limit of our recordings), but with a concentration of energy in a chirp at similar frequencies to the chital call.

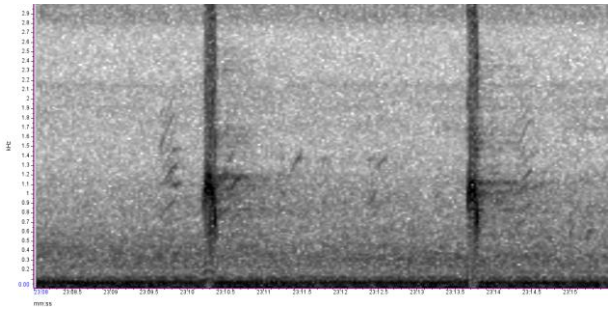

Figure 5. *Langur alarm calls, very broadband short bursts.*

We verified that these alarm calls were given in response to tigers by presenting wild chital deer with a tiger model in the form of a faux tiger skin (https://www.vidaxl.co.uk/e/vidaxl-tiger-carpet-plush-144-cm-brown/8718475509172.html), draped over one of the researchers. Previous studies have shown that prey animals respond strongly to artificial predator models [15].

For the experimental protocol, we first identified groups of the focal prey animal who were showing normal (non-stressed) behavior and who were close to the side of the road. One of the researchers would then descend from the vehicle and hide while putting on the tiger costume. They would then slowly approach the animals through the forest, attempting to imitate the motion of a tiger. (Figure 6). A presentation was considered successful if the animals did not bolt before seeing the tiger model. The predator presentation continued for 15 minutes, or until the prey animals had left the area. During this time, another researcher was recording the prey animal responses on a DR-44WL handheld recording device (TASCAM, CA, USA) with an AT8035 shotgun condenser microphone (Audio-Technica, OH, USA).

Figure 6. *Presentation of a faux tiger model to chital deer.*

3. Results

In total, we attempted 61 presentations to prey groups. In some of these cases (19.6%), the animals spotted us preparing the experiment and fled without making any vocalizations, leading to the presentation being aborted. However, we succeeded in carrying out 7 predator presentations to chital deer, 5 to macaques, and 2 to langurs. In each of these successful cases (100%) where the animals saw the tiger model before bolting, they produced characteristic alarm calls.

In addition to the 12 aborted predator model presentations, there were numerous occasions in which the prey animals encountered the researchers moving through the forest without the tiger costume. We succeeded in carrying out 14 human presentations to chital deer, 10 to macaques, and 11 to langurs. In all but one of these cases, the animals remained silent. There was one instance of a chital deer calling to a human, but in that case, the deer also saw the rest of the research team with shotgun microphones and camera equipment.

In total, the CARACAL devices recorded approximately 2800 hours of audio, of which, approximately 1300 hours has been analysed manually to date. Using the CARACAL devices recording ambient sound passively (i.e. without predator model presentations), chital alarm calls were recorded at a rate of approximately 0.1 per hour during the daytime hours, macaques 0.14 per hour, and langurs 0.008 per hour.

4. Discussion

Our results showed that a predator model is an effective way to elicit alarm calls in the target prey species, and that the alarm calls are highly specific to predator presence, as determined by the artificial model. This is a strong indication that interspecific eavesdropping allows humans to infer predator presence from prey alarm calls.

We did not have the opportunity to witness prey interactions with real tigers, however this is not surprising as such events are difficult to predict and infrequent. Nonetheless, the alarm calls generated by the predator model are acoustically very similar to opportunistic recordings made by local wildlife guides of prey alarm calls heard in the case of real encouters of prey with predators. Moreover, the use of prey alarm calls by rangers and wildlife guides is a strong indication that these are reliable signs of predator presence.

The absence of alarm calls in response to humans, while not unexpected given the specificity of many animal predator alarm calls [16], [17], [18], is an encouraging sign when designing a predator warning system for local people. The system must detect areas that prey animals consider high risk from predators, while not identifying areas as “dangerous” where humans themselves are the only potential predators present.

Specificity of the prey alarm calls to tigers is difficult to determine. Leopards also prey on both monkeys and deer, and it is likely that the prey animals do not distinguish in their alarm calls between different types of big cats. However, as leopards also pose a threat to humans working in the forest, the alarming of prey species in response to leopard presence is an advantage, rather than a disadvantage.

Preliminary analysis of the CARACAL recordings indicates that alarm calls are relatively infrequent, meaning that any warning system based on prey alarm calls is unlikely to be rendered unhelpful by being overwhelmed by large numbers of calls. The ability of the CARACAL to localize the sound source is an additional potential strength, as it would allow alarms to be further validated by their spatial correlation between recording devices.

Any future warning system will rely on effective automatic detection of alarm call on the CARACAL devices. Our project

is currently testing such a system, which, combined with the remote notification through sub-GHz radio, will allow the deployment of a widespread and novel tool to prevent loss of human life and enhance conservation efforts.

5. Conclusions

We have shown using a faux tiger model that three prey species - chital deer, grey langurs, and rhesus macaques - respond reliably to perceived tiger presence with distinctive alarm calls, which can be monitored and interpreted by humans to build a broadly deployed warning system to identify areas of high tiger risk, and to warn local villagers of areas of the forest to avoid.

In the next implementation of our system, we plan to deploy a first of its kind, fully-operational pilot system in southern Nepal, which will use the communication of other animal species to inform humans of potentially dangerous predator presence in the forest.

6. Acknowledgements

We would like to thank the Big Cat Initiative of the Great Plains Foundation for their generous ongoing support in funding this project. The Nepal Forestry Department has been supportive in allowing access to the community forests, and we would particularly like to thank the local rangers, Tengchu and Sukhna who helped us to deploy the equipment and kept us safe from real tigers. Thanks also to Suman, the proprietor of Freedom Bar.

7. References

[1] J. Goodrich *et al.*, 'Panthera tigris. The IUCN Red List of Threatened Species', 2022. [Online]. Available: https://dx.doi.org/10.2305/IUCN.UK.2022-1.RLTS.T15955A214862019.en.

[2] R. Dhungana, T. Savini, J. B. Karki, M. Dhakal, B. R. Lamichhane, and S. Bumrungsri, 'Living with tigers Panthera tigris: patterns, correlates, and contexts of human–tiger conflict in Chitwan National Park, Nepal', *Oryx*, vol. 52, no. 1, pp. 55–65, Jan. 2018, doi: 10.1017/S0030605316001587.

[3] P. N. Newton, 'Associations between Langur Monkeys (Presbytis entellus) and Chital Deer (Axis axis): Chance Encounters or a Mutualism?', *Ethology*, vol. 83, no. 2, pp. 89–120, 1989, doi: 10.1111/j.1439-0310.1989.tb00522.x.

[4] J. M. B. Fugate, H. Gouzoules, and L. C. Nygaard, 'Recognition of rhesus macaque (Macaca mulatta) noisy screams: evidence from conspecifics and human listeners', *American Journal of Primatology*, vol. 70, no. 6, pp. 594–604, 2008, doi: 10.1002/ajp.20533.

[5] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, and D. Llusia, 'Terrestrial Passive Acoustic Monitoring: Review and Perspectives', *BioScience*, vol. 69, no. 1, pp. 15–25, Jan. 2019, doi: 10.1093/biosci/biy147.

[6] A. Kershenbaum, J. L. Owens, and S. Waller, 'Tracking cryptic animals using acoustic multilateration: A system for long-range wolf detection', *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1619–1628, Mar. 2019, doi: 10.1121/1.5092973.

[7] H. Root-Gutteridge *et al.*, 'Not afraid of the big bad wolf: calls from large predators do not silence mesopredators', Jan. 2024, Accessed: Jan. 14, 2024. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/362859

[8] D. J. Clink, H. Bernard, M. C. Crofoot, and A. J. Marshall, 'Investigating Individual Vocal Signatures and Small-Scale Patterns of Geographic Variation in Female Bornean Gibbon (Hylobates muelleri) Great Calls', *Int J Primatol*, vol. 38, no. 4, pp. 656–671, Aug. 2017, doi: 10.1007/s10764-017-9972-y.

[9] F. H. HARRINGTON, 'Chorus Howling by Wolves: Acoustic Structure, Pack Size and the Beau Geste Effect', *Bioacoustics*, vol. 2, no. 2, pp. 117–136, Jan. 1989, doi: 10.1080/09524622.1989.9753122.

[10] E. Walsh, L. Wang, D. Armstrong, T. Curro, L. Simmons, and J. McGee, 'Acoustic Communication in *Panthera tigris*: A Study of Tiger Vocalization and Auditory Receptivity', *Durham School of Architectural Engineering and Construction: Faculty Publications*, Jan. 2003, [Online]. Available: https://digitalcommons.unl.edu/archengfacpub/38

[11] M. Wijers, A. Loveridge, D. W. Macdonald, and A. Markham, 'CARACAL: a versatile passive acoustic monitoring tool for wildlife research and conservation', *Bioacoustics*, vol. 30, no. 1, pp. 41–57, Jan. 2021, doi: 10.1080/09524622.2019.1685408.

[12] H. Nagendra, 'Tenure and forest conditions: community forestry in the Nepal Terai', *Environmental Conservation*, vol. 29, no. 4, pp. 530–539, Dec. 2002, doi: 10.1017/S0376892902000383.

[13] R. Joshi, R. Chhetri, and K. Yadav, 'Vegetation Analysis in Community Forests of Terai Region, Nepal', *Int. J. Environ.*, vol. 8, no. 3, pp. 68–82, Dec. 2019, doi: 10.3126/ije.v8i3.26667.

[14] K. Brown, 'The political ecology of biodiversity, conservation and development in Nepal's Terai: Confused meanings, means and ends', *Ecological Economics*, vol. 24, no. 1, pp. 73–87, Jan. 1998, doi: 10.1016/S0921-8009(97)00587-9.

[15] A. Dassow, 'Exploring the Interior Structure of White-handed Gibbon and Rat Vocal Communication - ProQuest', PhD, University of Wisconsin, Madison, 2014. Accessed: Jan. 14, 2024. [Online]. Available: https://www.proquest.com/openview/40b690edf5d6cc38431a0bc50bef07a8/1?pq-origsite=gscholar&cbl=18750

[16] C. Fichtel and P. M. Kappeler, 'Anti-predator behavior of group-living Malagasy primates: mixed evidence for a referential alarm call system', *Behav Ecol Sociobiol*, vol. 51, no. 3, pp. 262–275, Feb. 2002, doi: 10.1007/s00265-001-0436-0.

[17] B. Walton and A. Kershenbaum, 'Heterospecific recognition of referential alarm calls in two species of lemur', *Bioacoustics*, vol. 28, no. 6, pp. 592–603, Nov. 2019, doi: 10.1080/09524622.2018.1509375.

[18] J. M. Macedonia and C. S. Evans, 'Essay on Contemporary Issues in Ethology: Variation among Mammalian Alarm Call Systems and the Problem of Meaning in Animal Signals', *Ethology*, vol. 93, no. 3, pp. 177–197, 1993, doi: 10.1111/j.1439-0310.1993.tb00988.x.

On the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis

Eklavya Sarkar[1,2]*, Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Switzerland
[2]Ecole polytechnique fédérale de Lausanne, Switzerland
{eklavya.sarkar, mathew}@idiap.ch

Abstract

Marmoset monkeys encode vital information in their calls and serve as a surrogate model for neuro-biologists to understand the evolutionary origins of human vocal communication. Traditionally analyzed with signal processing-based features, recent approaches have utilized self-supervised models pre-trained on human speech for feature extraction, capitalizing on their ability to learn a signal's intrinsic structure independently of its acoustic domain. However, the utility of such foundation models remains unclear for marmoset call analysis in terms of multi-class classification, bandwidth, and pre-training domain. This study assesses feature representations derived from speech and general audio domains, across pre-training bandwidths of 4, 8, and 16 kHz for marmoset call-type and caller classification tasks. Results show that models with higher bandwidth improve performance, and pre-training on speech or general audio yields comparable results, improving over a spectral baseline.
Index Terms: bioacoustics, call-type and caller classification, speech and audio, bandwidth.

1. Marmoset Vocalizations

Non-human vocal communication, such as bioacoustics, i.e. the study of animal vocalizations, is rapidly advancing through the advent of machine learning and the correlated progress in human speech processing [1]. Common marmosets (*Callithrix jacchus*) are of particular interest due to their highly vocal nature, acoustically diverse call repertoire, and acute auditory capabilities. Their extensive vocalizations are rooted in a complex social system, and are thus able to encode a range of information, such as group affiliation, sex [2], population, dialect [3], and even individual caller identity [4, 5], over a number of social and emotional states [6, 7]. Their remarkable vocal adaptability also allows them to modify the duration [8], intensity [9], complexity [10], or timing [11] of their calls. These vocal characteristics align them closely with human speech properties, such as care-giving to infants, turn-taking [12], and categorical perception of sounds [13], and make them into a well-suited surrogate model for understanding the vocal communication of non-human primates among biologists [14] and neuroscientists [15].

In the literature, the automatic analysis of marmoset vocalizations, i.e. call-type, caller identity, or sex classification, has been conducted by leveraging signal processing features alongside traditional machine learning classifiers. Early work demonstrated that k-NN, SVM, and optimal path forest classifiers achieved notable success over multilayer perceptrons (MLPs), Adaboost, and logistic regression, especially with small, specific datasets [16]. Research exploring a variety of audio and spectral feature representations, such as signal energy, zero crossing rate, spectral rolloff, and MFCCs, indicated that integrating different feature could enhance the system's performance on synthetically augmented vocal datasets [17]. Recent studies have also explored leveraging deep learning based techniques. Using convolutional neural networks to process spectrograms for simultaneous vocalization detection, call-type classification, and caller identification was found to outperform separate models for each task [18]. Statistics of log-mel filter-bank energies used as input for recurrent neural networks (RNNs) were shown to improve the detection and classification of calls over SVM or MLPs [19]. Self-supervised learning (SSL) frameworks, which create surrogate labels from the data, were used with the aim of leveraging the large quantities of unlabeled data for birdsong detection [20] and bioacoustic event detection [21].

A novel study demonstrated that neural representations derived from models pre-trained on human speech through SSL could distinguish individual marmoset caller identities [22]. The authors argued that SSLs only learn the intrinsic structure of the unlabeled input signal, typically through a masking-based pre-text training task, to capture essential information independently of any domain-specific knowledge, such as human speech production, and thus can be cross-transferred across different acoustic domains, such as bioacoustics. Building on these findings, our paper investigates the utility and limitations of such pre-trained foundation models for the purpose of marmoset call analysis, with a focus on the following key points:

1. **Classification**: We investigate whether such models can be effectively leveraged for marmoset call analysis tasks, namely call-type and caller classification, which, to the best of our knowledge, has not yet been demonstrated. Additionally, while [22] focused solely on caller detection in a binary framework, we extend the scope to a multi-class approach.
2. **Bandwidth**: Given that these models are typically pre-trained at a bandwidth of 8 kHz, we address their mismatch with the biological vocalization and auditory range of marmosets, predominantly concentrated in the 5–10 kHz spectral region [23], and thus evaluate their capability to accurately represent marmoset calls. By examining models pre-trained across varying bandwidths, we aim to evaluate their effectiveness in adequately representing marmoset calls, and seek to clarify how model bandwidth influences their classification.
3. **Pre-training domain**: It remains unclear how models pre-trained on human speech compare to trained on other acoustic domains for accurately capturing marmoset call characteristics. We examine representations produced by different pre-training sources, such as human speech and general audio, across supervised and self-supervised learning frameworks, against a spectral baseline to identify the most suitable pre-

github.com/idiap/speech-utility-bioacoustics

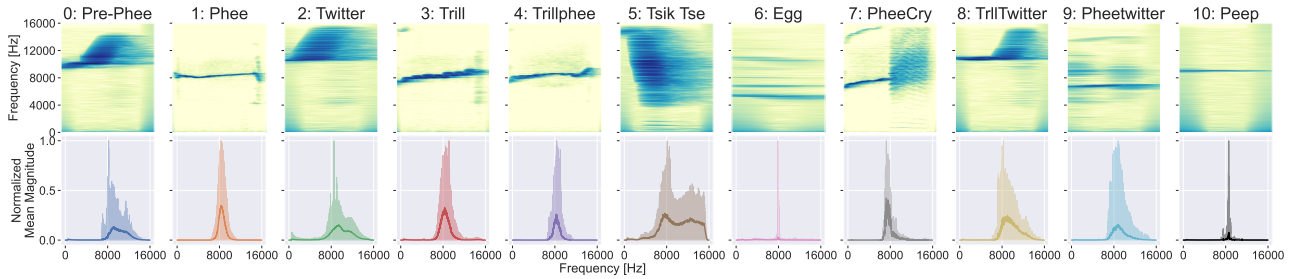

Figure 1: *Marmoset vocalizations with a 16 kHz bandwidth.* Top: *Spectrograms of a* single *call-type vocalization.* Bottom: *The mean spectrum for* all *vocalizations per call-type across the dataset, normalized. Shaded areas indicate ± 1 std from the mean spectrum.*

training source for cross-domain bioacoustic signal analysis. The rest of the paper is organized as follows: Section 2 gives the study's methodology, section 3 & 4 present a call similarity and classification analysis. Section 5 finally concludes the paper.

2. Methodology

2.1. Dataset and Tasks

For our study, we used the InfantMarmosetsVox (IMV) dataset [22], which contains 72,921 labelled marmoset vocalization segments (totalling to 464 minutes), sampled at 44.1 kHz, across ten marmoset individuals and contains eleven marmoset call-types. Table 1 presents the data distribution in function of the call-types and callers. For our experiments, we divide the dataset into a *Train*, *Val*, and *Test* sets, following a random 70:20:10 split. We denote call-type and caller identity multi-class classification as CTID and CLID respectively.

Table 1: *InfantMarmosetsVox dataset statistics.*

ID	Call-type	Count	Caller ID	Count
0	Peep (pre-phee)	1283	0	15521
1	Phee	27976	1	8648
2	Twitter	36582	2	13827
3	Trill	1408	3	5838
4	Trillphee	728	4	5654
5	Tsik Tse	686	5	3522
6	Egg	1676	6	4389
7	Pheecry (cry)	23	7	2681
8	TrllTwitter	293	8	6387
9	Pheetwitter	2064	9	6454
10	Peep	202	-	-
	Total	**72921**	**Total**	**72921**

Figure 1 gives the visualizations of all call-types as well the density distribution of the spectrums across the entire dataset. Frequencies below 500 Hz are nullified purely for visualization to eliminate any low-frequency noise. We can observe that information starts at around 7-8 kHz for most calls in this dataset.

2.2. Models and Feature Representations

For our study, we select four distinct frameworks for feature representations $\mathcal{F}$: hand-crafted (HC) features derived through signal processing techniques, neural representations obtained via self-supervised learning (SSL), pre-trained on either human speech or general audio, and features generated through supervised learning (SL) models pre-trained on general audio. These frameworks are summarized in table 2. We extract the features from these frameworks by giving the marmoset calls as input.

Table 2: *# Parameters P and feature dimension D of selected models, pre-trained on AudioSet (AS) or LibriSpeech (LS).*

$\mathcal{F}$	Corpus	P	D	Type
C22 [24]	-	-	24	HC
WavLM [25]	LS	94.38M	1536	SSL
BYOL [26]	AS	5.32M	2048	SSL
PANN [27]	AS	8.08M	2048	SL

Hand-crafted: The Highly Comparable Time-Series Analysis (HCTSA) framework, used for interpreting diverse time series data, extracts 7700 features through signal processing methods, such as LPC [28]. It has been applied to diverse tasks such as birdsong discrimination [29], ecosystem monitoring [30], and marmoset caller identification [5]. Despite its broad applicability, HCTSA's computational demands and feature redundancy are significant limitations. The CAnonical Time-series CHaracteristics (Catch22/C22), a steamlined subset of HCTSA, provides high performance with minimal redundancy across numerous classification problems [24]. We extend this feature set to a final dimension of $D = 24$ by appending the first and second order statistics, and use it as our spectral baseline.

SSL pre-trained on human speech: Following the approach in [22], we use feature representations from SSL models trained on human speech, extending it to both call-type and caller identity classification. We select the WavLM base model, pre-trained on the 960-hour LibriSpeech dataset, based on its effectiveness in marmoset call detection as well as its versatility in speech processing tasks as demonstrated in the SUPERB challenge [31]. For each layer, feature representations of length 768 are extracted for each frame. Then, they are transformed into fixed-length utterance-level representations by computing and aggregating first and second order statistics across the frame-axis, resulting in a final representation of length $D = 1536$.

SL pre-trained on general audio: Expanding marmoset call analysis literature, we utilize embeddings from models pre-trained on the AudioSet (AS) dataset, which includes audio event classes such as environmental sounds, musical instruments, and human and animal vocalizations. Specifically, we choose the *AudioNTT2020* model from the BYOL-A architecture [26], extracting embeddings from its final fully connected layer of length $D = 2048$. Inputs are processed into log-mel spectrograms, adhering to the spectral parameters detailed in the original study, i.e. a 8 kHz bandwidth, 64 ms window size,

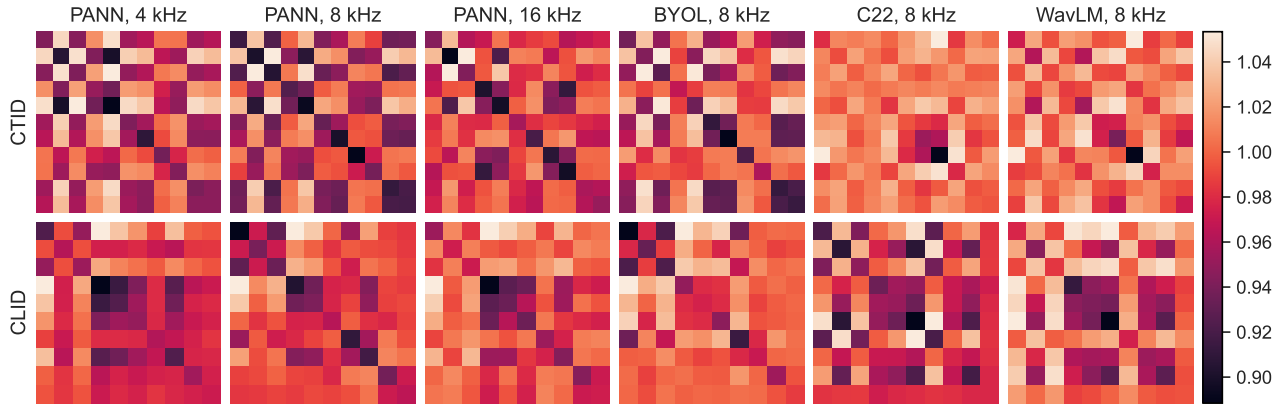

Figure 2: *Pairwise mean cosine distances matrices for features $\mathcal{F}$ at different bandwidths for call-types (CTID) and callers (CLID). Diagonal entries represent intra-class distances, and off-diagonal the inter-class. Darker regions indicate higher similarity.*

10 ms hop size, and 64 mel bins spanning from 60 to 7800 Hz.

SL pre-trained on general audio: We further investigate feature extraction from large-scale networks pre-trained for general audio pattern recognition. The *CNN14* model from the *PANN* network [27] is chosen, with pre-trained weights applied at three different bandwidths: 4, 8, and 16 kHz. This model employs a balanced sampling strategy across AudioSet's sound classes and also processes input vocalizations into spectrograms to extract log-mel filterbanks. For a bandwidth of 16 kHz, window and hop sizes are set to 1024 and 320 samples, respectively, and proportionally halved for 8 and 4 kHz. The model utilizes 64 mel bands, spanning from 50 Hz and to the Nyquist frequency. Embeddings of length $D = 2048$ are extracted from the linear layer preceding the final classification layer.

3. Call Similarity Analysis

This section presents a pairwise similarity analysis of the selected features on the *Train* set to identify any discernible patterns or correlations for given the vocalizations. Specifically, we investigate how variations in the bandwidth of the pre-trained models affect the similarity distribution of intra-class embeddings, and examine any distinctions between models pre-trained on speech against general audio. To compare the features, which are high-dimensional vectors, we use the cosine distance defined as $\text{sim}(\boldsymbol{x_1}, \boldsymbol{x_2}) = 1 - (\boldsymbol{x_1} \cdot \boldsymbol{x_2} / \|\boldsymbol{x_1}\| \cdot \|\boldsymbol{x_2}\|)$, bounded in $[0, 2]$. Two features are identical when their cosine distance is 0, orthogonal at 1, and opposite at 2. For WavLM, we select the first layer, and only use the first half of the extracted features, corresponding to the mean values averaged frame-wise.

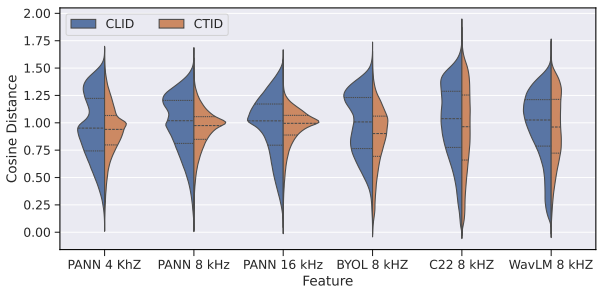

Figure 3: *Distribution of pairwise cosine distances.*

Figure 3 presents the overall distribution of pairwise distances. The distrtibutions are overlapping, centering around a median distance of 1 for all representations, suggesting a lack of clear correlation or similarity within the embeddings generated. Figure 2 further delineates the distributions into distance matrices for each feature set, where diagonal and off-diagonal entries correspond to intra-class and inter-class distances respectively. In an ideal scenario, embeddings from the same call-type or caller would exhibit closer distances, where as embeddings from different classes would have a higher dissimilarity.

We can observe that the models pre-trained on general audio datasets (BYOL and PANN) yield more distinct peaks and diagonals, on figures 3 and 2 respectively, compared to those pre-trained on human speech (WavLM) or the handcrafted baseline (Catch22). This distinction is more pronounced for call-types than for caller identification. This is expected, given that the call-types are spread across caller classes (a caller produces different calls, while a call can come from any caller). Although these patterns indicate some level of class-specific clustering, the distribution of distances largely show that the features are highly orthogonal. The similarity analysis thus indicates minimal feature correlation, and suggests that classifying these vocalizations with a simple linear classifier would be challenging, as there is no clear linear separability between the classes.

4. Classification Analysis

Based on the insights of our similarity analysis, we aim to evaluate the saliency of the extracted representations, and proceed to classify them using a simple, non-linear MLP, for the multi-class classification tasks. We implement three blocks of [Linear, LayerNorm, ReLU] layers, with 128, 64, and 32 number of hidden units respectively, followed by a final linear layer to obtain the posterior probabilities. To evaluate the performance we used Unweighted Average Recall (UAR) as the metric to account for any class imbalance. To obtain robust results, we employ the grid search methodology with *Val* UAR score as the optimization criterion. We train the classifier for 30 epochs with cross-entropy loss, and search for the optimal hyperparameters values of η and batch-size across $2^{[5-9]}$ and [1e-3, 1e-4] respectively for each feature–task permutation on *Train* and *Val*. The optimization consists of Adam and a η-scheduler of factor 0.1 and patience of 10 epochs. Lastly, for WavLM, we classify each of the encoder layers [0–13] to identify the optimal layer.

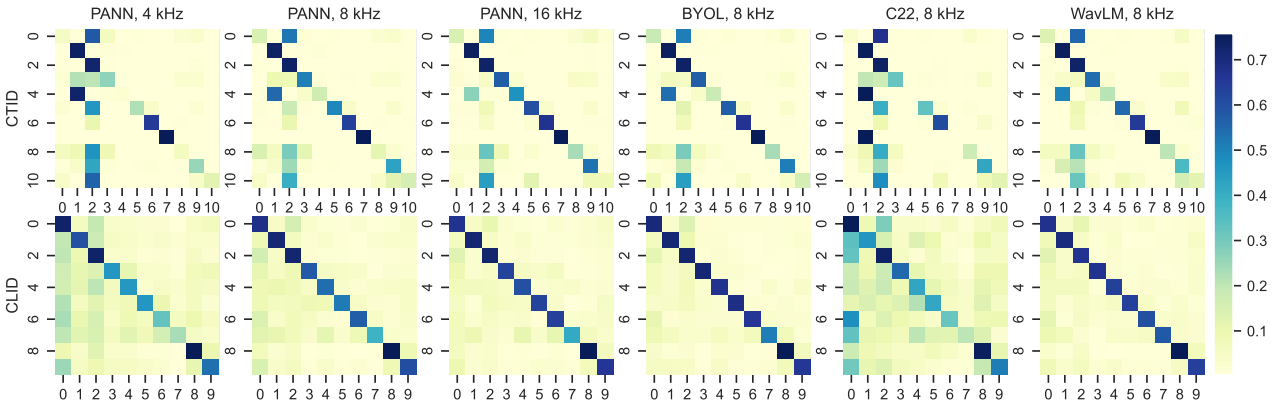

Figure 4: *Normalized confusion matrices with row indices representing true class labels. Darker diagonals signify higher performance.*

Figure 5 presents the layer-wise scores for WavLM, normalized per task to a [0, 1] range. We can observe that the lower layers are clearly much more salient representations for both tasks compared to higher layers. Based on these results, we use the best individual WavLM layers for our two tasks.

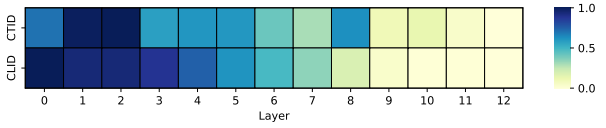

Figure 5: *Layer-wise UAR scores of WavLM features, normalized per task. Darker regions indicate a higher performance.*

Table 3a) summarizes the classification results of the different feature sets at an 8 kHz bandwidth (BW). Random performance is given as 100 over the number of classes. Notably, BYOL features outperform the other features, for both CTID and CLID, despite having fewer parameters than WavLM and PANN, while C22 proves to be the overall weakest representation. WavLM shows the highest difference in performance across tasks. Meanwhile, table 3b) highlights the impact of pre-training bandwidth for salient representations on PANN features. The results clearly show that the bandwidth size correlates directly with the performance, increasing monotonically. Particularly, PANN features at 16 kHz achieve the highest performance across all features and BWs for CTID. BYOL embeddings at 8 kHz notably outperform PANN at 16 kHz for CLID. The best scores for both tasks are also closely matched in value.

Figure 4 shows the classifier's performance through confusion matrices. We can again clearly observe the monotonic improvement in CTID classification performance for PANN features as the bandwidth increases. We also notice a prevalent trend of false positives for call-type ID 2 (Twitter) across all feature sets, especially against IDs 0, 8, and 10, attributable to its high occurrence in the dataset and broad spectral range [32, 33]. The CLID results contain distinctly fewer misclassifications, which aligns with expectations since the call-types are spread among the different callers classes. The exception is C22, which yields the weakest performance. Caller classes with higher data volumes (IDs 0 and 2) perform better compared to the others. Finally, a clear improvement in performance correlated with bandwidth is seen for PANN features, as with CTID.

Table 3: *UAR scores [%] on Test for pre-trained features $\mathcal{F}$. WavLM's best layer's score is given.*

Section	$\mathcal{F}$	BW	CTID	CLID
(a)	Random	-	9.09	10
	C22	8	41.96	35.62
	WavLM	8	59.99	67.47
	BYOL	8	**63.64**	**68.30**
	PANN	8	58.54	56.02
(b)	PANN	4	46.27	41.10
	PANN	8	58.54	56.02
	PANN	16	**69.09**	**65.39**

5. Summary and Conclusion

This paper investigated the utility and limitations of foundations models, pre-trained on human speech or general audio, which have not been demonstrated for marmoset call-type and caller identity multi-class classification. To that end, we conducted and validated two studies across three lines of investigation.

First we conducted a call similarity analysis, which revealed that the features extracted from these models lacked linear separability within or across classes. Then, we conducted a classification study which demonstrated that a non-linear classifier can still achieve substantial performance, and highlighted that a larger bandwidth directly correlates with improved performance. Classification of call-types also appeared to be more sensitive to bandwidth changes than caller identities. Additionally, the pre-training domain of speech and general audio showed comparable performances, with a distinct improvement over handcrafted features. Finally, we obtained close best performance for both call-type and caller classification tasks.

In conclusion, our findings underscore the potential of leveraging pre-trained foundation models for bioacoustic signals, particularly when the model's bandwidth aligns with the biological auditory and vocal range of the studied species. Future collaborative work with biologists and linguistics researchers could explore the biological implications of these results, especially in understanding the evolutionary aspects of marmoset vocal behaviour and their perceptual processing, to bridge the gap between computational models and biological insights in non-human vocal communication research.

6. Acknowledgements

This work was funded by Swiss National Science Foundation's NCCR Evolving Language project (grant no. 51NF40_180888).

7. References

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

[2] J. L. Norcross and J. D. Newman, "Context and gender-specific differences in the acoustic structure of common marmoset (callithrix jacchus) phee calls," *American journal of primatology*, vol. 30(1), p. 37–54, 1993.

[3] Y. Zürcher and J. M. Burkart, "Evidence for dialects in three captive populations of common marmosets (callithrix jacchus)," *International Journal of Primatology*, vol. 38, no. 4, pp. 780–793, 2017.

[4] J. BS, H. DHR, and C. CK, "The stability of the vocal signature in phee calls of the common marmoset, callithrix jacchus," *American journal of primatology*, vol. 31(1), pp. 67–75, 1993.

[5] N. Phaniraj, K. Wierucka, Y. Zürcher, and J. M. Burkart, "Who is calling? optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers," *Journal of Royal Society Interface*, 2023.

[6] G. Epple, "Comparative studies on vocalization in marmoset monkeys (hapalidae)," *Folia Primatol (Basel)*, vol. 8, no. 1, pp. 1–40, 1968.

[7] R. Seyfarth and D. Cheney, "Signalers and receivers in animal communication," *Annual review of psychology*, vol. 54, pp. 145–73, 02 2003.

[8] H. Brumm, K. Voss, I. Köllmer, and D. Todt, "Acoustic communication in noise: regulation of call characteristics in a new world monkey," *Journal of Experimental Biology*, vol. 207, no. 3, pp. 443–448, 01 2004.

[9] S. J. Eliades and X. Wang, "Neural correlates of the lombard effect in primate auditory cortex," *Journal of Neuroscience*, vol. 32, no. 31, pp. 10 737–10 748, 2012.

[10] T. Pomberger, C. Risueno-Segovia, J. Löschner, and S. R. Hage, "Precise motor control enables rapid flexibility in vocal behavior of marmoset monkeys," *Current biology*, vol. 28(5), p. 788–794, 2018.

[11] S. Roy, C. T. Miller, D. Gottsch, and X. Wang, "Vocal control by the common marmoset in the presence of interfering noise," *Journal of Experimental Biology*, vol. 214, no. 21, pp. 3619–3629, 11 2011.

[12] D. Takahashi, A. Fenley, and A. Ghazanfar, "Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, p. 20150370, 05 2016.

[13] M. S. Osmanski and X. Wang, "Perceptual specializations for processing species-specific vocalizations in the common marmoset (Callithrix jacchus)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, no. 24, p. e2221756120, 2023.

[14] K. Worley and al., "The common marmoset genome provides insight into primate biology and evolution," *Nature Genetics*, vol. Nat Genet. 2014 Aug;46(8):850-7., p. 850–857, 07 2014.

[15] H. Okano, A. Miyawaki, and K. Kasai, "Brain/minds: brain-mapping project in japan," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 370, 05 2015.

[16] H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque, "Machine learning algorithms for automatic classification of marmoset vocalizations," *PLOS ONE*, vol. 11, pp. 1–14, 09 2016.

[17] A. Wisler, L. J. Brattain, R. Landman, and T. F. Quatieri, "A Framework for Automated Marmoset Vocalization Detection and Classification," in *Proc. Interspeech 2016*, 2016, pp. 2592–2596.

[18] T. O. et al., "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 654–662, 2018.

[19] Y. Zhang, J. Huang, N. Gong, Z. Ling, and Y. Hu, "Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks," *The Journal of the Acoustical Society of America*, vol. 144, pp. 478–487, 07 2018.

[20] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. of ICASSP*, 2021, pp. 3875–3879.

[21] P. C. Bermant, L. Brickson, and A. J. Titus, "Bioacoustic Event Detection with Self-Supervised Contrastive Learning," *bioRxiv*, 2022.

[22] E. Sarkar and M. Magimai.-Doss, "Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?" in *Proc. of Interspeech*, 2023, pp. 1189–1193.

[23] M. S. Osmanski, X. Song, Y. Guo, and X. Wang, "Frequency discrimination in the common marmoset (callithrix jacchus)," *Hearing Research*, vol. 341, pp. 1–8, 2016.

[24] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics," *Data Mining and Knowledge Discovery*, 2019.

[25] S. C. et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, 2022.

[26] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul 2021.

[27] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[28] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of The Royal Society Interface*, vol. 10, no. 83, 2013.

[29] A. Paul, H. McLendon, V. Rally, J. T. Sakata, and S. C. Woolley, "Behavioral discrimination and time-series phenotyping of birdsong performance," *PLOS Computational Biology*, vol. 17, no. 4, pp. 1–21, 04 2021.

[30] S. S. Sethi, "Automated acoustic monitoring of ecosystems," Ph.D. dissertation, Imperial College London, UK, 2020.

[31] S. wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[32] A. L. Pistorio, B. Vintch, and X. Wang, "Acoustic analysis of vocal development in a new world primate, the common marmoset (callithrix jacchus)," *Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1655–1670, Sep 2006.

[33] J. Agamaite, C. Chang, M. Osmanski, and X. Wang, "A quantitative acoustic analysis of the vocal repertoire of the common marmoset (callithrix jacchus)," *Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2906–2928, Nov. 2015.

Exploratory Analysis of Early-Life Chick Calls

Antonella M.C. Torrisi, Ines Nolasco, Elisabetta Versace, Emmanouil Benetos

Queen Mary University of London, UK

{a.m.c.torrisi, i.dealmeidanolasco, e.versace, emmanouil.benetos}@qmul.ac.uk

Abstract

Animal calls are crucial for communication and key indicators of animal welfare. Early-life chick (*Gallus gallus*) calls are vital for hen-chick interactions and reveal their affective states. However, automated detection and recognition systems for chick vocalisations are lacking. Previous studies have identified various call types linked to internal states, but existing models lack systematic validation and are prone to human bias. To address this gap, we developed a computational framework for the automatic detection and feature extraction of chick calls. Using these features, we analysed the calls of one-day-old chicks using various soft and hard clustering techniques to determine whether distinct categories or a continuous spectrum better characterise their repertoire. This preliminary work provides a systematic approach to enhance the understanding and classification of chicks' vocal behaviour, with significant applications in behavioural studies and vocal interactive systems.

Index Terms: chick vocalisations, signal processing, feature extraction, clustering

1. Introduction

Vocalisations are critical for understanding animals' psychophysical states [16], especially in early life stages when calls signal physiological needs, encourage parental care [11], and help regulate internal states [28]. Vocalisations can also influence the affective states of other animals [9], making them essential for social regulation and livestock welfare. Given the growing interest in animal welfare policies [33], studying animal vocalisations has become increasingly important as a fundamental indicator of welfare due to its accessibility and potential for remote evaluation [29].

Poultry chicks (*Gallus gallus*) serve as a reference model for the study of vocalisations due to their precocial nature (i.e., self-sufficient after hatching) which aids the investigation of their vocal behaviour in different experimental conditions post-hatch, and the audible frequency range of their calls (2-6.5 kHz) is more easily translatable into other models [31]. Exploiting the simplicity of their vocal repertoire facilitates the development of algorithms for the automatic and systematic analysis of vocal behaviour. These methods could be adapted to understand and classify the vocal behaviours of different animal species, offering valuable insights for recognition systems and interactive interfaces.

Historically, chick vocalisation classification relied on human annotation and visual inspection of spectrograms [8, 19]. However, these methods often lead to subjective biases. Recent attempts using convolutional neural networks still encountered limitations as a significant number of calls remained undefined, highlighting the need for a more comprehensive and systematic classification system [32]. The current division of a chick's vocal repertoire[19] consists of four categories linked to different affective states and social contexts: distress calls (long, loud, descending frequencies), short peeps (brief, low energy), pleasure calls (ascending frequencies), and warbles (long, harmonic, and repetitive).

To address these gaps, this study proposes a computational framework for the automated detection and feature extraction of chicks' calls. Utilising signal processing methods to extract key acoustic features and employing unsupervised models to explore the vocal repertoire, this approach aims to improve significantly automatic classification [30, 21].

The choice and selection of features crucially impact the effectiveness of classification. This involves two key aspects: the use of manual or automatic methods for feature extraction, and the choice between one-dimensional versus multidimensional features. Manual methods are more accurate and provide a comprehensive understanding of the data [12] but are time-consuming. Conversely, automatic methods are faster and more efficient [25], especially with large datasets, though they may be less accurate and susceptible to noise and other environmental factors [26, 4].Regarding feature dimensionality, one-dimensional features are widely used in animal vocalisation studies [15] for their simplicity, interpretability, and suitability for real-time applications and large datasets [27]. These features are practical as they require fewer computational resources and can be effectively extracted and utilised even with smaller datasets. To ensure our framework incorporates robust and effective feature sets, we conducted a comprehensive review of optimal features used in animal vocalisations [12, 37, 27], and then more specifically focusing on chicks' calls [13, 36]. Key features include time-domain ones like Duration and frequency-domain features like Fundamental Frequency (Pitch) and Spectral Centroid. Thus, we decided to use one-dimensional features to enhance the classification accuracy and interpretability of chicks' vocalisations.

For cluster analysis, we explored both soft and hard clustering techniques to determine whether chick vocal repertoire is better characterised by distinct categories or a continuous spectrum [15]. Soft clustering techniques (like Fuzzy C-Means and Gaussian Mixture Model), that allow data points to belong to multiple clusters, capture the complexity of gradation within call types and handle overlapping clusters and noise better [35, 14]. Hard clustering techniques (like K-means, DBSCAN, and Hierarchical Clustering), which assign each data point to a single cluster, may be more suitable for well-separated clusters [14] and can yield faster results [10]. Furthermore, we explored the suitability of the extracted features to distinguish between female and male chicks. As a prerequisite for our analysis, we developed two methods for the onset and offset detec-

tion of chick vocalisations. While the detailed comparison of these methods is beyond the scope of this paper, we present results for the best-performing method to provide context for the subsequent clustering analysis.

By adopting a data-driven, unsupervised approach, this study aims to develop a systematic framework for classifying chick calls. This framework could uncover call variation among individual chicks, offering deeper insights into their vocal repertoire and achieving a more nuanced and accurate classification system for chick vocalisations.

2. Methods

2.1. Data Collection

The dataset includes 31 audio recordings of individual chicks from two experimental setups investigating tactile recognition in early-life chicks. In the first setup, chicks were tested immediately after hatching in a 90x90 cm arena with a spherical imprinting object. In the second setup, they were tested on their second day in a similar arena with two differently shaped objects. Audio recordings were sampled at 44.1 kHz.

Three experts annotated the sex of the chicks and the onsets and offsets of vocalisations using Sonic Visualiser[7], following an inter-observer agreement procedure. For the onset detection task, the data was divided into training (19 audio recordings), validation, and testing sets (6 audio recordings each), balanced across experimental conditions, recording days, and sex.

For feature extraction and clustering analysis, 12 audio recordings (5,633 calls) from the first experimental condition were selected due to low background noise, making them suitable for extracting noise-sensitive features such as the envelope and Root Mean Square (RMS). Examples of the chicks' calls can be found on the supplemental material page[1].

2.2. Preprocessing

Before analysis, all audio files underwent preprocessing. Each recording was normalised by scaling values to the maximum amplitude. For the feature extraction, a **Bandpass Filter (BPF)** was applied to each audio recording to attenuate background noise and focus on frequency bands relevant to chicks' vocalisations (2000 to 12600 Hz). This involved normalising cutoff frequencies using the Nyquist frequency, applying a Butterworth bandpass filter, and using two coefficients for zero-phase filtering to minimise phase distortion. For the clustering analysis, features were standardised using z-score scaling, ensuring a mean of 0 and a standard deviation of 1 for accurate computation in clustering algorithms.

2.3. Onset and Offset Detection

For onset detection, we evaluated algorithms that detect changes in frequency (High Frequency Content [22]), phase and amplitude (Thresholded Phase Deviation [22], Normalised Weighted Phase Deviation [22], Rectified Complex Domain [22]), and energy (Superflux [5]). For the offset detection task, a time window was defined based on the durations of calls using ground truth onsets and offsets. Then, three methods were tested to improve call offset accuracy in noisy environments. The first method, local minimum detection, identified the local minimum of the energy (a) within a predefined window for the call's offset. The second method, first-order difference combined with local minimum detection, enhanced accuracy by applying first-order differencing of energy ($\Delta a[i] = a[i+1] - a[i]$) before detecting the local minimum. The third method, second-order difference combined with local minimum detection, used second-order differencing ($\Delta^2 a[i] = a[i+2] - a[i]$) before detecting the local minimum, employing the second subsequent frame for subtraction.

2.4. Feature Extraction

The calls were segmented based on the annotated onsets and offsets, for detailed feature analysis. Different time and frequency domain features were computed and summarised using one-dimensional statistics to better describe and classify each call. The primary feature extracted was the **Call Duration**, which is the time between the onset and offset of each call. Next, we extracted the fundamental frequency (F0), representing the average number of oscillations per second[6]. To compute the **Fundamental Frequency (F0)**, we used the PYin algorithm[20]. PYin first estimates F0 values using a probabilistic thresholded distribution from the YIN algorithm, then applies Viterbi decoding to find the most probable F0 sequence. From F0, we derived the frequency bins for the first and second harmonics [22], **F1** and **F2**. These values were then used with the Fast Fourier Transform (FFT) and divided by the sample rate to access the magnitude values at those specific frequency bins. The **Root Mean Square (RMS)** [23] measures the energy contained within each call waveform, indicating the overall amplitude or intensity. The **Spectral Centroid** [17] is the weighted average of frequencies in a signal's spectrum, representing its centre of mass. Here, this feature was computed over the mel-spectrogram representation set on frequencies between 2000 Hz to 12,600 Hz (the range where the calls and harmonics occur). The **Envelope** [34] of a signal represents the magnitude of its instantaneous amplitude over time. To derive the envelope, the call waveform was segmented and the analytic signal was computed using the Hilbert transform. The envelope was then obtained by taking the absolute values of the analytic signal. The **Joint Time-Frequency Scattering Coefficient (JTFS)** [3] analyses audio signals by capturing temporal and spectral features through a cascade of wavelet transforms, including second-order scattering for non-stationary characteristics and higher-order interactions. The features from both domains are then combined into a joint representation, resulting in a multi-layered representation of the signal's structure across multiple time and frequency scales. For each call, JTFS features were computed using the first-order scattering and joint time-frequency scattering transforms, and then the energy values in three different orientations of the JTFS coefficients were computed to capture different patterns of frequency modulation over time. We obtained then a matrix of 26 features for all the calls of our study. This feature extraction captures a comprehensive array of characteristics essential for subsequent clustering analysis, ensuring a robust understanding of the underlying patterns in chick vocalisations. Table 1 summarises the extracted features, which encompass time-domain, frequency-domain, and time-frequency domain characteristics.

2.5. Clustering Analysis

We tested five clustering techniques to understand the structure of chicks' vocal repertoire, whether it forms a continuous spectrum or distinct categories. Both **soft clustering techniques**,

[1] https://github.com/antorr91/Chicks_exploratory_study/blob/main/Notebook_examples.ipynb

Table 1: *Extracted features for the exploratory study. (T) denotes time-domain features, (F) denotes frequency-domain features, and (T-F) denotes time-frequency domain features.*

Feature	Definition
Call Duration (T)	Time difference between the offset and onset of a call.
F0 Mean (F)	Average fundamental frequency within the call.
F0 Std Dev (F)	Variability of fundamental frequency values.
F0 Skewness (F)	Symmetry of the F0 distribution.
F0 Kurtosis (F)	Peakedness or flatness of the F0 distribution.
F0 Bandwidth (F)	Range or span of F0 values within a call.
Mean of F0's First-Order Difference (F)	Average of differences between consecutive F0 values within a call.
F0 Slope (F)	Rate of frequency change from onset to peak of a call.
F0 Magnitude Mean (F)	Average intensity of the fundamental frequency (F0).
F1 Magnitude Mean (F)	Average intensity of the first formant (F1).
F2 Magnitude Mean (F)	Average intensity of the second formant (F2).
Ratio F0/F1 Magnitude Mean (F)	Ratio of the intensity of F0 to F1.
Ratio F0/F2 Magnitude Mean (F)	Ratio of the intensity of F0 to F2.
Spectral Centroid Mean (F)	Average spectral centroid.
Spectral Centroid Std Dev (F)	Variability of spectral centroid values.
RMS Mean (F)	Average RMS value indicating overall energy.
RMS Std Dev (F)	Variability of RMS values.
Attack Magnitude of the Envelope (F)	Intensity of the attack of the call (from onset to peak).
Attack Time of the Envelope (T)	Duration from onset to peak amplitude.
Slope of the Envelope (T)	Rate of change of amplitude during the attack phase.
JTFS Energy Mean (up) (T-F)	Average energy of JTFS coefficients (upward).
JTFS Energy Std Dev (up) (T-F)	Variability of JTFS coefficients' energy (upward).
JTFS Energy Mean (down) (T-F)	Average energy of JTFS coefficients (downward).
JTFS Energy Std Dev (down) (T-F)	Variability of JTFS coefficients' energy (downward).
JTFS Energy Mean (flat) (T-F)	Average energy of JTFS coefficients (flat).
JTFS Energy Std Dev (flat) (T-F)	Variability of JTFS coefficients' energy (flat).

where each instance can belong to multiple clusters, such as Fuzzy C-Means and the Gaussian Mixture Model (GMM), and **hard clustering techniques**, where an instance can belong to just one cluster, such as K-means, Hierarchical Agglomerative Clustering (HAC), and Density-based Spatial Clustering of Applications with Noise (DBSCAN), were employed. Centroid-based clustering includes **K-means** [24], which partitions data into k distinct clusters aiming to minimise the sum of squared distances between data points and their corresponding cluster centroids or mean of the data point assigned to each cluster, and **Fuzzy C-Means**[24], which involves determining centroids as the central points of clusters and grouping data points based on their proximity (in our case given by the Euclidean distance) to these centroids. The **Gaussian Mixture Model (GMM)** [24] is an expectation-maximisation clustering technique, a probabilistic model that assumes data is generated from a mixture of several Gaussian distributions, assigning probabilities to data points for cluster membership. While primarily a soft clustering technique, GMM can also be used as a hard clustering method by assigning each instance to the cluster with the highest posterior probability. Density-based clustering is represented by **DBSCAN** [24], which defines clusters as continuous regions of high density and identifies clusters based on the density of data points. It can detect clusters of various shapes and is robust to noise. Lastly, **Hierarchical Clustering** [2], a connectivity-based method, builds a hierarchy of clusters using two approaches: the bottom-up (agglomerative) method, which merges individual clusters progressively, and the top-down (divisive) method, which splits a single cluster recursively. We used the agglomerative approach, starting with each data point in its cluster and merging the closest pairs iteratively until all points form a single cluster, represented as a dendrogram. The stability of each cluster was optimised by varying algorithm parameters through grid searches. The validity of the results was verified using metrics such as the Silhouette Score [2], which measures how similar a point is to its cluster compared to other clusters, Elbow Method (Within Cluster Sum of Squares, WCSS)[1] that identifies the optimal number of clusters by looking for the point where adding more clusters does not significantly decrease WCSS and the Calinski-Harabasz Index (or, Variance Ratio Criterion, VRC) [18] that evaluates the ratio of the sum of between-cluster dispersion and within-cluster dispersion. The soft clustering techniques were evaluated using additional metrics: the Fuzzy Partition Coefficient (FPC)[24] for Fuzzy C-Means, which assesses the quality of fuzzy separation, and the Akaike Information Criterion (AIC)[24] and Bayesian Information Criterion (BIC)[24] for Gaussian Mixture Models (GMM), which balance model fit and complexity. For visualisation, we used Uniform Manifold Approximation and Projection (UMAP) [15]. This non-linear technique preserves the manifold structure of high-dimensional data, making it ideal for visualising complex relationships and revealing intricate patterns within the vocalisation data.

2.6. Classification of Sex Differences in Chicks' Calls

To investigate sex differences in chicks' calls, we employed a Random Forest for feature selection, followed by a Support Vector Machine (SVM) and a Decision Tree for classification [2]. The dataset, comprising 3,764 calls from seven males and 1,869 calls from five females, was split into training (67.5%) and testing (32.5%) sets, ensuring no individual overlap. All classifiers were optimised using a grid search to find the best combination of parameters.

3. Results and Discussion

3.1. Onset and Offset Detection Performance

The HFC algorithm's performance for onset detection outperformed the other methods with an overall weighted F1 measure of 0.85. For the offset detection task, the method that combined first-order energy difference and local minimum detection proved to be the most effective, with a weighted F1 measure of 0.94.

3.2. Feature Analysis

The qualitative evaluation of features on a subset of calls, selected based on their distinctiveness in the spectrogram, revealed significant differences across call types. Specifically, features such as Call Duration, F0 standard Deviation, F0 Bandwidth, F0 Magnitude, RMS, Spectral Centroid, Envelope, and JTFS coefficient statistics were particularly informative. Correlation analysis showed the highest scores for call duration, RMS statistics, envelope features, and JTFS coefficient energy, indicating these features capture similar aspects of the calls. In contrast, fundamental frequency (F0) statistics, harmonics (F1 and F2), and spectral centroids exhibited low correlations. Additionally, mean F0 negatively correlated with F1 magnitude, the F0-F1 ratio, and other features related to call duration, RMS, envelope, and JTFS energy.

3.3. Clustering Results

Cluster analysis revealed that the optimal number of cluster divisions varies according to the method and metrics used. Table 2 summarises the findings for all the methods tested per number of clusters.

Table 2: *Summary results for all the methods tested per number of clusters*

Cluster	Method	Silhouette Score	CHI	WCSS
2	K-means	**0.3808**	**3171.5**	93689.3
2	Fuzzy C-means	**0.3768**	**3157.4**	65928.0
2	HAC	**0.3724**	**2931.4**	96316.6
2	DBSCAN	**0.6010**	**19.5**	131313.6
2	GMM	**0.335**	**2789.7**	**97938.0**
3	K-means	0.2094	2023.7	85205.3
3	Fuzzy C-means	0.2078	2008.5	42894.7
3	HAC	0.2083	1875.8	87890.2
3	DBSCAN	0.4701	26.9	**129258.5**
3	GMM	0.263	1394.2	97946.4
4	K-means	0.1753	1652.4	77877.2
4	Fuzzy C-means	0.1105	1314.6	**32269.5**
4	HAC	0.1482	1440.8	82842.9
4	DBSCAN	0.4930	17.0	128576.8
4	GMM	0.174	1060.9	93560.3
5	K-means	0.1803	1461.6	**71834.7**
5	Fuzzy C-means	0.0631	1222.7	25711.9
5	HAC	0.1509	1229.9	**78147.0**
5	GMM	0.143	1222.9	78357.5

Further, for Fuzzy C-Means, the highest FPC value of 0.695 occurs with 2 clusters, indicating an optimal balance between cluster clarity and fuzziness, while the GMM shows that dividing into 3 clusters yields the best fit with scores of -168790 (AIC) and -161271 (BIC), following the Elbow rule, suggesting an effective balance between data fit and model complexity. Taken together, these results converge towards an optimal division of our dataset into two clusters. However, statistical analyses for different clusters and features and the subsequent qualitative analysis through the extraction of random samples of calls from the different clusters suggest a better division into 3 clusters. Figure 1 presents a UMAP visualisation of the chicks' calls clustered into three groups using Gaussian Mixture Models (GMM). Additional visualisations examining the different methods and numbers of clusters can be found on the supplemental material page [1]. These initial results are not exhaustive in providing a complete picture of the composition of the chicks' vocal repertoire. However, considering the limited dataset and deriving from a single experimental condition of separation from the group, they do not seem in line with previous studies on chick vocalisations.

3.4. Results of Sex Classification Analysis

Based on the best features selection with the Random Forest classifier, SVM and Decision Tree were trained and tested using as predictors: Duration call, F0 slope, F2 Magnitude Mean, F2-F0 Ratio, Spectral Centroid Mean, JTFS Energy Mean (up), JTFS Energy Std Dev (up), JTFS Energy Mean (down) and JTFS Energy Std Dev (down). The results of predicting sex with SVM and Decision Tree are displayed in Table 3.

Both models performed below the chance level, with overall accuracies of 0.42 for SVM and 0.37 for Decision Tree. The low performance metrics (accuracy, precision, recall, and F1-score all below 0.5) indicate that these features are insufficient for accurate sex prediction in chicks' calls. Further research

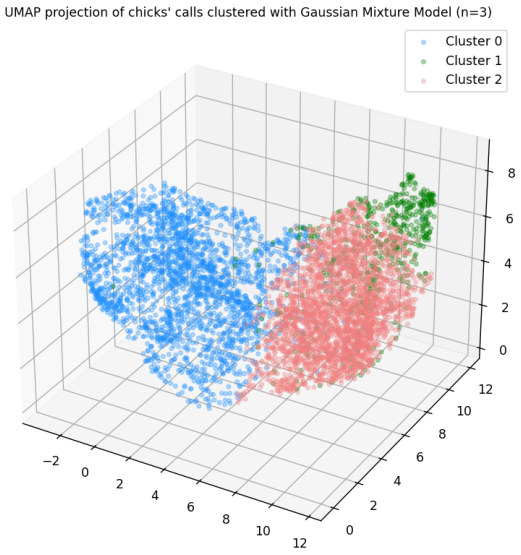

Figure 1: *UMAP- 3D Representation of Three Clusters from GMM*

Table 3: *Comparison of SVM and Decision Tree Results*

Metric	SVM		Decision Tree	
	Female	Male	Female	Male
Precision	0.14	0.43	0.23	0.40
Recall	0.01	0.92	0.06	0.75
F1-Score	0.02	0.59	0.10	0.52
Support	1199	984	1199	984

with additional features or alternative modelling techniques is necessary to improve predictive accuracy.

4. Conclusion

This study presents a computational framework for automatic detection, feature extraction and clustering analysis of chicks' vocalisations. Our approach provides a systematic method to improve the current classification of their vocal repertoire. The extracted features proved informative in identifying at least three distinct categories of chick calls. However, these features did not prove to be significant in discriminating between sexes. The clustering analysis serves as a crucial step towards developing a more robust classification system, potentially guiding the labelling process for subsequent supervised models. This data-driven approach may reveal vocal patterns not previously recognised through traditional human-annotated classification schemes. As a proof of concept, this study lays the groundwork for developing algorithms for automatic feature extraction and unsupervised analysis of early-life chick calls. Further investigations should analyse data from various experimental conditions and developmental stages to comprehensively understand chicks' vocal behaviour.

5. Acknowledgements

IN acknowledges support from EPSRC [EP/R513106/1]. EV is supported by a Royal Society Leverhulme Trust fellow-

ship [SRF\R1\21000155] and Leverhulme Trust research grant [RPG-2020-287]. EB is supported by RAEng/Leverhulme Trust research fellowship [LTRF2223-19-106]. The authors thank Christopher Mitcheltree for his contribution to the feature extraction step.

6. References

[1] Deepak Agrawal and Shikha Kushwaha. Centriod selection process using wcss and elbow method for k-mean clustering algorithm in data mining.

[2] Imran Ahmad. *40 Algorithms Every Programmer Should Know: Hone your problem-solving skills by learning different algorithms and their implementation in Python*. Packt Publishing Ltd, 2020.

[3] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. Joint time–frequency scattering. *IEEE Transactions on Signal Processing*, 67(14):3704–3718, 2019.

[4] Paul Best, Ricard Marxer, Sébastien Paris, and Hervé Glotin. Deep audio embeddings for vocalisation clustering. *PLOS ONE*, 18, 2023.

[5] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, volume 7, page 4, 2013.

[6] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, and Paavo Alku. *Introduction to Speech Processing*. 2 edition, 2022.

[7] Chris Cannam, Christian Landone, Mark B Sandler, and Juan Pablo Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *ISMIR*, pages 324–327, 2006.

[8] Nicholas Collias and Martin Joos. The spectrographic analysis of sound signals of the domestic fowl. *Behaviour*, pages 175–188, 1953.

[9] Frans BM De Waal. Putting the altruism back into altruism: the evolution of empathy. *Annu. Rev. Psychol.*, 59(1):279–300, 2008.

[10] Vibekananda Dutta, Krishna Kumar Sharma, and Deepti Gahalot. Performance comparison of hard and soft approaches for document clustering. *International Journal of Computer Applications*, 41:44–48, 2012.

[11] Joanne Edgar, Suzanne Held, Charlotte Jones, and Camille Troisi. Influences of maternal care on chicken welfare. *Animals*, 6(1):2, 2016.

[12] Julie E Elie and Frederic E Theunissen. The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal cognition*, 19:285–315, 2016.

[13] Ilaria Fontana, Emanuela Tullo, Andy Butterworth, and Marcella Guarino. An innovative approach to predict the growth in intensive poultry farming. *Computers and electronics in agriculture*, 119:178–183, 2015.

[14] Matthias E. Futschik and Bronwyn Carlisle. Noise-robust soft clustering of gene expression time-course data. *Journal of bioinformatics and computational biology*, 3 4:965–88, 2005.

[15] Jack Goffinet, Samuel Brudner, Richard Mooney, and John Pearson. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*, 10:e67855, may 2021.

[16] Karin A Laurijs, Elodie F Briefer, Inonge Reimert, and Laura E Webb. Vocalisations in farm animals: A step towards positive welfare assessment. *Applied Animal Behaviour Science*, 236:105264, 2021.

[17] Phu Ngoc Le, Eliathamby Ambikairajah, Julien Epps, Vidhyasaharan Sethu, and Eric HC Choi. Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, 53(4):540–551, 2011.

[18] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, 2010.

[19] G Marx, J Leppelt, and F Ellendorff. Vocalisation in chicks (gallus gallus dom.) during stepwise social isolation. *Applied Animal Behaviour Science*, 75(1):61–74, 2001.

[20] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 659–663. IEEE, 2014.

[21] Félix Michaud, Jérôme Sueur, Maxime LE Cesne, and Sylvain Haupert. Unsupervised classification to improve the quality of a bird song recording dataset. *ArXiv*, abs/2302.07560, 2022.

[22] Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer, 2015.

[23] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 7(1):155–166, 2005.

[24] Lior Rokach, Oded Maimon, and Erez Shmueli. *Machine Learning for Data Science Handbook*. Springer, 2023.

[25] Benjamin Rowe, Philip Eichinski, Jinglan Zhang, and Paul Roe. Acoustic auto-encoders for biodiversity assessment. *Ecol. Informatics*, 62:101237, 2021.

[26] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, Matt Robbins, Sara C. Keen, Holger Klinck, and Steve Kelling. Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PLoS ONE*, 11, 2016.

[27] Sebastian Schneider, Kurt Hammerschmidt, and Paul Wilhelm Dierkes. Introducing the software case (cluster and analyze sound events) by comparing different clustering methods and audio transformation techniques using animal vocalizations. *Animals : an Open Access Journal from MDPI*, 12, 2022.

[28] Karen A Spencer and Jeroen Minderman. Developmental programming via activation of the hypothalamic–pituitary–adrenal axis: a new role for acoustic stimuli in shaping behavior? *Advances in the Study of Behavior*, 50:87–126, 2018.

[29] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

[30] Dan Stowell and Mark D. Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2, 2014.

[31] Melaku Tefera. Acoustic signals in domestic chicken (gallus gallus): a tool for teaching veterinary ethology and implication for language learning. *Ethiopian Veterinary Journal*, 16:77–84, 2012.

[32] Pieter Thomas, Tomasz Grzywalski, Yuanbo Hou, Patricia Soster de Carvalho, Maarten De Gussem, Gunther Antonissen, Frank Tuyttens, Eli De Poorter, Paul Devos, and Dick Botteldooren. Using a neural network based vocalization detector for broiler welfare monitoring. In *10th Convention of the European Acoustics Association*, 2023.

[33] Belinda Vigors, Peter Sandøe, and Alistair B Lawrence. Positive welfare in science and society: differences, similarities and synergies. *Frontiers in Animal Science*, 2:738193, 2021.

[34] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.

[35] Philip Wadewitz, Kurt Hammerschmidt, Demian Battaglia, Annette Witt, Fred Wolf, and Julia Fischer. Characterizing vocal repertoires—hard vs. soft classification approaches. *PLoS One*, 10(4):e0125785, 2015.

[36] Changhong Wang, Emmanouil Benetos, Shuge Wang, and Elisabetta Versace. Joint scattering for automatic chick call recognition. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 195–199. IEEE, 2022.

[37] Jiangjian Xie, Yujie Zhong, Junguo Zhang, Shuo Liu, Changqing Ding, and Andreas Triantafyllopoulos. A review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecological Informatics*, 73:101927, 2023.

Bird Vocalization Embedding Extraction Using Self-Supervised Disentangled Representation Learning

Runwu Shi[1], *Katsutoshi Itoyama*[2], *Kazuhiro Nakadai*[1]

[1]Tokyo Institute of Technology, Japan
[2]Honda Research Institute Japan, Co., Ltd., Japan
{shirunwu, itoyama, nakadai}@ra.sc.e.titech.ac.jp

Abstract

This paper addresses the extraction of the bird vocalization embedding from the whole song level using disentangled representation learning (DRL). Bird vocalization embeddings are necessary for large-scale bioacoustic tasks, and self-supervised methods such as Variational Autoencoder (VAE) have shown their performance in extracting such low-dimensional embeddings from vocalization segments on the note or syllable level. To extend the processing level to the entire song instead of cutting into segments, this paper regards each vocalization as the generalized and discriminative part and uses two encoders to learn these two parts. The proposed method is evaluated on the Great Tits dataset according to the clustering performance, and the results outperform the compared pre-trained models and vanilla VAE. Finally, this paper analyzes the informative part of the embedding, further compresses its dimension, and explains the disentangled performance of bird vocalizations.
Index Terms: bioacoustics, bird song embedding, disentangled representation learning, self-supervised learning

1. Introduction

Automatic bioacoustic analysis requires the collection of various vocalizations within one species. Such vocal repertoires facilitate the diversity analysis of vocalization and quantitative analysis of vocal behavior. Typically, this process can be conducted by human experts, however, when the categories of repertoire come to hundreds and thousands, this work will be both time consuming and subject to bias. Such situations provide opportunities for self-supervised methods, which do not require large amounts of annotated data.

One type of such method uses pre-trained models to extract embeddings of specific layers [1, 2]. The advantage of this approach favors the case of a limited dataset, where a priori knowledge from other fields can be utilized. The other type considers using self-supervised learning based on the Autoencoder (AE) structure, after which the encoder output will be regarded as suppressed embedding [3]. In [4], a convolutional auto-encoder network is used to learn the abstract embedding of vocalization segments in 6 species, including birds and marine mammals, and performance is quantitatively evaluated using clustering results. In [5], the Variational Autoencoder (VAE) is adopted to learn the vocal embedding of syllables from laboratory mouse and zebra finch, and the results prove that such learned features outperform handpicked features in a variety of downstream tasks.

Despite recent work related to animal vocalization embedding using self-supervised learning has made significant progress, some practical issues remain. Most of the related methods focus on the note or syllable level of vocalization, while some bird songs contain various levels of elements. The songs of Great Tits can be divided into various hierarchical levels as shown in Figure 1 [6], which can be regarded as a special case of sound ontology [7]. The note is the most fundamental unit separated by silence. The syllable is the sequence of notes repeated in the same order in a song. Beyond this is the song that consists of the same syllables. Typically, to extract notes or syllables from a continuous song, some methods such as threshold detection are necessary [8], while it should be noticed that such methods are always sensitive to background noise and the characteristics of different notes, which require researcher's experience and inspection. Moreover, cutting notes inevitably omits original information, and the syntactic relationships among notes are also not fully considered. To build convincing vocalization repertoires on the song level, obtaining the corresponding embeddings directly from the entire song is necessary.

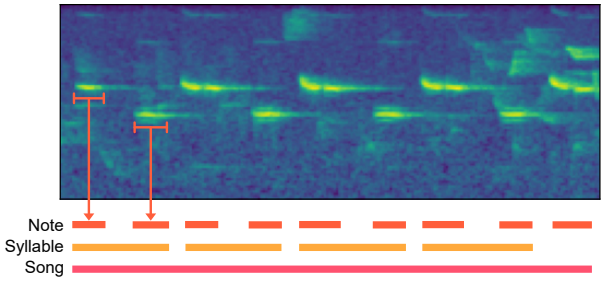

Figure 1: *Different elements in bird song (Great Tit).*

However, learning song embeddings from the entire song using the vanilla VAE structure is challenging. Firstly, it is common for the same type of song to have different numbers of notes repeated, which leads to different lengths of songs. This information will be mixed with the discriminative information and make the embeddings more ambiguous. For instance, embeddings of songs with different lengths can easily be clustered into different groups. Moreover, songs are prone to have syntax changes at the note level, such as 'A-B-A-B' and 'B-A-B' [6], combined with background noise, embeddings should learn to filter out these effects and focus only on critical syntax contents.

Considering such issues, this paper proposes a method to extract song embedding at the song level based on disentangled representation learning (DRL). Instead of using only one encoder, two encoders are adopted to learn the global feature and the local feature simultaneously, in which the global feature represents the temporal related information including fundamental elements such as the number and the position of the notes, and the local feature represents the discriminative information

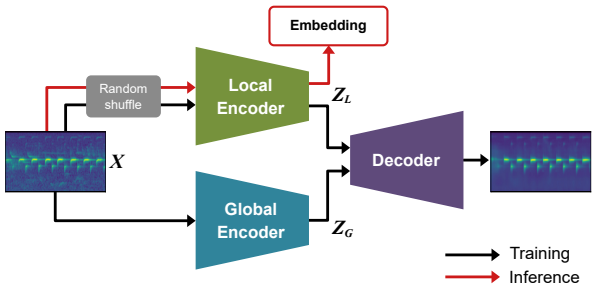

Figure 2: *The framework of the proposed method.*

of each song, as shown in Figure 2. These local features will be used as vocalization embedding of each song. For evaluation, the proposed method is compared with the embeddings of the pre-trained baseline model on clustering performance, and finally, the interpretation of the learned embeddings is discussed. Our main contributions can be summarized as:

- The first attempt to disentangle the structured bird vocalization, and the disentangled embeddings presents a better performance.
- Analyze the information amount learned by embeddings and give the interpretation and embedding compression method.

2. Proposed method

This section introduces the proposed method including the modeling motivation, model structure, and training strategy.

2.1. Structure

Most of the Great Tits' songs consist of repeated fundamental elements such as notes and syllables, beyond this, this paper considers that the features of such songs can also be divided into global features and local features, and such structured songs should also be the prerequisite for disentanglement. The overall framework with two separate encoders is shown in Figure 2.

The global features could be shared among different types of songs and different bird individuals such as the shape of each note, the repetition times related to the length of the song, and even the temporal position of background noise, etc. More importantly, such global features should not contain information related to discrimination, as it's really common for different types of songs to share the same global feature such as the same times of repeated notes. Denote the dataset $X = \{x_1, x_2, ..., x_N\}$ consisting of N Mel-spectrogram of each song segment, and denote $Z_G \in R^{d_G \times T}$ as the global latent representation, where T is the length of the spectrogram. Given the trainable parameters of encoders ϕ, the global feature can be written using posterior distribution as $q_\phi(Z_G|X)$.

For local features, such features should contain more discriminative information that can distinguish diverse songs from different individuals. For instance, the spectral information on the note level and the relationship among the nearby notes representing the syntax content should be contained in these features. This kind of high-level representation should be extracted from the original song inputs and separated from the global features that encode more generalized information. Such features with less redundant information should significantly improve the performance on downstream tasks such as clustering embeddings. The local features are one-dimensional vectors, denoted as $Z_L \in \mathbb{R}^{d_L}$, which can be represented using posterior distribution as $q_\phi(Z_L|X)$.

This idea of modeling comes from disentangled representation learning (DRL) for human speech and musical signals [9, 10, 11, 12, 13]. The DRL can be defined as the learning paradigm that aims to obtain representations capable of identifying and disentangling the underlying factors hidden in the observed data [14]. For instance, the speech can be disentangled into content, speaker, and prosody information. Such methods always assume that in continuous speech, the content information is temporal dependent and the speaker information is temporal invariant [15, 16, 17], these two kinds of latent information will be learned using two separate encoders under the VAE structure. Similar to these related works, we use a temporal encoder for the global feature, and a down sampling encoder with randomly shuffled spectrogram input for the local feature.

The structure of the model is adopted from the SpeechTripleNet [9]. The global encoder uses a 1D convolution layer with kernel size 1 to project the dimension of the input spectrogram into 256, followed by two 1D convolution layers and batch normalization layers. Then two self-attention layers are used to learn the temporal relationship. Finally, global features of shape [128, T] are sampled from a multidimensional Gaussian distribution using the reparameterization trick, where T is the length of the spectrogram. For the local encoder, the input spectrogram will be randomly shuffled with the fixed length segment 32, which can be regarded as ignoring the long term temporal relationship and focusing only on local information [17]. Then, a 1D convolutional layer with kernel size 1 is used to enlarge the dimension of the spectrogram into 256, followed by three 1D convolutional layers with the kernel size of 3, 3, and 5 respectively, and each convolutional layer is followed by an average pooling layer with the size of 2. After layer normalization, the reparameterization trick is used to sample local features, a one-dimensional vector containing 128 latent units. For the decoding process, the local features are firstly expanded to the same shape as the global feature and concatenated with it and then the concatenated feature with the shape of [256, T] is input into the decoder. The decoder has the same structure as the global encoder embedded with self-attention layers that can extract temporal information for better reconstruction. The trainable parameters of the decoder are represented as θ, and the decoder models the conditional probability $p_\theta(X|Z_G, Z_L)$ given the two latent features. The shape of the output is the same as the input spectrogram.

2.2. Training strategy

The training strategy determines the performance of disentangling the useful local features from songs. To achieve better disentanglement, the hyperparameters γ_G, γ_L, C_G, and C_L are adopted in the loss function compared to the original VAE, as shown in equation 1. The learning objective is to minimize this, equivalent to maximizing the lower bound of $\log p_\theta(X|Z)$.

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{X,q_\phi(Z_G,Z_L|X)}[\log p_\theta(X|Z_G, Z_L)] + \\ & \mathbb{E}_X[\gamma_G|D_{KL}(q_\phi(Z_G|X) \parallel p(Z_G)) - C_G|] + \\ & \mathbb{E}_X[\gamma_L|D_{KL}(q_\phi(Z_L|X) \parallel p(Z_L)) - C_L|] \end{aligned} \quad (1)$$

where the first term is the reconstruction loss, and the second and the third terms are the KL divergence of global and local latent features. The γ_G and γ_L are weight factors that control the disentangled extent, a larger value (larger than 1) al-

ways results in a more disentangled latent representation [18]. In addition, the value of γ_G/γ_L also influences the information flow between global and local features [9, 10]. As shown in equation 2, the mutual information among the input and the latent representation $I(X;Z)$ is the lower bound of the KL divergence term, which means that the KL divergence is larger than the information that Z can transmit about the input X [19].

$$\begin{aligned}&\mathbb{E}_X[D_{KL}(q_\phi(Z \mid X) \parallel p(Z))] \\ =&\mathbb{E}_{q(Z,X)}\left[\log \frac{q(Z,X)}{q(Z)p(X)}\right] + \mathbb{E}_{q(Z,X)}\left[\log \frac{q(Z)}{p(Z)}\right] \\ =&I(X;Z) + D_{KL}\left(q(Z) \parallel p(Z)\right)\end{aligned} \tag{2}$$

For instance, if the weight of the global encoder γ_G is given a much larger value than γ_L, the gradient will be dominated by the global encoder output, resulting in the faster vanishing of $D_{KL}\left(q_\phi(Z_G|X) \parallel p(Z_G)\right)$, which means there will be less information about input data X learned by Z_G, and more information will be encoded in Z_L. Typically, the value of γ_G/γ_L should be larger than 1 to let the Z_L be more informative. To more explicitly control the encoding capacities of the two encoders, the controllable parameters C_G and C_L are adopted in the learning objective. These two bounds pressure the KL divergence to converge to a certain information capacity [9, 18]. Since the local encoder should learn more discriminative information, its capacity bound C_L is much larger than C_G.

Empirically, γ_G and γ_L are set to 100 and 10, and C_G and C_L are set to 0.4 and 100, respectively, and we find that the capacity value significantly influences the final performance. During training, the capacity is linearly increased to the maximum value in the first 20K steps [18, 9]. The reconstruction term in equation 1 is realized by the negative log-likelihood among the reconstructed and ground truth samples. The Adam optimizer is used with a learning rate of $1e^{-4}$. The model is trained using a batch size of 64 for a total of 200K training steps.

3. Experiment

This section provides the details of dataset and preprocessing, experiments for model training, visualization of the embeddings, and quantitative performance evaluation and comparison.

3.1. Dataset

This paper focuses on the vocalization of the Great Tits (Parus major). Great Tits can produce varied songs that always consist of repeated notes and syllables. For plenty of diverse samples, this paper uses a publicly available dataset that contains the songs of many different individual Great Tits [6]. The song of Great Tits has individual specific repertoires, which means that each type of song from different bird individuals should be distinguished. The median amount of songs of each Great Tit individual is 4 and the largest amount is 13. The annotation information contains the individual label and the song label of each bird individual, and the labels are obtained using a semi-supervised method which means the labels are not totally accurate [8], which is ignored in this paper.

For the preprocessing of each song segment, each song is transformed into a Mel-spectrogram. A Fast Fourier Transformation with a window size of 1024 is adopted, and the window shift is set to 256. The number of Mel filters is set to 80. The sampling rate is 22 050 Hz, and the cutoff frequency is set to 1500 Hz and 10 000 Hz. To facilitate the training, the length of the spectrogram is limited from 100 to 400. After removing the samples that are too long and too short, there are a total of 98207 song segments used for subsequent experiments.

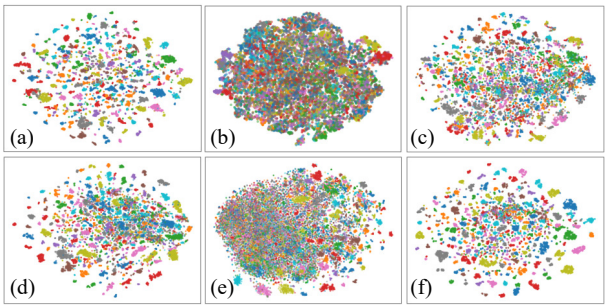

Figure 3: *T-SNE of embeddings of compared methods with different colors for different song types. (a) Local encoder of the proposed method, (b) VAE (Global Encoder with Decoder), (c) Wav2Vec2, (d) Hubert, (e) VQ-APC, (f) OpenL3.*

3.2. Experiment and Results

To comprehensively evaluate the performance of the proposed method, we randomly extracted 70% of the bird individuals as the training dataset, 10% as the validation dataset, and the remaining 20% for testing. In detail, the training, validation, and test dataset contains 244 individuals with 1112 song types, 35 individuals with 151 song types, and 71 individuals with 305 song types, respectively, and there are a total of 67425, 10764, and 20018 song segments in these three subdatasets. We evaluate the clustering performance of the embeddings from the individuals that the model has never seen, to verify if the model has truly learned the discriminative representation.

The process of clustering includes embedding extraction, dimension reduction, and clustering. Firstly, all the local features of each song segment will be collected as the song embedding with a length of 128, then the UMAP is used to reduce the dimension of the embeddings [1, 5, 20], and after that, the HDBSCAN algorithm is used to cluster these condensed embeddings, which is always used in bioacoustic field [4, 20]. The label information is only used to check the correctness of the clustering results. The method in [4] is used to search for optimized parameters of UMAP and HDBSCAN, resulting in the UMAP unit of 4, cluster size of 5, minimum samples of 3, and epsilon of 0.1.

To quantitatively evaluate the clustering of the obtained embeddings, Normalised Mutual Information (NMI) is used to represent the extent to which embeddings are correctly clustered. Given the clusters C and labels L, the NMI calculates the relative entropy between the joint distribution $P_{L,C}$ and the product of P_L and P_C, and normalized by the sum of the entropy of L and C, as shown in equation 3. The NMI will be 1 if the labels match the clusters perfectly. For all compared methods, we use the most optimized clustering parameters to calculate the NMI.

$$NMI(L;C) = \frac{D_{KL}\left(P_{L,C} \parallel P_L \otimes P_C\right) \times 2}{H(L) + H(C)} \tag{3}$$

For comparison, we use the embeddings of pre-trained baseline models and the latent representation of vanilla VAE. We choose several baseline models including Wav2Vec2 [21], Hubert [22], VQ-APC [23], and OpenL3 [24]. The first 3 baseline models are trained in human speech and have demonstrated

their ability in bioacoustic tasks [2, 4]. The OpenL3 provides audio embeddings which can be used for acoustic scene classification. For the vanilla VAE [25], the same structure and the training method as the global encoder is adopted, the only difference being that the embedding channel is compressed to 1 along the time dimension. The T-SNE map of the embeddings is shown in Figure 3. Different colors mean different song types. The embedding output of the local encoder clusters different song types more clearly than other methods, proving the local encoder learns more discriminative features. For comparison, the clustering of vanilla VAE using only one encoder is much more ambiguous, and the OpenL3 embeddings perform better than other baselines.

Table 1: *Performance comparison of different methods.*

Methods	**Parameter**	**Dimension**	**NMI ↑**
Ours	1.7M(Local Enc*) 7.2M(Global Enc) 7.3M(Decoder)	128	0.901
Ours(compress)		**27**	**0.902**
VAE	14.5M	128	0.426
Wav2vec2 [21]	95.0M	768	0.741
Hubert [22]	94.7M	768	0.827
VQ-APC [23]	4.6M	512	0.494
OpenL3 [24]	4.7M	6144	0.895

The NMI results of these methods are shown in Table 1. The NMI of the proposed method is 0.901, and after compression, our method obtains a higher score of 0.902 with a much lower dimension, which will be discussed in the next section. For comparison, the embedding of vanilla VAE only has an NMI of 0.426. The Wav2vec2 gets an NMI of 0.741 and the Hubert gets a higher value of 0.827, suggesting that the structural features learned from human speech can also be extended to bird songs. For the OpenL3 trained on video which provides multi-modality information, the NMI comes to an impressive value of 0.895. The OpenL3 trains on a large amount of data including almost 40M samples [24]. Such diversity of data inputs makes the method highly generalizable to bioacoustic tasks with advanced performance [4, 26].

4. Analysis and Discussion

This section analyzes the obtained embeddings and observes the imbalance of the amount of information in different embedding units, from which more informative embedding units can be extracted to dramatically decrease the dimension of embedding.

Each song embedding with the shape of [1, 128] is sampled from the multidimensional Gaussian distribution with 128 latent units, from which the KL divergence between each unit and the normal distribution can be calculated to estimate the amount of information learned by different units. As shown in Figure 4 (a), the x and y axis represent the mean and variance of each unit, and the color bar means different unit indexes. In Figure 4 (b), the x axis represents the unit index and y axis represents the value of KL divergence, and those units with much larger KL divergence correspond to the unit with lower variance and diverse mean values, as shown in the blue box in these two figures. These units can also be regarded as deterministic units since randomness is eliminated and is mainly controlled by the mean value. It should be noted that the KL divergence of these units at these specific positions always has larger values with different samples of input, indicating these units are

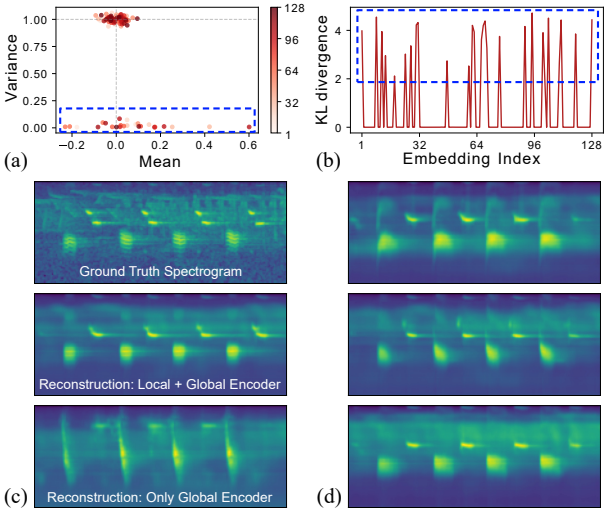

Figure 4: *(a) Each unit's element-wise mean and variance, (b) KL divergence of each unit in embedding, (c) Reconstruction results using all or only the global encoder, (d) Reconstruction results when changing informative units in embedding.*

more informative [19]. To verify this, we extract these 27 informative units and conduct the same test as above. These much shorter embeddings can achieve a higher NMI of 0.902, proving these key units represent discriminative parts of the embedding. Moreover, since the sum of the KL divergence is constrained by the total channel capacity in the learning objective, and empirically, a larger capacity usually leads to more informative units. This method is beneficial for more bioacoustic tasks since the informativeness of the repertoire is diverse in different species.

We also conduct reconstruction experiments to verify the knowledge learned by these informative units. In detail, the output of the global encoder remains unchanged, only one informative unit of the overall 128 units is chosen and adjusted, and then the concatenated features are fed into the decoder to reconstruct the spectrogram. Figure 4 (c) shows the original, reconstructed spectrogram, and reconstructed spectrogram with all 128 units set to 0. This shows that the global encoder learns the temporal information, such as the number and position of each note, but lacks the detailed shape of each note. Each row in figure 4 (d) presents the reconstruction results when one informative unit is adjusted, in which the discriminative features such as note and overtone are changed. However, due to the inconsistency of the element in spectrograms, the extracted informative units do not have isolated disentangled features such as color, direction, and shape as the vision toy dataset [27].

5. Conclusions

This paper demonstrates the feasibility of using DRL to extract bird vocalization embeddings. By adopting DRL, which utilizes two encoders to capture both global and local features of the songs of Great Tits, we achieve the extraction of embeddings from the whole song level, and the clustering performance surpasses the other methods. Furthermore, the analysis reveals the informativeness contained in embedding units, from which the compressed embeddings can be obtained. This approach enhances our understanding of bird vocalization patterns and provides a potential way to measure the information richness of vocal repertoires in more species in further research.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. JP19KK0260, JP20H00475 and JP23K11160.

7. References

[1] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, Dec. 2023.

[2] E. Sarkar and M. Magimai.-Doss, "Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?" in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1189–1193.

[3] R. Suzuki, S. Sumitani, Z. Harlow, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "Extracting Bird Vocalizations from a Complex Natural Soundscape in Forests Using Robot Audition Techniques," in *2023 IEEE/SICE International Symposium on System Integration (SII)*, Jan. 2023, pp. 1–6.

[4] P. Best, S. Paris, H. Glotin, and R. Marxer, "Deep audio embeddings for vocalisation clustering," *PLOS ONE*, vol. 18, no. 7, p. e0283396, Jul. 2023.

[5] J. Goffinet, S. Brudner, R. Mooney, and J. Pearson, "Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires," *eLife*, vol. 10, p. e67855, May 2021.

[6] N. M. Recalde, A. Estandía, L. Pichot, A. Vansse, E. F. Cole, and B. C. Sheldon, "A densely sampled and richly annotated acoustic dataset from a wild bird population," *bioRxiv*, p. 2023.07.03.547484, Jul. 2023.

[7] T. Nakatani and H. G. Okuno, "Sound ontology for computational auditory scence analysis," in *AAAI/IAAI*, 1998, pp. 1004–1010.

[8] N. Merino Recalde, "Pykanto: A python library to accelerate research on wild bird song," *Methods in Ecology and Evolution*, vol. 14, no. 8, pp. 1994–2002, Aug. 2023.

[9] H. Lu, X. Wu, Z. Wu, and H. Meng, "SpeechTripleNet: End-to-End Disentangled Speech Representation Learning for Content, Timbre and Prosody," in *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa ON Canada: ACM, Oct. 2023, pp. 2829–2837.

[10] J. Lian, C. Zhang, and D. Yu, "Robust Disentangled Variational Speech Representation Learning for Zero-Shot Voice Conversion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6572–6576.

[11] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," Jul. 2021.

[12] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, "Pitch-Timbre Disentanglement Of Musical Instrument Sounds Based On Vae-Based Metric Learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 111–115.

[13] Y. Wu, *Self-Supervised Disentanglement of Harmonic and Rhythmic Features in Music Audio Signals*, Sep. 2023.

[14] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled Representation Learning," Aug. 2023.

[15] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling Voice and Content with Self-Supervision for Speaker Recognition," Nov. 2023.

[16] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," Sep. 2017.

[17] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised Speech Decomposition via Triple Information Bottleneck," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 7836–7846.

[18] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," Apr. 2018.

[19] E. Dupont, "Learning Disentangled Joint Continuous and Discrete Representations," Oct. 2018.

[20] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLOS Computational Biology*, vol. 16, no. 10, p. e1008228, Oct. 2020.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," Jun. 2021.

[23] Y.-A. Chung, H. Tang, and J. Glass, "Vector-Quantized Autoregressive Predictive Coding," May 2020.

[24] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE, May 2019, pp. 3852–3856.

[25] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 2022.

[26] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, M. Henry, N. Pinto, C. Noufi, C. Clough, D. Herremans, E. Fonseca, J. Engel, J. Salamon, P. Esling, P. Manocha, S. Watanabe, Z. Jin, and Y. Bisk, "HEAR: Holistic Evaluation of Audio Representations," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, Jul. 2022, pp. 125–145.

[27] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *International Conference on Learning Representations*, Nov. 2016.

What Needs to be Known in Order to Perform a Meaningful Scientific Comparison Between Animal Communications and Human Spoken Language

Roger K. Moore

Speech & Hearing Research Group, Dept. Computer Science, University of Sheffield, UK

r.k.moore@sheffield.ac.uk

Abstract

Human spoken language has long been the subject of scientific investigation, particularly with regard to the mechanisms underpinning speech production. Likewise, the study of animal communications has a substantial literature, with many studies focusing on vocalisation. More recently, there has been growing interest in comparing animal communications and human speech. However, it is proposed here that such a comparison necessitates the appraisal of a minimum set of critical phenomena: i) the number of degrees-of-freedom of the vocal apparatus, ii) the ability to control those degrees-of-freedom independently, iii) the properties of the acoustic environment in which communication takes place, iv) the perceptual salience of the generated sounds, v) the degree to which sounds are contrastive, vi) the presence/absence of compositionality, and vii) the information rate(s) of the resulting communications.

Index Terms: animal communications, human spoken language, comparative communications

1. Introduction

Human spoken language has been the subject of scientific investigation for a considerable period of time, particularly with regard to the mechanisms underpinning the process of speech production [1, 2, 3, 4, 5]. Likewise, the study of animal communications has a substantial history [6, 7], with many studies focused on the particular role of vocalisation [8, 9, 10]. More recently, there has been growing interest in the similarities and differences between human speech and animal communications [11, 12, 13, 14, 15, 16, 17], especially aspects of spoken language that hitherto have appeared to be unique (or at least special) in comparison with the structure of vocal communication systems observed in the rest of the animal kingdom [18, 19].

Of course, there are many ways in which human and animal behaviour may be compared, and the approach taken very much depends on the interests of the individual researchers and their home fields of study. However, comparing animal communications and human speech requires a particularly careful appraisal of a number of phenomena relating to the production, transmission and reception of communicative signals, while simultaneously taking into account the social and pragmatic contexts within which such interactions take place (see Fig. 1). This is not easy to do – especially for animals! Nevertheless, an important methodological step is to identify a minimum set – a 'checklist' – of critical phenomena that need to be characterised in order to perform a meaningful scientific comparison between animal communications and human spoken language. This paper puts forward such a checklist.

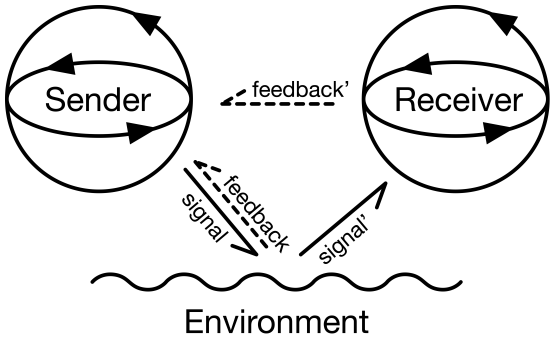

Figure 1: *Illustration (using an extension of Maturana & Varela's pictographs [20, 21]) of communication between sender and receiver 'cognitive unities' (human beings or animals) via a conditioning environmental context.*

2. What we need to Know

It is proposed that, in order to perform a scientific comparison between animal communications and human spoken language, the minimum that needs to be known is ...

i the number of degrees-of-freedom of the vocal apparatus,
ii the ability to control those degrees-of-freedom independently,
iii the properties of the acoustic environment in which communication takes place,
iv the perceptual salience of the generated sounds,
v the degree to which sounds are contrastive,
vi the presence/absence of compositionality, and
vii the information rate(s) of the resulting communications.

2.1. Degrees-of-freedom of the vocal apparatus

The term 'degrees-of-freedom' (DoF) was originally established in the field of statistics to characterise the number of *independent* pieces of information that contribute to estimating the value of a parameter. More recently, DoF has been used in robotics to refer to the number of independent articulators (which is usually directly related to the number of actuators or motors). The concept captures the *dimensionality* of the space of possible physical movements, and is thus highly relevant to characterising their potential use in signalling/communications.

In principle, the number of DoFs of the vocal apparatus is derivable from measurements of the anatomy and the physics of sound production. However, unlike an artificial device such as a robot, a natural living system possesses a very large (effectively, infinite) number of degrees-of-freedom. In this case,

statistical analysis of the movements – e.g. by 'principal components analysis' (PCA) – can provide an estimate of the underlying dimensionality (i.e. how the anatomy is used in practice) and thereby inform the structure of an appropriate mathematical model.

A common model of animal vocalisation (especially human speech) is the 'source-filter' model [1], in which the excitation provided by a sound source (such as the vibration of vocal folds in a larynx, or the actions of a syrinx) is modelled separately from the resonant properties of an acoustic-tube approximation of the vocal tract. Modelling the latter using 'linear prediction analysis' (LPA) [22][1], 'formant' resonators [23, 24] or an 'acoustic-waveguide' [25, 26] can provide accurate simulations of human speech, and can be extended to many other animals (especially non-cetacean mammals) [27, 28].

However, a malleable sound source or a deformable tube has many potential DoFs. Hence, a crucial factor for communication is the degree to which they are under active *control*.

2.2. Control

'Control Theory' is an established discipline in the field of engineering [29], and 'Perceptual Control Theory' (PCT) is the application of control theory to modelling the behaviour of living systems [30, 31]. Derived from 'cybernetics' [32], a key notion is the use of *feedback* to regulate an intended control action. In particular, *closed-loop* control using negative feedback provides a simple yet powerful mechanism for stabilising behaviour in the face of unknown disturbances (just as a thermostat is able to maintain the ambient temperature in a room despite doors and windows being constantly opened and closed).

Of particular interest here are: i) the *number* of DoFs under active control (i.e. the ability of an animal or human being to control those degrees-of-freedom independently), and ii) the *quality* of control for each DoF in terms of their temporal and positional precision as well as their resistance to disturbance. In other words, for communications it is not enough to know what DoFs are being controlled; it is also necessary to know how feasible it is for a *sender* to achieve particular motor targets in a reliable and timely manner, and whether such targets can be maintained in the presence of disturbances[2] [33, 34, 35].

2.3. Acoustic environment

Once a communicative signal has been generated by a sender, it has to propagate through the environment to a *receiver* (as illustrated in Fig. 1). Clearly, the acoustic characteristics of the environment will impact on how (or whether) the signal is perceived by the receiver. However, the environment may also have an impact on the sender, either through long-term (phylogenetic) adaptation [36, 37] or, more interestingly, via short-term (feedback) control [38, 39], in which case the environment may be viewed as a potential disturbance[3]. Hence, determining the level of dependency between emitted signals and the communicative environment is a crucial piece of information in the context of comparing animal communications with human speech.

[1]It may be interesting to note in passing that LPA allows the quality of the information present in the source and filter paths to be modified independently, and thus could facilitate a novel means for investigating animal communications (particularly with regard to the topics addressed in Sections 2.4 and 2.5).

[2]Note that this may be categorised as *sender-oriented* control.

[3]Note that this may be categorised as *receiver-oriented* control.

2.4. Perceptual salience

Once a communicative acoustic signal arrives at the ears of a listener, it is not only important that it is heard (above any ambient noise/interference), but also that any crucial distinctions are actually perceived as different. In other words, there is no value in a sender crafting subtle differences between signals if a receiver is unable to discriminate between them. Hence, it is necessary to understand the psychophysics of listeners' perceptual acuity, for example by characterising their ability to detect 'just-noticeable-differences' (JNDs). However, although measuring JNDs for human listeners is a well-established procedure [40], it is considerably more difficult to perform on animals [41].

Of course, senders are usually also receivers, which means that they are (in principle) able to assess the salience of their own communicative emissions. However, this strategy is only valid for communication between conspecifics; communication between different animals, between humans and animals or even between humans and artificial agents will inevitably be limited by any mismatch in perceptual capabilities [42, 43].

2.5. Contrastive signalling

While it is important for a sender to create perceptible distinctions between each item in their inventory of signals, for a living system there is another factor at play – 'energetics' – that is, the degree of physical and/or neurological *effort* involved in the process. This means that, in principle, by increasing the level of effort, it is not only possible for a sender to optimise perceptual salience by making signals louder, but also by making them *clearer* (i.e. more distinct from one another). Of course, the active management of effort is dependent on a sender's *motivation* to do so, and thus linked to their situational context, for example there may be a degree of urgency associated with the communications.

In human speech, the ability to vary the clarity of a signal along a continuum from *hypo*-articulation (mumbling) to *hyper*-articulation (clear speech) is described by 'H&H Theory' [44] which posits that sound production is actively managed using a closed-loop control process (as already discussed in Section 2.2 above). Not only does this facilitate the dynamic adjustment of speech intelligibility in the face of arbitrary environmental disturbances such as noise or reverberation (see Fig. 2), but it has also led to a system of communication in which sounds are used in a *contrastive* manner, i.e. to distinguish one meaning from another. This is manifest as the 'phonemic' structure of human speech whereby acoustically distinct speech sounds are *only* perceived as different (by native listeners) *if* they signal the difference between one word and another (in their language) [45]. Crucially, acoustically distinct speech sounds are perceived as the *same* if they do *not* signal the difference between one word and another. That is, the sounds listeners perceive - the 'phonemes' - are conditioned on the *meaning* of an utterance, not on a fixed set of acoustic properties. Unfortunately, this dual language-independent 'phonetic' (*physiophonic*) and language-dependent 'phonemic' (*psychophonic*) nature of speech is not always appreciated, with the consequence that the term 'phoneme' is often misused [46].

While these contrastive behaviours are an emergent consequence of the active management of energetic constraints, and thus would seem to reflect a general principle that could apply to *all* communicative behaviour, it has yet to be shown that this is the case – especially for animal communications – although some studies have addressed the issue [48, 49, 50].

"I! ... DO! ... NOT! ... KNOW!"
"I do not know"
"I don't know"
"I dunno"
"dunno"
[ə̃ə̃ə̃]

Figure 2: *An illustration of contrastive behaviour in everyday human conversation. On hearing a verbal enquiry from a family member as to the whereabouts of some mislaid object, the listener might reply with any of the utterances shown (all of which would be perceived as "I do not know") [47]. The particular utterance emitted would depend on the communicative context; the shouts would be necessary in a noisy environment, the nasal grunts would be sufficient in a quiet environment.*

2.6. Compositionality

One of the distinguishing features of human spoken language is the 'particulate' nature of the sound system [51]. That is, just as chemical elements do not blend together, but combine to form structures with quite different properties to their constituent parts, so sounds may be used in different combinations to signify completely unrelated *meanings*. For example, a vocal production systems with d independent DoFs each capable of producing s distinct signals can generate up to s^d different sounds, which means that a sequence of n sounds can support up to $(s^d)^n$ different meanings. As Alexander von Humboldt observed nearly two-hundred years ago: "*language makes infinite use of finite media*" [52].

Clearly, exploiting combinatorics through the efficient *reuse* of sub-structures is an effective means for expanding the expressive power of a communications system. It also provides a means for composing new meanings out of old meanings (which is, indeed, a blending operation). Whether meaningful sequences are formed by combination or by composition, these processes give rise to repetitive sound patterns. Hence, there is interest in algorithms for detecting such repetition in human speech [53, 54] and for determining whether such repetition occurs in animal communications [55, 56, 57].

2.7. Information rates

A prime concern in speech-based interaction is *what* people say, and considerable research resources have been devoted to characterising such behaviour at the traditional acoustic, phonetic, phonological, morphological, lexical, syntactic and semantic levels of description. Such studies involve a multitude of approaches to characterising the complexity of spoken language [58], but 'information theory' [59, 60] provides a particularly powerful paradigm for a single unified approach to *quantitative* measurement. For example, Coupé et al. have shown that *all* human languages have an information rate of ∼39 bits/sec at the phonetic level [61], and Bergey & DeDeo estimate that the information density at the lexical level is ∼13 bits/sec [62].

Similar principles have been applied to animal communications, e.g. entropic values have estimated for bottlenose dolphin whistles and squirrel monkey chucks [63]. Likewise, considerable effort has been devoted to understanding the appropriate methodology [64]. In the context of this paper, all of the considerations discussed in Sections 2.1 to 2.6 (especially, DoFs and JNDs) could be characterised using an information theoretic approach, thereby providing a unified method for comparison across different species.

In reality, human spoken language and animal communications is unlikely to be a fixed code with a *constant* information rate. The information that is communicated is inevitably going to be conditioned on critical causal variables [65] such as ...

- the situated and embodied context (i.e. *pragmatics*),
- the temporal evolution of events (i.e. *synchronics*), and
- the level of effort that participants are prepared to devote to communicative behaviour (i.e. *energetics*).

In other words, a key question in communication is not just *what* is communicated, but *why*, *when* and *how* it is communicated – and these factors will be reflected in a local variation in information rate, e.g. on encountering local minima in cooperative interaction [66], or as a function of cognitive load in unstructured human conversation [62]).

3. Summary and Conclusion

This paper has proposed that, in order to perform a meaningful scientific comparison between animal communications and human speech, the minimum that needs to be known is ...

i the number of degrees-of-freedom of the vocal apparatus,
ii the ability to control those degrees-of-freedom independently,
iii the properties of the acoustic environment in which communication takes place,
iv the perceptual salience of the generated sounds,
v the degree to which sounds are contrastive,
vi the presence/absence of compositionality, and
vii the information rate(s) of the resulting communications.

The claim that this list constitutes *the* minimal set of phenomena is a strong one, and it implies that these criteria are somehow unique. This claim is justified on the basis that each item in the list represents a crucial link in the communications chain between sender and receiver (as illustrated in Fig. 1). Failing to characterise one or more of these phenomena would render an overall comparison lacking in important details.

Having said that, as has been made clear in Section 2, it is not the case that these topics have hitherto been ignored. Quite the contrary, it is acknowledged that many of these areas have already been the subject of extensive investigation. However, it is suggested here that they have often been pursued somewhat independently. Hence, it is posited that there is value in reiterating the dependencies that exist within a communications chain, especially with regard to highlighting **closed-loop control** as a ubiquitous mechanism for regulating behaviour and **information theory** as a universal means for quantifying the relevant outcomes at all points along the chain.

It should also go without saying that none of these aspects of communications are particularly easy to characterise, particularly in animals. However, the claim being made here is that without knowing the answers to these questions, it will be next-to-impossible to draw meaningful comparisons across species. It is therefore hoped that this approach will stimulate productive interdisciplinary discussion.

Finally, although this paper has focused on *vocal* communications, the same principles apply to *multimodal* communications, i.e. gestures, body pose, facial expressions, eye gaze, etc. The principles expounded here would then encompass i) the distribution of information across the available modalities, and ii) the dynamics of shifting the emphasis from one modality to another as a function of the changing communicative context.

4. Acknowledgements

The issues addressed in this paper were inspired by an invitation to review a paper on the topic of animal vocalisation and its potential relation to human speech. I therefore wish to thank and acknowledge – albeit anonymously – the authors of the said manuscript for stimulating the particular train of thought that has been outlined here.

5. References

[1] C. G. M. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[2] P. Ladefoged, *Elements of Acoustic Phonetics*. London: University of Chicago Press, 1962.

[3] P. B. Denes and E. N. Pinson, *The Speech Chain: The Physics and Biology of Spoken Language*. New York: Anchor Press, 1973.

[4] D. B. Fry, *The Physics of Speech*. Cambridge: Cambridge University Press, 1979.

[5] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass.: MIT Press, 1998.

[6] J. W. Bradbury and S. L. Vehrencamp, *Principles of Animal Communication*. Sunderland: Sinauer Associates, 1998.

[7] S. L. Hopp and C. S. Evans, *Acoustic Communication in Animals*. New York: Springer Verlag, 1998.

[8] R. M. Seyfarth and D. L. Cheney, "Meaning and emotion in animal vocalizations," *Ann N Y Acad Sci.*, vol. 1000, pp. 32–55, 2003.

[9] W. T. Fitch, "Production of vocalizations in mammals," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed. Oxford: Elsevier, 2006, pp. 115–121.

[10] R. M. Seyfarth and D. L. Cheney, "Production, usage, and comprehension in animal vocalizations," *Brain and Language*, vol. 115, no. 1, pp. 92–100, 2010.

[11] W. T. Fitch, "The evolution of speech: a comparative review," *Trends in Cognitive Science*, vol. 4, no. 7, pp. 258–267, 2000.

[12] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: what is it, who has it, and how did it evolve?" *Science*, vol. 298, pp. 1569–1579, 2002.

[13] J. F. Prather, "Auditory signal processing in communication: Perception and performance of vocal sounds," *Hearing Research*, vol. 305, pp. 144–155, 2013.

[14] T. C. Scott-Phillips, "Meaning in animal and human communication," *Animal Cognition*, vol. 18, no. 3, pp. 801–5, 2015.

[15] R. K. Moore, R. Marxer, and S. Thill, "Vocal interactivity in-and-between humans, animals and robots," *Frontiers in Robotics and AI*, vol. 3, no. 61, 2016.

[16] S. C. Vernes, "What bats have to say about speech and language," *Psychonomic Bulletin & Review*, vol. 24, no. 1, pp. 111–117, 2017.

[17] S. M. ter Haar, A. A. Fernandez, M. Gratier, M. Knörnschild, C. Levelt, R. K. Moore, M. Vellema, X. Wang, and D. K. Oller, "Cross-species parallels in babbling: animals and algorithms," *Phil. Trans. R. Soc. B.*, vol. 376, no. 1836, pp. 1–11, 2021.

[18] T. Scott-Phillips, *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. London, New York: Palgrave MacMillan, 2015.

[19] M. D. Beecher, "Why Are no animal communication systems simple languages?" *Frontiers in Psychology*, 2021.

[20] H. R. Maturana and F. J. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications, 1987.

[21] R. K. Moore, "Introducing a pictographic language for envisioning a rich variety of enactive systems with different degrees of complexity," *Int. J. Advanced Robotic Systems*, vol. 13, no. 74, 2016.

[22] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[23] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[24] J. N. Holmes, "Formant synthesizers: cascade or parallel?" *Speech Communication*, vol. 2, pp. 251–273, 1983.

[25] J. Mullen, D. M. Howard, and D. T. Murphy, "Digital waveguide mesh modeling of the vocal tract acoustics," pp. 119–122, 2003.

[26] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3231–3254, 2005.

[27] R. K. Moore, "A real-time parametric general-purpose mammalian vocal synthesiser," in *INTERSPEECH*, San Francisco, CA, 2016, pp. 2636–2640.

[28] A. Anikin, "Soundgen: An open-source tool for synthesizing nonverbal vocalizations," *Behavior Research Methods*, vol. 51, pp. 778–792, 2019.

[29] J. J. DiStefano III, A. R. Stubberud, and I. J. Williams, *Feedback and Control Systems*, 2nd ed. New York: McGraw-Hill, 1990.

[30] W. T. Powers, *Behavior: The Control of Perception*. NY: Aldine: Hawthorne, 1973.

[31] W. Mansell and T. A. Carey, "A perceptual control revolution," *The Psychologist*, vol. 28, no. 11, pp. 896–899, 2015.

[32] N. Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, Mass.: The MIT Press, 1965.

[33] E. N. MacDonald, D. W. Purcell, and K. G. Munhall, "Probing the independence of formant control using altered auditory feedback," *Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 955–965, 2011.

[34] K. S. Kim and L. Max, "Estimating feedforward vs. feedback control of speech production through kinematic analyses of unperturbed articulatory movements," *Frontiers in Human Neuroscience*, vol. 8, 2014.

[35] M. S. Brainard and A. J. Doupe, "Auditory feedback in learning and maintenance of vocal behaviour." *Nature reviews. Neuroscience*, vol. 1, no. 1, pp. 31–40, 2000.

[36] J. A. Endler, "Signals, signal conditions, and the direction of evolution," *The American Naturalist*, vol. 139, pp. S125–S153, 1992.

[37] T. G. Forrest, G. L. Miller, and J. R. Zagar, "Sound propagation in shallow water: implications for acoustic communication by aquatic animals," *Bioacoustics*, vol. 4, no. 4, pp. 259–270, 1993.

[38] E. Lombard, "Le sign de l'élévation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[39] H. Brumm and P. J. B. Slater, "Animals can vary signal amplitude with receiver distance: evidence from zebra finch song," *Animal Behaviour*, vol. 72, pp. 699–705, 2006.

[40] P. G. Cheatham, *A Comparison of the Visual and Auditory Senses as Possible Channels for Communication*. U.S. Air Force, Matériel Command, Wright-Patterson Air Force Base, 1950.

[41] E. F. Ndez-Juricic, "The role of animal sensory perception in behavior-based management," *Conservation Behavior: Applying Behavioral Ecology to Wildlife Conservation and Management*, vol. 21, no. 149, 2016.

[42] R. K. Moore, "Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction," in *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*, K. Jokinen and G. Wilcock, Eds. Springer Lecture Notes in Electrical Engineering (LNEE), 2016, pp. 281–291.

[43] G. Huang and R. K. Moore, "Is honesty the best policy for mismatched partners? Aligning multi-modal affordances of a social robot: An opinion paper," *Frontiers in Virtual Reality, section Virtual Reality and Human Behaviour: Do we really interact with artificial agents as if they are human?*, vol. 3, no. 1020169, 2022.

[44] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990, pp. 403–439.

[45] D. Jones, "The history and meaning of the term 'phoneme'," in *Phonology: Selected Readings*, E. C. Fudge, Ed. Harmondsworth, UK: Penguin Books, 1973, ch. 1, pp. 17–34.

[46] R. K. Moore and L. Skidmore, "On the use/misuse of the term 'phoneme'," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, 2019, pp. 2340–2344.

[47] S. Hawkins, "Roles and representations of systematic fine phonetic detail in speech understanding," *Journal of Phonetics*, vol. 31, pp. 373–405, 2003.

[48] D. L. Bowling and W. T. Fitch, "Do animal communication systems have phonemes?" *Trends in Cognitive Sciences*, vol. 19, no. 10, pp. 555–557, 2015.

[49] S. Engesser, J. M. S. Crane, J. L. Savage, A. F. Russell, and S. W. Townsend, "Experimental evidence for phonemic contrasts in a nonhuman vocal system." *PLoS biology*, vol. 13, no. 6, p. e1002171, 2015.

[50] M. Lachmann, S. Szamado, and C. T. Bergstrom, "Cost and conflict in animal signals and human language," *Proc Natl Acad Sci USA*, vol. 98, no. 23, pp. 13 189–13 194, 2001.

[51] W. L. Abler, "On the particulate principle of self-diversifying systems," *Social Biological Structures*, vol. 12, no. 1, pp. 1–13, 1989.

[52] W. von Humboldt, "Uber die Verschiedenheit des menschlichen Sprachbaues und ihren Einfuss auf die geistige Entwickelung des Menschengeschlechts," *Berlin: Royal Academy of Science*, 1836.

[53] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorisation," *IEEE Signal Processing Letters*, vol. 15, pp. 131–134, 2008.

[54] G. Aimetti, R. K. Moore, and L. ten Bosch, "Discovering an optimal set of minimally contrasting acoustic speech units: a point of focus for whole-word pattern matching," Makahuri, Japan, 2010.

[55] A. Kershenbaum and E. C. Garland, "Quantifying similarity in animal vocal sequences: which metric performs best?" *Methods in Ecology and Evolution*, vol. 6, pp. 1452–1461, 2015.

[56] A. Kershenbaum, D. T. Blumstein, M. A. Roch, C. Akcay, G. Backus, M. A. Bee, K. Bohn, Y. Cao, G. Carter, C. Casar, M. Coen, S. L. DeRuiter, L. Doyle, S. Edelman, R. Ferrer-i Cancho, T. M. Freeberg, E. C. Garland, M. Gustison, H. E. Harley, C. Huetz, M. Hughes, J. H. Bruno, A. Ilany, D. Z. Jin, M. Johnson, C. Ju, J. Karnowski, B. Lohr, M. B. Manser, B. McCowan, E. Mercado III, P. M. Narins, A. Piel, M. Rice, R. Salmi, K. Sasahara, L. Sayigh, Y. Shiu, C. Taylor, E. E. Vallejo, S. Waller, and V. Zamora-Gutierrez, "Acoustic sequences in non-human animals: a tutorial review and prospectus," *Biological Reviews*, vol. 91, pp. 13–52, 2016.

[57] P. Sharma, S. Gero, R. Payne, D. F. Gruber, D. Rus, A. Torralba, and J. Andreas, "Contextual and combinatorial structure in sperm whale vocalisations," *Nature Communications*, vol. 15, no. 3617, 2024.

[58] D. Gibbon, R. K. Moore, and R. Winski, *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. K. Moore, and R. Winski, Eds. Berlin, New York: Mouton de Gruyter, 1997.

[59] C. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[60] C. E. Shannon, "Prediction and entropy of printed English," *The Bell System Technical Journal*, pp. 50–64, 1951.

[61] C. Coupé, Y. Oh, D. Dediu, and F. Pellegrino, "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche," *Science Advances*, vol. 5, no. 9, 2019.

[62] C. A. Bergey and S. DeDeo, "From "um" to "yeah": Producing, predicting, and regulating information flow in human conversation," *arXiv*, 2024.

[63] B. McCowan, S. Hanser, and L. Doyle, "Using information theory to assess the diversity, complexity, and development of communicative repertoires," *Journal of Comparative Physiology*, vol. 116, no. 2, pp. 166–172, 2002.

[64] A. Kershenbaum, "Entropy rate as a measure of animal vocal complexity," *Bioacoustics*, vol. 23, no. 3, pp. 195–208, 2014.

[65] R. K. Moore, "Pragmatics, synchronics and energetics in spoken language – an information theoretic perspective," in *Limits and Benefits of Information-Theoretic Perspectives in Spoken Communication*, Dublin, 2023.

[66] ——, "Local minima drive communications in cooperative interaction," in *Proceeding of the AISB Convention*, Swansea, 2023.

Proc. 4th Intl. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), Kos, GR, 6 and 9 Sep 2024

Towards Differentiable Motor Control of Bird Vocalizations

Vincent Lostanlen[1]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
vincent.lostanlen@ls2n.fr

Abstract

Machine learning is ready to transform the experimental protocol of birdsong acquisition and playback in ethology and integrative neuroscience . An emerging methodology, known as differentiable digital signal processing (DDSP), allows to train neural networks for machine listening so as to fit the synthesis parameters which correspond to unlabeled audio data. In this short article, I present the value and of extending DDSP, initially developed for speech and music processing, to avian bioacoustics. The main two challenges reside in the definition of a suitable decoder and learning objective. I review some prior publications in biomechanical models of vocal production for passerines, similarity computing, and differentiable solvers of ordinary differential equations. Together, these publications hint at the feasibility of a fully automated and unsupervised algorithm for biologically plausible resynthesis of birdsong.
Index Terms: birdsong, model-based deep learning, physical modeling synthesis

1. Extended abstract

Over the past decade, the renewed interest for deep learning in signal processing has led to a new generation of systems for passive acoustic monitoring [1]. For example, BirdNET is a deep neural network which detects bird vocalizations from acoustic sensor data and recognizes the corresponding species according to a predefined taxonomy [2]. Comparable solutions exist for flight calls [3] and for open taxonomies [4]. Yet, in these examples, the machine listening system reduces birdsong to a sequence of time segments whose boundaries align with the onset of offset of each song bout [5]. In doing so, it erases spectrotemporal patterns which are attributable to intraclass variability.

Although per-species timings may suffice for ecologists who study wild avian populations, ethologists and neuroscientists often depend on a richer description of birdsong content as part of their research protocols. There is abundant literature on the evolutionary and developmental aspects of vocal learning in songbirds: e.g., zebra finches, canaries, and budgerigars. Moreover, a well-known study by Pepperberg *et al.* has shown the exceptional abilities of an African gray parrot in terms of functional vocalizations when interacting with humans in English [6]. In these studies, automating species classification would be useless, since the specimens are known and kept in an aviary. Rather, a valuable source of information on animal behavior is found in the fundamental frequency (f_0) contours of animal vocalizations. Unfortunately, f_0 tracking is more difficult for birdsong than for solo music or speech, due to higher rates of amplitude and frequency modulation. Hence, if f_0 tracking of birdsong is to be automated in the future, it requires a dedicated approach.

Despite the proven merits of machine learning in bioacoustic detection and classification, the task of f_0 tracking comes with a challenge of its own: that of collecting training data. Indeed, the expert annotation of f_0 contours is even more costly and time-consuming than that of species-specific vocal activity detection. For lack of available ground truth, the task must be approached via unsupervised learning techniques. Historically, some of these techniques have been successfully applied to marine bioacoustics (e.g., [7]) but rarely ever to birdsong, with the notable exception of spherical k-means [8]. Still, up to recently, unsupervised representation learning algorithms were unsuitable for highly time-varying and spectrally rich signals such as birdsong.

The situation has changed recently with the introduction of a new methodological framework for unsupervised learning in speech and music, known as differentiable digital signal processing (DDSP). The key idea behind DDSP is to train an autoencoder whose encoder contains learnable parameters but whose decoder does not, while both are compatible with automatic differentiation. Minimizing the reconstruction of error of the autoencoder over a training set of unlabeled natural sounds is tantamount to solving an inverse problem whose associated direct problem is specified by the decoder [9]. In its earliest version, the DDSP decoder was a simple additive sinusoidal model with random Gaussian noise and reverberation. More recently, a broader range of decoders has been developed, directly mimicking the state of the art in acoustical simulation and virtual analog audio effects: let us refer to [10] for a review. Therefore, DDSP is a kind of "model-based deep learning" in the sense that it hybridizes physics-driven and data-driven insights so as to learn an informative representation of natural sounds [11].

I propose to adapt the DDSP framework to the long-standing problem of unsupervised representation learning of birdsong. DDSP has already been successfully applied to f_0 estimation in music signals, under the name of DDSP-inv [12]. My scientific hypothesis is that DDSP-inv has the potential to improve the state of the art in analysis–synthesis of birdsong, currently held by hidden Markov models (HMM) [13] and, more recently, WaveNet [14]. However, I believe that the standard formulation of DDSP, based on sinusoidal models and multiscale spectrogram loss (MSS), is not suitable to birdsong. Indeed, even so the authors of DDSP have presented a demonstration of birdsong analysis–resynthesis as part of their "Paint With Music" outreach project, the result does not sound naturalistic [1]. To serve the needs of ethologists and neuroscientists working on captive birds, the components of DDSP must be redesigned.

On one hand, the groundbreaking publications of Mindlin, Laje, Amador, Sitt, Perl, and colleagues have laid the ground-

[1] Link to "Paint With Music" project:
https://magenta.tensorflow.org/paint-with-music

work for a comprehensive physical description of the vocal apparatus in some well-studied songbirds, e.g., zebra finch and canary [15, 16, 17, 18, 19]. The commonality between these publications is to model the syrinx as a nonlinear dynamical system whose parameters have a biomechanical interpretation. For example, [18] apply the theory of Takens–Bogdanov bifurcations to present a dynamical system governed by the following second-order ordinary differential equation (ODE):

$$\ddot{\boldsymbol{x}} = \gamma^2\alpha + \gamma^2\beta\boldsymbol{x} + \gamma\boldsymbol{x}^2 - \gamma\boldsymbol{x}\dot{\boldsymbol{x}} - \gamma\boldsymbol{x}^3 - \gamma, \qquad (1)$$

where $\boldsymbol{x}$ represents the departure of the midpoint position of the oscillating labia in the syrinx, α and β are functions of the air sac pressure and the activity of the ventral syringeal muscle, and γ is a time scaling factor. Although a Python implementation is available[2] to compute $\dot{x}$ from $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$, it depends on NumPy; as such, it is not interoperable with neural network training. We propose to reimplement this synthesizer in PyTorch, a Python framework for differentiable computing. More precisely, the torchdiffeq library [20] allows to program solvers for ordinary differential equation in which the solution (x) may be differentiated with respect to the parameters ($\boldsymbol{\theta}$). Via reverse-mode automatic differentiation, it will be possible to evaluate the gradient of a function of x may with respect to neural network weights $\mathbf{W}$ where $\boldsymbol{\theta}$ is defined as $f_{\boldsymbol{W}}(\boldsymbol{x})$ and $f_{\mathbf{W}}$ is the encoder.

On the other hand, a new generation of differentiable time–frequency representations have the potential to improve the conditioning of the inverse problem in DDSP, which may accelerate gradient-based optimization when training the encoder. For example, a differentiable implementation of the joint time–frequency scattering transform (JTFS) has recently been released as part of the Kymatio package [21]. Prior work on synthetic chirps has confirmed that, with JTFS, parameter estimation is faster, more accurate, and less susceptible to random initialization than MSS [22]. Although there is a gap in acoustical complexity between synthetic chirps and real birdsong, this result is encouraging because it directly addresses the issue of unsupervised learning in the presence of fast spectrotemporal modulations. Another option would be to use a pretrained neural network as feature map for similarity computing between the natural signals and its autoencoded version.

In conclusion, I have described the promise and challenge of learning to control a physical model of birdsong without supervision and have outlined the necessary steps to get there. Beyond the fundamental interest of advancing differentiable digital signal processing (DDSP), its application to birdsong would unlock new research protocols in ethology and neuroscience.

2. Acknowledgements

This work is supported by ANR projet nIrVAna (ANR-23-CE37-0025). I thank Michael Newton and Yining Xie for helpful discussions and thank anonymous reviewers for their feedback.

3. References

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

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[3] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, p. e0214168, 2019.

[4] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin *et al.*, "Learning to detect an animal sound from five examples," *Ecological informatics*, vol. 77, p. 102258, 2023.

[5] V. Lostanlen and B. Mcfee, "Efficient evaluation algorithms for sound event detection," in *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.

[6] I. M. Pepperberg, "Functional vocalizations by an african grey parrot (psittacus erithacus)," *Zeitschrift für Tierpsychologie*, vol. 55, no. 2, pp. 139–160, 1981.

[7] P. Li, X. Liu, H. Klinck, P. Gruden, and M. A. Roch, "Using deep learning to track time× frequency whistle contours of toothed whales without human-annotated training data," *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 502–517, 2023.

[8] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

[9] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[10] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, p. 1284100, 2024.

[11] G. Richard, V. Lostanlen, Y.-H. Yang, and M. Müller, "Model-based deep learning for music information research," *arXiv preprint arXiv:2406.11540*, 2024.

[12] J. Engel, R. Swavely, L. H. Hantrakul, A. Roberts, and C. Hawthorne, "Self-supervised pitch detection by inverse audio synthesis," in *Proceedings of the ICML Workshop on Self-supervision in Audio and Speech*, 2020.

[13] L. Gutscher, M. Pucher, C. Lozo, M. Hoeschele, and D. C. Mann, "Statistical parametric synthesis of budgerigar songs," in *Proceedings of INTERSPEECH*, 2019.

[14] R. R. Bhatia and T. H. Kinnunen, "An initial study on birdsong re-synthesis using neural vocoders," in *International Conference on Speech and Computer*. Springer, 2022, pp. 64–74.

[15] G. B. Mindlin and R. Laje, *The physics of birdsong*. Springer Science & Business Media, 2005.

[16] A. Amador and G. B. Mindlin, "Beyond harmonic sounds in a simple model for birdsong production," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 18, no. 4, 2008.

[17] J. Sitt, A. Amador, F. Goller, and G. Mindlin, "Dynamical origin of spectrally rich vocalizations in birdsong," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 78, no. 1, p. 011905, 2008.

[18] Y. S. Perl, E. M. Arneodo, A. Amador, and G. B. Mindlin, "Nonlinear dynamics and the synthesis of zebra finch song," *International Journal of Bifurcation and Chaos*, vol. 22, no. 10, p. 1250235, 2012.

[19] A. Amador and G. B. Mindlin, "Low dimensional dynamics in birdsong production," *The European Physical Journal B*, vol. 87, pp. 1–8, 2014.

[20] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[21] J. Muradeli, C. Vahidi, C. Wang, H. Han, V. Lostanlen, M. Lagrange, and G. Fazekas, "Differentiable Time-Frequency Scattering On GPU," in *Digital Audio Effects Conference (DAFx)*, 2022.

[22] C. Vahidi, H. Han, C. Wang, M. Lagrange, G. Fazekas, and V. Lostanlen, "Mesostructures: Beyond spectrogram loss in differentiable time-frequency analysis," *Journal of the Audio Engineering Society*, vol. 71, no. 9, pp. 577–585, 2023.

[2] Python implementation: https://github.com/zekearneodo/syrinxsynth

Exploring bat song syllable representations in self-supervised audio encoders

Marianne de Heer Kloots[1], *Mirjam Knörnschild*[2]

[1]Institute for Logic, Language and Computation, University of Amsterdam; The Netherlands
[2]Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science; Germany
m.l.s.deheerkloots@uva.nl, mirjam.knoernschild@mfn.berlin

Abstract

How well can deep learning models trained on human-generated sounds distinguish between another species' vocalization types? We analyze the encoding of bat song syllables in several self-supervised audio encoders, and find that models pre-trained on human speech generate the most distinctive representations of different syllable types. These findings form first steps towards the application of cross-species transfer learning in bat bioacoustics, as well as an improved understanding of out-of-distribution signal processing in audio encoder models.

Index Terms: self-supervised models, computational bioacoustics, comparative analyses, interpretability

1. Introduction

Many researchers in bioacoustics would benefit from robust and accurate feature spaces that can handle graded vocalizations in real-world field recordings, for example for the purpose of automatic classification. In the domain of human speech and sound processing, much recent progress is driven by so-called self-supervised audio encoder models [1, 2], which learn rich representations of acoustic signals through a masked audio segment prediction task on unlabelled data. Training such models from scratch for non-human species is currently still infeasible, due to the limited size of most bioacoustic datasets [3, 4]. However, existing pre-trained models still offer promising opportunities through their use in *cross-species transfer learning*, providing a new tool to explore divergences and commonalities between species [5]. Here, we explore how a variety of self-supervised audio models trained on human and non-human generated sounds encode bat song syllable types in field recordings of one species' territorial song.

2. Data

We use a dataset of 20 territorial songs produced by males of the Greater Sac-Winged Bat (*Saccopteryx bilineata*), recorded in Costa Rica using an ultrasonic microphone (Avisoft USG 116Hme with condenser microphone CM16; frequency range 1–200 kHz). These multisyllabic vocalizations are acquired by imitation from tutor males during ontogeny [6] and encode personal information about the singer such as individual identity, group affiliation and regional origin [7]. Territorial songs are composed of up to six different syllable types [8], five of which are present in our dataset and manually labelled for analyses (420 syllables in total; including 135, 97, 92, 9, and 87 instances of syllable types A, B, C, D, and E, respectively).

3. Analyses

3.1. Data pre-processing

Several pre-processing steps were performed before feeding our dataset of *S. bilineata* territorial songs through the pre-trained audio encoder models. For denoising, we used the noise reduction algorithm implemented in the software Avisoft SASLab Pro, which automatically recognizes syllables and removes noise below a user-defined threshold in the frequency domain. Depending on the noise floor of each recording, threshold levels were between -60 to -75 dB. Detected noise was reduced by 90dB. We further applied a high-pass filter of 10 kHz.

Vocalizations of the recorded *S. bilineata* population have a species mean fundamental frequency (F0) around 15.5 kHz (SD: 2 kHz), but also contain much energy above 20 kHz. Such higher frequencies are mostly inaudible to humans and outside the training distribution of the pre-trained audio encoders studied here. After denoising, we therefore move the songs into the human auditory range by slowing down all recordings in our dataset by a factor of 8. In the slowed down recordings, mean syllable duration is 235 ms (SD: 135 ms) and most energy is contained within the 1-8 kHz frequency band for all syllable types (F0 mean: 2.3 kHz, SD: 900 Hz). Finally, we downsample all recordings to 16 kHz, as required for processing by the pre-trained audio encoders.

3.2. Feature extraction

Our set of four self-supervised models comprises two different architectures and three different sets of pre-training data (see Table 1). The AVES model is an audio representation model developed for encoding animal vocalizations; we here use the AVES-`bio`-base configuration pre-trained on a large set of animal sounds from various species. We also include another HuBERT-based model trained exclusively on human speech (Librispeech audiobooks [9]), as well as a Wav2Vec2.0 model trained on the same data, and a second Wav2Vec2.0 model trained exclusively on music (from the Free Music Archive [10]). Each model consists of a CNN-based waveform encoder followed by 12 Transformer layers, ultimately generating 768-dimensional feature sequences at a frame rate of 20 ms.

Table 1: *Self-supervised audio models included in our analyses*

Name	Architecture	Training data
AVES [11]	HuBERT	360h, animals
HuBERT (speech) [2]	HuBERT	960h, speech
Wav2Vec2 (speech) [1]	Wav2Vec2.0	960h, speech
Wav2Vec2 (music) [12]	Wav2Vec2.0	900h, music

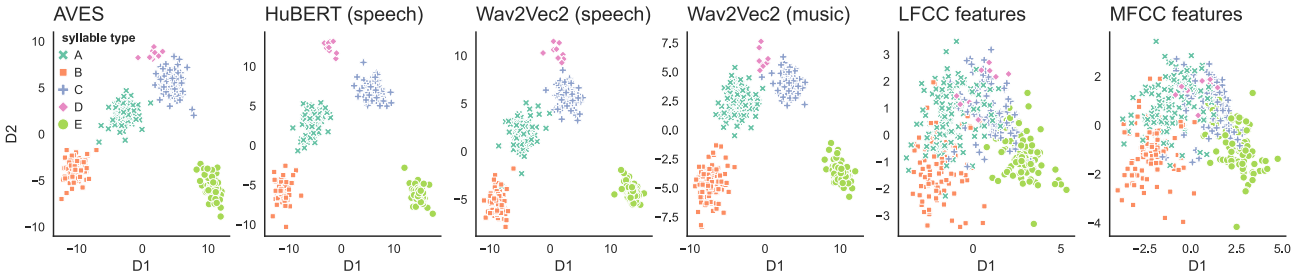

Figure 1: *Syllable projections along the two most discriminative directions in each LDA-transformed feature space.*

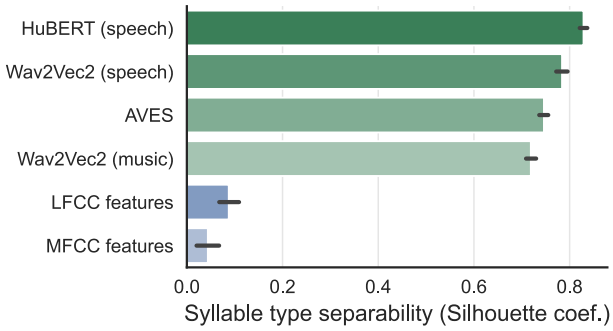

Figure 2: *Separability between syllable type clusters is highest in the self-supervised models trained on human speech (error bars show 95% confidence intervals).*

To create syllable representations using each of the self-supervised audio encoders, we pass a full song as input through the model, and extract frame representations from its final Transformer layer. We then average over all frames within each syllable, resulting in one 768-dimensional feature embedding for every syllable. Mean-pooling across time might seem overly simplistic for capturing distinct temporal dynamics between syllable types (e.g. upsweeps vs. downsweeps). However, similar mean-pooled Transformer-based embeddings have been shown to successfully capture information across several timescales in human speech processing (for example on the phoneme- [13, 14] and word-level [15]), and perform well on bioacoustic transfer learning tasks [11, 16].

Finally, we include two simpler feature sets of 13-dimensional linear- and Mel-frequency cepstral coefficients for comparison, each computed with a 400 sample FFT window.

3.3. Separability analyses

We aim to assess how distinctively *S. bilineata* territorial song syllables are encoded in each feature space. For this purpose, we first project each set of syllable features into its 4 most discriminative directions using Linear Discriminant Analysis (LDA). Figure 1 visualizes every syllable's location along the first two directions of each projected feature space. This reveals that the self-supervised audio encoder models encode each of the 5 syllable types into distinguishable subspaces, which are linearly decodable from their final layer representations. In contrast, the LFCC and MFCC features show much more entanglement between syllable types.

To more precisely quantify the separability between different syllable types in each feature space, we compute silhouette coefficients for each syllable type cluster based on Mahalanobis distances between samples.

The silhouette coefficient for each sample is defined as $(b - a)/\max(a, b)$, where a is the mean distance to all other points in the same cluster, and b is the mean distance to all other points in the next nearest cluster. The mean silhouette coefficients per LDA-projected feature space are visualized in Figure 2. This shows that syllable type separability is highest in the two self-supervised models trained on human speech, followed by the model trained on animal vocalizations, and finally the model trained on music.

4. Discussion & Conclusions

We find that the syllable types in our territorial song recordings, when slowed down to the human hearing range, are distinctively encoded by self-supervised audio encoders. Representations learned by such models thus encode useful features for *S. bilineata* syllable identification, even when only pre-trained on sounds generated by other species.

Interestingly, syllable types are most separable in the models pre-trained on human speech. This indicates that rich representations optimized for a single species' vocal repertoire might form a more promising basis for cross-species transfer learning than those optimized to encode a large variety of species, or non-vocal sound sources like musical instruments. However, the animal vocalization model included in our current comparison set was pre-trained on a substantially smaller amount of audio than the speech and music models (Table 1). A comparison against models pre-trained on fewer hours of speech would be needed to determine whether training dataset size could explain the difference between models trained on human speech vs. multiple species. Between the speech-trained models, the HuBERT architecture showed a slight syllable separability advantage compared to the Wav2Vec2 architecture. This could be due to the clustering objective that is part of the HuBERT training procedure [2], potentially driving the model's internal representations towards generally more separable subspaces.

Our current findings indicate that self-supervised audio encoders pre-trained on human speech generate useful representations for distinguishing between *S. bilineata* song syllable types. However, territorial songs in this species are known to also encode singer identity, and several other features [7] — models might differ in which features they most prominently encode. Representations from self-supervised models can be optimized by supervised fine-tuning to encode the most relevant features for specific classification and detection tasks. In future work, we aim to further investigate what interpretable features contribute to the distinctive syllable type representations across each of the audio encoders' internal layers, and test the applicability of our approach to other tasks in bat bioacoustics, such as syllable detection, species and dialect identification.

5. Acknowledgements

We thank Pierre Orhan and co-authors for sharing the weights of their music-trained Wav2Vec2 model with us, and two anonymous reviewers for very helpful comments.

6. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[3] R. Manriquez, S. Kotz, A. Ravignani, and B. De Boer, "Deep Learning on Small Datasets to Classify Mammalian Vocalizations," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, Turin, Italy, 2024, pp. 4687–4690. [Online]. Available: https://dael.euracoustics.org/confs/landing_pages/fa2023/001052.html

[4] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022. [Online]. Available: https://peerj.com/articles/13152

[5] J. Cauzinille, B. Favre, R. Marxer, and A. Rey, "From speech to primate vocalizations: Self-supervised deep learning as a comparative approach," in *The Evolution of Language: Proceedings of the 15th International Conference (Evolang XV)*, 2024, p. 57214. [Online]. Available: https://evolang2024.github.io/proceedings/paper.html?nr=60

[6] M. Knörnschild, "Vocal production learning in bats," *Current Opinion in Neurobiology*, vol. 28, pp. 80–85, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959438814001275

[7] M. Knörnschild, S. Blüml, P. Steidl, M. Eckenweber, and M. Nagy, "Bat songs as acoustic beacons - male territorial songs attract dispersing females," *Scientific Reports*, vol. 7, no. 1, p. 13918, 2017. [Online]. Available: https://www.nature.com/articles/s41598-017-14434-5

[8] O. Behr, O. von Helversen, G. Heckel, M. Nagy, C. C. Voigt, and F. Mayer, "Territorial songs indicate male quality in the sac-winged bat Saccopteryx bilineata (Chiroptera, Emballonuridae)," *Behavioral Ecology*, vol. 17, no. 5, pp. 810–817, 2006. [Online]. Available: https://doi.org/10.1093/beheco/arl013

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7178964

[10] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A Dataset For Music Analysis," *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 2017. [Online]. Available: https://archives.ismir.net/ismir2017/paper/000075.pdf

[11] M. Hagiwara, "AVES: Animal Vocalization Encoder Based on Self-Supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10095642/

[12] P. Orhan, Y. Boubenec, and J.-R. King, "Algebraic structures emerge from the self-supervised learning of natural sounds," 2024, bioRxiv/2024.03.13.584776s. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.03.13.584776v1

[13] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2021, pp. 914–921. [Online]. Available: https://www.doi.org/10.1109/ASRU51503.2021.9688093

[14] M. de Heer Kloots and W. Zuidema, "Human-like Linguistic Biases in Neural Speech Models: Phonetic Categorization and Phonotactic Constraints in Wav2Vec2.0," in *Proc. INTERSPEECH*, 2024. [Online]. Available: https://www.doi.org/10.21437/Interspeech.2024-2490

[15] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What Do Self-Supervised Speech Models Know About Words?" *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 372–391, Apr. 2024. [Online]. Available: https://doi.org/10.1162/tacl_a_00656

[16] J. Cauzinille, B. Favre, R. Marxer, D. J. Clink, A. H. Ahmad, and A. Rey, "Investigating self-supervised speech models ability to classify animal vocalizations: The case of gibbon's vocal signatures," in *Proc. INTERSPEECH*, 2024. [Online]. Available: https://www.doi.org/10.21437/Interspeech.2024-1096

Feature Representations for Automatic Meerkat Vocalization Classification

Imen Ben Mahmoud[1], *Eklavya Sarkar*[1,2], *Marta Manser*[3], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3]University of Zurich (UZH), Switzerland
{ibmahmoud, esarkar, mathew}@idiap.ch, marta.manser@ieu.uzh.ch

Abstract

Understanding evolution of vocal communication in social animals is an important research problem. In that context, beyond humans, there is an interest in analyzing vocalizations of other social animals such as, meerkats, marmosets, apes. While existing approaches address vocalizations of certain species, a reliable method tailored for meerkat calls is lacking. To that extent, this paper investigates feature representations for automatic meerkat vocalization analysis. Both traditional signal processing-based representations and data-driven representations facilitated by advances in deep learning are explored. Call type classification studies conducted on two data sets reveal that feature extraction methods developed for human speech processing can be effectively employed for automatic meerkat call analysis.

Index Terms: bioacoustics, feature representations, self-supervised learning, call type classification

1. Introduction

Meerkats are highly social animals with a complex social structure [1]. Featuring a dominant breeding pair and cooperative behaviors, they dig safe places through their foraging areas. Communication among a clan occurs through various vocalizations including barks, chirps, trills, and growls. They are essential in coordinating group activities, warning of potential dangers, and maintaining social cohesion. Researchers have identified and classified around 30 types of vocalizations in meerkats [2]. These vocalizations can be categorized into alarm calls emitted when a potential predator is encountered [3], contact calls used to maintain group cohesion [4], and dominance calls employed during a conflict to assert social hierarchy. Additional vocalizations serve to express various other emotions. These vocalizations are part of a complex communication system, influenced by the group's social organization and ecology [5].

Over the past two decades, there has been a notable improvement in understanding this communication system, particularly in decoding the context of calls. For example, in [6], it is demonstrated that meerkat alarm calls encode information about both predator type and the signaler's perception of urgency simultaneously. Additionally, in [7], it was found that close calls are used to adjust movement direction and maintain group cohesion, especially in low-visibility environments and during continuous movement. However, understanding the context precedes contextual analysis. The process of categorizing calls is mainly conducted by human listeners, who rely on their expertise. Nonetheless, even among these experts, varying interpretations may arise, highlighting the complexity inherent in the classification task [8].

Although previous research has provided insights into the social and contextual aspects of meerkat vocalizations, there remains a lack of computational methods for the automatic analysis of this language. Specifically, to the best of our knowledge, there has not been a formal study on the automatic classification of meerkat vocalizations. One of the main reasons being that biological level and linguistic level analysis of meerkat vocalizations has evolved more recently, leading to the availability of reliable data sets for automatic analysis. As a first step, the present paper aims to investigate feature representations for automatic meerkat vocalization analysis. The motivation for this arises from the important role feature representation plays in pattern analysis and classification systems. In the past, in the field of speech and audio processing, these representations were largely obtained by combining prior knowledge with signal processing. Even though meerkat vocalizations have been analyzed using signal processing, there is still a lack of reliable prior knowledge to extract feature representations for automatic analysis. In recent years, with advances in deep learning, data-driven feature representations have become more prominent and have been demonstrated useful for bioacoustic analysis. In this paper, we investigate both types of feature representations.

The remainder of the paper is organized as follows: Section 2 introduces the two types of feature representations, providing a detailed overview of the methods used. Section 3 delineates the experimental setup and workflow, including the dataset used during the study, the classification setup, and the evaluation metric. Section 4 presents the classification results with a comprehensive analysis of the findings. Finally, Section 5 concludes our study.

2. Feature representations

This section motivates and presents the different feature representations investigated in this paper. These representations are grouped as (a) knowledge-based/hand-crafted feature representations and (b) neural-based data-driven feature representations.

2.1. Knowledge-based/hand-crafted feature representations

Catch22: Highly Comparable Time-Series Analysis (HCTSA) is an interpretable signal processing-based framework, where a set of 7700 features are extracted by characterizing the signal by different time series analysis methods, such as linear correlation, modeling fitting (e.g., autoregressive moving average analysis, GARCH), wavelet analysis, and extraction of information theoretic measures. It is then combined with feature selection to build statistical models for the end task [9]. The efficacy of this framework has been demonstrated for bioacoustic analysis. For instance, these features have been investigated for behavioral birdsong discrimination [10], automated acoustic

monitoring of ecosystems [11], as well as marmoset caller identification [12]. One of the limitations of the HCTSA approach is computational complexity, as it involves the evaluation of many similar features. In recent work, CAnonical Time-series CHaracteristics (Catch22) features, a subset of 22 HCTSA features that are minimally redundant has been proposed, and its utility has been demonstrated across 93 real-world time-series classification problems [13]. These features fall into different conceptual grouping such as distribution shape, linear autocorrelation, incremental differences, and self-affine scaling. The dimension of the feature set is 24 including the mean and the standard deviation.

COMPARE: COMPARE features have been developed for paralinguistic speech processing. The COMPARE feature set of length 6373 consist of functionals of (a) energy related low level descriptors (LLDs), (b) spectral LLDs, and (c) voicing related LLDs estimated over an utterance [14].

eGeMAPS: extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is yet another feature set developed for paralinguistic speech processing [15]. The feature set consists of 88 different features. They are obtained by extracting (a) LLDs, namely, frequency-related parameters, energy/amplitude related parameters, and spectral (balance) parameters, and (b) temporal features consisting of the rate of loudness peaks, mean length and standard deviation of voiced and unvoiced regions, and number of continuous voices regions per second from the acoustic signal.

2.2. Neural-based data-driven feature representations

Self-supervised learning-based: In traditional supervised learning, models rely on labeled data, which is expensive and time-consuming to obtain. Thus, the emergence of self-supervised learning (SSL) techniques offers a powerful alternative to these learning methods by leveraging unlabeled data and designing pretext tasks involving human speech. By doing so, it allows models to learn meaningful representations without relying on explicit human annotations. In [16], the authors explored leveraging embedding spaces focusing on the Marmoset caller discrimination. The study demonstrated that representations pre-trained on human speech could be effectively applied to the bio-acoustics domain. Motivated by that study, we chose three popular SSL models, namely, WavLM [17], wav2vec2 [18] and HuBERT [19], pre-trained with 960 hours of audio from Librispeech corpus [20]. We extract embeddings from one of the layers or all layers of the SSL model and model it for call classification.

Supervised-learning based (denoted as CNN-crafted): In this part, we focus on the feature extraction phase within a classification framework. This involves directly inputting waveform data into a neural network using an end-to-end Convolutional Neural Network (CNN) architecture. The architecture is inspired by [21] and is presented in Table 4. The model is trained to perform call type classification. After training, we derive a feature set of dimension 80 from each call by extracting the output of the penultimate layer of the model, referred to as CNN handcrafted features throughout the study.

3. Experiments

This section presents the dataset of our study, consisting of two Sets (A and B) of meerkat calls used during the study, followed by a detailed breakdown of the study's workflow.

3.1. Meerkat calls dataset

Set A consists of 90 audio recordings of 9 different meerkat call types collected and labeled by Prof. Marta Manser, University of Zurich, following ethical approval: Aggression (agg), Sentinel (sen), Alarm (al), Chatter (ch), Grooming (gr), Close-call (cc), Submission (sub), Lead (ld) and Sunning (su). Every file was manually segmented using Koe [22]; an open-source software to visualize, segment, and classify acoustic units in animal vocalizations, amounting to a total of 1795 vocalization segments at a sampling rate of $44.1\,\mathrm{kHz}$, with a mean and median length of $161 \pm 118\,\mathrm{ms}$ and $102\,\mathrm{ms}$ respectively. Table 1 shows the distribution of the different call types of Set A. It is crucial to emphasize that this table reveals a significant imbalance within the dataset, mirroring the real-world scenario.

Table 1: *Distribution of the different call types present in Set A.*

agg	sen	al	ch	gr	cc	sub	ld	su
125	411	609	108	12	81	99	28	322

Set B is a public dataset [23]. The corpus consists of 6428 individual files, categorized into 7 call types, sampled at $48\,\mathrm{kHz}$ with a mean of $148 \pm 96\,\mathrm{ms}$ and a median of $124\,\mathrm{ms}$. Four classes seen previously in Set A are also present in Set B, with three additional ones: Short note (sn), Social call (sc), and Move (mv). Table 2 displays the distribution of the different call types in Set B.

Table 2: *Distribution of the different call types present in Set B.*

agg	cc	al	ld	sn	soc	mo
375	1477	645	164	1854	1154	759

3.2. Experimental set-up

As a preprocessig step, we downsampled all waveforms to $16\,\mathrm{kHz}$ and vocalizations shorter than $100\,\mathrm{ms}$ were systematically replicated until they reached the desired minimum duration of 100 ms. To compare the feature representations, we adopted a 5-fold cross-validation strategy by employing 80:20 train-test split. Figure 1 shows the call classification framework. As illustrated in the figure, a call-level fixed length representation is obtained for each feature type and fed as input to a support vector machine (SVM) based classifier. We compare the feature representations by evaluating the respective call classifiers in terms of unweighted average recall (UAR). We chose UAR as metric due to class imbalance in the datasets. Unlike weighted average accuracy (classification accuracy), UAR measure gives importance to recognition of all classes. Higher UAR means higher recall across classes. When training the SVM classifier, we applied a grid search methodology on the training set of each fold with the Unweighted Average Recall (UAR) as the optimization criterion to search space of the hyperparameters (presented in Table 3). In the reminder of the section, we explain the call-level fixed length representation obtained for each feature representation type.

In the case of knowledge-based feature representation, (a) *pycatch22* toolkit was employed for extracting 24 dimensional call-level Catch22 features and (b) openSMILE [24] tool is used to extract 6373 dimensional call-level COMPARE feature representation and to extract 88 dimensional call-level eGeMAPS feature representations.

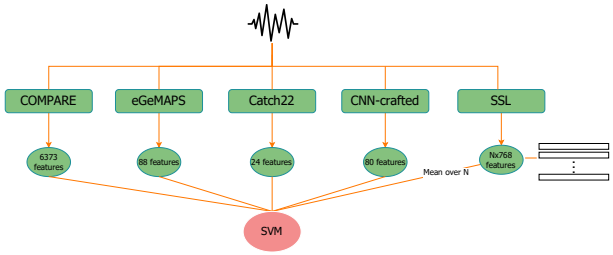

Figure 1: *Diagram of the workflow of the study. N denotes number of frames.*

Table 3: *SVM hyperparameters grid*

Parameter	**Values**
C	1e[-1, 0, 1, 2]
Gamma	1e[-3,-2,-1,0]
Kernel	[‘Linear’, ‘RBF’, ‘Polynomial’, ‘Sigmoid’]

In the case of SSL feature representations, the call-level 768 dimensional feature representation is obtained as follows: (a) 768 dimension output of CNN encoder, 1^{st}, 2^{nd}, 6^{th} or the last transformer layer is obtained per frame and averaged over frames, (b) the 768 dimension output of each of the 12 transformer layers are averaged per frame and then the resulting per frame representation is averaged over frames. The S3PRL toolkit [25] was used to extract the embeddings.

In the case of CNN-crafted feature representation, there is a need to train a CNN-based call classifier for feature extraction. As the data sets were small in size with severe class imbalance, as opposed to training a CNN feature extractor per fold, we employed stratified k-folds cross-validation strategy to get a single CNN feature extractor. This method constructs folds while maintaining class proportion integrity, i.e., ensuring consistent class proportions in both training and test sets, mirroring those of the original dataset. We set the number of folds to 5 and trained CNNs for each fold using the architecture presented in Table 4 using PyTorch. The adaptive average layer target size was set to one. This allows the network to handle variable length waveform inputs and yield fixed-length (80-dimensional) call level feature representation. We employed the cross-entropy error criterion to train the CNN. The CNN of the best performing fold was selected to extract 80 dimensional call-level CNN-crafted feature representation (from the output of the fully connected hidden layer).

4. Results and discussion

Table 5 presents an analysis of SSL neural embeddings. It can be observed the that lower layer transformer layer embeddings and CNN encoder representations yield better systems than higher layer transformer layer embeddings. Averaging the embeddings across the transformer layers, although yields better system than layer 6 and last layer embeddings, is not helpful when compared to layer 1 embedding, layer 2 embedding or CNN encoder output alone. Taken together, this indicates that lower transformer layer embeddings of SSLs pre-trained on human speech are more informative than higher transformer layer embeddings for meerkat call classification.

Table 6 compares the systems across different feature representations. For SSL feature representation wav2vec2, WavLM and HuBERT, we have reported the best system performance

Table 4: *CNN architecture for CNN-crafted feature extraction. n_f denotes number of filters. HU denotes number of hidden units.*

Block	Operation	Kernel	Stride	Padding	n_f/HU
1	Convolution	40	30	0	40
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
2	Convolution	7	1	0	40
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
3	Convolution	3	1	0	80
	Max Pooling	2	2	0	-
	ReLU Activation	-	-	-	-
4	Adaptive Avg Pooling	-	-	-	-
	Flatten	-	-	-	-
	Fully Connected	-	-	-	80

Table 5: *UAR scores of chosen representations using wav2vec2 (W2), WavLM (WL) and HuBERT (HT) models on Test set of Set A and B*

Model	Set A			Set B		
	W2	WL	HT	W2	WL	HT
CNN	0.71	0.68	0.74	0.78	0.77	0.78
1^{st} Transformer	0.71	0.72	0.73	0.79	0.82	0.78
2^{nd} Transformer	0.73	0.71	0.72	0.79	0.82	0.79
6^{th} Transformer	0.54	0.50	0.64	0.69	0.70	0.76
Last Transformer	0.35	0.38	0.55	0.52	0.53	0.67
Average of Transformers	0.63	0.59	0.61	0.75	0.72	0.76

from Table 5. In the case of hand-crafted features, it is observed that eGeMAPS and COMPARE feature based systems yield better system than Catch22 feature representation. In the case of SSL feature representations, the systems are comparable. The CNN-crafted feature representation yields the best systems. When comparing hand-crafted features and neural embeddings, COMPARE feature outperforms SSL features on Set A and performs slightly worse when compared to wav2vec2 and HuBERT. It is worth pointing out that the COMPARE feature largely outperforms higher transformer layer embedding based systems (layer 6 and last layer in Table 6). This indicates that, similar to neural embeddings from networks pre-trained on human speech, hand-crafted representations developed for speech processing applications can be useful for meerkat call classification.

Table 6: *UAR scores on Test set of Set A and B with 5-fold CV for call types classification*

Model	Set A	Set B
eGeMAPS	0.61	0.66
COMPARE	0.80	0.75
Catch22	0.61	0.56
wav2vec2	0.73	0.79
WavLM	0.72	0.82
HuBERT	0.74	0.79
CNN-crafted	**0.84**	**0.84**

The main distinction between Set A and Set B lies in the number of classes, the number of samples, and the class dis-

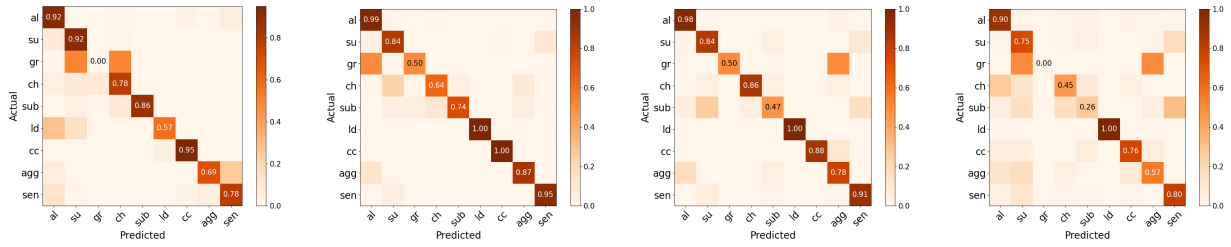

Figure 2: *Confusion matrices for SVM classifier using, from left to right, WavLM, CNN-crafted, COMPARE and Catch22 embeddings on the test set of Set A.*

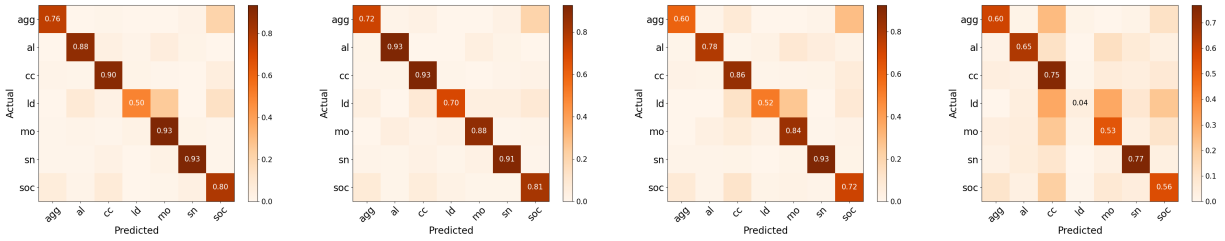

Figure 3: *Confusion matrices for SVM classifier using, from left to right, WavLM, CNN-crafted, COMPARE and Catch22 embeddings on the test set of Set B.*

tribution within the datasets. As discussed previously, Set B comprises more samples, fewer number of classes and exhibits better class balance than Set A. Therefore, our initial expectation was that Set B would yield superior performance. This hypothesis is confirmed with the SSL models, eGeMAPS, and the CNN model, where results with Set B perform better than Set A. Confusion matrices for WavLM, CNN-crafted, COMPARE and Catch22 are presented for Set A and Set B in Figure 2 and Figure 3. It can be observed that all the call types are mostly classified well except for "gr" in Set A which has the lowest amount of data.

For the case of CNN-crafted, Figure 4 shows the cumulative frequency response of the 40 first layer convolution filters. This is estimated by applying a DFT of 1024 points on filters of length 40 samples and taking logarithm of the summed magnitude responses. Although Set A and Set B have been collected independently and labeled, it can be observed that the cumulative filter responses of the CNNs of Set A and Set B are similar with a major emphasis between 0-2 kHz. This indicates that the CNNs are capturing information systematically for class classification across the two data sets. In our future work, we will investigate what kind of acoustic information does that frequency range carries in meerkat vocalizations for call analysis.

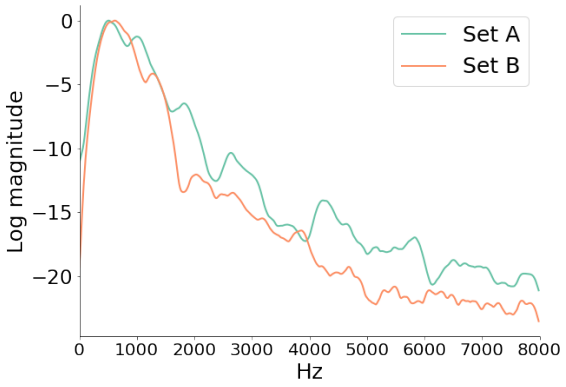

Figure 4: *Cumulative frequency responses of first layer filters of CNN*

5. Conclusions

Meerkats with their highly social nature and diverse vocal repertoire, provide an intriguing model system for investigating animal communication and, as an extension could help us better understand the evolution of human communication. One of the challenges in that direction is the lack of methods for automatic meerkat call analysis. In that direction, this paper explored feature representations for automatic analysis of meerkat vocalizations. We compared time-series analysis-based hand-crafted feature representation, hand-crafted feature representations developed for human speech processing, SSL-based feature representations obtained from neural networks trained on human speech, and feature representations automatically learned in a task-dependent manner from meerkat calls using CNNs. Our studies show that hand-crafted feature extractors and SSL feature extractors developed for human speech processing can be used for meerkat call classification. Similarly, we observe that the CNN-based method developed for automatic feature learning in a task-dependent manner for human speech processing can be scaled for meerkat call classification task (CNN-crafted). Our future work will focus on analyzing these diverse feature representations to tease out and explain the acoustic information that is relevant for meerkat call analysis.

6. Acknowledgement

The first author carried out a part of this work at Idiap as part of her Master-AI thesis at UniDistance, Switzerland. This work was partially funded by Swiss National Science Foundation's NCCR Evolving Language project (grant no. 51NF40 180888).

7. References

[1] J. R. Madden, J. A. Drewe, G. P. Pearce, and T. H. Clutton-Brock, "The social network structure of a wild meerkat population: 2. intragroup interactions," *Behavioral Ecology and Sociobiology*, vol. 64, no. 1, pp. 81–95, Nov 2009. [Online]. Available: https://doi.org/10.1007/s00265-009-0820-8

[2] S. W. Townsend, B. D. Charlton, and M. B. Manser, "Acoustic cues to identity and predator context in meerkat barks," *Animal Behaviour*, vol. 94, pp. 143–149, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003347214002413

[3] G. Moran, "Vigilance behaviour and alarm calls in a captive group of meerkats, suricata suricatta," *Zeitschrift für Tierpsychologie*, vol. 65, no. 3, pp. 228–240, 1984. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0310.1984.tb00101.x

[4] S. Engesser and M. B. Manser, "Collective close calling mediates group cohesion in foraging meerkats via spatially determined differences in call rates," *Animal Behaviour*, vol. 185, pp. 73–82, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003347221003997

[5] 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. Academic Press, 2014, vol. 46, pp. 281–310. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128002865000067

[6] M. B. Manser, R. M. Seyfarth, and D. L. Cheney, "Suricate alarm calls signal predator class and urgency," *Trends in Cognitive Sciences*, vol. 6, no. 2, pp. 55–57, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661300018404

[7] G. E. C. Gall and M. B. Manser, "Group cohesion in foraging meerkats: follow the moving 'vocal hot spot'," *R. Soc. Open Sci.*, vol. 4, p. 170004, 2017.

[8] A. Kershenbaum, D. T. Blumstein, M. A. Roch, Çağlar Akçay, G. Backus, and M. A. B. et al., "Acoustic sequences in non-human animals: a tutorial review and prospectus," *Biological Reviews*, vol. 91, no. 1, pp. 13–52, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12160

[9] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of The Royal Society Interface*, vol. 10, no. 83, p. 20130048, 2013. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0048

[10] P. Avishek, H. McLendon, V. Rally, J. T. Sakata, and S. C. Woolley, "Behavioral discrimination and time-series phenotyping of birdsong performance," *PLoS Comput. Biol.*, vol. 17, no. 4, p. e1008820, Mar. 2021.

[11] S. S. Sethi, R. M. Ewers, N. S. Jones, A. Signorelli, L. Picinali, and C. D. L. Orme, "Safe acoustics: An open-source, real-time eco-acoustic monitoring network in the tropical rainforests of borneo," *Methods in Ecology and Evolution*, vol. 11, no. 10, pp. 1182–1185, 2020. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13438

[12] N. Phaniraj, K. Wierucka, Y. Zürcher, and J. M. Burkart, "Who is calling? optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers," *Journal of The Royal Society Interface*, vol. 20, no. 207, p. 20230399, 2023. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2023.0399

[13] C. H. L. S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics," *Data Mining and Knowledge Discovery*, vol. 33, no. 6, pp. 1821–1852, Nov 2019. [Online]. Available: https://doi.org/10.1007/s10618-019-00647-x

[14] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. Interspeech 2016*, 2016, pp. 2001–2005.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, and C. B. et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[16] E. Sarkar and M. Magimai.-Doss, "Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?" in *Proc. INTERSPEECH 2023*, 2023, pp. 1189–1193.

[17] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, and Z. e. a. Chen, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[21] 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, 2019.

[22] Y. Fukuzawa, W. H. Webb, M. D. Pawley, M. M. Roper, S. Marsland, D. H. Brunton, and A. Gilman, "Koe: Web-based software to classify acoustic units and analyse sequence structure in animal vocalizations," *Methods in Ecology and Evolution*, vol. 11, no. 3, pp. 431–441, 2020. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13336

[23] M. Thomas, F. H. J. B. Averly, V. Demartsev, M. B. Manser, T. Sainburg, M. A. Roch, and A. Strandburg-Peshkin, "A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations," *Journal of Animal Ecology*, vol. 91, no. 8, pp. 1567–1581, 2022. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.13754

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[25] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, and Y. Y. L. et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

Western Jackdaw Call Classification in Noisy Environments Using CNNs

Bilal Sardar[1], *Lakshmi Babu Saheer*[1], *Sam Reynolds*[1]

[1]Anglia Ruskin University, Cambridge, United Kingdom

BS850@student.aru.ac.uk,lakshmi.babu-saheer@aru.ac.uk,srdr102@pgr.aru.ac.uk

Abstract

The western jackdaw (Corvus/Coloeus monedula) is a passerine bird found across Europe. Automated detection and classification of jackdaw calls from audio recordings can support population monitoring and behavioral studies. However, background noise presents significant challenges. This research presents multiple deep-learning approaches for jackdaw call classification robust to realistic environmental noise. Experiments are performed with multiple deep-learning models using different features including custom convolutional neural network (CNN) models using MFCC and spectrogram features, pre-trained BirdNet, InceptionV3, Xception, ResNet50 models using spectrograms, LSTM-based RNN models using MFCCs and pre-trained transformer models using raw waveforms(Wav2Vec2). Tests performed on a manually curated dataset of jackdaw calls and noise segments extracted from field recordings achieve the best performance of 98% accuracy on the validation set using custom CNN and BirdNet models. Further manual validation of extracted Jackdaw calls on unlabeled raw field data shows a precision score of 93%. This research also presents data balancing and aggressive noise filtering to improve model generalization under varying real-world noise.

Index Terms: bird call recognition, MFCCs, spectrogram analysis, deep learning

1. Introduction

Automated bird species recognition from recordings is critical for scalable avian ecology research and conservation efforts [1, 2]. Manual field surveys are limited in scope, time-intensive, and susceptible to observer biases. In contrast, automated methods based on deep neural networks now rival expert-level performance in biodiversity monitoring from audio data [3]. Specifically, convolutional neural networks (CNNs) demonstrate state-of-the-art capabilities in classifying bird vocalizations to species [4, 5, 6]. Key enablers include using spectrogram image representations of audio data as input [7], thereby leveraging transfer learning from extensive image recognition research[8, 9].

However, background noise remains the primary impediment to accurate classification, causing pervasive false detections and inaccuracies [10]. Interfering sounds like wind, rain, machinery, human activity, and calls from other species present in real-world field recordings introduce major challenges. Addressing this requires developing robust algorithms specifically tailored for noisy conditions frequently encountered in ecosystem monitoring. While prior innovations demonstrate promise in improving noise resilience on controlled single-species audio datasets, they struggle to bridge the gap to uncontrolled natural soundscapes with complex noise profiles [6, 11].

Effective noise reduction is crucial for accurate species identification. A novel tri-layered approach enhances bird call clarity by integrating three filtration techniques: high-pass filtering to remove low-frequency noise, Per-Channel Energy Normalization (PCEN) to emphasize key audio features, and spectral gating to eliminate background noise. Additionally, a thresholding mechanism based on the signal-to-noise ratio (SNR) discards noisy samples, ensuring clear audio suitable for precise bird-call classification [12]. Another innovative technique for acoustic bird species classification under low SNR conditions uses an adaptive threshold based on the constant false alarm rate (CFAR). This method dynamically adjusts the segment energy threshold to robustly detect bird sounds amidst varying noise levels, significantly improving classification performance in challenging environments [13].

These methods face limitations due to differences between training and real-world conditions. Reliable species classification in noisy environments requires more research, integrating ecology, acoustics, and machine learning. Identifying Corvid calls from raw, unlabelled field data can be improved using a semi-supervised model, where a small amount of labelled data trains advanced deep learning models[14].

Corvids (Family: Corvidae) contain the crows, rooks, jackdaws, ravens, jays, magpies, treepies and nutcrackers. They are known for their complex vocalizations and communication skills. Audio recognition technology, where it can be readily used to identify and analyze corvid vocalizations, may represent a cornerstone to unlocking further research into their ecology. Our research contributes towards this aim, developing a tailored neural network pipeline for recognizing western jackdaw (Corvus/Coloeus monedula) vocalizations under varying noise. We integrate state-of-the-art techniques from literature to design an accurate jackdaw classifier serving as a case study for conservation-focused bioacoustic monitoring. The novelty of this research apart from building the first ever dedicated robust jackdaw call recognition model comparing different state-of-the-art deep learning architectures is to test it on real-world noisy field recordings with very limited data labelled for training the models.

2. Methodology

Mel-frequency cepstral coefficients (MFCCs) are widely used audio features for many audio recognition tasks. This feature extraction technique applies the mel-scale filter-banks to amplify lower frequency components on the logarithmic power spectrum and then transforms these to cepstral coefficients using the inverse Fourier transform.

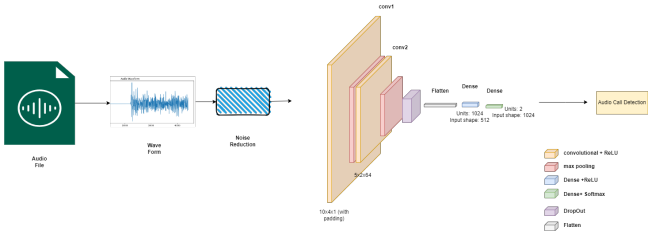

Figure 1: *Custom CNN Model Architecture with audio features*

2.1. Noise Reduction

The noise reduction technique employed in this research utilizes a threshold-based low-pass filtering approach in the frequency domain to improve the signal-to-noise ratio of audio signals by setting frequency components within ±30% of the Nyquist frequency to zero. Initially, the Fast Fourier Transform (FFT) is applied to convert the input audio signal from the time domain to the frequency domain. Subsequently, a copy of the FFT of the signal is created, and a threshold-based filtering operation is performed, where frequency components beyond a specified threshold are set to zero. This threshold is checked for values ranging from 0.1 to 1, and is set to 0.3 based on experiments showing optimal performance. This effectively eliminates high-frequency noise from the signal. Following this, the inverse FFT is applied to transform the filtered signal back to the time domain, yielding a cleaner version of the original audio signal. The benefits of this technique are manifold: it enhances signal quality by reducing distortion, improves feature extraction accuracy, particularly when using methods like Mel-frequency cepstral coefficients (MFCCs), and leads to better performance of subsequent processing tasks such as training Convolutional Neural Networks (CNNs). Additionally, the noise reduction(figure 2) enhances the robustness of models to environmental variations, contributing to better generalization on unseen data. Overall, this technique serves as a valuable preprocessing step in audio signal processing workflows, facilitating more accurate and effective analysis (check figure 3) and modeling of audio data. Given:

- Input audio signal: y
- Sample rate: sr
- Threshold for filtering: th

1. **Compute the Fast Fourier Transform (FFT) of the input signal:**

$$y_f = \text{FFT}(y)$$

2. **Define the frequency domain components:**

Total number of samples: $N = \text{int}(sr \times \text{DURATION})$

Frequencies for FFT bins: $xf = \text{fftfreq}(N, \frac{1}{\text{SAMPLE\_RATE}})$

3. **Apply a low-pass filter by setting the frequency components beyond a certain threshold to zero:**

Create a copy of the FFT of the input signal:

$\text{new\_yf} = y_f.\text{copy}()$

Define the middle index: $\text{middle} = \frac{\text{len}(y)}{2}$

Set the frequency components beyond the threshold to zero:

$\text{new\_yf}\,[\text{int}\,(\text{middle} - \text{len}(y) \times th) : \text{int}\,(\text{middle} + \text{len}(y) \times th)] = 0$

4. **Compute the inverse FFT to obtain the filtered signal in the time domain:**

$$\text{new\_y} = \text{IFFT}(\text{new\_yf})$$

$$\text{new\_y} = \text{real}(\text{new\_y})$$

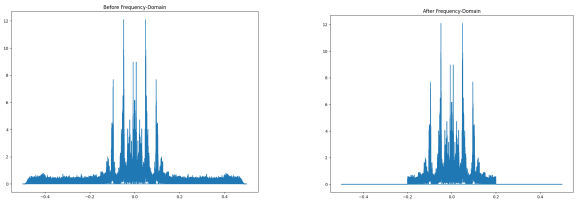

Figure 2: *Before And After Noise Reduction*

2.2. Model Architectures

The custom model used here is a Convolutional Neural Network (CNN) architecture(figure 1) designed for audio classification tasks. It consists of several convolutional layers that apply filters to the input data, extracting meaningful features through a series of convolutions and max pooling operations. The convolutional layers are followed by dropout layers to prevent overfitting during training. The extracted features are flattened and passed through fully connected (dense) layers, culminating in a final layer with two neurons and a softmax activation function. This output layer allows the model to classify the input data into one of the two categories. The model's strength lies in its ability to automatically learn hierarchical representations of the input data, making it well-suited for tasks like audio classification. CNNs are particularly effective for processing data with spatial or temporal dependencies, such as audio spectrograms or images. By leveraging local connectivity and weight sharing, CNNs can efficiently capture patterns and learn robust feature representations. Additionally, the use of max pooling layers helps to reduce computational complexity and introduces invariance to small shifts or distortions in the input data. Overall, the CNN architecture is a powerful and widely used approach for audio analysis tasks, capable of learning complex patterns and achieving high performance in various audio classification problems. However, the proposed CNN assumes an input with a fixed length, which means that one can not just feed the entire long audio signal to the model for detection. Detection is performed by using a moving window and classifying each window with the model. This should be specified for clarity.

The BirdNet sound classification model, which is a pretrained convolutional neural network (CNN) designed for bird audio, utilizes dual-spectrogram inputs to offer complementary perspectives of the raw audio waveform. It processes 48kHz audio, resampling if necessary, and generates two mel-scale spectrograms to capture both low and high-frequency details. Like the custom CNN model, BirdNet expects a fixed-length input, so detection is carried out by sliding a window over the audio signal and classifying each segment. This approach needs to be explicitly outlined.

The first spanning 0-3kHz uses a 2048-point FFT with 278 hop size and 96 mel bins. The second from 500Hz-15kHz uses a 1024-point FFT with 280 hop size, also with 96 mel bins. Both have a resulting dimensionality of 96x511. Raw audio is normalized between -1 and 1 before spectrogram conversion. A

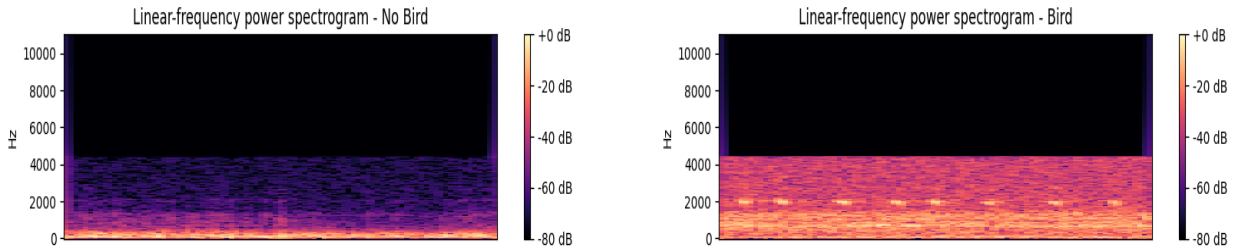

Figure 3: *Linear-frequency Power Spectrogram for No Bird and Jackdaw Call*

nonlinear mapping is applied to the magnitude spectra as described in [15] to improve sound event detection. The dual-spectral input provides a rich initial representation of the data to the deep neural network.

The backbone classification architecture is EfficientNetB0-inspired, employing inverted residual blocks with squeeze-and-excitation modules. It processes the multi-resolution input through several convolutional layers, followed by global average pooling to produce a 1024-unit embedding vector representing the input audio clip. This is finally classified using a linear layer to predict bird species. The dual-spectrogram input in conjunction with an efficient deep neural network allows accurate detection and classification of bird vocalizations from raw waveform audio recordings. The model is robust to sampling rate variation and represents both low and high-frequency content useful for identifying bird sounds.

3. Experiments and results

Experiments are performed on the field recordings using multiple deep-learning models.

3.1. Datasets

Models were trained and tested from data collected from a colony of western jackdaws in a rural setting in Sweden. The data were collected using Audiomoth [16] autonomous recording units between March - July 2023, as part of an ongoing bioacoustic study. Recordings were made at 48000Hz with a bit rate of 16 bits in wave format. Recording units were deployed at a range of 5-10m from occupied nest boxes. A part of the dataset ($\tilde{9}$ hours) was manually labeled with calls and other background classes to be modeled as a binary class task. The labeled data consists of 2576 samples (128 mins) of Jackdaw calls and 14867 samples (743 mins) of false positive background audio. This labeled dataset was split in the ratio of 80/20 for the train-test split and a separate cross-validation set within the train set to estimate the best-performing hyperparameters. Further, $\tilde{1}$7.7 hours of raw field recordings are used as unlabelled test data to manually validate the models for future deployment.

3.2. Experimental conditions

Multiple deep-learning models are being implemented for building a Jackdaw call recognition model. This includes both custom models and fine-tuned pre-trained models using both audio features and spectrogram-based image features. The custom CNN models were trained for MFCC audio features as a 1-D CNN model or for spectrogram image features as a 2-D CNN model. Similarly, multiple pre-trained CNN models (like ResNet50, InceptionV3, and Xception), including the popular BirdNet model are fine-tuned using the data. Experiments are also performed on sequence models like LSTMs and pre-trained transformer models. This work builds upon existing research that comprehensively covers deep learning methods for audio classification, including CNN-based, RNN-based, and hybrid models. The referenced paper provides a detailed overview of these techniques and their applications to various audio datasets, including discussions on feature extraction methods and model training[17].

3.3. Experimental result

As mentioned earlier, the model performances are reported on the held-out test set. The range of hyperparameters tuned for the models is depicted in Table 1. The initial experiments are performed using data balancing and noise reduction techniques as shown in table 2 on the custom CNN model. It is observed that noise reduction is a key step in improving performance and data balancing may not have as much impact as expected. Considering this, multiple models are trained using noise reduction on unbalanced data as shown in table 3. The table shows accuracy (as Acc), precision (as P), recall (as R) and F1-score (as F1) on the labeled test set. The results show that the spectrogram features show slightly better performance than MFCC features. BirdNet model and the custom CNN models both make use of spectrogram features and seem to be the best models that are comparable.

It can be observed from the table that overall the best-performing model is the custom CNN model using the spectrogram features which shows good performance across all metrics. BirdNet model has a similar performance with a slightly reduced precision and f1-score. CNN model using MFCC features (Fig 4) is comparable to the InceptionV3 and Xception models for all error metrics, with slightly lower recall scores. ResNet50 spectrogram models and the LSTM-based RNN models using MFCC features are both low in recall and f1-score (Fig 5) compared to other models. While the ResNet50 model seems to have a very high precision score. This model could be overfitting and hence unable to extract all the Jackdaw calls in the validation set. The transformer model using the raw waveforms (using wav2vec) also does not match the performance of the custom CNN and BirdNet models but has consistent scores across all metrics including accuracy, precision, recall, and f1-score(Fig 6). More fine-tuning and optimisation might be needed to improve the transformer model performance.

The model's detection accuracy is validated by an ornithologist specializing in western jackdaw vocalizations. The fine-tuned BirdNet model identified 471 Jackdaw calls from 17.7 hours of raw field data. Predictions with a confidence score above 50% were filtered to 248 calls for manual validation, yielding about 230 accurate calls and a precision of 93%.

Table 1: *Hyperparameters optimised*

Hyperparameter	**Range of Values**
Number of CNN-layers	[2 to 5]
Optimizer	['adam', 'rmsprop', 'sgd']
Learning Rate	[0.00001, 0.0001, 0.01, 0.1]
Batch Size	[32, 64, 128]
Epochs	[10, 50, 400]
Dropout Rate	[0.0, 0.25, 0.33, 0.5, 0.75, 0.9]
MaxPool2D Pool Sizes	[(2, 2), (3, 3)]
Conv2D Filter Sizes	[(3, 3), (5, 5)]
Dense Units	[512, 1024, 2048]

Table 2: *Noise Reduction and Data Balancing Experiments*

Model	**Technique**	**Data**	**NR**	**Acc(%)**
CNN	MFCC	Unbalanced	✓	96.23
CNN	MFCC	Balanced	✓	95.51
CNN	MFCC	Unbalanced	✗	90.11
CNN	MFCC	Balanced	✗	91.01

Misidentified clips were mostly caused by noise or wind. Future improvements will include training with additional false-positive sounds and adjusting confidence thresholds to balance missed detections and false positives.

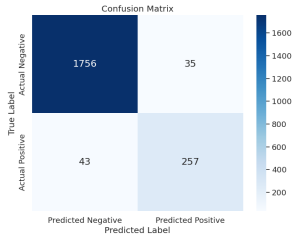

Figure 4: *CNN MFCC Confusion Matrix*

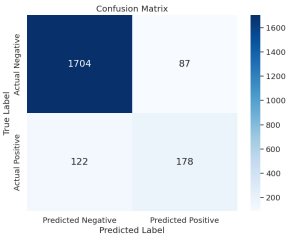

Figure 5: *LSTM MFCC Confusion Matrix*

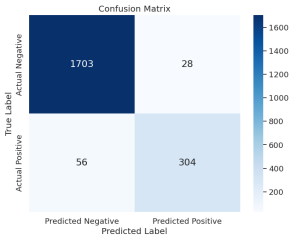

Figure 6: *Wav2Vec2 Transformer Confusion Matrix*

3.4. Performance on Popular Bird Classification Datasets

To further validate the effectiveness of our CNN model, we conducted additional experiments using two widely recognized bird classification datasets: BirdCLEF 2024 and warblrb10k. These datasets provide a diverse range of bird vocalizations, allowing us to assess the model's generalization capabilities beyond our primary western jackdaw dataset.

Table 3: *Comparing multiple deep learning models*

Model	**Feature**	**Acc(%)**	**P(%)**	**R(%)**	**F1**
CNN	MFCC	96.23	94.91	87.23	92.10
BirdNet	Spectrogram	98.31	97.94	98.03	97.98
CNN	Spectrogram	98.91	98.90	98.93	98.91
ResNet50	Spectrogram	89.45	1.0	65.51	79.16
InceptionV3	Spectrogram	95.36	98.49	88.43	93.19
Xception	Spectrogram	93.67	93.39	89.01	91.39
LSTM	MFCC	90.00	80.02	77.20	78.60
Transformers	Wav2Vec2	89.31	89.21	89.33	89.31

Table 4: *CNN Model Performance on Popular Bird Classification Datasets*

Dataset	**Mean ROC AUC Score(%)**	**Baseline Results**
BirdCLEF 2024	**78.69**	74.49[18]
Warblrb10k	81.51	**90.18**[19]

3.4.1. BirdCLEF 2024 Dataset

The BirdCLEF 2024 dataset, part of the annual BirdCLEF challenge, focuses on identifying 182 bird species from audio recordings. It includes diverse species from various regions, testing the model's ability to distinguish multiple species.

3.4.2. Warblrb10k Dataset

The warblrb10k dataset, a benchmark for bird audio detection, contains 8,000 UK smartphone recordings crowdsourced via the Warblr app. It includes diverse environmental noises, making it challenging to detect bird sounds amidst weather, traffic, and human noise.

3.4.3. Results

CNN model optimized for western jackdaw dataset is fine-tuned with the training sets of BirdCLEF 2024 and warblrb10k datasets to validate the architecture's effectiveness on standard datasets. Table 4 summarizes the test results of our model, alongside the baseline accuracy from GitHub [18, 19]. Our CNN model generalizes well to bird classification tasks, achieving a mean ROC AUC score of 78.69% on multi-species (BirdCLEF2024) and 81.51% on binary classification (warblrb10k).

4. Conclusion

Our study addresses the automated detection and classification of western jackdaw calls, crucial for population monitoring and behavioral research in Europe. With a limited amount of labelled raw field data, the research aims to build a semi-supervised few-short learning system to label the data. Through a convolutional neural network (CNN) approach robust to environmental noise, we achieve over 98% validation accuracy on a curated dataset of jackdaw calls and noise segments from field recordings. Multiple models are being evaluated using MFCC and spectrogram features. The deep learning models studied include custom CNN models and multiple pre-trained models, BirdNet, ResNet50, InceptionV3, and Xception models. LSTM-based sequence models and pre-trained transformer models are also experimented with.The BirdNet and custom CNN with spectrogram features performed best on labeled test data. Data balancing and noise filtering improved generalization in noisy conditions, advancing avian monitoring and setting a benchmark for future avian audio classification studies.

5. References

[1] S. D. H. Permana, G. Saputra, B. Arifitama, W. Caesarendra, R. Rahim *et al.*, "Classification of bird sounds as an early warning method of forest fires using convolutional neural network (cnn) algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4345–4357, 2022.

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[3] J. Stastny, M. Munk, and L. Juranek, "Automatic bird species recognition based on birds vocalization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–7, 2018.

[4] A. Incze, H.-B. Jancsó, Z. Szilágyi, A. Farkas, and C. Sulyok, "Bird sound recognition using a convolutional neural network," in *2018 IEEE 16th international symposium on intelligent systems and informatics (SISY)*. IEEE, 2018, pp. 000 295–000 300.

[5] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European signal processing conference (EUSIPCO)*. IEEE, 2017, pp. 1744–1748.

[6] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different cnn-based models for improved bird sound classification," *IEEE Access*, vol. 7, pp. 175 353–175 361, 2019.

[7] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio-based bird species identification using deep learning techniques," *LifeCLEF 2016*, pp. 547–559, 2016.

[8] M. Sankupellay and D. Konovalov, "Bird call recognition using deep convolutional neural network, resnet-50," in *Proc. Acoustics*, vol. 7, no. 2018, 2018, pp. 1–8.

[9] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin, F. H. Jensen, J. Morford, M. G. Emmerson, E. Versace, E. Grout, H. Liu, B. Ghani, and D. Stowell, "Learning to detect an animal sound from five examples," *Ecological Informatics*, vol. 77, p. 102258, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495412300287X

[10] J. Benesty, J. Chen, and Y. A. Huang, "Noise reduction algorithms in a generalized transform domain," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 6, pp. 1109–1123, 2009.

[11] Z. Zhao, L. Yang, R.-r. Ju, L. Chen, and Z.-y. Xu, "Acoustic bird species classification under low snr and small-scale dataset conditions," *Applied Acoustics*, vol. 214, p. 109670, 2023.

[12] W. Ansar, A. Chatterjee, S. Goswami, and A. Chakrabarti, "An efficientnet-based ensemble for bird-call recognition with enhanced noise reduction," *SN Computer Science*, vol. 5, no. 2, p. 265, 2024.

[13] Y. Zhang and J. Li, "Birdsoundsdenoising: Deep visual audio denoising for bird sounds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 2248–2257.

[14] L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 23, 2022.

[15] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1764–1768.

[16] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, "Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment," *HardwareX*, vol. 6, p. e00073, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468067219300306

[17] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, 2023.

[18] S. Jha, "Birdclef 2024 dataset," 2024. [Online]. Available: https://github.com/skj092/kaggle-BirdCLEF-2024/tree/main

[19] Microfaune, "warblrb10k dataset," 2024. [Online]. Available: https://github.com/microfaune/microfaune

Zero-shot Avian Species Detection from Unlabelled Field Audio Data

Gayathri Singaram[1], *Lakshmi Babu Saheer*[2], *Dena Jane Clink*[3], *Roeun Sala*[4], *Moeurk Hong*[5], *Hélène Birot*[6]

[1] [2] Anglia Ruskin University, United Kingdom, [3] K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, USA, [4] [5] [6] Jahoo, Angdoung Kraleung Village, Sen Monorom Orang, MondulKiri, Mondulkiri Province, Cambodia.

gs886@student.aru.ac.uk, lakshmi.babu-saheer@aru.ac.uk, djc426@cornell.edu, research@gibbon.life, booking@gibbon.life, helene.birot@worldhope.org

Abstract

The automated identification of bird species is a highly complex task, as environmental noise and variations in bird sounds are present in field recordings. This study focuses on identifying 7 bird species from Cambodian field recordings. Due to the lack of labelled in-domain data, zero-shot learning approach is used. Custom & pre- trained models: BirdNet, YAMNet, VGGish, InceptionV3, ResNet50,Xception,RNN+LSTM & CNN are trained on Xeno- Canto & e-bird datasets as audio (MFCC and Chroma features) & spectrogram-based image tasks. This research reveals that MFCC features boost the accuracy of CNN in audio tasks, highlighting its adaptability. Custom models outperformed pre-trained models in audio feature tasks with 99% accuracy, while pre-trained models excel in spectrogram tasks with 92% accuracy on the validation sets. All models had low performance on test data, which could be attributed to the bias towards oriental pied hornbill & great hornbill species due to data imbalance.

Index Terms: Bioacoustics, Signal Processing, Bird Sound Identification, Deep Learning, CNN, RNN, LSTM, BirdNet, MFCC, Chroma features.

1. Introduction

Avian acoustics research, essential for understanding bird behavior and conservation, has progressed from traditional, labor-intensive sound detection methods which involves manual validation to advanced automatic recognition using deep learning and advanced signal processing, enhancing accuracy and efficiency. Recent advancements in bird vocalization analysis have highlighted semi-supervised learning and deep clustering as emerging yet evolving techniques for identifying bird species in field recordings. A recent study explores zero-shot learning (ZSL) for bird species identification using field guide illustrations, [1] introduces a contrastive encoding method and Prototype Alignment to map illustrations and photographs to a shared space. Testing on the iNaturalist2021 dataset, they achieve 12% top-1 and 38% top-10 accuracy for unseen species, showcasing the effectiveness of illustrations in ZSL, another review [2] emphasizes the importance of accurate preprocessing like adaptive denoising while exploring the evolution of feature extraction methods from manual to automated approaches (like CRNN and WaveNet along with methods such as feature fusion and network architecture search to enhance model performance. The review suggests combining visual (spectrogram images) and acoustic data to improve bird species identification using generalised models with larger datasets to aid biodiversity conservation.

Right features and feature fusion methods help the classification process. Studies have considered different feature-based approaches to classify bird sounds like MFF-ScSEnet, a novel method for bird song identification that combines Mel-filters and Sincnet-filters (filters that process raw audio) with the ResNet18 network and a ScSEnet attention module to enhance spectrogram analysis[3]. This approach significantly improves accuracy in acoustic feature extraction but struggles with mixed or limited samples. [4] introduces Multi-scale CNN and Ensemble Multi-Scale CNN models that use the wavelet transform for spectrogram feature generation, outperforming traditional CNNs in bird species recognition by capturing detailed frequency information.

Bird sound detection faces challenges like environmental noise and the complexity of various vocalizations especially in a 24X7 field recording setting.Researchers [5, 6] have highlighted the acoustic diversity and complexity among bird species, noting variations due to geographical differences and recording distances, leading to high intra-species variation, which complicates accurate species characterization. Researchers [7, 8] address the issue of limited and unbalanced training data in bird sound recognition, emphasizing the need for large, representative datasets to prevent model overfitting and to capture species variability, pointing out the critical gap in the availability of verified datasets for classifier training.

Different studies on bird sound recognition have been performed using audio and spectrogram images: [9] presents ”Transound”,leveraging a vision transformer and MFCC for effective bird mitigation at airports. [10] combined CNNs with a transformer encoder, enhancing bird sound identification by using multiple acoustic features. [11] proposes the SFLN (spectrogram frame linear network) method for classifying bird sounds based on continuous frame sequences and spectrogram analysis, demonstrating superior performance compared to previous methods. [12] introduces a de-noising method combining wavelet packet decomposition with filtering, enhancing noise reduction in bird recordings and suggesting potential broader applications.

In the world of avian bioacoustics, the BirdCLEF challenges have consistently pushed the boundaries of bird call recognition technology with diverse methodologies like multiple data augmentation[13] and ensemble techniques[14] for robustness against environmental noise. [15] addressed challenges of memory management, species variety, and signal-to-noise ratio variance using pre-trained models and data augmentation, with Inception-v3 outperforming ResNet.

While existing studies focus on well-curated labeled data, this research aims to identify seven specific bird species (4 rare and 3 endangered) from unlabelled field recordings in Jahoo, Cambodia. The study introduces a novel deep learning approach, advancing towards self-

supervised learning. Pre-trained models are retrained with standard datasets to ensure comparability with custom models, thus setting new benchmarks in the field. Through extensive exploration of bird calls and songs using audio and spectrogram analysis, the research diverges from traditional methods that rely on segmented or controlled data. Analyzing raw data from complex environments, the study demonstrates the potential of deep learning models to handle the unpredictability of real-world data. By challenging existing methodologies, it establishes new standards for bioacoustic research, which is crucial for effective conservation strategies.

2. Multi-Modal Project Approach

In this project, bird species identification is divided into two separate tasks: Bird call and Bird song identification using two different approaches: audio features and spectrogram images for seven bird species namely, Germains peacock pheasant, Giant ibis, Great hornbill, Orange necked partridge, Oriental pied hornbill, White-shouldered ibis, White-rumped Shama. These 7 species are expected to be prominent in the recordings as per the field experts.

2.1. Dataset

Testing data for this project was provided by Cornell University in collaboration with Cambodian collaborators, where data was acoustically collected by deploying 10 SwiftOne units, programmed to continuously record with a 30 dB gain and a 32 kHz sample rate where data has been in continuous passive acoustic monitoring since 2022. The dataset comprises extensive field recordings of avian sounds, from which 53 hours of audio data was specifically chosen to align with bird breeding seasons, to ensure the diverse range of bird calls and songs because audio recordings were not categorized in calls and songs. The data was pre-processed to remove the noise and clean the data after acquiring the noise profile of raw data to understand ambient noise characteristics present in the audio files. The appropriate threshold factors were empirically determined for the power spectral density of audio from the range of noise threshold factors between '3.0' to '-12.5dB'. Following the noise reduction, audio files were segmented into 4-second clips, resulting in 38141 samples.

Training data for this project was collected from two sources "Xeno-canto" and "e-bird", for the aforementioned seven target species, including audios and sonograms of bird calls and songs. The collected audio files were examined using Audacity, and segmented into individual calls and songs, and labeled according to the species name. The audio data consists of 1187 calls and 253 songs, and spectrogram image data consists of 302 calls and 338 songs. Preprocessing for the audio task involved converting all samples to the sampling rate of 44KHz and using a single channel, for the image task spectrograms were resized to 224 $\times$ 224 pixels, and images were converted to RGB format. The train and validation data split was in the ratio of 80:20 to compare the model performances.

Two primary features were extracted:

MFCC: Parameters set included 13 MFCC features, with a Fast Fourier Transform window length of 2048 samples and a hop length of 512, resulting in each audio file being characterized by 13 MFCC features.

Chroma Features: Combined Chroma CQT and Chroma CENS features were used to capture pitch and harmony, with 12 features from each type per audio file and a hop length of 512. This led to each audio file with 24 Chroma features with feature fusion.

2.2. Modelling

Experiments are performed using custom and pre-trained CNNs (for both audio and spectrogram images) and RNN+LSTMs (for the audio task) after empirically optimising the architecture and hyper-parameters in each case.

The **custom CNN** architecture begins with a masking layer, followed by two convolutional layers (using 32 and 64 filters) for feature extraction, batch normalization, and max pooling for stability and dimensionality reduction, and dropout layers to prevent overfitting. It transitions to fully connected layers, ending with output layers for different data classes (6 for calls, 5 for songs), totaling 1,416,710 parameters.

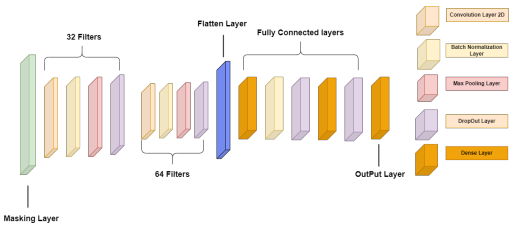

Figure 1: *CNN Architecture*

The **RNN+LSTM** architecture is designed for recognizing long-term dependencies in sequential data, with LSTM layers arranged to increase (64, 128, 256) and then decrease (128, 64) in units. It includes Dropout layers for preventing overfitting, followed by a flattened layer for transitioning to a dense network structure and ending with dense layers for feature refinement and classification (output of 5 and 6 classes for songs and calls respectively). This architecture boasts over 27 million trainable parameters, indicating its complexity and capability for multi-class classification tasks.

BirdNet analyzer was initially employed to label the test data, the preprocessed test data was fed to the analyzer and it resulted in the prediction of a species count of 5117, in which one of the target species *White Rumped Shama* was identified. Later, BirdNET Analyzer was utilized for

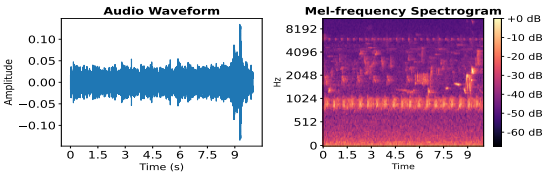

Figure 2: *Target Species: White Rumped Shama (visualizing waveform in amplitute and spectrogram in frequency aligns with human perception and facilitates clear understanding)*

both training and generating predictions on test data. Some modifications to the original BirdNET architecture like the 'Adam' optimizer were employed with specific optimised learning rates and accuracy metrics.

YAMNet and **VGGish** models involved preprocessing audio files according to model requirements and resampling to model standard rate. Audio for these two models were extracted using YAMNet and VGGish embeddings. Data augmentation and custom layers (neural network with

dense, dropout regularization and softmax) were applied to YAMNet, A Custom CNN classifier was designed and optimised to use the VGGish extracted embeddings.

ResNet-50, Inception V3, and **Xception** models were adapted by removing their top layers, with Global Average Pooling layer and customized with a dense layer of 1024 neurons using ReLU activation, Inception V3 was enhanced with two dense layers and a 50% dropout rate. Softmax layer finalized the architectures, ensuring the pre-trained weights remained frozen. Adam or RMSprop optimizers alongside categorical cross-entropy loss were selected for model compilation, optimizing performance.

3. Experiments

Models are evaluated using the unseen labelled validation set split out of the training data. Based on the validation performance, one or more models could be used to label the test data using a voting scheme.

3.1. Audio Task

Task	Features	CNN				RNN+LSTM			
		A	P	R	F1	A	P	R	F1
Calls	**MFCC**	**99**	**99**	**100**	**99**	**88**	**88**	**85**	**86**
Calls	Chroma	81	86	85	85	71	69	61	62
Songs	MFCC	94	77	79	76	78	62	66	63
Songs	Chroma	66	44	49	45	70	61	64	61

Table 1: *Deep Learning Audio Model Results, A=Accuracy, P=Precision, R=Recall, F1=F1 score*

Deep CNN and RNN+LSTM models were trained on MFCC and chroma features separately(table 1). It was revealed that models utilizing MFCC features markedly surpassed those with chroma features in audio classification. Specifically, the CNN model reached a 99% validation accuracy for call classification with high precision and recall, which slightly decreased for songs, in contrast to 94% for calls and 66% for songs with chroma features. Similarly, the RNN+LSTM model exhibited enhanced performance with MFCC, achieving 88% accuracy for calls and 78% for songs, with performance decreasing when trained on chroma. This trend highlights the importance of feature selection, with MFCC features significantly boosting model accuracy and validation accuracy. Germain's peacock pheasant was consistently recognised in both (calls & songs) task(figure 3), along with giant ibis and great hornbill in calls and songs task with very few mis-classifications and higher precision and recall rates, emphasizing role of MFCC's in audio classification.

In the evaluation of pre-trained models for audio classification, BirdNet, even with achieving 99% training accuracy for both calls and songs on the train set, could not generate even a single validation label correctly. YAMNet, using its embeddings, demonstrated robust performance with a 94% accuracy and high precision and recall rates for calls, though it experienced a significant performance drop in songs task(table 2). VGGish, while underperforming on its own, significantly improved when integrated with a CNN-2D model, reaching validation accuracies of 95% for calls and 85% for songs, with good precision and recall scores for both tasks, highlighting the combination of VGGish features and the CNN architecture. It was also noted that the great hornbill was recognised across all pre-trained models, with very few misclassifications.

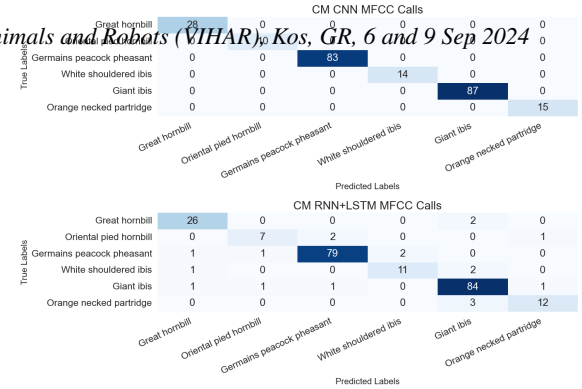

Figure 3: *Confusion matrix - Best performing deep learning audio models(CNN & RNN+LSTM in Calls task with MFCC) (Results represent audio task where training set has 6 species).*

Models	Calls				Songs			
	Acc	P	R	F1	Acc	P	R	F1
YamNet	94	89	88	88	80	79	68	69
CNN+Vggish	95	84	83	83	85	74	76	72

Table 2: *Pretrained Models in Audio Classification, Acc=Accuracy, P=Precision, R=Recall, F1=F1 score*

3.2. Spectrogram Image Task

Feature selection in image processing: CNNs learn visual features through multiple layers and filters. Inception V3 targets features at multiple scales with its varied layers in the inception module, while ResNet50 uses residual connections to aid in training deeper networks by overcoming the vanishing gradient issue. Xception improves upon Inception with depthwise separable convolutions for efficient, channel-wise feature learning. In spectrogram-based image classification tasks, 8 models were employed in total for songs and calls.

Model	Calls				Songs			
	Acc	P	R	F1	Acc	P	R	F1
CNN	91	40	41	41	89	88	33	24
InceptionV3	88	72	66	67	92	76	58	49
ResNet50	80	48	51	49	86	98	25	27
Xception	88	71	75	71	91	65	44	61

Table 3: *Spectrogram Image Model Results, Acc=Accuracy, P=Precision, R=Recall, F1=F1 score*

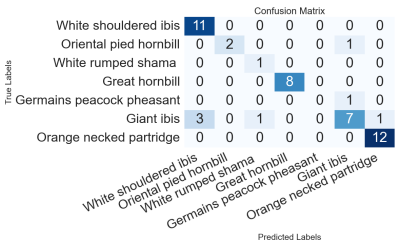

Figure 4: *Xception Spectrogram classification in calls (Results represent Spectrogram image task trained with 7 species)*

It is observed that CNN models maintained consistency in terms of training accuracy of 87% across both calls and songs tasks, with a slightly higher validation accuracy for calls 91% compared to songs 89% (table 3). This suggests better generalization for calls within the validation dataset. In terms of class identification, the model performed better on calls, correctly identifying 4 out of 7 classes, compared to 2 out of 6 for songs. Precision and recall rates were the lowest.

Model	Dataset	Task	ACC	Reference
Transound	Xeno-Canto	Audio	95	[9]
Inception	Xeno-Canto	Audio	94	[15]
Custom CNN	Xeno-Canto	Audio	99	This Study

Table 4: *Comparison of Previous studies with this study on Xeno-canto dataset using MFCC features*

In a comparison of state-of-the-art pretrained models, Xception and InceptionV3 demonstrated superior performance on the calls task, with 88% validation accuracy, both identifying 6 out of 7 classes but with lower precision and recall rates. Xception also excelled in the songs task, maintaining 91% validation accuracy, showcasing its strong feature extraction for image data. Conversely, ResNet50 was less effective, with 86% accuracy for songs and 80% for calls, identifying the fewest classes in both tasks. These results highlight the significance of selecting models based on task specificity, with Xception emerging as the top performer (figure 4) for both tasks due to its high accuracy and class identification capabilities.

4. Results & Evaluation

The comparative analysis with previous studies on Xeno-Canto dataset (Table 4) highlights the effectiveness of deep learning models, particularly custom CNN architecture, in leveraging MFCC features for accurate audio classification. It can be observed that the proposed model outperforms state-of-the-art models. This emphasizes the importance of model selection and customization in optimizing performance.

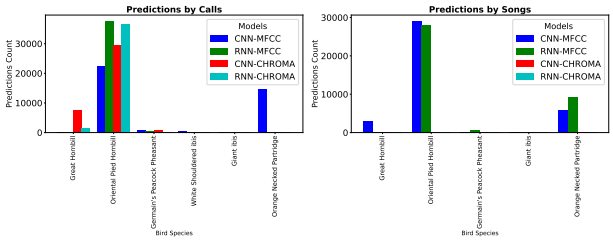

Figure 5: *Audio task deep models predictions calls & songs (Results represent audio task where training set has 6 species) (Low representation of species like giant ibis and white-rumped shama in training data significantly impact model predictions)*

Post-training, models were tested on unlabeled data to evaluate how many species can be identified from field recordings(Figure 5). Custom deep learning models outperformed pre-trained ones, especially in audio classification. However, biases were noted. For calls, CNN models with MFCC features identified species like the Oriental Pied Hornbill with high confidence, but less represented species were often missed, indicating overfitting. RNN+LSTM models detected a broader range of species but with lower confidence.

For song tasks, pre-trained models like YAMNet predicted a wider spectrum of species but with moderate confidence, needing further refinement. Spectrogram image tasks revealed additional challenges: custom CNN and pre-trained models like InceptionV3 & ResNet50 often identified the Great Hornbill in calls, and heavily favored the Giant Ibis in songs, showing a tendency to overfit. These issues highlight the need for balanced data representation and improved model tuning to ensure equal species recognition.

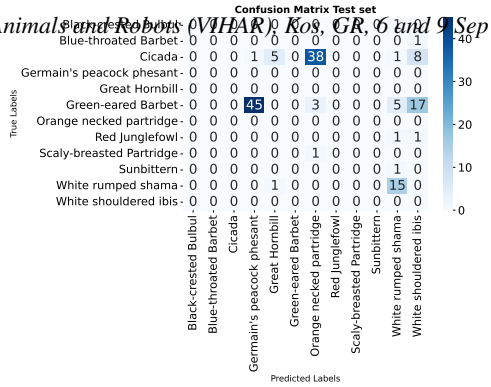

Figure 6: *Unseen test set(raw field data) Results*

Evaluation of Model Performance on Unlabeled Data The predicted labels must be validated once the species are identified in the field recordings. The evaluation of model performance on unlabeled data was conducted using ensemble voting approach, manual validation, and comparison with a reference model (BirdNet). This multi-step process ensured a comprehensive assessment of the model's accuracy and reliability. To determine species predictions, an ensemble voting method was employed to generate consensus labels for each recording by implemented models. A subset of 50 samples per species was then selected for manual validation. These samples were cross-verified with labels generated by the BirdNet model. The manual validation involved listening to audio recordings and analyzing spectrogram images to match frequencies and other acoustic features. This process revealed that most recordings initially classified as bird sounds were infact insect sounds (primarily cicadas) with significant environmental noise. For species like the White-rumped Shama, BirdNet labels were more accurate compared to the ensemble model's predictions. A secondary comparison of these labels identified six instances of misclassification,(figure 6) highlighting areas for further model improvement. The BirdNet model proved useful for certain avian species, but overall, the findings highlight the importance of a balanced and well-represented dataset and in-domain data for reliable species classification.

5. Conclusion

This study demonstrates that MFCC features and custom CNNs outperform pre-trained models in avian audio classification, especially in identifying bird calls with high precision using labeled data. However, identifying bird species in field recordings using zero-shot learning was impossible due to a lack of domain adaptability, despite using a voting scheme with multiple models. The challenges of species bias and overfitting highlight the need for better model refinement and balanced training datasets. Pre-trained models like YAMNet and Xception performed well in the validation set but failed to generalize to the test set. Manual validation with BirdNet showed high reliability and precision in avian species identification. This approach was the first step towards self-supervised learning, using models trained on other labeled datasets to identify calls and songs from seven specific species in raw, unlabeled field recordings. Future work could use clustering methods to group similar bird sounds and label test data, potentially enhancing accuracy by focusing on nuanced differences in bird sounds using few-shot learning techniques.

6. References

[1] A. C. Rodríguez, S. D'Aronco, R. C. Daudt, J. D. Wegner, and K. Schindler, "Recognition of unseen bird species by learning from field guides," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1742–1751.

[2] J. Xie, Y. Zhong, J. Zhang, S. Liu, C. Ding, and A. Triantafyllopoulos, "A review of automatic recognition technology for bird vocalizations in the deep learning era," *Ecological Informatics*, vol. 73, p. 101927, 2023.

[3] S. Hu, Y. Chu, Z. Wen, G. Zhou, Y. Sun, and A. Chen, "Deep learning bird song recognition based on mff-scsenet," *Ecological Indicators*, vol. 154, p. 110844, 2023.

[4] J. Liu, Y. Zhang, D. Lv, J. Lu, S. Xie, J. Zi, Y. Yin, and H. Xu, "Birdsong classification based on ensemble multi-scale convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 8636, 2022.

[5] A. Kershenbaum, D. T. Blumstein, M. A. Roch, Ç. Akçay, G. Backus, M. A. Bee, K. Bohn, Y. Cao, G. Carter, C. Cäsar *et al.*, "Acoustic sequences in non-human animals: a tutorial review and prospectus," *Biological Reviews*, vol. 91, no. 1, pp. 13–52, 2016.

[6] S. D. Hill, W. Ji, K. A. Parker, C. Amiot, and S. J. Wells, "A comparison of vocalisations between mainland tui (prosthemadera novaeseelandiae novaeseelandiae) and chatham island tui (p. n. chathamensis)," *New Zealand Journal of Ecology*, pp. 214–223, 2013.

[7] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, and A. Joly, "Lifeclef bird identification task 2016: The arrival of deep learning," in *CLEF: Conference and Labs of the Evaluation Forum*, no. 1609, 2016, pp. 440–449.

[8] P. Gavali and J. S. Banu, "Bird species identification using deep learning on gpu platform," in *2020 International conference on emerging trends in information technology and engineering (ic-ETITE)*. IEEE, 2020, pp. 1–6.

[9] Q. Tang, L. Xu, B. Zheng, and C. He, "Transound: Hyper-head attention transformer for birds sound recognition," *Ecological Informatics*, vol. 75, p. 102001, 2023.

[10] S. Zhang, Y. Gao, J. Cai, H. Yang, Q. Zhao, and F. Pan, "A novel bird sound recognition method based on multifeature fusion and a transformer encoder," *Sensors*, vol. 23, no. 19, p. 8099, 2023.

[11] X. Zhang, A. Chen, G. Zhou *et al.*, "Spectrogram-frame linear network and continuous frame sequence for bird sound classification. eco inform 54: 101009," 2019.

[12] N. Priyadarshani, S. Marsland, I. Castro, and A. Punchihewa, "Birdsong denoising using wavelets," *PloS one*, vol. 11, no. 1, p. e0146790, 2016.

[13] M. V. Conde and U.-J. Choi, "Few-shot long-tailed bird audio recognition," *arXiv preprint arXiv:2206.11260*, 2022.

[14] M. V. Conde, K. Shubham, P. Agnihotri, N. D. Movva, and S. Bessenyei, "Weakly-supervised classification and detection of bird sounds in the wild. a birdclef 2021 solution," *arXiv preprint arXiv:2107.04878*, 2021.

[15] C.-Y. Koh, J.-Y. Chang, C.-L. Tai, D.-Y. Huang, H.-H. Hsieh, and Y.-W. Liu, "Bird sound classification using convolutional neural networks." in *Clef (working notes)*, 2019.

Data Ethics and Practices of Human-Nonhuman Sound Technologies and Ecologies

Petra Jääskeläinen[1]*, Elin Kanhov*[1]

KTH Royal Institute of Technology, Sweden

ppja@kth.se, ekanhov@kth.se

Abstract

Human-nonhuman sound interaction and technologies aim to bridge the gap of inter-species communication. While they emerge from attempts to understand and communicate with nonhumans, they also raise questions on the ethics of nonhuman data use, for example regarding the unintended consequences such data extraction can have to nonhumans. In this paper, we discuss power relations and aspects of representation in nonhuman data practices, and their potential critical implications to nonhumans. Drawing from prior research on data ethics and posthumanities, we conceptualize two challenges of nonhuman data ethics for the design of Human-Nonhuman Interaction (HNI) and technologies in sound ecologies. We provide takeaways for how sensitivities toward nonhuman stakeholders can be considered in the design of HNI in the context of sound ecologies.

Index Terms: nonhuman data ethics, data ethics, human-nonhuman interaction, human-animal interaction, data extractivism, technological mediation

1. Introduction

While research on human-nonhuman interaction is developing in many domains, including animal communication [1] and human-computer interaction [2, 3], so far little focus has been placed on the ethical aspects of nonhuman data use and data practices in the context of such interactions in sound ecologies. These ethical aspects has been raised previously in nonhuman philosophy and ethics, for example in terms of power structures between humans and nonhumans [4], nonhuman representation [5], and labour [6]. All of these concerns can be directly projected to examine nonhuman data ethics and practices.

By *human-nonhuman interaction (HNI)*, we refer in this paper broadly not only to inter-species communication and design of technologies for such purposes, but also to the actuality of humans living, both passively and actively, in constant interaction and relationality – or entanglement – with nonhumans [7]. In this paper, we explicitly focus on living nonhuman entities (e.g. animals, plants, ecosystems) rather than, for example, technological companions [8]. Humans interact with nonhumans simply by entering their habitat and observing their ways of life, without attempts of communicating. In this way, we take a relational [9] environmental posthumanist perspective [10, 7, 11, 12] on the kinship between humans and nonhumans [8], and focus on the role of sound in such relational ecologies [13, 14, 15]. Within this context of HNI, our specific focus is therefore to examine *the ethics of data practices* in nonhuman sound ecologies. These questions arise, for example, when we enter environments in which nonhumans reside; introduce technology into them; design technologies for inter-species interaction; and generally when we collect and use nonhuman data. There is a distinction to be made between ethics of data use and ethics of entering nonhuman environments for data collection, and we discuss both of these in this paper under the term data practices. Not all processes require both of these, and it is likely that they raise different ethical issues in practice.

In this work, we draw from data ethics, posthumanities, and sound ecologies literature to inform the use of data in the context of human-nonhuman interactions, asking the question: *how do sonic entanglements relate to questions of power dynamics between humans and nonhumans, and how may technological mediation affect such dynamics?* By drawing from existing literature [16], we outline two ethical challenges of nonhuman data use and practices: 1. Examining and Challenging Human-Nonhuman Power Structures, and 2. Examining the Nonhuman Data Representation and Labour. Thus, this paper contributes with providing critical perspectives on data ethics and power relations of human-nonhuman interactions in sound ecologies. We discuss potential benefits and concerns in how research in this domain can configure power relations between humans and nonhumans through data practices [17, 18, 16], and how technology plays an active role in configuring these relations through the mediation of sound between humans and nonhumans. Lastly, we urge for further critical reflection on nonhuman data ethics in HNI and sound ecologies.

We first begin by situating HNI into relational sound ecologies. Subsequently, we place practices of data extraction into the wider context of knowledge production through sound, to then conceptualize what nonhuman data ethics can implicate.

2. Background

2.1. Relational Sound Ecologies

Sound is part of relational [9, 10, 19] ecologies that involve both humans and nonhumans. In the era of the Anthropocene, these entangled and relational more-than-human ecologies are often discussed in terms of how they contribute to sustainability, such as biodiversity and maintaining healthy ecosystems [9]. We consider "sound ecologies" to be any more-than-human system in which sound plays a role in the relationality between entities. A more-than-human onto-epistemology in posthumanist research advocates relational thinking [9] and a decentralization of humans [10, 19], for example in relation to other living entities. These perspectives have been informed by both de-colonial research on indigenous environmental relations (e.g. place-based onto-epistemologies) [20], and feminist science and technology studies (STS) [10, 7].

[1] Both authors have contributed equally to this paper.

However, in the Western modernist scientific paradigm [21, 9, 22, e.g.] there is a strong tendency to study "measurable" and "modellable" aspects, often with insufficient sensitivity to nonhuman subjectivities [23]. When we think of these human-nonhuman relations – specifically in the context of sound-technology-mediation – we need to consider questions of how human interactions with nonhumans, and technologies that mediate these interactions, shape the nonhumans' reality, rather than approaching them from an anthropocentric perspective. In an attempt to de-centralize these anthropocentric perspectives, we can begin to "de-colonize" and reconfigure our relation to nonhumans – an effort that has become increasingly explored in technology interactions in recent years in the form of more-than-human technology design [24, 25, 18, e.g.].

Humans are deeply entangled with other species in sound ecologies, and this involves a constant configuration of power relations between various human and nonhuman entities. This becomes particularly evident in studies of noise pollution [26, 27], where human ways of life not only affects the physical environment of living nonhuman entities, but also silences their sonic expressions, capabilities and realities. However, it is not only through such destructive practices that humans are engaged in sound ecologies. Turning to indigenous cultures, it is clear that humans have long been sonically entangled with nonhumans. For example, the Kaluli people of Papua New Guinea have a deep sonic and musical connection with their environment [28]. Such ways of knowing have, not least in the Western world, been undermined by rationalization in the modernist scientific paradigm.

Furthermore, due to the differences in our make compared to living nonhuman entities, in certain aspects we are also very concretely detangled from each other's sonic realities. For example, humans are incapable of hearing infrasounds of breaking icebergs, whales and elephants, and ultrasounds of bats, mice and corals, as frequencies of such sounds lie outside of the human range of hearing [29, 30]. By using technological tools and mediation, however, humans can become able to hear these sonic realities of other living entities.

2.2. Knowledge Production and Data in Sound Ecologies

Knowing the world through sound offers information and sensory input that widely differ from visual inquires, which are often dominating the ways of knowing for humans. Thus, sonic imagination in itself can help us think beyond the visually dominated human-centred world [14, 31, 32, 33]. These visual ways of knowing are central also to other primates, which thus are naturally advantaged from sharing the same senses as humans in this human-centred world. As such, sonic perspectives can be understood as part of an embodied, embedded and situated knowledge practice [22, 11], where an "acoustemology" [34], or acoustic epistemology, affords sensitivities towards nonhuman subjectivities beyond normative (Western rational) ways of knowing. This notion of embodied knowledge has been generally acknowledged in the design of technology in past decades [35, 36, 37], changing the way how technology design is approached.

To access the world of sound beyond using our ears, which, as we have already established, are limited in terms of range and sensitivity to certain levels of sound, we can turn to technology to "enhance" and "decode" sound ecologies. In fact, the digital revolution has offered new tools and methods for accessing nonhuman sound ecologies that has provided understandings for how complex such ecologies are [29]. This affords not only new incentives for environmental conservation but also possibilities for inter-species communication. For such practices to be possible, however, the data that is recorded, or *extracted* from ecological sites, must be manipulated so as to be intelligible to humans.

A critical question therefore arises regarding what such processes of technologically enhanced entanglements induce, if we examine the power relations and focus on the subjectivities of nonhumans. While technological developments and capabilities enable further exploration of the sonic world and provide insight and deeper understanding of nonhuman realities, they also have the potential to disturb and change the natural habitats and behaviours of the nonhumans studied [26]. As such, there is a danger that technology becomes a tool for extractivist practices toward nonhumans, serving the anthropocentric worldview and enforcing the contemporary power configurations that place humans as the locus. It is essential, then, that ethical reflection is directed toward the potential critical impact on nonhumans when such technologies are designed and introduced in these more-than-human configurations.

Furthermore, it is important to note that different research fields have varying motivations and intentions for their sonic data collection. This can be due to cultural, geopolitical, and institutional differences, and their ethical guidelines and practices often vary. For example, while animal behavioral research has the intent to understand nonhumans, technology engineering research has a primary interest in advancing technological development, and artistic practice might work with nonhuman data in creative dialogue with society. In summary, the human-nonhuman sound interaction is a very diverse field of practices, and the data ethics practices of each specific case should be examined carefully.

3. Conceptualizing Nonhuman Data Ethics

Exploring these critical questions and impacts on nonhumans further, we turn to feminist data ethics literature [16, e.g.] as a perspective to understand how power relations are constructed through data and data practices. Bringing this together with other literature that examines power relations between humans and nonhumans (such as speciesism [4] and human-animal media studies [5]), we argue that data practices involving nonhumans are actively configuring inter-species power relations. In this section, we draw on this research to conceptualize important dimensions that need to be examined in terms of ethics of nonhuman data and sound technology practices.

The principles of *data feminism* are intended to re-think and reconfigure power relations in the context of human data practices. In regards to more-than-human sound ecologies, we can apply the same principles to examine power relations of nonhuman data use and practices – a connection that feminist environmental posthumanities research has more widely built on to examine questions that relate to human-nonhuman relations [12]. There are seven feminist principles for working with data, which we will examine in the context of nonhuman data in sound ecologies. These are; 1. Examine power, 2. Challenge power, 3. Elevate emotion and embodiment, 4. Rethink binaries and hierarchies, 5. Embrace pluralism, 6. Consider context, 7. Make labour visible [16]. Examining power concerns the need to critically investigate the power configurations that relate to data and data practices, and challenging power means taking concrete steps of re-configuring the identified power imbalances. Elevating embodiment highlights the earlier discussed need to expand the knowledge-making to its embodied situat-

edness. Rethinking binaries and hierarchies can help change the way information is conceptualized, leading into embracing pluralism which encourages diverse ways of knowing, communicating and being. Consideration of context refers to acknowledging the situated context of each case, and lastly, making labour visible concerns tracing and exposing all the labour that takes place in data practices. We now project these principles onto the case of HNI in sound technologies and ecologies.

3.1. Examining and Challenging Human-Nonhuman Power Structures

Feminist data ethics advocate for firstly examining prevailing power structures, to then actively challenge them. Transferring this principle onto the design of HNI sound technologies, researchers should consider how power is configured between various human and nonhuman stakeholders with these technologies, and how sound technologies can be re-imagined in ways that the nonhuman stakeholders gain more power and agency. These aspects urge the designers and developers of the technologies to think about on whose terms the technology is designed and who is benefiting from it in the long term. Relevant questions to ask in this context are: *how is power configured between various human and nonhuman stakeholders in the technological configurations*, and *how can these technologies be radically re-imagined in a way that the nonhuman stakeholders gain more power*? These questions probe designers and developers of the technologies to think about on whose terms the technology is designed and who is benefiting from it in the long term. In a practical sense, approaches such as mapping the critical and positive stakeholder (nonhuman) concerns can be incorporated in processes of reflecting on such questions. These types of methods have recently started to emerge in HCI research [38, e.g.]. Thus, there surfaces a need to explore more methods that can be used in technology and data practices for developing sensitivities to nonhuman stakeholders.

Considering on whose terms the technology is designed, it is important to study data practices on a larger scale. This concerns, for example, what kind of practices and types of data are dominating the landscape in HNI. One of the central aspects that characterizes human data practices and, more widely, practices of designing technology, is the aspire to *decode*, *systematize*, and *model* [21, 22]. Designers and developers should consider how these processes of technological mediation affect the type of information that is mediated, and what is gained or lost when we try to organize nonhuman sounds in "human ways". Prior studies have explored data surveillance and data extraction in the context of various nonhumans, for example discussing how modeling and rationalizing can lead to harmful outcomes for the nonhumans [39]. Also, studies demonstrate how data practices configure new environments and nonhuman-environmental relations, and how such practices give voice to various "monitored" nonhumans (e.g. animals, plants) [40, 41]. As humans attempt to monitor, record, decode, analyse and even communicate with nonhumans, we need to ask on whose terms these (inter-)actions are practiced.

It is also crucial to examine processes of intervention, and how human and technological presence in nonhuman habitats may affect the nonhuman ecologies, related to the third principle of elevating emotion and embodiment. As discussed, sound and particularly vocalization plays a role in the power dynamic between humans and nonhumans. For instance, cats vocalize in a particular way when engaging with humans, and animals that are taken out of their natural habitats can start vocalizing more intensely as an sign of dependence on human caretakers. By practicing empathy toward the nonhuman and fully engaging in sensitive and embodied listening, designers and developers can "make kin", e.g. reflect and reconfigure our relation to nonhumans [8]. Furthermore, we need to fully understand the long-term implications of placing technological artefacts (mics, sensors, transmitters, etc.) in nonhuman sound ecologies, and how the nonhumans change and adapt to these. In the posthumanist literature, it has been explored how the human has coevolved with technology through the concept of the cyborg [42]. Like humans, nonhuman entities are not immune to technological influence, and it can be argued that they are also in a cyborg relationship with their (technological) environments [40]. Yet, they have less power in giving consent to being so. From a sound ecology perspective, data collection practices can also involve introducing sounds to wild environments, which calls for ethical reflection on the impact of our data practices on sound ecologies. For instance, researchers may purposefully introduce sounds to lure birds or other species into communication, or simply produce sounds by talking, walking, and using vehicles.

Following de-colonial science and technology practices [43, 44], researchers can further ask whether we always have the right to enter a nonhuman habitat for the sake of scientific and technological advancement. This question urges us to examine our human privileges, and our role as "nonhuman colonizers" that use technology as a tool for colonization. While these issues have surfaced often in de-colonial data studies [45], they have not been examined in depth when it comes to HNI. Thus, we urge these questions to become an integral part of data ethics in HNI sound ecologies.

3.2. Examining Nonhuman Data Representation and Labour

Turning to the principles of rethinking binaries and hierarchies, as well as embracing pluralism, another relevant dimension of data practices relates to representation, which is commonly discussed in human data ethics [16]. This concerns what and who is represented in data collection and analysis, which in human terms is discussed in aspects of gender and race, for example. Transferring this notion to the context of nonhuman data practices, we can ask: *which species are studied and which are not*, and *what data is dominating in the data practices*? This also raises questions on what the critical implications are for various species when they are represented in differing ways. For example, a lack or excess of representation of certain species may affect their everyday life and experiences, as some species might be considered more "worth" studying than others (e.g. [30]). Furthermore, we can also ask what implications there are if the species are represented and discussed in a certain normative way. As an example, when animals are represented in human culture (media), people might be more likely to approach and interact with certain familiar species in the wild or sympathize more with such species which can have direct consequences for their livelihoods and environments. Similarly, nonhumans that are deemed "hostile" can be treated in very different and noncaring ways by humans – or even completely disregarded and excluded from conservation.

Diversifying representations not only applies to the data itself, but also to multiple ways of knowing and making knowledge (as discussed in Section 2.2). This can be done by challenging the predominant ways of doing research in HNI and seeking to diversify such practices. These remarks urge the designers and developers of HNI technologies to reflect carefully

on data practices, collection, and use in terms of how the data is manipulated; what forms it takes; what ways of knowing it promotes; and ultimately, what ways of knowing are prioritized and dominating the data practices. Furthermore, such diversifying can be cultivated by attuning to ways of being and knowing that are currently overlooked or underrepresented. These aspects urge us to fine-tune into and examine more carefully the contexts in which the data exists, is produced, and understood.

Related to considerations of context and making labour visible, we wish to emphasize the need to acknowledge nonhuman labour in collection of data and design of the technology. Most often in HNI, nonhumans are contributing their data without having a choice to do so. It is therefore also relevant to consider whether they should be compensated for that data extraction, and whether there are ways of asking nonhumans for consent of use. In animal ethics [4] and environmental ethics [46] it has been argued that ethical consideration should be attributed to nonhumans following their unique needs. For example, species with similar needs call for similar consideration and care, as a principle of equal treatment. When this is applied to labour and data ethics, we can consider different nonhuman species to do differing types of labour – actions or behaviors – in producing data and interacting with humans and technology. Furthermore, we can anticipate a need for them to be compensated differently from this labour, following each species' unique needs and interests. This raises challenging questions about how such compensation should take place. For example, if we compensate zebra finches species members with plant seeds, it can be seen as their species-specific interest. At the same time, we might contribute to domestication of the species and further inter-species colonization. Furthermore, we might overlook the individual preferences and variability of specific species members [47].

Lastly, the labour that both humans and nonhumans engage in is actively shaping the earlier discussed representations of nonhumans by rendering some species more visible than others. Reflecting on how such nonhuman data labour can be practiced on ethical terms is therefore of critical importance – in a similar way to how the handling of human data is becoming an increasing concern in all parts of digital society [45, 18, 16].

4. Conclusion

In this paper, we have discussed critical questions in regard to the data ethics of human-nonhuman in sound technologies and ecologies. Drawing from feminist and de-colonial data ethics, posthumanities, and sound ecologies literature, we have conceptualized sound ecologies as relational sites in which knowledge production and data practices coincide. We have provided two concrete areas to examine when it comes to nonhuman data ethics (1. Examining and Challenging Human-Nonhuman Power Structures, and 2. Examining Nonhuman Data Representation and Labour). We discussed related challenges through concrete examples, and reflected on what unintended consequences such data practices can have to nonhumans. We aim for this paper to spark discussion on data and sound technology practices in the communities that design human-nonhuman interactions, and urge for the VIHAR community to examine these data ethics questions in further depth in the future.

5. Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP- HS) funded by the Marianne and Marcus Wallenberg Foundation, and a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864189).

6. References

[1] M. D. Beecher, "Why are no animal communication systems simple languages?" *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.602635

[2] C. Mancini, "Animal-computer interaction: a manifesto," *Interactions*, vol. 18, no. 4, p. 69–73, jul 2011. [Online]. Available: https://doi.org/10.1145/1978822.1978836

[3] I. Hirskyj-Douglas and S. Webber, "Reflecting on methods in animal computer interaction: Novelty effect and habituation," in *Proceedings of the Eight International Conference on Animal-Computer Interaction*, ser. ACI '21. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3493842.3493893

[4] P. Singer, *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York Review, 1975.

[5] E. M. Cesaresco, *The place of animals in human thought*. Literary Licensing, Mar. 2014.

[6] E. Barron and J. Hess, "Non-human labour: the work of Earth Others," in *The Handbook of Diverse Economies*, ser. Chapters, J. K. Gibson-Graham and K. Dombroski, Eds. Edward Elgar Publishing, 2020, ch. 17, pp. 163–169. [Online]. Available: https://ideas.repec.org/h/elg/eechap/18372_17.html

[7] K. Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press, 2007.

[8] D. J. Haraway, *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, 08 2016. [Online]. Available: https://doi.org/10.1215/9780822373780

[9] S. West, L. Jamila Haider, S. Stålhammar, and S. Woroniecki, "Putting relational thinking to work in sustainability science – reply to raymond et al." *Ecosystems and People*, vol. 17, no. 1, pp. 108–113, 2021. [Online]. Available: https://doi.org/10.1080/26395916.2021.1898477

[10] D. Haraway, "A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s," *Australian Feminist Studies*, vol. 2, pp. 1–42, 3 1987.

[11] R. Braidotti, *Posthuman Knowledge*. Medford, MA: Polity Press, 2019.

[12] C. Åsberg, "Ecologies and technologies of feminist posthumanities," *Women's Studies*, vol. 50, no. 8, pp. 857–862, 2021.

[13] B. Truax, *The World Soundscape Project's Handbook for Acoustic Ecology*. A.R.C. Publications, 1978, google-Books-ID: jKNlswEACAAJ.

[14] R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. Simon and Schuster, Oct. 1993.

[15] M. Droumeva and R. Jordan, Eds., *Sound, Media, Ecology*. Cham, SWITZERLAND: Palgrave Macmillan, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/kth/detail.action?docID=5803020

[16] C. D'Ignazio and L. F. Klein, *Data feminism*. London, England: MIT Press, Oct. 2023.

[17] S. Mezzadra and B. Neilson, "On the multiple frontiers of extraction: excavating contemporary capitalism," *Cultural Studies*, vol. 31, no. 2-3, pp. 185–204, 2017.

[18] K. Crawford, *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021, OCLC: on1111967630.

[19] B. Latour, *Politics of nature : how to bring the sciences into democracy*. Cambridge, Mass: Harvard University Press, 2004.

[20] R. W. Kimmerer, *Braiding sweetgrass*. Minneapolis, MN: Milkweed Editions, Aug. 2015.

[21] B. Latour, *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, 1987, google-Books-ID: sC4bk4DZXTQC.

[22] D. Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies*, vol. 14, no. 3, pp. 575–599, 1988, publisher: Feminist Studies, Inc. [Online]. Available: https://www.jstor.org/stable/3178066

[23] D. Debaise, "What is a Non-Human Subjectivity?" *Archives de Philosophie*, vol. 75, no. 4, pp. 587–596, Nov. 2012, bibliographie_available: 0 Cairndomain: www.cairn-int.info Cite Par_available: 0 Publisher: Facultés Loyola Paris. [Online]. Available: https://www.cairn-int.info/journal-archives-de-philosophie-2012-4-page-587.htm

[24] R. Wakkary, *Things We Could Design: For More Than Human-Centered Worlds*. MIT Press, Aug. 2021.

[25] I. Nicenboim, E. Giaccardi, M. L. J. Søndergaard, A. V. Reddy, Y. Strengers, J. Pierce, and J. Redström, "More-than-human design and AI," in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*. New York, NY, USA: ACM, Jul. 2020.

[26] H. Slabbekoorn, R. J. Dooling, A. N. Popper, and R. R. Fay, Eds., *Effects of Anthropogenic Noise on Animals*, ser. Springer Handbook of Auditory Research. New York, NY: Springer, 2018, vol. 66. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-8574-6

[27] R. Sordello, O. Ratel, F. Flamerie De Lachapelle, C. Leger, A. Dambry, and S. Vanpeene, "Evidence of the impact of noise pollution on biodiversity: a systematic map," *Environmental Evidence*, vol. 9, no. 1, p. 20, Sep. 2020. [Online]. Available: https://doi.org/10.1186/s13750-020-00202-y

[28] S. Feld, *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression*, 3rd ed. Durham [N.C.]: Duke University press, 2012.

[29] K. Bakker, *The Sounds of Life: How Digital Technology Is Bringing Us Closer to the Worlds of Animals and Plants*. Princeton University Press, 2022. [Online]. Available: http://www.jstor.org/stable/j.ctv2hnkcc3

[30] T. Nagel, *What Is It Like to Be a Bat?* Oxford University Press, Incorporated, 2024. [Online]. Available: https://books.google.se/books?id=1hX5EAAAQBAJ

[31] K. Wrightson, "An introduction to acoustic ecology," *Soundscape: The journal of acoustic ecology*, vol. 1, no. 1, pp. 10–13, 2000.

[32] V. Tkaczyk and L. van der Miesen, "Sonic Things: Knowledge Formation in Flux," *Sound Studies*, vol. 6, no. 2, pp. 105–113, Jul. 2020, publisher: Routledge _eprint: https://doi.org/10.1080/20551940.2020.1794651. [Online]. Available: https://doi.org/10.1080/20551940.2020.1794651

[33] H. Twidle and A. Eloff, "Sounding Environments," in *The Routledge Handbook of Environmental History*, E. O'Gorman, W. S. Martín, M. Carey, and S. Swart, Eds. London: Routledge, 2023.

[34] S. Feld, "Acoustemology," in *Keywords in Sound*, D. Novak and M. Sakakeeny, Eds. Duke University Press, Apr. 2015, pp. 12–21. [Online]. Available: https://read.dukeupress.edu/books/book/166/chapter/106258/

[35] *Where the action is: the foundations of embodied interaction*. Cambridge, MA, USA: MIT Press, 2001.

[36] K. Höök, "Soma design - intertwining aesthetics, ethics and movement," in *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, ser. TEI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1. [Online]. Available: https://doi.org/10.1145/3374920.3374964

[37] K. Höök, S. Benford, P. Tennent, V. Tsaknaki, M. Alfaras, J. M. Avila, C. Li, J. Marshall, C. D. Roquet, P. Sanches, A. Ståhl, M. Umair, C. Windlin, and F. Zhou, "Unpacking non-dualistic design: The soma design case," *ACM Trans. Comput.-Hum. Interact.*, vol. 28, no. 6, nov 2021. [Online]. Available: https://doi.org/10.1145/3462448

[38] C. Núñez-Pacheco and A. Poikolainen Rosén, "Articulating felt senses for more-than-human design -a viewpoint for noticing," 06 2024.

[39] M. Tironi and D. I. Rivera Lisboa, "Artificial intelligence in the new forms of environmental governance in the chilean state: Towards an eco-algorithmic governance," *Technology in Society*, vol. 74, p. 102264, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X23000696

[40] J. Gabrys, *Program earth: Environmental sensing technology and the making of a computational planet*. U of Minnesota Press, 2016, vol. 49.

[41] S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions," *BioScience*, vol. 68, no. 8, pp. 563–576, 07 2018. [Online]. Available: https://doi.org/10.1093/biosci/biy068

[42] D. J. Haraway, *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York : Routledge, 1991. [Online]. Available: http://archive.org/details/simianscyborgswo0000hara

[43] S. Costanza-Chock, *Design justice*. The MIT Press, 2020.

[44] A. Alvarado Garcia, J. F. Maestre, M. Barcham, M. Iriarte, M. Wong-Villacres, O. A. Lemus, P. Dudani, P. Reynolds-Cuéllar, R. Wang, and T. Cerratto Pargman, "Decolonial pathways: Our manifesto for a decolonizing agenda in hci research and design," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411763.3450365

[45] U. A. Mejias and N. Couldry, *Data Grab: The New Colonialism of Big Tech and How to Fight Back*. Chicago, IL: University of Chicago Press, 2024. [Online]. Available: https://press.uchicago.edu/ucp/books/book/chicago/D/bo216184200.html

[46] K. E. Goodpaster, "On being morally considerable," *The Journal of Philosophy*, vol. 75, no. 6, pp. 308–325, 1978. [Online]. Available: http://www.jstor.org/stable/2025709

[47] E. Aaltola, *Animal Individuality: Cultural and Moral Categorisations*, ser. Reports from the Department of Philosophy. University of Turku, 2006. [Online]. Available: https://books.google.se/books?id=_Z3WAAAAMAAJ

Introducing LeVI-imit – self-supervision based articulatory model imitating speech on a web browser

Heikki Rasilo[1], *Yannick Jadoul*[2]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium
[2]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
Heikki.rasilo@vub.be, Yannick.Jadoul@uniroma1.it

Abstract

We introduce a web browser-based articulatory synthesizer and a novel proof-of-concept feature that inverts speech recorded by the user into articulatory movements, and imitates the input speech. Inversion is based on a neural network that is pretrained using a self-supervised actor-critic reinforcement learning approach. In the training phase, the actor-network learns articulatory gestures, that when performed by the given vocal tract model, leads to acoustic output that is as close as possible to the speech features given as its input. After training, the actor-network is converted to a tensorflow-javascript model that is run on a web browser. The inverted speech is plotted as continuous articulatory movements as well as played back as acoustic output. To our knowledge this is the first easily accessible web browser-based speech inversion system, that may work as a demonstrator for speech phenomena, as well as a quick tool for subjective evaluation of speech inversion performance.

Index Terms: speech recognition, human-computer interaction, speech inversion, demonstration, multimodality

1. Introduction

Speech inversion, or acoustic-to-articulatory inversion, refers to mapping of acoustic speech back to the physical articulations from which it originates. Human language learners learn to invert speech, enabling speech imitation, i.e. using one's own vocal tract (VT) to articulate heard speech of others. Techniques to perform speech inversion in engineering solutions or research have been studied for decades, and recent advances in machine learning has brought the community closer and closer to accurate and efficient solutions for the problem.

Evaluating and experimenting with speech inversion systems is often out of reach for a casual user. In research, its evaluation is often based on reported objective measures of inverted articulations' closeness to measured articulations [1,2] or similarity of input and imitated audio features [3,4]. The quality of the inverted speech has also been evaluated in automatic speech recognition tests or by human listeners [3]. Inversion results of example utterances are also often presented in pre-recorded video clips [4], or audio files [6,7], leaving an interested person little possibility to freely experiment with the inversion tools. Such experimentation would require installation of software packages (e.g. VocalTractLab (VTL) [7]), knowledge of programming, proper set up of VT models and their parameters, and access to the trained models.

A good real-time web browser-based demonstration of the forward VT model exists in Pink Trombone (PT)[1], but it is for now not capable of imitating user input. Recent research reports speech inversion into the control parameters of PT [6], but the inversion is not present in the browser implementation. In this paper we introduce LeVI-imit[2], a novel fast speech inversion system included in a web browser-based articulatory model that can be easily used to invert and imitate any recorded audio input. To our knowledge this is at the moment the only available easily accessible acoustic-to-articulatory inversion visualization tool. Even though the inversion performance of our first implementation is far from perfect, having such a tool is ideal to demonstrate and visualize speech articulation and related phenomena. Moreover, due to its interactive nature, it provides a fast means to test and evaluate the underlying speech inversion method, investigate potential issues, as well as to subjectively compare it to alternative methods.

1.1. Speech inversion and previous research

Over the last decades, various methods have been used to invert speech to the articulatory parameters of several different articulatory models. [8] used the articulatory synthesizer of The Haskins Laboratories [9], based on the Mermelstein articulatory model [10]. DIVA [12,13], a model for vocal control learning, as well as the vocal learner of Howard and Messum [13] use versions of the articulatory synthesizer by Maeda [14]. Drayton [15] inverted vowel sounds to parameters of the articulatory synthesizer of Praat [16]. Lately, Shibata, Zhang and Shinozaki [2] have used the VTL-model in their unsupervised imitation learning study. The scopes, assumptions and performances of the mentioned systems vary, but no system offers a user-friendly interactive inversion tool.

There are several aspects that contribute to the difficulty of solving the inversion task in machines. To mention but a few of the associated difficulties, this task involves finding a suitable VT model that is close enough to the human articulatory system, dealing with the many-to-one property of speech production (i.e. different articulations may lead to similar acoustic outputs - think of ventriloquists for example), addressing with variation in people's VT morphologies (i.e. a child imitating adult speech will end up using a smaller VT, whose acoustic outcomes are also different), dealing with the high dimensionality of the articulatory space (and thus vast search spaces for optimal solutions), tailoring acoustic features

[1] https://dood.al/pinktrombone

[2] https://ai.vub.ac.be/levi-demo/LeVI.html

suitable for the inversion task, and evaluating the inversion results.

Our focus in the field is to find methodology that allows articulatory learning from input that is similar to what a human learner would have access to. Supervised learning methods that learn to map acoustic speech to measured articulatory data (obtained for example by using Electromagnetic midsagittal articulography) exist (e.g. [1,2,16]), but human learners do not have access to exact articulations behind heard speech – they have to explore articulations using their own VTs and find which ones' acoustic output somehow matches with heard speech. Many studies of articulatory learning (e.g. [12,9,17]) are based on assumptions of some already acquired units of auditory perception before articulatory learning takes place, whereas in human speech learning articulation and acoustic perception are likely to coevolve [12,18,19].

In order to solve the speech inversion problem in a realistic setting with such limited information, or without exact articulatory targets for supervised learning, approaches requiring weaker supervision are needed. One option for such weakly-supervised algorithms is Reinforcement Learning (RL), aiming to improve imitation performance based on reward signals. In the last couple of years a few studies have shown that an actor-critic RL approach can be used to teach articulatory models to imitate speech ([3,20,4]. These approaches are based on the distal learning principle of [23], where a learned forward model can be used to backpropagate an acoustic error back to error in articulation, that can be again used to train the inverse model[1]. We have used this approach in our recent study ([26]), and in this paper, we have adapted the trained actor network as well as the articulatory synthesizer to perform audio-visual speech imitation in a web browser, based on input recorded by the user.

2. Components of the synthesizer

LeVI VT-model, developed based on the model of Mermelstein [10] is used for the synthesis, visualization, as well as the learning phase of speech inversion. The model is introduced in detail in ([27]). Positions of articulators are controlled with 9 parameters: tongue base (2 parameters), tongue tip (2 parameters), hyoid bone position, velum opening, jaw angle, lip protrusion and lip length. The control of these parameters is gesture-based (see e.g. [28]): every parameter can be given an individual *target position* (or target value, e.g. for the jaw angle) that will be reached in a given *target time*. The trajectory from the existing state of the VT parameters (their position, velocity and acceleration at the given moment) to the given target follows a minimum-jerk trajectory, known for example from human arm movements [29]. In the reported simulations new target positions and target times are updated every 150ms, i.e. speech is processed in 150ms frames. The VT-model updates the actual positions dynamically inside each frame in 10ms steps. The glottal excitation to the model takes one of three discrete states (silence, noise or voiced) for each 10ms step. Using airflow as a continuous variable has not yet been explored. The VT-model is originally implemented in Matlab, then automatically converted to C++ with Matlab Coder, and further compiled into a WebAssembly (WASM) library using emscripten[2]. As such it can be integrated with JavaScript (JS) into a web application which can be run fully independently in the browser (i.e., without any additional requests to any server). The graphical user interface is built completely in HTML and JS, and the drawing of the VT-configuration is based on articulator positions set by the user or by the randomly babbling or imitating VT-model.

In order to attempt to invert speech input recorded by the user, a previously trained ([26]) Tensorflow [30] model of the actor (inversion) network (see section 4) is converted into a Tensorflow.js [3] model that can be used directly from the JavaScript code. In the current version of the model, mel-frequency cepstral coefficients (MFCC) and harmonics-to-noise ratio (HNR) are used as the acoustic features input to the inversion network. Whereas during training of the Tensorflow RL model, these features were obtained with Parselmouth (v0.4.3; [31]), the Python library could not be used during imitation in the browser. Instead, the exact same features are calculated through a WASM-compiled version of the MFCC and HNR algorithms implemented in the corresponding version of Praat (v6.1.38; [32]).

3. LeVI vocal tract model on web browser

The audio-visual web browser based VT-model consists of an SVG object that represents the midsagittal image of the human vocal tract. The contour of the VT is loaded from an image file, but the movable parts are represented as SVG objects, such as lines and a circles. The parameters can be manually moved with the mouse (or touch), and the synthesizer automatically plays a short synthesis based on the new vocal tract configuration after doing so. It is also possible to start continuous voiced glottal excitation, as well as random movement of the articulators, in which case random targets and their times are periodically updated for all the articulatory parameters.

The audio processing is performed using the Web Audio API[4]. Two audio buffers of 150 ms in length, with a sampling frequency of 16kHz are created, as well as two audio buffer sources. In order to obtain continuous audio output, the playback of the two buffers is alternated: while one buffer source is playing, the other source is filled with synthesizer output samples for the following 150 ms, and scheduled to start playback 150ms after the first source started. The WASM implementation of the synthesis works faster than real-time to ensure gapless audio. While an audio buffer is filled, the VT-model is also requested to output the parameters of the VT for every 10ms frame during the synthesis, and every 10ms, the SVG visual is updated based on the current parameter values.

4. Training the inversion model

Recent research has shown that actor-critic RL-approaches can learn approximate solutions for speech inversion. In this approach the actor-network performs speech inversion, i.e. takes acoustic speech as input, and outputs parameter values controlling an articulatory model. In traditional RL-approaches the critic network learns to output the expected quality of a performed action, and is trained for example by using the

[1] Human speech learning does consist of other signals as well, such as non-vocal parental feedback [24], parental imitations [25] or visual cues, but these reward signals are out of the scope of this study.

[2] https://emscripten.org
[3] https://www.tensorflow.org/js
[4] https://www.w3.org/TR/webaudio/

temporal difference loss (see [33]). In the recent studies however, the critic is adjusted to reconstruct the output of the vocal tract: i.e. it effectively learns a differentiable version of the forward articulatory model that can then be used to backpropagate the error gradients between the imitation output and the speech to be imitated to the actor network ([2,22,3,26]).

We follow the approach of [2], where the actor learns a deterministic policy, i.e. learns a deterministic (as opposed to stochastic) VT action to be taken in order to imitate given input speech. The algorithm used is the Deterministic Policy Gradient algorithm [34], with the critic adjusted to output acoustic features in the same dimensionality as the LeVI VT-model.

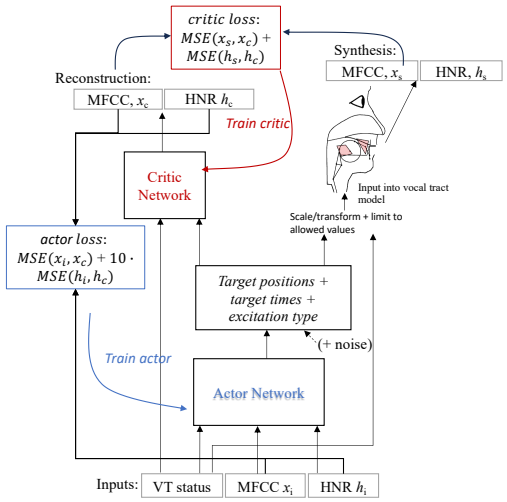

Figure 1. *Architecture of the actor-critic RL-model, used to train the inversion (actor) network.*

4.1. Speech data for training

The speech data used to train the model consists of all the utterances of the Caregiver Y2 UK corpus [35], synthesized with Microsoft Azure speech synthesizer, voice "en-US-GuyNeural". White gaussian noise at -80 dBFS is added to all utterances to avoid sequences of zeros in the resulting signals. In principle, any speech input of any language can be used to train the model. We chose to use this synthesized speech for the reason that the inversion performance can be later evaluated by using the pronunciation assessment feature of Azure using the same language locale "en-US" as used for the input speech. As the used LeVI VT-model is configured to produce speech with an average adult male VT length of 17.5 cm, we expect the current inversion demonstration to work better for male voices. The demonstrated methodology could however be perfectly adapted to different lengths of VTs, or even completely different morphologies, such as animal VTs. We have not yet explored techniques for VT normalization in order to adapt the inversion training for speakers of different VT lengths. MFCCs are extracted in 10ms steps with a Gaussian window of 25 ms. HNRs are extracted with 10ms step size, minimum pitch of 60Hz, silence threshold of 0.1 and number of periods per window of 4.5. The energy coefficient of the MFCCs is discarded.

4.2. Training

2000 randomly selected utterances of the created data are used for training. On every episode, training of the network proceeds as follows. A random utterance is sampled from the training set and gone through from the beginning to the end in 150ms steps. The actor model is used to predict targets and target times for the 9 VT parameters. Its input (or the *state*) at time t consists of the 12-dimensional MFCC-vectors and HNR values from a time interval [t-150ms, t+150ms], as well as the positions, velocities and accelerations of the 9 vocal tract parameters at time t. Random gaussian noise is added to the predicted actions during training in order to encourage exploration (noise is not added during testing or imitation). The predicted (noisy) action is fed to the vocal tract model. States, actions and VT outputs from each step are stored in a replay buffer of size 10,000.

After every performed action (when the buffer has at least 62 samples in our implementation), a minibatch of 32 exemplars are randomly sampled from the buffer and used to train the critic and the actor networks. The stored (noisy) action and the VT-status are first input to the critic, critic loss L_c (see below) is calculated using the stored synthesizer output, and the critic network is trained. Next, the stored state is fed through the actor network, the resulting action and the stored VT-status is fed through the critic network (without added noise), loss L_a and its gradients related to the actor network parameters is calculated, and the actor network is updated.

The critic network is configured to output an MFCC-feature matrix and a HNR vector for [t, t + 150ms]. LeVI is also set to synthesize 150ms of speech. Resultingly, the task of the actor network becomes to find actions that imitate the input speech in the period [t, t + 150ms]. The critic network is trained to minimize the critic loss $L_c = MSE(x_s, x_c) + MSE(h_s, h_c)$, where x_s and h_s are the synthesized MFCC and HNR features correspondingly, and x_c and h_c are the MFCC and HNR features output by the critic. MSE is the mean squared error. The actor network is trained to minimize the actor loss $L_a = MSE(x_i, x_c) + 10 \cdot MSE(h_i, h_c)$, where x_i and h_i are the corresponding input features. A heuristically found factor of 10 for the HNR component when training the actor network is given to balance its contribution to the total error. Adam-optimizer with learning rate of 0.001 are used for both networks. The architectures of the actor and critic models are shown in the supplementary material. The actor and critic networks have 38,329 and 103,415 trainable parameters correspondingly. Tensorflow version 2.15.0 was used for training and testing. Training for 50,000 episodes took approximately 12 hours of time.

Every 300 training episodes, the performance of the inversion is tested with a separate test set of 50 words not used in training. For this phase only the actor network and the VT-model are used. $MSE(x_s, x_i)$ and $MSE(h_s, h_i)$ between the test set input and the inverted and synthesized speech features are shown in Figure 2 during the training. It can be seen that the inversion system learns articulatory actions that lower both error measures, thus the acoustic features of the imitations approach the original speech. The decreasing error curve is very similar to the study of [3], using similar training methodology.

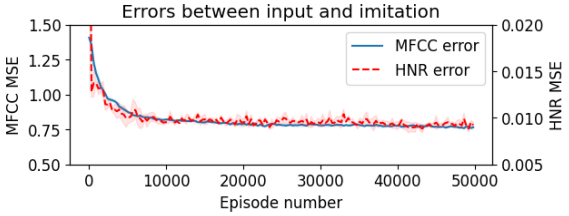

Figure 2. *Mean squared error between the input and imitated MFCC and HNR during the training phase.*

5. Using the inversion model on a web browser

After the actor model has been trained, it is converted into a tensorflow.js model to be used in JS. Since the JS model does not support a Lambda-layer used for slicing the excitation type output, the slicing layer is manually removed from the resulting json-file, and slicing is added on top in JS. The web application includes a simple recording tool with which the user can record speech audio which the model aims to imitate. Since the recorded audio (with the MediaStream Recording API) is delivered to the application in "chunks" of varying size, the audio processing can be significantly sped up by performing the WASM-based feature extraction on 0.5 second segments already during the recording, as soon as enough speech data is available (as opposed to waiting for the whole recording to be finished before extracting features). After audio has been recorded, clicking the "Imitate input" button triggers the speech inversion based on the extracted features.

Since the predicted vocal tract action does not only depend on the input acoustic features, but also on the current status of the VT, the synthesizer needs to predict and also perform the action before the next prediction can be made, in order to have access to the following VT status. To ensure fluent audio output and thus to avoid the requirement to perform both the actor model prediction and the articulatory synthesis inside the rather short 150ms frame, the imitation in this version is performed in two steps. In the first step, every 150ms a VT action (+ excitation) is predicted, and the synthesizer is run using the predicted action in the background for 150 ms (without graphical or audio output) to obtain the next vocal tract status. The performed action and excitation are saved in an action buffer. This process is repeated until the complete feature vector is processed, resulting in a bank of VT actions. In the second step, these actions are performed in a sequence, inside the playback loop (see section 3), the obtained audio is output, and the vocal tract positions are drawn. Prediction of a VT-action based on the tensorflow-js model takes usually less than 100ms for each 150ms frame to be processed.

6. Discussion

Since the purpose of this study is mainly to introduce a proof-of-concept of a web browser-based speech inversion application, we do not at this point provide a detailed objective measure of its performance. However, the reader is encouraged to test the model in the link given on the first page. Figure 2 demonstrates that the used methodology learns to approach the acoustic goal it was given, i.e. minimize the error between the acoustic features between input audio and the synthesizer output. This means that in terms of the used features the output of the imitation should be close to the recorded audio, but its subjective likeness to natural speech sounds or its articulation may not always meet one's expectations. However, here a few observations based on testing the web-browser based implementation are discussed.

Pronunciation assessment of Azure on 50 utterances from the imitated test set give an average score (in the scale of 0-100) of 48.5 for vowel and 32.6 for consonant phonemes, when the original annotations are used as the reference text. For a sanity check, for the original Azure synthesized sentences the corresponding values are 99.1 and 99.2. The model generally imitates silence as silence, and voiced sounds as voiced, thus the rhythm of imitations compared to the recorded speech remains quite natural. With recorded silence as input, the articulators are moving quite restlessly, and the excitation type seems to be noise, resulting to hiss or whisper type output. This is presumably due to the HNR of recorded silence resembling noise, and without having signal energy as a feature in the optimization goal, the model imitates this by producing a random noisy output signal. For future experiments, overall signal energy could be included as a feature to reproduce silence more accurately. As such, the model seems to be also sensitive to background noises in the recording, leading to unnecessary voiced articulations in order to "imitate" unwanted sounds. The system might benefit from additional features that separate speech sounds in the recordings from unwanted noise.

Imitating a long vowel /i/ seems to be approximately correctly imitated as a frontal vowel, and /a/ as a back vowel, but again the unnecessary movements of some articulators distort the clarity of the vowels. Transitions between vowels can be perceived similar to the input, whereas consonant sounds can be articulated quite incorrectly. Sometimes long vowels seem to have some distortion due to individual frames of the excitation signal shifting to noise instead of voiced. Adding a cost for vocal tract effort or for change in the glottal excitation type should be explored in the future to stabilize the performance further.

Based on the first experiments, the training of the inversion network could be improved in several ways. More training data, and more network complexity, such as inclusion of convolutional layers or regularization techniques, might help to reduce the acoustic error further. Measure of energy should be included in the features to be optimized, as well as some penalty on articulatory effort. Acoustic features could be tailored to give more weight to features of consonant sounds that tend to be shorter in duration than vowels and thus weighing less in the final loss measures. VT-normalization methods suitable for the training procedure should be developed to accommodate speakers with differing VT morphologies.

The discussed inversion performance is achieved with the learner having only the acoustic features of heard speech as a target for learning. Human infants have access to other types of feedback as well, such as smiling or touching encouraging correct pronunciations [24], visual perception of speakers' faces, imitative feedback by their caregivers [25], or direct corrective feedback. Since the tensorflow models are also trainable from JS, some of these types of feedback would be in principle possible to include in the web application and allow the actor-network to continue learning based on user input.

We have introduced the first prototype of an acoustic-to-articulatory speech inversion system running on a web browser. Even though the imitation performance is not very accurate, some basic aspects of input speech are correctly imitated. For the development of functioning speech applications such implementations that work in an everyday environment with real, recorded speech signals are crucial to point out problems that may remain hidden when systems are evaluated based on tailored acoustic/articulatory metrics or in clean or simulated test environments. Based on fast experimentation of the prototype, problem points in the training procedure could already be identified to be taken into account in future versions.

7. Acknowledgements

HR was funded by a Senior Postdoctoral Fellowship (1258822N) of the Research Foundation – Flanders (FWO).

8. References

[1] P. Wu *et al.*, "Speaker-Independent Acoustic-to-Articulatory Speech Inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096796.

[2] H. Shibata, M. Zhang, and T. Shinozaki, "Unsupervised Acoustic-to-Articulatory Inversion Neural Network Learning Based on Deterministic Policy Gradient," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Jan. 2021. doi: 10.1109/slt48900.2021.9383554.

[3] M.-A. Georges, J. Diard, L. Girin, J.-L. Schwartz, and T. Hueber, "Repeat after Me: Self-Supervised Learning of Acoustic-to-Articulatory Mapping by Vocal Imitation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022. doi: 10.1109/icassp43922.2022.9747804.

[4] S. Medina *et al.*, "Speech Driven Tongue Animation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 20374–20384. doi: 10.1109/CVPR52688.2022.01976.

[5] I. Howard and M. Huckvale, "Training a vocal tract synthesiser to imitate speech using distal supervised learning," in *Proc. SpeCom: 10th International Conference on Speech and Computer*, University of Patras, Wire Communications Laboratory, 2005, pp. 159–162.

[6] D. Südholt, M. Cámara, Z. Xu, and J. D. Reiss, "Vocal Tract Area Estimation by Gradient Descent," 2023, doi: 10.48550/ARXIV.2307.04702.

[7] P. Birkholz, "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, Apr. 2013, doi: 10.1371/journal.pone.0060603.

[8] K. L. Markey, "The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development," Citeseer, 1994. Accessed: Mar. 25, 2013.

[9] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *The Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 321–328, Aug. 1981, doi: 10.1121/1.386780.

[10] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, Apr. 1973, doi: 10.1121/1.1913427.

[11] F. H. Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production.," *Psychological Review*, vol. 102, no. 3, pp. 594–621, 1995, doi: 10.1037/0033-295x.102.3.594.

[12] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, Aug. 2011, doi: 10.1080/01690960903498424.

[13] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011.

[14] S. Maeda, "Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Springer Netherlands, 1990, pp. 131–149. doi: 10.1007/978-94-009-2037-8_6.

[15] J. Drayton, E. Miranda, and A. Kirke, "A comparison of fitness functions in a genetic algorithm for acoustic-articulatory parameter inversion of vowels," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Berlin Germany: ACM, Jul. 2017, pp. 271–272. doi: 10.1145/3067695.3076112.

[16] P. Boersma and V. Van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[17] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015. doi: 10.1109/icassp.2015.7178812.

[18] S. Hiroya and M. Honda, "Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, Mar. 2004, doi: 10.1109/tsa.2003.822636.

[19] D. R. van Niekerk *et al.*, "Simulating vocal learning of spoken language: Beyond imitation," *Speech Communication*, vol. 147, pp. 51–62, Feb. 2023, doi: 10.1016/j.specom.2023.01.003.

[20] J. I. Skipper, J. T. Devlin, and D. R. Lametti, "The hearing ear is always found close to the speaking tongue : Review of the role of the motor system in speech perception," *Brain and Language*, vol. 164, pp. 77–105, Jan. 2017, doi: 10.1016/j.bandl.2016.10.004.

[21] H. Mitterer, O. Scharenborg, and J. M. McQueen, "Phonological abstraction without phonemes in speech perception," *Cognition*, vol. 129, no. 2, pp. 356–361, Nov. 2013, doi: 10.1016/j.cognition.2013.07.011.

[22] Y. M. Siriwardena, C. Espy-Wilson, and S. Shamma, "Learning to Compute the Articulatory Representations of Speech with the MIRRORNET," in *INTERSPEECH 2023*. ISCA, Aug. 2023. doi: 10.21437/interspeech.2023-562.

[23] M. I. Jordan and D. E. Rumelhart, "Forward Models: Supervised Learning with a Distal Teacher," *Cognitive Science*, vol. 16, no. 3, pp. 307–354, Jul. 1992, doi: 10.1207/s15516709cog1603_1.

[24] M. H. Goldstein, A. P. King, and M. J. West, "Social interaction shapes babbling: Testing parallels between birdsong and speech," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 13, pp. 8030–8035, Jun. 2003, doi: 10.1073/pnas.1332441100.

[25] T. Kokkinaki and G. Kugiumutzakis, "Basic aspects of vocal imitation in infant-parent interaction during the first 6 months," *Journal of Reproductive and Infant Psychology*, vol. 18, no. 3, pp. 173–187, Aug. 2000, doi: 10.1080/713683042.

[26] H. Rasilo, Y. Jadoul, and B. de Boer, "Distal Learning vs. Temporal Difference Policy Gradient Algorithms in Self-Supervised Speech Inversion Tasks," *submitted*, 2024.

[27] H. Rasilo and O. Räsänen, "An online model for vowel imitation learning," *Speech Communication*, vol. 86, pp. 1–23, Feb. 2017, doi: 10.1016/j.specom.2016.10.010.

[28] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, Dec. 1989, doi: 10.1207/s15326969eco0104_2.

[29] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *The Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, Jul. 1985, doi: 10.1523/jneurosci.05-07-01688.1985.

[30] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016, doi: 10.48550/ARXIV.1603.04467.

[31] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, Nov. 2018, doi: 10.1016/j.wocn.2018.07.001.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." Feb. 03, 2018.

[33] R. S. Sutton and A. Barto, *Reinforcement learning: an introduction*, Second edition. in Adaptive computation and machine learning. Cambridge, Massachusetts London, England: The MIT Press, 2020.

[34] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning.," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2016.

[35] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, and H. van den Heuvel, "A Speech Corpus for Modeling Language Acquisition: CAREGIVER," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Robot Language Acquisition Modelling via Cross-Situational Learning with Little Data

Xavier Hinaut[1,2,3]

[1]Inria centre of Bordeaux University.
[2]LaBRI, Bordeaux University, Bordeaux INP, CNRS UMR 5800.
[3]Bordeaux University, CNRS, IMN, UMR 5293, Bordeaux, France
xavier.hinaut@inria.fr

Abstract

How do children bootstrap language through noisy supervision? Most prior works focused on tracking co-occurrences between individual words and referents. We model cross-situational learning (CSL) at sentence level with few (1000) training examples. We compare two recurrent neural network architectures often use as cognitive models: reservoir computing (RC) and LSTMs on three datasets including complex robotic commands. Surprisingly, reservoirs demonstrate robust generalization when increasing vocabulary size: the error grows slowly compared to an LSTM of fixed size. This suggests that that random projections used in RC helps to bootstrap generalization quickly. How robots acquire basics of language like in child-caregiver (Human-Human) interactions could give hints of how to link animal vocalisations with behaviour in ambiguous context. Cross-statistics between sequence of vocalisations and various contexts could probably be learnt in few trials by such Reservoir architecture.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Discussion

Comparison to other non-recurrent architectures It is likely that Transformers architecture [1] would require more data for training, thus the comparison at this tiny data scale (1000 examples) does not seem relevant. However, their attention mechanism is interesting, in particular to parse long sentences in some of the more challenging datasets that we tried [2]. In future work we will explore how such attention mechanisms can help reservoir computing to scale to much bigger datasets, enabling to have an architecture able to generalize from tiny to big datasets.

2. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] S. R. Oota, F. Alexandre, and X. Hinaut, "Cross-situational learning towards robot grounding," *HAL preprint*, 2022.

[3] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. Juven and X. Hinaut, "Cross-situational learning with reservoir computing for language acquisition modelling," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[6] A. Variengien and X. Hinaut, "A journey in esn and lstm visualisations on a language task," *arXiv preprint arXiv:2012.01748*, 2020.

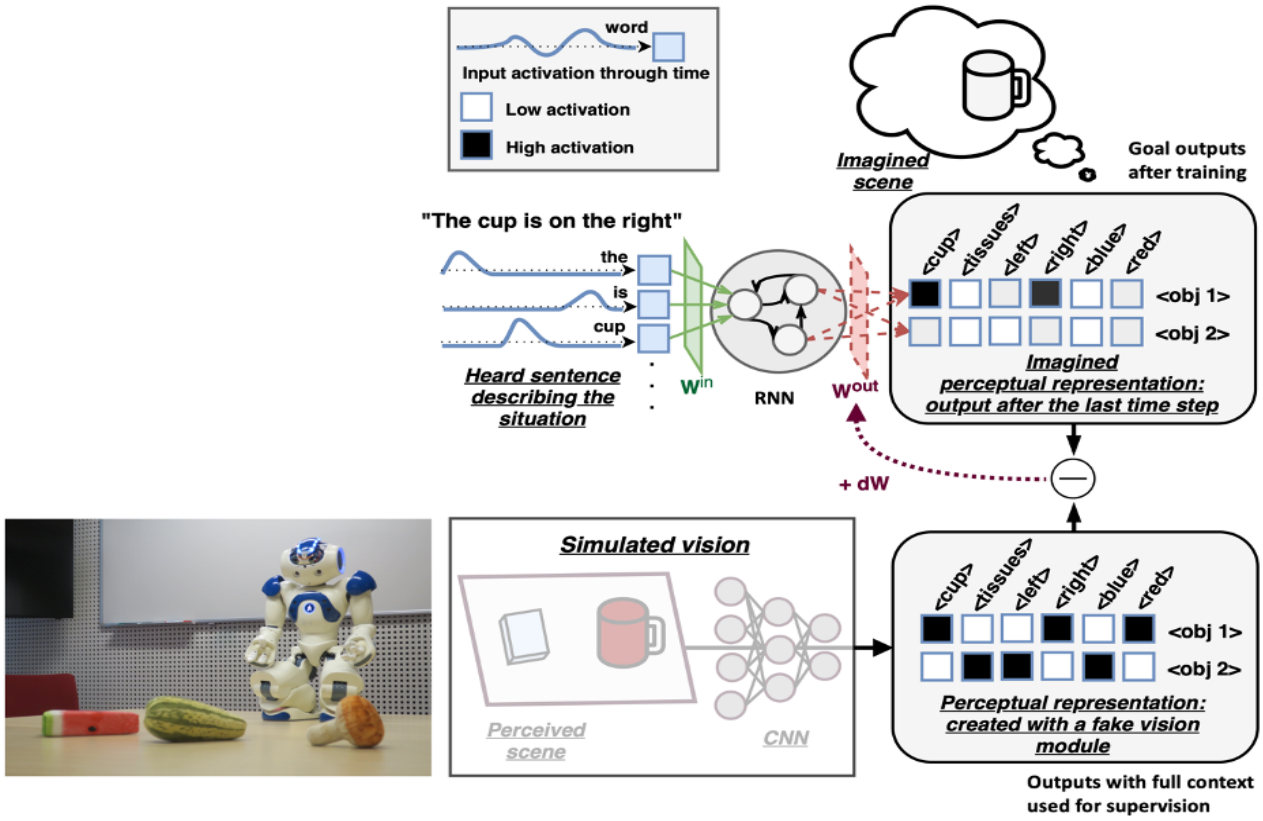

Figure 1: *The Cross-Situational Learning (CSL) learning procedure for a Recurrent Neural Network (RNN) architecture. We compare two RNNS: Reservoir Computing (RC) [3] and Long Short-Term Memory network (LSTM) [4]. The model has to reconstruct an imagined scene from the sentence given word by word. The simulated vision creates a perceptual representation corresponding to the full description of objects in the scene. This representation is used as target outputs for the reservoir, even if the sentence only partially describes the objects in the scene, or if it describes only one object. This particular set-up creates cross-situational learning conditions similar to the ones children are facing. The set-up, input and target outputs were the same for the LSTM experiments. (Image adapted from [5].*

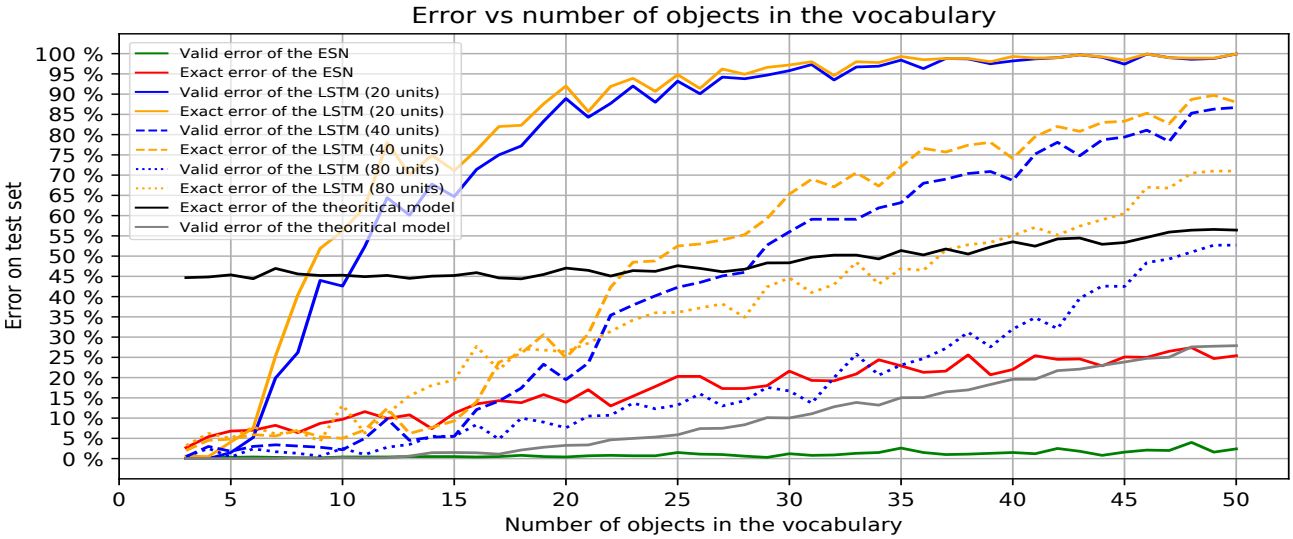

Figure 2: *Comparison of the performance of 5 models (1 ESN + 3 LSTMs + theoretical) for different number of objects in the dataset. Echo State Network (ESN) is a particular instance of the Reservoir Computing paradigm. The small LSTM (20 units), optimized to perform well on a dataset with 4 objects, is not able to keep good performance with a higher number of objects. The medium LSTM (40 units) trained for longer with dropout is able to outperform the ESN until 15 objects. The bigger LSTM (80 units) limits the rise of the error compared to the other LSTM. However, it comes with poorer performances even for a small number of objects. The ESN is able to keep an error below the theoretical model and all the LSTMs despite the fact that its hyper-parameters were optimized for the 4-object dataset. Image from [6].*

Vocal, Visual, and Tactile Signals in Cat–Human Communication: A Pilot Study

Elin N Hirsch, Joost van de Weijer, Susanne Schötz

Lund University, Sweden

elin.hirsch@med.lu.se, joost.van_de_weijer@humlab.lu.se, susanne.schotz@med.lu.se

Abstract

To investigate multimodal signals in cat–human communication we recorded 36 cat–owner interactions in everyday situations that were judged by the owners for valence (negative, mixed or positive). We then coded the videos for behaviour using an ethogram including vocal, visual and tactile (multimodal) signals. Vocalisations were segmented and acoustic measures of duration and F0 obtained. In cats, common behaviours were tail up/halfway up and ears forward, while vocal signals were more common in owners. The distribution of all behaviours was compared across the three levels of valence. In negatively judged interactions, cat tail position was frequently vertical. In interactions judged as mixed, cats remained passive to their owners trying to interact with them. Frequent cat behaviours in positively judged interactions were sniff/lick, rub, and soft gaze. The acoustic variables did not show clear variation that could be attributed to judged valence.

Index Terms: cat–human interactions, interspecific communication, multimodal signals

1. Introduction

1.1. Human–animal communication

Communication is the transmission of a signal (e.g., vocal, visual or tactile behaviour) from an emitter to a receiver [1]. Animals (including humans) communicate – not exclusively with conspecifics, but also with individuals from other species. They do this to share information, and to express emotions and needs. For instance, interactions between companion animals and their human caretakers are common, and research findings suggest that these interactions are socially and semantically meaningful and therefore beneficial for our health and wellbeing [2]. Domestic cats (*Felis silvestris catus*) are one of our most popular pets, and communicative interactions between cats and their owners are common. Yet very little is known about the nature and the successfulness of these interactions. Well-functioning communication is crucial for a meaningful relationship with our companion animals and can help mitigate the risk for development of undesired behaviours.

1.1.1. Cat–human communication

Domestic cats, hereafter cats, and humans communicate using multimodal signals [e.g., [3]], and bimodal communication (visual and vocal) seems to be more attractive to cats than unimodal (vocal) [4]. With their origin as a solitary territorial species, cats primarily rely on olfactory signals, especially in the social interaction with other cats [5]. However, since their domestication approximately 10 000 years ago [6], cats have lived in the proximity of both conspecifics and humans and adapted their communicative repertoire to include visual (body postures and movements) and tactile (body contact) signals as well as vocalisations [see [7]]. The visual and tactile signals used by cats in interactions with humans have probably evolved from social signals in interactions with other cats – primarily mother–young [8]. For example, cats typically interact with humans by rubbing and head bunting, and signal affiliative intent by using *tail up* (tail raised vertically, sometimes with the tip bent). Cats have also developed a large and highly varied vocal repertoire to get the attention of their human caretakers in order to reach their goals (e.g., receive food, access to outdoors). Although vocal cat–cat communication is common in sexual, territorial (agonistic) and social (e.g. mother–young) interactions [9], cats seem to use vocalisations – mainly meowing – more frequently in communication with humans [10]. The number of vocalisation (or call) types described in cats varies between three and 21 [11], [12]. Based on phonetic features, there seem to be at least nine major pure types (i.e., meowing, trilling, growling, hissing, howling/yowling, snarling/crying, purring, chirping, and chattering) with numerous subtypes and several combinations (e.g. trill-meowing) [9], [13]. Many cat owners, on the other hand, regularly talk to their cats, often using cat-directed speech, a speech style which is characterized by a high fundamental frequency, short phrase duration, and repetitiveness, similar to child-directed speech [e.g. [14], [15]].

As previous studies on communication between humans and cats have focused on unimodal (vocal or visual) or bimodal (vocal and visual) signals [e.g., [4], [16]], the present exploratory study, in contrast, investigates multimodal (vocal, visual and tactile) signals in cat–human interactions. The aims were to examine 1) which multimodal cat and human communicative signals are the most frequent ones in interspecific interactions, and 2) how signals differ between interactions judged by owners as negative, mixed or positive. We expected humans to talk much to their cats and cats to use many visual (tail, ears) and tactile (rub, touch) signals, but also vocalisations to communicate with humans.

2. Material and methods

2.1. Ethical approval

The study was approved by the Swedish Ethical Review Authority (no. 2022-04514-01). Participants received information, oral and written, about the project's purpose, approach, and use of recordings and signed a written informed consent form before participation.

2.2. Subjects and materials

Seven owners (5 female, 2 male) of 15 (6 female, 8 male) cats (13 domestic shorthair, 2 mixed) in 15 dyads recruited from personal connections participated in the study. Three owners had one cat, two owners had two cats, one owner three and one

five cats. All cats were at least one year old and had no known health issues which may have interfered with their behaviour (e.g., pain).

Recordings took place in the cats' home environment during everyday interactions between the cats and owners, such as cuddle, feeding, and care. At the end of each recording, the owners judged the valence of the recorded interaction as either negative, mixed or positive. All recordings were made with two Insta360 Go 2 wide-angle cameras [17]. One camera was mounted on a tripod placed in the room and recorded the interaction from a distance. The other camera was head-mounted and recorded the interaction from the owner's perspective. A total of 36 interactions were recorded and judged by the owner.

2.3. Behavioural coding

Cat and owner behaviour was manually annotated by the first author using the Behavioural Observation Research Interactive Software (BORIS) [18] using an ethogram based on previously described behaviour in cat–human interactions [12], [19], [20], [21], [22], [23], [24] (Table 1).

Five randomly selected recordings were independently annotated also by the second author. The proportion agreement between the two labellers varied from 0.84 to 1.00 for the behavioural categories separately. The overall interrater reliability of the five clips was measured as the intraclass correlation coefficient. The value was 0.95, suggesting that the reliability was satisfactory.

2.4. Acoustic-phonetic coding

All vocal signals were segmented from the recordings of the head-mounted camera by the third author using the speech analysis software Praat [25]. Cat vocalisations were labelled for vocalisation (call) type (e.g., meowing, trilling, hissing, growling) based on phonetic features according to [26]. Owner speech was labelled as either human-directed (the valence judgments at the end) or cat-directed. Acoustic measures of duration and fundamental frequency (F0) of the cat and owner vocal signals were obtained in Praat.

Table 1: Ethogram of coded behaviours.

Behaviour	Description
Human only	
Care	Human handles cat with a caring intent or to provide medicine
Lift/hold	Human lifts and/or actively holds cat
Reach/invite	Human invites interaction by reaching body part or object towards cats or using inviting signals
Stroke/scratch	Human strokes, pets and/or scratches the cat using hand or tool
Touch other	Human initiates body contact with cat not using the hands
Cat only	
Crouching	Cat positions body close to the ground, all four legs are bent, and the belly is touching (or raised slightly of) the ground
Ears back/angled/ flattened	Cat holds ears facing backwards (rotated) and/or flattened
Ears forward	Cat holds ears in neutral and/or forward-facing position
Ears other	Other ear positions not described in the ethogram including combination of ear positions
Lip licking	Cat licks its lip(s)
Locomotion	Cat is moving in a forward, sideways or backward motion
Lying	Cat has body placed in a horizontal position, on its side, back, belly, or curled in a circular formation
Rub	Cat rubs head or body on human or wraps tail around the human's body
Sitting	Cat is in an upright position, hind legs are flexed and resting on the ground, while front legs are extended and straight
Sniff/lick	Cat smells or licks human
Soft gaze	Cat gazes softly at human
Standing	Cat is immobile, with only paws on the ground and legs extended, supporting the body
Stretch	Cat extends either front or back legs away from the body or arches back with legs extended
Tail other	Other tail positions not described in the ethogram including combination of tail positions
Tail wrapped	Cat holds tail wrapped close to body, with/without around or under body
Tail down	Cat holds tail down, in a relaxed manner, with/without with end kinked out
Tail fast	Cat moves tail, or tip of tail, fast in a lashing, thrashing or wagging sideways motion
Tail parallel	Cat holds tail straight or slightly curved parallel to ground, standing, sitting or lying down
Tail slow	Cat moves tail, or tip of tail, slowly in a soft wagging sideways motion or soft quivering
Tail up/ halfway up	Cat holds tail in an upright, or half-way up, position, with/without with a small curve of the tip
Tail vertical	Cat holds tail rigid and facing down, with/without the tail base turned up
Touch/knead/ tread	Cat initiates/is in body contact with human or pushes forepaws into the ground near (<50 cm) or at human in a rhythmic, kneading motion
Both human and cat	
Approach	Cat/human moves body towards human/cat
Feed/eat	Human offers food or treat(s) to cat or cat ingests food or treat(s)
Interaction other	Other behaviours relating to interactions not described in the ethogram
Leave/dodge	Cat or human actively avoids interaction
Passive	Cat or human remains passive towards interaction initiation by the other
Play	Cat and human interact together with an object in a "non-serious" playful manner, or cat or human interacts with object in a "non-serious" playful manner
Slow blink	Cat or human performs an intentional blink or series of half-blinks followed by a half or full closing of eyes
Vocalisation	Cat or human produces a sound with the voice

2.5. Analysis

The focus of the analysis is the distribution of the behaviours in relation to the owners' judgments of the interactions (i.e., negative, mixed or positive). For this purpose, we created a contingency table of the two variables and performed a correspondence analysis to visualize their associations. Additionally, we compared acoustic characteristics of the owner and the cat vocalisations also in relation to the owners' judgments. Below, we present vocalisation duration and F0 modulation, a measure that correlates with the amount of variability in an utterance, and which is independent of a speaker's average fundamental frequency. It is calculated as the F0 standard deviation divided by the F0 mean [27]. The analysis was performed in R [28]. For the correspondence analysis, we used the package "ca" [29].

3. Results

The durations of the 36 recordings varied from about 0.7 to 8.0 minutes, with an average of 2.4 minutes. The owners judged 27 interactions as positive, 7 as mixed, and 2 as negative. Table 1 shows an overview of the contexts in which the 36 interactions were recorded including owner judgements of valence.

Table 2: *Recording contexts and judged valence.*

Context	n	Negative	Mixed	Positive
Call	4	-	-	4
Care	3	-	1	2
Cuddle	4	1	-	3
Food	14	-	3	11
Groom	3	-	2	1
Lift	2	1	-	1
Obstacle	3	-	1	2
Play	2	-	-	2
Treat	1	-	-	1
Total	36	2	7	27

3.1. Behaviours

A total of 2 154 behavioural events were annotated, 1 044 by the cats, and 1 110 by the owners. There were 872 owner vocalisations addressed to their cats, ranging from 5 to 66 vocalisations per interaction with an average of 24.2. Cat vocalisations were considerably less frequent, ranging from 0 to 6 with an average of 0.9. Cat tail position and movement was annotated 275 times. Most of the time (n = 82), the position was *tail up/halfway up*. Other frequent tail positions were *tail parallel* (n = 64) and *tail down* (n = 20). Tail movements included *tail slow* (n = 57) and *tail fast* (n = 37), Similarly, the ear positions of the cats were annotated 170 times. This position was predominantly *ears forward* (n = 117) or *ears back/angled/flattened* (n = 45). Other frequent behaviours were *locomotion* (n = 112), *feed/eat* (n = 44), *standing* (n = 95), and *stroke/scratch* (n = 69).

3.2. Associations between behaviours and judged valence

Figure 1 shows the association between the behaviours and the owner judgements as a biplot resulting from a correspondence analysis of the two variables. The positions of the judgements suggest that the dimensions of the plot represent the difference in valence. Positive interactions are plotted towards the lower right end, negative interactions towards the upper left end, and mixed interactions towards the lower left end. Stronger associations are suggested by the distance from the behaviours from the crossing of the horizontal and vertical zero lines. Behaviours that are close to this crossing (e.g., *owner vocalisation, tail slow*, and *locomotion*) are not strongly associated with the judgements while behaviours far away are. The plot suggests that behaviours associated with positive interactions are close to the horizontal line in the right part of the graph. Examples of these are *sniff/lick, soft gaze, cat approach, rub, touch/knead/tread,* and *ears other*, but also with *human leave/dodge* and *human passive.* Behaviours typically associated with the negative interactions are in the higher left part of the graph. These are *tail vertical* and *human interaction other*. Behaviours associated with mixed interactions, finally, are in the lower left part of the graph, including *touch other* and *cat passive*. Many of the remaining behaviours are located rather close to the zero-crossing, and therefore not strongly associated with the judged valence.

3.3. Acoustic characteristics

Figure 2 shows boxplots of the durations and F0 modulations in cat and owner vocalisations. Since the number of negative interactions was comparatively small, and there were relatively few cat vocalisations, there was only one cat vocalisation in negative interactions. Typical durations of owner and cat vocalisations were approximately one second and half a second, respectively. The durations of the vocalisations do not appear to differ drastically between the interactions, even though cat vocalisations were longer on average in positive interactions than in mixed interactions. The F0 modulations were also larger in owner vocalisations than in cat vocalisations and did not appear to differ systematically by the interaction evaluation.

4. Discussion

In this pilot study we examined multimodal signals used during 36 interspecific cat–human interactions. In cats, the most common signals were visual. The most frequent tail positions was *tail up/halfway up*, and the most frequent ear position was *ears forward.* This is in line with [15], [30] who reported that the majority of cat–human interactions were initiated using these tail and ear positions. Cat vocalisations were not very frequent in our sample. This was somewhat surprising, as cats have learned to use vocal signals when communicating with humans [10, pp. 67–93]. Our results may be explained by the recording situation which was unfamiliar to both owners and cats. Cats are extremely sensitive to changes in their environment, including unfamiliar objects and individuals as well as the behaviour of their owners [31]. However, no owner stated that their cat was bothered by the recording equipment.

As expected, owners frequently talked to their cats [15]. However, the duration and F0 modulation of the owner vocalisations did not vary with interaction valence.

We also investigated how communicative signals differ between contexts judged by the owner as positive, mixed or negative. The correspondence analysis of behaviour and judged valence showed that most cat and owner behaviours occurred in all three levels of judged valence. Some behaviours, however, were associated more strongly with judged valence. In cats, for instance, *sniff/lick, soft gaze, cat approach, rub, touch/knead/tread,* and *ears other* were more strongly associated with positive interactions, while *tail vertical* was more strongly associated with negative interactions. It is possible that *ears other* can be explained as a positive behaviour as many cats turned one or both ears towards their owners when they were talking to them from behind. *Rubbing* has been reported in positive interactions by [32]. In owners, the behaviour *touch other* was strongly associated with interactions judged as mixed. This may be explained by that many owners kissed their cats on the head or back, perhaps to encourage the cats to interpret an interaction as more positive.

The results from this study should be regarded as tentative for several reasons. First, the material and number of participants was small. Secondly, the material was unbalanced as most of the interactions were judged by the owners as positive. Third, there was an uneven distribution of contexts, valence, and cats per owner which could have influenced the results. In future studies, we will record a larger material and investigate which visual and tactile signals are combined with vocalisations in interactions of different valence.

5. Conclusions

Because of the small number of interactions and the unbalanced judgements of valence in this material is difficult to draw any general conclusions. However, there seem to be differences in both cat and owner behaviours in interactions judged as negative, mixed and positive.

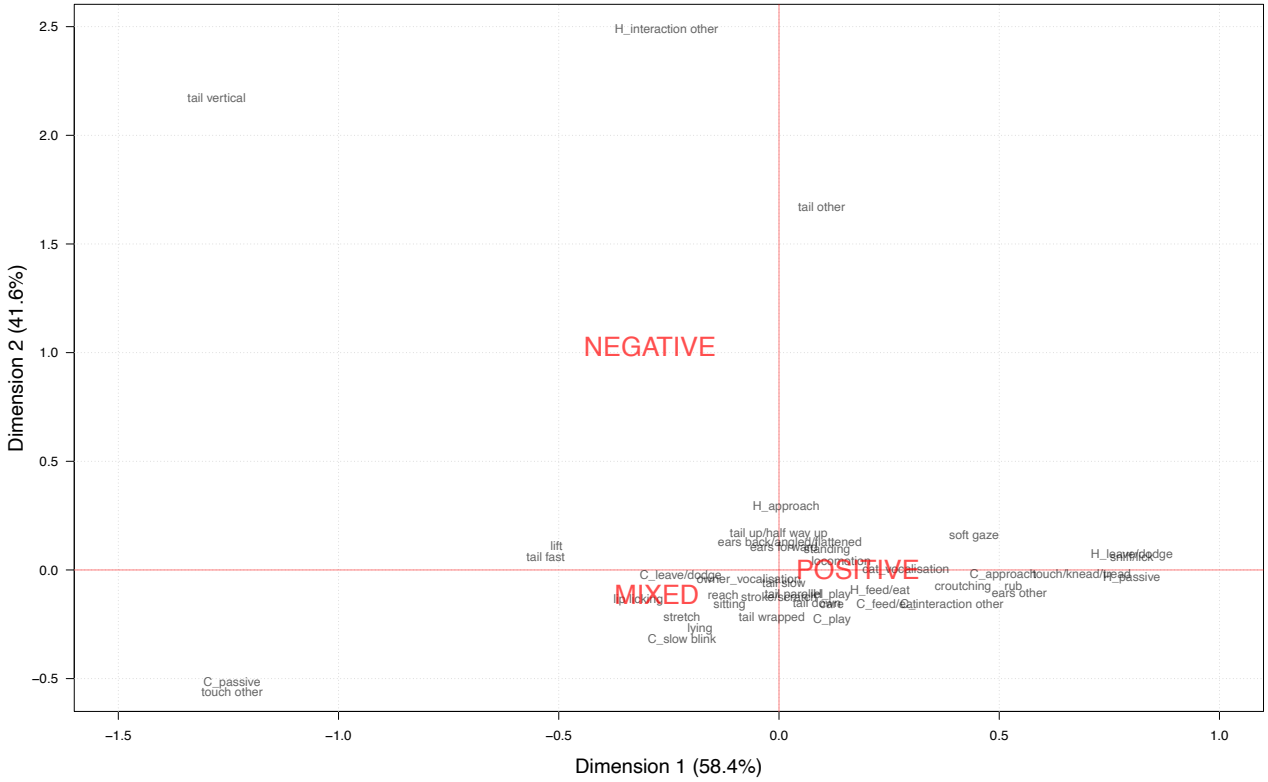

Figure 1: Correspondence analysis biplot of the annotated cat and human behaviours and owner judgements (two negative, seven mixed and 27 positive). Behaviours which may be found in both species are marked with an initial C for cats, e.g., C_play, and an initial H for humans, e.g., H_approach.

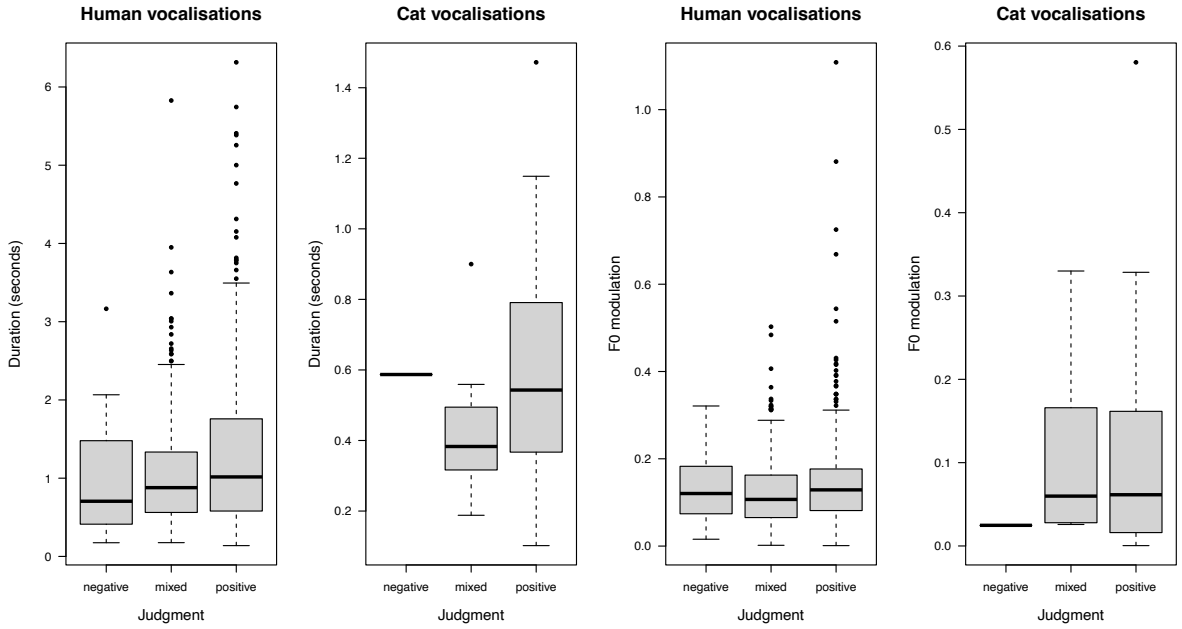

Figure 2: Box plots of the acoustic characteristics duration and F0 modulation of cat and human vocalisations in the owner judgements (two negative, seven mixed and 27 positive). Please note that the scales differ in order to better show all the data used in each plot.

6. Acknowledgements

The authors wish to acknowledge Agria och Svenska Kennelklubben Forskningsfond (grant number N2021-0001) for financial support. A special thanks goes to all participating cats and their owners.

7. References

[1] S. J. Shettleworth, 'Communication and Language', in *Cognition, Evolution, and Behavior*, Oxford University PressNew York, NY, 2009, pp. 508–547. doi: 10.1093/oso/9780195319842.003.0014.

[2] M. R. Pastorinho and A. C. A. Sousa, *Pets As Sentinels, Forecasters and Promoters of Human Health*. Cham: Springer, 2020.

[3] A. Quaranta, S. d'Ingeo, R. Amoruso, and M. Siniscalchi, 'Emotion Recognition in Cats', *Animals*, vol. 10, no. 7, p. 1107, Jun. 2020, doi: 10.3390/ani10071107.

[4] C. De Mouzon and G. Leboucher, 'Multimodal Communication in the Human–Cat Relationship: A Pilot Study', *Animals*, vol. 13, no. 9, p. 1528, May 2023, doi: 10.3390/ani13091528.

[5] K. R. Vitale Shreve and M. A. R. Udell, 'Stress, security, and scent: The influence of chemical signals on the social lives of domestic cats and implications for applied settings', *Appl. Anim. Behav. Sci.*, vol. 187, pp. 69–76, Feb. 2017, doi: 10.1016/j.applanim.2016.11.011.

[6] J.-D. Vigne, J. Guilaine, K. Debue, L. Haye, and Gérard, 'Early taming of the cat in Cyprus', *Science*, vol. 304, no. 5668, Art. no. 5668, Apr. 2004, doi: 10.1126/science.1095335.

[7] J. W. S. Bradshaw, 'Sociality in cats: A comparative review', *J. Vet. Behav.*, vol. 11, pp. 113–124, Jan. 2016, doi: 10.1016/j.jveb.2015.09.004.

[8] C. L. Cameron-Beaumont, 'Visual and tactile communication in the Domestic cat (Felis silvestris catus) and undomesticated small felids', PhD Thesis, University of Southampton, 1997. [Online]. Available: https://eprints.soton.ac.uk/463206/

[9] S. Schötz, *The Secret Language of Cats*. in How to understand your cat for a better, happier relationship. Toronto: Hanover Square Press, 2018.

[10] J. Bradshaw and C. Cameron-Beaumont, 'The signalling repertoire of the domestic cat and its undomesticated relatives', in *Turner, D.C. and Bateson, P. (eds), The Domestic Cat: the Biology of its Behaviour*, Cambridge: Cambridge University Press, 2000.

[11] M. Moelk, 'Vocalizing in the House-Cat; A Phonetic and Functional Study', *Am. J. Psychol.*, vol. 57, pp. 184–205, 1944, doi: 10.2307/1416947.

[12] C. Tavernier, S. Ahmed, K. A. Houpt, and S. C. Yeon, 'Feline vocal communication', *J. Vet. Sci.*, vol. 21, no. 1, p. e18, 2020, doi: 10.4142/jvs.2020.21.e18.

[13] S. Schötz, 'Phonetic Variation in Cat–Human Communication', in *Pets as Sentinels, Forecasters and Promoters of Human Health*, A. Sousa and M. Pastorinho, Eds., Switzerland: Springer International Publishing AG, 2020, pp. 319–347.

[14] D. Burnham, C. Kitamura, and U. Vollmer-Conna, 'What's New, Pussycat? On Talking to Babies and Animals', *Science*, vol. 296, no. 5572, Art. no. 5572, May 2002, doi: 10.1126/science.1069587.

[15] C. De Mouzon, M. Gonthier, and G. Leboucher, 'Discrimination of cat-directed speech from human-directed speech in a population of indoor companion cats (Felis catus)', *Anim. Cogn.*, vol. 26, no. 2, pp. 611–619, Mar. 2023, doi: 10.1007/s10071-022-01674-w.

[16] C. De Mouzon, R. Di-Stasi, and G. Leboucher, 'Human perception of cats' communicative cues: human-cat communication goes multimodal', *Appl. Anim. Behav. Sci.*, vol. 270, p. 106137, Jan. 2024, doi: 10.1016/j.applanim.2023.106137.

[17] 'Insta360 Go 2 video camera'. [Online]. Available: https://www.insta360.com/product/insta360-go2/

[18] O. Friard and M. Gamba, '"BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations', *Methods Ecol. Evol.*, vol. 7, no. 11, pp. 1325–1330, Nov. 2016, doi: 10.1111/2041-210X.12584.

[19] L. A. Stanton, M. S. Sullivan, and J. M. Fazio, 'A standardized ethogram for the felidae: A tool for behavioral researchers', *Appl. Anim. Behav. Sci.*, vol. 173, pp. 3–16, Dec. 2015, doi: 10.1016/j.applanim.2015.04.001.

[20] B. Navarro Rivero, 'Cat-Human Interactions in a cat café: implications for health and welfare', MSc Degree project in Ethology, Dept. Of Biology Education, Stockholm University, 2021.

[21] A. L. Podberscek, J. K. Blackshaw, and A. W. Beattie, 'The behaviour of laboratory colony cats and their reactions to a familiar and unfamiliar person', *Appl. Anim. Behav. Sci.*, vol. 31, no. 1–2, pp. 119–130, Jul. 1991, doi: 10.1016/0168-1591(91)90159-U.

[22] M. Wedl *et al.*, 'Factors influencing the temporal patterns of dyadic behaviours and interactions between domestic cats and their owners', *Behav. Processes*, vol. 86, no. 1, pp. 58–67, Jan. 2011, doi: 10.1016/j.beproc.2010.09.001.

[23] J. M. Loberg and F. Lundmark, 'The effect of space on behaviour in large groups of domestic cats kept indoors', *Appl. Anim. Behav. Sci.*, vol. 182, pp. 23–29, Sep. 2016, doi: 10.1016/j.applanim.2016.05.030.

[24] C. Haywood, L. Ripari, J. Puzzo, R. Foreman-Worsley, and L. R. Finka, 'Providing Humans With Practical, Best Practice Handling Guidelines During Human-Cat Interactions Increases Cats' Affiliative Behaviour and Reduces Aggression and Signs of Conflict', *Front. Vet. Sci.*, vol. 8, p. 714143, Jul. 2021, doi: 10.3389/fvets.2021.714143.

[25] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2023.

[26] S. Schötz, J. van de Weijer, and R. Eklund, 'Phonetic Methods in Cat Vocalisation Studies: A report from the Meowsic project', in *Proceedings from Fonetik 2019 10-12 June 2019, Stockholm Sweden*, M. Heldner, Ed., 2019, p. 55. doi: 10.5281/zenodo.3245999.

[27] C. D. Mouzon, C. Gilbert, R. Di-Stasi, and G. Leboucher, 'How's my kitty? Acoustic parameters of cat-directed speech in human-cat interactions', *Behav. Processes*, vol. 203, p. 104755, Nov. 2022, doi: 10.1016/j.beproc.2022.104755.

[28] R Core Team, 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: http://www.R-project.org

[29] O. Nenadic and M. Greenacre, 'Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package', *J. Stat. Softw.*, vol. 20, no. 3, pp. 1–13, 2007.

[30] B. L. Deputte, E. Jumelet, C. Gilbert, and E. Titeux, 'Heads and Tails: An Analysis of Visual Signals in Cats, Felis catus', *Animals*, vol. 11, no. 9, p. 2752, Sep. 2021, doi: 10.3390/ani11092752.

[31] M. Amat, T. Camps, and X. Manteca, 'Stress in owned cats: behavioural changes and welfare implications', *J. Feline Med. Surg.*, vol. 18, no. 8, pp. 577–586, Aug. 2016, doi: 10.1177/1098612X15590867.

[32] L. R. Finka, 'Conspecific and Human Sociality in the Domestic Cat: Consideration of Proximate Mechanisms, Human Selection and Implications for Cat Welfare', *Animals*, vol. 12, no. 3, p. 298, Jan. 2022, doi: 10.3390/ani12030298.

On production mechanisms of group howling by *Canis lupus*: A case study

Axel G. Ekström[1], *Manon Delaunay*[2], *Linda Oña*[2]

[1]KTH Royal Institute of Technology, Sweden [2]Leipzig University, Germany

axeleks@kth.se

Abstract

We present early work on the production of howls by Hudson Bay wolves (*Canis lupus hudsonicus*). During vocalizations, jaw height appears mostly constant at a distinctly lowered position. We computed predicted first formants for a vocal tract length of appropriate size, with a flared oral cavity. Results are consistent with the grey wolves engaging in formant tuning, matching the fundamental frequency of phonation with the first resonant frequency of the vocal tract, amplifying the signal and increasing its loudness.

Index Terms: animal vocalization, phonetics, biomechanics, fundamental frequency, vocal tract

1. Introduction

In speech, the voice source from the vocal folds of the larynx is filtered by the shape of the supralaryngeal vocal tract, resulting in changes to its resonant properties or *formants,* F_n [1]. When the fundamental frequency (f_0) of a sound source (such as the human voice) is close to or matches the frequency of a formant, the perceived loudness of the sound can increase, making the sound seem more resonant and powerful. This phenomenon is referred to as formant tuning and is primarily known for its use in singing [2].

In nature, too, animals may engage in similar f_0–formant tuning to increase the loudness of a call or utterance. This behavior has been observed in other species, such as gibbons (*Hylobatidae* spp) [3] and Common marmosets (*Callithrix jacchus*) [4]. Here, we provide an interim report on work on the production mechanisms of affiliative howling in a Hudson Bay wolf (*Canis lupus hudsonicus*). We report on early results of a grey wolf articulator computational model, which provides support for the hypothesis that grey wolves may actively alter their jaw positions to tune F_1 to f_0.

2. Methods

2.1. Sample

The grey wolf (*Canis lupus*) is the largest extant canid species. The Hudson Bay wolf is one of over 30 extant subspecies of the genus, and is native to the tundra landscapes of the Queen Elizabeth Islands, northern Canada. Our data was collected by LO from a captive pack housed at Osnabrück, Germany during a period in January, 2023. Here, we sampled a single isolated howl as the focus of this case study.

2.2. Acoustics

f_0 was assessed manually using correlograms [5], a method based on waveform matching, known for its robustness to noise.

Figure 1: *Hudson Bay wolf (Canis l. hudsonicus) howling. Note that, the oral tract is visibly flared, as opposed to narrowed or rounded. In sustained howls, jaw height appears stable with little change throughout the utterance, suggesting non–randomness.*

This was done because the presence of multiple vocalizing individuals in the recordings, renders reliable estimation from automatic methods unrealistic.

2.3. Articulator model

To model vocal tracts, we used the *TubeN* software [6], based on [7], which computes vocal tract transfer functions based on the circuit theory established in [1] with wall losses by [8]. The mathematical bases of the program are described in [7].

2.3.1. Grey wolf vocal tract data

To our knowledge, there is no reported vocal tract length for any non–domestic dog (*Canis l. familiaris*) subspecies of grey wolf in the relevant literature. However, values reported for domestic dogs allow for a rough estimation. In particular, there is a near uniform correlation between the length of the skull and vocal tract length (r = .962), which is highly statistically significant at $p < .001$ [9]. The skull length for a Grey wolf has been measured at 23.6 cm [10], within the ranges reported for German shepherd specimens by [9]. These data allow for a rough estimate of a Grey wolf VTL at $\approx$ 22 cm; we assumed an otherwise linear tube at 4 cm^2. These length values may be overestimates as grey wolves are the largest extant canid, and Hudson Bay wolves are a medium-sized subspecies.

2.3.2. Oral tract flaring

Most mammals do not appear to move their tongue to affect formants [11]. However, in speech, F_1 is tied to jaw height [12]. Visual inspection of our howling strongly illustrates that howls

are produced with flared oral cavities (Figure 1), suggesting exploitation of a similar phenomenon. Namely, Flaring has the effect of shortening the effective length of the tract. Here, we estimated the effect of flared tubes according to the equation provided by Lindblom and colleagues in their work on the acoustics of spread lips and "notched" tubes [13]. In their framework, the effect of a "notched" segment can modeled as a shorter uniform segment, added to the length of the un–notched tube sequence.

Because attaining measurements of the length of the oral cavity flare from in–vivo vocalizing subjects is not feasible, we posited a "floor" at 3 cm, and a "ceiling" at 5 cm. According to computations by [13], a notch of 3 cm is approximated as a new segment of roughly 1.25 cm, added to the length of the "short" tube; a notch of 5 cm is approximated as a new segment of roughly 1.75 cm. The relationship is mostly consistent across segments of different diameter settings. Ultimately, it will be necessary to attain these measurements from the animals directly (i.e., by measuring the distance in cm from the labial commissures to the anteriormost portion of the face in a diseased specimen). Finally, as a control condition, we also computed F_1 for schwa for a vocal tract length of 22 cm.

3. Results

3.1. Fundamental frequency

For our selected howling utterance (approximately 3.36 *s*), we observed a largely consistent f_0 maintained throughout the utterance (M = 459 Hz, SD = 16.57 Hz).

3.2. Effect of flaring

Our computer models predicted an upward–shifting effect of flaring on F_1. For the "floor" (flare = 3 cm) F_1 = 436 Hz; for the "ceiling" (flare = 5 cm), F_1 = 471 Hz. Consistent with the hypothesis that howling involves tuning f_0 and F_1, the formant frequencies predicted by the flared models closely approximated the estimated f_0 – markedly more so than F_1 predicted for a schwa at vocal tract length = 22 cm (Figure 2).

4. Discussion

Much remains unexplored about how sounds are produced by animals. The present work contributes to this emergent picture by positing a framework capable both of reconstructing (or reverse engineering) animal vocalization resonance frequencies, and explaining them as factors of mammalian articulation. In this case study, we made several simplistic assumptions informing our vocal tract models. In the future iterations, letting real–life anatomical data inform our models would provide for more reliable results. Finally, howling is a stereotypically social behavior that typically engages several members of a pack. Our results, if verified, may indicate that a pack of howling grey wolves maintain territorial boundaries [14] by engaging in simultaneous formant tuning.

5. Acknowledgements

The results of this work will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017-00626).

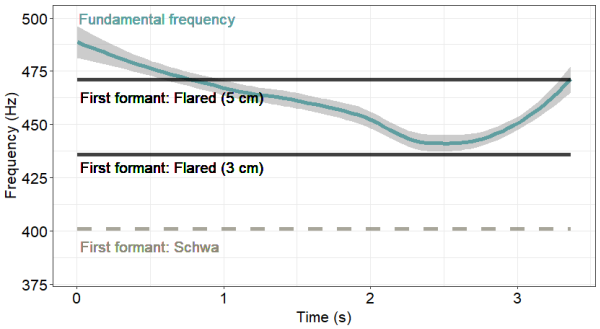

Figure 2: *Assuming a "flare" of 3 to 5 cm results in a computer model predicted F_1 which more closely approximates f_0 than that predicted for schwa. This is consistent with the hypothesis that grey wolves tune the resonant properties of the oral tract to the f_0.*

6. References

[1] G. Fant, *The acoustic theory of speech production*. The Hague: Mouton, 1960.

[2] J. Sundberg, *The science of the singing voice*. Northern Illinois University Press, 1987.

[3] H. Koda, T. Nishimura, I. T. Tokuda, C. Oyakawa, T. Nihonmatsu, and N. Masataka, "Soprano singing in gibbons," *American Journal of Physical Anthropology*, vol. 2, no. 149, pp. 347–355, 2012.

[4] H. Koda, I. T. Tokuda, M. Wakita, T. Ito, and T. Nishimura, "The source-filter theory of whistle-like calls in marmosets: Acoustic analysis and simulation of helium-modulated voices," *The Journal of the Acoustical Society of America*, vol. 6, no. 137, pp. 3068–3076, 2015.

[5] S. Granqvist and B. Hammarberg, "The correlogram: A visual display of periodicity." *The Journal of the Acoustical Society of America*, vol. 114, pp. 2934–2945, 2003.

[6] K. Zhang, R. Song, R. Tu, J. Edlund, J. Beskow, and A. G. Ekström, "Modeling, synthesis and 3D printing of tube vocal tract models with a codeless graphical user interface," in *Proceedings from FONETIK 2024*, Stockholm, Sweden, June 2024, pp. 155–160.

[7] J. Liljencrants and G. Fant, "Computer program for VT-resonance frequency calculations," *STL-QPSR*, pp. 15–21, 1975.

[8] G. Fant, "Vocal tract wall effects, losses, and resonance bandwidths," *STL–QPSR*, vol. 2, pp. 28–52, 1972.

[9] T. Riede and W. T. Fitch, "Vocal tract length and acoustics of vocalization in the domestic dog (canis familiaris)," *Journal of Experimental Biology*, vol. 202, pp. 2859–2867, 1999.

[10] B. van Valkenburgh, B. Pang, D. Bird, A. Curtis, K. K. Yee, C. J. Wysocki, and B. A. Craven, "Respiratory and olfactory turbinals in feliform and caniform carnivorans: The influence of snout length," *The Anatomical Record Advances in Integrative Anatomy and Evolutionary Biology*, vol. 297, pp. 2065–2079, 2014.

[11] W. T. Fitch, "The phonetic potential of nonhuman vocal tracts: comparative cineradiographic observations of vocalizing animals," *Phonetica*, vol. 2-4, no. 57, pp. 205—-218, 2000.

[12] B. E. Lindblom and J. E. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," *The Journal of the Acoustical Society of America*, vol. 4B, no. 50, pp. 1166–1179, 1971.

[13] B. Lindblom, J. Sundberg, P. Branderud, and H. Djamshidpey, "On the acoustics of spread lips," in *In Proceedings of Fonetik 2007*. Stockholm, Sweden: TMH-QPSR, 50, 2007, pp. 13–16.

[14] F. H. Harrington, "Aggressive howling in wolves," *Animal Behaviour*, vol. 35, pp. 7–12, 1987.

Deciphering Asian Elephant Rumble Calls to Classify Mahout and Social Interactions

Seema Lokhandwala[1], *Rohan Kumar Gupta*[1], *Priyankoo Sarmah*[1], *Rohit Sinha*[1]

[1]Indian Institute of Technology Guwahati, India

(seema176155001,rohan_kumar,priyankoo,rsinha)@iitg.ac.in

Abstract

This paper explored the possibility of classifying elephant vocalizations based on their contextual associations, with a specific focus on rumble calls. The study aimed to differentiate acoustic variation of rumble calls in two contexts: interactions with other elephants and interactions with human caretakers (mahouts). The data was collected through on-field work to gather context-specific elephant vocalization data. We developed a support vector machine-based classifier using both conventional and deep learning-based features. These deep learning-based features were generated using a speech encoder trained through a self-supervised learning method. The classifier achieved an accuracy of 66.6% for conventional features and 84.9% for deep learning-based features in subject-independent scenario. We observed that even with subject-dependent training and testing of the data, the approach utilizing deep learning-based features outperformed the conventional features.

Index Terms: animal communication, human-elephant interaction, rumbles

1. Introduction

The enduring relationship between mahouts and elephants has spanned centuries, characterized by a profound bond of trust and companionship. However in contemporary times, this dynamic connection has changed arising from shifting societal norms and conservation challenges. This evolving relationship mirrors the intricate interplay of cultural heritage, environmental pressures, and emotional intelligence essential for the care of these animals. Despite the declining tradition of mahoutship, it still holds a rich historical significance. Across the diverse regions of South and Southeast Asia, one can observe commonalities in training methods, command words, and management practices among various traditions [1]. Asian elephants *(Elephas maximus)* form intimate, social and working relations with their mahouts and can comprehend an impressive set of command words.

Asian elephants are social and geographically dispersed species [2]. Therefore, the ability to communicate acoustically over long and short distances is crucial for mating, cooperation and maintaining group cohesiveness [3, 4]. They produce a wide variety of calls, including low-frequency rumbles and growls, as well as high-frequency chirps, roars, trumpets, and barks, along with a range of imitation and combination calls [5, 6].

Studies have related the acoustic structure of elephant vocalizations to individual identity factors like sex, age, and emotional state. Stoeger *et al.* [7]. showcased that elephants can emit various call types, including rumbles, trumpets, and snorts, in reaction to verbal cues from trainers. They also observed that rumbles produced during interactions with conspecifics displayed a variation in acoustic structure, differing from rumbles elicited by trainer cues [7]. Furthermore, Lokhandwala *et al.* [8] demonstrated that Asian elephants' trumpet calls vary depending on whether they are interacting with mahouts vs conspecifics. However, the trumpet call was predominantly observed to be emitted during negative interactions between mahouts and elephants [8].

Drawing from the findings of the studies mentioned above, we can deduce that the structure of rumble call types varies depending on the context. These context-specific acoustic structures may indicate the emotional or motivational state of elephants. In this study, our objective is to explore the variation in rumble calls produced during handler interaction and those produced during interactions with conspecifics.

2. Methodology

2.1. Study subjects and Context-specific data recording

The Kaziranga National Park and Tiger Reserve (KNP) in Assam, India, is where the elephant vocalization and associated behaviour data was collected. For this study, 25 elephants of different ages were selected from KNP which are used for various activities such as tourism, patrolling and anti-poaching efforts.

Recording sessions were conducted across the field site, encompassing elephants' bathing, browsing, and nighttime tethering areas. A round-robin method was employed, with approximately 4 hours dedicated to monitoring each subject. Observations of behavior, lasting from 15 minutes to an hour, were recorded every 30 seconds. In this study, social interactions and handler interactions were categorized based on these distinct contexts:

Social Context: This behavior was noted during interactions between elephants, predominantly characterized by contact calls and communication among conspecifics.

Handler Interaction (HI) Context: This behavior was identified during interactions between elephants and their mahouts, the human caretakers responsible for their care and management. Mostly when mahouts were feeding the elephants or while tying chains on their feet. Additionally, rumbles were not observed during negative interactions with mahouts.

2.2. Experimental setup

Based on field notes, every acoustic recording underwent initial visual inspection using PRAAT 6.2.03 software [9]. Rumbles were subsequently identified in the raw data, and calls were

marked and trimmed from start to finish to extract pertinent data for subsequent analysis. In this study, the classifier was developed on both conventional features and deep-learning features.

To extract conventional features, we downsampled the rumbles to 600 Hz and used noise reduction low-pass filters with a cutoff of 200 Hz, a Hanning window, and 10 Hz smoothing with a pitch floor. The "To Formants (keep all)" function in PRAAT was then used to extract the formants of the rumbles. Subsequently, a stop Hann band (0–10 Hz) and pass Hann band (11–150 Hz) were employed for each rumble to target specific frequency ranges. We extracted 15 parameters out of 99 rumble vocalizations in both contexts. The conventional parameter set includes fundamental frequency parameters (mean F0, minimum F0, maximum F0, standard deviation F0), filter-related parameters (maximum and mean of first three formant locations), and temporal parameters (time to minimum F0, call duration, time to maximum F0).

For extracting deep-learning features, we downsampled the rumbles to 16000 Hz and employed the problem-agnostic speech encoder (PASE) proposed in [10]. The encoder comprises multiple convolutional layers and is trained on the LibriSpeech database [11] while solving multiple self-supervised tasks. The encoder takes raw speech as input and generates features of dimension 100. For this study, we derived a global features by applying the mean function over all generated features for a subject, which is henceforth denoted as PASE features in the subsequent text.

To assess the model's performance, we employed the 5-fold cross-validation methodology for subject-independent scenario. We utilized a Support Vector Machine (SVM) with a linear kernel to build the classification model due to its computational efficiency for both subject-dependent and subject-independent scenario.

3. Results and Discussion

The classification model was built using Python, where for each fold of the k-fold evaluation, a separate SVM model was trained with the respective training set. The performance of the classification model was evaluated using accuracy and F1-score metrics. The accuracy and F1-score from the 5-fold cross-validation were computed and presented in Table 1.

Table 1: *Classification performances for distinguishing between HI and Social behavior classes in terms of accuracy (%) and F1-score (%), using 5-fold cross-validation.*

	Conventional Features		PASE Features	
	Accuracy	F1-score	Accuracy	F1-score
Fold 1	85	85	90	91
Fold 2	60	60	80	80
Fold 3	60	63	80	80
Fold 4	65	62	85	85
Fold 5	63.1	63	89.5	89
Average	66.6	66.6	84.9	85

The SVM model for subject-independent scenario achieved an average accuracy of 66.6% and an average F1-score of 66.6% for conventional features, while for PASE features, it achieved an average accuracy of 84.9% and an average F1-score of 85%. To investigate the impact of subject-dependent scenario on classification we trained and tested a separate model, in which the conventional features resulted in an accuracy of 56.5%, whereas PASE features achieved an accuracy of 73.9%. It was noted that even with subject-dependent training and testing, the approach utilizing a PASE features outperformed the conventional features. In the conventional approach, we examined the features that are crucial for this classification by analyzing feature importance. The top five features identified were the mean locations of F3 and F1, the minimum location of F3, and the maximum and standard deviation of F0. In conclusion, our observations primarily highlight two key findings: firstly, the PASE features approach outperformed the conventional features approach in all classification task, and secondly, elephants produce rumbles that exhibit variations while interacting with mahouts compared to with conspecifics.

4. Acknowledgements

The authors acknowledge the Forest Department of Assam, India for permission of data collection and express gratitude to forest guards and elephant handlers for their support during fieldwork in Kaziranga National Park and Tiger Reserve.

5. References

[1] T. R. Trautmann, *Elephants and Kings: An Environmental History*. University of Chicago Press, 07 2015. [Online]. Available: https://doi.org/10.7208/chicago/9780226264530.001.0001

[2] T. N. Vidya and R. Sukumar, "Social organization of the Asian elephant *(Elephas maximus)* in southern india inferred from microsatellite DNA," *Journal of Ethology*, vol. 23, pp. 205–210, 8 2005.

[3] S. Nair, R. Balakrishnan, C. S. Seelamantula, and R. Sukumar, "Vocalizations of wild Asian elephants *(Elephas maximus)*: Structural classification and social context," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2768–2778, 2009. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3224717

[4] S. D. Silva, "Acoustic communication in the Asian elephant, *Elephas maximus maximus*," *Behaviour*, vol. 147, pp. 825–852, 2010.

[5] A. S. Stoeger, D. Mietchen, S. Oh, S. D. Silva, C. T. Herbst, S. Kwon, and W. T. Fitch, "An Asian elephant imitates human speech," *Current Biology*, vol. 22, pp. 2144–2148, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2012.09.022

[6] M. A. Pardo, J. H. Poole, A. S. Stoeger, P. H. Wrege, C. E. O'Connell-Rodwell, U. K. Padmalal, and S. de Silva, "Differences in combinatorial calls among the 3 elephant species cannot be explained by phylogeny," *Behavioral Ecology*, pp. 1–12, 2019.

[7] A. S. Stoeger and A. Baotic, "Operant control and call usage learning in African elephants," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1836, p. 20200254, 2021.

[8] S. Lokhandwala, P. Sarmah, and R. Sinha, "Classifying mahout and social interactions of Asian elephants based on trumpet calls," in *Speech and Computer*, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. Cham: Springer International Publishing, 2022, pp. 426–437.

[9] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[10] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 161–165.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

Toward integrating evolutionary models and field experiments on avian vocalization using trait representations based on generative models

Reiji Suzuki[1], *Zachary Harlow*[2], *Kazuhiro Nakadai*[3], *Takaya Arita*[1]

[1]Nagoya University, Japan
[2]University of California, Berkeley, USA
[3]Tokyo Institute of Technology, Japan
reiji@nagoya-u.jp

Abstract

We propose a novel agent-based evolutionary model for audible animal vocalizations based on generative models and discuss further possibilities of combining such models and their results with field experiments, exemplified by a case study of the Spotted Towhee. We constructed a sexual selection model in which male and female song genotypes are vectors in the latent space of the variational autoencoder of the focal species, and the spectrogram images generated from the vectors are regarded as songs and song preferences. The model results suggested that clear and moderately complex vocalizations tended to be selected. In addition, we conducted a preliminary playback experiment to investigate the effects of generated songs on wild birds in the field. We used a robot audition technique, HARKBird, to track the spatial patterns of response songs from resident birds to the playback. Experiments suggested that even generated sounds that are noisy to the human ear may have a salience to wild birds.

Index Terms: sound source localization, sexual selection, generative models, evolutionary models, robot audition, artificial life

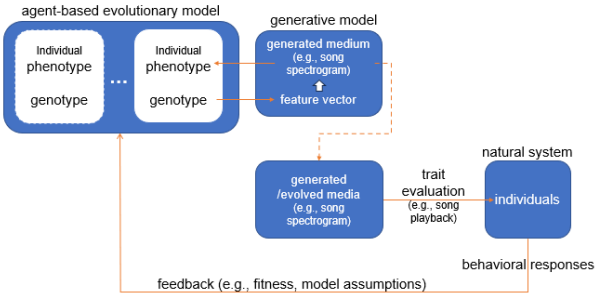

Figure 1: *Schematic of the relationship between the agent-based model, the generative model, and the natural system. Song genotypes in the evolutionary model are converted into song phenotypes (spectrograms) by the generative model. The biological relevance of song phenotypes can be evaluated in the natural environment through playback experiments. The bottom arrow indicates how future work could incorporate feedback from experiments in the natural system.*

1. Introduction

Agent-based modeling is a suitable computational method for studying the evolution and interactions of organisms and social groups. Models modify genotypes through mutation and genetic recombination and select phenotypes based on simple pre-defined rules of interaction and evolution.

Deep learning techniques [1] are contributing to computational bioacoustics, and generative models have recently been used in the study of animal communication and ecoacoustics for exploring spectrogram data by mapping these data onto low-dimensional latent spaces [2]. They are especially powerful for automatic sound identification and feature analysis. In particular, variational autoencoders (VAEs) have been used to generate complex nonlinear speech features that can be represented linearly [3], thoroughly analyze clustered animal vocalizations [4], and have been applied to attribute analyses of sounds generated beyond the scope of original data [5].

This paper proposes a novel agent-based evolutionary model for animal vocalizations based on generative models. It discusses further possibilities of combining such models and results with field experiments [6], and presents a proof of concept playback experiment on the Spotted Towhee (*Pipilo maculatus*). VAEs can define low-dimensional feature vectors that are suitable to be input as genes in agent-based evolutionary models (Fig. 1 (top)). Feature vectors in the generative model determine the complex phenotypes that are input to the agent-based model, in this case spectrograms.

In our proposed model, inspired by Higashi et al. [7], genes for male songs and female song preferences are coded as feature vectors in the latent space of a generative model, and spectrogram images represent male song traits and female preference traits. We use a generative model based on bird songs from the Spotted Towhee, and explore the properties of bird songs that were evolved based on sexual selection. We expect that the emerging properties of realistic sounds can bring insights into the significance of vocalizations under the assumed evolutionary context (e.g., sexual selection).

Bird songs are the perfect medium for considering direct or indirect interactions between agent-based evolutionary models and real-world environments because the evolved vocalizations can be presented to wild birds and their responses evaluated directly (Fig. 1 (bottom)). Acoustic interactions between animals and artificial agents have been studied [8], and there is increasing interest and discussions in realizing human-animal communication using generative AIs [9]. Linking agent-based evolutionary models with real-world environments via generative models opens new research directions by following hypotheses based on evolved phenotypes and constructing artificial systems that interact with natural systems in real time.

As an initial approach, we examine whether and how generated and audible sounds can affect the behavioral responses of wild birds. We present preliminary playback experiments in which Spotted Towhees received songs sampled at differ-

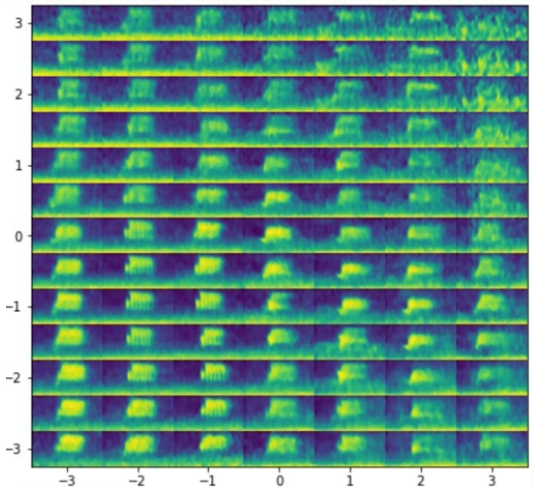

Figure 2: *Song distribution within the latent space of Spotted Towhee songs with feature vectors represented by spectrograms. Songs near the origin in the center are most similar to natural songs.*

ent distances from the origin of the latent space in the generative model. We used HARKBird, a bird song localization and analysis tool [10, 11], to extract fine-scale spatial patterns that might reflect subtle differences in the behavioral responses against generated sounds.

2. Creation of latent space of bird songs using a generative model

Field recordings of the Spotted Towhee were made at Blue Oak Ranch Reserve, on May 14th 2023, with an 8-channel microphone array (TAMAGO-03; System in Frontier, Inc.). Sound sources were localized and separated using HARKBird[1], a bird song recording, localization, and annotation software using a microphone array and the open-sourced robot audition software HARK [12] (see [10, 11] for detail). The process resulted in approximately 1000 2-second songs suitable for training data. Songs consist of a short introductory phrase followed by a trill phrase. The recordings were converted into 496 x 128 pixel gray-scale sound spectrogram images. A generative model from the recorded songs was constructed using a convolutional variational autoencoder (VAE). This model consisted of an encoder with eight convolutional layers, three fully connected layers that compressed information into two dimensions, and a decoder symmetric to the encoder network configuration. See Sainburg et al. [2] for more information about this type of VAE. A representative spectrogram was generated from the corresponding coordinate position in the 2-dimensional latent space and mapped onto the same coordinate system (Fig. 2).

Generated songs reproduced the main features of recorded songs with several variations and exhibited an increasingly noisy song structure further away from the origin of the latent space.

[1]https://sites.google.com/view/alcore-suzuki/home/harkbird

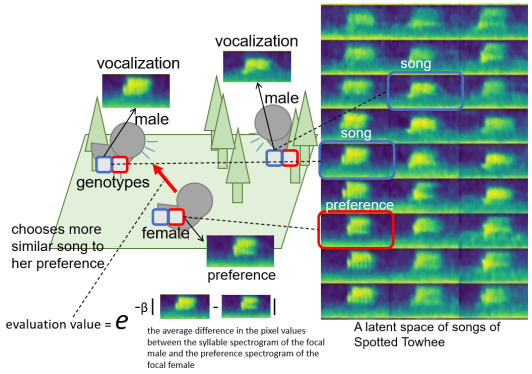

Figure 3: *Evolutionary model of male song genotypes and female preference genotypes. A female preference genotype (red) is compared with male song genotypes (blue) created from the latent space. Selection acts on spectrograms, stochastically choosing the most similar pairs for the next generation.*

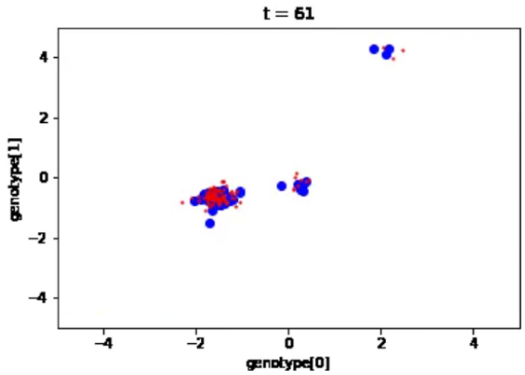

Figure 4: *An example of male gene segregation after 61 generations. Blue dots indicate male song genes and red dots indicate female preference genes.*

3. Evolutionary model of birdsongs and preferences using a generative model

The proposed model (Fig. 3) is inspired by the mathematical model of sympatric speciation by sexual selection in [7]. We consider a male and female population consisting of N individuals each. Each individual has two real-valued genes. Each gene represents a 2D vector (or position) in the latent space (a pair of (x, y) coordinates) (Fig. 2). One is a gene used to generate a song spectrogram vocalized by a male, and the other is a gene used to generate a female preference spectrogram. Both genes' x and y coordinates were generated randomly within the $[-W, W]$ range in the initial population.

In the model, we assume that females select males with a probability proportional to $exp^{-\beta \times x}$ where x is the average difference in pixel values between the male song spectrogram and the female's preference spectrogram, and β is a coefficient. Females select the male with the song stochasticlly closest to their preference spectrogram. One male and one female offspring are produced from the parental genes, including the effects of recombination and mutation. Recombination is modeled as BLX-α crossover [13] that occurs with probability p_c, which is a

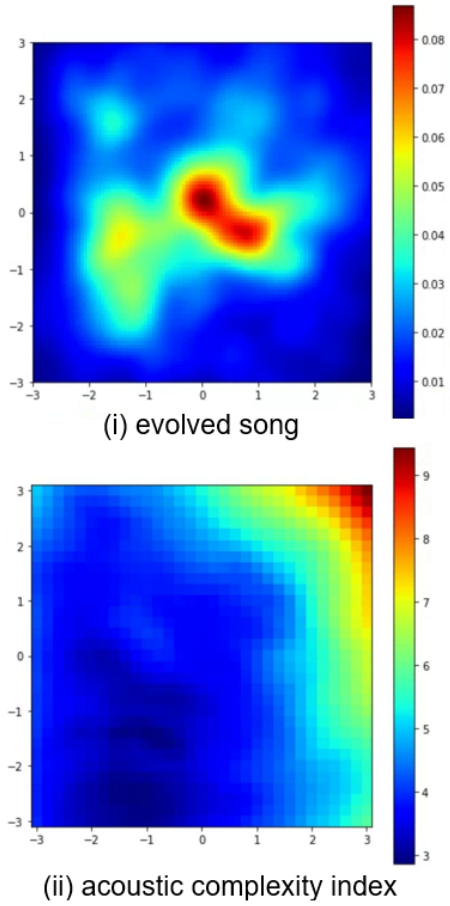

Figure 5: *Evolutionary experiments for the Spotted Towhee. Distribution across the latent space of both models for (i) evolved song genes showing the heavily selected genotypes in red and (ii) acoustic complexity index (ACI) across the latent space with more complexity shown in red. Image: @2024 Airbus, Maxar Technologies, Google.*

crossover method designed to produce offspring genes by combining the characteristics of real-valued genes of parents within a defined range. Mutation is modeled as a normal random value with mean 0 and standard deviation σ that occurs in each gene with probability p_m. Trials are conducted over T generations, with the value of each gene in the initial population determined randomly from $[-W, W]$.

4. Evolutionary experiments

We used the parameter settings as follows: N=100, W=5.0, β=0.3, α=1.1, T=100, p_c=0.5, p_m=0.15 and σ=0.2. The analysis of several trials showed that the song and preference genes tended to correlate and differentiate into several groups from the initial population in each trial (Fig. 4).

The genes that converged as a result of differentiation differed greatly from trial to trial. Therefore, additional experiments were conducted to extract the overall trend. Fig. 5 (i) shows the frequency distribution of the vocalization genes of the last generation of males in 2000 trials, using a kernel density estimation (KDE) distribution. Fig. 5 (ii) shows a measurement of the Acoustic Complexity Index (ACI) for the spectrograms in the latent space [14].

The vocalization gene distribution shows that the selected songs are generally distributed over a wide area around the origin. Compared to the spectrogram distribution (Fig. 2), the vocalizations tend to be distributed in the range where there is less noise and the vocalizations are generated relatively clearly. The acoustic complexity index (ACI) is a quantitative measure of the biological sound in a recording under the assumption of no environmental noise. ACI increases as the temporal variation of power increases across frequency bins. High ACI values were associated with noisy and unclear songs, while low ACI values indicated simple songs with very little frequency variation (Fig. 2). The model tended to avoid these areas of the latent space and selected songs of intermediate ACI, selecting songs with low noise and variable sound elements. We observed a similar tendency of the generated and evolved vocalizations for Blue-and-white Flycatcher (*Cyanoptila cyanomelana*) in Japan [15].

We tested a comparative model in which female ratings were determined by the difference in x and y coordinates between genes instead of the difference between spectrograms, and the population rapidly converged to the origin, resulting in a unimodal distribution centered at the origin in repeated experiments. This indicates an inherent selection pressure on the center of the latent space. Nevertheless, the complex distribution in Fig. 5 (i) indicates the influence of generated vocalizations and preferences on the evolutionary process.

5. Playback experiments using generated songs

We performed preliminary playback experiments to test whether birds would respond to songs generated from the generated model used in the evolutionary model. This experiment was approved by the Animal Care and Use Committee and the Graduate School of Informatics at Nagoya University, Japan (no. I230002-002). HARKBird was used to track the 2D locations of responding birds during the playback experiments using recordings with multiple microphone arrays (e.g., [16]).

Playback experiments were conducted on Spotted Towhees at Blue Oak Ranch Reserve located in Santa Clara County, California. The reserve comprises angled valleys and ridges of mixed oak woodlands and grasslands with a wide, flat valley supporting meadows and riparian forests. We selected three generated songs from different distances to the origin of the latent space, representing high, medium, and low levels of song structure. This pattern of well-defined song structure near the origin and increasingly noisy songs farther into the latent space is characteristic of the two-dimensional latent space generated by VAE (Fig. 2).

Experimental equipment consisted of a loudspeaker and two microphone arrays (TAMAGO-03, System in Frontier) situated 40 m apart along the territorial boundary of two male birds. Playbacks were conducted between 09:00 and 12:00 on May 17th, 2023, with a separation of at least 30 minutes between experiments. Recordings were captured in 20-minute files for analysis. Songs were detected and 2D-localized using HARKBird by triangulating the direction of arrival of sound sources estimated from each microphone array. Field notes were used to corroborate the reconstructed movement patterns.

All three playbacks provoked a response from the birds, i.e., vocalizations of multiple conspecific individuals around the

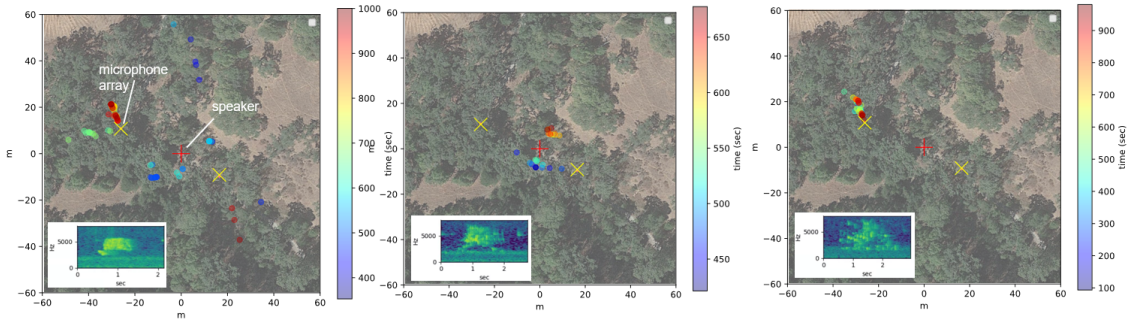

Figure 6: *Spotted Towhee playback experiments of evolved songs that had (a) high structure (b) medium structure and (c) low structure. The red cross indicates the location of the playback speaker and the yellow x's are the locations of the microphones. The color of the dots indicates the timing of localized songs.*

speaker according to the observations. The most substantial response came from the song generated at the origin position of the latent space (0, 0) (Fig. 6 (a)). This playback was closest to the structure of a wild bird with well-defined introductory and trill phases. This playback drew in singing territorial males from both adjacent territories. One came within 15m of the playback speaker, wandered around it, and then stayed about 40m to the northwest of the speaker for a long time. The other approached from the southeast and came within 40m of the speaker.

The medium-structure playback generated from (-8, 0) in the latent space drew in one individual to within 10m of the loudspeaker and sang from several locations in the area (Fig. 6 (b)). The introductory and trill segments of this playback were unseparated and had a similar duration and bandwidth to a natural sound. This playback was noisy but still elicited a strong response from this male.

Even the low-structure playback generated from (-20, 0) appeared to elicit a response from a territorial male (Fig. 6 (c)). This song had very little structure and resembled a harsh metallic noise. It had a similar duration to a natural song, but there was no separation between song elements. An individual with a territory on the west side (left side of the figure) came to the edge of the forest road near its boundary, sang while maintaining a position far from the loudspeaker, and left shortly after the playback ended. The individual then returned about a minute later and sang from a similar position, intermittently pausing. The response to this metallic noise-like playback suggests that biologically relevant information is present.

Playbacks that originated nearest the origin of the VAE latent space provoked the greatest response from wild birds, whose vocalizations often reflect the characteristics of the sound. However, even the distant, almost noisy vocalizations elicited a response.

We further conducted preliminary playback experiments at the same site in June 2024. We used several vocalizations that were frequently selected in the evolutionary experiments. We obtained some responses from the wild birds, but we still need further investigation to see whether and how the variations in their patterns can affect their responses.

6. Conclusion

This study presents a novel approach to avian communication research that uses agent-based evolutionary models and generative AI models to produce novel vocalizations that retain biologically relevant information. The experimental results suggested that clear and moderately complex songs tended to be selected by the model. A preliminary playback experiment presenting generated songs to wild birds showed that even noisy songs to the human ear were recognizable to the birds. Further research is needed to determine if evolved songs' noise differs from the noise of naturally degraded playbacks and how responses to regional dialects compare to evolved songs. Modeling the songs of species with complex song repertoires could go further to produce novel song types that are quite different from natural songs but also salient to the species.

As generative AI such as ChatGPT and Stable Diffusion permeate society, surprising people, making work more efficient, and entertaining them, it also presents challenges such as flooding society with generative content and unintentionally biasing people's behavior. Similar issues may arise at the interface between nature and AI society. At the same time, the thoughtful use of AI could create new points of contact between agent-based evolution models, artificial systems, nature, and ecology. In the future, we would like to consider these possibilities and continue our attempts to integrate evolutionary models and field experiments using robot audition techniques [11].

7. Acknowledgements

This study is supported in part by JSPS KAKENHI JP19KK0260, JP21K12058, JP20H00475, JP24K15103. We thank Hao Zhao's support for annotation of the data of Spotted Towhee.

8. References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, 10, e13152, 2022.

[2] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS Computational Biology*, e1008228, 2020.

[3] T. Sainburg and T. Q. Genter, "Toward a computational neuroethology of vocal communication: From bioacoustics to neurophysiology, emerging tools and future directions," *Frontiers in Behavioral Neuroscience*, 15, 811737, 2021.

[4] P. Best, S. Paris, H. Glotin, and R. Marxer, "Deep audio embeddings for vocalisation clustering," *PLoS ONE*, 18, e0283396, 2023.

[5] A. L. Gašper Beguš and S. Gero, "Approaching an unknown communication system by latent space exploration and causal inference," *arXiv-eprint*, 2303.10931, 2023.

[6] R. Suzuki, S. Sumitani, C. Ikeda, and T. Arita, "A modeling and experimental framework for understanding evolutionary and ecological roles of acoustic behavior using a generative model," in *Proceedings of ALIFE 2022: The 2022 Conference on Artificial Life (ALIFE2022)*, isal_a_00542, 2022.

[7] M. Higashi, G. Takimoto, and N. Yamamura, "Sympatric speciation by sexual selection," *Nature*, 402 (6761), 523–526, 1999.

[8] R. K. Moore, R. Marxer, and S. Thill, "Vocal Interactivity in-and-between Humans, Animals, and Robots," *Frontiers in Robotics and AI*, 3, 61, 2016.

[9] Y. Yovel and O. Rechavi, "AI and the doctor dolittle challenge," *Current Biology*, 33, R781–R802, 2023.

[10] R. Suzuki, S. Matsubayashi, R. W. Hedley, K. Nakadai, and H. Okuno, "HARKBird: Exploring acoustic interactions in bird communities using a microphone array," *Journal of Robotics and Mechatronics*, 27, 213–223, 2017.

[11] R. Suzuki, S. Sumitani, Z. Harlow, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "Extracting bird vocalizations from a complex natural soundscape in forests using robot audition techniques," in *Proceedings of 2023 IEEE/SICE International Symposium on System Integrations (SII2023)*, pp. 728–733, 2023.

[12] K. Nakadai and H. G. Okuno, "Robot audition and computational auditory scene analysis," *Advanced Intelligent Systems*, 2, 9, 2000050, 2020.

[13] L. J. Eshelman and J. D. Schaffer, "Real-coded genetic algorithms and interval-schemata," *Foundations of Genetic Algorithms*, vol. 2, pp. 187–202, 1993.

[14] N. Pieretti, A. Farina, and D. Morri, "A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI)," *Ecological Indicators*, 11, 868–873, 2011.

[15] R. Suzuki, R. F. an Zachary Harlow, K. Nakadai, and T. Arita, "An approach to integrating evolutionary models and field experiments on avian vocalization using trait representations based on generative models," in *Proceedigns of the 63th Workshop of the Special Interest Group on AI-challenge in Japanse Society of Artificial Intelligence*, SIG-Challenge-06, 2023.

[16] S. Sumitani, R. Suzuki, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "Fine-scale observations of spatio-spectro-temporal dynamics of bird vocalizations using robot audition techniques," *Remote Sensing in Ecology and Conservation*, 7, 18–35, 2020.

A single formant explicates the ubiquity of "meow"

Axel G. Ekström[1], *Laura Cros Vila*[1], *Susanne Schötz*[2], *Jens Edlund*[1]

[1]KTH Royal Institute of Technology, Sweden [2]Lund University, Sweden

axeleks@kth.se

Abstract

Across languages, the species–typical vocalization by domestic cats (*Felis catus silvestris*) is transcribed similarly, typically corresponding to [miau:] or [waʊ:]. Such consistent and ubiquitous cross-linguistic transcription is apparently onomatopoetic. However, in humans, these qualities make unique use of the tongue; in comparison, most nonhuman mammals do not appear to employ their tongues while vocalizing. The purpose of this work was to explore whether tube models modeled after the buccolabial oral tract morphology of the domestic cat, may be used to reverse engineer the apparent diphthong-like quality typically perceived in cat meows (the "au" in meow). For cats specifically, the short vocal tract is likely a causal factor, as the contribution of higher formants to vowel quality in the front–to–back dimension is significantly reduced. Results of computational models and perception tests suggest that the shift in apparent vowel quality may be driven by F1, corresponding in our model to raising of the mandible.

Index Terms: animal vocalization, vowel quality, vocal tract, speech acoustics, source/filter theory

1. Introduction

Across languages, the domestic cat (*Felis catus sylvestris*) "meow" is transcribed with remarkable ubiquity. In languages as diverse as Japanese, Welsh, and Mi'kmaq (Tab. 1), descriptions of the utterance denote an apparent transition corresponding to a consonant–vowel–vowel, consonant–vowel–consonant, or consonant–vowel–vowel–consonant cycle, where the first and last would-be consonants (and vowels) indicate (partial or complete) mouth closure and nasalized phonation ([m n]), and the sequence of would-be vowel phonemes indicates diphthong–like changes in the open–to–close and front–to–back dimension of phonetic space ([eo], [au]). Such ubiquity is suggestive of onomatopoetic properties. However, while the study of animal vocalizations occupies a small yet fruitful niche in phonetic sciences, remarkably little such research is occupied with making direct parallels with human speech production.

There are several works concerned with the phonetics of cat vocalizations. Most focus on pitch or fundamental frequency (f_0) contours employed in vocalizations [1, 2, 3, 4, 5, 6, 7] though studies of call production [8] and apparent vowel-like quality in calls [9] have also been performed. However, these works have rarely considered the nature of the feline vocal tract (i.e., the "filter"). In contrast with humans, who possess relatively long pharynges, flat faces and tongues partially descended into the pharynx, most nonhuman mammals including cats have a "primitive" vocal tract configuration, with a high larynx, short–and–narrow pharynx, and a tongue contained primarily in the oral cavity [10], constraining vowel production capacities [11, 12, 13, 14]. Even so, however, most mammals do not appear to move the tongue to affect formant frequencies like humans do in speech [13]. Finally, the cat vocal tract is markedly shorter than those of adult humans [15]. Given that shorter vocal tracts produce higher formant frequencies, which contribute less to vowel quality than lower ones [16], it is unlikelier still that the apparent diphthongs transcribed from cat vocalizations reflect human–like articulatory postures. This combination of observations prompts an intriguing question; namely, in the absence of human speech production biomechanics, how are these vowel–like qualities produced?

Table 1: *Transcription of the standard vocalization by domestic cats in 10 unrelated languages.*

Language	Transcription
Afrikaans	*miaau*
Basque	*mau*
Bengali	*miu*
Finish	*miau*
Hebrew	*miàw*
Japanese	*nyān*
Korean	*yaong*
Mi'kmaq	*mia'wj*
Vietnamese	*meo*
Welsh	*miaw*

A view from comparative anatomy and speech acoustics offers a proposal, with the shape of the cat buccolabial oral tract (SVT_{BL}) as tentative explanatory variable. While human faces are flat, with an acute angle from the anterior cranial fossa (which houses the frontal lobes) to the anteriormost section of the oral tract, non-human mammals have "long" faces, with potentially significant phonetic consequences. Classic phonetics research demonstrates the impact on formats of mouth closure [16, 17, 18]. In particular, the first formant (F1), typically exhibits a distinct negative relationship with jaw height; as the jaw is raised, F1 tends to decrease. The short SVTs of cats [15] thus imply that higher formats play a less substantive role in determining the perceptual vowel–like quality of calls, and that F1 is the sole or primary determinant. The combination of relevant articulatory factors – short supralaryngeal vocal tract (SVT), long face, and an preestablished jaw-F1 contingency – present an alternative model of articulation: **H1**: *for domestic cats, apparent vowel–like quality may be driven by F1, resulting from changes in jaw height*. In this paper, we explore empirical support for this view on the basis of computational modeling with reference to cat physiology.

Figure 1: *Like most mammals, domestic cats are prognathic, with characteristically long faces, with labial commissures are positioned on either side of the pointed face. The subject (a 2-year-old female Neva Masquerade domestic cat) is sleeping. The owner is holding the fur back from around the labial commissures to increase visibility. The animal is not held or restrained against her will. Photo credit @ Axel G. Ekström.*

2. Methods

2.1. Computing vocal tract area transfer functions

We sought to predict plausible vocal tract transfer functions for open and narrowed jaw states. To predict formants we used a custom software [19] – based on [20] – which computes vocal tract transfer functions based on the circuit theory established by [16], with wall losses by [21]. The same computational approach has previously been used to successfully model human vocal tract transfer functions [22]. Mathematical bases of the program are not described further here; interested readers are referred to [20, 23].

2.2. Cat model

2.2.1. Cat vocal tract data

We sought to implement realistic vocal tract lengths for our models. [15] report a total vocal tract length (VTL) for domestic cats at approximately 8 cm. The measurement reported is "glottis to lips". It is not clear if, by "lips", the intended landmark is the anteriormost portion of the oral tract, or the labial commissures (where inferior and superior labia meet at the corner of the mouth). In humans, the impact of either alternative on total VTL is relatively minor [24], resulting from humans' uniquely orthognathic (flat) faces [25]. Cats, like most mammals have long faces, reflecting an evolutionary selection pressure for olfaction [26]. We make the assumption that, in apparent continuity with phonetics research [15] intended the former alternative, and that the domestic cats VTL are $\approx$ 8 cm. We assumed an otherwise linear tube at 2 cm^2. There are no indications of an abrupt discontinuity in the cat vocal tract [15, 26], where articulatory changes have stark and disproportionate effects on acoustic outcomes [27, 28]. Thus, while simplistic, this approach likely does not undersample sensitive regions in the domestic cat vocal tract.

2.2.2. Cat buccolabial oral tracts

In order to ascertain the length of "flare potential" (i.e., the length of the vocal tract section that may be spread as the animal lowers its jaw) of the domestic cat oral cavity, we measured the outwardly observable length of the lips back-to-front in five adult cats (Fig. 1, Tab. 2). Measurements of SVT_{BL} were taken unobtrusively, by holding a ruler from the right-hand side labial commissure of the animal while sleeping/resting. Because cat craniofacial morphology is anteriorly narrow (i.e., pointed), simply applying the ruler against the skin was likely to artificially inflate values. Therefore, we positioned the ruler against the masseteric fossa, on the lateral surface of the ramus of the mandible, and measurements taken as a factor of the ruler pointed straight along the protrusion of the face. We performed this exercise on four domestic short-hairs and one Neva Masquerade, which while a pedigree breed is not known for any distinguishing craniofacial features that would appear to impact our measurements (unlike e.g., Persian cats, which are characterized by comparatively flat faces and pinched nares). These values are likely conservative, as labial commissures may be pulled back further during jaw lowering. Values for a narrowed mouth opening were estimated as the distance from behind the mandibular incisors to the front of the lips. Measurements were rounded to the nearest half-integer (in cm).

Table 2: *Breed, sex, heights at the withers, and length of buccolabial vocal tract length (SVT_{BL}) in five domestic cats. Our sample included four house cats (Domestic Shorthairs, DSH) and one Neva Masquerade (NEM). All sampled individuals were adults. Values in cm.*

Breed	Sex	Height at withers (cm)	SVT_{BL} (cm)
DSH	F	28	2.8
DSH	F	29	2.8
NEM	F	29	3.1
DSH	M	30	3.3
DSH	M	34	3.2

2.2.3. Flared states

[8] estimated tube flaring for cat vocalizations in the shape of a Bessel horn, based on [29]. However, we argue that a mid-sagittal view of a vocalizing or yawning cat (i.e., a cat with lowered jaw position) is more consistent with a deep and angular "notch", rather than the flare of a Bessel horn. For this reason, we instead applied a method for estimating the effect of "notched" tubes according to the equation provided by [24] in their work on the acoustics of spread lips. Effectively, a tube model including a notched segment can be conceptualized as "long" (the length of the whole sequence) and "short" (the length of the unnotched section); the notched section can be algorithmically replaced with a different uniform segment, added to the length of the "short" segment. According to this work, a notch of 3 cm is approximated as a new segment of approximately 1.25 cm, added to the length of the "short" (i.e., unnoteched) tube. In the experimental settings explored by [24], this is largely consistent across diameter settings.

2.2.4. Narrowed states

A constriction was assumed at the anteriormost portion of the oral tract, at 1 cm x .2 cm^2. We do not claim these models are anatomically accurate models. Our approach was strictly computational; we sought to investigate if a diphthong–like quality like that indicated by universal transcriptions of domestic cat vocalizations, could be emulated through a model moving from open to narrow mouth closure.

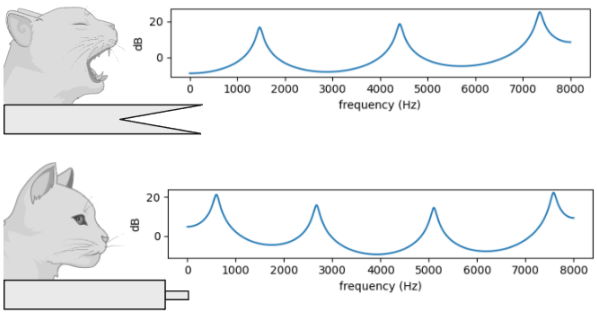

Figure 2: *Predicted vocal tract response curves for flared (top) and narrow (bottom) states, modeled after domestic cat vocal tract data [15] and SVT_{BL} data. Transfer functions from [19]. Image created with BioRender.*

2.3. Listening experiments

We sought to test our models in a pilot perception test. The purpose of this exercise was to determine whether an [au]–like diphthong could be synthesized from an articulator model based on cat jaw movements. Assumptions implemented here will be expanded upon in future work.

2.3.1. Stimuli

We sought to test the hypothesis that jaw movements are a primary articulator and origin of vowel–like quality in domestic cat meows. We synthesized diphthongs with [30], with initial states corresponding to lowered jaw positions or flared tubes, and final states corresponding to raised jaw positions or narrowed tubes).[1] We tested two setups. In the first, we synthesized only F1. In the second, we synthesized F1–F3. In order to ascertain human listeners' perception of the synthesized diphthongs, we programmed a set of two experiments. All subjects participated in both experiments, one after the other. The purpose of the two experiments was to identify the strongest candidate cause of the perceived "eow" in meow. We tested participants' perception of formant changes with constant f_0 – i.e., "steady-state" vowels. [6] report various f_0 curves for cat vocalizations across both individuals and contexts. These calls may be more or less consistent with a "prototypical meow"; here, we are concerned with vowel–like quality, and not with replicating natural meows per se. For this reason, diphthongs were synthesized with with f_0 = 110 Hz, 220 Hz, 330 Hz, 440 Hz, descending to –5 Hz. In addition, all sounds were synthesized as 1 second and .5 seconds in length, for a total of 64 trials (2 synthetic formant trajectories x 4 f_0 conditions x 2 length conditions, and 6 foil trajectories x 4 f_0 conditions x 2 length conditions).

2.3.2. Interpretation

Because transcriptions are often dissimilar between subjects, even when transcribers are native speakers of the same language, we interpreted listener data liberally, according to whether input indicated perceived phonetic closure (e.g., "auw" and "eoo") were both interpreted as indicating closure in the open–to–close dimension.

[1] An example sound file is available here: http://sndup.net/w9wjq

Table 3: *Predicted peaks for uniform, narrowed, and flared SVT_{BL}. Values are computed for a VTL = 8cm, with SVT_{BL} at 3 cm. Values in Hz.*

SVT_{BL}	F1	F2	F3
Schwa	1103	3309	5516
Flared	1305	3922	6537
Narrowed	602	2675	5105

Table 4: *Rater (N=7) outcomes for synthesized diphthong–like qualities, predicted by open–to–closed articulator models. Sounds were synthesized from F1, and from F1–F3.*

Model	Length	[au]–like
F1	.5 *s*	75%
	1 *s*	92.86%
F1–F3	.5 *s*	92.86%
	1 *s*	85.71%

2.3.3. Participants

Experiments were programmed in the Cognition online behavioral research platform (cognition.run). Participants were instructed to write what they heard. To avoid biasing results in favor of our "jaw–based" hypothesis, participants were not explicitly informed that presented stimuli would be "vowels", "vowel–like", "diphthongs", or "diphthong–like". However, they were informed that the sounds would not be words; this done to avoid biasing perception experiment results against our hypothesis, by encouraging generous interpretations of the sounds.

3. Results

3.1. Acoustic models

In the "flared" condition (low jaw position), all formants were shifted up from schwa (Fig. 2, Tab. 3). In the "narrow" condition (high jaw position), formants were shifted down. The change to F1 differed from schwa. Changes to formants predicted for the two conditions were (from notched–low jaw to narrow–high jaw) –703 Hz for F1, –1247 Hz for F2, and –1432 Hz for F3. Results are consistent with the principle that closure induces a decrease in F1 [16, 17].

3.2. Perception

3.2.1. Participants

In total, 7 participants (5 female) aged 26–55 (M = 34.43, SD = 12.44) completed the experiment. No participants reported any significant hearing difficulties.

3.2.2. Results

Results indicate that listeners typically perceived open–to–close diphthong–like sounds as "au" (or [au]–like) (Tab. 4). We did not observe any marked difference between .5 *s* and 1 *s* utterances.

4. Discussion

4.1. Origin of the meow

Consistent with **H1**, our data suggests that the would–be vowel quality in cat meows may be contingent on jaw movements, with the apparent quality to listeners being mainly driven by F1, and provides an explanation for the ubiquity of the "meow" form observed across languages of the world (Fig. 3). In particular, this exercise aimed to determine whether vowel–like contrasts reasonably transcribed similarly to the domestic cat "meow" form could be predicted through movements by a computational tube vocal tract model inspired by cat jaw movements. Based on the early data, we propose a programmatic sketch. Namely, in the course of articulating a "prototypical meow":

1. a transition from closed to open jaw facilitates the perception of an [m]–like utterance, as airflow is partially diverted from the nasal tract to the oral tract;
2. the velum may be closed off during sustained loud calls [13], facilitating non–nasalized vowel–like qualities;
3. the high frequency of higher formants means their contribution to quality is relatively minor.
4. a low jaw position may facilitate an [a]–like quality, as F1 is shifted up with decreasing jaw height;
5. (a high jaw position with a narrow anteriormost opening facilitates the perception of an apparently [u]–like quality, as F1 is shifted down;
6. a narrowing mouth opening achieved through raising of the jaw facilitates the impression of an open–to–close change.

4.2. Vowel–like qualities with short vocal tracts

The last few decades have seen a substantial increase in bioacoustics research – but much remains unexplored about how sounds are produced by living animals. The current state of research is not consistent with an ability in non–human animals to produce the range of human vowels, or produce them in the same way as human speakers [14]. In particular, much of the relevant research has been concerned with primate vocalizations [11, 14, 31]; the calls of other species have been subjected to comparatively little scrutiny [9, 13, 32].

The present work contributes to this emergent picture by positing a framework capable both of reconstructing (or reverse engineering) animal vocalization resonance frequencies, and explaining them as factors of mammalian articulation. [9] has noted that apparently back vowel-like vocalizations are seemingly produced with raised jaw, while apparently open and front vowel-like portions are produced with lowered mandible. Such differences may be relatively negligible in human vocalizations [24, 18]. Our finding that the vowel–like quality in cat vocalizations are heavily driven by F1 alone may have important implications for understanding the apparently vowel–like quality of other small animals, including dogs and monkeys.

4.3. Limitations

4.3.1. On acoustics of nasal tracts

Our models assumed a sequence of tube segments where the only changes to the shape of the resonator were at its anteriormost section. In speech, humans readily close off the velum and nasal passages, redirecting airflow from the nasal tract to the oral tract and facilitating non–nasalized sounds. In comparison, the vocal tracts of non–human animals may be less readily capable of closing off the velum [10, 11]. Notably, however, [13] reports that a variety of animals may do so, at least during effortful loud calls. Whether cat meows are one such call is not known. If not, however, there is a likely interference of the resonance of the nasal tract on vocal tract response curves [33]. To our knowledge, there are currently no descriptions of feline nasal tracts that permit acoustic modeling – though [15, 26] provide relevant data. Methods designed for estimating the effects of the nasal tract on speech behavior [33] may be useful for this purpose.

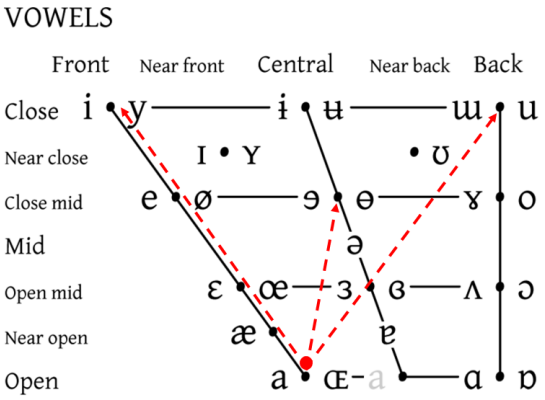

Figure 3: *Moving the jaw may facilitate an shift in perceived vowel–like quality, driven by changes to F1. The auditory trajectory is superimposed on the International Phonetics Alphabet vowel chart. Note that the overlap is illustrative only; we do not suggest that cats produce phonemes, nor produce phoneme–like sounds in the same manner, or for the same purpose, as humans.*

4.3.2. Fundamental frequency

In our synthetic stimuli, we constrained f_0. [6] report a range of distinct f_0 contours for cat vocalizations across a range of behavioral contexts. It is conceivable that as a call produced in a given context (e.g., in front of a closed door) differs significantly from that produced in a different one (e.g., when contained in a carrier), these differences may ultimately correspond to disparate articulatory innervations, and produced with an intended purpose or to achieve a given outcome (i.e., the door being opened, or being let out of the carrier). Future endeavors may be designed to reverse engineer calls holistically and seek to uncover the disparate articulatory and behavioral underpinnings of contextually distinct call properties.

4.4. Future directions

In this work, we have explored several tentative assumptions for modeling cat meows, and assessing the validity of models. We intend to base future iterations directly on a comprehensive description of *Felis c. sylvestris* vocal anatomy. We also intend to investigate a larger range of variable synthetic sounds, and evaluate their validity using a greater and more diverse sample of listeners.

5. Acknowledgements

The results of this work will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017-00626).

6. References

[1] M. Moelk, "Vocalizing in the house-cat; a phonetic and functional study," *The American Journal of Psychology*, vol. 2, no. 47, pp. 184–205, 1944.

[2] K. A. Brown, J. S. Buchwald, J. R. Johnson, and D. J. Mikolich, "Vocalization in the cat and kitten," *Developmental Psychobiology*, vol. 2, no. 11, pp. 559–570, 1978.

[3] G. R. Farley, S. M. Barlow, R. Netsell, and J. V. Chmelka, "Vocalizations in the cat: behavioral methodology and spectrographic analysis," *Experimental Brain Research*, no. 89, pp. 333–340, 1992.

[4] N. Nicastro, "Perceptual and acoustic evidence for species-level differences in meow vocalizations by domestic cats (Felis catus) and african wild cats (Felis silvestris lybica)," *Journal of Comparative Psychology*, vol. 118, p. 287–296, 2004.

[5] S. Schötz, J. van de Weijer, and R. Eklund, "Phonetic characteristics of domestic cat vocalizations," in *Proceedings of the 1st International Workshop on Vocal Interactivity in-and between Humans, Animals and Robots (VIHAR) 2017*, Skövde, Sweden, 2017, pp. 5–6.

[6] S. Schötz, J. van de Weijer, and R. Eklund, "Context effects on duration, fundamental frequency, and intonation in human-directed domestic cat meows," *Applied Animal Behaviour Science*, vol. 270, p. 106146, 2024.

[7] M. A. Schnaider, M. A. Heidemann, A. H. P. Silva, C. A. Taconeli, and C. F. M. Molento, "Cat vocalization in aversive and pleasant situations," *Journal of Veterinary Behavior*, no. 55, pp. 71–78, 2022.

[8] C. Shipley, E. C. Carterette, and J. S. Buchwald, "The effects of articulation on the acoustical structure of feline vocalizations," *The Journal of the Acoustical Society of America*, vol. 89, pp. 902–909, 1991.

[9] S. Schötz, "Phonetic variation in cat–human communication," in *Pets as Sentinels, Forecasters and Promoters of Human Health*. Springer, 2020, pp. 319–347.

[10] V. E. Negus, *Comparative anatomy and physiology of the larynx*. Heinemann, 1949.

[11] P. Lieberman, *Uniquely human: The evolution of speech, thought, and selfless behavior*. Harvard University Press, 1991.

[12] R. Carré, B. Lindblom, and P. F. MacNeilage, "Rôle de l'acoustique dans l'évolution du conduit vocal humain," *Comptes Rendus de l'Académie des Sciences, Série IIb*, vol. 320, p. 471–476, 1995.

[13] W. T. Fitch, "The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals," *Phonetica*, vol. 57, pp. 205–218, 2000.

[14] A. G. Ekström, "Correcting the record: Phonetic potential of primate vocal tracts and the legacy of Philip Lieberman (1934–2022)," *American Journal of Primatology*, p. e23637, 2024.

[15] G. E. Weissengruber, G. Forstenpointner, G. Peters, A. Kübber-Heiss, and W. T. Fitch, "Hyoid apparatus and pharynx in the lion (Panthera leo), jaguar (Panthera onca), tiger (Panthera tigris), cheetah (Acinonyx jubatus) and domestic cat (Felis silvestris f. catus)," *Journal of Anatomy*, vol. 201, pp. 195–209, 2002.

[16] G. Fant, *The acoustic theory of speech production*. The Hague: Mouton, 1960.

[17] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements." *The Journal of the Acoustical Society of America*, vol. 41, pp. 1002–1010, 1967.

[18] B. E. Lindblom and J. E. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," *The Journal of the Acoustical Society of America*, vol. 4B, no. 50, pp. 1166–1179, 1971.

[19] K. Zhang, R. Song, R. Tu, J. Edlund, J. Beskow, and A. G. Ekström, "Modeling, synthesis and 3D printing of tube vocal tract models with a codeless graphical user interface," in *Proceedings from FONETIK 2024*, Stockholm, Sweden, June 2024, pp. 155–160.

[20] J. Liljencrants and G. Fant, "Computer program for VT-resonance frequency calculations," *STL-QPSR*, pp. 15–21, 1975.

[21] G. Fant, "Vocal tract wall effects, losses, and resonance bandwidths," *STL–QPSR*, vol. 2, pp. 28–52, 1972.

[22] J. Sundberg, B. Lindblom, and A. M. Hefele, "Voice source, formant frequencies and vocal tract shape in overtone singing. a case study." *Logopedics Phoniatrics Vocology*, vol. 48, pp. 75–87, 2023.

[23] J. Sundberg, B. Lindblom, and J. Liljencrants, "Formant frequency estimates for abruptly changing area functions: A comparison between calculations and measurements." *The Journal of the Acoustical Society of America*, vol. 91, pp. 3478–3482, 1992.

[24] B. Lindblom, J. Sundberg, P. Branderud, and H. Djamshidpey, "On the acoustics of spread lips," in *In Proceedings of Fonetik 2007*. Stockholm, Sweden: TMH-QPSR, 50, 2007, pp. 13–16.

[25] D. E. Lieberman, *The evolution of the human head*. Harvard Belknap Press, 2011.

[26] B. van Valkenburgh, B. Pang, D. Bird, A. Curtis, K. K. Yee, C. J. Wysocki, and B. A. Craven, "Respiratory and olfactory turbinals in feliform and caniform carnivorans: The influence of snout length," *The Anatomical Record*, vol. 297, pp. 2065–2079, 2014.

[27] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.

[28] R. Carré, P. Divenyi, and M. Mrayati, *Speech: A dynamic process*. De Gruyter, 2017.

[29] A. H. Benade, *Fundamentals of musical acoustics*. New York: Oxford University Press, 1976.

[30] S. Barreda, *phonTools: Functions for phonetics in R.*, 2015, r package version 0.2-2.1.

[31] S. Grawunder, L. S. N. Uomini, T. Bortolato, C. Girard-Buttoz, R. M. Wittig, and C. Crockford, "Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage," *Philosophical Transactions of the Royal Society B*, vol. 377, p. 20200455, 2022.

[32] A. G. McElligott, M. Birrer, and E. Vannoni, "Retraction of the mobile descended larynx during groaning enables fallow bucks (Dama dama) to lower their formant frequencies," *Journal of Zoology*, vol. 270, pp. 340–345, 2006.

[33] M. Havel, J. Sundberg, L. Traser, M. Burdumy, and M. Echternach, "Effects of nasalization on vocal tract response curve." *Journal of Voice*, vol. 37, pp. 339–347, 2023.

Towards a Universal Method for Meaningful Signal Detection

Louis Mahon

University of Edinburgh, UK

lmahon@ed.ac.uk

Abstract

It is known that human speech and certain animal vocalizations can convey meaningful content because we can decipher the content that a given utterance does convey. This paper explores an alternative approach to determining whether a signal is meaningful, one that analyzes only the signal itself and is independent of what the conveyed meaning might be. We devise a method that takes a waveform as input and outputs a score indicating its degree of 'meaningfulness'. We cluster contiguous portions of the input to minimize the total description length, and then take the length of the code of the assigned cluster labels as meaningfulness score. We evaluate our method empirically, against several baselines, and show that it is the only one to give a high score to human speech in various languages and with various speakers, a moderate score to animal vocalizations from birds and orcas, and a low score to ambient noise from various sources.

Index Terms: speech processing, animal vocalizations, complexity, meaningfulness

1. Introduction

Humans are highly proficient at auditory pattern recognition, at least for certain subsets of audio signals, such as those of human speech. We are are able to interpret a wide variety of meanings across a wide variety of waveforms. Similarly, when we regard an animal vocalizations as meaningful, it is by finding the behaviour or information that we believe it signals. However, even independently of interpreting what a given signal means, humans also possess some ability to detect whether a signal is meaningful. For example, if one hears speech in a language they do not understand, or certain animal vocalizations, they may still get a sense that this is the sort of signal that could convey meaning. We do not feel the same if we hear other types of signal, such as ambient noise or white noise on the radio. Some signals, and data more generally, exhibit a systematic structure that suggests the potential to convey a meaning. Humans have a degree of intuition for recognizing this sort of structure. The goal of this paper is to make first steps towards articulating what it is, and how we might measure it automatically.

There exist several classic approaches to measuring complexity. Kolmogorov complexity takes the length of the shortest program that generates the given data. It is uncomputable, but there are computable approximations, such as file-compression ratio under some compression algorithm. A similar approach is found in the minimum description length principle [1] (MDL), though this is mostly concerned with fitting models. Entropy is closely related to MDL, and is often used as a measure of complexity for data of various sorts [2, 3]. One problem that arises with these methods, and their many derivatives, is that they give a low score to simple, highly regular data, and a high score to random, noisy data, with the data we consider most meaningful, such as human speech, falling somewhere between the two. Thus, we can say neither that a high score nor a low score indicates the data is meaningful. This fundamental issue has been identified by various authors [4, 5, 6, 7, 8].

The basis for our method is that length of the smallest representation of a piece of data indicates how complex it is. However, unlike prior methods, either algorithmic such as Kolmogorov complexity, or statistical such as entropy, we make a division of the description into a 'meaningful' and 'meaningless' portion. This is similar, on a high level, to some theoretical work to divide the Kolmogorov complexity into meaningful information and noise, using, e.g., 'sophistication' [8, 6] or 'effective complexity' [5, 7]. When selecting the shortest overall description of the data, we include both the meaningful and the meaningless portion, but, after making this selection, we ignore the meaningless portion and take only the length of the meaningful portion as contributing to the complexity. Put semi-formally, if d is a description of data X, and $m(d)$ is the meaningful portion of d, then our proposed meaningfulness score is given by $|m(d^*)|$, where $d^* = \text{argmin}_d\, |d(X)|$. Note that the meaningless portion still plays an important role, as it helps select the optimal description.

Of course, the meaning of a signal depends not just on the structure of the signal itself, but also on the surrounding social context. What we investigate here might more accurately be called "potential [for a signal] to be meaningful given the right context". With this caveat in mind, for the sake of concision, we refer to this just as "meaningfulness".

The contributions of this paper are the following:

- the articulation of the problem of characterizing meaningfulness and why existing methods are inadequate;
- the description of a method that avoids these shortcomings and is able to give a high score to data we know to be meaningful, and a low score to random or simple uniform data;
- the empirical evaluation of this method, compared against several baselines, on a variety of signal types.

In the remainder of this paper, Section 2 gives an overview of related work, Section 3 describes our method in detail, Section 4 presents the empirical evaluation, and Section 5 outlines future work and summarizes.

2. Related Work

The problem of measuring the complexity of data has mostly been studied in the visual domain, that is, in measuring the complexity of an image. Some methods use file compression ratio, GIF and TIFF in [9] and JPEG in [10], claiming that a

lower ratio means high complexity. Others have used the gradient of pixel intensities [11] or fractal dimension [12]. In [4], it was shown that these approaches fail to distinguish meaningful complexity from noise, and give a maximum score to white noise images. Instead, they propose to cluster patches of the image, using the MDL principle to select outliers and the number of clusters, then take the entropy of cluster indices that appear inside each patch. Our method is inspired by that of [4], but differs in two respects. Firstly, it applies to the one-dimensional case of signal processing, rather than the two-dimensional case of images, which means the patch-based recursive procedure used in [4] cannot apply. Secondly, we omit the complicated calculation of entropy from regions of cluster indices that [4] uses, and instead invoke the distinction between the meaningful and meaningless portions of the description.

In the signal processing domain, several works have proposed to measure complexity using some variant of entropy, such as evaluating on multiple timescales [3], using Tsallis q entropy [2], or replacing sections of the waveform with discrete symbols [13]. Unsupervised analysis of speech and animal vocalizations has mostly focused either on combining clustering and deep learning for feature extraction [14, 15], or on acoustic unit discovery. Some works have applied the deep learning plus clustering approach specifically to animal vocalizations, such as distinguishing different species' vocalizations [16], or distinguishing call types within a single population of orcas [17]. In terms of calculating time series complexity, one method that has been used by several works [18, 19] is to take the fractal dimension, as calculated by the Katz method [20]. In Section 4, we show empirically that our method is better able to distinguish different signal types than the Katz fractal dimension, as well as entropy and compression-ratio.

3. Method

We assume we are presented with some set of data points, and want to assign it a meaningfulness score. We cluster the data and represent each point by first specifying its assigned cluster, and then specifying where it falls in that cluster's distribution. The former, which we take as the meaningful portion, comprises an index from $1, \ldots, K$. The latter, the meaningless portion, could admit many different coding schemes, but by the Kraft-McMillan inequality, we know that, under the optimal coding scheme, the length will be bounded by, and close to $-\log p(x)$, where $p(x)$ is the probabilty under the cluster's distribution.

Alternatively, a data point can be specified directly, independently of its assigned cluster. For example, if the data consists of 64-dimensional vectors of 32-bit floats, it can be specified directly with $64 \times 32 = 2048$ bits. In this case, we regard the entire description for that data point as meaningless, as it is far away from its cluster centroid, suggesting it does not fit into a coherent pattern alongside the rest of the data points.

For a given clustering partition and given data point, we choose either the cluster-based description, or the cluster-independent description, whichever is smaller. We also take into account the number of bits needed to directly specify the clustering model itself, such as the cluster centroids (the exact parameters depend on the clustering method used). This imposes a slight additional cost on having a larger number of clusters. The total description length under a given partition is the description length of the model plus the sum of the lengths of the description of each data point under that partition, and the partition is selected to minimize this overall description length. Once the optimal partition has been found, we add together the length of the meaningful portion of the description for each data point, which amounts to taking the Shannon information content of the cluster labels assigned to all data points that use the cluster-based description, plus the description length of the model itself. The resulting sum is the final meaningfulness score.

3.1. Formal Description

Let $X \subset \mathbb{R}^m, X = x_1, \ldots, x_n$ be the input data. Let c be the numerical precision, e.g. $c = 32$ in the case of representing real numbers with 32-bit floats. Let $p(x; \mu, \Sigma)$ be the multivariate normal probability of data point $x \in \mathbb{R}^m$ given cluster centroid $\mu \in \mathbb{R}^m$ and diagonal covariance matrix $\Sigma \in \mathbb{R}^m$ (we consider only diagonal covariances to speed up search). Let g be the function that takes as input a partition function $f : \mathbb{R}^m \to \{0, \ldots, n-1\}$, and a data point $x \in X$, and returns the centroid of x under partition f. That is $g(f, x) = \frac{1}{|C|} \sum_{y \in C} y$, where $C = \{y \in X | f(y) = f(x)\}$. Let h be the analogous function that returns the diagonal of the covariance matrix of the cluster of point x under f, and let $q(x, f) = p(x; g(x, f), h(x, f))$. Let $l(i, f)$ give the number of points assigned to the ith cluster under the partition f. Then the fit clustering model is given by

$$f^* = \underset{f:\mathbb{R}^m \to \{1,\ldots,n\}}{\operatorname{argmin}} \sum_{i=0}^{n} \min(cm, -\log q(x_i, f)) \quad (1)$$

$$+ \sum_{i=1}^{n} \log \frac{n}{l(f(i), f)} \mathbb{1}(-\log q(x_i, f) < cm)\,, \quad (2)$$

where the last term uses the indicator function $\mathbb{1}$ to select only those points whose cluster-based description cost is less than their cluster-independent description cost, and the sum represents the Shannon information content of the cluster labels of those points. This selection of the partition that allows the shortest overall description of the data follows the MDL principle. Then, the complexity score is given by

$$\sum_{i=1}^{n} \log \frac{n}{l(i, f^*)} \mathbb{1}(-\log q(x_i, f^*)) \quad (3)$$

$$+ 2cm \sum_{i=1}^{n} \mathbb{1}(l(i, f^*) > 0)\,, \quad (4)$$

where the second sum is for the description of the model itself: the mean and a covariance diagonal vector of each cluster.

3.2. Implementation Details

We use a Gaussian mixture model (GMM) for clustering The GMMs are initialized with k-means, use diagonal covariance matrices, and are optimized with the usual expectation-maximization algorithm, with tolerance $1e-3$, capped at 100 iterations. They are fit 10 times with random initializations and the we select the one with the highest data probability. In order to optimize the number of clusters, we fit a separate GMM with K clusters for $K = 1, \ldots, 8$, and keep the partition from the one with the lowest cost, as given by (1). We take a single waveform as input, and form a spectrogram (window size = fft size = 30, overlap = 3). The fft for each window (i.e. column of the spectgrom) is then taken as a data point, and so we treat the single waveform as a dataset on which to run our method.

We then repeat our method twice more, where instead of taking each individual segment as a separate data point, we take contiguous chunks of several segments, 2 on the first repeat and

Table 1: *Comparison of the mean (with std in brackets) scores given by our method for each type of signal, compared with four baseline methods. Only ours gives speech a very high score, animal vocalizations a high score, and other sounds a low score.*

	ours	katz	ent	zl comp ratio	wav comp ratio
walla	61.3 (1.78)	12.9 (1.18)	100.0 (0.01)	21.7 (1.51)	14.8 (0.63)
tuning-fork	43.6 (7.93)	48.4 (8.31)	86.1 (9.79)	21.3 (4.80)	15.0 (3.86)
birdsong	72.8 (4.11)	5.1 (0.75)	100.0 (0.01)	10.8 (1.10)	22.6 (0.84)
birdsong-background	18.5 (6.84)	0.0 (0.01)	99.9 (0.03)	1.7 (0.48)	28.3 (0.07)
orcavoc	75.1 (2.95)	35.6 (8.15)	100.0 (0.01)	19.1 (4.40)	10.8 (2.36)
orcavoc-background	29.0 (5.15)	6.2 (1.19)	100.0 (0.01)	17.5 (7.36)	21.3 (0.70)
irish-m-speech	83.8 (2.14)	10.0 (1.68)	100.0 (0.01)	40.7 (4.45)	83.0 (3.36)
irish-f-speech	83.1 (2.84)	12.5 (2.69)	98.0 (1.83)	35.2 (2.84)	54.6 (10.00)
german-m-speech	84.1 (3.81)	12.4 (1.55)	100.0 (0.01)	67.6 (7.25)	29.5 (2.20)
german-f-speech	88.3 (1.89)	17.4 (1.18)	100.0 (0.01)	35.1 (4.87)	50.4 (1.40)
english-m-speech	84.6 (2.43)	16.4 (2.31)	100.0 (0.01)	35.5 (3.51)	20.2 (1.36)
english-f-speech	85.0 (2.71)	25.7 (4.53)	100.0 (0.01)	37.6 (5.02)	21.4 (2.26)
rain	2.1 (0.24)	25.4 (0.78)	100.0 (0.01)	7.4 (0.24)	4.7 (0.30)

4 on the second. At levels two and three, the vector of each data point is not the frequency spectrum, but rather the multi-set of cluster indices, from the previous level, found in chunk centred at that point, e.g. if the chunk contained two points that were assigned to the first cluster, none to the second and one to the third, the vector would be $[2, 0, 1]$. This is to allow the method the potential to capture higher-level compositional structure, such as found in language.

4. Experimental Evaluation

In this section, we show the scores from applying our method to different types of signals. Typical machine learning methods target only a particular domain and aim to distinguish different classes within that domain, e.g. distinguish between different phonemes or different speakers from human speech in a given language. Ours, in contrast, is a general method, designed to apply to any waveform with no restrictions on domain. Therefore, we evaluate it on various different types of signal, and report the average score for each signal type. Our method operates separately on each waveform, and requires no training data.

4.1. Datasets

The signal types we consider are birdsong, orca vocalizations, the background noise in these recordings, human speech in English, Irish and German, and two types of ambient noise: rainfall and muffled overlapping human conversations, a.k.a. 'walla'. We also consider recordings of tuning forks, which are physical musical devices designed to give a pure tone when struck. Walla and rainfall are public recordings from `https://www.soundjay.com`. The orca vocalizations we use comprise discrete calls only and are taken from public domain recordings by the National Park Service (NPS), available at `https://archive.org/details/KillerWhaleorcinusOrcaSoundsVocalizations`. Birdsong recordings are from the Powdermill acoustic dataset recorded in the Powdermill Nature Reserve, PA, and comprise Black-throated Green Warbler (BTNW), Ovenbird (OVEN), and Eastern Towhee (EATO). Speech recordings are taken from the common voice project, with one male and one female speaker of each language. We take time intervals of 1s at 44100 Hz, for all signal types. To show that our method is not simply responding to vocalization having greater amplitude than background noise, we normalize all waveforms to the same mean amplitude. For all datasets, we randomly pick 10 utterances per class, and manually select sections with vocalizations (ot without vocalizations, in the 'without' setting).

4.2. Comparison Methods

Several existing methods purport to measure data complexity. Some authors have proposed variations of entropy for this purpose [21, 22]. Here we compare our method to a baseline that takes the Shannon entropy of the spectrogram of the input signal. Another approach is to use the file compression ratio, This has been used in opposite senses, [23] claim that a more 'complex' signal will be less compressible, whereas [24] claim that a 'communication' signal will be *more* compressible. We evaluate the file-compression ratio, both of the waveform and of the spectrogram, in order to see if there is a pattern in either direction. We use FLAC compression for audio and Zempel-Liv on the image of the spectrogram. These are deliberately selected to be lossless because we are interested in a universal measure of meaningfulness so do not want to introduce assumptions about what parts of the information can be discarded, as e.g. in MP3 compression which invokes human audio perception and psychoacoustics. Finally, we also compare to the Katz fractal dimension, as used in [18].

4.3. Main Results

Table 1 shows the scores produced by our method, and the four comparison methods, for each signal type. Our method gives the highest score to the three human languages, English, Irish and German, followed by bird and orca vocalizations, which both get a similar score, and the lowest score to all the sounds that are not vocalizations: the background noise of the birds and orcas, tuning forks, rainfall and muffled human conversation.

This aligns with our existing understanding of the amount of information conveyed by each of these signal types. We know human speech is highly meaningful, and it is interesting that our method gives a very similar score to the three different languages, and six different speakers. Our scoring is consistent

with the general principle that all human languages are roughly equally efficient at conveying meaning [25, 26, 27].

We can be relatively sure that tuning forks, rainfall, and ambient noise are low in meaningfulness. Animal vocalizations are less well understood. They is strong evidence that orca vocalization [28, 29] convey meaning but whether it is as rich as human speech remains an open question. It is therefore correct of our method to give them a higher score than the meaningless baselines, and not unreasonable to be lower than human speech. We also note that the available recordings of vocalizations may not reflect the full meaning being conveyed. One second of human speech contains several phonemes and so spans some diversity of its sound inventory, but the right scale at which to interprete orca vocalizations is not clear. In the absence of the field's understanding of the semantic units in orca vocalization, we should regard the figures from Table 1 as a lower bound on their amount of meaningful content.

The comparison methods do not produce the same distinction of the different human speech signals as highly meaningful. The 'katz' and 'entropy' methods completely fail to show a systematic distinction, with 'entropy' assigning all methods essentially the same score, and 'katz' giving scores that appear random. The file compression ratio methods, especially, Zempel-Liv, fare better, and generally give speech a high score. This supports the claim of [24] that meaningful signals are more compressible, vs [23] who claimed they were less compressible. However, Zempel-Liv compression ratio is also high for uniform simple inputs, and we can see this in it giving a higher score to the tuning fork than to the animal vocalizations. Compression ratio is an approximation to inverse Kolmogorov complexity, which we have argued is the wrong theoretical approach to quantifying meaningfulness. The overly high score given to simple signals such as the tuning fork is a manifestation of this.

4.4. Ablation Studies

Table 2 shows the results of removing two main parts of our method. In 'no-mdl', we do not use the minimum description length to select the number of clusters K, instead we fix $K = 5$ for all inputs and all levels. In this setting, the scores are similar for all signal types. Aside from English speech, which is slightly higher than the others, it fails to distinguish more from less meaningful signals. In 'just-one-level', we omit the recursive clustering procedure described in Section 3, and only the score from the first level. This performs similarly to the full method, still managing to roughly group the signal types into human speech as one group, animal vocalizations as another group, and background/meaningless noise as a third. This shows that the higher levels are only minimally utilized. We expect that future extensions, perhaps with segment lengths that depend on the input, will show a benefit from the higher layers.

4.5. Signal Length

In order to show how the score of our method depends on the size of the input, Figure 1 plots the scores for each signal type as a function of the number of samples. The sample rate is held constant, so fewer samples equats to spanning a smaller time window. The rightmost point for each line corresponds to the main results presented in Table 1, of 1s at 44100 Hz.

For very low numbers of samples, the method gives all inputs a similar score. However, when the number of samples increases to roughly 20000, it is largely able to distinguish speech, animal vocalizations and ambient noises from each other. This shows that, with as little as 0.5s of audio, our method can assign reasonable meaningfulness scores to the signal types presented.

Table 2: *Ablation studies, removing the recursive clustering procedure ('just one level') and the MDL-based selection of outliers and the number of clusters ('no mdl').*

	ours	just one level	no mdl
walla	62.5 (2.05)	29.6 (2.02)	73.1 (1.74)
tuning-fork	44.9 (8.11)	27.0 (5.47)	67.5 (6.60)
birdsong	75.6 (3.76)	42.8 (4.21)	77.6 (2.18)
birdsong-background	19.4 (6.91)	7.8 (3.17)	68.3 (2.72)
orcavoc	74.9 (2.76)	46.4 (3.10)	60.8 (6.69)
orcavoc-background	29.7 (5.17)	12.5 (2.36)	47.5 (6.65)
irish-m-speech	84.8 (2.26)	63.8 (3.77)	81.4 (1.80)
irish-f-speech	87.9 (2.36)	65.3 (4.04)	80.3 (3.21)
german-m-speech	89.8 (2.31)	70.2 (4.45)	76.5 (4.99)
german-f-speech	89.3 (1.54)	68.6 (2.95)	82.2 (1.72)
english-m-speech	86.9 (2.47)	63.8 (4.49)	82.1 (4.62)
english-f-speech	87.7 (2.72)	70.3 (4.59)	66.5 (7.74)
rain	2.6 (0.31)	0.2 (0.01)	70.9 (2.99)

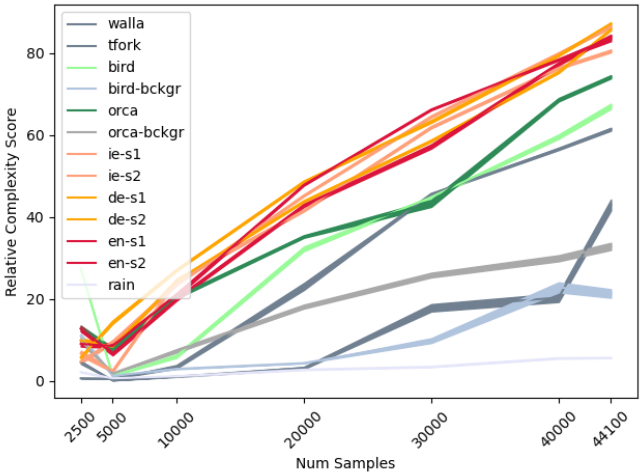

Figure 1: *Scores of our method for each signal type, as a function of the number of samples (sample rate 44100 Hz). Colours are rougly grouped by signal type, red-orange for speech, green for animal vocalizations and blue-grey for others. For each language, the first speaker ('-s1') is male, and the second female.*

5. Conclusion

This paper presented a novel method for quantifying meaningfulness of data, and applied this to method in the domain of signal processing. The meaningfulness score was very high for human speech, across six speakers and three languages, high for birdsong and orca vocalizations, and low for various other signal types, including ambient noise and the pure tone of tuning forks. The method involves clustering segments of the input signal, so as to minimize the total description length of the data under that clustering, and then taking the Shannon entropy of the cluster labels. To our knowledge, this is the first metric to give a low score to both simple uniform signals, and random noisy signals, while still giving a high score to sounds we know to be meaningful such as human speech. Future work includes, allowing variable length sound segments to adapt to different timescales in vocalization, and testing on a wider variety of animals, speakers, languages and other sound sources.

6. References

[1] P. D. Grünwald, *The minimum description length principle*. MIT press, 2007.

[2] H. V. Ribeiro, M. Jauregui, L. Zunino, and E. K. Lenzi, "Characterizing time series via complexity-entropy curves," *Physical review E*, vol. 95, no. 6, p. 062106, 2017.

[3] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical review letters*, vol. 89, no. 6, p. 068102, 2002.

[4] L. Mahon and T. Lukasiewicz, "Minimum description length clustering to measure meaningful image complexity," *Pattern Recognition*, vol. 145, p. 109889, 2024.

[5] N. Ay, M. Müller, and A. Szkola, "Effective complexity and its relation to logical depth," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4593–4607, 2010.

[6] P. M. Vitányi, "Meaningful information," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4617–4626, 2006.

[7] M. Gell-Mann and S. Lloyd, "Information measures, effective complexity, and total information," *Complex.*, vol. 2, no. 1, pp. 44–52, 1996.

[8] M. Koppel, "Complexity, depth, and sophistication," *Complex Systems*, vol. 1, no. 6, pp. 1087–1091, 1987.

[9] M. M. Marin and H. Leder, "Examining complexity across domains: Relating subjective and objective measures of affective environmental scenes, paintings and music," *PloS One*, vol. 8, no. 8, p. e72412, 2013.

[10] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[11] C. Redies, S. A. Amirshahi, M. Koch, and J. Denzler, "Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects," in *Proceedings of the European Conference on Computer Vis.* Springer, 2012, pp. 522–531.

[12] W. Sun, G. Xu, P. Gong, and S. Liang, "Fractal analysis of remotely sensed images: A review of methods and applications," *International J. of Remote Sens.*, vol. 27, no. 22, pp. 4963–4990, 2006.

[13] X. Liu, A. Jiang, N. Xu, and J. Xue, "Increment entropy as a measure of complexity for time series," *Entropy*, vol. 18, no. 1, p. 22, 2016.

[14] L. Mahon and T. Lukasiewicz, "Efficient deep clustering of human activities and how to improve evaluation," in *Asian Conference on Machine Learning*. PMLR, 2023, pp. 722–737.

[15] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[16] M. J. Guerrero, C. L. Bedoya, J. D. López, J. M. Daza, and C. Isaza, "Acoustic animal identification using unsupervised learning," *Methods in Ecology and Evolution*, vol. 14, no. 6, pp. 1500–1514, 2023.

[17] C. Bergler, M. Schmitt, R. X. Cheng, A. K. Maier, V. Barth, and E. Nöth, "Deep learning for orca call type identification-a fully unsupervised approach." in *INTERSPEECH*, 2019, pp. 3357–3361.

[18] Z. Ali and M. Talha, "Innovative method for unsupervised voice activity detection and classification of audio segments," *Ieee Access*, vol. 6, pp. 15 494–15 504, 2018.

[19] S. Yazdi-Ravandi, D. M. Arezooji, N. Matinnia, F. Shamsaei, M. Ahmadpanah, A. Ghaleiha, and R. Khosrowabadi, "Complexity of information processing in obsessive-compulsive disorder based on fractal analysis of eeg signal," *EXCLI journal*, vol. 20, p. 642, 2021.

[20] M. J. Katz, "Fractals and the analysis of waveforms," *Computers in biology and medicine*, vol. 18, no. 3, pp. 145–156, 1988.

[21] M. S. Hughes, "Analysis of digitized waveforms using shannon entropy," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 892–906, 1993.

[22] D. Mateos, R. Guevara Erra, R. Wennberg, and J. Perez Velazquez, "Measures of entropy and complexity in altered states of consciousness," *Cognitive neurodynamics*, vol. 12, pp. 73–84, 2018.

[23] N. Nagaraj, K. Balasubramanian, and S. Dey, "A new complexity measure for time series analysis and classification," *The European Physical Journal Special Topics*, vol. 222, no. 3, pp. 847–860, 2013.

[24] A. R. Rosete, K. R. Baker, and Y. Ma, "Using lzma compression for spectrum sensing with sdr samples," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2018, pp. 282–287.

[25] S. P. Liversedge, D. Drieghe, X. Li, G. Yan, X. Bai, and J. Hyönä, "Universality in eye movements and reading: A trilingual investigation," *Cognition*, vol. 147, pp. 1–20, 2016.

[26] J. A. Hawkins, *Cross-linguistic variation and efficiency*. OUP Oxford, 2014.

[27] P. Rubio-Fernandez, F. Mollica, and J. Jara-Ettinger, "Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation." *Journal of Experimental Psychology: General*, vol. 150, no. 3, p. 583, 2021.

[28] S. Sandholm, "Do orcas have semantic language? machine learning to predict orca behaviors using partially labeled vocalization data," *arXiv preprint arXiv:2302.10983*, 2023.

[29] E. L. Saulitis, C. O. Matkin, and F. H. Fay, "Vocal repertoire and acoustic behavior of the isolated at1 killer whale subpopulation in southern alaska," *Canadian Journal of Zoology*, vol. 83, no. 8, pp. 1015–1029, 2005.

On Feature Learning for Titi Monkey Activity Detection

Aditya Ravuri[1], *Jen Muir*[2], *Neil D. Lawrence*[1]

[1]University of Cambridge, UK, [2]Anglia-Ruskin University, UK

ar847@cam.ac.uk

Abstract

This paper, introduces a robust machine learning framework for the detection of vocal activities of Coppery titi monkeys. Utilizing a combination of MFCC features and a bidirectional LSTM-based classifier, we effectively address the challenges posed by the small amount of expert-annotated vocal data available. Our approach significantly reduces false positives and improves the accuracy of call detection in bioacoustic research. Initial results demonstrate an accuracy of 95% on instance predictions, highlighting the effectiveness of our model in identifying and classifying complex vocal patterns in environmental audio recordings. Moreover, we show how call classification can be done downstream, paving the way for real-world monitoring.

Index Terms: voice activity detection, self-supervised learning, representation learning

1. Introduction

Acoustic data analysis provides valuable insights into the ecological, behavioural, and health aspects of animal species. Manual processing of large volumes of acoustic data is challenging, leading to the adoption of machine learning methods in bioacoustic research. This study focuses on Coppery titi monkeys (Plecturocebus cupreus), an accessible species in our local zoos, to explore machine-learning techniques for vocalization analysis. The primary challenge is the development of a framework for voice activity detection using large volumes of passively collected titi monkey data, as relatively small amounts of expert-annotated data is usually available.

In this work, we outline a robust and performant model that appears to be robust and performant in identifying calls, that was first published as part of [1]. This companion abstract is intended to serve as a technical summary and further exposition on why it was chosen for our use-case.

2. Activity Detection Methodology

We break down the problem of activity detection into first modelling the probability of an active call by time segment $\mathbb{P}(c_t = 1|\mathbf{a})$ given an audio sequence $\mathbf{a}$, and then finding the $\text{argmax}_{\mathbf{c}}\mathbb{P}(\mathbf{c}|\mathbf{a})$ to find the most likely activity sequence given the audio.

Initial algorithms that segmented calls using spectrograms with energy between a pre-specified band were successful at identifying calls, albeit with a very high false positive rate, the need to tune hyperparameters based on context (e.g. zoo) and most importantly, led to non-smooth boundaries (e.g. noise can lead to single points in time identified as calls).

After manually labelling a significant amount of data (but still only a fraction of the collected data), we fit a simplistic model (illustrated in fig. 1) to the data, which consisted of around 500 manually labelled files, each of a 10-minute duration.

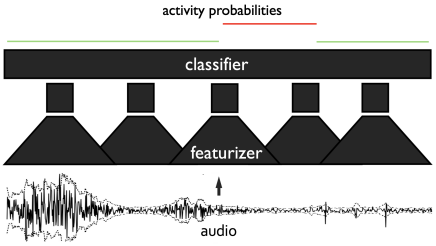

Figure 1: *Illustration of model architecture.*

MFCCs are very good representations

We found that using an MFCC featurizer with 40 MFCCs, used alongside a bidirectional LSTM-based classifier (with three layers and sixteen hidden units, and a single linear layer that compresses the hidden representation to a single probability of activity at an instance in time) works remarkably well at call detection.

The probabilistic model concretely, is, given a segment of five-second $\mathbf{a}$, we model $\forall t : c_t|\mathbf{a} \sim \text{Bernoulli}(\sigma(f(\mathbf{a})_t))$, where f denotes the featurizer and classifier.

We split up our labelled audio into five-second segments, and train the classifier on segments with calls. We achieve around a 95% accuracy on instance prediction on a validation set (i.e. $\sum_t \mathcal{I}(c_t = \lceil\sigma(f(a_t))\rceil)/n \approx 95\%$). The model has a conditional accuracy of about 82% (i.e. $\sum_{t:c_t=1}\lceil\sigma(f(a_t))\rceil/n \approx 82\%$).

An interesting consequence of our model architecture is that there's an inductive bias towards smooth outputs, illustrated in fig. 2, unlike linear or transformer architectures. We found that using a linear classifier, as is typical with wav2vec (1.0 and 2.0, [2, 3]) based ASR models, with either an MFCC or a wav2vec featurizer are non-performant, and a wav2vec featurizer with an LSTM classifier does not perform better than

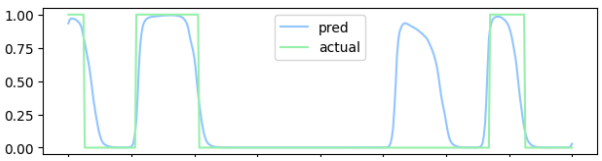

Figure 2: *Illustration of model predictions, showing smoothness of output probabilities.*

the MFCC version (and is more expensive).

Finally, we use a beam-search decoder implemented in torchaudio [4, 5] to identify segments corresponding to calls (although this is, in practice, similar to a unique consecutive search, it offers the possibility to include a language model as part of the search algorithm).

How much data is needed?

Results of re-training the MFCC-LSTM model on varying amounts of data is shown in fig. 3, showing that model metrics rise significantly until at least 250 files (half of the available data). "cond_preds" refers to the conditional accuracy of the previous section, and "hits_corr" refers to the correlation between the number of calls identified within one audio segment and the number of calls that were labelled by an expert.

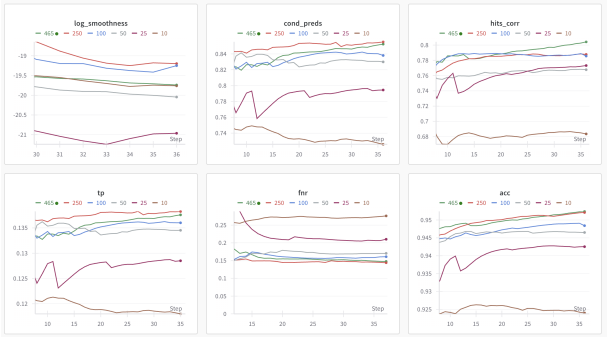

Figure 3: *Classification metrics over 1000 epochs of model training, as the number of files used for training are varied.*

If a linear classifier is needed...

Wav2vec2 based models can be trained (without labels, or using an initial classifier to pick out segments without silence) to obtain classification accuracies of about 92% (as opposed to 88.95% using MFCCs), although both of these methods are not very performant.

3. Call Classification

Given identified calls, we found that wav2vec-based features (averaged across the calls) are (somewhat marginally) better than average MFCCs, visualised using a tSNE dimensionality reduction [6] in fig. 4.

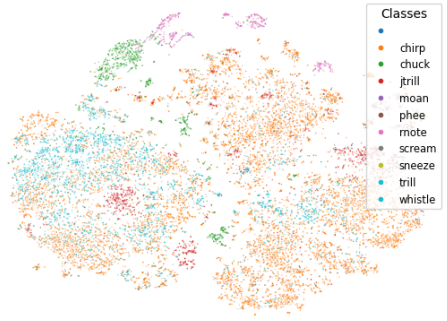

Figure 4: *tSNE of average wav2vec-based features coloured by call type.*

Moreover, in practice, we're interested in a specific pattern, known as non-linear phenomena (which has been explored for other mammals, for example, [7]). We fit a second classification model to calls identified using our first MFCC-LSTM classifier, to then do a second classification step that grouped identified calls as:

- true positive calls without non-linear phenomena
- true positive calls with non-linear phenomena
- false positives.

The model architecture for the second classifier was similar to our first, although, instead of outputting a label per time point, we do audio level classification, expecting that the second model will only take in calls, and this is done by averaging the latent representation of the LSTM before feeding it into a linear layer. The accuracy of this second step classifier is about 75%, and when deployed to a totally unseen data file and context, the model hits were all found to be cases of non-linear phenomena, and included cases where the expert found difficult to identify manually.

4. Conclusion

In conclusion, our study presents a highly effective machine learning framework for detecting the vocal activities of Coppery titi monkeys using a combination of MFCC features and a bidirectional LSTM classifier. The model demonstrated a robust capability in reducing false positives and achieving a high accuracy rate in vocal call detection, with promising applications in real-world bioacoustic monitoring. Moreover, the adaptation of our framework to distinguish specific call patterns, including non-linear phenomena, shows potential for enhancing ecological and behavioral studies. This work lays the groundwork for future research in applying advanced machine learning techniques to bioacoustics, potentially extending to a wider range of species and environmental conditions, thereby contributing significantly to wildlife conservation and ecological studies.

5. References

[1] J. Muir, A. Ravuri, E. Meissner *et al.*, "Anonymous paper," 2024, under review.

[2] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[4] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[5] J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, A. Kumar, C.-Y. Yu, C. Zhu, C. Liu, J. Kahn, M. Ravanelli, P. Sun, S. Watanabe, Y. Shi, Y. Tao, R. Scheibler, S. Cornell, S. Kim, and S. Petridis, "Torchaudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch," 2023.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[7] W. Fitch, J. Neubauer, and H. Herzel, "Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production," *Animal Behaviour*, vol. 63, no. 3, pp. 407–418, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003347201919128

Index of Authors

VIHAR 2024

http://vihar-2024.vihar.org/