Proceedings of the 3rd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots

VIHAR 2021

Paris, France, 13-15 October 2021

Virtual conference

Published by:
Ricard Marxer

ISBN: 978-2-9562029-2-9

Credits:
Editors: Mohamed Chetouani, Elodie Mandel-Briefer, Angela Dassow, Ricard Marxer, Roger K. Moore, Nicolas Obin, Dan Stowell
Cover photo: Yann Caradec, La Tour Eiffel vue de la Tour Saint-Jacques, Paris août 2014, Recolored and cropped by Ricard Marxer, CC BY-SA 2.0,
https://en.wikipedia.org/wiki/File:La_Tour_Eiffel_vue_de_la_Tour_Saint-Jacques,_Paris_ao%C3%BBt_2014_(2).jpg
Proceedings assembled by: Ricard Marxer

Workshop took place virtually in Paris, France — October 13-15, 2021

Published online at http://vihar-2021.vihar.org/ — October 11, 2021

Workshop Organisation

Organising Committee

Mohamed Chetouani Sorbonne University, FR

Dan Stowell Tilburg University / Naturalis Biodiversity Centre, NL

Angela Dassow Carthage College, US

Ricard Marxer Université de Toulon, Aix Marseille Univ, CNRS, LIS, FR

Nicolas Obin IRCAM, Sorbonne University, FR

Roger K. Moore University of Sheffield, UK

Elodie Briefer University of Copenhagen, DK

Scientific Committee

Sabrina Engesser University of Vienna

Elodie Mandel-Briefer ETH Zürich

Roger K. Moore University of Sheffield

Julie Oswald University of St. Andrews

Dan Stowell Tilburg University / Naturalis Biodiversity Centre

Ricard Marxer Université de Toulon, Aix Marseille Univ, CNRS, LIS

Workshop supported by

SORBONNE
UNIVERSITÉ

Conference Program

Keynotes

Papers 1a

Papers 1b

Papers 2a

Papers 2b

Papers 3a

Papers 3b

Animal-Computer Interaction: Future Directions for Animals Interacting with Computers

Ilyena Hirskyj-Douglas

Biography

Ilyena Hirskyj-Douglas is a Lecturer of Animal-Computer Interaction at the University of Glasgow. She has studied and built novel systems for animals for over ten years speaking widely on equity, design, and methods and theory behind animals interacting with computer-enabled systems globally. Her initial contribution to the field started with methods and approaches for dogs to control screen systems in their home. Today, Ilyena focuses on creating an animal-to-animal internet, animal controlled systems, and musical and video systems for monkeys with user-centred values at the core.

“Who’s a good dog?” The influence of dog directed speech on human-dog interactions

Katie Slocombe

Biography

I completed my PhD in Chimpanzee Communication at the University of St Andrews, UK. After a brief post-doc at St Andrews I joined the University of York as a lecturer in 2007. Since establishing my own research group at the University of York I have continued to examine communication and social cognition in chimpanzees, but I have also investigated communication in other primates, including humans and dogs.

Verbal and Non-Verbal Human-Robot Interaction: Where are we and what's next?

Nikolaos Mavridis

Biography

Dr. Nikolaos Mavridis, PhD from the Massachusetts Institute of Technology, is an academic and consultant specializing in Robotics and Artificial Intelligence. He is the founder and director of the Interactive Robots and Media Lab (IRML) and has served in various professor positions, including NYU Poly and AD, Innopolis University, and UAEU, all the way to Full Professor rank. He has been a four-time TEDx speaker, as well as a speaker at Singularity University, and is also a pro-bono contributor to a number of organizations. His interests include human-robot interaction, machine learning, cognitive systems, and he has more than 80 peer-reviewed academic publications. Among other appointments, Nikolaos has served as a judge for the UAE Prime Minister's Office "Drones for Good" and "Robotics and AI for Good" competition, is a member of EU Cognition action and the MIT Educational Council, a mentor for the MIT Enterprise Forum, has given 4 TEDx talks (including Geneva and Athens) and has been a Singularity University speaker, and is frequently invited for keynote talks worldwide.

Representation learning for orca calls classification

Paul Best[a], Ricard Marxer[a], Sébastien Paris[a], Hervé Glotin[a]

[a] *Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon, France*

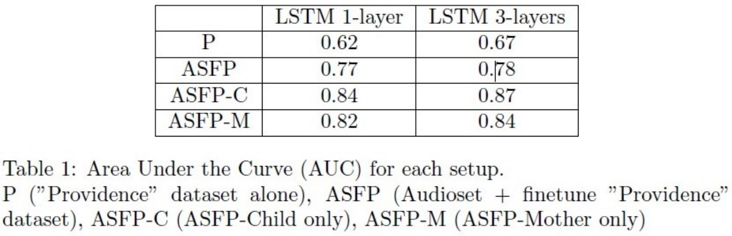
We study several semi-supervised representation learning approaches to improve orca call classification. Orca calls are stereotyped tonal vocalizations used for communication. We tackle the task with convolutional neural networks taking mel spectrograms as input. Robust deep learning models traditionally require large amounts of annotated data to obtain satisfactory performance. Starting from no labels, we first show how an auto-encoder reconstruction loss yields a meaningful representation of the data, enabling an efficient manual annotation procedure leveraging clusters in the latent space. With the resulting annotations, we study how supervised and unsupervised losses can be combined to make use of the large amounts of unlabelled data and enhance generalization performances. We use the triplet loss approach, augmenting the unlabelled data in the Fourier space (like SpecAugment). The effect of several distance metrics are studied for this loss (cross entropy, mutual information, euclidean distance, cross correlation and cosine similarity). This work provides an overview of numerous methods of representation learning, how they can be used to annotate data efficiently and/or as a regularization loss. This study is applied to the classification of orca calls which is key to understanding their communication system. The methods employed are not species specific and can be applied to other animal communication systems.

Investigating Automatic Audio Laughter Detection in a Mother-Child Interaction

Kevin El Haddad[a], Gabriel Meunier[b], Chiara Mazzocconi[c], Abdellah Fourtassi[d]

[a]*ISIA Lab, University of Mons*
[b]*Aix-Marseille University*
[c]*ILCB, LPL, Aix-Marseille University*
[d]*ILCB, LIS, Aix-Marseille University*

Laughter is one of the earliest means that an infant has to convey meaning,practising turn-taking, attention sharing, directing other's attention and con-tribute to interaction at the same level of an adult. Through development its use becomes more and more sophisticated both from a semantic and pragmatic perspective, and can give us important insights into the child's cognitive, linguistic and pragmatic development on different levels of observation (e.g. Reddy et al.(2002); Mireault and Reddy (2016), Mazzocconi and Ginzburg (2020). There is though a dearth of work in the study of laughter development and manual annotation is extremely time consuming. We will present our work in progress aimed at automatically detect laughter in the context of mother-child interaction in an ecological familiar context. Al-though some work can be found on automatic laughter detection, these tend to not focus exclusively on a mother-child context, and so, perform quite poorly when confronted with such data. Among the main reasons for this, count the inter-variability in the children's laughs as well as the similarity between children's laughs and other types of sounds the children may utter. As a starting point we are exploiting exclusively audio data, leaving to further work the elaboration of a model able to exploit both audio and visual data according to their availability.We thus present, here, first promising results on automatic mother-child laughter detection based uniquely on audio data along with analyses which will help to develop a better understanding of the problem faced and so, paving the way towards developing a fully functional, precise and robust uni/multi-modal mother-child laughter detection system. We also present the data collected, annotated, processed and used to develop the machine learning and deep learning-based systems used here. From this work, we can draw so far two main recommendations for future work. First, compared to traditional ma-chine learning algorithms like Support Vector Machines, K-Nearest Neighbours and Decision Trees combined with traditional features like the pitch and the Mel-Frequency Cepstrum Coefficients (MFCCs), our work shows that the best results on our data were obtained using Long-Short Term Memory layers trained on VGGish embeddings (team at Google, 2017) (with AUCchildren= 0.87,AUCmothers= 0.84 and AUCall= 0.78 - refer to Table 1 for the complete results). Second, increasing the amount and variability of the data usually improves the robustness and precision of deep learning systems. Our experiments also confirms this theory on our application as well. Our work thus suggests that i) increasing and augmenting our data and ii) to use deep learning-based architectures and pre-trained models are the right approaches towards improving our results.Reaching a good performance on this task will allow us to compare on a large scale typical and atypical neuro-psychological laughter and mother-child interactional dynamics development, as well as studying in more detail mother and child alignment and reciprocal influence on interactional dynamics. Laughter behaviour could indeed be an early marker of atypicalities in communicative development (e.g. Autism).



	LSTM 1-layer	LSTM 3-layers
P	0.62	0.67
ASFP	0.77	0.78
ASFP-C	0.84	0.87
ASFP-M	0.82	0.84

Table 1: Area Under the Curve (AUC) for each setup.
P (”Providence” dataset alone), ASFP (Audioset + finetune ”Providence” dataset), ASFP-C (ASFP-Child only), ASFP-M (ASFP-Mother only)

An Interactive Human—Animal—Robot Approach to Distance Sampling in Bioacoustics

Vincent Lostanlen[a], Pierrick Arnaud[b], Marc du Gardin[b], Laurent Godet[c], Mathieu Lagrange[a]

[a] *Université de Nantes, Centrale Nantes, CNRS, LS2N*
[b] *École navale*
[c] *Université de Nantes, CNRS, LETG*

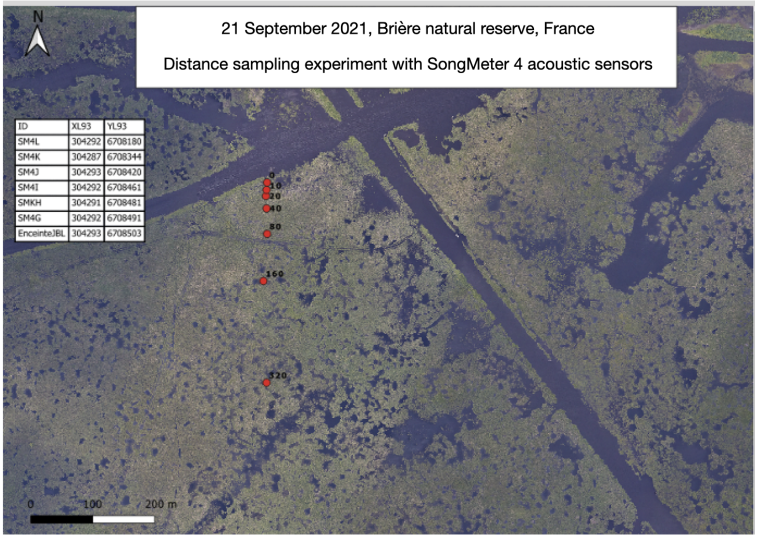
In the context of bird conservation, the measurement of relative species abundance often depends on human surveys: experts on the field would draw a list of the species they see and hear over a fixed duration, typically of the order of a few minutes. In addition to the ecological disturbance caused by the presence of humans, we note that this method is costly and time-consuming. As such, it lacks the scalability to monitor birds continuously over large spatiotemporal scales. Another approach consists in deploying a network of acoustic sensors which record continuously at fixed locations. Then, analyzing the resulting audio data via state-of-the-art machine listening techniques, such as BirdNET, allows to automate the process of checklisting species.

However, the detection radius of these techniques is unknown and probably species-dependent. For this reason, passive acoustic monitoring with autonomous recording units cannot yet replace human surveys. Against this problem, we propose an simple method to estimate the coverage area of bioacoustic event detectors in situ. This method resembles distance sampling in population ecology, in which the distance between listener and source serves to explain variations in detectability.

Recent publications in bio-acoustics have proposed to implement distance sampling by recording bird calls at known distances and then regressing those distances via some audio-based measurements, typically relative sound level with respect to background noise (Sebastian-Gonzalez et al., 2018; Yip et al., 2019). Although these new publications offer a statistical estimator of detection radius, they require the presence of vocalizing birds in situ for each species of interest, thus limiting the speed of the calibration process.

We propose to address this issue by playing back pre-recorded bird vocalizations from a portable loudspeaker instead of relying on actual birds. To this end, we download vocalizations from online digital audio archives (Macaulay Library and Xeno-Canto), denoise them manually, and amplify them to match the known vocal intensity of the species of interest.

In our presentation, we report the detection functions of ten species which are common in the natural park of Brière, a special area of conservation nearby Nantes, France. While doing so, we study the effect of recording equipment: high-quality (SongMeter 4) versus low-cost (AudioMoth). We also compare the detection abilities of machines versus an expert human operator on the field. The main goal of our method is to come up with a per-species estimate of vocal activity over the whole protected area of the natural park from a representative sample of sensor locations.



21 September 2021, Brière natural reserve, France

Distance sampling experiment with SongMeter 4 acoustic sensors

ID	XL93	YL93
SM4L	304292	6708180
SM4K	304287	6708344
SM4J	304293	6708420
SM4I	304292	6708461
SMKH	304291	6708481
SM4G	304292	6708491
EnceinteJBL	304293	6708503

Vocal interactions in meerkat groups on the move

Vlad Demartsev[a,b,c], Mara Thomas[a,b,c], Baptiste Averly[a,b,c], Marta Manser[d,c], Ariana Strandburg-Peshkin[a,b,c]

[a]*Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany*
[b]*Biology Department, University of Konstanz, Konstanz, Germany*
[c]*Kalahari Research Centre, Kuruman River Reserve, Van Zylsrus, South Africa*
[d]*Department of Evolutionary Biology and Environmental Studies, University of Zurich*

Animal vocal communication research traditionally focuses on the acoustic features and ecological context of calls in order to estimate their function and informational content. However, there are additional informational layers derived from response selectivity, timing and distance between individuals during vocalization events, which have the potential to both reflect and mediate social relations. By examining the spatio-temporal dynamics of signaling, we can potentially distinguish between two different signaling modes: a "broadcast" mode where signals convey information about the environmental context or signaler's state but are not contingent on signals from other individuals (e.g. alarm calls in response to predators), and a "conversational" mode where signals are given in response to other signals, leading to the development of an interactive vocal exchange. Here we investigated vocal interactions in meerkats (Suricata suricatta), social mongooses that live in cohesive groups which move together throughout the day and use a complex, well-characterized vocal repertoire to coordinate their behavior. Using collars combining GPS and acoustic sensors, we collected data on the movements and vocalizations from most individuals simultaneously within meerkat social groups, capturing a timeline of vocal events coupled with the positions and movements of individuals.

Audio recorders on the collars sometimes picked up calls not just from the collar-wearing individual but also from vocalizing neighbors nearby, making reliable caller identification difficult. We developed a semi-automated method to disentangle which individual produced each call by pairwise comparison of amplitude and acoustic structure of overlapping calls with ambiguous caller assignments. We then analyzed call-response dynamics between pairs of individuals for different call types to distinguish between call-type dependent transmission modes and to determine the spatio-temporal organization of interactive call exchanges.

We investigated the typical call-response dynamics between pairs of individuals associated with two particular types of frequently-emitted calls - "close calls" which are given while foraging to maintain cohesion and "short note" calls which are given in various contexts (e.g. sentinel, fast movement). We found that for both types of call, the average call rate for a given focal individual (henceforth the "responder") was highly distance dependent, with the responder more likely to produce calls when they were within a short distance of a vocalizing conspecific (henceforth the "initiator"). However, the temporal dynamics revealed different patterns for the two call types (Figure 1). For short note calls, the responder's peak call rate occurred concurrently with calls of the initiator, whereas for close calls this peak occurred at a lag of approximately 200 ms, typical of auditory processing and response time. These results suggest that while both types of call reflect local context, only close calls show evidence of a call-response dynamic, and are a part of an interactive signal exchange rather than a general contextual response. We also explore the possibility of short-term vocal convergence or "copying" within these vocal exchanges by investigating whether sequential calls are more similar to one another than are calls of the same type given further apart in time.

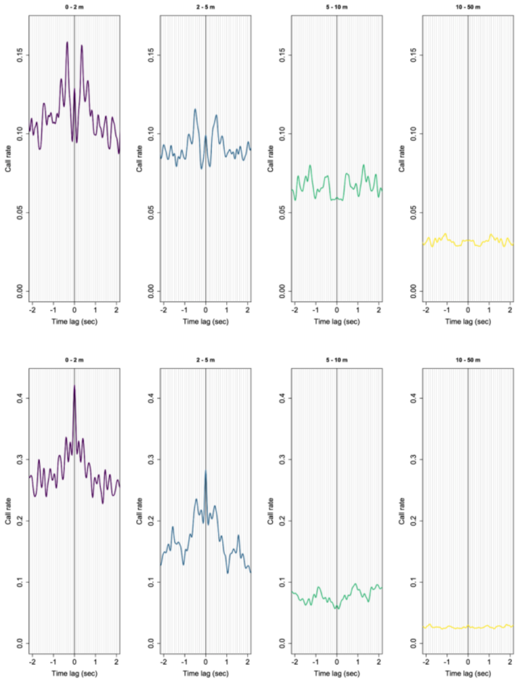


Figure 1. Call rate of a given individual (the "responder") as a function of time prior to and after a conspecific (the "initiator") call at t = 0. Panels represent average responder call rates for close calls (top row) and short note calls (bottom row) located at increasing distances (left - right) from the initiator.

Decadal frequency shift in blue whale song could be explained by dynamical social network analysis or flocking models

Franck Malige[a], Julie Patris[b], Maxime Hauray[c,b], Pascale Giraudet[a], Hervé Glotin[a]

[a] *Université de Toulon, Aix Marseille Univ, CNRS, LIS, DYNI, Toulon, France*
[b] *Université d'Aix-Marseille*
[c] *Institut de mathématiques de Marseille*

Blue whales emit very strong, complex and low sounds, repited rythmically for hours, called songs.

Several very distinct types of songs have been registered worldwide for this species (McDonald 2006). All these blue whale song types undergo a linear decrease of the emitted frequencies over time which has been documented for more than sixty years for some of them (McDonald 2009). This slow but constant evolution is an unexplained phenomenon. These last ten years, various hypothesis have been emitted to explained it from recovering from hunting to an increase of anthropogenic noise (McDonald 2009, Leroy 2018). However none of the hypothesis presented seem decisive to explain this global and very stable phenomenon and it remains an open question.

To model this frequency shift, we build a theoretical framework applying technics from dynamical social network analysis and from flocking behaviour. This model is based on very few biological hypothesis of conformity and sexual competition. We then compare the results given by this theorical tool to the measured shifts. We finally draw the conclusion that these models are compatible with all the data available, including yearly variations in the frequency (Gavrilov 2012). Finally, we propose a way to test these models in the future.

- McDonald, M., Mesnik, S. Hildebrand, J. Biogeographic characterization of blue whale song worldwide: using song to identify populations J. Cetacean Res. Manage., 55–65 (2006).
- McDonald, M., Hildebrand, J. Mesnick, S. Worldwide decline in tonal frequencies of blue whale songs. Endangered species research 9, 13–21 (2009).
- Gavrilov, A., McCauley, R. Gedamke, J. Steady inter and intra-annual decrease in the vocalization frequency of antarctic blue whales. J. Acoust. Soc. Am. 131, 4476–4480 (2012).
- Leroy, E. C., Royer, J.-Y., Bonnel, J. Samaran, F. Long Term and Seasonal Changes of Large Whale Call Frequency in the Southern Indian Ocean. Journal of Geophysical Research: Oceans 123, 8568–8580 (2018).

TamagoPhone: Augmented incubator to maintain vocal interaction between bird parents and egg during artificial incubation.

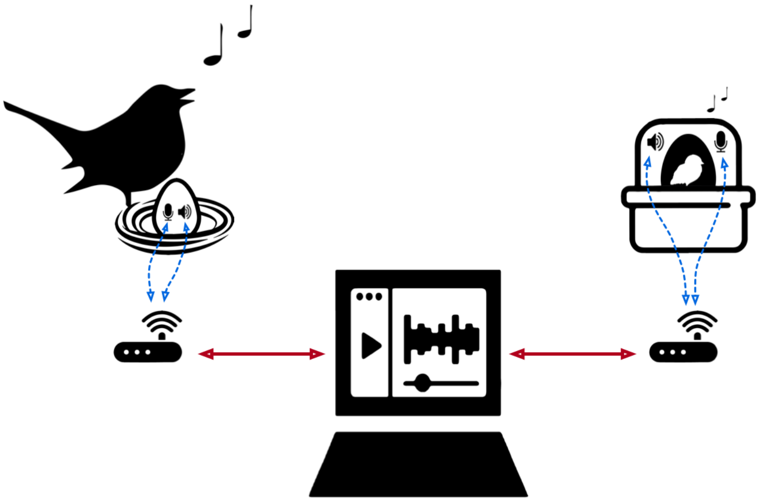
Rebecca Kleinberger[a], Janelle Sands[b], Sareen Harpreet[c], Janet M. Baker[a]

[a]*Massachusetts Institute of Technology Media Laboratory*
[b]*Massachusetts Institute of Technology*
[c]*Parson School of Design The New School*

It is common practice for zoos and bird conservationists to incubate bird eggs in artificial incubators to maximize hatching rates. In the wild or even in zoos within enclosures, eggs can be vulnerable to predators, diseases or temperature problems, and the use of artificial incubators during part or the entirety of the incubation time can greatly improve the chances of survival of the chicks and support preservation efforts for endangered species. However, some species exhibit important prenatal vocal interaction while within the egg. Indeed, parent birds often produce vocalisations directed to their eggs and in some species, chicks produce calls from within the egg a few days before hatching. Standard artificial incubation techniques deprive embryonic chicks of integral parent-offspring vocal communications during early development. Recent research has shed light on specific behavioral contexts associated with vocal pre-hatching events for specific species. Led by such research, there have been attempts of using curated static recordings during artificial incubation and hand-rearing to alleviate the lack of vocal interactions. However, our understanding of those vocal interactions is still in its infancy, and might never be understood well enough to synthesize meaningful replacement or select relevant recordings. Instead, we propose supplementing existing egg incubation techniques with a two-way, real-time audio system to allow parent birds and unhatched eggs to communicate with each other remotely in real-time during incubation. To achieve this, we designed a framework and approach called TamagoPhone. With this approach, the real egg removed from the nest is replaced by an augmented "dummy" egg, embedding a microphone and speaker, which is then cared for by the parent. The artificial incubator is also augmented with microphones and speakers, in addition to the traditional temperature, humidity, and motion control systems. Both sides, parent and egg, are connected by a two-way audio streaming platform, with the audio components inconspicuously integrated.

Previous research on responsiveness and vocalization in bird embryos supports our investigations. It is now well established that avian prenatal sensory experience affects development and has long-term consequences on postnatal behavior, which varies between altricial and precocial species. Species of birds who are able to feed themselves, are covered with down, have their eyes open, and leave the nest days following hatching, are classed as precocial. Birds with closed eyes, very little down, who are unable to leave the nest for some time are classified as altricial. Previous work has highlighted the development of auditory sensitivity of bird embryos prior to hatching.

In this work, we first review previous research on known functions of both parental and embryonic vocal signals, we then describe the TamagoPhone intervention and its potential application in various contexts including preservation, research, and farming. Finally, we propose a taxonomy of success criteria and evaluation metrics for each application context. We believe that such a system could be a way to use technology to increase vocal connectivity between the mother and her young while acknowledging our human limitations in understanding the possible meanings and functions of the vocal signals exchanged.



Could Animals be Taught to Communicate their Emotions? A Potential Behavior Training Protocol Modeled in Parrots

Jennifer Cunha[a]

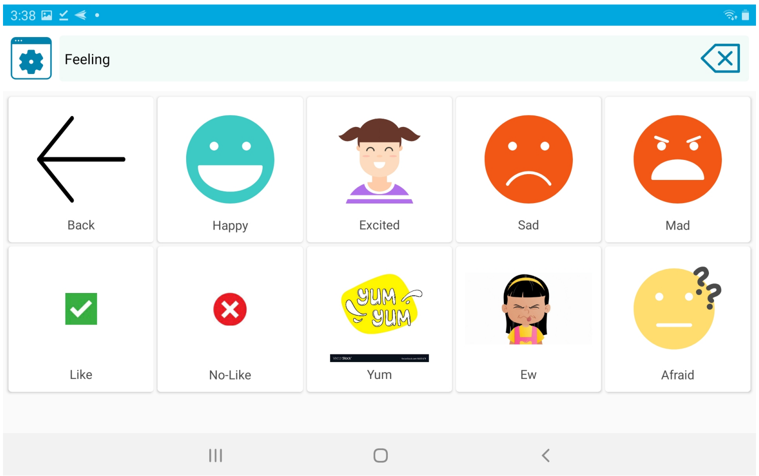
[a] *Parrot Kindergarten, Inc.*

Based on Dr. Paul Eckman's work, the American Psychological Association has long identified five standard emotion states: happy, sad, afraid, surprised, and mad. These five emotion states were used as the basis to design a communication system in which parrots were trained to identify and label all five of those emotion states and press buttons on the CommBoard, a commercially-available Augmentative Alternative Communication Device, in order to express those states (labeled "feelings") to their human caregivers.

The training involved identifying the individually-expressed behavior repertoire associated with the emotion states in each parrot. For instance, when one of the cockatoos, Isabelle, was "excited," she spread her wings and bobbed her head and torso with focused eyes while vocalizing at a high pitch. When she was "happy," her feathers and eyes were relaxed and she often vocalized at a more steady, quiet pitch, like a "song" or she talked quietly. Both behavior repertoires were categorically distinct from one another and had differing triggers. The trainer then identified triggers for each "feeling" behavior repertoire. For instance, the presence of a particular human friend triggered "excitement" as did engagement in particular play activities, and specific music genres. Thus, the trigger labels were "Grandma," "Learning" and "Dance Music."

The triggers for "happy" included particular food items, certain play activities, and new toys, resulting in that particular behavioral repertoire. Thus, the "happy" trigger labels were "Banana," "Book," and "New Toy."

Using a two-choice forced task discrimination training for associative conditioning under voluntary conditions, the parrots were taught trigger word labels (i.e., grandma, learning, dance music, banana, book, and new toy). They were then taught the concept of "Feeling" with adapted categories for feeling labels, (excited, happy, afraid, mad, and sad). The trigger words were associated with behavior labels and then the birds were given discrimination tasks to gauge accuracy with the conditioned word associations. Finally, the birds were asked questions generalizing the feelings into novel (previously trained) world knowledge word associations (i.e., other birds, people, experiences) to analyze corroboration between their stated feeling states and their behavior repertoires with those particular triggers. In this session, I will discuss the training protocol to teach trigger-label vocabulary, emotion-label vocabulary, and the conditioning techniques to associate the trigger labels to the emotion categories and behavior repertoire.

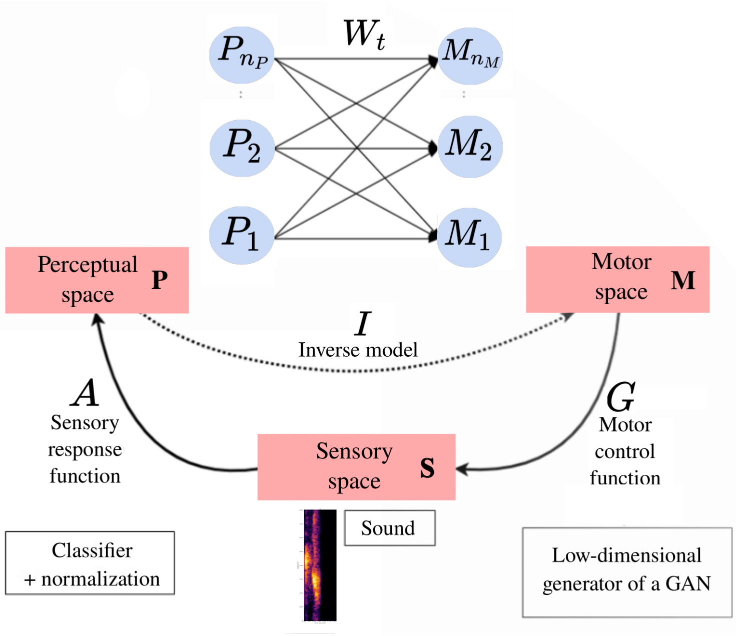

3:38
Feeling
Back
Happy
Excited
Sad
Mad
Like
No-Like
Yum
Ew
Afraid

A complete vocal learning model for canary syllables

Silvia Pagliarini[a]

[a] *University of California Los Angeles*

Birds are an ideal model for exploring the representation of vocal learning by imitation of tutors. Indeed, the behavioral studies and the neuroanatomical structure of the vocal control circuit in humans and birds provide the basis for bio-inspired models of vocal learning (Brainard 2002, Kuhl 2004, Chakraborty 2015). During the first period of their life, babies and juvenile birds listen to their parents/tutors in order to build a neural representation of the experienced auditory stimulus. Then, they start to produce sound and progressively get closer to reproducing their tutor song. This phase of learning is called sensorimotor phase and is characterized by the presence of babbling, in babies, and subsong, in birds. It ends when the song crystallizes and becomes similar to the one produced by the adults. Previous studies have attempted to implement imitative learning in computational models and share a common structure (Oudeyer 2005, Pagliarini 2020). These learning architectures include the learning mechanisms and, eventually, exploration and evaluation strategies. A motor control function enables sound production and sensory response models either how sound is perceived or how it shapes the reward. The inputs and outputs of these functions lie (1) in the motor space (motor parameters' space), (2) in the sensory space (real sounds) and (3) either in the perceptual space (a low dimensional representation of the sound) or in the internal representation of goals (a non-perceptual representation of the target sound). To obtain such a complete vocal learning model represents a challenge: each of the components described above needs to be tested and calibrated in order to obtain a suitable vocal learning model. Moreover, the learning architecture must be defined and embedded in the model, taking into account the biological constraints if desidered. The model we propose is a complete vocal learning model with a full action-perception loop (i.e., it includes motor space, sensory space, and perceptual space, and it is enriched by a motor control function capable of reproducing sounds). In our model, the sound production is performed by a low-dimensional WaveGAN generator (Donahue 2018, Pagliarini 2021). In this generator model, the input space becomes the latent space after training and allows the representation of a high-dimensional dataset in a lower-dimensional manifold. We obtained realistic canary sounds using only three dimensions for the latent space, and provided quantitative and qualitative analyses that demonstrate the interpolation abilities of the model. A recurrent neural network classifying syllables serves as the perceptual sensory response. We tested whether or not (1) a low-dimensional generator model is able to produce realistic syllables, and (2) the mapping between the perceptual space and the motor space can be learned via an inverse model. We focused on the comparison between different motor space dimensions and different learning rates. As general perspectives of this work, one could (1) test the same generator with different datasets to assess its capability of reproducing realistic outputs, and (2) use the same model structure for vocal learning in artificial agents' communication or humans.

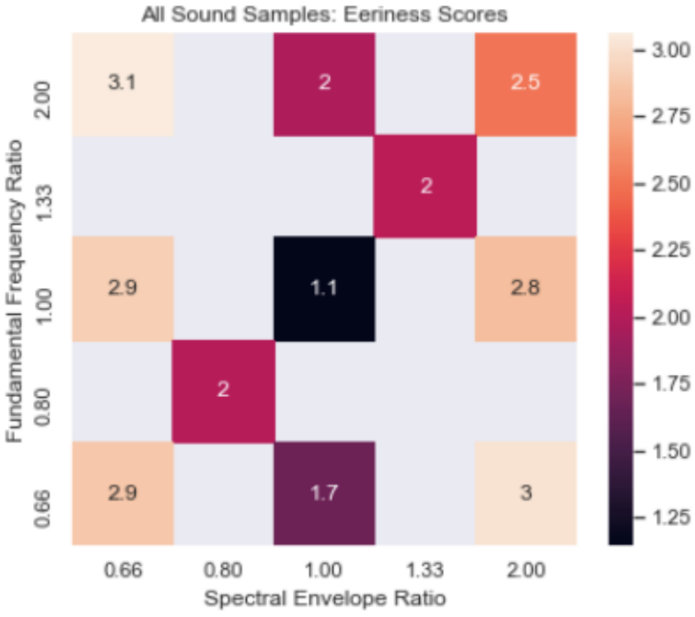

W_t
P_{n_P}
P_2
P_1
M_{n_M}
M_2
M_1
Perceptual space **P**
Motor space **M**
I
Inverse model
A
Sensory response function
G
Motor control function
Sensory space **S**
Sound
Classifier + normalization
Low-dimensional generator of a GAN

Can a Voice be Uncanny?

Ian T. Coldren[a], Roger K. Moore[a]

[a] *University of Sheffield*

It is well established in robotics that near-human artefacts can engender feelings of uncanniness and even repulsion [1], and clearly this has important implications for human-robot interaction [2]. There are numerous descriptive explanations of why this so-called 'uncanny valley' effect occurs [3-5], but the only quantitative model posits that the perception of uncanniness arises from category uncertainty when perceptual cues are misaligned [6]. This hypothesis has been confirmed for faces [7,8], and for mismatched faces and voices [9]. However, until now, it has not been tested on voices alone, i.e. is it possible for a voice to be perceived as uncanny? Following the same experimental protocol developed for assessing uncanniness in faces [10], we independently varied the length of the vocal tract and voice pitch to create a range of different voices. Subjective tests revealed that it is indeed possible to configure vocal combinations that are perceived by listeners as uncanny, thereby confirming the importance of creating 'appropriate' voices when designing voice-enabled artefacts in human-robot interaction.



[1] Mori, M. (1970). Bukimi no tani (The Uncanny Valley). Energy, 7, 33–35.

[2] Walters, M. L., Syrdal, D. S., Dautenhahn, K., Boekhorst, R. te, & Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. Autonomous Robots, 24(2).

[3] Pollick, F. E. (2010). In search of the uncanny valley. In P. Daras & O. M. Ibarra (Eds.), User Centric

Media (pp. 69–78). Berlin Heidelberg: Springer.

[4] Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2011). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Social Cognitive and Affective Neuroscience, 7(4), 413–422.

[5] Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Frontiers in Psychology, 6(390), 1–16.

[6] Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Nature Scientific Reports, 2(864).

[7] Seyama, J. (2007). The uncanny valley: effect of realism on the impression of artificial human faces. Presence, 16(4), 337–351.

[8] MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. Computers in Human Behavior, 25, 695–710.

[9] Mitchell, W. J., Szerszen Sr., K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. I-Perception, 2(1), 10–12.

[10] MacDorman, K. F., & Chattopadhyay, D. (2015). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition, 146, 190–205.

Should I stay or should I go? – Humans' perception of social valence in artificially generated sounds

Beáta Korcsok[a], Tamás Faragó[b], Bence Ferdinandy[c], Ádám Miklósi[b,c], Péter Korondi[d], Márta Gácsi[b,c]

[a]*Department of Mechatronics, Optics and Mechanical Engineering Informatics, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Budapest, Hungary*
[b]*Department of Ethology, Eötvös Loránd University, Budapest, Hungary*
[c]*MTA-ELTE Comparative Ethology Research Group, Budapest, Hungary*
[d]*Department of Mechatronics, University of Debrecen, Hungary*

Specific emotionally expressive vocalizations can elicit approach-avoidance reactions in humans and non-human animals serving as a social valence dimension, complementing the differentiation of the biological functions of vocaliations with similar perceived emotional valence and intensity. In the framework of human-robot interactions, emotionally expressive non-verbal vocalizations are an important, albeit somewhat neglected aspect of communication in regards to the social dimension. Our aim was to investigate whether humans are able to attribute social valence to artificially generated sounds, and whether the social valence attribution is linked to acoustic attributes or valence-intensity perception. We used 343 artificial sounds generated via Praat with differing call lengths, fundamental frequencies and multiple levels of complexity achieved by adding acoustic parameters prevalent in animal vocalizations. The sounds were a subset of samples generated in a previous study investigating valence and intensity. We created an online questionnaire with a manikin task in which the participants could indicate if they would approach an object emitting the sample sound or withdraw from it, by moving a human-like figure closer or farther away from the sound source. More than 170 participants filled out the questionnaire; their answers were analysed with Mixed effects linear regressions and with Simple slope analysis post-hoc tests. The results show that participants exhibited different approach or avoidance reactions to the sounds based on various acoustic parameters: short sounds evoked approach in all categories, as well as lower pitch sounds, except in the most biologically complex sound categories in which higher pitch elicited approach. In interaction with call length, short, less loud and high pitched sounds elicited approach. Loud sounds elicited avoidance. Valence and intensity perception was also associated with approach-avoidance reactions, as positive valence sounds elicited approach, while negative ones withdrawal. High intensity sounds also elicited avoidance in case of negative valence sounds. We conclude that these approach-avoidance tendencies relating to acoustic parameters together with their variations in complexity categories can provide an advantage in designing sounds for social robots with specific approach or avoidance eliciting functions.

From vocal prosody to movement prosody, from HRI to understanding humans

Philip Scales[a], Véronique Aubergé[a,b], Olivier Aycard[a]

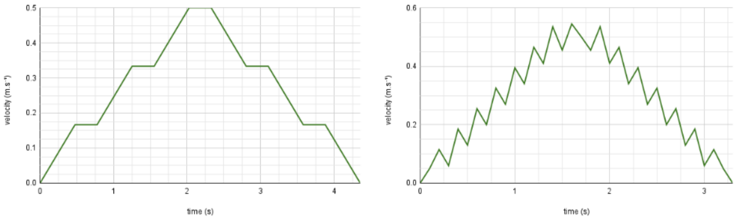
[a] *LIG, Université Grenoble Alpes*
[b] *CNRS*

Vocal interaction is one of many communication vectors which are employed when Humans, Animals, and Robots interact. Prior works by Aubergé et al. have aimed to study and characterize vocal prosody, and to understand its role and impact on interactions not only among humans, but also between humans and robots. Now, in our line of work, we aim to explore another dimension and communication vector which is that of spatio-temporal interaction. Thus, we aim to study how a mobile robot's kinematics, movement profiles, paths and timing can impact interaction with people. Essentially, we aim to study the concept of "prosody of movement". This is a critical point to understand, given that many projects and companies are aiming to deploy mobile robots around humans in the near future.

In a recent work, we took a first step towards understanding the link between motion parameters and human's interpretations of robot's behaviour. In order to achieve this, we designed a corpus of robot motions which cover a large scope of feasible motions. We then filmed a mobile robot performing a large number of movement trajectories with varying movement prosody and appearance parameters selected according to the motion corpus design. The parameters of the motion corpus are hypothesized to be responsible for eliciting reactions in people, who tend to interpret robot motions in terms of intentions or attitudes. We also believe that certain motion parameters are analogous to vocal prosody parameters. The first goal of the corpus construction was to systematically categorize and distinguish types of movement. The second goal was to build and provide open access to a novel video corpus, which can be re-used by other researchers to conduct their own studies. The third goal was to design a perception experiment, using the videos as the stimuli.

The design and deployment of the online perception study is ongoing. The aim of the study is to discover relationships between robot motion parameters, and people's subjective perception of the robot. In order to accomplish this, we aim to show various videos to the participants, while asking them to answer a questionnaire based on their perception of the robot. Various adjectives will be proposed to describe the robot and its motion in the video, and participants will rate the appropriateness of each one.

Results from our current and future experiments should allow us to further our understanding of Human-Robot Interaction, as well as give some insight into how Humans process various types of cues in interaction. We also aim to integrate the data and knowledge acquired during these studies into the design of a mobile robot's decision and navigation control algorithms, in order to allow designers to choose the way in which the robot navigates according to the influence it may have on people.



Two velocity profiles drawn from the robot motion corpus. The base motion consists in a linear acceleration, followed by a 300ms constant speed plateau, and linear deceleration. Here, we see two variants of the base motion (left: hesitant "increment" variant, right: jittery "saccade" variant).

Early-life stress affects Mongolian gerbil interactions with conspecific vocalizations in a sex-specific manner

Kate A. Hardy[a,b], Denise Hart[c], Merri J. Rosen[d,e]

[a]*Hearing Research Group, Department of Anatomy and Neurobiology, Northeast Ohio Medical University (NEOMED)*
[b]*Biomedical Sciences, Kent State University*
[c]*Hiram College*
[d]*Hearing Research Group, Department of Anatomy and Neurobiology, NEOMED*
[e]*Brain Health Research Institute, Kent State University*

Early-life stress (ELS) is known to affect cognition, learning, and emotional regulation. The mechanisms that induce ELS-derived dysfunction are broad and include alterations in anatomy and physiology during critical periods, a time at which the environment has a substantial impact on development.

Despite the existence of similar critical period mechanisms in the auditory system, it is unknown whether ELS affects simple sensory perception. This may be the case, as children from low socio-economic status environments (a proxy for ELS) have increased risk for impaired speech perception. Recent studies from our laboratory have indicated that ELS impairs gerbil auditory perception of temporally-varying features, such as those in gerbil vocalizations. Here we tested whether ELS affects gerbil behavioral responses to conspecific vocalizations.

Methods To induce ELS, Mongolian gerbil pups were intermittently maternally separated and restrained during a time window encompassing the critical period for auditory cortex maturation. These animals and age-matched controls were then tested as juveniles in a Y-shaped maze (A) with speakers at the ends of two arms to allow approach behavior, and a decoy speaker at the end of the third arm to allow avoidance behavior. This design allows animals to approach or withdraw from a sound source. The sound stimuli were exemplars from either alarm (B) or contact calls (C) from conspecifics. Time spent in each arm and in a central monitoring position was scored offline.

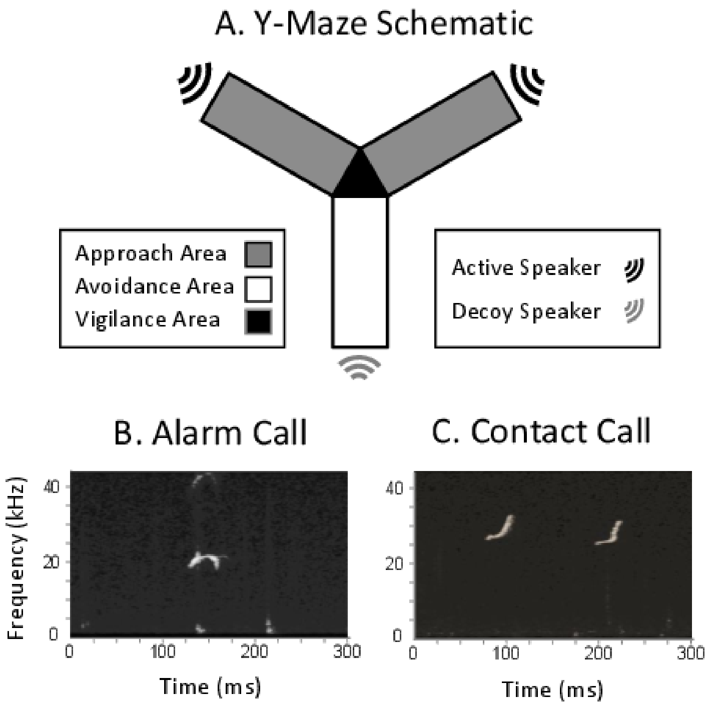
Results During alarm call playback, control gerbils increased time spent near the sound source and decreased time spent farthest away from the source, especially in males. Regardless of sex, ELS increased caution in the presence of alarm calls. Additionally, ELS males showed more avoidance of the alarm call while ELS females showed less avoidance relative to controls.

During contact call playback, all animals decreased time spent avoiding the sound. Regardless of treatment, males monitored more in the center than females. In females, ELS led to much stronger investigation relative to controls. Females also greatly decreased time spent monitoring. ELS did not alter contact call behavior in males.

To account for potential influences from higher order brain regions, we assessed behaviors related to cognition and anxiety. We found no treatment effects in any of our conventional behavioral tests, suggesting the effects of sex and ELS in the Vocal Preference Test were not due to altered anxiety or cognition, raising the possibility that the auditory system itself is altered by ELS.

Discussion These findings indicate that alarming stimuli may be more salient to male gerbils than female gerbils, but that this salience is reduced when males experience ELS. Further, ELS leads to strong investigation of contact calls in females only, which suggests either heightened interest in social interaction or possibly a lack of impulse control.

Implications Our results indicate that ELS influences behavioral responses to ethologically relevant sounds, which may arise from differences in auditory perception. These data are among the first to demonstrate auditory dysfunction emerging from early-life stress, and will contribute to understanding how ELS in children shapes communication.



Give a Robot a Voice

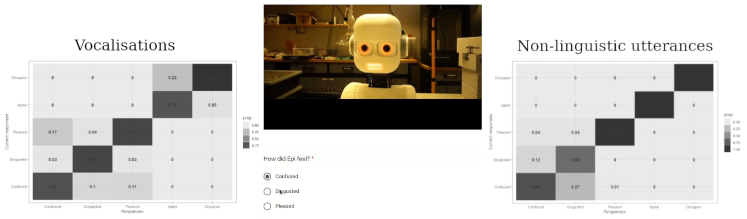
Pierre Klintefors[a], Simon Grendeus[a], Edvin Boyner[a], Alexander Pettersson[a]

[a]*Lund University*

This project aimed to investigate how non-verbal vocalizations, sometimes referred to as semantic-free utterances, could be implemented in a humanoid robot to facilitate communication in human-robot interaction. There are potential benefits with this form of communication over natural language. Computers are still limited in their ability to generate and parse spoken language. Further, if the participants in the interaction do not understand the language of the robot, the interaction might fall apart. Semantic-free utterances, while less precise in their informational value, potentially offer a universal form of communication that might be easier to implement. Previous research on semantic-free utterances has focused on "typical" robot sounds, like beeps and boops, referred to as non-linguistic utterances. However, these are potentially perceived as less ecologically valid. In contrast, communicative utterances based on sounds from the vocal tract, referred to as paralinguistic utterances, could be easier to interpret since humans evolved to interpret such sounds and because they are more congruent with a human-looking robot. Thus, we compared paralinguistic utterances with non-linguistic utterances in their ability to communicate emotion.

We collected non-verbal reactions for various affective categories from human participants, which were validated using the survey platform Prolific. This validation resulted in five categories of affective vocalisations: agreeing, disagreeing, pleased, disgusted and confused. Five prototypes for each category were then synthesized, using the R package Soundgen, based on the human recordings with the most accurate classification on Prolific. These paralinguistic utterances were then implemented in the humanoid robot Epi. The synthesized voice was modulated to have the fundamental pitch and formant frequencies within a child's vocal range. This was done to make the voice match the look of the robot and to lower the human participants' expectations of the social and cognitive abilities of the robot. The synthesised vocalisations were then tested and compared to non-linguistic utterances in a classification survey involving video clips of the robot.

When rated on a 5 point Likert-scale, the paralinguistic utterances were rated as less creepy, Mdn = 1, compared to the non-linguistic utterances, Mdn = 3, U = 140.5, p = .022. On the other hand, a two-sample Wilcoxon test showed that the classification accuracy of the paralinguistic utterances, Mdn = 84%, was lower compared to the non-linguistic utterances, Mdn = 96 %, U = 144, p = .034. However, this difference could be explained by a few poorly performing prototypes that lowered the average. There might also have been a practice effect when classifying the non-linguistic utterances. These results show that it is possible to implement paralinguistic utterances for human-robot interaction and that these could be perceived as more organic and less creepy than non-linguistic utterances. In the future, we plan to test the effectiveness of paralinguistic utterances in real life interactions, perhaps where the participants are not tied to set categories and can interact freely with the robot. We also plan to compare the utterances with other forms of communication, like natural language. This project was a multidisciplinary collaboration between engineering students and cognitive science students at Lund University.



How did Epi feel? *

◉ Confused

○ Disgusted

○ Pleased

Quietly angry: declared customer satisfaction vs. automatically detected emotion in contact center calls

Eric Bolo[a], Muhammad Samoul[a], Nicolas Seichepine[a], Mohamed Chetouani[a,b]

[a]*Batvoice AI*
[b]*Sorbonne Université*

Context and objectives In spite of digitalization, phone calls remain an essential communication channel in today's contact centers, but they are more difficult to analyze than written or form-based interactions. In an increasingly customer-centric business environment, gathering insight from interactions is needed to improve customer experience. To that end, companies have traditionally used surveys to gather feedback and gauge customer satisfaction, while AI-based tools have more recently been designed to automatically analyze the content of calls. This ongoing work compares both approaches via a case study using real-world call center data. We study the relationship between 1/ various affective indicators - arousal, valence, and a binary "anger" flag - produced by machine learning algorithms from customer speech and 2/ a satisfaction (CSAT) score optionally attributed by the customer as the call ends. In this abstract we report results for one affective indicator, namely anger.

Method We consider calls with one or more utterances classified as containing anger, and compare the CSAT scores of such calls with the scores for calls with no anger detected (total number of calls=118 048, with one or more anger utterances=9 341). We test the significance of observed differences with a Mann-Whitney U test. We further compare the proportion of detected anger in calls with high/low CSAT score using a z-test. Finally, we analyze the relationship between detected anger and whether a CSAT score was attributed by the customer, given that for attribution to occur, the customer agent must first transfer the customer to the questionnaire, and the customer must then decide to answer (number of calls with CSAT score = 55 838, with one or more anger utterances= 2 048). Main results We found the distributions of calls with detected anger and calls without to differ significantly with respect to CSAT score (Mann Whitney U: $p<0.001$). The results indicate that the proportion of calls with detected anger (0.18) to be respectively higher/lower in calls with low/high CSAT scores ($z=40.31$, $p<0.01$). Finally, we found that calls with detected anger had a significantly lower CSAT response rate ($z = 23.27$, $p<0.001$). These results coincide with the expectation that angry customers will report lower satisfaction. However, they also highlight a selection bias involved in CSAT scoring, since CSAT for calls containing anger appears to be under-reported.

Vocal response of the male serin, Serinus serinus, to interactive playback

Ana Mamede[a]

[a] *FCTUC, UC, PT*

Song overlapping and alternating in birds has been studied over the past few decades and, more recently, spawned some controversy over its communicative value. Proposed hypotheses to explain the function of this vocal behavior leads to quality and/or motivation signaling. Results of previous experiments showed that male serin react to songs with shorter inter-syllable intervals but not to frequency variations. We analyze vocal response to interactive playback attempting to rate alternate and overlap song stimuli influence. Analyzing overall differences in responses between experimental treatments, namely song length and interval between songs and syllables, we found a decrease in male song length with playback overlapping and alternating. On the other hand, the decrease in inter-syllabic range during alternating playback may indicate higher aggressiveness. The results suggest that singing in overlap and alternate can be considered a threat but male reaction is significantly different to both stimuli.

Variations and Behavioral Correlates of the Usage of Territorial and Threat Calls in a Tropical Songbird, the Pied Bush Chat (Saxicola caprata): A field study

Navjeevan Dadwal[a], Dinesh Bhatt[b], Vinay Kumar Seth[c]

[a]*Department of Life Sciences and Biotechnology, Chandigarh University, Punjab, India*
[b]*Avian Diversity and Bioacoustics Laboratory, Department of Zoology and Environmental Science, Gurukula Kangri University, Haridwar, Uttarakhand, India*
[c]*Department of Environmental Science, Faculty of Modern Studies, Uttarakhand Sanskrit University, Haridwar, Uttarakhand, India*

A large portion of avian behavior is expressed via their songs and calls. The song indeed dominates the significant part of avian vocalizations; however, there are some situations where a specific behavior needs to be expressed via calls rather than songs. For instance, a prey-predator interaction where the territory holders protect their nest against a potential predator or when an individual is trying to locate its mate/offspring via contact calls. In the present study, we devised a predator-prey interaction simulation where the nesting pair responded via territorial and threat calls. This field study was carried out in the natural habitats of a tropical songbird, the Pied Bush Chat, during their breeding seasons (Feb–May 2017-21), at Haridwar, Himalayan Foothills, India. The present study results indicated the two distinct call types used by the study species during their interactions with the decoy predator, territorial calls, and threat calls. According to the experimental setup, the display of threat calls was inevitable. However, we also recorded a huge sum of vocalizations which usually correspond to the territorial disputes (male rivalry or mate attraction). Therefore, it seems that the territory holders uttered these calls to show readiness to defend their nests and to display their reproductive quality. The combinations of both calls indicated that the acoustic response was directed not only to the decoy predator but it also had an element of the display of the male quality towards their mates. Keywords: Tropical songbird, Call repertoire, Pied Bush Chats

Universal emotional translators: a machine learning adventure to explore acoustic correlates of emotional valence in domestic animals

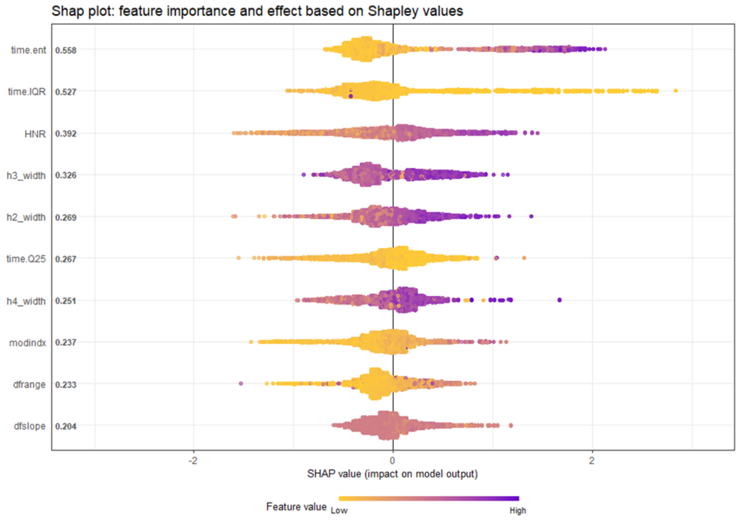
Romain Adrien Lefèvre[a], Ciara Cleo Rose Sypherd[a], Elodie Floriane Mandel-Briefer[a]

[a]*Behavioural Ecology Group, Section for Ecology & Evolution, Department of Biology, University of Copenhagen, 2100 Copenhagen Ø, Denmark*

Emotions form an integral part of an animal's behaviour as they facilitate responses elicited by internal (e.g., sensations) or external events (e.g., a change of situation) of significance for the organism. While positive emotions trigger approach behaviours towards the stimuli that enhance an animal's fitness, negative emotions elicit avoidance behaviour when encountering fitness-threatening stimuli. These short-lived affective reactions are characterized by their valence (negative vs. positive) and their arousal (calming vs. exciting), and are reflected in neural, physiological, behavioural, and cognitive changes. Since the acoustic structure of vocalizations is linked to the context in which they are produced, vocal signals may be a promising immediate and non-invasive indicator of animals' affective states.

To this date, vocal indicators of the emotional valence experienced by the sender have been poorly investigated. Among other factors, the lack of easily reproducible statistical methods and technical limitations in selecting appropriate features to be analysed may have severely constrained research findings. Automation tools and services provided by recent advances in machine learning methods and signal processing, computer science, and information technology motivated the emergence of computational bioacoustics, a cross-disciplinary discipline combining techniques and knowledge from computer science biology and acoustics. To the best of our knowledge, despite some advances in acoustic ecology and conservation, automatic acoustic classification of emotional valence has never been explored in multiple species simultaneously. However, if the vocal expression of emotions has been conserved throughout evolution, as Darwin suggested (1956), direct between-species comparisons using the same set of acoustic indicators should be possible.

The present study aimed at discriminating emotional valence based on vocalisations produced during different positive and negative contexts across cows, horses, pigs, wild boars, and goats. To provide valid and consistent indicators of emotional expression to be explored in future studies, we aimed at using a scalable decision-tree-based algorithm named eXtreme Gradient Boosting (XGBoost) and rank acoustic features importance based on their Shapley value. Our model reached 86% of accuracy with 87% of negative calls being correctly classified against 75% for the positive calls. Results showed that high values in the time entropy (R = 0.91), harmonics-to-noise ratio (R = 0.85), mean bandwidth of the second (R = 0.73), third (R = 0.62) and fourth harmonics (R = 0.59), in the range of the dominant frequency measured across the acoustic signal (R = 0.61), and in the modulation index (R = 0.60) were more likely to predict positive emotions, while high values in the first quartile time (R = -0.41) and in the interquartile time range of energy (R = -0.31), were more likely to predict negative emotions, although being moderately associated. While the domain of animal welfare could benefit from automatic processing of vocal expression, we believe this study's outcome will shape a first basis to contribute to developing a universal and non-invasive tool to assess animals' affective states based on modulations of their vocal signals.


Shap plot: feature importance and effect based on Shapley values
time.ent 0.558
time.IQR 0.527
HNR 0.392
h3_width 0.326
h2_width 0.269
time.Q25 0.267
h4_width 0.251
modindx 0.237
dfrange 0.233
dfslope 0.204
-2
0
2
SHAP value (impact on model output)
Feature value
Low
High

Dogs react to the social valence of conspecific but not of human vocalizations

Tamás Faragó[a], Irene Rojas Atares[b], Paula Pérez Fraga[b], Lilla Kocsis[a], Morgane Audiguier[a], Soufiane Bel Rhali[a], Attila Andics[a,b]

[a]*Department of Ethology, Eötvös Loránd University, Budapest, Hungary*
[b]*MTA-ELTE 'Lendület' Neuroethology of Communication Research Group, Hungarian Academy of Sciences – Eötvös Loránd University (ELTE), Budapest, Hungary*

Besides the emotional valence that reflects positive or negative inner state, vocalizations also encode social valence: this basic biological meaning reflects whether the listeners' adaptive reaction would be approaching (positive) or avoiding (negative) the caller. Emotional and social valence can conflict: while agonistic calls both evoke avoidance and indicate negative inner state of the caller, distress calls reflect a negative inner state but might also carry a positive social meaning evoking approach from listeners. Due to simple general rules of acoustic valence and arousal encoding, emotion communication is possible not just within but across species too, but less is known about the social valence in this regard. Dogs living in the human environment and adapted to interact with them on a daily basis are an excellent species to study cross-species emotion communication. We were curious whether dogs' approach/avoidance reactions to emotionally loaded calls were evoked by emotional or social positivity/negativity, and whether these were affected by the caller species. Thus, to test how the emotional and social valence of con- and heterospecific calls affect dogs' reactions, we played back dog and human agonistic (growls/roars) distress (whines/cries) and playful/comfort calls (growls, grunts, moans/laughs) from a hidden speaker to 110 dogs in 2*3 groups, respectively. We coded their initial approach or avoidance reaction and compared their latencies and occurrences between groups. Dogs that heard dogs' distress or playful/comfort calls approached the speaker more likely and sooner, while those that heard agonistic growls reacted more likely and sooner with withdrawal. Dogs' approach and avoidance reactions thus reflect the basic biological meaning (social valence) rather than the emotional valence of conspecific calls. To human calls however, we found no effect of social or emotional valence on the approach or avoidance reactions, dogs likely approached the speaker even in case of aggressive roars. This difference between the reactions to dog and human calls might be due to dogs' strong general preference towards humans, overshadowing the effect of the emotional information. Further tests using other species' calls and artificial emotion expressing sounds will clarify and unravel the cross-species effects of emotional and social valence.

The Efficacy of Choice and Control in Developing Human-Dolphin Communicative Transactions

Diana Reiss[a], Rita Kanagat[a], Marcelo Magnasco[b]

[a]*Hunter College, City University of New York*
[b]*The Rockefeller University*

We designed and implemented an interactive touchscreen system for dolphins in human care, to provide them with some degree of choice and control over obtaining different objects, activities and video stimuli, and as a means of two-way communication with humans. Obtaining an object or activity required the dolphins to touch a visual symbol or 2-D representation of that object or activity. A novel computer synthesized whistle (CSW) would be paired with the visual stimuli to enable us to further investigate vocal learning and productive use of novel whistles by dolphins. In preparation for the touchscreen interactions, we conducted sessions to teach dolphins that they had some degree of choice and control. To accomplish this, we conducted a total of 15 sessions, 10-15 minutes in duration, from February-May 2018 with two male captive born bottlenose dolphins (Tursiops truncatus) at the National Aquarium in Baltimore, MD. In these sessions, toys objects (balls and small Polyform G Series boat fenders) were held just above the pool surface by an experimenter. A touch by a dolphin to an object resulted in a specific CSW being played into the dolphin pool as the object was given to the dolphin. The dolphins touched the offered objects and thus were exposed to the corresponding CSWs a total of 95 times, showing a strong preference for balls (n=89) over boat fenders (n=6). Although signal attribution to a specific dolphin was not feasible, we recorded rapid and spontaneous vocal imitation (n=2) and production (n=11) of facsimiles of the CSW "ball" signal by the dolphins following a minimal number of prior exposures (n=9) to the model sound. The majority of the dolphins' facsimiles were productions (defined as facsimiles occurring ¿ 1.0 s after the CSW); the dolphins emitted fewer imitations, defined at <1.0 s following the CSW with no intervening whistles. Notably, during the latter sessions, one of the dolphins initiated bringing the balls back to the experimenter, either tossing or gently pushing them back and engaging in more interactive ball transactions. In at least one case, the dolphin produced the ball facsimile concurrent with tossing the ball to the experimenter. These findings are consistent with previous studies using similar procedures conducted in our lab (Reiss & McCowan, 1993). Our results suggest that providing dolphins with choice and control in obtaining objects paired with novel whistles during dolphin-human social transactions results in spontaneous vocal imitation and productive use of CSW facsimiles by participating dolphins, may be instrumental in developing learned associations, and ultimately may facilitate interspecies communication.

Why Socially Interactive Instruction Is Necessary for Nonhuman Acquisition of Communicative Competence

Irene M. Pepperberg[a]

[a] *The Alex Foundation*

I will review the Model/Rival (M/R) technique that has been used to establish referential interspecies communication with Grey parrots (*Psittacus erithacus*). I will describe the original format developed by Todt, the influence of other forms of observational learning outlined by other researchers, and the adaptations that I devised. I will describe how my undergraduate trainers and I isolated the various components that constitute the technique and explain how each is necessary, but how only the combination of all components is sufficient for successful implementation—and how improper implementation can lead to failure. The findings should be of interest to any human intending to teach a referential communicative code to nonhumans.

Behavioral experiments for gathering labeled animal vocalization data

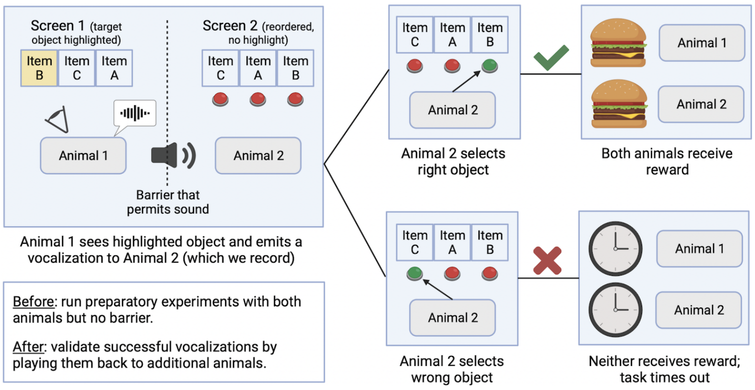
Andrew Ross[a], Su Jin Kim[b]

[a] *Courant Institute of Mathematical Sciences, New York University*
[b] *Department of Psychological and Brain Sciences, Johns Hopkins University*

Across taxa as diverse as insects, amphibians, birds, and mammals, animals use vocalizations as functional references to components of their environment (Pika et al. 2018). While some referential vocalizations are known at birth and never structurally modified, others are learned socially and vary across populations and time (Janik et al. 2000). For example, captive bottlenose dolphins can learn to associate vocalizations with novel referents such as objects or conspecifics, and obey syntactic rules (Herman et al. 1984, Janik et al. 2013, King et al. 2013). Such studies raise questions of whether wild populations' vocalizations share these and other features with human language, and whether it might be possible to translate between them.

Motivated by such questions as well as the success of statistical and machine learning (ML) techniques in processing human languages, many works have started applying similar methods to animal vocalization datasets (Coffey et al. 2019, Kohlsdorf et al. 2020, Poupard et al. 2019, Vradi et al. 2021). However, the datasets these works analyze are generally unlabeled, which rules out applying the best-performing supervised methods for translation. Although unsupervised methods for modeling and translating between human languages are improving quickly (Artetxe et al. 2018, Devlin et al. 2019), there is no guarantee that they will work for non-human animal vocalizations, and also no simple way of evaluating their accuracy. If we had a reasonable amount of labeled data, or vocalizations we knew were references to well-defined objects or concepts, we could apply (and evaluate) a much wider array of machine translation techniques, including semi- and fully supervised methods.

In this work, we propose a system for designing behavioral experiments to gather labeled animal vocalization data. We describe two experiments: (1) two animals are rewarded when one can efficiently communicate to the other the identity of a highlighted object in an unordered collection on a digital display, and (2) a single animal is played back recordings of another animal from a previous trial, allowing for validation that vocalizations reliably identify the same objects in new contexts. Highlighted and surrounding objects can be varied to test whether animals have existing "vocabulary" for specific concepts, and the procedure can be automated to allow for efficient data collection and validation over larger populations. These experiments have the potential not only to generate large volumes of labeled data for future analysis, but also to shed light on whether, when, and how different species of animals communicate referentially.

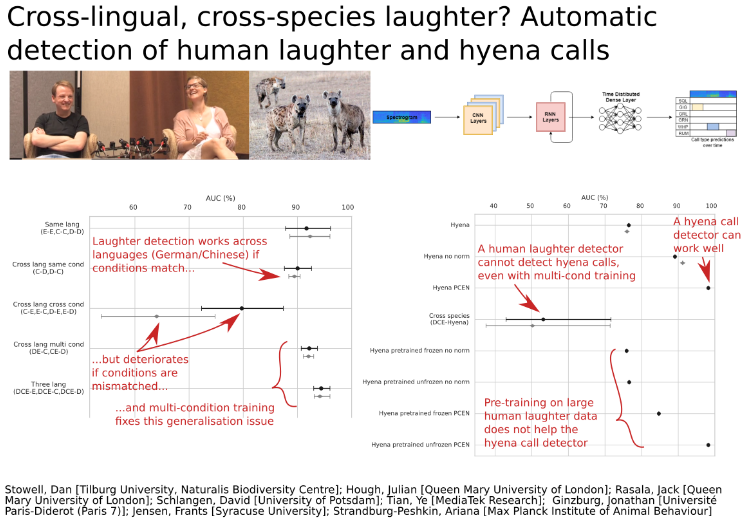

Screen 1 (target object highlighted)
Item B
Item C
Item A
Screen 2 (reordered, no highlight)
Item C
Item A
Item B
Animal 1
Animal 2
Barrier that permits sound
Animal 1 sees highlighted object and emits a vocalization to Animal 2 (which we record)
Before: run preparatory experiments with both animals but no barrier.
After: validate successful vocalizations by playing them back to additional animals.
Item C
Item A
Item B
Animal 2
Animal 2 selects right object
Animal 1
Animal 2
Both animals receive reward
Item C
Item A
Item B
Animal 2
Animal 2 selects wrong object
Animal 1
Animal 2
Neither receives reward; task times out

Cross-lingual, cross-species laughter? Automatic detection of human laughter and hyena calls

Dan Stowell[a], Julian Hough[b], Jack Rasala[b], David Schlangen[c], Ye Tian[d], Jonathan Ginzburg[e], Frants Jensen[f], Ariana Strandburg-Peshkin[g]

[a] *Tilburg University, Naturalis Biodiversity Centre*
[b] *Queen Mary University of London*
[c] *University of Potsdam*
[d] *MediaTek Research*
[e] *Université Paris-Diderot (Paris 7)*
[f] *Syracuse University*
[g] *Max Planck Institute of Animal Behaviour*

Mammals, including humans, have basic vocalisation (call) types that are considered ”universal” across a given species. In humans this is true of sounds such as laughter, which are believed to have originated earlier in mammalian evolution before the speciation of Homo sapiens. Inspired by these factors, as well as the ability of machine learning to make use of the structure in separate but related tasks (so-called transfer learning/multi-task learning), we investigate the development of an acoustic detector for laughter sounds across three different human languages (German, English, Chinese) in conversational data as well as for eight types of call of the spotted hyena (Crocuta crocuta), using a common deep learning architecture. In accord with the universality of laughter, we find strong performance of laughter detectors in cross-lingual situations, providing that differences in dataset conditions are accounted for. However, incorporating human datasets did not improve detection of hyena sounds. We also uncovered issues in data normalisation which contradict received wisdom in machine learning, which for hyena data could be remedied by applying a recently-proposed ”per-channel energy normalisation” method. Nevertheless, the same deep learning architecture can be used for robust high quality detection separately within each species, human and hyena.



Cross-lingual, cross-species laughter? Automatic detection of human laughter and hyena calls

Laughter detection works across languages (German/Chinese) if conditions match...

...but deteriorates if conditions are mismatched...

...and multi-condition training fixes this generalisation issue

A hyena call detector can work well

A human laughter detector cannot detect hyena calls, even with multi-cond training

Pre-training on large human laughter data does not help the hyena call detector

Stowell, Dan [Tilburg University, Naturalis Biodiversity Centre]; Hough, Julian [Queen Mary University of London]; Rasala, Jack [Queen Mary University of London]; Schlangen, David [University of Potsdam]; Tian, Ye [MediaTek Research]; Ginzburg, Jonathan [Université Paris-Diderot (Paris 7)]; Jensen, Frants [Syracuse University]; Strandburg-Peshkin, Ariana [Max Planck Institute of Animal Behaviour]

Index of Authors

VIHAR 2021

http://vihar-2021.vihar.org/

9 782956 202929