Proceedings of the
2nd International Workshop on
Vocal Interactivity in-and-between Humans, Animals and Robots

VIHAR 2019

London, UK, 29-30 August 2019

Published by:
Ricard Marxer

ISBN: 978-2-9562029-1-2

Credits:
Editors: Angela Dassow, Ricard Marxer, Roger K. Moore, Dan Stowell
Cover photo: Colin, The Palace of Westminster from the dome on Methodist Central Hall, Recolored by Ricard Marxer, CC BY-SA 4.0,
https://commons.wikimedia.org/wiki/File:Palace_of_Westminster_from_the_dome_on_Methodist_Central_Hall.jpg
Proceedings assembled by: Ricard Marxer

Workshop took place in London, UK — August 29-30, 2019

Published online at http://vihar-2019.vihar.org/ — September 23, 2019

Workshop Organisation

Organising Committee

Dan Stowell Queen Mary University of London, UK

Angela Dassow Carthage College, US

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

Julian Hough Queen Mary University of London, UK

Roger K. Moore University of Sheffield, UK

Elisabetta Versace Queen Mary University of London, UK

Emmanouil Benetos Queen Mary University of London, UK

Jessie Wand The Alan Turing Institute, UK

Scientific Committee

Elodie Mandel-Briefer ETH Zürich

Roger Moore University of Sheffield

Julie Oswald University of St. Andrews

Serge Thill Radboud University

Robert Eklund Linköping University

Dan Stowell Queen Mary University of London

Emmanouil Benetos Queen Mary University of London

Julian Hough Queen Mary University of London

Kaspar Althoefer Queen Mary University of London

Ildar Farkhatdinov Queen Mary University of London

Matthew Purver Queen Mary University of London

Elisabetta Versace Queen Mary University of London

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

Workshop supported by

Queen Mary University of London

The Alan Turing Institute

PeerJ

Conference Program

Keynotes

Posters

Oral session 1

Oral session 2

Oral session 3

Proc. 2nd Intl. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), London, UK, 29-30 Aug 2019

Studying bats to shed light on speech and language

Sonja Vernes

Abstract

Vocal production learning - the ability to modify vocal signals based on auditory feedback - is an essential component of human speech and spoken language. Comparative studies of vocal learning in animals will be valuable for understanding the biology underlying this trait. Bats are highly social animals that have developed sophisticated vocal systems for navigation and communication. Their capacity for vocal learning, small size, amenability to neurogenetic manipulations, and the long history of studying the neuroethological traits in bats, makes them an excellent system to model vocal learning. I will present work including highly controlled behavioural paradigms, genomic approaches, and neuro-molecular studies that aim to dissect out the biological mechanisms underlying vocal learning in bats. These approaches aim to show how neuro-genetic mechanisms contribute to a complex behaviour like vocal learning and may ultimately shed new light on the biology and evolution of human speech and language.

Biography

Sonja Vernes is the leader of the Neurogenetics of Vocal Communication group. Sonja obtained her PhD in Neurogenetics from The University of Oxford and is currently a 'Max Planck Research Group Leader' holding a W2 position in the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. She is also an affiliated Principal Investigator at the Donders Institute for Brain, Cognition and Behaviour, Radboud University. Her research interests focus on the genetic bases of human language and the use of bats as a model for vocal communication that can inform the biological encoding and evolution of this trait. Sonja is a founding director of the Bat1K genome sequencing consortium (http://www.bat1k.com) and a FENS-Kavli Network of Excellence Scholar (http://fenskavlinetwork.org).

How machines learn to talk. Machine Learning for Conversational AI

Verena Rieser

Abstract

Conversational Artificial Intelligence (AI) makes interaction with machines possible through voice and text platforms, and is a rapidly growing area of research and commerce. These Conversational AI Systems have experienced a revolution over the past decade, moving from being completely handcrafted to using data-driven machine learning methods. In this talk, I will review these current developments including my work on using reinforcement learning and deep learning models, and evaluate these methods in the light of recent results from two large-scale studies: First, I will summarise results from the End-to-End NLG Challenge for presenting information in closed-domain, task-based dialogue systems. Second, I will report our experience from experimenting with these models for generating responses in open-domain social dialogue as part of the Amazon Alexa Prize challenge.

Biography

Verena Rieser is a Professor in Computer Science at Heriot-Watt University, Edinburgh, where she is affiliated with the Interaction Lab. Verena holds a PhD from Saarland University (2008) and worked as a postdoctoral researcher at the University of Edinburgh (2008-11). Her research focuses on machine learning techniques for spoken dialogue systems and language generation, where she has authored almost 100 peer-reviewed papers. She has served as area chair for ACL for both generation and dialogue. For the past two years, Verena and her group were the only UK team to make it through to the finals of the Amazon Alexa Prize.

Interpersonal speech-based interaction

Mohamed Chetouani

Abstract

Analysing human behaviours during social interactions requires to explicitly take into account all the participants. By doing so, researchers in various domains such as psychology, psychiatry, neuroscience, affective computing and human-machine interaction have developed methodologies and tools for analysing and modelling human-human and human-machine interpersonal interactions. Within this context, the challenge is to develop machines that can decode social interaction by assessing individual and interpersonal dynamics of behaviours, with the goal of analysing and predicting human's implicit social signals and emotional expressions. In this talk, we will show how jointly analysing individual and inter-individual behaviours offers the opportunity to capture relevant makers of pathologies in particular in autism spectrum disorders. We will present our works on (i) modelling parent-infant interaction using non-verbal features and the role of infant-directed speech in engagement, (ii) computational models of multimodal emotional contagion and (iii) applications in robotics and services. Analysing human behaviours during social interactions requires to explicitly take into account all the participants. By doing so, researchers in various domains such as psychology, psychiatry, neuroscience, affective computing and human-machine interaction have developed methodologies and tools for analysing and modelling human-human and human-machine interpersonal interactions.

Biography

Mohamed Chetouani is Full Professor in Signal Processing and Machine Learning for the Human Machine Interaction Institute for Intelligent Systems and Robotics (CNRS UMR7222) at Sorbonne University, and CSO at Batvoice Technologies. He received the M.S. degree in Robotics and Intelligent Systems from the UPMC, Paris, 2001. He received the PhD degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling (UK). Prof. Chetouani was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona (Spain). He is currently a Full Professor in Signal Processing, Pattern Recognition and Machine Learning at the UPMC. His research activities, carried out at the Institute for Intelligent Systems and Robotics, cover the areas of social signal processing and personal robotics through non-linear signal processing, feature extraction, pattern classification and machine learning. He is also the co-chairman of the French Working Group on Human- Robots/Systems Interaction (GDR Robotique CNRS) and a Deputy Coordinator of the Topic Group on Natural Interaction with Social Robots (euRobotics). He is the Deputy Director of the Laboratory of Excellence SMART Human/Machine/Human Interactions In The Digital Society. Website: http://people.isir.upmc.fr/chetouani/

Evolution of the Speech Apparatus: Monkey Vocal Tracts are Speech Ready

Tecumseh Fitch

Abstract

The capacity to produce a sufficient set of acoustically distinct phonemes lies at the heart of our ability to communicate linguistically. For four decades, the inability of nonhuman primates to produce human speech sounds has been claimed to stem from limitations in their vocal tract anatomy, a conclusion based on plaster casts made from the vocal tract of a monkey cadaver. We used x-ray videos to quantify vocal tract dynamics in living macaques during vocalization, facial displays, and feeding. We demonstrate that the macaque vocal tract could easily produce an adequate range of speech sounds to support spoken language, showing that previous techniques based on postmortem samples drastically underestimated primate vocal capabilities. Our findings imply that the evolution of human speech capabilities required neural changes rather than modifications of vocal anatomy. Macaques have a speech-ready vocal tract but lack a speech-ready brain to control it.

Biography

Tecumseh Fitch is the head of the Department of Cognitive Biology at the University of Vienna. His research has followed two main paths: the evolution of cognition, and the bioacoustics of vocal production. He studies both topics from a broad comparative perspective. Initially trained in evolutionary and behavioral biology, he did a PhD in cognitive science at Brown University, after deciding to study language evolution from a biological perspective. He taught in both biology and psychology departments at Harvard and St Andrews before moving to Vienna in 2009 to co-found the new Department of Cognitive Biology, within the Life Sciences Faculty at the University of Vienna. He has recently published a book 'The Evolution of Language' (CUP, 2010) and is a recipient of an ERC Advanced Grant. He has worked on a wide variety of species, including whooping cranes, deer, elephants, dogs and many primate species, and much of his work features direct experimental comparisons of such species with human beings.

Play Vocalizations in White-handed Gibbons (*Hylobates lar*)

Angela Dassow[1]

[1]**2001 Alford Park Drive, Kenosha, Wisconsin, United States**

ABSTRACT

This work explores a previously undocumented call produced by white-handed gibbons (*Hylobates lar*) during a specific type of behavioral activity. The sound produced is akin to a goat bleat and is structurally unique from the rest of the gibbons' vocal repertoire. The behavior, which is largely playful in nature, is modulated by the utterance of this unique call. While the development of the call production is unclear, the function of the call appears to be a stop signal to prevent any physical harm to each other during play bouts. This call varies in amplitude and the loudness of the call is correlated with the forcefulness of the encounter. Early, quiet versions of this call are a precursor to later, louder calls when lighter play behavior escalates to rougher play behavior.

INTRODUCTION

The majority of the species in the Hylobatidae family are known for their ability to produce sex-specific vocalizations (Geissmann, 2002; Marshall and Marshall, 1976). Females produce a sequence known as a great call, while males produce a sequence known as a coda (Raemaekers and Raemaekers, 1985). The function and structure of these calls and the rest of the gibbons' vocal repertoire has been the subject of numerous studies (Clarke et al., 2006, 2012, 2015; Cowlishaw, 1992; Bartlett, 2003; Marler and Mitani, 1989; Raemaekers and Raemaekers, 1984). These studies have revealed distinct, stereotypical waveforms that are commonly heard in wild and captive gibbon populations. Though the vocal repertoire has been well studied in *H. lar*, a unique and functionally important call type has been overlooked until recently. That this call type is so distinct from the rest of the vocal repertoire, suggests that a re-examination of wild populations is needed to further understand the vocal interactivity between closely related individuals. Additionally, the suggested function of this call may be evolutionarily conserved across all Hylobatids and merits further study.

METHODS

To date, these unique calls have only been recorded in a single, related pair of captive gibbons. The relationship between these gibbons and the methods used to capture this call and associated behavior are described in the following sections. Though this call has yet to be recorded in another pair of gibbons, the frequency with which the current pair utilizes this call type suggests that further examination of individuals housed in a similar family dynamic will reveal additional support for this call as a part of the regular vocal repertoire.

Data collection

Zoo gibbon vocalizations were gathered from August 2012 to September 2013 at the Racine Zoological Society in Racine, Wisconsin, United States. This zoo houses a father-daughter pair of adult white-handed gibbons. The father was born in the wild and the daughter was born in captivity. Both animals were habituated to the presence of the observer and recording equipment for a minimum of one hour prior to recording each day.

Recording equipment was set up outside of the gibbon enclosures to record their natural vocalizations. The recording equipment included a Vidpro XM-55 Condenser Shotgun Microphone, a Blue Yeti Pro USB Condenser Microphone, both recording at 44.1 kHz, and a shock mount to reduce ambient noise interference. Recordings were taken 2-3 times per week for approximately 7 hours per day. During this time, the gibbons were free to move around their enclosure and were not exposed to any toxic agents

or restrained for any invasive medical procedures. These recordings were used to establish the vocal repertoires of the captive gibbons. This work was approved by the University of Wisconsin-Madison's Research Animal Resource Center (RARC) and Institutional Animal Care and Use Committee (IACUC protocol number is L00452-0-08-12), Gibbon Species Survival Plan (SSP) Coordinator and Director of the Racine Zoological Society which is an Association of Zoos and Aquariums (AZA) accredited zoo.

RESULTS

Twenty-eight hours of vocalizations and nineteen hours of video footage were obtained from over 450 hours of observations.

Description of bleats

A total of 74 bleats were isolated from the Racine Zoological Society pair that range in duration from .4-1.8 seconds. The average minimum frequency is 595 +/- 23.5 Hz and the average maximum frequency is 9,891.5 +/- 277.7 Hz. Figure 1 provides a spectral view of the complexity of a single bleat uttered by the male gibbon.

Description of behavior

The initial behavior associated with this call moves from rougher grooming of one another to both gibbons locking their hands and feet together. Upon grasping one another, they roll around and bare their teeth. During this time, the behavior moves from a more playful nature to a more aggressive, but still playful, nature and the more submissive individual utters the bleating sound. As the roughness of the play behavior escalates, the amplitude of this call increases until the aggressor backs down and moves away.

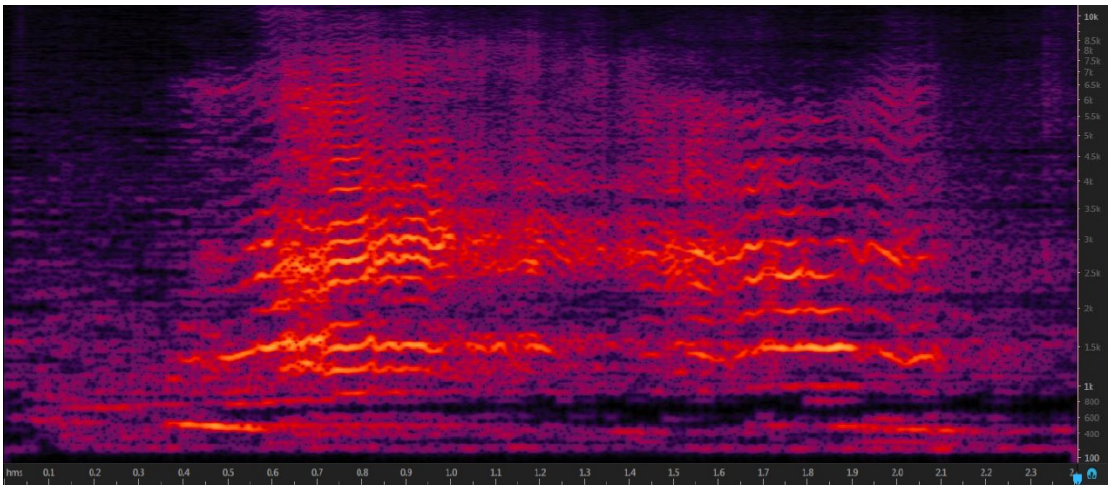

Figure 1. A single bleat call produced by the male gibbon at the Racine Zoological Society.

DISCUSSION

A previously undescribed sound from a father-daughter pair of *H. lar* at the Racine Zoological Society was recorded and described. The father, who was born in the wild, and the daughter, who was born in captivity, produce a quiet call for a very brief period of time daily. This bleat-like call is so quiet that it is likely only conveying potential information content to individuals within 3-4 meters of the vocalizer. The function of such a call is likely to inform the aggressor that the interaction has escalated too far. This call does vary in amplitude and the loudness of the call is correlated with the aggressiveness of the interaction.

The origin of this call is still unclear, but there are several possibilities for why this has not been observed previously. First, the amplitude of this call is very low and as such an observer would have to be within 3-4 meters of the gibbons to hear it. In the wild, this is highly unlikely. Second, unlike most of their other calls, this bleat was typically observed either mid-day or in the afternoon. This is generally not the best time to record in the wild due to increased ambient noise and decreased sound transmission

quality from humidity and warmer air (Larom et al., 1997). Third, this call may be a function of familial play or aggression (Cowlishaw, 1992). In the wild, the offspring would have dispersed to a new territory away from their parents.

Describing these bleats as novel has its drawbacks, especially when the categorization is based off of a data set that is as small as one pair of gibbons. It is unlikely that anything is damaged in their supralaryngeal vocal tract because the rest of their vocal repertoire is the same as the wild gibbons. Additionally, a large number of experts across several other zoos and sanctuaries have been consulted and they have confirmed that they have not observed this behavior or vocalization. These experts, which are the main caretakers of their respective gibbons, were given a video of the Racine gibbons bleating and fighting and they were asked whether this behavior or sound was found in their pairs of gibbons. Currently no observations of the call or behavior which leads up to this call have been made with 40 other gibbons under the care of 5 other zoos and sanctuaries. One other pair of gibbons does engage in similar chasing behavior and unusual call production though how similar the sounds are to one another will require further research. One contributing factor that may be relevant to why the Racine gibbons consistently exhibit play behavior and vocalizations, but most of the other gibbons do not is the fact that they are related to one another whereas most of the other pairs are not. The one pair that does appear to behave somewhat similarly is a pair of half-brothers who have also been housed together for several decades. Given the rarity of housing related gibbons together into adulthood, it would not be expected to witness such behavior. Functionally, it is plausible that all gibbons can produce these calls and behavior and this familial play is part of normal gibbon development. As individuals reach maturity, this play behavior may become more physical and it could mark a point in which the offspring need to leave the group and search for their own territory. In captive settings, the ability to disperse is restricted and therefore this behavior is more persistent. Future work should focus on related pairs of gibbons to further test the hypothesis that this behavior is typical play behavior between related individuals.

ACKNOWLEDGMENTS

I would like to thank the director and staff at the Racine Zoological Society for their support in conducting this research. I would also like to thank the Computer Science Department and Zoology Department at the University of Wisconsin-Madison and Carthage College for funding support.

REFERENCES

Bartlett, T. Q. (2003). Intragroup and intergroup social interactions in white-handed gibbons. *International Journal of Primatology*, 24(2):239–259.

Clarke, E., Reichard, U. H., and Zuberbühler, K. (2006). The syntax and meaning of wild gibbon songs. *PloS one*, 1(1):e73.

Clarke, E., Reichard, U. H., and Zuberbühler, K. (2012). The anti-predator behaviour of wild white-handed gibbons (hylobates lar). *Behavioral ecology and sociobiology*, 66(1):85–96.

Clarke, E., Reichard, U. H., and Zuberbühler, K. (2015). Context-specific close-range "hoo" calls in wild gibbons (hylobates lar). *BMC evolutionary biology*, 15(1):56.

Cowlishaw, G. (1992). Song function in gibbons. *Behaviour*, 121(1-2):131–153.

Geissmann, T. (2002). Duet-splitting and the evolution of gibbon songs. *Biological Reviews*, 77(1):57–76.

Larom, D., Garstang, M., Payne, K., Raspet, R., and Lindeque, M. (1997). The influence of surface atmospheric conditions on the range and area reached by animal vocalizations. *Journal of experimental biology*, 200(3):421–431.

Marler, P. and Mitani, J. C. (1989). A phonological analysis of male gibbon singing behavior. *Behaviour*, 109(1-2):20–45.

Marshall, J. T. and Marshall, E. R. (1976). Gibbons and their territorial songs. *Science*, 193(4249):235–237.

Raemaekers, J. J. and Raemaekers, P. M. (1984). The ooaa duet of the gibbon (hylobates lar). *Folia Primatologica*, 42(3-4):209–215.

Raemaekers, J. J. and Raemaekers, P. M. (1985). Field playback of loud calls to gibbons (hylobates lar): territorial, sex-specific and species-specific responses. *Animal Behaviour*, 33(2):481–493.

Robot-pet vocal interactions: Domestic chicks as a model system

Elisabetta Versace[1,2], Michael Mcloughlin[1,3], Joshua Brown[3,5], Dan Stowell[2,3], Ildar Farkhatdninov[2,3,5], Kaspar Althoefer[2,3,4,5]

[1] School of Biological and Chemical Sciences, Department of Biological and Experimental Psychology, Queen Mary University of London, London, United Kingdom
[2] Alan Turing Institute, British Library, 96 Euston Road, London, United Kingdom
[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom
[4] School of Engineering and Material Science, Queen Mary University of London, London, United Kingdom
[5] Centre for Advanced Robotics @ Queen Mary, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Elisabetta Versace
Mile End Road 327, London, E1 4NS, United Kingdom
Email address: e.versace@qmul.ac.uk

Abstract

The use of socially engaging robots with functions as different as education, rehabilitation and companionship is spreading in everyday life to assist human beings. The case of artificial pets – robotic agents that exhibit pet-like behaviour – is peculiar in the sense that the assistance provided by artificial pets is the attachment elicited in the human agent towards the robot, rather than a service provided to the partner. This social bond is sustained by three features: artificial pets appear to act autonomously (to possess animacy, the property of being alive), they respond to the partner's actions (with social engagement) and they depend on the partner. Here we use vocal interactions between artificial agents and domestic chicks (*Gallus gallus*) to investigate the role of animacy and social engagement in establishing social attachment between artificial pets and non-human animals. Domestic chicks exhibit several advantages as a model system: they promptly attach to artificial objects through the mechanism of filial imprinting, they are precocial and can be easily tested soon after birth, they are spontaneously attracted by cues of animacy, they have a strong social motivation, they are a model of physiological development and neurodevelopmental disorders such as autism, they are a species with large economical relevance.
This approach has potential applications (a) for neuroscience and biomedical research interested in the development of social attachment and vocal interaction (b) in farming (c) for the development of robots that assist pets in the absence of their owners.

The acoustic correlates of *Aegyo* (애교) speaking style in South Korea

Ji-eun Kim[1], Carolina Baslino[2], and Volker Dellwo[3]

[1] Department of Korean Linguistics and Literature, Seoul National University, Seoul, South Korea
[2] English Department, University of Zurich, Zurich, Switzerland
[3] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Corresponding Author:
Ji-eun Kim[1]
Nambusunhwanro 218, 1, Seoul, 08787, South Korea
Email address: smart173@snu.ac.kr

Abstract

It is unclear why some voices are perceived as more attractive than others. Here we investigated the acoustic correlates of *Aegyo* (애교), a popular Korean speaking style used to appeal others by enhancing one's 'cuteness'. Fourteen Seoul Korean speakers (8F, 6M) were recorded uttering numbers from one to ten, in both native Korean and Sino Korean words. Pitch, intensity, and duration were measured and statistically analyzed. We found that Aegyo speaking style is phonetically more variable compared to non-Aegyo conversational speech. Aegyo significantly increased pitch range ($\chi^2(1)=86.966$, p<2.2e-16***), mean pitch ($\chi^2(1)=6.805$, p=0.0091**), intensity range ($\chi^2(1)=5.354$, p=0.0207*), and duration ($\chi^2(1)=63.675$, p=1.457e-15***). This is in agreement with previously reported auditory impressions. In the future it will be important to understand the perceptual effects of Aegyo on attractiveness in listeners and non-listeners of Korean.

Introduction

Defining what makes a voice attractive or not is a goal long pursued but still far from being achieved. Although voice attractiveness has become an interdisciplinary focus of interest, there has been a considerable amount of research done from an evolutionary approach (for a detailed review, see Pisanski & Feinberg, 2019), with many studies exploring sexually dimorphic correlates such as fundamental and formant frequencies (e.g. Borkowska & Pawlowski, 2011; Pisanski & Rendall, 2011). Despite the observation of distinct patterns in men's and women's vocal preferences, voice attractiveness highly relies upon the interaction between speaker and listener and their individual preferences, as well as possible within-subject variation. Moreover, research suggests that speakers manipulate acoustic features of their voices depending on the conversational context and partner, for instance, to sound more attractive and in seductive interactions (Fraccaro et al., 2013; Hughes, Mogilski & Harrison, 2014; Leongómez et al., 2014).

In these lines, there is a widespread practice among young speakers in South Korea to enhance their vocal appeal by performing, sometimes even professionally, a speaking style known as *Aegyo*. Aegyo (=애교) can be defined, according to Puzar & Hong (2018), as "a layered articulation of behaviours, gestures, vocal and linguistic adjustments, narratives and fashions that serve to enact child-like charm and infantilised cuteness" (333). Although it has been argued to present strong similarities to baby-talk in Korean language (McGuire, 2015; Puzar & Hong, 2018), Aegyo is a complex phenomenon that has been rarely examined and few studies provide only social and demographic descriptions of it (Park, 2010; Puzar, 2011; McGuire, 2015; Puzar & Hong, 2018). In contrast, linguistic aspects of this speaking style have been solely introduced by Strong (2012), who claims that mean pitch is

key in defining Aegyo voice, but other several cues remain still unexplored. Perceptually, six acoustic features have been reported as possible indicators of Aegyo: palatalization (Strong, 2012); duration and nasalization (Strong, 2012; Puzar & Hong, 2018); whispering voice and high-pitched voice (McGuire, 2015); and consonant strengthening (Puzar & Hong, 2018). Additionally, there is anecdotal evidence suggesting that, although speakers using *Aegyo* are often perceived as attractive and desirable in South Korea, it seems that most people from western countries (whose reactions to Aegyo are very popular on *YouTube*) find this speaking style annoying or even ludicrous, which means that Aegyo may provide us a window to observe the variation of vocal preferences and voice attractiveness across languages and cultures. Therefore, we decided to carry out an acoustic profile of the Aegyo performance as a first step towards an understanding of the acoustic-phonetic characteristics in this speaking style. We compiled the first systematic data collection on Aegyo, including recordings of both read and spontaneously elicited speech under situations in which Aegyo is common and uncommon and with instructions to either use or not use Aegyo. With this material we are planning to understand the attractiveness of Aegyo by Korean and non-Korean listeners in the future. Here, we provide a first description of the acoustic-phonetic characteristics of Aegyo.

Materials & Methods

Fourteen Seoul Korean speakers, eight females and six males (age range: 25 to 31), were recorded. Recordings took place in the sound-treated booth in the Phonetics Laboratory at Seoul National University using Tascam DR-100MKIII and SHURE 10A head-worn microphone (sampling rate = 44.1 kHz, quantization level = 16 bit). The recording consisted in the participants uttering both native Korean and Sino-Korean numbers from 1 to 10, first in 'Aegyo speaking style' and second in 'Non-Aegyo speaking style'. Participants were explicitly and constantly asked either to perform Aegyo to full (Aegyo speech style) or not to perform Aegyo to least (Non-Aegyo speech style), by visual and oral instructions.

The collected data was automatically aligned and annotated into three tiers (phoneme, word, and utterance) by using a Korean Phonetic Aligner (Yoon & Kang 2013), followed by manual inspection and correction. The three tiers were separately analyzed in order to take the difference between prosodic units into account. When defining prosodic units larger than a segment, we relied on the orthographic and morphological standard: word breaks to define 'word' and line breaks to define 'utterance'. Five acoustic correlates were analyzed for each tier: mean pitch, pitch range, mean intensity, intensity range, and duration. As large variability of speaker and item was observed, the mixed effects model was chosen for statistics, defining speaker and item as random variables. Speaking style and the acoustic correlates were used as fixed variables.

Results and Discussion

A phonological analysis was first conducted to provide an overall perspective on Aegyo. To put it in a nutshell, the collected Aegyo was prosodically variable, but segmentally less variable, compared to non-Aegyo. The change of tone and loudness was more frequent and extreme in Aegyo, while different segments were merged into one, reducing the intelligibility. Vowels showed centralization (e.g. [ʌ]→[o]) and monothongization, whereas consonants showed glottalization and frontalization (e.g. [t̪ɕ, l, s]→[t]; [t̪hɕ]→[t̪h]; [t̪ɕ', s']→[t']).

The phonological impressions were supported by our phonetic analysis that showed the significantly larger range of pitch and intensity in Aegyo speaking style. The mean pitch was also significant, as Strong (2012) claimed, albeit less significant than pitch range. Meanwhile, the significance varied from prosodic units as below:

dep. var.	**utterance**			**word**			**phoneme**		
	$\chi^2(1)$	**p-value**	**sig.**	**$\chi^2(1)$**	**p-value**	**sig.**	**$\chi^2(1)$**	**p-value**	**sig.**
pitch range	9.852	0.0016	***	132.27	<2.2e-16	***	86.966	<2.2e-16	***
int. range	2.258	0.1329	.	3.857	0.0495	*	5.354	0.0207	*

mean pitch	1.939	0.1637	.	8.106	0.0044	**	6.805	0.0091	**
mean int.	0.343	0.558	.	2.206	0.1375	.	0.085	0.7704	.
duration	13.934	0.0002	***	142.03	<2.2e-16	***	63.675	1.457e-15	***

When it comes to segmental parameters, significantly longer duration was observed in Aegyo speaking style. However, further analysis on the segmental level is desired to explore the reason behind the low speech intelligibility of Aegyo. Our subsequent study will also include the remaining data of longer elicited speech and spontaneous speech.

Aegyo is a speaking style of Korean applied in situations in which voices need to gain in attractiveness. In this study we investigated some acoustic-phonetic characteristics of Aegyo. We found that – in line with auditory impressions – Aegyo shows a higher variability of segmental and suprasegmental characteristics. As such, Aegyo is similar to other speaking styles with higher than average acoustic variability such as infant directed speech or clear speech, two speaking styles of very different intentions, although auditorily, Aegyo is clearly distinguishable from such speaking styles and other speaking styles – in particular 'clear speech' – which are not targeted at making voices more attractive. This may be due to the less segmental variability of Aegyo, which still needs to be verified by further acoustic analysis. Meanwhile, this may also mean that the type of average variability results obtained in Aegyo in the present study is bound to vary systematically between other phonetically variable speaking styles for it to become distinctive. In future research it will be interesting to understand (a) the characteristic acoustic-phonetic patterns of Aegyo and (b) how such patterns contribute to make speech more attractive in Korean. In respect to (b) it will further be interesting to understand whether such attractiveness patterns are universal or whether they are a result of a culturally shaped understanding of voice attractiveness in Korean.

Acknowledgements

This research was supported by grant number 2018.0218 from the Swiss Government Excellence Scholarship.

References

Borkowska B, Pawlowski B. 2011. Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour* 82:55-59.

Fraccaro PJ, O'Connor JJM, Re DE, Jones BC, DeBruine LM, Feinberg DR. 2013. Faking it: deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour* 85:127-136.

Hughes SM, Mogilski JK, Harrison MA. 2014. *Journal of Nonverbal Behavior* 38:107-127.

Leongómez JD, Binter J, Kubicová L, Stolařová P, Klapilová K, Havlíček J, Roberts SC. 2014. Vocal modulation during courtship increases proceptivity even in naive listeners. *Evolution and Human Behavior* 35:489-496.

McGuire ML. 2015. I Don't Need Feminism, I Have Aegyo: Cuteness and Heteronormativity in South Korea.

Park S. 2010. (The) preference of Korean 'Ae-gyo' to adult attachment. M.A. thesis. Korea University. (in Korean)

Pisanski K, Feinberg DR. 2019. Vocal Attractiveness. In: S Frühholz & P Belin, eds. *The Oxford Handbook of Voice Perception*. Oxford: Oxford UP, 607-625.

Pisanski K, Rendall D. 2011. The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America* 129: 2201-2212.

Puzar A. 2011. Asian dolls and the westernized gaze: Notes on the female dollification in South Korea. *Asian Women* 27:81-111.

Puzar A, Hong Y. 2018. Korean Cuties: Understanding Performed Winsomeness (Aegyo) in South Korea. *The Asia Pacific Journal of Anthropology* 19:333-349.

Strong S. 2012. Too Cute for Words: An Investigation of Prosody's Role in the Construction of Aegyo as a Speech Style. *Washington University Undergraduate Research Digest* 8:29-43.

Yoon T, Kang Y. 2013. Korean Phonetic Aligner Program Suite.

Vocal expression of emotional valence in pigs across multiple call types and contexts

Elodie F Briefer[1], Pavel Linhart[2], Richard Policht[2], Marek Špinka[2], Lisette Leliveld[3], Sandra Düpjan[3], Birger Puppe[3], Mónica Padilla de la Torre[4], Andrew M. Janczak[4], Cécile Bourguet[5], Véronique Deiss[6], Alain Boissy[6], Carole Guérin[7], Eva Read[7], Marjorie Coulon[8], Edna Hillmann[9], Céline Tallet[7]

[1] Institute of Agricultural Sciences, ETH Zürich, Zürich, Switzerland
[2] Institute of Animal Science, Prague, Czech Republic
[3] Leibniz Institute for Farm Animal Biology, Germany
[4] Department of Production Animal Clinical Sciences, Norwegian University of Life Sciences, Norway
[5] Bureau E.T.R.E., France
[6] INRA UMR 1213 Herbivores, France
[7] INRA UMR 1348 PEGASE, France
[8] Cabinet EASIER, France
[9] Humboldt-Universität zu Berlin, Germany

Corresponding Author:
Elodie F Briefer[1]
Present address: University of Copenhagen, Copenhagen Ø, 2100, Denmark
Email address: elodie.briefer@bio.ku.dk

Introduction

Emotions, unlike mood, are short-lived reactions associated with specific events. They can be characterized by two main dimensions, their arousal (bodily activation) and valence (negative versus positive) (Mendl et al. 2010). Knowledge of the valence of emotions experienced by domestic and captive animals is crucial for assessing and improving their welfare, as it enables us to minimize the negative emotions that they might experience and to promote positive ones. Emotions can affect vocalizations directly or indirectly through the brain, lungs, larynx or vocal tract. As a result, vocal expression of emotions has been observed across species (Briefer 2012), and could serve as a non-invasive and potentially very reliable tool to assess animal emotions. In pigs (*Sus scrofa*), vocal expression of emotions has been relatively well studied (e.g. Leliveld et al. 2016; Briefer et al. 2019). However, it is not known if the vocal indicators revealed in previous studies are valid across call types and contexts. To find this out, we conducted a meta-analysis of the effects of emotional valence on pig vocalizations, including calls recorded in the most common emotional situations encountered by pigs throughout their lives, from birth to slaughter.

Materials & Methods

Recordings

Pigs of various ages (piglets to finishing pigs) were recorded in 22 contexts triggering both negative emotions (e.g. crushing, missed nursing, castration, fear conditioning, isolation, restraint, barren environment, and slaughter), and positive emotions (e.g. nursing, huddling, social reunion, exposition to an enriched arena, and running) (for more details see Briefer et al. 2019; Illmann et al. 2013; Tallet et al. 2013; Linhart et al. 2015; Leliveld et al. 2016). The putative valence of the various contexts was based on the function of emotions to trigger avoidance (negative emotions) or approach (positive emotions) and the behavior of the pigs (Mendl et al. 2010).

Vocal analyses

In order to exclude very short sounds, in which parameters might not be accurately measured, only high quality calls with a duration > 0.05 s were selected for the acoustic analysis (n = 6017 calls). We used the

acoustic features of the calls to classify them as low-frequency stable, modulated or tonal calls, high-frequency stable or modulated calls, or mixed calls (6 types), based on Tallet et al. (2013). Then, depending on the call type, we extracted 11 to 18 vocal parameters using a custom-built script in Praat, which batch-processed the analyses and the exporting of output data. The measured parameters belonged to the six following categories: source-related (fundamental frequency, "F0"), energy spectrum distribution, duration, amplitude modulation ("AM"), noise, filter-related (vocal tract resonances).

Statistical analyses

To eliminate redundancy, we used a principal component analysis to select one vocal parameter within each category, which explained most of the variance in the data across all call types, for further analyses. Since the minimum formant dispersion ("DFmin"), originally categorized along with the linear predictive coding ("LPC") coefficients never associated (i.e. loaded highly (r >= l0.5l) on the same PC) with these parameters, it was analyzed separately. These selected seven parameters (i.e. one for each of the six categories and DFmin; Table 1) were then used as outcome variables in linear mixed-effects models (lmer function in R software), to assess if they were affected by the valence of the contexts (positive or negative; fixed factor). The models included as control factors the age category and the call type. The context of production nested within the identity of the pig, nested within the experiment number, nested within the team who performed the recording was added as a random factor to control for repeated measurements and dependencies. The p-values were calculated with parametric bootstrap tests.

Results and Discussion

Five of the seven tested vocal parameters were affected by the valence of the context (Table 1). After controlling for the type of call and the age category (control factors), our analyses revealed that pigs produced calls characterized by a higher center of gravity, a shorter duration, less noise (lower Wiener entropy), lower formants (measured using the formant dispersion) and LPC coefficients in positive compared to negative contexts.

Table 1. Model estimates, lower (lo.ci) and upper (up.ci) 95% confidence intervals for the vocal parameters included in the linear mixed-effect models, as a function of the valence of the contexts (*p<0.05; **p < 0.01; "NS" Non significant).

Parameter	Valence	estim	lo.ci	up.ci	P value
Mean F0 (Hz)	Pos	132.91	124.81	141.13	NS
	Neg	138.19	130.49	146.26	
Spectral centre of gravity (Hz)	Pos	967.70	877.57	1084.53	*
	Neg	895.54	806.46	996.58	
Duration (s)	Pos	0.17	0.14	0.21	**
	Neg	0.42	0.34	0.51	
AM extent (dB)	Pos	5.77	4.60	7.30	NS
	Neg	5.67	4.50	7.24	
Wiener entropy	Pos	-1.63	-1.84	-1.44	**
	Neg	-1.52	-1.71	-1.33	
DFmin (Hz)	Pos	846.64	778.01	921.82	**
	Neg	964.72	899.40	1035.78	
4th LPC coefficient (Hz)	Pos	3913.22	3742.58	4069.90	**
	Neg	4185.93	4020.41	4334.90	

Some of these changes are in line with previous findings (e.g. spectral center of gravity, Leliveld et al. 2016; duration, Briefer et al. 2019). In particular, shorter durations in positive contexts have been observed across multiple species and could be a feature conserved throughout evolution (Briefer 2012). Overall, our results suggest that some parameters change with the valence experienced by pigs in a similar way across call types. These vocal parameters could be very useful for developing automated methods to monitor pig welfare on-farm.

Acknowledgements

This research is funded by the ERA-Net ANIHWA project SOUNDWEL.

References

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. Journal of Zoology, 288:1–20

Briefer, E. F., Vizier, E., Gygax, L., Hillmann, E. (2019). Expression of emotional valence in pig closed-mouth grunts: Involvement of both source- and filter-related parameters. Journal of the Acoustical Society of America, In Press

Illmann, G., Hammerschmidt, K., Špinka, M. and Tallet, C. (2013). Calling by domestic piglets during simulated crushing and isolation: a signal of need? PLOS ONE, 8:e83529.

Leliveld, L. M. C., Düpjan, S., Tuchscherer, A., Puppe, B. (2016). Behavioural and physiological measures indicate subtle variations in the emotional valence of young pigs. Physiology & Behavior, 157:116–124.

Linhart, P., Ratcliffe, V. F., Reby, D., Špinka, M. (2015). Expression of emotional arousal in two different piglet call types. PLoS ONE, 10:e0135414.

Mendl, M., Burman, O. H. P., Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. Proceedings of the Royal Society B, 277:2895–2904.

Tallet, C., Linhart, P., Policht, R., et al (2013). Encoding of situations in the vocal repertoire of piglets (*Sus scrofa*): A comparison of discrete and graded classifications. PLoS ONE, 8:e71841.

Learning How to Sing: Developing a Virtual Bird to Probe Zebra Finch Vocal Interactivity

Julia Hyland Bruno[1], Seth Cluett[2], Ben Holtzman[3], George E. Lewis[4]

[1] Center for Science and Society, Columbia University, New York, NY, USA
[2] Computer Music Center, Columbia University, New York, NY, USA
[3] Lamont Doherty Earth Observatory, Columbia University, New York, NY, USA
[4] Department of Music, Columbia University, New York, NY, USA

Corresponding Author:
Julia Hyland Bruno[1]
511 Fayerweather Hall, MC2509, 1180 Amsterdam Avenue, New York, NY, 10027, USA
Email address: jhb2202@columbia.edu

Abstract

Vocal learning is a vital ingredient in the acquisition of culturally-transmitted communication systems such as human language or birdsong. Whereas developmental language acquisition is recognized to be a mess of processes both sensorimotor and social, students of birdsong learning have largely focused on vocal imitation—i.e., the problem of learning *what* to sing. However, for certain songbirds, such as the gregarious zebra finch, the question of how juveniles learn to use their vocalizations in social interactions—i.e., how birds learn *how* to sing—may be equally salient. The relative simplicity and temporal precision of the adult male zebra finch's crystallized song has made it a fruitful model system for studying behavioral and neurobiological mechanisms of imitative vocal learning. However, beyond the core sequence of song elements copied from a tutor during development, zebra finch song bouts exhibit within-individual diversity in both sequence and timing, as well as variability between individuals in the degree of this structural diversity. How such performance-level plasticity is acquired and whether it serves any social function is not known. Zebra finches are group-living, and males sing abundantly in general social as well as female-directed settings. Beyond courtship, song may also play a role in mated pair maintenance. Before exploring these various functions, let alone how the young bird acquires them, a fuller characterization of zebra finch singing plasticity is required. Previously we have observed that familiar males very rarely sing at the same time, suggesting that zebra finches can control the timing of their songs. Here we provide an introduction to the diversity and potential flexibility of zebra finch song performance, including a method of phenotyping individual birds' repertoires according to temporal consistency and repertoire size. We then describe a biomimetic approach for probing zebra finch vocal interactivity by means of a software-based virtual avian interlocutor—employing three different instantiations which vary in their mechanisms for attending and responding to real-time acoustic input from a live zebra finch—aimed at reproducing patterns of turn-taking as observed in real birds.

Introduction

Vocal learning, the capacity to map sounds produced and sounds heard, is a necessary precondition for culturally-transmitted vocal traditions such as human languages or birdsong dialects. Because this capacity is seemingly rare in the animal kingdom as well as unobserved in our nearest primate relatives, songbirds have become a model system for exploring its neurobiological and behavioral mechanisms. An important parallel between vocal learning in language and birdsong is the presence of sensitive periods for learning during development. The analogy between birdsong and language acquisition is limited, however, by a focus on vocal imitation in avian vocal learning and a general conceptualization of birdsong learning as a process of learning *what* to sing. This conceptualization is yoked to a picture of birdsong

function that is derived mainly from the behavior of territorial species, which need to acquire the appropriate signals for mate attraction and repelling rivals. For communal-living birds, on the other hand, learning *how* to sing may present an additional challenge. In this respect, the zebra finch—perhaps the most studied songbird (Griffith & Buchanan, 2010)—has been underutilized, if not misrepresented. Zebra finches are not territorial, but instead live in nomadic groups of fluctuating size, breeding opportunistically around water availability, forming tight, often lifelong pair bonds but also associating with other individuals of both sexes over extended periods of time (Zann, 1996). Little is known about how vocal interactions accompany or mediate such ordinary features of zebra finch life.

The male zebra finch has been celebrated for the simplicity and predictability of his signature song motif, a stereotyped sequence of crystalized vocal gestures acquired during development. Each male typically produces a single, idiosyncratic motif throughout his adult life. Yet birds do not emit isolated motifs when they sing. Rather, zebra finches 'compose' song bouts consisting of multiple motif repetitions which may be linked by variable pauses and a variety of para-motif vocalizations, which altogether give singing performances variability in both rhythm and sequence. These variations tend to be dismissed as production noise. However, analyses of large samples of song recordings (Hyland Bruno & Tchernichovski, 2017) reveal that individual birds have characteristic song-performance repertoires which occupy only small subspaces of all possible stochastic variations. Moreover, birds can be phenotyped along two dimensions of plasticity, according to both the temporal consistency vs. jitter and the sequential stereotypy vs. complexity of their songs (Figure 1).

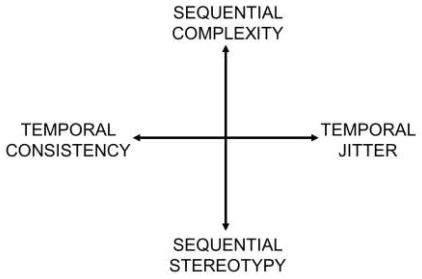

Figure 1. Two-dimensional space for phenotyping individual zebra finches according to the plasticity of their song performance repertoires.

Here we take a biorobotic approach to characterizing how zebra finch singing plasticity may be differentially expressed in social interactions. Guided by the empirical observation that birds with established social bonds take turns at singing and appear to coordinate the timing of their songs so as to avoid overlap (Hyland Bruno & Tchernichovski, 2017), we set out to attempt to reproduce a pattern of turn-taking in dyadic encounters between virtual acoustic agents and live birds representing the different song repertoire phenotypes shown in Figure 1.

Materials & Methods

Since we don't know a priori how zebra finches coordinate their songs, we are developing three versions of our virtual bird (VB) in parallel, using the flexible music programming software Max/MSP: 1) a sampler playback system, in which a human "plays" zebra finch; 2) an autonomous interactive agent that tracks the live bird's vocalizations and responds in real time according to a predetermined set of rules; and 3) an autonomous interactive agent whose real-time responses are determined via machine learning. The preprogrammed repertoire is the same for all VBs, and consists of a typical zebra finch song motif (a 750-ms sequence of discrete harmonic, broadband, and frequency-modulated sounds separated by short silent intervals) plus innate "calls," which birds of both sexes produce singly in rapid exchanges with conspecifics, and which males also produce during singing, interspersed with their learned song motifs. Previous work has demonstrated that zebra finches readily engage in calling exchanges with a vocal robot (Benichov et al., 2016). In trials with VB #1, the human operator will aim to "sing" as much as possible without overlapping the live bird, selecting performance types (Figure 1) or composing songs out of motifs and calls (and silences) in real time. VB #2 incorporates rudimentary machine "listening" in order to respond autonomously. The VB tracks the amplitude envelope of the live zebra finch's vocalizations

and extracts a running record of sound onsets and offsets. It then uses this information to determine both when and how to produce vocal responses. The VB singing style can be programmed in various configurations, either to always produce exemplars of a specific phenotype (one of the quadrants in Figure 1), or to mimic the exemplar or phenotype produced by the live bird. Finally, VB #3 is also an autonomous agent, but one which recycles and improvises on inputs from the live bird, an adaptation of real-time sequence modeling and statistical learning methods developed for human-computer musical improvisation (Assayag et al., 2006; Collins, 2014).

Results and Discussion

Building biomimetic machines and studying the resultant phenomena is a promising way of probing our understanding of animal behavior (Webb, 2008). Presented here is an iterative approach, inspired by explorations in improvised computer music (Lewis, 1999), to developing a virtual acoustic agent capable of simulating the overlap avoidance/turn-taking that we observe in zebra finch vocal interactions (which is also a hallmark of human communication systems [Levinson, 2006; Pelz-Sherman, 1998]). That empirical observation prompts a host of questions: How is such coordination achieved? Is successful coordination related to the singing phenotypes of individual birds? Are repertoires fixed or do they change over time, and, if so, why? How are dyadic interactions related to group dynamics? Our ultimate goal with this project is to generate new hypotheses that can be tested in future animal experiments.

Acknowledgements

We thank the Columbia Presidential Scholars in Society and Neuroscience program for fostering this interdisciplinary collaboration.

References

Assayag, G., Bloch, G., Chemillier, M., Cont, A., & Dubnov, S. (2006). OMax brothers: a dynamic topology of agents for improvisation learning. In *AMCMM '06*. New York: ACM

Benichov, J.I., Benezra, S.E., Vallentin, D., Globerson, E., Long, M.A., & Tchernichovski, O. (2016). The forebrain song system mediates predictive call timing in female and male zebra finches. *Curr. Biol., 26,* 309-318.

Collins, N. (2014). Virtual musicians and machine learning. In K. Collins, B. Kapralos & H. Tessler (Eds.), *The Oxford Handbook of Interactive Audio*. DOI: 10.1093/oxfordhb/9780199797226.013.021

Griffith, S.C., & Buchanan, K.L. (2010). The zebra finch: the ultimate Australian super model. *Emu, 110,* v-xii.

Hyland Bruno, J., & Tchernichovski, O. (2017). Regularities in zebra finch song beyond the repeated motif. *Behav. Processes*. DOI: 10.1016/j.beproc.2017.11.001.

Levinson, S.C. (2006). On the human "interaction engine." In S.C. Levinson & N.J. Enfield (Eds.), *Roots of human sociality: culture cognition and interaction* (pp. 39-69). Oxford: Berg.

Lewis, G.E. (1999). Interacting with latter-day musical automata. *Contemp. Music Rev.,* 18, 99-112.

Pelz-Sherman, M.L. (1998). *A framework for the analysis of performer interactions in Western Improvised Contemporary Art Music*. Unpublished doctoral dissertation. University of California, San Diego.

Webb, B. (2008). Using robots to understand animal behavior. *Adv. Study Behav., 38,* 1-58.

Zann, R. (1996). *The zebra finch: a synthesis of field and laboratory studies.* Oxford: University Press.

Initial Observation of Human-Bird Vocal Interactions in a Zoological Setting

Rébecca Kleinberger[1], Janet Baker[1], and Gabriel Miller[2]

[1]**MIT Media Lab, 75 Amherst Street, Cambridge MA**
[2]**Institute for Conservation Research - San Diego Zoo, 15600 San Pasqual Valley Rd, Escondido, CA**

Corresponding author:
Rébecca Kleinberger[1]

Email address: rebklein@media.mit.edu

ABSTRACT

Vocal interactions between humans and non-human animals are pervasive, but studies are often limited to communication within species. Here, we conducted a pilot exploration of vocal interactions between visitors to the San Diego Zoo Safari Park and Sampson, an 18-year-old male Hyacinth Macaw residing near the entrance. Over the course of one hour, 82 vocal and behavioral events were recorded, and various relationships between human and bird behavior were noted. Analyses of this type, applied to large datasets with assistance from artificial intelligence, could be used to better understand the impacts, positive or negative, of human visitors on animals in managed care.

INTRODUCTION

Interaction and communication between humans and domesticated animals (such as cows, horses, dogs, and cats) are well documented (eg. Saito et al. (2013)). However, interaction between humans and threatened or endangered wild species (like Hyacinth Macaws) are generally less characterized. Nevertheless, at zoological institutions, humans have extensive contact with rare species, and human interaction is an important element of the lives of animals in managed care.

When entering the San Diego Zoo Safari Park, a Hyacinth Macaw (Anodorhynchus hyacinthinus) named Sampson is the first visible animal. The largest (head to tail) flying parrot, his species is classified as vulnerable on the International Union for Conservation of Nature RedList (BirdLifeInternational, 2019). The San Diego Zoo Safari Park was visited by 1.5 million guests in 2018, meaning Sampson is passed by an average of 4,000 guests daily. We aimed to begin characterizing vocal interaction between Sampson and guests in order to better understand the dynamics of human-bird vocal communication in a zoological setting. Sampson's enclosure is about 4x8m and is surrounded by a waist-high fence (Fig. 1) which excludes guests from approaching too closely. Sampson can freely move within his enclosure.

Figure 1. Sampson, the Hyacinth Macaw, in his enclosure near the San Diego Zoo Safari Park entrance

BACKGROUND

The effects of the presence and behaviors of humans on animals at zoological organizations has been studied for several decades. The significance of humans for animals is important in at least five different ways: as enemies, prey, symbionts, pieces of the inanimate environment, or members of its own species (Hediger, 1969). The first and last cases have been studied to evaluate the contexts in which the presence of humans elicits positive, neutral, or negative effects on captive wildlife in zoos.

Humans (and human-generated sounds) sometimes have negative effects on other species. For example, visitors increased distress levels of wolfs (Pifarré et al., 2012), pandas (Owen et al., 2004), orangutans (Birke, 2002), and construction noises increased stress in of big cats (Chosy et al., 2014).

However, certain human-animal interactions clearly benefit particular animals. The uniqueness of each human-animal diad helps explain the rich connections between animals and their primary caregivers (Ward, 2015). Claxton (2011) also explores the effects of daily contact with both familiar and unfamiliar people and concludes that such interactions can lead to positive outcomes if human contact is intentionally designed to address environmental enrichment aims. She also emphasises the importance of tailoring human contact on a species-by-species basis. Understanding the interactions between animals and zoo visitors can allow determination of the visitor characteristics and behaviors which are most appealing to animals, and lead to higher levels of animal-human interaction (Cook, 1995), playfulness (Owen, 2004), and energy expenditure (Nimon and Dalziel, 1992).

In the specific context of the vocal interactions, researchers have recently explord the concept of zoo voices and the characteristics of the merged soundscape created from animal and human voices, as well as animal and human generated sounds. Tunnicliffe and Scheersoi (2012) discuss ways in which zoos make their voices available to visitors, creating dialogues and active listening. In the specific example of parrots, we interpret physical and vocal behaviors from the animal based on interviews with the bird's expert caregivers as well as previous research on parrots, in particular the important work from Pepperberg (1994) that revealed parrots abilities. Parrots are known to be vocal and social, and related parrot species exhibit head-bobbing as part of courtship behavior (Symes et al., 2004) or as a sign of playfulness.

METHODS

In this exploratory study, we recorded one hour of audio and video of a Hyacinth Macaw's enclosure, and analyzed interactions between humans and the bird. Six different types of behaviors were scored: three were human vocalizations (adult speaking to bird, child speaking to bird, and adult whistling to bird) and three were bird behaviors (bird vocalisation, bird head nodding, and bird moving toward a visitor). We recorded instances when visitors vocally addressed the bird (but did not score instances where humans conversed solely with one another nor when they stopped to look at the bird without talking to him). When guests vocalized toward the bird, they commonly (but not always) raised the tone of their voice and faced the bird. Throughout the trials, background noise (from entry gates and human-human conversation) was continual. These preliminary observations were obtained in the course of implementing a larger project exploring ways to provide audio enrichment for animals in managed care.

RESULTS

During one hour, we recorded 82 instances of the six target events. Between 300 and 400 visitors entered the park during this hour. Many visitors stopped to look at the bird and were engaged in human-human discussions while doing so, and 34 (approximately 10%) verbally addressed the bird. We recorded 41 unique human vocal events, including 3 instances of whistling and 14 instances of children talking to the animal. The macaw vocalized 16 times, nodded 19 times and physically approached visitors 7 times.

63% of the bird vocalisations (10/16) were preceded by less than 10 seconds by a visitor addressing the bird (2 adult whistles, 3 child vocalizations, 5 adult vocalizations). 43% of these times (3/7), the bird approached the visitor who just vocally addressed him. 58% of head nodding behaviors (11/19) were preceded by less than 10 seconds by a visitor addressing the bird (2 adult whistles, 4 child vocalizations, 5 adult vocalization). The bird often responded to visitors with a combination of several different behaviors (for example, five occurrences of vocalizing + head bobbing and two occurrences of approaching + head bobbing + vocalizing). In four instances, the interactions between visitor and bird contained turn-taking, dialog-like characteristics during which neither the bird not the visitor would vocalise during the other's turn and the bird would vocalise or nod more than once.

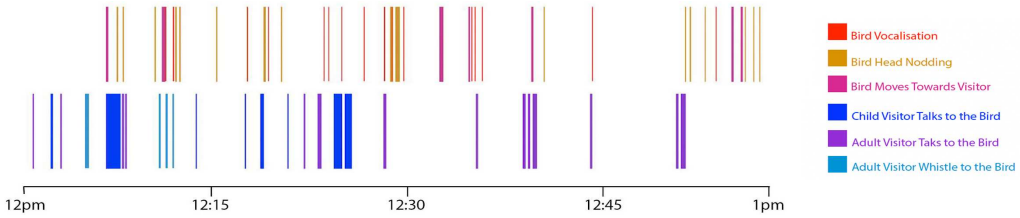

Figure 2. timeline of all scored behaviors of the bird (top row) and of the human visitors (bottom row)

DISCUSSION AND CONCLUSION

Generally, animal behavior studies focus on single species. However, in managed care, the interactions of animals with humans is of paramount importance. To our knowledge, this is the first description of unstructured human-bird vocal interaction in a zoological setting. Unstructured communication (in the absence of goal-oriented training, etc) comprises the majority of captive animals' experience, and yet these sessions are largely uncharacterized. Here we observe that Sampson's experiences are highly interactive, and that his vocalizations appear to be correlated with (and tend to follow) those of guests.

This study is extremely preliminary; by examining the temporal relationships between human and bird behaviors, we can begin to draw inferences about how guests influence the behavior and engagement levels of animals under managed care. Further studies could subdivide Sampson's vocalizations into various types, and the influence of variables like time-of-day, weather, and particular individual humans could be investigated. The influence of human speech on bird vocalization and behavior is likely to depend not only on the species of bird but also on the tendencies of individual birds. Sampson, in particular, is described by his keepers as less 'talkative' than other parrots in the collection, yet clearly still has vocal dialogue with guests. In the future, using large datasets comprised of many hours of audio recordings, algorithms can annotate bird and human vocalizations. Autocorrelation functions can then reveal the temporal dependence of these signals. AI approaches using deep learning could provide more in-depth ethological understanding. Previous instances of the use of deep learning for the recognition of animal behavior such as DeepLabCut (Mathis, 2018) could be used as a starting point. Better standards of care will be reached as we continue to understand the impact of human visitors on animal behavior.

We wish to thank Jenna Duarte and Michelle Handrus for their extensive support. All procedures described were approved by the Zoological Society of San Diego IACUC under proposal 19-002.

REFERENCES

BirdLifeInternational (2019). Anodorhynchus hyacinthinus. the iucn red list of threatened species 2016.

Birke, L. (2002). Effects of browse, human visitors and noise on the behaviour of captive orang utans.

Chosy, J. et al. (2014). Behavioral physiological responses in felids to exhibit construction. *Zoo biology*.

Claxton, A. M. (2011). The potential of the human–animal relationship as an environmental enrichment for the welfare of zoo-housed animals. *Applied Animal Behaviour Science*, 133(1-2):1–10.

Cook, S. (1995). Interaction sequences between chimpanzees and human visitors at the zoo. *Zoo Biology*.

Hediger, H. (1969). Man and animal in the zoo.

Mathis, A. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning.

Nimon, A. and Dalziel, F. (1992). Cross-species interaction and communication: a study method applied to captive siamang and long-billed corella contacts with humans. *Applied Animal Behaviour Science*.

Owen, C. (2004). Do visitors affect the asian short-clawed otter in a captive environment. In *Proceedings of the 6th Annual Symposium on Zoo Research-BIAZA*, pages 202–211.

Owen, M. A. et al. (2004). Monitoring stress in captive giant pandas: behavioral and hormonal responses to ambient noise. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*.

Pepperberg, I. (1994). Vocal learning in grey parrots: effects of social interaction, reference, and context.

Pifarré, M. et al. (2012). The effect of zoo visitors on the behaviour and faecal cortisol of the mexican wolf (canis lupus baileyi). *Applied Animal Behaviour Science*, 136(1):57–62.

Saito, A. et al. (2013). Vocal recognition of owners by domestic cats (felis catus). *Animal cognition*.

Symes, C. et al. (2004). Behaviour and some vocalisations of the grey-headed parrot poicephalus fuscicollis suahelicus (psittaciformes: Psittacidae) in the wild. *Durban Museum Novitates*.

Tunnicliffe, S. D. and Scheersoi, A. (2012). Voices in zoos and aquariums. *IZE Journal*, (48).

Ward, S. (2015). Keeper-animal interactions: Differences in the behaviour of animals affect stockmanship.

Sex dimorphic phrase combinatorics in the song of the indris (*Indri indri*)

Anna Zanoli[1], Chiara De Gregorio[1], Daria Valente[1], Valeria Torti[1], Giovanna Bonadonna[1], Rose Marie Randrianarison[2], Cristina Giacoma[1] & Marco Gamba[1]

[1] Department of Life Sciences and Systems Biology, University of Torino, Torino, Italy
[2] Groupe d'étude et de recherche sur les primates de Madagascar (GERP), Antananarivo, Madagascar

Corresponding Author:
Marco Gamba[1]
Via Accademia Albertina 13, Torino, 10123, Italy
Email address: marco.gamba@unito.it

Abstract

We used a logic distance to investigate intra and inter-individual variation in the phrase combinatorics of a singing primate, the indri, which inhabits the montane rain forests of Madagascar. Indris combine long notes, short single notes, and phrases consisting of two, three, four, or five units with slightly descending frequency. We calculated the similarity across different individual songs using the Levenshtein distance. We then analyzed the degree of similarity within and between individuals and found that: i) the phrase structure of songs varied between reproductive males and females; ii) male contributions to the song are overall more similar to those of other males; iii) male contributions are more stereotyped than females' ones. The picture emerging from phrase combinatorics in the indris is in agreement with previous findings of rhythmic features and repertoire size, which also suggested that female songs are potentially more distinctive than those of males.

Introduction

Communication between conspecifics often involves the use of vocalizations because acoustic signals allow encoding a considerable amount of information in a short time (Bradbury and Vehrencamp 2011). Animal vocal signals can be emitted in the form of short vocalizations or given in sequences of variable length (Catchpole and Slater 2008) as it happens in insects, amphibians, and mammals (Kershenbaum et al. 2016). There are several methods for investigating different levels of structural information in acoustic displays. The Levenshtein distance is a quantitative method for measuring the similarity of sequences (hereafter LD; Margoliash et al. 1991). The LD is a logical distance commonly used to quantify the difference between two strings of data (e.g., human words, sequences of visual movements or sequences

of song themes; Gooskens 2004). This technique has often been used to measure similarity in human dialects (Wieling 2014), and it has been applied to animal vocal sequences, but for a very limited number of species (*Passerina cyanea*, Margoliash et al. 1991; 1994; *Phylloscopus trochilus*, Gil and Slater 2000; *Megaptera novaeangliae*: Helweg et al. 1998; Tougaard and Eriksen 2006; Garland et al. 2012). When seen in comparison with humans, animals showed a limited combinatory ability to concatenate vocal emissions in phrases, at least in the acoustic domain (Berwick et al. 2011), but the information available on the variability within a species is very little (Honda and Okanoya 1999; Takahasi et al. 2010). Moreover, few investigations on primate vocal sequences are currently available and none of them are evaluating the stereotypy of song structure between sexes using LD (Gustison et al. 2016).

Indris (*Indri indri*, Gmelin 1788) represent a distinctive species for studying vocal communication because of its rich repertoire (Maretti et al. 2010) and the impressive long-distance songs, which are unique among lemurs (Gamba et al. 2016; Torti et al. 2017). The song of the indri consists of a long series of modulated notes, organized in phrases (Gamba et al. 2011). Male and female indris within a group, including juveniles, take part in a chorusing song, which lasts 40-250 s (Maretti et al. 2010). Previous research showed that the indris can emit songs in different context and that the song can elicit different behaviors depending on their acoustic structure. Cohesion songs, emitted when the animals were dispersed in the territory, were followed by a displacement of the emitters significantly higher than that following the advertisement songs, which were usually given when the animals where in visual contact (Torti et al. 2013). Other studies have shown that male and female contributions to the song differ, both quantitatively and qualitatively, in the temporal and frequency structure of units, and repertoire size (Giacoma et al. 2010; Sorrentino et al. 2012). Sex dimorphism is also present in the modulation of the frequency of vocal emissions, in the duration of note types and the rhythmic structure of a contribution (Gamba et al. 2016; De Gregorio et al. 2018). Because group encounters in the indris are rare (Bonadonna et al. 2014; Bonadonna et al. 2017), it has been suggested that songs may play a role in finding a partner and mediate pair formation.

Since previous work (De Gregorio et al. 2018) shows that females adjust their contributions in order to achieve the synchronization with males, we hypothesize that this adjustment can be also reflected in a sexually dimorphic use of phrases combination. Studies of song structure in bird duets suggested that females' song would be more acoustically variable than that of males accordingly to the territorial model of duet evolution, which is consistent with socially monogamous pairs that actively defend their territory. The active role of females of Australian magpies (*Gymnorhina tibice*) in territorial defense was correlated with a song repertoire more elaborate in comparison to that of the male. Repertoires of females were as large or larger and more complex than those of males, on the level of both the syllable and the song (Brown and Farabaugh 1991). Like Australian magpies, indri groups occupy non-overlapping areas in the forest (Pollock 1979) and use the songs to inform neighboring groups about the occupation of a territory and to actively defend the territory during group encounters (Torti et al. 2013). As the indris utter advertisement and cohesion songs (Torti et al. 2013), by which they inform neighbors about the sex, age, and status of singing individuals (Giacoma et al. 2010; Sorrentino et al. 2012) and bring together the members of a group (Torti et al. 2013), we predicted that the female contribution to the song would be structurally different than that of males.

Materials & Methods

Observations and recordings

We studied 8 groups (N_{tot}= 36 individuals) living in the Maromizaha Forest (18°56'49"S, 48°27'53"E). We recorded the animals between 2011 and 2017. We observed a social group per week, approximatively from 6 AM to 1 PM. All recordings were carried out without the use of playback stimuli, and nothing was done to modify the behavior of the indris. We recorded 142 songs, consisting of duets and choruses with a maximum of five individuals singing in the same song. For the analysis, we considered a total of 17 focal animals: nine reproductive adult males, and eight reproductive adult females. The different number of males and females is motivated by the fact that, during the study period, the reproductive male of a group changed. All the songs were recorded using solid-state recorders (Olympus LS05, Tascam DR-100, Tascam DR-05) at a distance comprised between 2 and 20m. We always kept the visual contact with the vocalizing animals and maximized our efforts to face the focal animals during the emission of the song. Sequences from multiple years were present in the sample, but the songs were all labeled as advertisement songs and were recorded in the same context (Torti et al. 2013). Using the focal animal sampling technique (Altmann 1974), we were able to attribute each vocalization to its signaler. We will refer to an individual' singing within a song or a chorus as an 'individual contribution.'

Acoustic and statistical analyses

We edited segments containing indri's songs using Praat 6.0.30 (Boersma and Weenink 2008) and BORIS 5.1 (Friard and Gamba 2016). We saved each recorded song in a single audio file (in WAV format). We saved the information related to the identity of each singer in a Praat textgrid. We then labeled all the vocal units according to their belonging to a song portion (long notes or descending phrases, see Torti et al. 2013 for details) and to a descending phrase (hereafter, DP; see Torti et al. 2017 for details). We considered phrases consisting of two (DP2), three (DP3), four (DP4), five (DP5), and six (DP6) units. This information was saved in Praat and exported to a Microsoft© Excel spreadsheet (Gamba et al. 2012).

To understand whether there were differences in song structure between sexes, we investigated the DPs combinatorics in each individual contribution. We transformed each contribution in a string of labels separated by a break symbol (e.g., DP2|DP3|DP4|DP3). We obtained 142 strings for females, and 119 strings for males (with an average of 13.2 songs per individual, SD = 5.91). We calculated the Levenshtein distance (LD) for each pair of strings (package StringDist 0.9.4.2 in RStudio) because this methodology provides a robust quantitative approach for the study of animal acoustic sequences (Kershenbaum and Garland 2015). It calculates the minimum number of necessary changes (insertions, deletions, and substitutions) to transform one string into another (Kohonen 1985). We obtained a squared matrix consisting of the distances between each pair of strings. We then averaged LDs to calculate within- and between-individual means and to investigate whether females and males differed in their degree of variation. For this purpose, we ran Mantel tests (9999 randomizations) using a matrix featuring the average individual means against a model matrix consisting of 0 when the corresponding individuals were of the same sex (Krull et al. 2012), and 1 when they were opposite sexes (package *vegan* in RStudio). When investigating differences at the group level or within-sex, we used the non-parametric paired samples Wilcoxon test to compare the average individual LDs of each member of a pair or the within- versus between-individual LDs. In the case of such a small sample size, the Mantel test is not recommended (Legendre and Fortin 1989). Only for the Wilcoxon test, the group in which the male changed was entered twice, considering the two pairs as different groups.

Results and Discussion

We analyzed 260 individual contributions consisting of a total of 2018 phrases. We obtained 77 ± 21 phrases per male and 78 ± 23 phrases per female. We found that average phrase duration was 1.285 s (range: 0.380 - 3.000 s). The number of phrases in the individual song ranged between 2 and 27 phrases.

We found a significant difference between the LDs calculated for males and females, where females showed higher average individual means than males (Mantel test: r = 0.167, p-value = 0.002). In all groups, the females had higher LDs (LD = 6.497 + 1.674) than those of males (LD = 3.946 + 0.814) showing that female contribution to the song was less stereotyped (Fig. 1, Wilcoxon paired test: V = 0, df = 7; p-value = 0.007813). Both females and males showed a higher variability at the between-individual ($LD_{females}$ = 7.386 ± 0.709, LD_{males} = 4.885 ± 0.325) than at the within-individual level (Fig. 1), except for the females of groups 4 and 8. Overall, we found a significant difference between within- and between-individual LDs (Wilcoxon paired test: V = 0, df = 7; p-value = 0.007813).

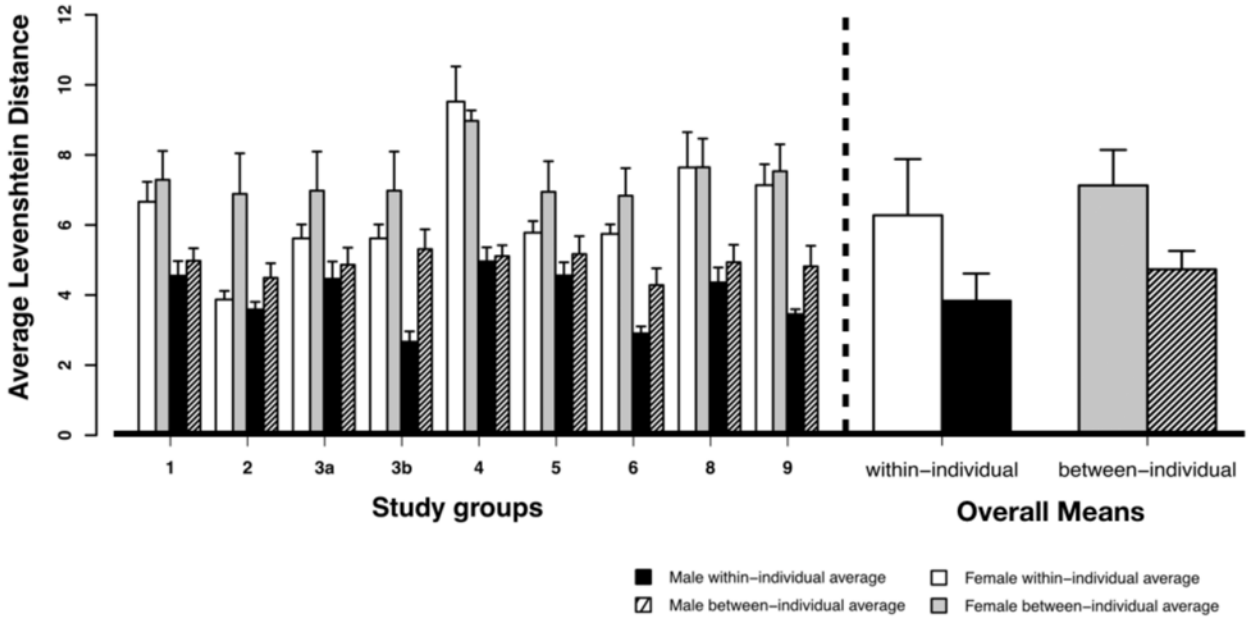

Figure 1: Comparison of Average Levenshtein Distance among sexes and individuals, in the nine studied groups. Bar plot describing the individual and overall degree of stereotypy and variability expressed by the average Levenshtein Distances (LDs). Within-individual LDs are reported for females (white bars) and males (black bars), as well as between-individual LDs (grey bars for females, striped bars for males). Group 3 is reported twice because the male of the reproductive pairs changed in 2014. Capped lines represent Standard Deviation

We found support for our prediction that the phrase structure of songs varied between reproductive males and females. The LDs showed that the between-individual stereotypy of male contributions is much higher than females' one. Males, therefore, appeared to produce songs that are overall more similar to those of other males and showing higher stereotypy when compared to females. In agreement with previous studies that reported sexual dimorphism in the overall timing and repertoire size (Giacoma et al. 2010), and the frequency modulation, duration and the

rhythm (Gamba et al. 2016; Torti et al. 2017, De Gregorio et al. 2019), we found that male and female indris also differed in the phrase combinatorics of their songs. This result is in line with the hypothesis that female components of the song were more complex than that of males, suggesting that singing for females may serve to advertise the mated status of their partner and prevent extra-pair copulations and male desertion, as it happens in birds (Levin 1996). In agreement with previous findings on the different role of males and females during the song (Giacoma et al. 2010), we found that female song is potentially more distinctive than that of males. We expanded the findings of Sorrentino and colleagues (2012) showing that females not only have a broader repertoire of units, but they also emit descending phrases that we did not observe in males (e.g., descending phrases of six units).

These results are in agreement with previous finding on birds (Brown and Farabaugh 1991) confirming that in those species in which females are involved in territorial defense, their repertoires are as large or larger than those of males, on the level of both units and phrases. In support of the higher variability in female song structure, there is the recent evidence that genetics plays a critical role in determining the characteristics of DPs in males, whereas it may have a lesser impact on female songs (Torti et al. 2017). A more variable song structure may, in fact, add up to a more flexible structuring of the phrase notes, but further investigations are needed.

This work also expands on and complements previous studies on humpback whales (Helweg et al. 1998; Tougaard and Eriksen 2006), showing that the Levenshtein distance is simple, efficiently computable and highly applicable to any behavioral data that are produced in a sequence. Our results confirmed that the Levenshtein distance method is a simple but powerful technique that can be applied to assess stereotypy or divergence between sexes.

Acknowledgements

This research was supported by Università degli Studi di Torino and by grants from the Parco Natura Viva—Centro Tutela Specie Minacciate. We are grateful to GERP and Dr Jonah Ratsimbazafy. We thank Dr Cesare Avesani Zaborra and Dr Caterina Spiezio for helping us with the organization of the field station in Maromizaha. We are grateful to the researchers and the international guides, for their help and logistical support. We also thank San Diego Zoo Global, LDVI, Dr Chia L. Tan. The contents of this document are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the position of the European Union. We have received permits for this research, each year, from "Direction des Eaux et Forêts" and "Madagascar National Parks": 2011 - N° 274/11/MEF/SG/D GF/DCB.SAP/SCB, 2012 - N°245/12/MEF/SG/DGF/DCB.SAP/SCB, 2014 - N°066/14/MEF/SG/DGF/DCB.SAP/SCB, 2015 - N° 180/ 15/ MEEMF/ SG/ DGF/ DAPT/ SCBT; 2016 - N° 98/ 16/ MEEMF/ SG/ DGF/ DAPT/ SCB.Re and N° 217/ 16/MEEMF/ SG/ DGF/ DSAP/ SCB.Re, 2017 - 73/17/MEEF/SG/DGF/DSAP/SCB.RE.

References

Altmann, J. (1974) Observational study of behavior: sampling methods. Behaviour, 49: 227-267. doi: 10.1163/156853974X00534

Berwick, R. C., Okanoya, K., Beckers, G. J. L., Bolhuis, J. J. (2011). Songs to syntax: the linguistics of birdsong. Trends in Cognitive Sciences, 15: 113-121. doi: 10.1016/j.tics.2011.01.002

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer [Computer program].
Bonadonna, G., Torti, V., Sorrentino, V., Randrianarison, R. M., Zaccagno, M., Gamba, M., Giacoma, C. (2017). Territory exclusivity and intergroup encounters in the indris (Mammalia: Primates: Indridae: *Indri indri*) upon methodological tuning. The European Zoological Journal, 84: 238-251. doi: 10.1080/24750263.2017.1318184
Bradbury, J.W., Vehrencamp, S.L. (2011). Principles of animal communication. Sinauer, Sunderland.
Brown, E. D., Farabaugh, S. M. (1991). Song sharing in a group-living songbird, the Australian magpie, *Gymnorhina tibicen*. Part III. Sex specificity and individual specificity of vocal parts in communal chorus and duet songs. Behaviour, 118: 244-274.
Catchpole, C.K., Slater, P.J.R. (2008). Bird song: biological themes and variations. Cambridge University Press, Cambridge.
De Gregorio, C., Zanoli, A., Valente, D., Torti, V., Bonadonna, G., Randrianarison, R. M., Giacoma, C., Gamba, M. (2018). Female indris determine the rhythmic structure of the song and sustain a higher cost when the chorus size increases. Current Zoology, 65: 89-97. doi: 10.1093/cz/zoy058
Friard, O., Gamba, M. (2016) BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods in Ecology and Evolution, 7: 1325–1330. doi:10.1111/2041-210X.12584
Gamba, M., Favaro, L., Torti, V., Sorrentino, V., Giacoma, C. (2011). Vocal tract Flexibility and variation in the vocal output in wild indris. Bioacoustics, 20: 251-265. doi: 10.1080/09524622.2011.9753649
Gamba, M., Friard, O., Giacoma, C. (2012). Vocal tract morphology determines species-specific features in vocal signals of lemurs (*Eulemur*). International Journal of Primatology, 33.6: 1453-1466. doi:10.1007/s10764-012-9635-y
Gamba, M., Torti, V., Estienne, V., Randrianarison, R. M., Valente, D., Rovara, P., Giacoma, C. (2016). The indris have got rhythm! Timing and pitch variation of a primate song examined between sexes and age classes. Frontiers in neuroscience, 10:249. doi: 10.3389/fnins.2016.00249
Garland, E. C., Lilley, M. S., Goldizen, A. W., Rekdahl, M. L., Garrigue, C., Noad, M. J. (2012). Improved versions of the Levenshtein distance method for comparing sequence information in animals' vocalisations: tests using humpback whale song. Behaviour, 149: 1413–1441. doi: 10.1163/1568539X-00003032
Garland, E. C., Noad, M. J., Goldizen, A. W., Lilley, M. S., Rekdahl, M. L., Garrigue, C., Constantine, R., Hauser, N. D., Poole, M. M., Robbins, J. (2013). Quantifying humpback whale song sequences to understand the dynamics of song exchange at the ocean basin scale. The Journal of the Acoustical Society of America, 133: 560-569. doi: 10.1121/1.4770232
Giacoma, C., Sorrentino, V., Rabarivola, C., Gamba, M. (2010). Sex differences in the song of *Indri indri*. International Journal of Primatology, 31: 539-551. doi: 10.1007/s10764-010-9412-8
Gil, D., Slater, P. J. B. (2000). Song organisation and singing patterns of the willow warbler, *Phylloscopus trochilus*. Behaviour, 137: 759-782. doi: 10.1163/156853900502330
Gooskens, C. and Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Language variation and change, 16: 189-207. doi: 10.1017/S0954394504163023
Gustison, M. L., Semple, S., Ferrer-i-Cancho, R., Bergman, T. J. (2016). Gelada vocal sequences follow Menzerath's linguistic law. Proceedings of the National Academy of Sciences, 113: E2750-E2758. doi: 10.1073/pnas.1522072113

Helweg, D. A., Cato, D. H., Jenkins, P. F., Garrigue, C., McCauley, R. D. (1998). Geographic variation in South Pacific humpback whale songs. Behaviour, 135: 1-27. doi: 10.1163/156853998793066438
Honda, E. and Okanoya, K. (1999). Acoustical and syntactical comparisons between songs of the white-backed Munia (*Lonchura striata*) and its domesticated strain, the Bengalese finch (*Lonchura striata* var. *domestica*). Zoological Science, 16: 319-326. doi: 10.2108/zsj.16.319
Kershenbaum, A. and Garland, E. C. (2015). Quantifying similarity in animal vocal sequences: which metric performs best? Methods in Ecology and Evolution, 6: 1452-1461. doi: 10.1111/2041-210X.12433
Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., et al. (2016.) Acoustic sequences in non-human animals: a tutorial review and prospectus. Biological Reviews, 91:13-52. doi: 10.1111/brv.12160
Kohonen, T. (1985). Median strings. Pattern Recognition Letters, 3: 309-313. doi: 10.1016/0167-8655(85)90061-3
Krull, C. R., Ranjard, L., Landers, T. J., Ismar, S. M., Matthews, J. L., Hauber, M. E. (2012). Analyses of sex and individual differences in vocalizations of Australasian gannets using a dynamic time warping algorithm. The Journal of the Acoustical Society of America, 32: 1189-98. doi: 10.1121/1.4734237
Levin, R. N. (1996). Song behaviour and reproductive strategies in a duetting wren, *Thryothorus nigricapillus*: I. Removal experiments, Animal Behaviour, 52: 1093-1106. doi: 10.1006/anbe.1996.0257
Levin, R. N. (1996). Song behaviour and reproductive strategies in a duetting wren, *Thryothorus nigricapillus*: II Playback experiments, Animal Behaviour, 52: 1107-1117. doi: 10.1006/anbe.1996.0258
Maretti, G., Sorrentino, V., Finomana, A., Gamba, M., Giacoma, C. (2010). Not just a pretty song: an overview of the vocal repertoire of *Indri indri*. Journal of Anthropological Sciences, 88: 151-165.
Margoliash, D., Staicer, C. A., Inoue, S. A. (1991). Stereotyped and plastic song in adult indigo buntings, *Passerina cyanea*. Animal Behaviour, 42: 367-388. doi: 10.1016/S0003-3472(05)80036-3
Pollock, J.I. (1979). Spatial distribution and ranging behavior in lemurs. The study of prosimian behavior, 359-409.
Sonnenschein, E., and Reyer, H.U. (1983). Mate-guarding and other functions of antiphonal duets in the slate-coloured boubou (*Laniarius funebris*). Zeitschrift für Tierzuchtung und Zuchtungsbiologie, 63:112-140. doi: 10.1111/j.1439-0310.1983.tb00083.x
Sorrentino, V., Gamba, M., Giacoma, C. (2012). A quantitative description of the vocal types emitted in the indri's song. Leaping ahead: advances in prosimian biology. Springer, New York 315-322. doi: 10.1007/978-1-4614-4511-1_35
Takahasi, M., Yamada, H., Okanoya, K. (2010). Statistical and Prosodic Cues for Song Segmentation Learning by Bengalese Finches (*Lonchura striata* var. *domestica*). Ethology, 116: 481–489. doi: 10.1111/j.1439-0310.2010.01772.x
Torti, V., Gamba, M., Rabemananjara, Z. H., Giacoma, C. (2013). The songs of the indris (Mammalia: Primates: Indridae): contextual variation in the long-distance calls of a lemur. Italian Journal of Zoology, 80: 596-607. doi: 10.1080/11250003.2013.845261
Torti, V., Bonadonna, G., De Gregorio, C., Valente, D., Randrianarison, R. M., Friard, O., Pozzi, L., Gamba, M., Giacoma, C. (2017). An intra-population analysis of the indris' song

dissimilarity in the light of genetic distance. Scientific reports, 7: 10140. doi: 10.1038/s41598-017-10656-9

Tougaard, J and Eriksen, N. (2006) Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure. Behaviour, 143: 239-252. doi: 10.1163/156853906775900685

Wieling, M., Montemagni, S., Nerbonne, J., Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. Language, 90: 669-692. doi: 10.1353/lan.2014.0064

Melody Matters: An Acoustic Study of Domestic Cat Meows in Six Contexts and Four Mental States

Susanne Schötz[1], Joost van de Weijer[2], Robert Eklund[3]
[1] Department of Logopedics, Phoniatrics and Audiology, Lund University, Sweden
[2] Lund University Humanities Lab, Sweden
[3] Department of Culture and Communication, Linköping University, Sweden

Corresponding Author:
Susanne Schötz
Dept. of Logopedics, Phoniatrics, and Audiology, Lund University, SE-221 85 Lund, Sweden
Email address: susanne.schotz@med.lu.se

Abstract
This study investigates domestic cat meows in different contexts and mental states. Measures of fundamental frequency (f0) and duration as well as f0 contours of 780 meows from 40 cats were analysed. We found significant effects of recording context and of mental state on f0 and duration. Moreover, positive (e.g. affiliative) contexts and mental states tended to have rising f0 contours while meows produced in negative (e.g. stressed) contexts and mental states had predominantly falling f0 contours. Our results suggest that cats use biological codes and paralinguistic information to signal mental state.

Introduction
Acoustic cues to paralinguistic information like a human speaker's physical and emotional state can be found in fundamental frequency (f0), intensity and duration (see e.g. Gangamohan, Kadiri, & Yegnanarayana, 2016). Some of these cues are related to so called biological codes, which can be observed in humans as well as nonhuman species. An example is that according to the 'frequency code' high f0 indicates smallness, submission, friendliness, and uncertainty, while low f0 signals largeness, dominance, aggressiveness, and certainty (Morton, 1977; Ohala, 1983; Gussenhoven, 2016). Animals are able to experience and express emotions (Bekoff, 2007, p. 42; Briefer, 2012), and as a consequence, it is reasonable to expect that their physical and mental state influences their vocalisations to include paralinguistic information found in f0 and duration.

Domestic cats (*Felis catus*) are – next to dogs (*Canis lupus familiaris*) – the most common companion animals in the world. Over 600 million cats are said to live with humans worldwide (Saito, Shinozuka, Ito, & Hasegawa, 2019). Cats have developed an extensive, variable and complex vocal repertoire, probably best explained by their social organisation, their nocturnal activity and the long period of association between mother and young (Bradshaw, Casey, & Brown, 2012). Moreover, as a consequence of their interaction with human beings, cats have learned to vary and nuance their voices ever since they were domesticated, approximately 9500 years ago (Vigne, Guilaine, Debue, Haye, & Gérard, 2004).

Cat–human communication is considered to be understudied (Saito et al., 2019). The findings of only a few studies on the topic suggest that the acoustics of cat vocalisations vary depending on the context, and the cats' emotional state. Brown, Buchwald, Johnson, & Mikolich (1978) compared sounds from kittens and adult cats in isolation, food deprivation, pain, threat, acute threat and kitten deprivation and found differences in duration, initial and peak f0. Nicastro (2004) found acoustic differences (duration and mean and max f0, first and second formant, and spectral tilt) between meows produced by domestic cats and African wild cats (*F. silvestris lybica*) in food-related, agonistic, affiliative, obstacle and distressing contexts. Yeon et al. (2011) analysed domestic cat vocalisations (growls, hisses and meows) produced by domestic and feral cats in one affiliative and four agonistic contexts and found differences in duration, mean fundamental and peak frequency. Schötz and van de Weijer (2014), finally, compared f0 of domestic cat meows in food- and vet-related contexts and found a predominance of rising contours in food-related contexts, and of falling contours in vet-related contexts, as well as larger f0 standard deviation in food-related meows.

In the present study we compare duration and f0 in meow vocalisations by domestic cats in six different contexts and four mental states. We hypothesised that cats use biological codes to convey paralinguistic-like information like emotion and intention depending on the context in which the cat was recorded and on their mental state.

Materials and Methods

The collected material consisted of audio and video recordings of 58 cats interacting in everyday contexts with humans (mainly their owners, but occasionally with one of the experimenters). The recordings were made using a GoPro Hero 4 Session video camera and a Roland R-09HR WAVE/MP3 recorder with Sony ECM-AW4 Bluetooth wireless microphones attached to collars worn by the cats. In addition, whenever a cat did not accept to wear the collar or when owners recorded and sent us videos recorded by them privately, other equipment (e.g. cell phones) was occasionally also used. Care was always taken to place or hold the microphone as close to the cats' mouths as possible without disturbing their natural behaviour. Audio files (unless recorded using the Roland R-09HR) were extracted from the video files as 44.1 kHz, 16 bit WAV files.

The material used in this study was recorded in one of the following six contexts: while waiting at a door (or a window) (*door*), while approaching a befriended human or cat (*greeting*), while soliciting or receiving food (*food*), while soliciting or during play (*play*), while being lifted (*lifting*) or while being in a cat carrier (*transport box*). Of these, the first five were relatively positive contexts while the last one generally was relatively negative for the cats. The mental state of the cats was classified as *attention seeking, content, discontent* or *stressed* based primarily on visual cues of the body, head and tail posture and movements (see e.g. Bradshaw & Cameron-Beaumont, 2000, pp. 73–74). Finally, each vocalisation was classified as either a *meow, trill, growl, hiss, howl, snarl, purr* or *chirp* (or a combination of two types), as described in Schötz (2018, pp. 254–257). Naturally, not all cats produced vocalisations in all contexts or mental states.

The type of vocalisation, recording context and mental state were all annotated with the speech analysis tool Praat (Boersma & Weenink, 2019) by the first author. A randomly selected sample of the files was independently annotated by the second author to estimate agreement in the type of vocalisations that the cats produced. Results showed varying degrees of agreement between the two labellers with kappa values ranging from 0.43 to 0.97 with an average of 0.70.

The most common human-directed vocalisation in our recording collection was the meow, defined as a voiced sound generally produced with an opening-closing mouth and containing a combination of two or more vowel sounds (e.g. [eo] or [iau]) with an occasional initial [m] or [w] (after Schötz, 2018). A total of 780 meows produced by 40 cats (22 females and 18 males, aged 1–12;6 years) were selected for acoustic analysis in this study. For all tokens, measures of f0 (maximum, minimum, mean and standard deviation (sd)) as well as duration were obtained. Additionally, F0 contours were generated using Praat Pitch Objects and manually corrected when necessary. To facilitate between-cat comparison, the contours were normalised by setting the minimum f0 for every meow to 0 semitones (st). Mean contours were obtained for each context and mental state by averaging f0 measured at 100 evenly distributed points in each meow. Differences between meows produced in different contexts and mental states were compared through visual inspection of the mean f0 contours as described below. Figure 1 shows an example of individual f0 contours and the corresponding mean f0 contour for the context *play*.

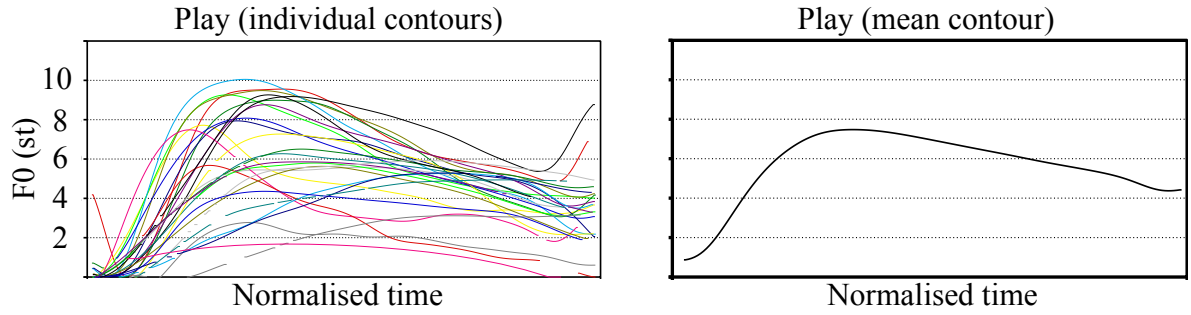

Figure 1.: Individual and average f0 contours for the context *play*.

Results

Duration and f0

Table 1 shows mean acoustic values in the different contexts and mental states. Differences between contexts and mental states were analysed for f0 mean, f0 sd, and duration (f0 minimum and maximum were not analysed as they highly correlated with f0 mean, and f0 range was not analysed as it highly correlated with f0 sd). The analysis was done in two steps. First, we performed mixed effects regression analyses to obtain an overall typical value for each cat across all contexts. Subsequently, these estimated values were subtracted from the values for each meow resulting in a positive number for a meow produced with a relatively high parameter value and a negative number for a meow with a relatively low parameter value. The resulting values were analysed using mixed effects regression with context and mental state as fixed effects and random intercepts for the different cats.

Table 1: Acoustic measurements (mean values).

		duration	*f0 (Hz)*				
Context	*n*	*(ms)*	*min*	*max*	*mean*	*range*	*sd*
door	75	754	601	712	661	111	30
food	341	728	501	641	581	140	38
greeting	61	670	395	542	484	148	44
lifting	20	724	575	720	654	145	39
play	27	561	318	444	393	124	36
transport box	165	932	484	617	546	133	33
Mental state							
attention	487	719	478	618	559	140	40
content	52	545	414	551	495	137	40
discontent	150	843	493	609	554	117	29
stressed	78	912	520	671	579	151	39

For contexts, we found that meows produced in *food* contexts were characterized by relatively high mean f0 (EST = 13.914, SE = 4.863, $t = 2.861$, $p = 0.006$) and short duration (EST = –31.94, SE = 13.27, $t = –2.407$, $p = 0.023$). On the contrary, meows produced by cats in a *transport box* were characterized by low mean f0 (EST = –26.988, SE = 6.846, $t = –3.942$, $p = 0.000$) and long duration (EST = 71.84, SE = 19.11, $t = 3.759$, $p = 0.001$). Meows produced in *door* contexts were relatively high in mean f0 (EST = 20.105, SE = 9.833, $t = 2.045$, $p = 0.044$), and meows produced in *play* contexts were characterized by low f0 variability (EST = –9.248, SE = 4.134, $t = –2.237$, $p = 0.026$). The remaining effects were all not significant.

For mental states, meows produced by *stressed* cats showed low average f0 (EST = –29.329, SE = 8.080, $t = –3.630$, $p = 0.000$), and long durations (EST = 99.727, SE = 27.307, $t = 3.652$, $p = 0.000$). Finally, meows produced by *discontent* cats were (marginally) significantly lower in f0 variability (EST = –3.475, SE = 1.777, $t = –1.956$, $p = 0.051$). All remaining effects were not significant.

F0 contours

Figure 2 shows mean f0 contours for the six contexts and the four mental states. The f0 contours for the meows in the positive (affiliative) contexts *door, greeting, food*, *play* and *lifting* all display rising patterns — the clearest can be seen in *greeting* — sometimes combined with a later fall. In contrast, the average contour produced by cats in a *transport box* is falling. Similarly, the f0 contours for the positive mental states *attention* and *content* are rising, while those produced by cats who were *discontent* or *stressed* display falling patterns

Discussion and future studies

The results from this study suggest that cat vocalisations are influenced by the context in which they were recorded or the mental state of the cat. We found effects on average f0, f0 variation, duration and on the melody (f0 contours). Roughly summarized, we observed that meows produced in positive contexts (by cats with a positive mental state) were high in pitch, short in duration and had a rising melody, while those produced in negative contexts (by cats with a

negative mental state) were low in pitch, long in duration and had a falling melody. It should be noted that some contexts contained meows by very few cats, e.g. *play* (2 cats) and *lifting* (4 cats). In future studies a larger number of cats will be analysed in each context and mental state.

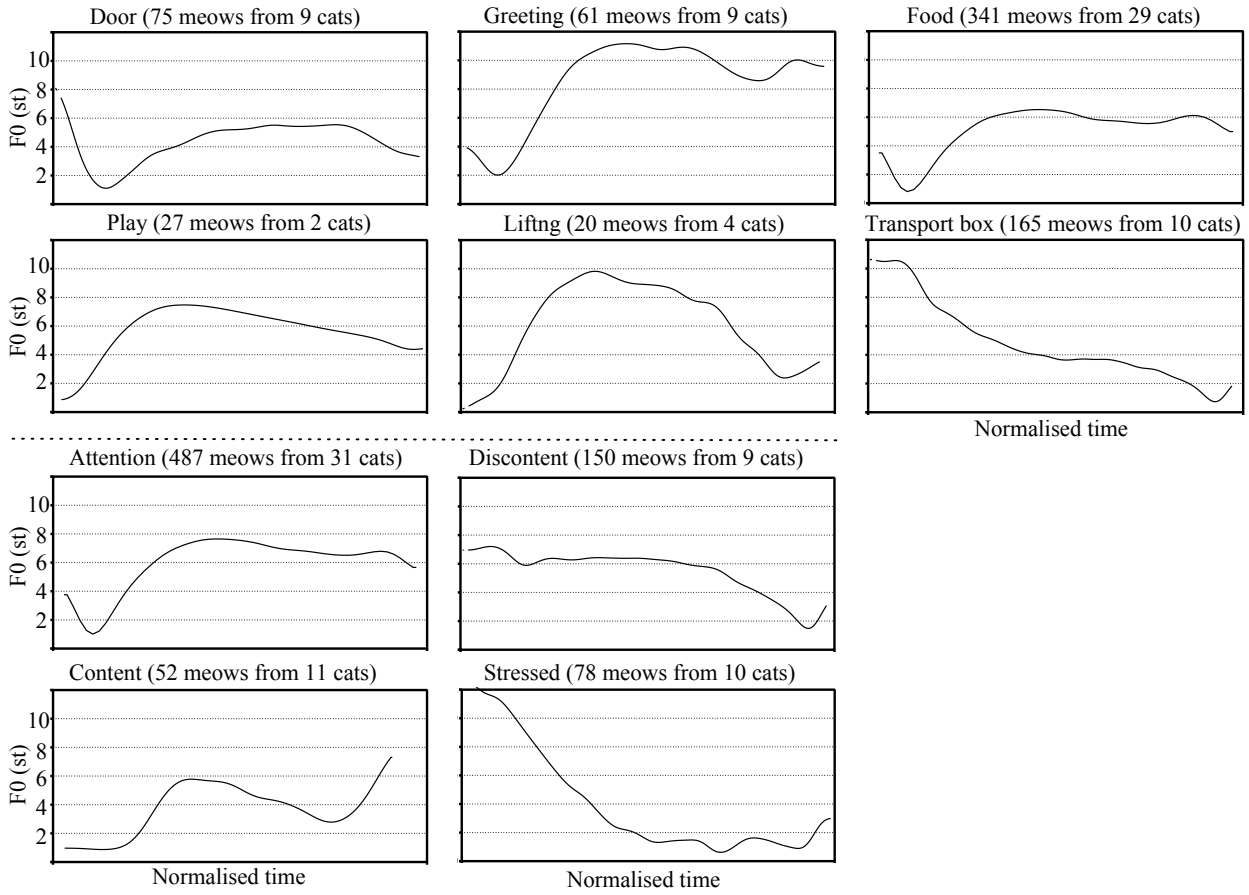

Figure 2. Mean f0 contours of meows from six contexts and four mental states (st: semitones).

A possible explanation of our findings is that cats use biological codes like the frequency code to vary the meaning of their vocalisations. Whether this is innate or a learned behaviour used mainly with humans is still unclear. We will investigate this in a future study by comparing human–directed and cat–directed vocalisations.

In order to understand the exact mechanism behind the paralinguistic variation in acoustic characteristics of meows we will need to explore the data further and include measures of intensity and voice quality. Other factors that potentially influence the acoustics of cat vocalisations need to be taken into consideration. Possible candidates are sex, age, weight, breed and level of emotional arousal. Environmental factors, such as the number of cats in a household, may also play a role.

Whether or not variation in f0 and duration can be used to assess the mental or emotional well-being of cats remains to be tested. Rising patterns, in that case, are likely to indicate contentment, while falling patterns signal stress or discontentment. Additionally, meows were far from the only type of vocalisation in our collection, which also included trills, growls, hisses,

howls, snarls, purrs or chirps, and also combinations of two vocalisation types. Our next step in trying to chart the vocal system of the cat will be to subject these other vocalisation types to similar acoustic analyses to see whether we find effects of context and mental state there as well.

Acknowledgements

The authors gratefully acknowledge the Marcus and Amalia Wallenberg Foundation and Lund University Humanities Lab. A special thanks goes to all participating cats and their owners.

References

Bekoff, M. (2007). Are you feeling what I'm feeling? *New Scientist*, *26 May 2007*, 42.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer (Version 6.0.46) [Computer program]*.

Bradshaw, J., & Cameron-Beaumont, C. (2000). 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 University Press.

Bradshaw, J., Casey, R. A., & Brown, S. L. (2012). *The behaviour of the domestic cat*. CABI Publishing.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence: Vocal communication of emotions. *Journal of Zoology*, *288*(1), 1–20.

Brown, K. A., Buchwald, J. S., Johnson, J. R., & Mikolich, D. J. (1978). Vocalization in the cat and kitten. *Developmental Psychobiology*, *11*, 559–570.

Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2016). Analysis of Emotional Speech—A Review. In A. Esposito & L. C. Jain (Eds.), *Toward Robotic Socially Believable Behaving Systems - Volume I* (Vol. 105, pp. 205–238).

Gussenhoven, C. (2016). Foundations of Intonational Meaning: Anatomical and Physiological Factors. *Topics in Cognitive Science*, *8*(2), 425–434. https://doi.org/10.1111/tops.12197

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, *111*(981), 855–869.

Nicastro, N. (2004). 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*, *118*(3), 287–296.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica*, *40*(1), 1–18.

Saito, A., Shinozuka, K., Ito, Y., & Hasegawa, T. (2019). Domestic cats (Felis catus) discriminate their names from other words. *Scientific Reports*, *9*(1).

Schötz, S. (2018). *The secret language of cats: How to understand your cat for a better, happier relationship*. Hanover Square.

Schötz, S., & van de Weijer, J. (2014). A study of human perception of intonation in domestic cat meows. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of Speech,Prosody, 23–23 May, 2014, Dublin, Ireland*. Dublin.

Vigne, J.-D., Guilaine, J., Debue, K., Haye, L., & Gérard. (2004). Early taming of the cat in Cyprus. *Science*, *304*(5668), 259–259.

Yeon, S. C., Kim, Y. K., Park, S. J., Lee, S. S., Lee, S. Y., Suh, E. H., … Lee, H. J. (2011). Differences between vocalization evoked by social stimuli in feral cats and house cats. *Behavioural Processes*, *87*(2), 183–189.

Call overlapping signals sexual status in Darwin's frogs

José Manuel Serrano[1,2], Noé Guzmán[3], Mario Penna[1], Marco Méndez[4], Claudio Soto-Azat[5]

[1] Programa de Fisiología y Biofísica, Facultad de Medicina, Universidad de Chile, Santiago, Chile
[2] Laboratorio de Comunicación Animal, Facultad de Ciencias Básicas, Universidad Católica del Maule, Talca, Chile.
[3] Laboratorio de Fisiología Animal, Facultad de Ciencias, Universidad de Chile, Santiago, Chile
[4] Laboratorio de Genética y Evolución, Facultad de Ciencias, Universidad de Chile, Santiago, Chile
[5] Sustainability Research Centre, Life Sciences Faculty, Universidad Andres Bello, Santiago, Chile

Corresponding Author:
José M. Serrano[2]
San Miguel 3605, Talca, Región del Maule, 3480112, Chile.
Email address: jserrano@ucm.cl

Abstract

Background. In animal reproductive contexts, calling behaviour is mostly performed by males but in species in which females call, it is not known how vocal interaction occurs between sexes, particularly when sexual dimorphism in signals is low, as in cases in which call repertoire is identical but acoustic properties differ. In Darwin's frog (*Rhinoderma darwinii*), a species in which males brood larvae inside their vocal sacs, females have higher dominant frequency and shorter calls and notes than males. Since in this species males persist calling after getting pregnant with larvae, different vocal interaction patterns are expected to occur among animals having dissimilar reproductive status.
Methodology. To explore the mechanisms underlying vocal recognition among the different sexual status of *R. darwinii*, we recorded natural duets between non-pregnant males (NPM), pregnant males (PM) and females (F) and evaluated their evoked vocal response to natural playback stimuli of each sexual status from November to February 2015-2016 in Chiloé island, Chile. Call rate, phase angles, sound pressure level (SPL), number of overlapping calls and delay of overlapping calls were measured to determine differential responses between natural duets and in response to stimuli consisting of natural calls of individuals of different sexual status.
Results. Spontaneous duet interactions occurred mainly between males and no clear differences between duets were detected. In playbacks, call ratios in response to calls of different sexual status were similar. Females decreased their SPL in response to F calls, while F and PM had longer call delays and lower call overlaps between each other. Major differences were observed

in call overlap, as the occurrence of this phenomenon was larger in playback experiments than during natural duets. The number of calls overlapped during natural duets was fewer (10.9 %) than during playback experiments (36.8 %).
Conclusions. Our results suggest that in *R. darwinii*, PM and F signalize their sexual status by decreasing their call overlap and that NPM respond indistinctly to the other sexual status. In general, these differences in selective call overlap between Darwin's frogs arise as a novel mechanism for signal recognition between animal vocal interactions.

Introduction

The display of sexual signals has been mostly considered an exclusive feature of males (Price 2015), however, growing evidence has shown that females can display sexual signals in various taxa (e.g. Serrano and Penna 2018), questioning their exclusive role as mediators of female choice and competition between males (e.g. Tobias et al., 2012). In addition, the study of duets and choruses formed by males and females may contribute new explanations about the role of signal exchanges in social and sexual processes (e.g. Janik and Slater 1998; Cui et al. 2010; Fishbein et al. 2018). In this regard, a largely unexplored issue in animal communication is how the timing of acoustic signals involved in recognition of conspecifics contribute to group cohesion in complex societies (Sheehan and Bergman 2016).

Darwin's male frogs brood in their vocal sac larvae collected from eggs laid by females and fertilized by males (Goicoechea et al. 1986). In the field, Darwin's frogs usually call isolated, in pairs or in small groups on moss mounds on undergrowth in temperate forest environments (Crump 2002). However, the occurrence of sexual and social interactions within and between sexes in Darwin's frogs have not yet been determined. Recently advertisement call of this species has been shown to possess a sexual dimorphism related to body size differences between males and females but lacks clear differentiation between males with different pregnancy status (Serrano 2019). The aim of the current study is to understand the role of vocal signalling for sexual recognition in a social environment conformed by males and females. It also expects shed light to understand the role of vocal interaction in a social environment conformed by male individuals with distinct sexual status. In this study we evaluate the hypothesis that Darwin's frogs recognize their sexual identity by means of their calls, by recording natural vocal interactions between individuals of different sexual status and conducting evoked vocal response (EVR) experiments with stimuli representing the diverse sexual status.

Materials & Methods

We describe patterns of vocal interactions in a social environment conformed by pregnant males PM), non-pregnant males (NPM) and females (F), recording natural duets between animals having these three status during the reproductive season lasting five months (October 2015 to February 2016) in a population located on the Island of Chiloé, Chile (43° 21′ S; 74° 6′ W). In addition, we evaluated EVR to playbacks of natural calls of individuals of the three sexual status.

Duet recordings

Vocalizations of subjects calling in duets were recorded with a digital recorder (Tascam DR-100) at a sampling rate of 44.1 kHz and 16-bit resolution and two directional microphones (Sennheiser ME-66) plugged to each recording channel. The distance separating the two subjects intervening in the duet was measured and sound pressure level (SPL re 20 µPa, C frequency weighting and fast time weighting) of calls of one individual conforming the duet was recorded placing a sound level meter microphone (Extech 407780) adjacent to the tip of the directional microphone. Latency (registered as phase angle of call onsets between the calls of the two individuals; Klump and Gerhardt 1992), number of call overlaps and delay between the onset of overlapping calls between interacting subjects were measured. To discard that call overlap was occurring by chance between pairs of individuals composing a duet, number of overlaps and overlap delay between duets was compared using generalized linear models (GLM).

Playback experiments

Call bouts of playback stimuli were composed of 10 natural calls of individuals of the three sexual status having a high signal to noise ratio. The amplitude of call bouts was standardized at 64 dB SPL at the position of the subjects and time intervals between successive calls within a call bout were generated with random intervals of silence lasting 5 – 60 s. These values approximate those occurring in natural interactions between individuals of Darwin's frog. Following this procedure, bouts of calls having different call rates and lasting 138 – 399 s resulted. This randomization in call timing allowed to evaluate the temporal relationship of the EVR to the stimuli, independent of potential rythmic calling behaviour based on an internal oscillator (Zelick and Narins 1985). Three-minute silent intervals spaced call bouts of the different stimuli and the order of presentation of call bouts of each sexual status was randomized. Stimuli were presented with a Samsung J1 WAV player connected via Bluetooth to a portable loudspeaker (i.Sound 5464). Spontaneous vocal activity of the experimental subjects was recorded and thereafter playbacks of calls of the three sexual status were presented sequentially through a loudspeaker placed on moss vegetation at 1 m and at an angle of about 90 degrees relative to the focal subject. Upon completion of each playback experiment, identity of focal individuals was registered.

Responses to natural stimuli were analyzed measuring call rate, latency, SPL and number of overlaps of response calls with the stimuli as for duets. Call rate and SPL were computed for periods of silence and stimuli presentation, while latency and number of overlaps were measured only for stimuli presentations. GLM, ANOVA and post-hoc tests were used to compare responses to stimuli of the three sexual status.

Results and Discussion

Duet recordings

Thirteen interactions were recorded between individuals belonging to the three sexual status: between NPM (N= 5), PM (N= 3), NPM and F (N = 4), and NPM and PM (N= 1). SPLs of the calls were not affected by the distance from the focal individual at which this variable was measured (range= 18 to 68 cm), as no significant correlation between this amplitude measure and recording distance occurred (n= 18; r= -0.13; df= 16; p= 0.612). No clear pattern was observed between the different kinds of duets regarding call rates, phase angles and SPL. Ratios between overlapping relative to non-overlapping calls were different from chance in duets composed by NPM, by NPM and F, and by PM (Table 1). As NPM was the only sexual status observed interacting with the two other sexual status, its responses to the three sexual status were

compared, showing that calls of this sexual status overlapped at an earlier time with F than with both types of males (ANOVA test, Chi^2= 6.972; p <0.05; Fig. 1).

Table 1. Percentage of overlapping calls in vocal interactions between non-pregnant males (NPM), pregnant males (PM) and females (F) of Darwin's frog in natural duets and with calls of these three sexual status in playback experiments. Significant differences between the numbers of overlapping and non-overlapping calls: *= p<0.05; **= p<0.01; nr= duets not recorded.

Interaction	% of overlapping calls between sexual status								
Focal individual	NPM	NPM	NPM	PM	PM	PM	F	F	F
Preceding caller or stimuli	NPM	PM	F	NPM	PM	F	NPM	PM	F
Natural duets	11.4**	16.3	12.3**	3.7*	10.6**	nr	22	0	nr
Playback experiments	29.8	48.8	42.1	42.1	36.4	5*	53.3	39.5	33.9

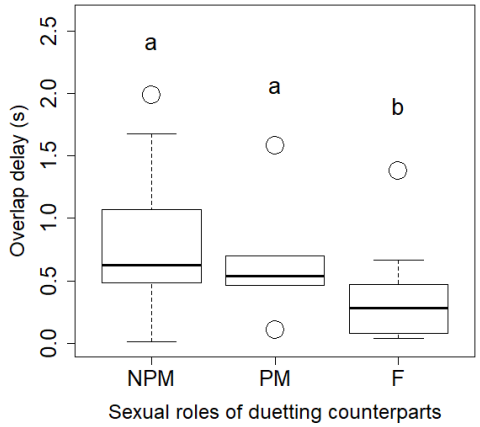

Fig. 1. Delays between the onsets of overlapping calls for duets between non-pregnant males in response to non-pregnant males (NPM), pregnant males (PM) and females (F). Different low-case letters (a, b) indicate significant differences in post-hoc analyses (Tukey tests, p <0.05).

Playback experiments

Thirty-two individuals were stimulated with natural calls, 14 of which were NPM, 12 PM and six F. Call ratios in response to calls of different sexes were similar. However, F decreased their SPL in response to F calls relative to the initial silent period (GLM test, t=3.136; p<0.01; Fig. 2A) and had longer latency to PM relative F and NPM stimuli (GLM test, t= -2.573; p< 0.05; Fig. 2B), while PM had lower number of overlapping calls to F relative to NPM and PM (GLM test, z= -1.691; p<0.05; Fig. 2C).

The occurrence of call overlaps was larger in playback experiments relative to duet interactions (GLM test, z= 8.11; p< 0.001; Table 1) and overlap delay in response to all stimuli differed between sexual status, as F responded with a shorter overlap delay to all the stimuli combined relative to both types of males (ANOVA test, Chi^2= 7.107; p< 0.05; Fig. 3). Such short

overlap delay of F during playback constrasts with the lower overlap delay observed also in F duettings (Fig. 1; see in Table 1 that F were observed overlapping calls with NPM only).

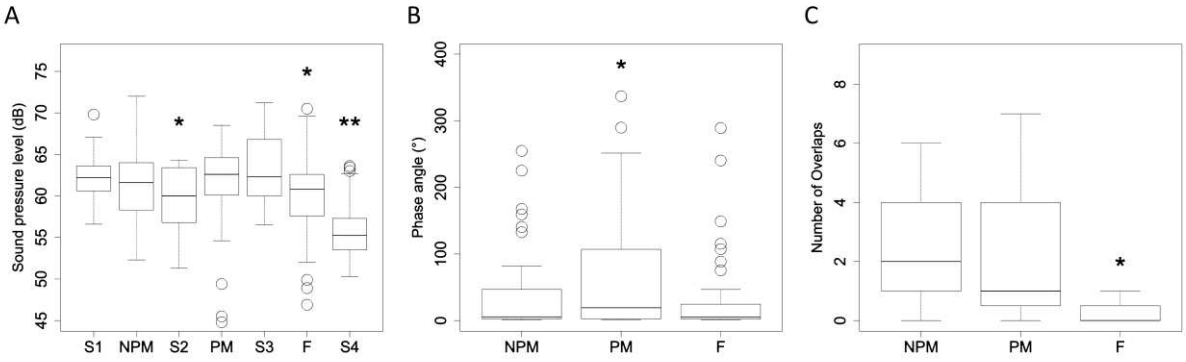

Fig. 2. Sound pressure level (A) and phase angle (B) in evoked calls of females, and number of overlaps (C) in evoked calls of pregnant males in response to natural stimuli of the three sexual status. Stimuli abbreviations: NPM: non-pregnant males, PM: pregnant males, F: females. S1, S2, S3 and S4: silent intervals between stimuli presentations. Asterisks indicate significant differences in post-hoc analyses relative to S1 in A, and between stimuli in B and C (Tukey tests, *: p<0.05, **: P<0.01)

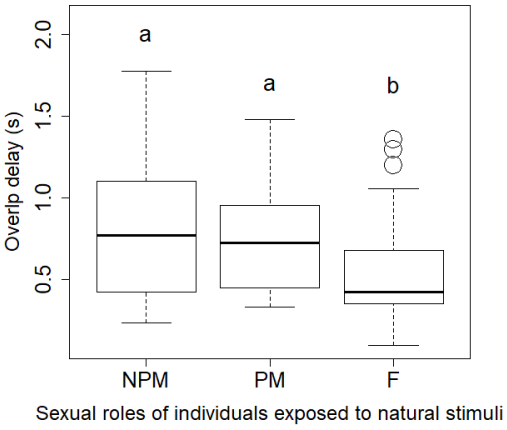

Fig. 3. Overlap delays observed for non-pregnant males (NPM), pregnant males (PM) and females (F) in response to playbacks of all natural stimuli combined. Different low-case letters (a, b) indicate significant differences in post-hoc analyses (Tukey test, p <0.05).

Our results suggest that in Darwin's frogs signal recognition is not evinced in gross measures of vocal activity such as call rate like it occurs in other species (e.g. Cui et al. 2010; Fishbein et al. 2018). However, subtle differences in call overlap apparently indicate dissimilar readiness to interact vocally between individuals of different sexual status. PM and F of *R. darwinii* are relatively selective in their modes of synchronization with calls of different sexual status and NPM interact similarly with all the sexual status, a strategy likely to favour spatial tolerance of potential breeding partners. These differences in selectivity in call overlap between

duetting pairs may contribute a novel mechanism of sexual recognition that could be relevant for acoustic interactions among other organisms and artificial devices.

Acknowledgements

Mara Santoyo, Nicolette Thompson, Matías Muñoz, María Luisa Estay and Jaime Beltrand helped in obtaining field data.

References

Cui, J., Wang, Y., Brauth, S. and Tang, Y. (2010). A novel female call incites male-female interaction and male-male competition in the Emei music frog, *Babina daunchina*. Animal Behaviour, 80: 181–187.

Crump, M. L. (2002). Natural history of Darwin's frog, *Rhinoderma darwinii*. Herpetolical Natural History 9: 21–30.

Fishbein, A. R., Löschner, J., Mallon, J. M. and Wilkinson, G. S. (2018). Dynamic sex-specific responses to synthetic songs in a duetting suboscine passerine. PLoS ONE 13: e0202353.

Goicoechea, O., Garrido, O. and Jorquera, B. (1986). Evidence for a trophic paternal-larval relationship in the frog *Rhinoderma darwinii*. Journal of Herpetology, 20:168–178.

Janik, V. M. and Slater, P. J. (1998). Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls. Animal Behaviour, 56: 829-838.

Klump, G. M. and Gerhardt, H. C. (1992). Mechanisms and function of call-timing in male-male interactions in frogs. In: McGregor, P. K. (ed). Playback and studies of animal communication. New York, USA: Springer Science & Business Media, p. 153-174.

Price, J. J. (2015). Rethinking our assumptions about the evolution of bird song and other sexually dimorphic signals. Frontiers in Ecology and Evolution, 3: 40.

Serrano, J. M. (2019). El rol de las señales acústicas en las interacciones sexuales y la estructura social de la ranita de Darwin (*Rhinoderma darwinii*). Doctoral Dissertation, Universidad de Chile.

Serrano, J. M. and Penna, M. (2018). Sexual monomorphism in the advertisement calls of a Neotropical frog. Biological Journal of the Linnean Society, 123: 388–401.

Sheehan, M. J. and Bergman, T. J. 2016. Is there an evolutionary trade-off between quality signaling and social recognition? Behavioral Ecology 27: 2–13.

Tobias, J. A., Montgomerie, R. and Lyon, B. E. (2012). The evolution of female ornaments and weaponry: social selection, sexual selection and ecological competition. Philosophical Transactions of the Royal Society B, 367: 2274–2293.

Zelick, R. and Narins, P. M. (1985). Characterization of the advertisement call oscillator in the frog *Eleutherodactylus coqui*. Journal of Comparative Physiology A, 156: 223–229.

Development and application of a robotic zebra finch (RoboFinch) to study multimodal cues in vocal communication

Ralph Simon[1,2], Judith Varkevisser[2], Ezequiel Mendoza[3], Klaus Hochradel[4], Constance Scharff[3], Katharina Riebel[2] & Wouter Halfwerk[1]

[1]Department of Ecological Science VU University Amsterdam, Amsterdam, The Netherlands
[2]Institute of Biology Leiden, Leiden University, Leiden, The Netherlands
[3]Institut für Biologie, Freie Universität Berlin, Berlin, Germany
[4]Institute of Measurement and Sensor Technology, UMIT-Private University for Health Sciences, Medical Informatics and Technology GmbH, Hall in Tirol, Austria

Corresponding Author: ralph.simon@vu.nl

Abstract

Understanding animal behaviour through psychophysical experimentation is often limited by insufficiently realistic stimulus representation. Important physical dimensions of signals and cues, especially those that are outside the spectrum of human perception, can be difficult to standardize and control separately with currently available recording and displaying techniques (e.g. video displays). Accurate stimulus control is in particular important when studying multimodal signals, as spatial and temporal alignment between stimuli is often crucial. Especially for audiovisual presentations, some of these limitations can be circumvented by the employment of animal robots that are superior to video presentations in all situations requiring realistic 3D presentations to animals. Here we report the development of a robotic zebra finch, called RoboFinch, and how it can be used to study vocal learning in a songbird, the zebra finch.

Introduction

The use of stationary and animated, robotic animal models knows a long tradition. Such models have been used in contexts as diverse as e.g. mate attraction signalling, predator-prey interactions or cooperation to investigate which stimulus properties trigger animals' reactions[1,2,3]. Advances in technology involving new materials, small-sized actuators, 3D printing techniques and more computational power have greatly increased the possibilities for a new generation of even more realistic robotic animal models[4], which can take the study of animal communication signals to the next level, especially in the context of multimodal signalling. This form of signalling where signals in one modality are either facultatively or obligatorily accompanied by signalling in one or more additional modalities, is widespread in animals especially in the context of mate attraction: Birds sing and dance, mammals show visual display, acoustic and chemical signals, many insects combine acoustic, vibratory and chemical signalling[5]. However, these signals are usually studied in one modality only, often owing to the technical problems involved in controlling more than one modality during stimulus presentations. Robotic models allow multimodal signal components to be controlled independently. This allows to expand the stimulus range and to produce artificial stimulus combinations testing receivers' reaction to different combinations of signal components. Robotic applications have already helped to understand multimodal stimulus processing in the

context of territory defence or sexual signalling (e.g. in some frog and bird species [6,4,7]) but could also open a research window into the development of the perception of multimodal signals. Here we outline the progress in designing a robotic bird, looking and singing like a songbird, the zebra finch *Taeniopygia guttata*. This species is an important model to study the behavioural and neurobiological aspects of vocal learning. A robotic zebra finch would allow studying the potential role of multimodal cues such as beak movements in vocal learning. To enable such studies, our goal was to create a realistic moving 3D-model of a singing bird showing the fast and sound-specific beak movements accompanying male song.

Development of the RoboFinch

To develop the basic form of the RoboFinch, we 3D scanned a taxidermic model of a zebra finch with a handheld 3D scanner (Eva, Artec3D, Luxembourg, Luxembourg). The beak was scanned with high resolution from a prepared skull (ATOS 5X, gom, Braunschweig, Germany, Fig. 1 A, B). These scans were combined in the program Catia V5R20 (Dassault Systèmes), where we also implemented the inner mechanics (Fig. 1 C). We printed the RoboFinch with stereolithography 3D printing (Form 2, Formlabs, Somerville, Massachusetts, US), which uses a laser to cure solid isotropic parts from a liquid photopolymer resin (Grey Pro, Formlabs Resin). The movement of the head and beak was controlled by coils we got from dismantling DigiBirds (Silverlit Toys Manufactory, Hongkong, China). The advantage of using those coils is that they are cost-effective, small and allow fast movements up to 100 Hz. The coils were controlled via a custom build controller board. All the movements and the sound were controlled via a data acquisition card (Measurement Computing USB-3101), which was connected to a small desktop PC (Intel NUC i5). All movements, sound and the schedule of the stimulus presentation was controlled by a custom made LabView (National Instruments) Program. We painted the 3D printed models by hand with mixes of Revell colours (Revell, Bünde, Germany), which we tried to closely fit to the colours of the plumage of live zebra finches. We measured the colour spectra of the RoboFinch and live males (n = 6) with a spectrometer (Flame, Ocean Optics, Largo, Florida, US). To create movement files, we did

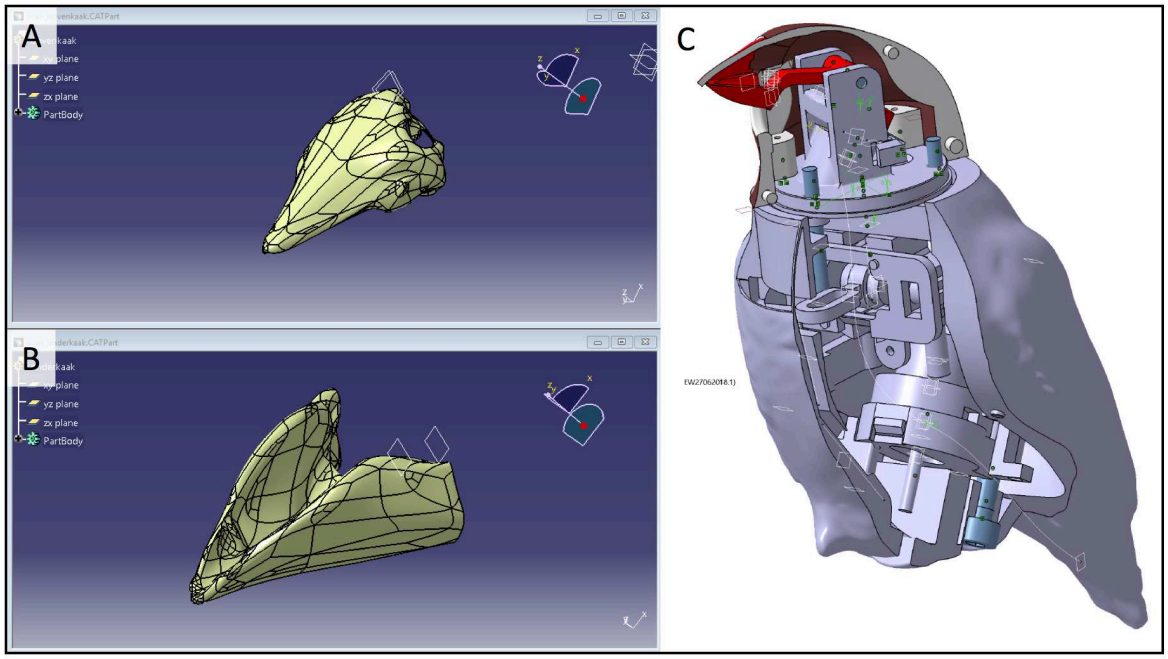

Figure 1. 3D Models of the RoboFinch. High resolution scan of (A) the upper and (B) the lower beak. (C) Catia Model of the RoboFinch showing the inner mechanics

high-speed video recordings (120fps) of singing male finches and deduced their head and beak movements with tracker software and played the recorded movements on the robot. The corresponding sound was played via a loudspeaker placed next to the robot.

Application of the RoboFinch in a tutoring situation

We tested the acceptance of the robots with two groups of young zebra finches, from 45-75 post-hatching, each group consisting of one male and one female. The setup consisted of a big cage with a black wall on one side. The black wall had a mesh window (20 x 15 cm), the robot was placed directly behind this window and the birds inside the cage could sit on a perch directly at that window. The birds were accustomed to the non-moving robot model for about 12 hours (afternoon and night) and then in the following morning (8 am) the robot started moving for the first time. The robot was programmed to move 6 times a day for half an hour displaying head and beak movements associated with short calls and song. We observed the birds with webcams (10fps), recorded their behavior and exemplarily analyzed sequences of video frames (600 greyscale frames, 1min) before and after the robot started moving. We did that for two sessions per day over the 6 initial days. As a measure of movement we used a frame differencing method and calculated the mean difference between adjacent frames (see Figure 2). Before the robot started moving, the young birds moved preferentially in the horizontal direction between the upper perches (the two upper images in Figure 2) indicating that the birds were not particularly interested in the robot when it wasn't moving and singing.

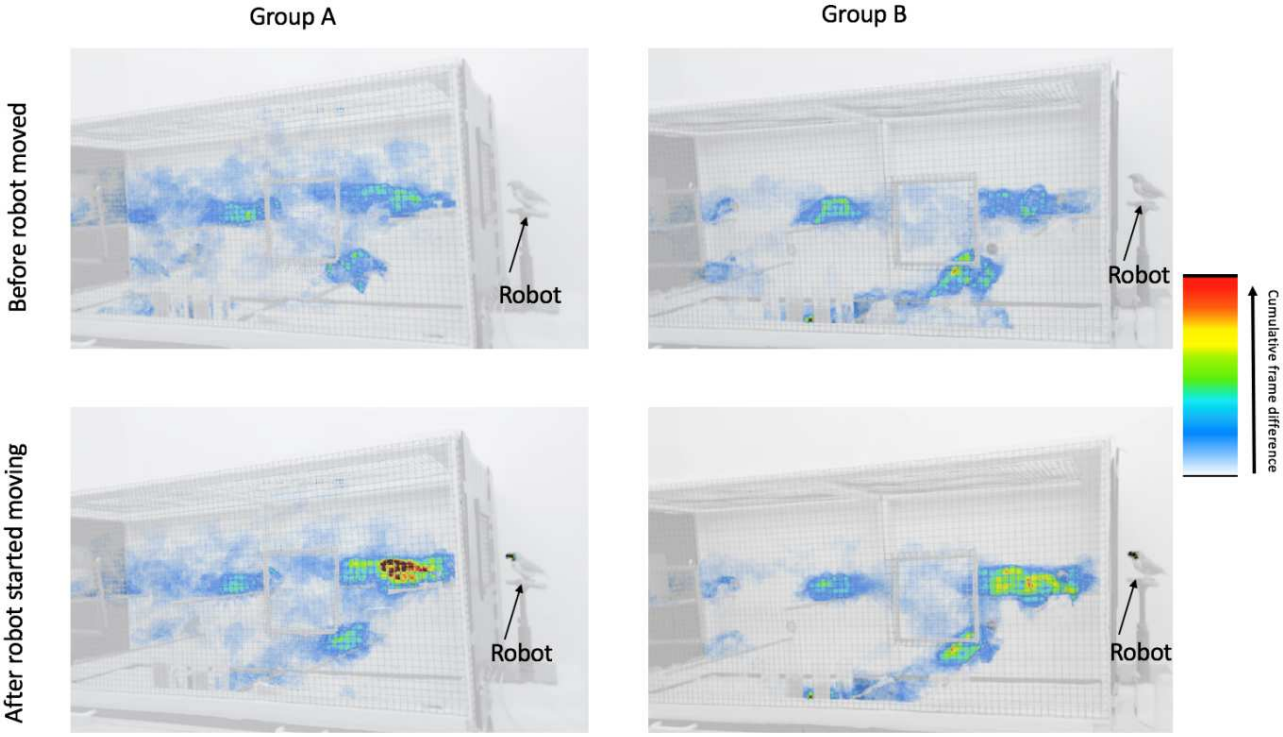

Figure 2. Images of the experimental cages of the two experimental groups of young birds summarising their movements in the cage with frame differencing. We plotted the frame differences in a colour gradation, the darker the red, the stronger were the cumulative pixel differences of adjacent frames and the stronger the movement. Each of the four plots is based on 12 events (2 tutor sessions on each of the 6 first exposure days) and for each event we analysed 600 frames (60s, 10fps). The images above are based on sequences right before the robot moved and vocalized, below directly after the robot started moving and vocalizung. Note that the movement of the robot is also visible in the lower images as dark dots.

As soon as the robot started moving and vocalizing however, the birds showed more interest and were approaching the robot as can be seen by the green to dark red pixels closer to the robot (Fig. 2, lower images, right end of the cage). Throughout the song tutoring, during the sensitive phase for song learning, the robot remained placed outside the cage, but at day 74 and 75 we also tested the acceptance of the robot in the cage and noticed the birds interacting with it, in particular throughout the phases when it was moving and singing (see Fig. 3).

Conclusion

Our work builds upon and adds to previous work suggesting that robotic models help to uncover the role of visual cues in song learning in zebra finches[8]. At least one behavioural experiment with zebra finches and a non-moving robot showed that the birds vocally interact with this model[9]. Moreover, plastic models are already used as tutors in zebra finch song learning experiments[10]. Our preliminary data demonstrates that our RoboFinch is accepted by young zebra finches nearly instantaneously and that they also interact with it. These observations suggest that the RoboFinch could successfully be used for song tutoring experiments. As we now are able to control visual and acoustic stimuli independently and find out how the combination of these can influence vocal learning, our experiments could help to uncover general principles of multimodal sensory integration in animal communication.

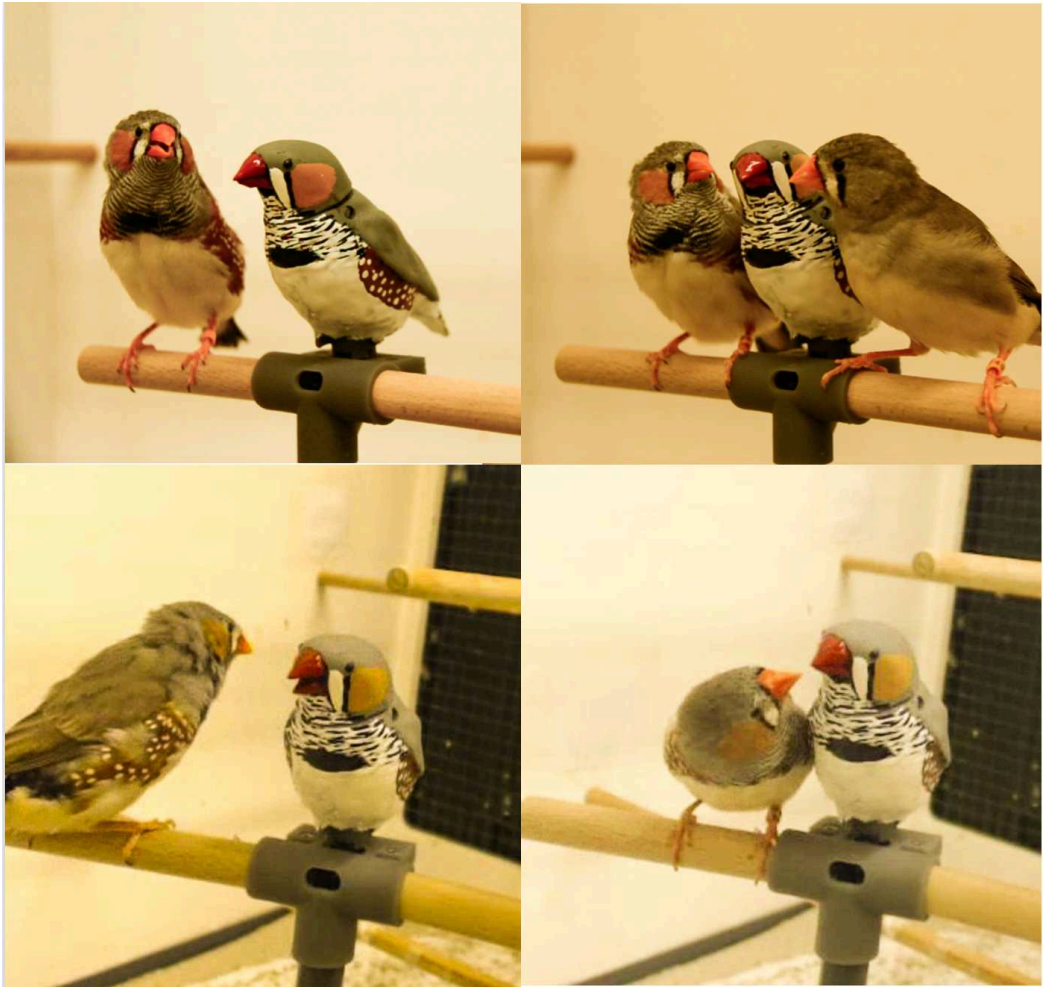

Figure 3. Young zebra finches interacting with the RoboFinch inside the cage.

Aknowledgements

We greatly aknowledge the help of Dré Kampfraath, Rogier Elsinga, Peter Wiersma and Wesley Delmeer during the development of the RoboFinch.

References

1 Krause, J., Winfield, A. F. & Deneubourg, J.-L. Interactive robots in experimental biology. *Trends in Ecology & Evolution* **26**, 369-375 (2011).

2 Webb, B. What does robotics offer animal behaviour? *Animal Behaviour* **60**, 545-558 (2000).

3 Webb, B. Using robots to understand animal behavior. *Advances in the Study of Behavior* **38**, 1-58 (2008).

4 Taylor, R. C., Klein, B. A., Stein, J. & Ryan, M. J. Faux frogs: multimodal signalling and the value of robotics in animal behaviour. *Animal Behaviour* 76, 1089-1097 (2008).

5 Halfwerk, W. *et al*. Toward Testing for multimodal perception of mating Signals. *Frontiers in Ecology and Evolution* **7**, 124 (2019).

6 Narins, P. M., Grabul, D. S., Soma, K. K., Gaucher, P. & Hödl, W. Cross-modal integration in a dart-poison frog. *Proceedings of the National Academy of Sciences* **102**, 2425-2429 (2005).

7 Ręk, P. & Magrath, R. D. Multimodal duetting in magpie-larks: how do vocal and visual components contribute to a cooperative signal's function? *Animal behaviour* **117**, 35-42 (2016).

8 Derégnaucourt, S., Poirier, C., Van der Kant, A., Van der Linden, A. & Gahr, M. Comparisons of different methods to train a young zebra finch (*Taeniopygia guttata*) to learn a song. *Journal of Physiology-Paris* **107**, 210-218 (2013).

9 Benichov, J. I. *et al*. The forebrain song system mediates predictive call timing in female and male zebra finches. *Current Biology* **26**, 309-318 (2016).

10 Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* **291**, 2564-2569 (2001).

A System for Robot-Chick Vocal interactions

Michael Mcloughlin[1], Shuge Wang[2], Emmanouil Benetos[1], Dan Stowell[1], Elisabetta Versace[2]

[1]Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom

[2]Department of Biological and Experimental Psychology, School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Michael Mcloughlin
Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom
Email address: mclacademia@gmail.com

Abstract

Precision Livestock Farming (PLF) is a field that seeks to optimise the output product of an animal by using advanced technologies, such as robots, to monitor individual animals. This means that robots are becoming more common on farms. This presents a unique opportunity for researchers to detect vocalisations between animals and act upon them accordingly. Animal welfare is a field that seeks to bring about the highest possible quality of life for an animal. By using robots to monitor vocal interactions between animals, we can glean unique insight into animal lives, and thus their needs. To this end, we present a framework that describes how a robot can detect domestic chick (*Gallus gallus*) vocalisations, classify them, and then carry out different actions depending on the vocalisations. In the future we will deploy these robots into laboratory and commercial farming settings to verify their effectiveness in a production environment.

Introduction

While Vocal Interaction in-and-between Humans And Robots (VIHAR) is in its infancy, discussion regarding farm animals, their welfare, and their relationship to the field have been only mentioned in passing (Moore et al., 2016; Morovitz et al., 2017). There is a need to address this issue, as vocalisations in-and-between humans, animals, and robots are highly present in animal welfare (Manteuffel et al., 2004). Animal welfare is a field that seeks to bring about the highest possible quality of life for an animal. However, there is some disagreement on the best possible

way to do this, mainly due to the different perceptions of what constitutes good animal welfare across the globe. To date, the most commonly agreed on aspect of animal welfare is the 'Five Freedoms'. These ensure an animal's freedom from hunger and thirst; discomfort; pain, injury and disease; fear and distress; and freedom to express normal behaviour. Some researchers also argue that we must look beyond the five freedoms in order to address all animal needs (Mellor, 2016).

Precision Livestock Farming (PLF) is a field that seeks to optimise the output product of an animal by using advanced technologies to monitor individual animals (Banhazi et al., 2012). It is a multi-disciplinary field, with engineers, veterinarians, and animal scientists collaborating together. Robots are already being integrated into the field of PLF, with cow milking machines automatically monitoring the levels of milk being taken from a cow (Houstiou et al., 2017). Robots are already being investigated in regard to how they may interact with poultry, using methods such as beamforming to identify the position and acoustic activity of chicks (Gribovskiy et al., 2010; Gribovskiy & Mondada, 2008, 2010). While it is possible to simply install microphones into animal housing, robots are useful as they are free to move around the housing unrestricted. Furthermore, robots do not require any specialist installation. This is important, as the farmer may not be able to use their barn while waiting for the system to be installed. By using robots to monitor vocal interactions between animals, we can glean unique insight into animal lives. Given the risks associated with working on farms (Danuser et al., 2001), it is likely that we will see more and more robots in commercial farming to carry out tasks previously done by humans. To this end, we present a framework that allows a robot to detect domestic chick (*Gallus gallus*) vocalisations, classify them, and then carry out different actions depending on the vocalisations.

The framework

Domestic chickens have a repertoire and their vocalisations are associated with specific behaviours (Collias & Joos, 1953). Some vocalisations are associated with distress (Sufka et al., 2006), notifying other members of the flock about the presence of food (Evans & Evans, 1999; Evans & Marler, 1994), or alerting the other members of the flock to aerial and ground predators (Evans et al., 1993; Kokolakis et al., 2010). While this repertoire is generally discreet, it is worth noting that the fundamental frequency of the call will decrease as the animals grow larger (Fontana et al.,

2015). The acoustic characteristics of events known as 'peckouts', where one member of the flock is attacked and killed by all others, have also been identified (Bright, 2008).

We propose a system where a robot can detect and classify these calls and act accordingly. To achieve this system, we are building a database to describe the overall repertoire of chicks. As this project is still in its infancy, it is not possible to specify which type of feature extraction methods and classifiers are best suited to the task at hand. However, methods in ecology show promise for identifying what may work best (Mcloughlin et al., 2019), and methods are already being deployed in order to detect disease in poultry houses (Carpentiera et al., 2019). Depending on the output of the classifier, we can notify a robot on how to act according to the type of vocalisations it hears. The robot will be equipped with a microphone that can record chick sounds. After recording a sound, an onset detection function (Dixon, 2006) can be applied to detect the presence of a chick call in the recording. Following this, feature extraction can be applied before being passed on to the pre-trained classifier. Depending on the type of vocalisation that is produced, the robot can act accordingly. This process is summarized in Figure 1. For example, if it detected a peckout, it could move towards the animals and try and prevent the other members of the flock from killing the target of the peckout. It could also be used to identify the needs of the animals. For example, a lack of food calls in a recording may indicate that the animals may need food, and the robot can release food in response.

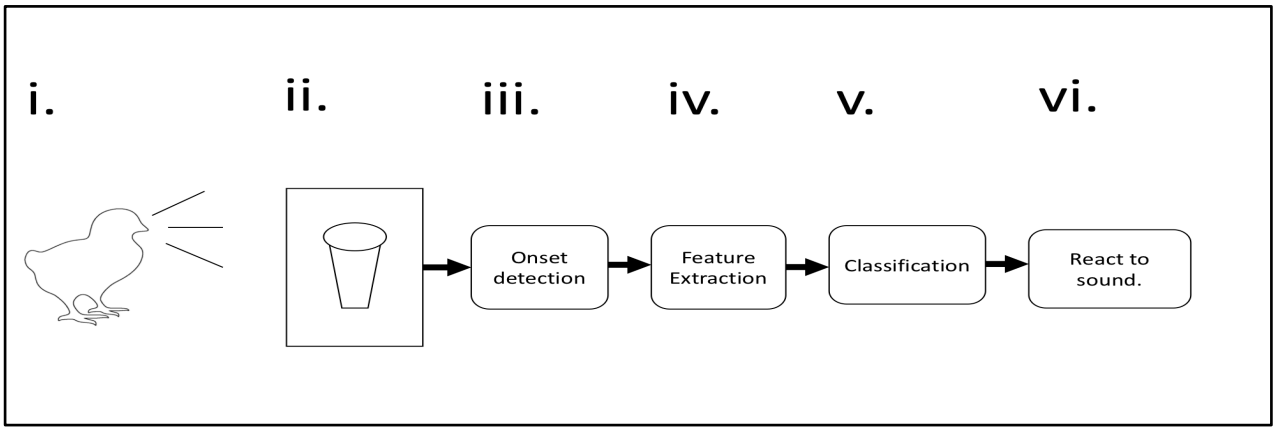

Figure 1: This shows the overall structure of the system. i) the chick produces a sound; ii) a robot equipped with a microphone records this sound; iii) Onset detection is used to identify where the sound occurred in the recording; iv) audio features are extracted from the recording; v) the

features are then classified by a pre-trained classifier; vi) the robot carries out a specific action depending on how the sound is classified.

Conclusion

We have presented a framework for the development of classifying chick vocalisations via robots, and how they can be used to facilitate chick-robot interaction. Future work will involve lab trials of this system, before deploying the robots to farm trials to investigate their effectiveness in the field.

References

Banhazi, T. M., Lehr, H., Black, J. L., Crabtree, H., Schofield, C. P., Tscharke, M., & Berckmans, D. (2012). Precision Livestock Farming: An international review of scientific and commercial aspects. *International Journal of Agricultural and Biological Engineering*, *5*, 1–9. https://doi.org/10.3965/j.ijabe.20120503.00

Bright, A. (2008). Vocalisations and acoustic parameters of flock noise from feather pecking and non-feather pecking laying flocks. *British Poultry Science*, *49*, 241–249. https://doi.org/10.1080/00071660802094172

Carpentiera, L., Vrankenab, E., Berckmans, D., Paeshuysea, J., & Norton, T. (2019). Development of sound-based poultry health monitoring tool for automated sneeze detection. *Computers and Electronics in Agriculture*, *162*, 573–581. https://doi.org/10.1016/j.compag.2019.05.013

Collias, N., & Joos, M. (1953). The spectrographic analysis of sound signals of the domestic fowl. *Behaviour*, *5*, 175–188. https://doi.org/10.1163/156853953X00104

Danuser, B., Weber, C., Künzli, N., Schindler, C., & Nowak, D. (2001). Respiratory symptoms in Swiss farmers: An epidemiological study of risk factors. *American Journal of Industrial Medicine*, *39*, 410–418. https://doi.org/10.1002/ajim.1032

Dixon, S. (2006). Onset detection revisited. In *Proceedings of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*.

Evans, C. S., & Evans, L. (1999). Chicken food calls are functionally referential. *Animal Behaviour*, *58*, 307–319. https://doi.org/10.1006/anbe.1999.1143

Evans, C. S., Evans, L., & Marler, P. (1993). On the meaning of alarm calls: functional reference

in an avian vocal system. *Animal Behaviour*, *46*, 23–38. https://doi.org/10.1006/anbe.1993.1158

Evans, C. S., & Marler, P. (1994). Food calling and audience effects in male chickens, Gallus gallus: Their relationships to food availability, courtship and social facilitation. *Animal Behaviour*, *47*, 1159–1170. https://doi.org/10.1006/anbe.1994.1154

Fontana, I., Tullo, E., Butterworth, A., & Guarino, M. (2015). An innovative approach to predict the growth in intensive poultry farming. *Computers and Electronics in Agriculture*, *119*, 178–183. https://doi.org/10.1016/j.compag.2015.10.001

Gribovskiy, A., Halloy, J., Deneubourg, J. L., Bleuler, H., & Mondada, F. (2010). Towards mixed societies of chickens and robots. In *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*. https://doi.org/10.1109/IROS.2010.5649542

Gribovskiy, A., & Mondada, F. (2008). Audio-visual detection of multiple chirping robots. In *Intelligent Autonomous Systems 10, IAS 2008*. https://doi.org/10.3233/978-1-58603-887-8-324

Gribovskiy, A., & Mondada, F. (2010). The PoulBot : a mobile robot for ethological studies on domestic chickens. *Proceedings of the International Symposium on AI-Inspired Biology*.

Houstiou, N., Fagon, J., Chauvat, S., Turlot, A., Klin-Eveillard, F., Boivin, X., & Allain, C. (2017). Impact of precision livestock farming on work and human- animal interactions on dairy farms. A review. *Biotechnol. Agron. Soc. Environ. BASE*, *21*, 268–275.

Kokolakis, A., Smith, C. L., & Evans, C. S. (2010). Aerial alarm calling by male fowl (Gallus gallus) reveals subtle new mechanisms of risk management. *Animal Behaviour*, *79*, 1373–1380. https://doi.org/10.1016/j.anbehav.2010.03.013

Manteuffel, G., Puppe, B., & Schön, P. C. (2004). Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science*, *88*, 163–182. https://doi.org/10.1016/j.applanim.2004.02.012

Mcloughlin, M. P., Stewart, R., & McElligott, A. G. (2019). Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of The Royal Society Interface*, *16*. https://doi.org/10.1098/rsif.2019.0225

Mellor, D. J. (2016). Updating Animal Welfare Thinking: Moving beyond the " Five Freedoms " towards " A Life Worth Living ". *Animals*, *6*, 21. https://doi.org/10.3390/ani6030021

Moore, R., Marxer, R., & Thill, S. (2016). Vocal Interactivity in-and-between Humans, Animals, and Robots. *Frontiers in Robotics and AI*. https://doi.org/https://doi.org/10.3389/frobt.2016.00061

Morovitz, M., Mueller, M., & Scheutz, M. (2017). Animal-Robot Interaction: The Role of Human Likeness on the Success of Dog-Robot Interactions. In *Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)* (pp. 22–26).

Sufka, K. J., Feltenstein, M. W., Warnick, J. E., Acevedo, E. O., Webb, H. E., & Cartwright, C. M. (2006). Modeling the anxiety-depression continuum hypothesis in domestic fowl chicks. *Behavioural Pharmacology*, *17*, 681–689. https://doi.org/10.1097/FBP.0b013e3280115fac

Matching human vocal imitations to birdsong: An exploratory analysis

Kendra Oudyk[1,2], Yun-Han Wu[3], Vincent Lostanlen[3,4], Justin Salamon[5], Andrew Farnsworth[4], and Juan Bello[3,6]

[1]**Department of Music, Art, and Culture Studies, University of Jyväskylä, Finland**
[2]**Integrated Program in Neuroscience, McGill University, Canada**
[3]**Music and Audio Research Lab, New York University, USA**
[4]**Cornell Lab of Ornithology, Cornell University, USA**
[5]**Adobe Research, San Francisco, USA**
[6]**Center for Urban Science and Progress, New York University, USA**

Corresponding author:
Kendra Oudyk[1]

Email address: kendra.oudyk@mail.mcgill.com

ABSTRACT

We explore computational strategies for matching human vocal imitations of birdsong to actual birdsong recordings. We recorded human vocal imitations of birdsong and subsequently analysed these data using three categories of audio features for matching imitations to original birdsong: spectral, temporal, and spectrotemporal. These exploratory analyses suggest that spectral features can help distinguish imitation strategies (e.g. whistling vs. singing) but are insufficient for distinguishing species. Similarly, whereas temporal features are correlated between human imitations and natural birdsong, they are also insufficient. Spectrotemporal features showed the greatest promise, in particular when used to extract a representation of the pitch contour of birdsong and human imitations. This finding suggests a link between the task of matching human imitations to birdsong to retrieval tasks in the music domain such as query-by-humming and cover song retrieval; we borrow from such existing methodologies to outline directions for future research.

INTRODUCTION

Humans often find bird sounds beautiful and interesting, and appear naturally inclined to imitate them. We can find bird imitations in various cultural contexts such as music and birdwatching. These imitations span the whole semiotic range from verbal description to verbatim copy, through mnemonics, onomatopoeia, whistling, and instrumental decoy (Taylor, 2017; Pieplow, 2017).

Having a machine match human and bird sounds is a multimodal problem for which there is no well-established computational framework. As of today, it is unclear whether this problem should be approached as speech recognition, as birdsong classification, or as melody extraction. Furthermore, variations within and between individual birds of a given species, as well as variations within and between humans in their imitation strategies, raise challenging research questions.

Machine listening research on human imitations of birdsong may play an important role in the emerging field of vocal interactivity in-and-between humans, animals, and robots (VIHAR). Indeed, this topic naturally involves all three agents. In particular, it investigates the ability of birds to produce songs which broadcast the acoustic signature of their species; the ability of humans to communicate the identity with their own voice; and the ability of robots (here, digital audio recording devices) to unify birdsong and human voice into a shared metric space of pairwise similarity. There is a growing body of machine listening research on vocal imitation in other areas, such as musical instruments (Kapur et al., 2004; Mehrabi et al., 2018), non-vocal sounds (Lemaitre et al., 2016a), basic auditory features (Lemaitre et al., 2016b), and audio concepts (Cartwright and Pardo, 2015). However, it appears that research on vocal imitations of non-human animal vocalizations is a novel area for VIHAR research.

The purpose of this paper is to explore the problem space of matching birdsong and imitations, in

order to guide the design of systems for classification and retrieval. To this end, we begin by describing our paradigm for collecting birdsong and human imitations. Then, we explore the data using various methods for matching human vocal imitations to birdsong, by assessing measures in the spectral, temporal, and spectrotemporal domains. We conclude by discussing potential approaches to this problem.

DATA COLLECTION

Imitations. Imitations were collected from a convenience sample of 17 participants (20-68 years; 4 female), including 10 with musical training and 11 with birding experience. Participants were seated alone in a sound-attenuated room. They were presented with a birdsong recording, and then immediately imitated what they heard. The sound of a clap marked the end of the birdsong excerpt and the beginning of the recording period, which lasted 2 seconds longer than the given birdsong stimulus. We used a MATLAB script to present stimuli and record imitations, using the internal speakers and microphone of a Dell Latitude E6420 laptop. Participants pressed a key to proceed to the next recording. Before data collection, there was a practice round with three birdsong recordings from outside the dataset. Participants were told that they could imitate in any manner they would choose.

Stimuli. In order to obtain birdsong for stimuli, field recordings of birdsong were scraped from Xeno-Canto.org, a citizen-science platform for sharing bird sounds (Vellinga and Planqué, 2015). The search was limited to a) the 'song' vocalization type (as opposed to, e.g., 'call'), b) a quality rating of A or B (on a scale from A to E, A being highest), and c) 10 specific species: black-capped chickadee (*Poecile atricapillus*), black-throated blue warbler (*Setophaga caerulescens*), common yellowthroat (*Geothlypis trichas*), mourning dove (*Zenaida macroura*), northern cardinal (*Cardinalis cardinalis*), prairie warbler (*Setophaga discolor*), red-eyed vireo (*Vireo olivaceus*), sora (*Porzana carolina*), veery (*Catharus fuscescens*), and white-throated sparrow (*Zonotrichia albicollis*). In order to obtain 'clean' birdsong excerpts that are suitable for imitation, we used Sonic Visualizer (Cannam et al., 2010) to manually annotate excerpts that a) had relatively high signal-to-noise ratio, b) contained song from the target species, and c) lasted approximately 2-10 seconds. From each of the 10 species, we randomly selected 10 recordings, and then selected the longest excerpt in each of those recordings to be used as stimuli for eliciting imitations, thus amounting to $10 \times 10 = 100$ stimuli per trial. Figure 1 shows a spectrogram that illustrates the data acquisition process for imitations.

This dataset [1] and the code [2] for this project are will be available will be available online.

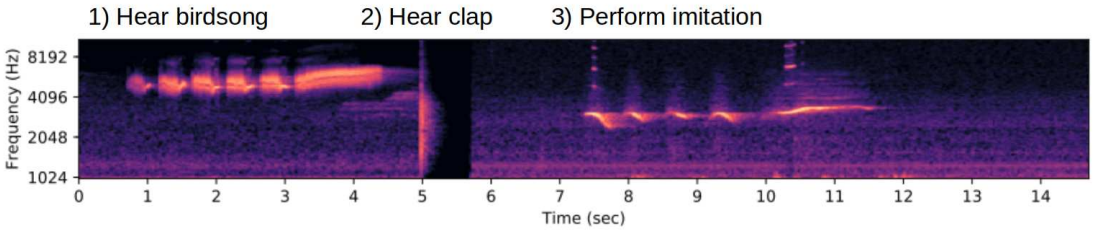

Figure 1. Spectrogram representation of one instance of data collection, comprising the playback of one stimulus the playback of a clap to alert the subject; and the live acquisition of the human imitation.

DATA EXPLORATION

Spectral analysis and results

If the goal in this problem space is to match human imitations to the imitated birdsong, an intermediate goal could be to match imitations to a species category. In previous research, Kapur et al. (2004) had success classifying human imitations of instruments (in beat boxing) using the feature space of the mel-frequency cepstral coefficients (MFCCs). In basic terms, MFCCs measure the overall shape of the acoustic energy spectrum over a frequency scale that is perceptually uniform. This feature is commonly used in speech recognition and music processing. The purpose of this section was to visually explore the separability of species in the MFCC space in order to see whether such features might be useful for species classification.

[1] https://birdvoximitation.weebly.com
[2] https://github.com/BirdVox/oudyk_vihar2019

For each imitation, we located the two spectrogram frames with the highest energy and calculated their 12 MFCCs of lowest quefrency. This resulted in a dataset of MFCC vectors which is exactly twice as large as the total number of imitations. In order to visualize how well species cluster in the space of the MFCCs, we used Principal Components Analysis (PCA) to reduce the dimensions from 12 to 2. PCA groups together dimensions (MFCCs here) in linear combinations that are maximally correlated, while minimizing the correlation between the groupings (i.e., principal components, PCs). PCA was performed in python with Scikit-learn (Pedregosa et al., 2011) using a full singular value decomposition with the standard LAPACK solver, with no rotation. The first two PCs respectively explained 30% and 24% of the variance in the full 12-MFCC space.

In the space of these two PCs, species appear to overlap with each other (see Figure 2A), so this feature does not look promising for species classification. The exception is the mourning dove (red dots), whose imitations are less distributed. This species may have elicited less-varied imitations because its song is slow, low-pitched, and memorable, and so may be easier to imitate (Pieplow, 2017).

We then explored what other information may be captured in this feature space. First, we visualized participants (see Figure 2B); while participants do not have striking separability, they appear to have greater separability than species. Next, in order to determine a simpler explanation for these two components, we performed *k*-means clustering on the imitations. This is a data-driven, non-deterministic method of grouping together data points based on their proximity to centroids ('means') in the given space (here, the reduced MFCC space). *K*-means was performed using the "elkan" variation (using the triangle inequality for efficiency) in Scikit-Learn with k=2 (i.e., 2 clusters), 10 runs with different centroids, a maximum of 200 iterations for a single run, and a tolerance of 0.0001 for inertia to declare convergence. The model took 2 iterations to converge, and the solution is visualized in Figure 2C. Manual inspection of a sample of data points within each cluster indicates that these clusters roughly correspond to *imitation strategy*: 86% of the sampled points in one cluster were whistled, and 83% in the other were not whistled.

Together, these results suggest that MFCCs are useful for identifying vocal strategy of birdsong imitations. They did not prove useful for classifying the imitated species, but there may be more information in a higher-dimensional representation of this space, with other settings for the analyses, or in other spectral features. These results are in line with previous research on vocal imitations of basic auditory features (Lemaitre et al., 2016b) and non-vocal sounds (Lemaitre et al., 2016a), showing that vocal imitation goes beyond simple mimicry, as features are adapted to human vocal abilities. While this spectral analysis does not appear to be useful for matching birdsong and imitations, clustering imitations by strategy may be useful if different matching methods prove more useful for different strategies.

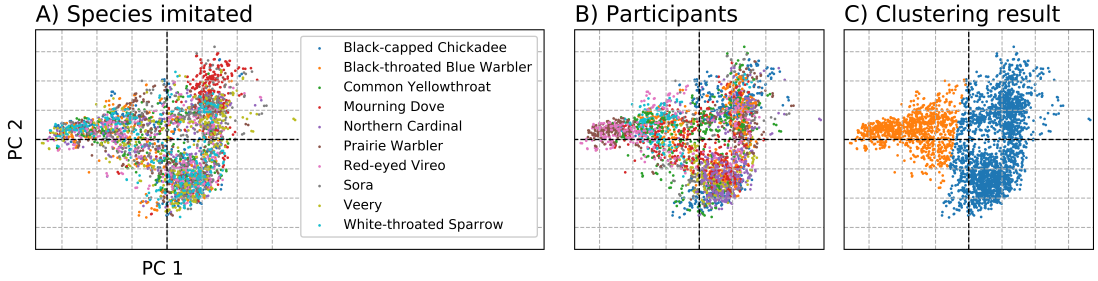

Figure 2. Results of the spectral analysis. The first two components from the PCA on the imitations' 12 MFCCs, overlaid with A) species, B) participants, and C) the result of *k*-means clustering.

Temporal analysis and results

In other areas of vocal imitation, humans are fairly accurate at reproducing the rhythmic or relative temporal structure of an audio sequence (Kapur et al., 2004). Therefore, we investigated whether a simple temporal feature — the number of sound events — could be useful for matching imitations and birdsong.

In order to count sound events, we used the following method, as illustrated in Figure 3A and B:

1. We used per-channel energy normalization (Wang et al., 2017; Lostanlen et al., 2018) as a pre-processing step to suppress background noise and emphasize foreground sounds, resulting in a spectrogram-like representation of the sound (see code for PCEN parameter specification).
2. We calculated an approximate signal-to-noise ratio (SNR) for each time point by subtracting the power of the minimum frequency bin from the maximum frequency bin, dividing by the median

frequency bin, then median-smoothing the SNR over 50 ms, giving a SNR curve ranging from 0-1.

3. We performed vocal activity detection with an initial peak threshold on the SNR of 0.45, and then followed the SNR curve in both directions to where it crossed the activity threshold of 0.2. These two crossings were taken as the onset and offset time for each detected sound activity.
4. We counted the number of sound events as the number of segment onsets.

We then visualized the relationship between the number of sound events in the stimuli and their imitations; as can be seen in Figure 3C, they roughly correspond. However, there was a tendency for imitations to overshoot low stimulus counts and undershoot high stimulus counts. Further, there are more outliers above zero than below zero, suggesting that participants more often drastically overshot than undershot the true number of events in the stimulus.

The correspondence between the number of events in stimuli and their imitations indicates that the number of sound events may be useful for matching imitations to the exact instance of birdsong being imitated. These results also suggest that our vocal activity detection technique is performing above chance, since there is high variance within modalities (bird vs. human), but still a positive correlation across modalities. In the future, this technique could be assessed more effectively with manually-segmented audio as the ground truth, and then more-confident conclusions could be drawn from the analysis. The parameters used performed well based on visual inspection, but may be optimized in the future as well.

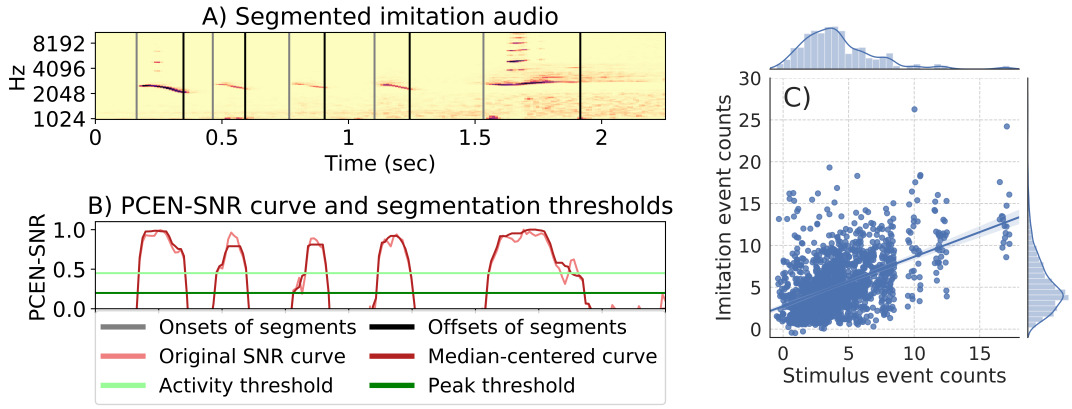

Figure 3. Illustration of temporal analysis. A) and B) illustrate the method of segmentation based on the signal-to-noise ratio in a per-channel-energy-normalized mel-spectrogram (PCEN-SNR). C) shows the relationship between stimulus and imitation event counts. The line and shaded area respectively denote linear regression and their 95% confidence intervals. Counts are jittered up to 0.5 for visibility.

Spectrotemporal analysis and results

We then addressed the problem using spectrotemporal information in the form of pitch contour classes. Contour classification has been used in musical analyses (Adams, 1976) and in music information retrieval (Bittner et al., 2017, 2015; Kako et al., 2009; Salamon and Gómez, 2012; Panteli et al., 2017; Salamon et al., 2013). Here, we borrow aspects of several methods, estimating the pitch contour using a polynomial fitted to pitch time series (Bittner et al., 2017), classifying the pitch contour by quantizing the space defined by polynomial features (Adams, 1976; Salamon et al., 2012), and then comparing the contours of stimuli and imitations using the Levenshtein distance (e.g., Lemström and Ukkonen, 2000).

As noted in the section on the spectral analysis, participants used various imitation strategies. Some strategies do not have a discernible pitch (e.g., imitations consisting of noisy or percussive vocalizations). Thus, for this analysis, we decided to restrict the study to four bird species (mourning dove, sora, white-throated sparrow, and northern cardinal) and 6 participants that produced the most whistling performances. This brought the number of imitations down to 240.

In order to extract a pitch contour from each active segment, we applied a fundamental frequency estimation algorithm. This algorithm consists in locating, for every frame in a per-channel energy normalized (PCEN) spectrogram, the mel-frequency bin of highest magnitude. Based on preliminary analyses, this simple frequency-domain procedure appeared more robust to octave errors than well-established time-domain algorithms, such as YIN (De Cheveigné and Kawahara, 2002). Then, we fit a second-degree polynomial of the form $f = \alpha t^2 + \beta t + \gamma$, as measured on a mel-frequency scale. Although

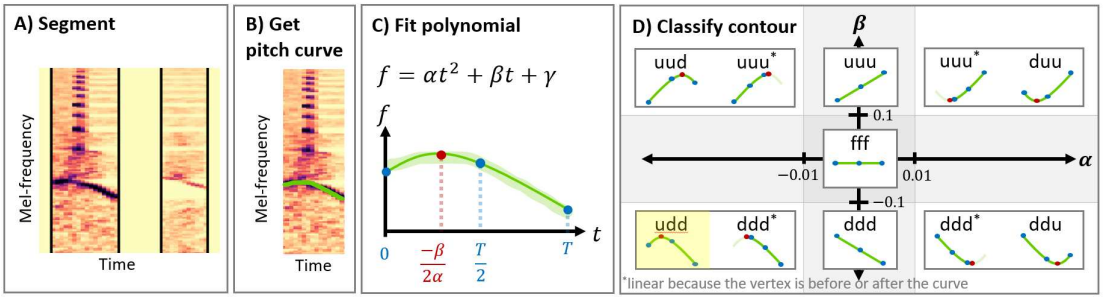

Figure 4. The steps involved in classifying pitch contours.

the intercept γ varies monotonously with frequency transposition, the quadratic term α and the linear term β are transposition-invariant. We can quantize the feature space of these terms into seven regions by thresholding α and β (see Figure 4D). However, since the vertex of the curve may be located before or after the actual curve, we also considered the location of the polynomial vertex relative to the beginning, middle and the end of the recording (see Table 1 and Figure 4C and D). Each curve is labelled with a 3-letter string, where *u* stands for up, *d* for down, and *f* for flat (e.g., *udd* is *up-down-down*).

We compared the pitch contours for the imitations and birdsong using the Levenshtein distance between a stimulus and *a*) its corresponding 6 imitations, *b*) 6 imitations of a similar song from the same species, and *c*) 6 randomly-chosen imitations from other species. If this measure is useful for matching birdsong and its imitations, then the Levenshtein distances between these pairs should be $a <= b < c$. In other words, the stimulus should be most similar to its imitations, then equally or less similar to imitations of birdsong from the same species, and least similar to a different species.

Contour class	Quadratic term α	Linear term β	Time location of vertex $v = \frac{-\beta}{2*\alpha}$
uuu	$-0.01 < \alpha < 0.01$	$0.1 < \beta$	(no vertex)
	$0.01 < \alpha$	$0.1 < \beta$	$v < 0$
	$\alpha < -0.01$	$0.1 < \beta$	$T < v$
duu	$0.01 < \alpha$	$0.1 < \beta$	$0 < v < \frac{T}{2}$
ddu	$0.01 < \alpha$	$\beta < -0.1$	$\frac{T}{2} < v < T$
ddd	$-0.01 < \alpha < 0.01$	$\beta < -0.1$	(no vertex)
	$0.01 < \alpha$	$\beta < -0.1$	$T < v$
	$\alpha < -0.01$	$\beta < -0.1$	$v < 0$
uud	$\alpha < -0.01$	$0.1 < \beta$	$\frac{T}{2} < v < T$
udd	$\alpha < -0.01$	$\beta < -0.1$	$0 < v < \frac{T}{2}$
fff	$-0.01 < \alpha < 0.01$	$-0.1 < \beta < 0.1$	(no vertex)

Table 1. Definitions of pitch contour classes.

Results showed that $a <= b < c$ was true for 71% of the selected data. This proportion rises to 79% if we only required the distance between stimulus and imitations of the same species to be smaller than those between it and imitation of a different species (i.e., $a < c$ and $b < c$). This indicates the usefulness of pitch contours for matching birdsong and imitations. Future analyses target the participants and species that did not primarily produce whistled sounds, as this analysis may be more effective for imitations that are predominantly tonal.

CONCLUSION AND FUTURE DIRECTIONS

The purpose of this study was to explore spectral, temporal, and spectrotemporal methods for matching birdsong and human imitations. The spectral space of the MFCCs was not sufficient to move beyond classifying imitation strategy. The temporal analysis revealed that the number of events roughly corresponds between imitations and original birdsong. However, the most promising results were found with the subsequent spectrotemporal analysis, in which we used the melody contour to match imitations to birdsong. Together, these results suggest that the problem of retrieval-by-imitation for birdsong is more akin to a melody recognition problem than a speech recognition problem. This suggests that this problem may be addressed using established methods in music information retrieval for query-by-imitation or imitation classification, and future work will follow these research directions.

ACKNOWLEDGEMENTS

This project was supported by the Leon Levy Foundation, the National Science Foundation's Big Data grant 1633206, and a travel grant from the University of Jyväskylä (KO).

REFERENCES

Adams, C. R. (1976). Melodic contour typology. *Ethnomusicology*, pages 179–215.

Bittner, R. M., Salamon, J., Bosch, J. J., and Bello, J. P. (2017). Pitch contours as a mid-level representation for music informatics. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society.

Bittner, R. M., Salamon, J., Essid, S., and Bello, J. P. (2015). Melody extraction by contour classification. In *ISMIR*, pages 500–506.

Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy.

Cartwright, M. and Pardo, B. (2015). VocalSketch: Vocally Imitating Audio Concepts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 43–46, Seoul, Republic of Korea. ACM Press.

De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Kako, T., Ohishi, Y., Kameoka, H., Kashino, K., and Takeda, K. (2009). Automatic identification for singing style based on sung melodic contour characterized in phase plane. In *ISMIR*, pages 393–398.

Kapur, A., Benning, M., and Tzanetakis, G. (2004). Query-by-beat-boxing: Music retrieval for the dj. In *Proceedings of the International Conference on Music Information Retrieval*, pages 170–177.

Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., and Susini, P. (2016a). Vocal imitations of non-vocal sounds. *PloS one*, 11(12):e0168167.

Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., and Susini, P. (2016b). Vocal imitations of basic auditory features. *The Journal of the Acoustical Society of America*, 139(1):290–300.

Lemström, K. and Ukkonen, E. (2000). Including interval encoding into edit distance based music comparison and retrieval. In *Proc. AISB*, pages 53–60.

Lostanlen, V., Salamon, J., Cartwright, M., McFee, B., Farnsworth, A., Kelling, S., and Bello, J. P. (2018). Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1):39–43.

Mehrabi, A., Choi, K., Dixon, S., and Sandler, M. (2018). Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 356–360. IEEE.

Panteli, M., Bittner, R., Bello, J. P., and Dixon, S. (2017). Towards the characterization of singing styles in world music. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 636–640. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pieplow, N. (2017). *Peterson Field Guide to Bird Sounds of Eastern North America*. Houghton Mifflin Harcourt Publishing Company, New York, NY.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Salamon, J., Rocha, B., and Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–84. IEEE.

Salamon, J., Serra, J., and Gómez, E. (2013). Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58.

Taylor, H. (2017). *Is Birdsong Music?: Outback Encounters with an Australian Songbird*. Indiana University Press.

Vellinga, W.-P. and Planqué, R. (2015). The xeno-canto collection and its relation to sound recognition and classification. In *CLEF (Working Notes)*.

Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A. (2017). Trainable frontend for robust and far-field keyword spotting. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5670–5674. IEEE.

Challenges for Integrating Multimodal Information into an Open-Source Human-Robot [Speech] Interaction System

Michael Connolly Brady

American University of Central Asia

Email address: brady_m@auca.kg

ABSTRACT

This report presents an open-source human-robot interaction system under development. Anyone with access to a 3D printer and an internet shopping cart should be able to build this robot for well less than the cost of a laptop computer. The system is meant to bridge the divide between closed-source children's toys types of robots, and the high-end expensive humanoid robots found in elite research laboratories. The focus of this report is on the development of the robot's software control system, especially in the area of non-linguistic and multimodal interactional cues. Such a control system goes beyond linguistic information transfer as the basis of communication. Visual cues, non-linguistic vocal gestures, and motor control are important aspects of interaction that need to be incorporated. From this explicit foundation, we consider some challenges faced and decisions made in the practical design of such a system.

INTRODUCTION

In the late 1990s and early 2000s, VoiceXML characterized the cutting edge of interactive speech technology (Abbot, 2002). A user could make a telephone call to a menu-driven speech system to e.g. find out the weather in a specific city, or to confirm airplane reservations. The machine would ask the questions, and speech recognition was based on the best match of the caller's reply with a menu of valid responses. Much has changed since those days in terms of artificial speech technologies. Today a person may ask questions to 'voice assistants,' such as Siri or Alexa, in a more open-ended manner to retrieve information (Hoy, 2018). Yet much has also remained the same. Linguistic information transfer is still generally the de facto goal of human-machine speech interaction. A more natural and efficient communication interface will involve the integration of visual information, affective and emotional cues, and forms of shared attention between the user and the machine. That is, linguistic information transfer based on audio is but a mere portion of the challenge for the human-machine speech interface.

One reason that interactive speech technologies today still lag in multimodal integration may relate to specialization pressures on engineers. Speech engineers tend to leave computer vision to the computer vision community, and to leave motor control to the robotics engineers. Likewise, computer vision and robotics engineers tend to treat speech as a black-box problem best left to speech engineers and computational linguists. As a result, multimodal integration is generally left as an issue for the interaction designer to solve. Furthermore, one reason that speech technologies are mostly still confined to linguistic information transmission may be that this is the low-hanging fruit. It is attractive and comfortable and routine to treat human communication as an elaboration of text processing. Yet, there is a demand to go beyond the comfortable. For instance, an exciting frontier in developing intervention therapies for children on the autism spectrum involves the use of humanoid robots (A.P. Costa, 2018). In short, the humanoid robot intermediary for autism therapy should be able to visually direct attention, should be able to perform and recognize communicative gestures, and should be able to convey [non-linguistic] social cues. In light of all this, we propose that: 1) a human-robot communication control system's design should not assume linguistic information transfer as The primary (and somewhat solved) problem, 2) the robot's control design should induce collaborations between researchers in vocal interaction, vision, and motor control, among others, and 3) such a robot's software and hardware need to be open, easy, serious, and affordable if the platform is to attract a community of developers to contribute to it.

MATERIALS AND METHODS

This section introduces the open-source robot hardware system. Then the design approach to the software is introduced. The software design entails some practical goals and constraints. One software design goal is that once someone has assembled the robot, they should be able to quickly test the robot and program it to do something interesting, without a steep learning curve. A second goal is that the control system's design should be flexible enough to address unforeseeable demands of users. The design should also allow for developers to share their work with each other. In an effort to address these issues, we provide something of a "browser - operating system" or *interpreter* that renders what we call "FluidScript" files. A FluidScript file is an XML file that elaborates on the VoiceXML protocol. The Fluidscript design also borrows from other protocols, such as Speech Synthesis Markup Language (https://www.w3.org/tr/speech-synthesis11), Behavior Markup Language (H. Vilhjálmsson, 2012), and Perception Markup Language (S. Scherer, 2012). The interpreter is developed for the Linux operating system (i.e. for Raspberry Pi), but Windows and Apple versions could be made available on demand.

Hardware System

The open-source robot hardware is composed of 3D printed parts, hobby servo motors, two USB-based cameras, standard audio based microphone(s) and loudspeaker, and a USB-based motor control system (Arduino). Anyone with a 3D printer and access to internet shopping should be able to build this robot for less than a few hundred dollars. The robot head's size is about that of an adult person's head size. The two cameras are fitted as the robot's eyes, actuated by a pan-tilt mechanism. A user may readily see where the cameras are pointed – or where 'the robot is looking.' The robot's eyecams each have lids and brows for conveying visual affective cues, i.e. valence (brows) and arousal (lids). The loudspeaker for speech and vocal sound production is placed at the back of a short tube, and two mechanical 'lips' are placed at the front of the tube. As the lips move, the radiation characteristic of the resulting sound prompts the listener to perceptually localize the sound to the robot's vocal cavity. In total, the robot head has 12 separate degrees of freedom, including: (1) right eyecam pan, (2) left eyecam pan, (3) eyecams tilt, (4) right eyecam lids, (5) left eyecam lids, (6) right eyecam brows, (7) left eyecam brows, (8) lips open-close, (9) head rotate, (10) head tilt, (11) neck rotate, (12) neck tilt. The hardware is designed so that motors and components are easily accessible, and so that parts may easily be replaced or customized, depending on potential user needs. Figure 1 presents the robot head in four poses to illustrate the various degrees of freedom. Development of the robot also includes an optional set of arms, though we focus on the robot's head here. Because the robot's electronics use computer standards (USB webcams, standard computer audio in-out, serial port for motor control), software for the robot can be designed from scratch by any skilled developer or team. For example, some users may prefer to use ROS (https://www.ros.org). Here, "form follows function." The priority is to provide a platform that is affordable and easy to maintain for the purpose of interaction research and software development. Though the robot may not look as realistic as some other humanoid robots, a visual appearance that is indistinguishable from that of a real person does not provide enough function to warrant the cost, in our humble opinion. See (http://www.fluidbase.com) for videos and more information on this hardware system.

Figure 1. open source robot hardware system

Building Behavior with 'Fluidscript'

Next we consider the general design of a FluidScript file. Like with VoiceXML, the interpreter starts with an initial script that contains global variables and boots the system. When the interpreter is started, this 'boot.fxml' Fluidscript file is read to initialize the robot. A short example boot script might look something like this:

```
01 <?xml version = "1.0"?>
02 <fluidscript version = "1.0", application = "boot.fxml"/>
03 <meta name = "author", content = "Michael Brady"/>
04 <meta name = "voice", content = "steve"/>
05 <form id = "boot">
06   <audio src="hello.wav"/>
07   <motors moveall = "rest"/>
08   <pause> 2.0 </pause>
09   <motors gesture = "cyclemotors.mpx"/>
10   <pause> 10.0 </pause>
11   <field name = "cond", type = "boolean">
12     <prompt video = "head.vgx", audio = "true", interrupt = "false">
13          Do my motors look to be working okay? <brows = "up"/>
14     </prompt>
15     <filled>
16       <if cond = "true">
17         <goto next = "greet.fxml"/>
18       </if>
19       <speak> Oh no. I will power off. </speak>
20       <goto next = "shutdown.fxml"/>
21     </filled>
22   </field>
23 </form>
24 </fluidscript>
```

Let us walk through this script. Lines 1 and 2 establish that this is an XML-based Fluidscript file named 'boot.fxml.' Some example meta-information is then provided, such as the voice the robot will speak with. Many types of meta-tags may be used here to specify global parameters and to initialize the robot with some background behaviors when not processing scripts. One Line 5, the script then enters into a 'form.' Forms are the building blocks of behavior for the robot. On Line 6, the robot says: "hello" by playing a .wav file. The <motors/> tag is used for moving motors. On Line 7, the script specifies that all motors should move to the 'rest' or start position (position where the robot should relax to when there is no power to the motors). Line 8 makes the script pause for two seconds while the motors move. The interpreter now has the motors supposed positions. On Line 9, the <motors/> tag is then used to play an '.mpx' file. This 'cycle.mpx' file contains the commands for a predefined sequence of motor moves – to test if the robot's motors are powered and calibrated properly. The .mpx files are developed separately and a library of .mpx files is provided with the interpreter, (new .mpx files may be created by advanced users). From there, the robot prompts the user with the question: "Do my motors look to be working okay?" When the <prompt/> tag is used, the field waits to be filled with a response. Here the robot is looking for a boolean (yes-no or true-false) response. As specified in this specific <prompt/> tag example, the response may come from both the visual system (using the program 'head.vgx' - computer vision that works to detect an affirmative head nod and-or thumbs-up hand gesture) and the auditory system (a spoken "yes" or "no," "affirmative" or "negative," from the default speech recognizer). If the response is 'yes,' the robot leaves the boot.fxml script and proceeds to run the specified next script, called 'greet.fxml.' If the response is not 'yes,' the robot speaks an error message and powers off by running 'shutdown.fxml.' The user can then try to correct the motor problem and re-boot the system. Here we can begin to see some dilemmas. What if the vision system detects a "thumb's up" affirmative hand gesture, but the speech system detects that the user said "no!" Perhaps in this case the solution is rather simple. However, how might we integrate a local lip-reading vision program with a cloud-based speech recognition program?

Let's take a look at a second Fluidscript file. Once the robot has booted, the interpreter may run local Fluidscripts, or (eventually) it may run scripts that are hosted remotely on the World Wide Web. A local script might be e.g. a home security monitoring system, a sassy chess opponent, or a script written by

an autism therapist. A remote script may be an airplane reservation system or a telepresence script that allows a remote user to become the 'Wizard-of-Ozz' controller of the robot. For now consider the local script of a greet.fxml, involving computer vision for face recognition:

```
01 <?xml version = "1.0"?>
02 <fluidscript version = "1.0", application = "greet.fxml"/>
03 <meta name = "affect", content = "sleepy"/>
04 <form id = "detect">
05   <field name = "face", type = "boolean"/>
06     <prompt video = "detectFace.vgx", interrupt = "false"/>
07     <filled>
08       <if face = "true">
09         <speak><excited> Oh, hi there <blink/></excited></speak>
10         <meta name = "affect", content = "wakeful"/>
11         <goto next = "#recognize"/>
12       </if>
13       <speak><sad>
14          I don't see anyone. I will keep looking.
15       </sad></speak>
16       <goto next = "#detect"/>
17     </filled>
18   </field>
19 </form>
20 <form id = "recognize">
21   <field name = "who", type = "modal"/>
22    <prompt video = "recognizeFace.vgx", interrupt = "false"/>
23     <filled>
24       <if who = "true">
25         <speak><excited>
26             It is good to see you <blink/> $who.name
27         </excited></speak>
28         <goto next = "#address($who.name)"/>
29       </if>
30       <speak><brows = "down"/> I don't recognize you. </speak>
31       <goto next = "shutdown.fxml"/>
32     </filled>
33   </field>
34 </form>
35 </fluidscript>
```

This script starts off as before. Then on Line 3, the affect parameter of the robot is set to "sleepy." The robot will remain "sleepy" until this parameter is changed again. The script then enters a form with an id of "detect," where a field is initialized to look for a face. If the computer vision program "detectFace.vgx" detects a face, the robot says "oh, hi there," the global affect is set to "awake," and control is passed to the form "recognize," (a script may contain multiple forms). If a face is not detected, the robot says: "I don't see anyone, let me keep looking," and re-runs the "detect" form. Perhaps there is a hardware problem that needs to be solved. The "recognize" form is similar to other forms we have now seen. It has a field called "who" that needs to be filled. If the face is recognized, the vision program "recognizeFace.vgx" is run to fill the field with a person's name. Control then passes to a form called "address" (not included here), that receives the name of the person recognized, where the robot may address the person with a greeting based on stored memories of that person. For the sake of simplicity in this example, if the face is not recognized, the robot is shut down. A better solution might be to ask the unrecognized person for their name, do some error checking, and if it is a new person, ask for permission to add that person to the database.

Challenges for Development

Hopefully at this point the reader is able to imagine building some basic behaviors and robot interactions using Fluidscript. Now let's consider some issues. One issue is that the Fluidscript interpreter is built to use external programs. Users write (and share) Fluidscript files, stand-alone computer vision programs, stand-alone audio processing programs, and motor output sequences. A long discussion might go here

about rationale and strategy. In short, the interpreter is mainly responsible for motor control. The interpreter owns the serial port that controls the robot's motors. However, the interpreter does not 'own' the USB cameras used by vision programs, nor does it 'own' the audio channels used by audio programs. That is, the interpreter is merely a traffic controller that manages audio and video channel usage. This allows developers to build and test audio and video programs independent of the interpreter.

With such a VoiceXML-derived control system, two difficulties with developing robot behaviors present themselves: 1) behavioral interrupts, and 2) sensor sharing and integration. 'Behavioral interrupts' refers to when one script is running and another script needs to jump in and override the first script. An example comes to mind from a preliminary study with the robot at a school for autistic children. When a child quickly thrusts his hand in front of the robot's eyecams, the robot might best react with an avoidance maneuver and vocal sound. Or, when a child tickles the robot's chin, a therapist might want the robot to stop and e.g. giggle. Better yet, the robot might giggle while continuing with its previous behavior. How to specify behavioral interrupts is a challenge for the rigid turn-taking nature of the Fluidscript protocol. With the <prompt/> tag, the use of behavioral interrupts is anticipated (e.g. Line 22 of greet.fxml), but how to specify interrupted and merged behaviors using Fluidscript is still in debate. 'Sensor sharing and integration' relates to how to run one stand-alone program that needs to use the video cameras and-or audio channels while another stand-alone program is using them. For instance, say a developer builds a face expression recognition computer vision program where results should change the character of the vocal output of an audio program (that has already launched). How might the vision program pass its information to the audio program? Writing a single program that uses both audio and video is feasible, but would e.g. lock out other programs from using the video camera(s) during speech production.

An important question for the vocal interaction community is: how might the Speech Synthesis Markup Language (SSML) be modified to better address the multimodal demands of human-robot communication? For instance, in marking up text for output from a speech synthesizer, a user may want to control eyecam lids, eyecam brows, lips, and even hand gestures in synchrony with speech. Should this motor control be based on behavior-generating parameters, or is it better to specify all of the details of behavior through specific tags (or some combination of the two)? Another challenge is in how to "mark up" speech when there is no linguistic message to be marked up? Grunts and laughs and snorts don't fare so well on most speech synthesizers. How might affect such as sarcasm, or dominance and submission be encoded in a notational system? We seek feedback from the vocal interaction community on this.

DISCUSSION

This report has introduced a specific open-source platform for human-robot interaction under development. For practical reasons and to provide a system that is accessible to non-expert users, the control system is broken into two parts. Non-expert users may write Fluidscript files, as exemplified in this report, to quickly develop robot behaviors. A library of computer vision and speech processing applications is made available for this. Advanced users may then write custom stand-alone audio and computer vision applications to run under Fluidscript flow-control. We are optimistic that a large number of such stand-alone open-source programs may be contributed, and that the Fluidscript interpreter may eventually grow into a full-fledged operating system. In the meantime, it is hoped that the vocal interaction community (and other communities) may offer useful feedback in developing a robust Fluidscript or communication protocol that will overcome the current rigid turn-taking paradigm. The robot is still in its beta testing phase, and is not ready for the public. But it is generally available to researchers and developers. Please contact the author if you are interested in contributing or in working with the robot.

REFERENCES

Abbot, K. R. (2002). *Voice Enabling Web Applications: VoiceXML and Beyond*. APress Media.

A.P. Costa, e. a. (2018). More attention and less repetitive and stereotyped behaviors using a robot with children with autism. *IEEE Intl Conf on Robot and Human Interactive Communication*.

H. Vilhjálmsson, e. a. (2012). The behavior markup language: Recent developments and challenges. *International Workshop on Intelligent Virtual Agents*.

Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants.

S. Scherer, S. Marsella, e. a. (2012). Perception markup language: towards a standard representation of perceived non-verbal behaviors. *12 International Conference on Intelligent Virtual Agents*.

Vocal emotion recognition in school-age children: normative data for the EmoHI test

Leanne Nagels[1,2], Etienne Gaudrain[3,2], Debi Vickers[4], Marta Matos Lopes[5,6], Petra Hendriks[1], and Deniz Başkent[2]

[1]**Center for Language and Cognition Groningen (CLCG), University of Groningen, Groningen, The Netherlands**
[2]**Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands**
[3]**CNRS, Lyon Neuroscience Research Center, Université de Lyon, Lyon, France**
[4]**Clinical Neurosciences Department, University of Cambridge, Cambridge, UK**
[5]**Hearbase Ltd, Hearing specialists, Kent, UK**
[6]**The Ear Institute, University College London, London, UK**

Corresponding author:
Leanne Nagels[1,2]

Email address: leanne.nagels@rug.nl

ABSTRACT

Traditionally, emotion recognition research has primarily used pictures and videos while audio test materials have received less attention and are not always readily available. Particularly for testing vocal emotion recognition in hearing-impaired listeners, the audio quality of assessment materials may be crucial. Here, we present a vocal emotion recognition test with non-language specific pseudospeech productions (based on Bänziger & Scherer, 2010) of multiple speakers expressing three core emotions (happy, angry, and sad): the EmoHI test. Recorded with high sound quality, the test is suitable to use with populations of children and adults with normal or impaired hearing, and across different languages. In the present study, we obtained normative data for vocal emotion recognition development in normal-hearing school-age (4-12 years) children using the EmoHI test. In addition, we tested Dutch and English children to investigate cross-language effects. Our results show that children's emotion recognition accuracy scores improved significantly with age from the youngest group tested on (mean accuracy 4-6 years: 48.9%), but children's performance did not reach adult-like values (mean accuracy adults: 94.1%) even for the oldest age group tested (mean accuracy 10-12 years: 81.1%). Furthermore, the effect of age on children's development did not differ across languages. The strong but slow development in children's ability to recognize vocal emotions emphasizes the role of auditory experience in forming robust representations of vocal emotions. The wide range of age-related performances that are captured and the lack of significant differences across the tested languages affirm the usability and versatility of the EmoHI test.

INTRODUCTION

Children's development of emotion recognition has been studied extensively using visual stimuli, such as pictures or sketches of facial expressions, or audiovisual materials (e.g., Nowicki and Duke, 1994), and particularly with clinical groups, such as autistic children (e.g., Harms et al., 2010). However, not much is known about the development of vocal emotion recognition (Scherer, 1986). Children have been reported to reliably recognize vocal emotions already from the age of 5 years on, but this ability continues to develop to adult-like levels throughout childhood (Tonks et al., 2007; Sauter et al., 2013). Based on earlier research on the development of voice perception (Mann et al., 1979; Nittrouer et al., 1993), children's performance may be lower compared to adults due to differences in their weighting of acoustic cues and a lack of robust representations of auditory categories. For instance, Morton and Trehub (2001) showed that, when acoustic cues and linguistic content contradict the emotion they convey, children mostly rely on linguistic content to judge emotions, whereas adults mostly rely on affective prosody. In addition, children and adults both perform better in facial than vocal emotion recognition tasks (Nowicki and Duke,

1994). All of these observations combined indicate that the formation of robust representations for vocal emotions is highly complex and possibly a long-lasting process even in typically developing children.

Research with hearing-impaired children has shown that they do not perform as well on vocal emotion recognition compared to their normal-hearing peers (Dyck et al., 2004; Hopyan-Misakyan et al., 2009; Nakata et al., 2012; Chatterjee et al., 2015). Hopyan-Misakyan et al. (2009) showed that children with cochlear implants (CIs) performed as well as their normal-hearing peers on facial emotion recognition but scored significantly lower on vocal emotion recognition. Facial emotion recognition seems to generally develop faster than vocal emotion recognition (Nowicki and Duke, 1994), particularly in hearing-impaired children (Hopyan-Misakyan et al., 2009), which may indicate that visual emotion cues are perceptually more prominent or easier to categorize than vocal emotion cues. A higher reliance on visual emotion cues as compensation for degraded auditory input, as emotion recognition in daily life is usually multimodal for which visual emotion cues can often be sufficient, may lead to less robust representations of vocal emotions. Furthermore, Nakata et al. (2012) found that children with CIs had difficulties primarily with differentiating happy from angry vocal emotions. This difference may be related to a higher reliance on differences in speaking rate to categorize vocal emotions, as this cue differentiates sad from happy and angry vocal emotions but is similar for the latter two emotions. Therefore, hearing loss also seems to influence the weighting of different acoustic cues, and hence likely also affects the formation of representations of vocal emotions.

As most research on the development of emotion recognition has used visual materials such as pictures or videos, good-quality audio materials are scarce. For normal-hearing listeners, the audio quality may only have a small effect on performance, but for testing hearing-impaired populations it may be highly important. Hence, we recorded high sound quality vocal emotion recognition test stimuli produced by multiple speakers with three basic emotions (happy, angry, and sad) that are suitable to use with hearing-impaired children and adults: the EmoHI test. We aimed to investigate how school-age children's ability to recognize vocal emotions develops with age and to obtain normative data for the EmoHI test for future applications, for instance, with clinical populations. In addition, we tested children of two different native languages, namely Dutch and English, to investigate potential cross-language effects.

METHODS

Participants

Fifty-eight Dutch children and 25 English children between the ages of 4 to 12 years, and 15 Dutch adults and 15 English adults participated in the study. All participants were monolingual speakers of Dutch or English and reported no hearing or language disorders. Normal hearing (hearing thresholds at 20 dB HL) was screened with pure-tone audiometry at octave-frequencies between 500 and 4000 Hz. The study was approved by local ethics committees of the participating institutions. A written informed consent form was signed by the parents of children and adult participants before data collection.

Stimuli and Apparatus

We made recordings of six native Dutch speakers producing two non-language specific pseudospeech sentences using three core emotions (happy, sad, and angry), and a neutral emotion (not used in the current study). All speakers were native monolingual speakers of Dutch without any discernable accent and did not have any speech, language, or hearing disorders. Speakers gave written informed consent for the distribution and sharing of the recorded materials. To keep our stimuli relevant to emotion perception literature, the pseudospeech sentences that we used, *Koun se mina lod belam* [kʌun sə minaː lɔt beːlɑm] and *Nekal ibam soud molen* [neːkɑl ibɑm sʌut moːlɛn], were taken from the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus by Bänziger and Scherer (2010). Speakers were instructed to produce the sentences in a happy, sad, angry, or neutral manner using emotional scripts that were also used for the GEMEP corpus stimuli (Scherer and Bänziger, 2010). The stimuli were recorded in an anechoic room at a sampling rate of 44.1 kHz. We selected the productions which received the highest accuracy scores of the four highest-rated speakers based on an online survey with Dutch and English adults. Table 1 shows an overview of these four selected speakers' demographic information and voice characteristics. The neutral productions and the productions of the other two speakers were part of the online survey, and are available with the stimulus set, but were not used in the current study to simplify the task for children. Our final set of stimuli consisted of 36 experimental stimuli with three items (combinations of two times

one sentence and one time the other sentence) per emotion and per speaker (3 items x 3 emotions x 4 speakers) and 4 practice stimuli with one item per speaker that were used for the training session.

Speaker	Age	Gender	Height	Average F0	F0 range
T2	36	F	1.68 m	302.23 Hz	200.71 - 437.38 Hz
T3	27	M	1.85 m	166.92 Hz	100.99 - 296.47 Hz
T5	25	F	1.63 m	282.89 Hz	199.49 - 429.38 Hz
T6	24	M	1.75 m	167.76 Hz	87.46 - 285.79 Hz

Table 1. Overview of the speakers' demographic information and voice characteristics.

Procedure

Children were tested in a quiet room at their home, and adults were tested in a quiet testing room at the two universities. The present experiment is part of a larger project (PICKA) on voice and speech perception conducted by the UMCG for which data were collected from the same population of children and adults in multiple experiments (Nagels et al., in review). The experiment started with a training session consisting of 4 practice stimuli and was followed by the test session consisting of 36 experimental stimuli. The total duration of the experiment was approximately 6 to 8 minutes. All items were presented to participants in a randomized order.

The experiment was conducted on a laptop with a touchscreen using a child-friendly interface that was developed in Matlab (Figure 1). The auditory stimuli were presented via Sennheiser HD 380 Pro headphones and calibrated to a sound level of 65 dBA. In each trial, participants heard a stimulus and then had to indicate which emotion was conveyed by clicking on one of three corresponding clowns on the screen. Visual feedback on the accuracy of responses was provided to motivate participants. Participants saw confetti falling down the screen after a correct response, and the parrot shaking its head after an incorrect response. After every two trials, one of the clowns in the back went one step up the ladder until the experiment was finished to keep children engaged and to give an indication of the progress of the experiment.

Figure 1. The experimental interface of the EmoHI test.

Data analysis

Children's accuracy scores were analyzed using the lme4 package (version 1.1.21, Bates et al., 2014) in R. A mixed effects logistic regression model with a three-way interaction between *language* (Dutch and English), *emotion* (happy, angry, and sad), and *age* in decimal years, and random intercepts per participant and per item was computed to determine the effects of language, emotion, and age on children's ability to recognize vocal emotions. We used backward stepwise selection with ANOVA Chi-Square tests to select the best fitting model, starting with the full factorial model, in lme4 syntax: $accuracy \sim language \cdot emotion \cdot age + (1|participant) + (1|item)$, and deleting one fixed factor at a time based on its significance. In addition, we performed Dunnett's tests on the Dutch and the English data with *accuracy* as an outcome variable and *age group* as a predictor variable using the DescTools package (version 0.99.25, Signorell et al., 2016) to investigate at what age Dutch and English children showed adult-like performance.

RESULTS AND DISCUSSION

Model comparison showed that the full model with random intercepts per participant and per item was significantly better than the full model with only random intercepts per participant [$\chi^2(1) = 393$, $p < 0.001$] or only random intercepts per item [$\chi^2(1) = 51.9$, $p < 0.001$]. Backward stepwise selection showed that the best fitting and most parsimonious model was the model with only a fixed effect of *age*, in lme4 syntax: $accuracy \sim age + (1|participant) + (1|item)$. This model did not significantly differ from the full model [$\chi^2(10) = 12.90$, $p = 0.23$] or any of the other models while being the most parsimonious. Figure 2 shows the data of individual participants and the median accuracy scores per age group for the Dutch and English participants. Children's ability to correctly recognize vocal emotions increased as a function of age [z-value = 8.91, estimate = 0.30, SE = 0.034, $p < 0.001$]. We did not find any significant effects of language or emotion on children's accuracy scores. Finally, the results of the Dunnett's tests showed that the accuracy scores of Dutch children of all tested age groups differed from Dutch adults [4-6 years difference = -0.47, $p < 0.001$; 6-8 years difference = -0.31, $p < 0.001$; 8-10 years difference = -0.19, $p < 0.001$; 10-12 years difference = -0.15, $p < 0.001$], and the accuracy scores of English children of all tested age groups differed from English adults [4-6 years difference = -0.43, $p < 0.001$; 6-8 years difference = -0.27, $p < 0.001$; 8-10 years difference = -0.20, $p < 0.001$; 10-12 years difference = -0.12, $p < 0.01$].

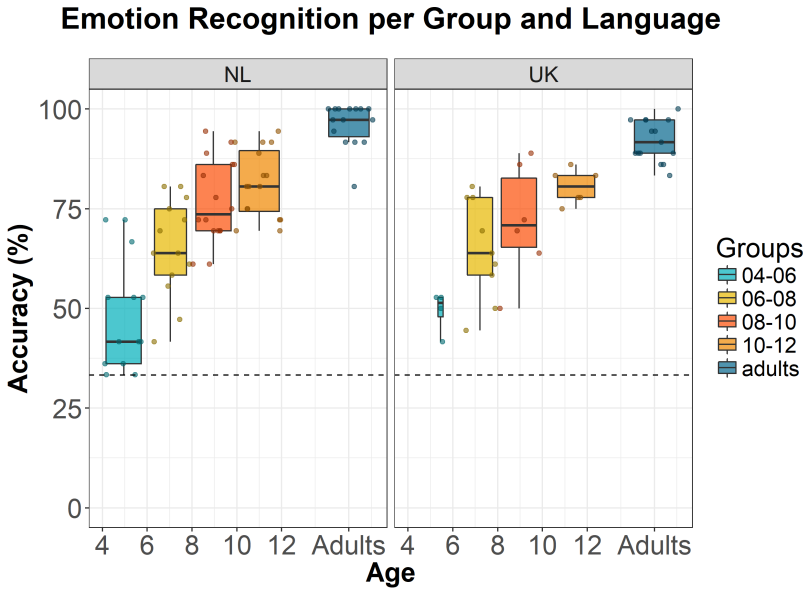

Figure 2. Accuracy scores of participants for emotion recognition per age group and per language (Dutch in the left panel; English in the right panel). The dots show individual data points at participants' decimal age (Netherlands (NL): $N_{children} = 58$, $N_{adults} = 15$; United Kingdom (UK) : $N_{children} = 25$, $N_{adults} = 15$). The boxplots show the median per age group, and the lower and upper quartiles. The whiskers indicate the lowest and highest data points within plus or minus 1.5 times the interquartile range.

Age effect

As shown by our results and the data displayed in Figure 2, children's ability to recognize vocal emotions improved gradually as a function of age. In addition, we found that, on average, even the oldest age group of 10- to 12-year-old Dutch and English children did not show adult-like performance yet. The 4-year-old children that were tested performed at or above chance level while adults generally showed near ceiling level performance, indicating that our test covers a wide range of age-related performances. Our results are in line with previous findings that children's ability to recognize vocal emotions improves as a function of age (Tonks et al., 2007; Sauter et al., 2013). It may be that children require more auditory experience to form robust representations of vocal emotions or rely on different acoustic cues than adults, as was shown for the development of sensitivity to voice cues (Mann et al., 1979; Nittrouer et al., 1993).

It is possible that the visual feedback caused some learning effects, although the correct response was not shown after an error, and learning would pose relatively high demands on auditory working memory, as there were only three items per speaker and per emotion presented in a randomized order.

Language effect

We did not find any cross-language effects between Dutch and English children's development of vocal emotion recognition, even though the materials were produced by Dutch native speakers. Earlier research has demonstrated that although adults are able to recognize vocal emotions across languages, there still seems to be a native language benefit (Van Bezooijen et al., 1983; Scherer et al., 2001). Listeners were better at recognizing vocal emotions that were produced by speakers of their native language than another language. However, these studies used five (Scherer et al., 2001) and nine (Van Bezooijen et al., 1983) different emotions which is likely considerably more complex then differentiating three basic emotions. In addition, the lack of a native language benefit may also be due to the fact that Dutch and English are closely related languages. We are currently collecting data from Turkish children and adults to investigate whether there are any detectable cross-language effects for typologically and phonologically more distinct languages.

Future directions

The results of the current study provide a baseline for the development of vocal emotion recognition for normal-hearing typically developing school-age children using the EmoHI test. Our results show that there is a large but relatively slow development in children's ability to recognize vocal emotions which also brings up the question on which specific acoustic cues children are basing their decisions and how this differs from adults. Future research using machine-learning approaches may be able to further explore such aspects. We are currently collecting data from children with CIs for whom the amount of auditory exposure is reduced due to degraded auditory input. The reduction of auditory exposure may delay or even limit the development of vocal emotion recognition in children with CIs, as some acoustic cues may not be available to hearing-impaired children due to degraded auditory input (Nakata et al., 2012). To conclude, the evident development in children's performance as a function of age and the generalizibility across the tested languages show the EmoHI Tests' suitability for future applications with hearing-impaired or other clinical populations of children and adults across different languages.

ACKNOWLEDGMENTS

We are grateful to all children, parents, and students that participated in the study, the speakers of our stimuli, and Basisschool de Brink in Ottersum, Basisschool de Petteflet, and BSO Huis de B in Groningen for their help with recruiting child participants. We would also like to thank Iris van Bommel, Evelien Birza, Paolo Toffanin, Jacqueline Libert, Jemima Phillpot, and Jop Luberti (illustrations) for their contribution to the development of the game interfaces, and Monita Chatterjee for her advice on recording the sound stimuli. This work was funded by the Center for Language Cognition Groningen (CLCG), a VICI Grant from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw) (Grant No. 918-17-603), the Medical Research Council (Senior Fellowship Grant S002537/1), and framework of the LabEx CeLyA ("Centre Lyonnais d'Acoustique", ANR-10-LABX-0060/ANR-11-IDEX-0007), and the French National Research Agency.

REFERENCES

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., and Grothendieck, G. (2014). Package 'lme4'. *R Foundation for Statistical Computing, Vienna*, 12.

Bänziger, T. and Scherer, K. R. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. In *A Blueprint for Affective Computing: A Sourcebook and Manual*, pages 271–294.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., and Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing research*, 322:151–162.

Dyck, M. J., Farrugia, C., Shochet, I. M., and Holmes-Brown, M. (2004). Emotion recognition/understanding ability in hearing or vision-impaired children: do sounds, sights, or words make the difference? *Journal of Child Psychology and Psychiatry*, 45(4):789–800.

Harms, M. B., Martin, A., and Wallace, G. L. (2010). Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. *Neuropsychology Review*, 20(3):290–322.

Hopyan-Misakyan, T. M., Gordon, K. A., Dennis, M., and Papsin, B. C. (2009). Recognition of Affective Speech Prosody and Facial Affect in Deaf Children with Unilateral Right Cochlear Implants. *Child Neuropsychology*, 15(2):136–146.

Mann, V. A., Diamond, R., and Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology*, 27(1):153–165.

Morton, J. B. and Trehub, S. E. (2001). Children's Understanding of Emotion in Speech. *Child Development*, 72(3):834–843.

Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., and Başkent, D. (in review). School-age children's development in sensitivity to voice gender cues is asymmetric.

Nakata, T., Trehub, S. E., and Kanda, Y. (2012). Effect of cochlear implants on children's perception and production of speech prosody. *The Journal of the Acoustical Society of America*, 131(2):1307–1314.

Nittrouer, S., Manning, C., and Meyer, G. (1993). The perceptual weighting of acoustic cues changes with linguistic experience. *The Journal of the Acoustical Society of America*, 94(3):1865–1865.

Nowicki, S. and Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, 18(1):9–35.

Sauter, D. A., Panattoni, C., and Happé, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology*, 31(1):97–113.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165.

Scherer, K. R., Banse, R., and Wallbott, H. G. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology*, 32(1):76–92.

Scherer, K. R. and Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In *A Blueprint for Affective Computing: A Sourcebook and Manual*, pages 271–294.

Signorell, A., Aho, K., Alfons, A., Anderegg, N., and Aragon, T. (2016). DescTools: Tools for descriptive statistics. R package version 0.99. 18. *R Found. Stat. Comput., Vienna, Austria*.

Tonks, J., Williams, W. H., Frampton, I., Yates, P., and Slater, A. (2007). Assessing emotion recognition in 9–15-years olds: Preliminary analysis of abilities in reading emotion from faces, voices and eyes. *Brain Injury*, 21(6):623–629.

Van Bezooijen, R., Otto, S. A., and Heenan, T. A. (1983). Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics. *Journal of Cross-Cultural Psychology*, 14(4):387–406.

A Study on the Lombard Effect in Telepresence Robotics

Ambre Davat[1,2], Gang Feng[1], Véronique Aubergé[2]

[1] GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP*, Grenoble, France
[2] LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP*, Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes

Corresponding Author:
Ambre Davat[1,2]
11 rue des Mathématiques, Saint Martin d'Héres, 38400, France
Email address: ambre.davat@gipsa-lab.fr

Abstract

In this study, we present a new experiment in order to study the Lombard effect in telepresence robotics. In this experiment, one person talks with a robot controled remotely by someone in a different room. The remote pilot (R) is immersed in both environments, while the local interlocutor (L) interacts directly with the robot. In this context, the position of the noise source, in the remote or in the local room, may modify the subjects' voice adaptations. In order to study in details this phenomenon, we propose four particular conditions: no added noise, noise in room R heard only by R, virtual noise in room L heard only by R, and noise in room L heard by both R and L. We measured the variations of maximum intensity in order to quantify the Lombard effect. Our results show that there is indeed a modification of voice intensity in all noisy conditions. However, the amplitude of this modification varies depending on the condition.

Introduction

When social entities are interacting, they automatically adapt their behavior to adverse conditions which could jeopardize their communication. In particular, vocal modifications due to noise are known as the Lombard effect. This phenomenon was first documented in humans (Lombard, 1911), then in other animal species, especially birds and mammals (Zollinger & Brumm, 2011). Moreover, the Lombard effect has implications for the design of virtual agents and social robots. Indeed, it can impair speech recognition systems, which are generally based on speech corpus recorded in quiet conditions (Hanson & Applebaum, 1990; Junqua, 1993). Implementing the Lombard effect in dialog systems would also increase their adaptation to noisy environment in a biomimetic way.

The Lombard effect consists mainly in an increase of speech intensity. It is an automatic phenomenon, which can only partly be controlled by the speaker (Pick et al., 1989). It varies with the type of noise and differs greatly from one individual to another. It also depends on the speaker's involvement. Indeed, there is some evidence showing that the Lombard effect is stronger during more interactive tasks: It increases during story-telling vs. labelling (Amazi Deborah K. & Garber Sharon R., 1982), communication with another vs. reading (Junqua, Fincke & Field, 1999), or vs. self-talk (Garnier, Henrich & Dubois, 2010). The sound immersion

techniques used in such experiments are also important, as speech modifications may be stronger when noise is played through headphones than over loudspeakers (Garnier, Henrich & Dubois, 2010). Furthermore, the intensity is well known as one of the prosodic parameters much implied for producing different socio-affective attitudes (authority, surprise etc.) in many languages (Aubergé, 2015). In face to face interaction, the intensity variations due to the Lombard effect are almost never confused with those having a socio-affective meaning. However, in the case of remote interaction, an ill-formed Lombard effect could be perceivably confused with social affect cues (Aubergé, 2017).

The paper will focus on a study concerning the Lombard effect in telepresence robotics. The specificity of this context is that all interlocutors are not present in the same room: one of them interacts remotely through a robot, which embodies its pilot in the "local" space. It is therefore an asymmetric system because the pilot needs to be acoustically immersed in both remote and local environment, while the interlocutors talk directly with the robot in the local environment. Because the Lombard effect happens automatically, it should affect the pilot's voice regardless of where the noise comes from. On the opposite, if the noise occurs in the remote space, the interlocutors cannot hear it as loud as the pilot, potentially not at all. However, they can hear the pilot's vocal adaptations, which could modify their behavior. In this paper, we propose an experiment in order to study these assumptions and present some results.

Materials & Methods

Methodology

In order to realize this study, we designed an experiment involving two subjects separated in two different rooms as shown on Figure 1. One of them played the role of the remote pilot (R), and the other one was the local interlocutor (L), interacting with the telepresence robot. Several configurations were tested in order to isolate the features of the Lombard effect in remote communication.

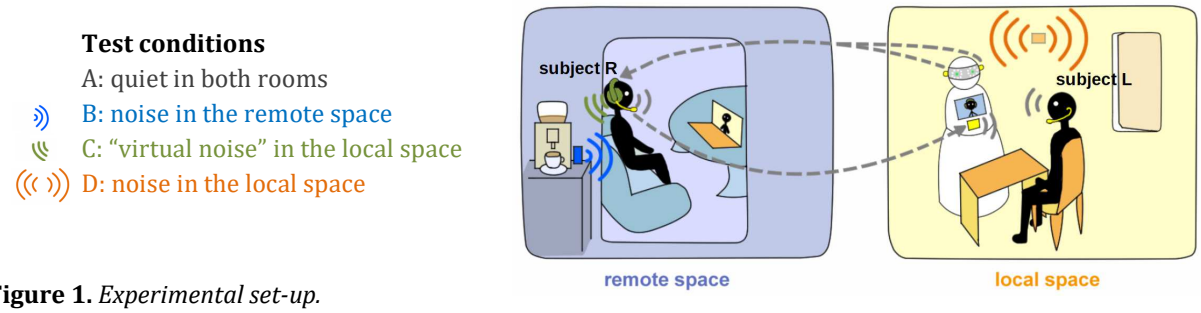

Figure 1. *Experimental set-up.*

The **condition A** was used in order to have a reference with no added noise. In all other conditions, R heard a noise, but this noise was not always audible by L. In **condition B**, a noise was played in R's room. As R voice was recorded with a headworn microphone, this noise was largely attenuated for the other subject, thus there was no need for R to speak louder to improve speech intelligibility. On the contrary, even if L did not hear the noise in this condition, L could need to speak louder in order to be understood by R. In **condition D**, a noise was played in the local space, and both subjects could hear it. This case is close to the standard communication with noise, where interlocutors are sharing the same environment. Finally, in **condition C**, a pre-recorded noise was injected in the headphones of R, in order to make R believe that this sound was existing in the local space and was heard by L.

Procedure

The experiment was introduced to the subjects as a test of the visual interface of a telepresence robot. The subject L was sitting in front of the robot and asked a list of simple questions to the subject R, who was the pilot of the robot. These questions were designed in order to be very simple, and trigger a finite set of answers. Examples of these questions (translated in English) are: "Which is the color of the grass?", "How many legs does a horse have?", "How much are 2 x 2?". Every 10 questions, there was a special instruction, requesting R to handle the robot via the interface. We pretended that we were measuring the answering time, in order to quantify the accessibility of the pilot interface. The subjects were informed that the test has to be done without a break. Indeed, during our first test, the subject L stopped speaking when the noise D was playing. The three different noises were regularly triggered by an experimenter.

To prevent bias in the experimental results, we wanted to nudge the subjects into thinking that noise occurrences had no link with the experiment. The noise sources were therefore hidden and diffused occasionally. We used pre-recorded realistic noises of a coffee machine and a drill. A ladder and a tool box were placed in the corridor at the entrance of the platform to suggest that there were building works in progress. Before the debriefing, most of the subjects believed that the noises were incidental, and none of them guessed that the aim of the experiment was to measure their vocal adaptation to noise. This confirms that our scenario was credible. Fulfilling the list of 140 questions took approximately 15 minutes long.

Technical specifications

Robot

The telepresence robot used was RobAIR Social Touch. It was co-constructed with the fablab of the LIG (*Laboratoire d'Informatique de Grenoble*). It uses a ROS architecture, and the teleconferencing interface is based on WebRTC. Contrary to most of telepresence robots (Kristoffersson, Coradeschi & Loutfi, 2013), it does not look like a screen on wheels, but is closer to a slightly anthropomorphic robot carrying a tablet computer. On either side of the "head" of the robot, one omnidirectional microphone Behringer B5 is placed. Signals recorded by both microphones are digitalized by an audio interface UR22MKII (Steinberg), and sent to R's headphones in order to reproduce a pseudo binaural hearing. R's voice is emitted by a loudspeaker JBL GO+ which is placed under the tablet computer.

Echo cancelation

A local wireless router was used, the network latency was therefore negligible. However, visual and vocal signals cannot be transmitted instantaneously by WebRTC applications. As (Počta & Komperda, 2016), we obtained a mouth-to-ear delay of around 150 ms, which is pretty good according to telecommunication standards (ITU, 2003). This delay means, though, that there is an echo effect: the pilot R can hear her/his voice emitted by the loudspeakers of the robot and recorded by its microphones. This echo effect can greatly affect the quality of communication, so we used an algorithm of echo-cancelation. It consists simply in reducing the volume of R's headphones when s/he is talking, and waiting 150 ms before returning to the usual volume.

Noise sources

We used two loudspeakers JBL GO+ as sound sources. One was placed next to a coffee machine in the corridor adjacent to R's room. It diffused an amplified audio recording of the coffee machine startup (condition B). The other loudspeaker was placed in a room adjacent to L's room and diffused drill noises (condition D), which were pre-recorded with the microphones of the robot (condition C). Both sound sources were calibrated at 55 dB(A) with a sound level meter (Lutron SL-4001) placed at the location of the subjects. Moreover, the local space was particularly reverberant, with a reverberation time of about 0.8 s.

Recordings and calibration

The voices of both subjects were captured by two wireless headworn microphones Sennheiser HSP4 and digitalized with an audio interface UR22MKII (Steinberg). R's signal was sent to L through the loudspeakers of the robot. However, L's signal was only used for measurement purpose. What R heard of the local space was recorded by the microphones of the robot. In addition, the signal of the internal microphone of the computer, as well as the monitor signal of the stream heard by R were also recorded, in order to be able to track noise occurrences.

The headphones worn by R were an AKG K242. They were calibrated with an artificial ear (Brüel & Kjær), in order to make the sounds perceived by R as loud as those in L's room. The sound attenuation through these headphones is negligible. The loudspeaker of the robot was also calibrated with a sound level meter (Lutron SL-4001), to ensure that the loudness of R's voice was faithfully transmitted.

Analysis of the results

14 groups of 2 subjects participated in this experiment. Most of them were native French speakers (25/28), two were fluent in French and one has a basic level, but sufficient to read the questions. Noise sequences were annotated and we extracted each keyword answered by subjects R (ex: "green"), and each pattern of questions read by subjects L (ex: "What is the color of…?"). We studied only the keywords / questions which were repeated at least 50 times among all the tests.

Recordings were filtered with a A-weighting digital filter (Zhivomirov, 2019) and voice intensity was computed by segments of 20 ms. Global results for maximum intensity can be seen on **Figure 2**. However, these results hide a great variability between subjects and between key-words / questions which could bias the analysis. Indeed, the number of each keyword in each condition varies from on experiment to another, because the answers of subjects R are not constrained. Therefore, in order to properly quantify the differences between each condition, we implemented a linear mixed effect model with *R*, following the tutorial of (Winter, 2013). This method allows to build a very simple model from the data, and at the same time, it provides measures of statistical significance. Written in *R* formula, this model is:

$$max\ intensity \sim noise\ condition + (1|keyword) + (1|subject)$$

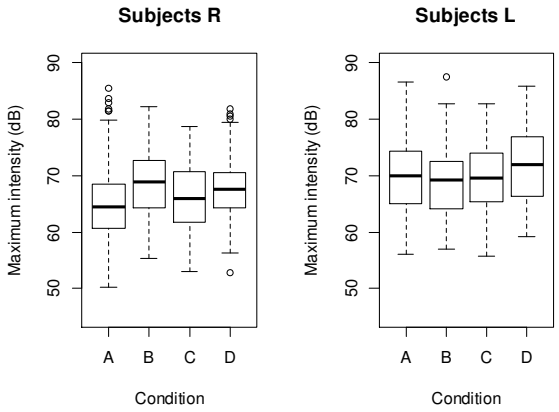

Figure 2. *Maximum intensity in each noise condition*

Summary of the models are shown in Table 1. They are coherent with the previous visualization, which means that we can base our observations on both figures. First, we note that the subjects R increased their maximum intensity in the 3 noisy conditions, while the subjects L spoke louder only in condition D, that is, when they

could hear the noise. This shows that the Lombard Effect can be observed during telepresence robotics. However, it seems that for everyday noise with a moderate intensity level and when the subjects are focused on a question/answer task, this effect is very small, in the range of 1 to 3 dB. By way of comparison, (Winkworth Alison L. & Davis Pamela J., 1997) found an increase up to 10 dB SPL with a Cocktail party noise at 55 dB SPL for reading and monologue tasks.

	Subjects R				**Subjects L**			
Condition	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Maximum intensity (dB)	64.61	+ 2.98	+ 1.19	+ 2.38	69.19	- 0.06	- 0.30	+ 2.24
Standard error (dB)	1.11	0.36	0.37	0.36	1.86	0.21	0.21	0.20
Number of extracts	1263	160	146	150	1125	118	129	143
Statistical significance	$\chi^2(1) = 98.47\ ;\ p < 2.2\ e^{-16}$				$\chi^2(1) = 132.59\ ;\ p < 2.2\ e^{-16}$			

Table 1. *Results of the linear mixed effect model (applied separately for R-data and L-data).*
Statistical significance was obtained by comparison with the null-model: $max\ intensity \sim 1 + (1|keyword) + (1|subject)$

Another interesting result is that the increase of intensity for subjects R depends on the noise condition. It is higher in condition B vs C and D, namely when the noise was played in the same room as the pilot, and not in the headphones, which is the opposite of the results expected from (Garnier, Henrich & Dubois, 2010). However, the noises used in condition B were very different from the ones used in C/D. Indeed, they presented some intensity spikes up to 61 dB(A). Moreover, a posteriori recordings with the robot in R's room also showed that the intensity of the noises B was about 3 dB louder during stationary phase than the noises C/D. Besides, it is worth noticing that while the subjects R were talking, the volume of their headphones was reduced from 50% to 10%, in order to perform echo cancelation, which means that the intensity of noises C/D was greatly reduced at the time they spoke, which was not the case in condition B.

The most interesting result concerns the difference between condition C and D. The subjects R heard the same noises in both conditions, but their increase of intensity was greater in condition D, when the subjects L could also hear the noise and adapt to it. This may highlight an effect of entrainment: the subjects R increased their voice intensity not only because of the noise, but also because their interlocutors L were speaking louder. Such observations were also made by (Székely, Keane & Carson-Berndsen, 2015). However, this effect was not observed for the subjects L, who did not increase their voice intensity in condition B and C. This may be explained by the nature of their task (reading) or because they were not able to hear the noise which made their interlocutors speak louder.

Discussion and conclusion

In order to study the Lombard effect in the context of telepresence robotics, we performed an experiment with pairs of subjects (R and L), focused on a question/answer task. Four noise conditions were tested: A - without noise, B - only R hears the noise, C - only R hears the noise over headphones, and D - R and L hear the noise. The noise occurrences were very short and perceived as accidental by the participants. However, whenever they were able to hear the noise, they had a tendency to speak a bit louder. An entrainment effect was also noted for the subjects R, who spoke louder when L could also hear the noise. These increases of voice intensity being very subtle, they could easily be mistaken as expressive variations indicating socio-affects. Further work will involve comparing the pitch and durations of vocal productions in the four different noise conditions.

Acknowledgements

We would like to thank Coriandre Villain (GIPSA-lab) for his technical assistance during the calibration. Thanks also to Emeline Le Goff and Zoé Giorgis (Univ. Grenoble Alpes) for their support during the development of the experimenter's interface and for annotating the data.

References

Amazi Deborah K., Garber Sharon R. 1982. The Lombard Sign as a Function of Age and Task. *Journal of Speech, Language, and Hearing Research* 25:581–585. DOI: 10.1044/jshr.2504.581.

Aubergé V. 2015. Gestual-facial-vocal prosody as the main tool of the socio-affective "glue": interaction is a dynamic system. Presented at *International workshop on audio-visual affective prosody in social interaction.* Bordeaux, France.

Aubergé V. 2017. The socio-affective glue: how to manage with the empathic illusion of human for robot? Presented at *VIHAR-2017 workshop*. Skövde, Sweden.

Garnier M, Henrich N, Dubois D. 2010. Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research* 53:588–608. DOI: 10.1044/1092-4388(2009/08-0138).

Hanson BA, Applebaum TH. 1990. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech. In: *International Conference on Acoustics, Speech, and Signal Processing*. 857–860 vol.2. DOI: 10.1109/ICASSP.1990.115973.

ITU. 2003. *UIT-T G.114: One-way transmission time*. International Telecommunication Union.

Junqua J. 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America* 93:510–524. DOI: 10.1121/1.405631.

Junqua J-C, Fincke S, Field K. 1999. The Lombard effect: a reflex to better communicate with others in noise. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. Phoenix, AZ, USA: IEEE, 2083–2086 vol.4. DOI: 10.1109/ICASSP.1999.758343.

Kristoffersson A, Coradeschi S, Loutfi A. 2013. A Review of Mobile Robotic Telepresence. *Adv. in Hum.-Comp. Int.* 2013:3:3–3:3. DOI: 10.1155/2013/902316.

Lombard É. 1911. Le signe de l'élévation de la voix [The sign of voice raising]. *Annales des Maladies de l'Oreille et du Larynx* XXXVII:101–109.

Pick HL, Siegel GM, Fox PW, Garber SR, Kearney JK. 1989. Inhibiting the Lombard effect. *The Journal of the Acoustical Society of America* 85:894–900. DOI: 10.1121/1.397561.

Počta P, Komperda O. 2016. A Black Box Analysis of WebRTC Mouth-to-Ear Delays. *Communications – Scientific Letters of the University of Zilina*.

Székely É, Keane MT, Carson-Berndsen J. 2015. The Effect of Soft, Modal and Loud Voice Levels on Entrainment in Noisy Conditions. *Interspeech.*

Winkworth Alison L., Davis Pamela J. 1997. Speech Breathing and the Lombard Effect. *Journal of Speech, Language, and Hearing Research* 40:159–169. DOI: 10.1044/jslhr.4001.159.

Winter B. 2013. Linear models and linear mixed effects models in R with linguistic applications.

Zhivomirov H. 2019. *A-weighting Filter with Matlab*. MATLAB Central File Exchange.

Zollinger SA, Brumm H. 2011. The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* 148:1173–1198. DOI: 10.1163/000579511X605759.

"Where do you go, Trico?": Talking to Animal Companions in the Video Game *The Last Guardian*

Hiloko Kato

Department of German Studies, University of Zurich, Zurich, Switzerland

Corresponding Author:
Hiloko Kato
Schönberggasse 9, Zurich, 8001, Switzerland
Email address: hiloko.kato@ds.uzh.ch

Abstract

With the rapid development of technology used in creating video games, the use of non-player character (NPCs) — characters controlled by the game that truly interact with the gamer—have become a well-established feature of video games in general. On the one hand, NPCs are state of the art and therefore cannot be left aside. On the other hand, they are very useful in terms of game mechanics and narrative, as they can be used—for example—as companions to the player's characters, often acting as helpful guides in vast open-world scenarios. These NPCs are predominantly human beings or anthropomorphised animals. Nonetheless, many games feature authentic animal characters that the player is allowed to interact with. It is not unusual for a player to form bonds with such NPCs, or even develop a kind of a relationship with them.

Against this background, this paper aims to look at different examples of interaction between players and their (authentic) animal companion in the highly lauded video game *The Last Guardian*, where the player's companion is a giant mythical man-eating beast called Trico. How is bonding or relationship building established and enhanced in this unusual scenario? How is the resulting attachment made perceivable? And what kind of strategies of interactivity—particularly verbal and vocal interactivity—occur between player and companion? In order to answer these questions, this study analyses data from YouTube videos of so-called 'Let's Players', who make their actions and decisions accountable by constantly commenting their experiences during the playthrough. The methodology used to analyse the data is based on the methods of ethnomethodolgy and conversation analysis, including detailed transcriptions of the analysed material.

Background

In 2009, a preview of the computer game *The Last Guardian* was released at the premier trade event of the video game industry, E3. The trailer shows a gigantic mythical half-bird half-mammal creature and a boy interacting with each other, hinting in certain key moments at the co-operational gameplay and, furthermore, their apparently fond relationship (i.e., the boy patting the creature's head or sleeping snuggled up in his feathers, cf. G4VideogameTrailers, 2009). Being the third instalment of a highly acclaimed saga created by game designer Fumito Ueda, the trailer whipped fans of the two first titles—*Ico* (Team Ico, 2001), and *Shadow of the Colossus* (Team Ico, 2005/2011/2018)—into a state of frenzy. However, *The Last Guardian* was not released until seven years later (mainly due to the change of hardware from PS3 to PS4), in 2016, causing devoted YouTubers to announce their enthusiasm with Let's Play titles such as "THE LONG WAIT IS FINALLY OVER!" (Pewdiepie) and "I've WAITED SO LONG FOR THIS!!!" (Jacksepticeye). As the trailer hinted, the gameplay consists of the player controlling the unnamed boy, cooperating with the mythical beast called Trico—a non-player character (NPC) controlled by the computer game—unravelling riddles from their mysterious past by solving puzzles and exploring the vast area of the game world. *The Last Guardian* focuses exclusively on the interaction between the player and an authentic animal NPC, bringing it to a whole new level: Despite the fact that Trico is introduced as a dangerous, man-eating beast, an attachment—and even deeply emotional relationship—forms between the player and the animal companion, which can be followed already in the first thirty minutes of the playthrough (averaging approximately twelve hours in total).

Though the still-young field of human-animal studies (HAS) has increasing scholarly output, research concerning virtual animals is still scarce. In game studies, as well, subjects concerning animals in video games, thematising their representation, their role within the game, or even their emotional value are difficult to find. It seems that in virtual worlds, the topics of ethics and emotions are still reserved for human beings—i.e., the players and their avatars (cf. Sicart, 2009; Isibister, 2013)—with the exception of Chittaro and Sioni (2012), who focus on aspects of violence against animals—namely insects (against which violence seems more acceptable than against other species)—and Wilson (2009), who, in contrast, describes the ethical use of farm animals according to the rules of *Castles of Burgundy*, even though it is an analogue boardgame. Janski (2016) provides a categorisation for animals in video games, emphasising the ethically questionable fact that they still act mostly as enemies or background entities. The intriguing aspect is the transformation suggested by the game creators: Referred to as man-eating beast, Trico is in fact introduced as an enemy (this is underlined more pointedly in the original Japanese game title 人喰いの大鷲トリコ, *The Man-Eating Giant Eagle, Trico*). At the same time, he is designated—mostly by the first insights in the trailer from 2009—to become a dear companion.

Against this multi-layered background, this paper presents a preliminary study on how the bonding or relationship between players and their mythical companion in *The Last Guardian* is established and enhanced. This was done by analysing how this attachment is made perceivable, primarily through means of verbal and vocal interactivity.

Materials & Methods

In order to answer the question of how bonding—or even a relationship—between the player and their animal companion is established and enhanced, it is necessary to closely look at the interaction between the two. Considering this, the method of ethnomethodology and of conversation analysis are being used, as they "provide an analytic resource through which we can begin to exploit the opportunities provided through video" (Heath 2004, 279). Instead of observing players by asking them to constantly commentate their actions while playing, the chosen material is "naturally occuring data" (ten Have, 2007: 9) and consists of YouTube videos of so called 'Let's Plays'. Being both performance and entertainment, these videos primarily demand continual commentary by the player(s) while playing through a game. By doing so, their actions are—in the terminology of ethnomethodology—made *accountable*: "I mean observable-and-reportable, i.e., available to members as situated practices of looking-and-telling" (cf. Garfinkel, 1967: 1). This making-accountable is done using all aspects of verbal and non-verbal communication. It is a peculiar and fascinating—because it is mostly accurate—claim that everything of importance and relevance for the interaction is brought forward by the participants, and thus can be observed. In our case, the method of ethnomethodology and of conversation analysis very much suits the setting of the analysis, as it forces us to really stay focused on the interaction itself by operating "closer to the phenomenon" and by working "on detailed renderings of interactional activities" (ten Have, 2007: 9).

The Let's Plays were chosen semi-randomly—the high profile of the Let's Player (i.e., Pewdiepie) or the quantity of viewers (i.e., Jacksepticeye) being determining factors. Admittedly, other videos were chosen randomly—emphasising the ethnomethodological approach that any material has to make its issues accountable—from the vast amount of Let's Play videos uploaded daily. The material consists of videos by the following four Let's Players: devotees Pewdiepie [P] and Jacksepticeye [J] and freshmenChristopherOdd [C] and TetraNinja [N]. (The letters in the brackets will be used in the transcripts, along with [T] for Trico). The lack of female Let's Players is deliberate, because the subject of gender and gaming—game studies being aware that female gamers are much rarer than their male counterpars—cannot be tackled in this paper.

First Insights

This section presents first insights on how bonding with animal companions in video games is established and made perceivable in *The Last Guardian*. For this purpose, a set of initial sequences within the first thirty minutes of video material of the four Let's Plays mentioned above was analysed, focusing on the most prominent and important features in verbal and vocal interaction.

The resulting major points are linked here with further implications and interpretation:

a. direct speech to address Trico – enabled through engaging into empathic actions (injuries, food)
b. reference to authenticity of appearance and behaviour of Trico – ensuring him to be an equal interactional partner
c. comparison to known animal behaviour or even own pets – as Trico is an unknown mythical creature
d. development and expanding of Trico's interactional repertoire – inviting the player to react accordingly

These aspects apply to all of the Let's players, even though they have differing background in knowledge and even in stance on companionship as such. It is interesting to see, how the two devotees Pewdiepie and Jacksepticeye have to adjust their initial most positive attitude towards Trico, which initially behaves not as a fond companion but turns out to be utterly hostile. They notwithstanding follow the same strategies as the others.

The next two aspects are more specialised features of some of the Let's players, but seem nontheless important:

e. talking as Trico and imitating kind of creaturesque voice – immerging in the mindset of Trico
f. talking inside the game world as the player character (although shortly) – immerging into of the game world by using its language

Pewdiepie, who compares his pet dog Edgar with Trico regarding behaviouristic issues (something he calls "dumb" without explaining in detail in this material what is meant), is also the one who extensively talks as Trico. The way he is doing this prosodically appears to express his playful attitude, but there seems to be more that meets the eye here: Talking as Trico is also a possible attempt to make its intentions perceivable and by thus all the more

controllable. This is in a sense similar to the function of the narrator of the game, who being the older self of the boy lets the player access his thoughts, making the whole situation more comprehensible (cf. Brotherson, 2016). Trico's highly mentioned authenticity (b.) implies him having a will of its own, making him and his behaviour often difficult to comprehend and to control. The statement by ChristopherOdd illustrates this aspect perfectly: "What do we know, we've never met a man-eating beast before" (ChristopherOdd, 2016, at 00:09:40). Therefore, lending him a voice seems to be a way to overcome his uncertain otherness and to be able to bond with him more easily. Two last important aspects shall be pointed out. At the same time, they represent two main fundamental concepts commonly found in video game analysis (cf. Kato/Bauer 2018).

g. game mechanics provide certain key moments, where gradual enhanced bonding relationship is made perceivable, but where the player merely reacts to what is given, not setting rules and guidelines himself
h. on the contrary, the player is able to act as he thinks best when labelling Trico (pet, friend, pal), despite the predefined conceptions considered by the game creators

To give a more detailed insight in the analysis, two examples may be given. They illustrate the mentioned points a) and g) from the list above. The transcripts, which are provided for better comprehension of how things are said, or in context of what is happening in the game, have been created on the basis of the German transcription system GAT 2 (cf. Selting et al., 2009). For the transcription conventions of this system and the meaning of individual symbols, please refer to the key at the end of this article.

a) First Evidence of Interaction as Turning-Point: From talking about to talking to Trico

One of the most intriguing aspects concerns the way Trico is addressed by the players. During the first encounter between the player and the mythical beast, all except ChristopherOdd switch from third-person narrative to a more direct way of addressing the beast. Pewdiepie is the first who switches to direct speech, resuming his concern about the unexpected beginning, when Trico awakes, but immediately turns hostile ("Trico, oh my god, what happened?", Pewdiepie, 2016, at 00:06:05). TetraNinja is the second, who instantly tries to pet Trico ("Friend? Friend? Let me pet you!", TetraNinja, 2016, at 00:05:30), which nourishes the supposition that doing something with the interactional partner incites the use of direct speech. The turning point in Jacksepticeye's playthrough seems to confirm this as he switches to direct speech [J1_08] after he realises that Trico is injured [J1_02; Jacksepticeye, 2016, transcript begins at 00:06:22]:

```
            Trico awakens and turns hostile, roaring at the boy, Jacksepticeye looks concerned
J1_01  J:   ! 'OH!,
             boy backs off, turning to face Trico from a distance, stands still, Trico calms down,
             talk by the narrator, boy and Trico face each other steadfastly
J1_02       <<h> OH he is !^HURT!>;
J1_03       oh GOD.
J1_04       <<ff> I am gonna !AB!solutely (-) FALL in lOve with this crEAture.>
            Jacksepticeye looks agitatedly into the camera, facing the viewer directly
J1_05       by the WAY;
J1_06       the this is what this games is ALL gonna be abOUt;
            Jacksepticeye looks back onto his screen
J1_07  T:   wincing sound, looks at his side
J1_08  J:   ↑hel'LO?
            boy starts to approach Trico slowly
J1_09       ↑hey ¯TRIco?
            boy starts to run towards Trico
J1_10       ↑you_re good ¯BUDdy?
J1_11       he got that THING in his lEg;
```

Significantly, Jacksepticeye inserts a (fortissimo) meta-statement about his future relationsship with Trico at this very turning-point [L J1_04– J1_06]. Even if the reason for placing his commitment against all odds right at this point is not made accountable, the fact that Trico is hurt (L J1_02) supports the positive stance of Jacksepticeye as a helping person in this game. While he performs this meta-statement, his player character and Trico face each other steadfastly. What happens next is a beautiful establishment of initial hellos, even though, of course, only the player is performing the addressing: Trico quits the staring by wincing and looking at his side [J1_07], causing Jacksepticeye to immediately react to this action with a high-pitched "hello" [J1_08]. It seems that he insinuates a conversational opening, with the behaviour of Trico functioning as the *summons* (Schegloff 1968). This is also reasonable to assume as at this point he (aka the boy) starts to move towards the verbal conversational partner. The physical interaction during the following removal of a spear results in an increase in the direct speech by all three players, who had already started to address Trico before. ChristopherOdd is the last to switch in direct

speech [L C1_01], after five long minutes of playthrough. Notably, it coincides with a statement about authenticity, which is prosodically complex by rising and falling fluctuation in pitch [L C1_03–C2_04; ChristopherOdd, 2016, transcript begins at 00:10:06]:

```
C1_01   C:  so you_re not sUre if you WANT that one?
C1_02       is THAT the deal?
C1_03       he_s sO ´LIfe `like.
C1_04       it_s !^CRA!zy.
```

This appealing combination underlines the assumption that authenticity of the animal's appearance and behaviour very much incites the acceptance of the virtual other as a partner in interaction, as the player begins treating him as a partner in communication as well.

g. Key moments implemented by the game (mechanics)

It is the game mechanics that provide certain key moments, where gradual enhanced bonding or even relationship is made perceivable and where the player merely reacts to what is given, rather than setting rules and guidelines himself. In these key moments, the player is confronted with tasks, which turn out to be one of the main driving forces leading to interaction with Trico and at the same time, strongly invite the player to act in an empathic way: The injured animal has to be treated and fed, and freed from his chains at last. Thoughtfully, this moment of liberation is connected to the possibility to call Trico by pushing a certain controller button, tying back the freedom aspect in favour of that of companionship. Or in other words: Trico is not set free to be free, but to make him stay and follow. Pewdiepie choses an other way to make Trico stay with him: He climbs on him and refuses to let go [L P1_03], emphasizing on their togetherness by stressing the "we" and purposefully rendering a grammatically incorrect sentence (L P1_04, Pewdiepie, 2016, transcript begins at 00:16:05).

```
P1_01  P:  your_re !FREE! trIco-
P1_02      you look a little NAked but that_s ok.
           sits in the beast's neck holding on to the fur on its head, head-up-display
           indicates how to let go by pressing a controller button
P1_03      <<breathy voice> i_m NOT gonna let go.>
P1_04      `WE_re gonna go.
```

By doing this, Pewdiepie misses another key moment of bonding within the liberation scene, which is provided by a cutscene: When standing on the ground, the boy is confronted with Trico lowering down to him. The boy carresses Trico's nose, saying laughingly "So that's what you look like! Pleased to meet you, Trico!", engaging into a proper introduction. At this scene, clearly staged as highlight by camera work and use of music, Jacksepticeye is shouting out loud in excitement. For ChristopherOdd, this is the moment to start talking directly to Trico whenever they have to cooperate, making the impression that they actually interact with each other (i.e. "Where do you go, Trico?", ChristopherOdd, 2016, at 00:15:07). It seems telling, that ChristopherOdd, who went into the game as a freshman, whose attachment grew steadily (i.e. being the last to change into direct speech) and who reads the cues of animal behaviour best (i.e. by finding the key to pet him at will), is the only player who make his perception in this aspect accountable:

```
C2_01  C:  HE: doesn_t seem to fond of the wAter situation so- (---)
C2_02      <<smiling voice> you_re kInd to start to sEE our our relAtionship with hIm (--) b BUILding>
C2_03      like we are both in this BAD situation togEther (.) but-
C2_04      Obviously there_s some things he can dO that I can NOT?
C2_05      A:ND-
C2_06      vice VERsa;
```

What ChristopherOdd expresses here–also on a prosodic level by stressing the words–is first of all, that he actually is "building" a "relationship" with Trico [L C2_02], even making explicit, that this is something that is obviously perceivable for the viewer ("start to see"). In addition, he makes clear, that this mentioned relationship has to be understood as equal as they both have to help each other [L C2_3–C2_06], underlining the aspects of companionship once more.

Outlook

This paper constitutes only a small range of preliminary findings making way for a more extensive discussion of the analysed material. First, the major points brought up in this paper have to be comprehended more thoroughly.

Second, further insight into the interaction between the player and an authentic animal NPC has to be gained. To this end, more data of the already selected four Let's plays has to be analysed in detail. Also the time span of data has to be expanded beyond the first thirty minutes of the playthrough: The analysis of the development throughout the course of the game seems interesting as the already established bond and trust is put to the test by the game in various ways later on, i.e. when Trico's stubbornness annoys the player or in the even more pressing situation, when Trico actually becomes a man-eating beast. This yields the opportunity to gain a deeper understanding of the player-beast interaction in situations of conflict.
After a detailed analysis of the present data, extending the dataset with additional Let's plays is a mandatory next step. In doing so, gender related aspects are worth exploring in a more thorough manner as well: Are female players bonding in a different way than their male counterparts, especially considering previously discussed aspects like verbal or non-verbal expressions, making their bonding accountable and explaining their emotional relationship?
In order to further support the findings presented in this paper a quantitative study could be conducted allowing for more in depth analysis like comparing different outcomes of the Let's plays by tying them back to the various backgrounds (i.e. playing the game with previous knowledge or playing it 'blind').
An other very important point is to expand the different foci by analysing other video games featuring authentic (non-anthropomorphised) animal companions. The idea for the *The Last Guardian* was given to game designer Fumito Ueda by experiencing the emotional bonding of the players to the player character's horse in *Shadow of the Colossus*, which was unintended and therefore all the more surprising. It proves that with the increasing authenticity of the digitally depicted animals, the willingness of the player to bond with his companion is heightened. Of course, a video game like *The Last Guardian* carries this topic to extremes. But also other video games with much lower threshold invite the player to bond with their companion, revealing much of their own attitude towards animals in general.

Key to GAT2 transcriptions

(the list below only contains the conventions relevant to this article)

robert_s	words joined together within units
<<breathy voice> >	para- and extralingustic actions and events accompanying speech
acCENT	focal stress, accentuation
accEnt	secondary stress
ac!CENT!	pronounced stress
Fluctuations in pitch at the end of intonational phrases:	
?	steep rise
,	medium rise
–	even level
;	medium drop
.	steep drop
Intralinear notation of fluctuations in stress and pitch	
^SO	rising-falling
¯SO	rising-falling
`	rising
´	falling
↑	leap to high-pitched voice
<<h> >	high-pitched voice
Changes in volume and pace of speech:	
<<ff> >	fortissimo, very loud

References

Chittaro L., Sioni R. 2012. Killing Non-Human Animals in Video Games: A Study on User Experience and Desensitization to Violence Aspects. PsychNology Journal, Volume 10, Number 3: 215-243.
Darwin C. 1872/1989. The Expression of the Emotions in Man and Animals. New York: Oxford University Press.
Brotherson C. 2016. The Last Guardian: 5 Storytelling Secrets. PlayStation Blog. Retrieved 21 June 2016. Available at https://blog.us.playstation.com/2016/06/21/the-last-guardian-5-storytelling-secrets/ [accessed 18.05.2019].
Garfinkel H. 1967. Studies in Ethnomethodology. Englewood Cliffs/NJ: Prentice Hall.
Goffman E. 1959. The Presentation of Self in Everyday Life. New York: Doubleday.

Heath Ch. 2004. Analysing face-to-face-interaction: video, the visual and material. Silverman, D. (ed.): Qualitative Research. Theory, Method and Practice. London: SAGE: 266-282.
Isibister K. 2013. How Games Move Us. Emotion by Games. Cambridge: The MIT Press.
Janski K. 2016. Towards a Categorisation of Animals in Video Games. Homo Ludens 1 (9): 85-102.
Kato H., Bauer R. 2018. The Player as Puppet? Visualised Decisions as a Challenge for Computer Games. In: Kocher M., Bauer R., Suter B. (eds.): Game Mechanics - Rules for the Magic Circle. Bielefeld: transcript Verlag: 217-241.
Schegloff E. 1968: Sequencing in Conversational Openings. American Anthropologist, 70, 1075-1095.
Selting M. et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion. 10: 353–402.
Sicart M. 2009. The Ethics of Computer Games. Cambridge: The MIT Press.
ten Have P. 2007. Doing Conversation Analysis. A Practical Guide. Los Angeles: SAGE.
Wilson D. 2017. The Eurogame as Heterotopia. In: Torner E., Waldron E.L., Trammell, A. (eds.): Analog Game Studies. Volume 2. Pittsburgh: ETC Press: 43-49.

Let's Plays and Youtube Video

ChristopherOdd. 2016. The Last Guardian Gameplay Part 1 - TRICO! - Lets Play Walkthrough. https://www.youtube.com/watch?v=XH4f-GRRhUA [accessed 18.05.2019].
Jacksepticeye. 2016. I'VE WAITED SO LONG FOR THIS!!! | The Last Guardian - Part 1. https://www.youtube.com/watch?v=2euZWNuZRTk&t=99s [accessed 18.05.2019].
PewDiePie. 2016. THE LONG WAIT IS FINALLY OVER! - The Last Guardian - Part 1. https://www.youtube.com/watch?v=0wuBbymxkMI&t=1s [accessed 18.05.2019].
TetraNinja. 2016. The Last Guardian Walkthrough Part 1 - Trico (PS4 Pro Let's Play Commentary). https://www.youtube.com/watch?v=Rt3F8VY0wME&t=1785s [accessed 18.05.2019].
G4VideogameTrailers. 2009. The Last Guardian E3 2009 Trailer. https://www.youtube.com/watch?v=GCQ-RZrz5rI [accessed 18.05.2019].

Large-scale unsupervised clustering of Orca vocalizations: a model for describing Orca communication systems

Marion Poupard[1,5], Paul Best[1], Jan Schlüter[1], Helena Symonds[2], Paul Spong[2], Thierry Lengagne[3], Thierry Soriano[4], and Hervé Glotin[1]

[1]**Univ. Toulon, Aix Marseille Univ., CNRS, LIS, DYNI, Marseille, France**
[2]**Orcalab Alert Bay, Canada**
[3]**LENHA, Univ. Lyon 1, CNRS, France**
[4]**Univ. Toulon, COSMER EA 9378, Toulon, France**
[5]**Biosong SARL, France**

Corresponding author:
Marion Poupard, Hervé Glotin

Email address: marion.poupard@univ-tln.fr, glotin@univ-tln.fr

ABSTRACT

Killer whales (*Orcinus orca*) can produce 3 types of signals: clicks, whistles and vocalizations. This study focuses on Orca vocalizations from northern Vancouver Island (Hanson Island) where the NGO Orcalab developed a multi-hydrophone recording station to study Orcas. The acoustic station is composed of 5 hydrophones and extends over 50 km^2 of ocean. Since 2015 we are continuously streaming the hydrophone signals to our laboratory in Toulon, France, yielding nearly 50 TB of synchronous multichannel recordings. In previous work, we trained a Convolutional Neural Network (CNN) to detect Orca vocalizations, using transfer learning from a bird activity dataset. Here, for each detected vocalization, we estimate the pitch contour (fundamental frequency). Finally, we cluster vocalizations by features describing the pitch contour. While preliminary, our results demonstrate a possible route towards automatic Orca call type classification. Furthermore, they can be linked to the presence of particular Orca pods in the area according to the classification of their call types. A large-scale call type classification would allow new insights on phonotactics and ethoacoustics of endangered Orca populations in the face of increasing anthropic pressure.

1 INTRODUCTION

The Orca (*Orcinus orca)* is a top-predator of the marine food chain (Jefferson et al., 1991). The Northern Resident Orcas community is composed of several "pods" composed of matrilines (Bigg et al., 1990). This cetacean can produce 3 different types of signals: clicks, whistles and pulsed calls (Ford, 1989). This study focuses only on vocalizations (pulsed calls). Some biological studies describe the communication of Orcas (Ford et al., 1987; Tyson et al., 2007; Weiß et al., 2007; Filatova et al., 2012), based on manual methods. Related work by Deecke et al. (1999) compared dialects of Orcas using artificial neural networks and showed that acoustic similarities are significantly correlated with the group association patterns. In order to analyze the animal's communication in different spacial and temporal contexts, automated analysis for captured sound is crucial. For that purpose, the field of bioacoustics offers numerous approaches using neural networks and deep learning (Glotin et al., 2013). The latter methods were explored to automatically detect orca calls emitted throughout 3 years of continuous recordings from 2015 to 2017 (Poupard et al., 2019a). In this study we build on these detections, and compute each vocalisation's pitch over time. This pitch analysis serves to differentiate vocalisations. In particular, we extract pitch features and cluster the vocalizations, partly recovering different call types as annotated by human experts.

2 MATERIAL

For 20 years, the NGO Orcalab developed and has maintained a unique multi-hydrophone recording station around Hanson Island (Northern Vancouver Island, Canada) to study Orcas. This acoustic station is composed of 5 hydrophones and extends over 50 km^2 of ocean (Fig. 1). In 2015, we have set up a continuous recording of all the hydrophones of this station (Fig. 1). The aim is to allow observation and modelling of bioacoustic activities for various species, at large spacial and temporal scales, including all details of their ecoacoustic niche, under various geophysical and anthropophonic conditions, more particularly in order to build new knowledge about Orcas.

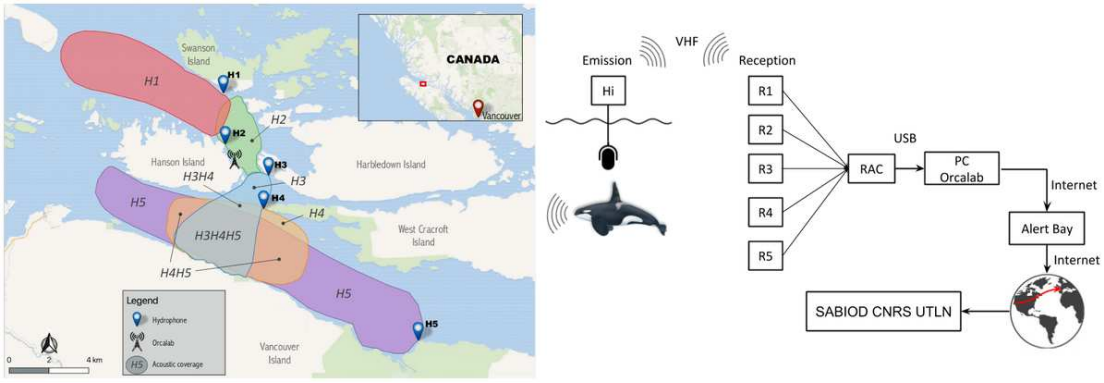

Figure 1. Left: Map of the area and the listening range of the 5 hydrophones H1 to H5. Detection zones indicate which hydrophones can record Orca calls in a given area, w.r.t. thirty years of observations by Orcalab. Right: Representation of the data acquisition, from recording until storage at Toulon.

3 DATA ANALYSIS

3.1 Vocalization Detector

We first designed an automatic acoustic event extractor (presented in (Poupard et al., 2019c)). Its output helped us build a dataset composed of 872 Orca vocalization samples and 6801 noise samples (boats, rain, void…), which we split randomly with 20% for the test set, 60% for the training set and 20% for the validation set.[1] With that in hand, we trained a CNN (originally designed for a bird detection task (Grill and Schlüter, 2017) to distinguish Orca vocalizations (not clicks) from boats and background noise (Poupard et al., 2019a). The model was trained with weakly annoted data (one label per file), originally using global max pooling to aggregate local predictions for comparison against the global label. After training, the global pooling was removed to obtain local probabilities for pitch and vocalization analysis. Max pooling lead to spiky local predictions (high precision, but low recall), which were unsuitable for our purposes. We found that training the model with global mean pooling instead gave much higher recall, covering the full length of each vocalization without sacrificing precision. The resulting Area Under the receiver operating characteristic Curve (AUC) of this detector is 89% (Poupard et al., 2019a).

Running this model on the summers of 2015, 2016 and 2017 results in 421,879 detected vocalizations across all five hydrophones.

3.2 Pitch Analysis

In order to describe call characteristics, the pitch (fundamental frequency) is often used (Berthommier and Glotin, 1999). The pitch is a property that describes the fundamental frequency of the speech wave (Houtsma, 1997; Babacan et al., 2013). Like humans, Orcas produce vocalizations that have several harmonic frequencies, combining into a multi-layered wave (Ford, 1989). Foote and Nystuen (2008) used pitch to differentiate different ecotypes of killer whales and Shapiro and Wang (2009) developed their own pitch tracker algorithm (PDA) based on human voice.

In this study, the Parselmouth Python library (Jadoul et al., 2018) was chosen as pitch estimator. It relies on the autocorrelation (AC) (Boersma, 1993; Berthommier and Glotin, 1999). It is illustrated

[1] A random split may sample train and test segments from nearby locations, giving an overly optimistic test error. We did not have enough annotated data for a chronological split avoiding this.

on a recording of Orca calls in Fig. 2. Computing all the pitches for one day of vocalizations on the 5 hydrophones took half an hour in average on GPU.

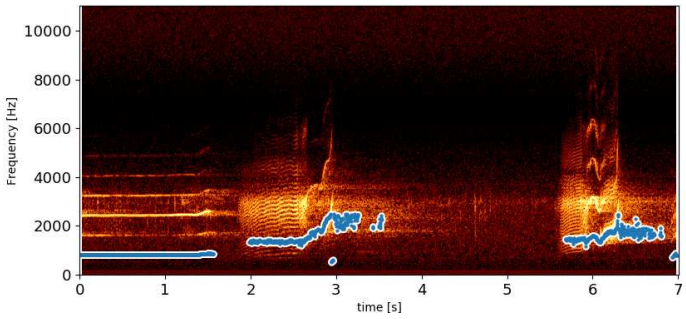

Figure 2. Example of a pitch extraction (pitch floor=300, pitch ceiling=2500, voicing threshold=0.2).

The AC only outputs a pitch point if it has a certain confidence in it (using a threshold on the strength of the unvoiced candidate relative to the maximum possible AC). Thus, with some detected vocalizations, fewer points were output. This property allowed us to filter out false positives and too low Signal to Noise Ratio vocalization detections. Keeping only vocalizations with more than 200 points filtered out 284,791 noisy vocalizations and false detections.

We thus extracted the pitch of 137,088 vocalizations.

3.3 Unsupervised Clustering

Unsupervised clustering is often the solution to solve classification tasks for unannotated data. Our intuition was that the Orcas' call types (Ford et al., 1987; Root-Gutteridge et al., 2014) could be automatically clustered by similarity in their pitch shape. A first step was thus to define the input to our unsupervised clustering algorithm. Thus features of the previously extracted pitch were selected to best describe the shape of the vocalizations with a minimum dimensionality.

The following features were chosen (Ford, 1984): argminFreq, argmaxFreq, minVel, maxVel, meanVel, startVel, endVel, minAccel, maxAccel, argminAccel, argmaxAccel, deltaFreq. Here argmin/max stand for the position in time of the maximum/minimum relative to the total duration. Min/max stand for minimum/maximum values. Mean stands for the average value. Start/end stand for the mean of the first/last 5% of the call. Delta stands for the minimum value substracted from the maximum value. Freq stands for frequency values (the estimated pitch), Vel stands for velocity (the derivative of the pitch), and Accel stands for acceleration (the derivative of the velocity).

Having extracted those features, they were used as an input for the HDBSCAN algorithm (McInnes et al., 2017), which is a hierarchical implementation of the Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). Several minimum cluster sizes and minimum sample sizes were explored, to optimize the number of output clusters, and the strictness of the clustering. Eventually 30 and 3 were chosen for the latter parameters respectively.

4 RESULT

The clustering algorithm hardly works when applied to all the collected vocalisations together (coming from the 5 different hydrophones with different depth and sensitivity), whereas it works decently when applied to each hydrophone separately. Here we present the results for the H1 hydrophone (see map in Fig. 1), which represents 6796 vocalizations. Further work will focus on generalizing the clustering method to any hydrophone after some normalization.

The HDBSCAN found 13 clusters (0 to 12; Fig. 4). The '-1' cluster is the algorithm's output of classifying what it considers as noise. To measure the clustering's relevance, 2 trained persons annotated 50 samples (picked randomly) from each cluster, according to the Oca call types as defined in the literature (Ford et al., 1987). We selected a subset of 6 call types (N1,N2,N4,N7,N9,N47): the ones most commonly found in our dataset (Fig. 3).

The distributions of call types among clusters (Fig. 4) show that our model was able in some clusters to isolate some type (N4 in clusters 0, 1, 4, and 5), to group calls with roughly similar upward types (N2,

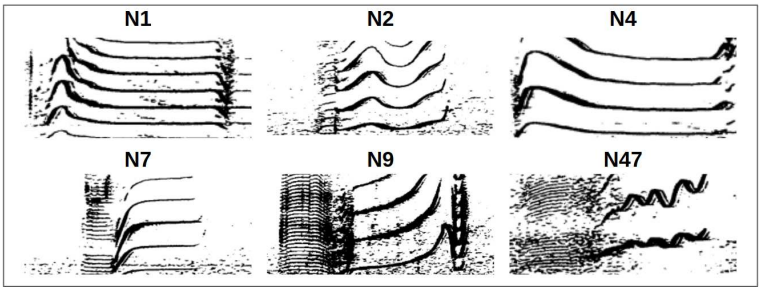

Figure 3. Selected subset of call types as defined in the literature (Ford et al., 1987).

N7, N9, N47 in clusters 2 and 3), and to classify boat noise (clusters 9 and 12). Those results demonstrate a promising approach to classifying orca vocalizations, in approximately 20 days of computation for 3 years of pentaphonic continuous recording.

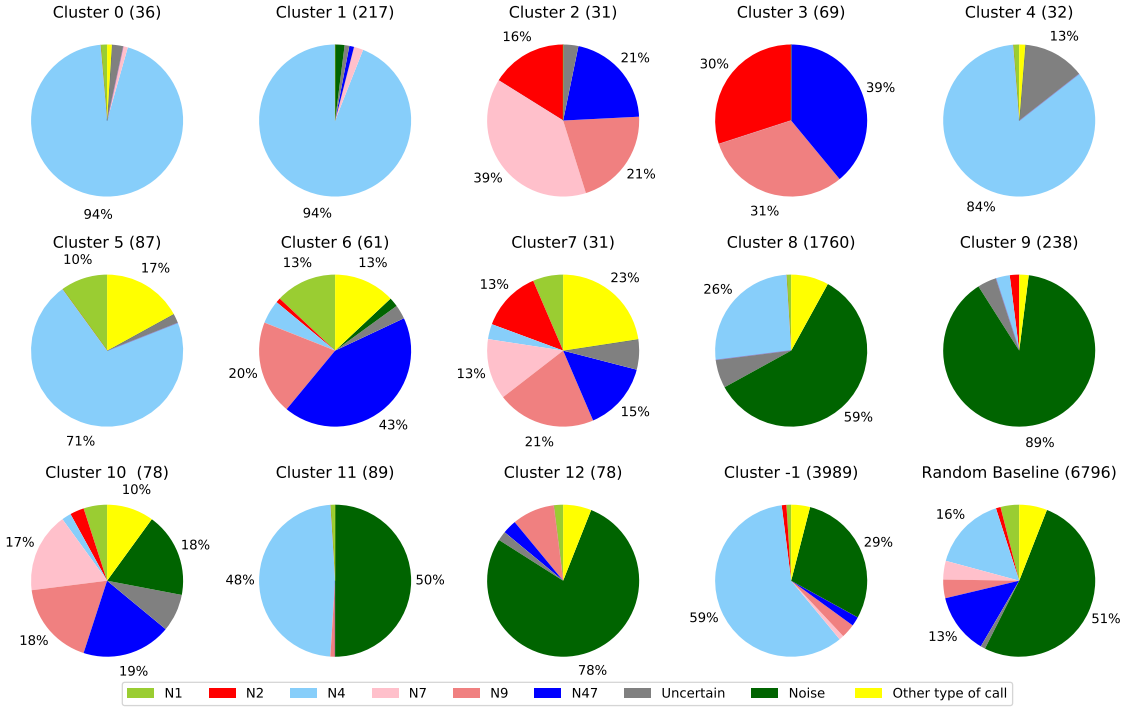

Figure 4. Distribution of call types among clusters found by the HDBSCAN algorithm. The numbers next to the cluster name show the amount of cluterized samples. The % are from the annotated subset.

5 CONCLUSION AND DISCUSSION

Our primary results can be linked with the presence of particular pods in the area. In fact, British Columbia (BC) is composed of different "acoustic clans". An acoustic clan is a group of Orcas that share particular types of calls known as discrete calls (Ford and Fisher, 1982). In the Northern BC, there are 4 main acoustic clans: the J, R, G and A Clan. For now, we will focus on the A Clan (Fig. 5), composed of several pods, themselves composed of groups of lineages called matrilines.

As shown in Fig. 5, there are 7 different pods in the A clan, having different call types (Ford, 1984). For example the A4 pods can produce N2, N4, N7 and N9 call types. Some types of calls are shared among multiple pods within the clan. For example, the N7 call extends to all pods, however each pod produces an unique N7 call. By recognizing pods and recording the different calls, we can establish a link between the pods and our clusters. In fact clusters 0, 1, 4, 5, and -1 have a high proportion of the N4 call

type (Fig. 4), we can thus expect that the A1, A4, and A5 pod vocalizations are present in these 5 clusters. The N47 call type is produced only by the A1 pods and this type is very present in 2 clusters (3 and 6), so we can state the hypothesis that these 2 clusters correspond to the A1 pod.

With such reasoning, these clusters represent a first approach to acoustically classify pods in BC, and in the future, matrilines (Weiß et al., 2006) and individual vocal signatures (Weiß et al., 2007).

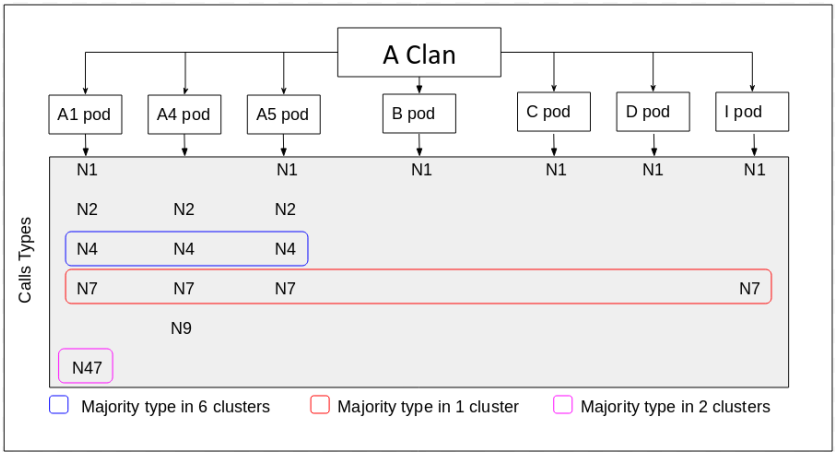

Figure 5. Selection of the 6 Call types produced by pods of the A clan inspired from (Ford, 1984).

Future work will improve the model at each of the 3 main steps: the learned vocalization detection, the pitch estimation that could be trained specifically to detect Orcas' pitch (Kim et al., 2018), and the unsupervised clustering of calls. An obvious improvement would consist in annotating more data for training. Parameter optimization is another possible enhancement, especially for the pitch estimation and the unsupervised clustering. For this purpose, relevant objective functions and accurate metrics need to be found. One could consider a global objective cost function to maximise the normalised mutual information of the bivariate distribution (Type, Cluster).

Once such an improved system is at hands, having a fully autonomous reliable Orca type call detector and classifier will open doors to many studies on Orca's communication and phonotactic regularities and divergence like in Malige et al. (2019). It would also allow behavioural studies (ethoacoustics), within various environments, including increasing anthropophony or whale watching pressure like in Poupard et al. (2019b).

ACKNOWLEDGMENTS

We thank first the Orcalab direction and collaborators for their incredible inspired work. We thank Biosong SA for the PhD funding of M. Poupard. This research is partly funded by FUI 22 Abyssound, ANR-18-CE40-0014 SMILES, ANR-17-MRS5-0023 NanoSpike and MARITTIMO EUR GIAS projects on advanced studies on cetaceans. We thank MI CNRS MASTODONS SABIOD.org and EADM MADICS CNRS scaled bioacoustic research groups, IUF for Glotin's chair 2011-16 during which he installed the remote recording at Orcalab to UTLN DYNI, and SEAMED PACA project and CNRS platform support for J. Schlüter's Post doc grant.

REFERENCES

Babacan, O., Drugman, T., d'Alessandro, N., Henrich, N., and Dutoit, T. (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *IEEE ICASSP*, pages 7815–7819.

Berthommier, F. and Glotin, H. (1999). A measure of speech and pitch reliability from voicing. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 61–70.

Bigg, M., Olesiuk, P., Ellis, G., Ford, J., and Balcomb, K. (1990). Social organization and genealogy of resident killer whales (orcinus orca) in the coastal waters of british columbia and washington state. *Report of the International Whaling Commission*, 12:383–405.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110.

Deecke, V., Ford, J., and Spong, P. (1999). Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale dialects. *J. ASA*, 105(4):2499–2507.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, KDD, pages 226–231.

Filatova, O., Deecke, V., Ford, J. K., Matkin, C., Barrett-Lennard, L., Guzeev, M., Burdin, A., and Hoyt, E. (2012). Call diversity in the north pacific killer whale populations: implications for dialect evolution and population history. *Animal Behaviour*, 83(3):595–603.

Foote, A. D. and Nystuen, J. A. (2008). Variation in call pitch among killer whale ecotypes. *The Journal of the Acoustical Society of America*, 123(3):1747–1752.

Ford, J. (1989). Acoustic behaviour of resident killer whales (orcinus orca) off vancouver island, british columbia. *Canadian Journal of Zoology*, 67(3):727–745.

Ford, J. et al. (1987). *A catalogue of underwater calls produced by killer whales (Orcinus orca) in British Columbia*. Department of Fisheries and Oceans, Fisheries Research Branch, Pacific ....

Ford, J. K. and Fisher, H. D. (1982). Killer whale (orcinus orca) dialects as an indicator of stocks in british columbia. *Rep. Int. Whal. Commn*, 32:671–679.

Ford, J. K. B. (1984). *Call traditions and dialects of killer whales (Orcinus orca) in British Columbia*. PhD thesis, University of British Columbia.

Glotin, H., LeCun, Y., Artières T. Mallat, S., Tchernichovski, O., and Halkias, X. (2013). Proc. nips4b : Neural information processing scaled for bioacoustics, from neurons to big data, joint to int. *Conference on Neural Information Processing Systems (NIPS)*. http://sabiod.org/nips4b.

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

Houtsma, A. J. (1997). Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, 26(2):104–115.

Jadoul, Y., Thompson, B., and De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Jefferson, T., Stacey, P., and Baird, R. (1991). A review of killer whale interactions with other marine mammals: Predation to co-existence. *Mammal review*, 21(4):151–180.

Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (2018). Crepe: A convolutional representation for pitch estimation. In *IEEE Int. conf. Acoustics, Speech Sig. Proc. (ICASSP)*, pages 161–165.

Malige, F., Djokic, D., Patris, J., Sousa-Lima, R., and Glotin, H. (2019). Use of recurrence plots to automatically extract songs in humpback whale recordings. *Submitted to Bioacoustics*.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Poupard, M., Best, P., Schlüter, J., Prevot, J., Spong, P., Symonds, H., and Glotin, H. (2019a). Deep learning for ethoacoustics of orcas on three years pentaphonic continuous recording at orcalab revealing tide, moon and diel effects. In *IEEE OCEANS*.

Poupard, M., DeMongolfier, B., and Glotin, H. (2019b). Ethoacoustic by bayesian non parametric and stochastic neighbor embedding to forecast anthropicpressure on dolphins. In *IEEE OCEANS*.

Poupard, M., Ferrari, M., Schluter, J., Astruch, P., Schohn, B., Rouanet, B., Goujard, A., Lyonnet, A., Giraudet, P., Barchasz, V., Gies, V., Best, P., Dominici, D., Lengagne, T., Soriano, T., and Glotin, H. (2019c). Passive acoustics to monitor flagship species near boat traffic in the UNESCO world heritage natural reserve of scandola. In *Input Academy : Conf. Innovation in Urban & Regional Planning*.

Root-Gutteridge, H., Bencsik, M., Chebli, M., Gentle, L. K., Terrell-Nield, C., Bourit, A., and Yarnell, R. W. (2014). Improving individual identification in captive eastern grey wolves (canis lupus lycaon) using the time course of howl amplitudes. *Bioacoustics*, 23(1):39–53.

Shapiro, A. D. and Wang, C. (2009). A versatile pitch tracking algorithm: From human speech to killer whale vocalizations. *The Journal of the Acoustical Society of America*, 126(1):451–459.

Tyson, R., Nowacek, D., and Miller, P. (2007). Nonlinear phenomena in the vocalizations of north atlantic right whales and killer whales. *J. Acoustical Soc. America*, 122(3):1365–1373.

Weiß, B. M., Ladich, F., Spong, P., and Symonds, H. (2006). Vocal behavior of resident killer whale matrilines with newborn calves: The role of family signatures. *J. ASA*, 119(1):627–635.

Weiß, B. M., Symonds, H., Spong, P., and Ladich, F. (2007). Intra-and intergroup vocal behavior in resident killer whales, orcinus orca. *J. Acoustical Soc. America*, 122(6):3710–3716.

Wave Propagation in the Biosonar Organ of sperm whales using Finite Difference Time Domain

Ferrari Maxence[1,2], Ricard Marxer[2], Mark Asch[1], and Hervé Glotin[2]

[1]**LAMFA, CNRS, U. Picardie**
[2]**Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France**

ABSTRACT

The sperm whale bio-sonar presents many specificities, such as its size, its loudness or its vocalization abilities. Furthermore it fulfills several roles in their foraging and social behaviour. However our knowledge about its operation remains limited to the main acoustic path that the emitted pulse may take. We still ignore the precise mechanisms that shape the wave and on which parts the sperm whale is able to act. In this paper, we describe a technique to simulate the sperm whale click generation from a physical perspective. Such an approach aims at unveiling the processes involved in their vocal production, as a stepping stone towards a better understanding of their interaction with peers and the environment.

INTRODUCTION

Sperm whales (*Physeter macrocephalus*, *Pm*) have the loudest bio-sonar in the animal kingdom (230 dB re: 1 μPa rms, Møhl et al. (2003)). The clicks produced by this sonar are not only used for their echolocation during dives, but also in their social interactions. During dives sperm whales emit trains of clicks, much like those of bats, whereas for socialization, they will emit small rhythmic groups of clicks. Since Norris and Harvey (1972) first theorized the way their sonar worked, it has been broadly accepted that *Pm* creates an initial pulse at the front of its head, in the *"museau de singe"* (aka. monkey lips), which will then bounce back and forth in its head. However, the details of such a mechanism and which parameters the sperm whale can act on, remain unknown.

Since the 90's (Aroyan et al. (1992)), scientists have been modeling the propagation of vocalized sound waves in marine mammals heads. The ability to model wave propagation in marine mammals allows to better understand the interaction between all the organs responsible for the sound creation, or the molding of the sound wave, to achieve the highly directive beam pattern of such species (Cranford et al. (2008), Wei et al. (2014)). To the best of our knowledge these type of simulations have not been performed on the bio-sonar of sperm whales.

Most of these simulations are based on anatomic data derived from computed tomography scans. This information enables the construction of the model geometry, and obtain the mechanical parameters for each material and their location (up to the CT scan resolution). However, most of the employed scans were performed on postmortem individuals. Cranford et al. (2014) compared data between dead and live specimens and their effects on the simulations. Dead specimen are prone to introducing artifacts in the model, such as air-filled blood vessels, but will not suffer from scanning errors due to the movement of a living specimen. However these deviations are likely not to change the mechanical parameters of the various tissues, and thus the Hounsfield unit that the CT-scan will measure, which has been shown (Soldevilla et al., 2005) to be correlated to the density and speed of sound.

In this work we describe a physical simulation of a *Pm* click using geometry and materials from dissection data and Finite Difference Time Domain (FDTD) for the wave propagation calculation.

BUILDING THE GEOMETRY

Unlike the other small marine mammals, sperm whales cannot be CT-scanned by normal means due to their size and weight. The only tomography data available have been performed on postmortem

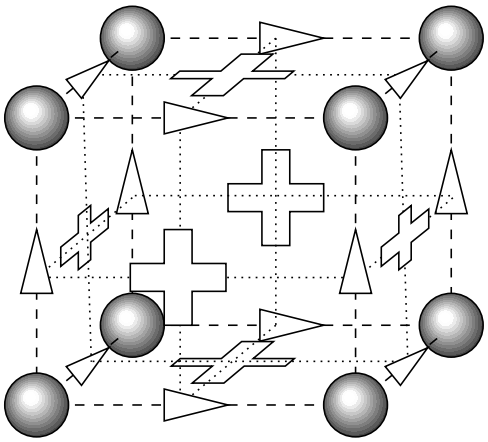

Figure 1. Part of the FDTD grid. Sphere : normal stress. Triangle : velocity. Cross : cross-stress

neonate sperm whales (Cranford (1999), Huggenberger et al. (2016)). However, those models cannot be simply scaled up since some anatomical elements do not match those of adult individuals obtained from dissections, such as the one shown in Clarke (1978). In order to shape our model we have used dissection data. We model each organ using Computer Assisted Design (CAD) software based on the slices from Clarke (1978). Since single blueprints did not match each other exactly, we had to scale some of them, or take the mean shape.

FINITE DIFFERENCE TIME DOMAIN

The method we used to simulate the sound propagation is the Finite Difference Time Domain (FDTD), but unlike Aroyan et al. (2000), we use the stress-velocity equations which allows to model shear waves that propagate through cross-stress.

As usual for these sort of simulations, the aim is to simulate the target body inside an infinite medium. The standard way of getting rid of reflections from the border of the simulation, and thus simulating a infinite medium while having a finite box, consists either of having multiple dampening layers near the border, or having special equations for the border that will make them 'invisible' to waves. All of those methods are always an approximation and will still produce some reflections in certain cases. We have used the Absorbing Boundary Condition (ABC) from Higdon (1986), with angles of 2.86° and 65°.

One iteration of the FDTD consists of the update of the speed grids, then the stress grids (including cross stress) and the ABC. The computation time of the boundary update is negligible compared to the stress and speed update (two orders of magnitude), and we could increase the number of angle of incidence with perfect absorption without any perceptible decrease in performance. However we consider this number of absorption angles to be enough.

EXPERIMENT

For our experiment we had to chose the mechanical parameters for each of the simulated media (skin, bones, *spermaceti*, water, etc.). While FDTD and our model are able to cope with anisotropic coefficients, for the sake of simplicity, in this first approach we have made an isotropic assumption. We have combined the measurements of Goold et al. (1996) (assuming a temperature of 30°C and atmospheric pressure), Clarke (1978) and the measurements done on the *Kogia breviceps* in Song et al. (2015). For the parameters not found in the literature, we have used values from the human body, based on the observation that the other parameter values are shared between the species (*Physeter macrocephalus, Kogia breviceps, Ziphius cavirostris, Homo sapiens sapiens*). The little variation introduced by the values borrowed from the other species will not have a significant impact on the results, since even a change of the order of 5% to 10% has little effect on the resultant beam (Cranford et al., 2008). The most important factor for the position of the various focal points is the geometry of the organs.

We simulated a sperm head in a $520 * 240 * 220$ cm^3 volume, with 1 cm resolution, and the materials were averaged following Toyoda et al. (2012). The simulation was implemented using PyTorch (a Deep

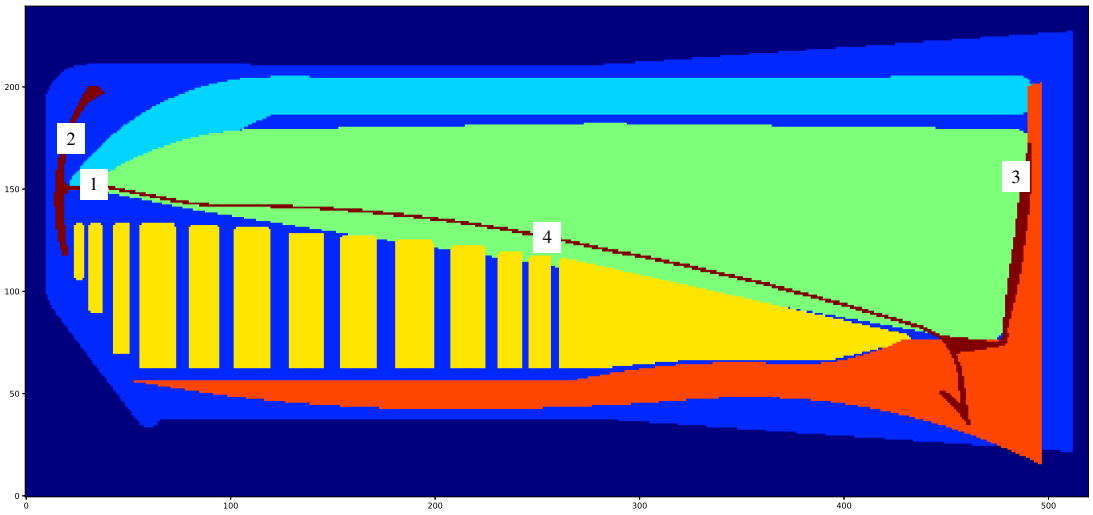

Figure 2. Material in the sagittal plane. Deep blue: water, blue: blubber and skin, cyan: muscle, yellow: junk, green: spermaceti, orange: bone, dark red: air. 1: museau de singe, 2: distal sac, 3: frontal sac, 4: right nasal passage.

Learning Python library) and run on an NVidia Titan X. The implementation performs at 4.6 iterations per second. Thus for a simulation of 20 ms with a time step of 1 μs, the computation time is 1 h 12. The simulation starts at rest. We then add to pressure points located next to the *museau de singe* in the spermaceti the difference of a 10 kHz sinusoidal wave during one period.

Figure 3 shows a recorded sound wave of a sperm whale click and the simulated pressure at the *museau de singe*. In both the recorded and simulated sounds we observe three pulses of a sperm whale click, in the simulated case these correspond to P0, P1 and P2. In the simulation we measure an offset of 6662 bins (or μs) between each of these pulses. These intervals are known as the inter pulse interval (IPI) and have often been used to estimate the total body length of the sperm whale (Clarke (1978), Gordon (1991), and Growcott et al. (2011)).

While the proposed model still fails to reproduce individual pulse wave shapes, such as those found in recorded vocalisations, it does produce a signal with a valid IPI. By using the three different methods cited above to estimate the body size from the IPI, we obtain sizes of 14.97 m, 14.47 m and 14.12 m respectively, which match the length of the actual sperm whale that the model is based on (14.2 m). This result mainly depends on three parameters: the bulk modulus, the density, and the length of the *spermaceti*. Yet, it is still a comforting proof that this part of the model is working.

In Figure 4, we can see the evolution of the simulation, with the sound wave propagating from the *museau de singe* to the frontal sac, then being reflected by it, and going back to the *museau de singe* to be reflected by the distal sac.

FUTURE WORK

The model presented here remains a rough approximation and requires further tuning to better reflect the real phenomena. The geometry of the right nasal passage needs to evolve, in its current form it acts as a perfect mirror and prevents the energy reflected from the frontal sac to reach the junk. The next stages of this research will focus on the fluid-filled knobs present in the frontal sac described by Norris and Harvey (1972). During dives, they might act as a filter, thus modifying the response of the sonar.

ACKNOWLEDGMENTS

We thank Direction Générale de l'Armement and Région Hauts de France for the PhD grant of M. Ferrari.

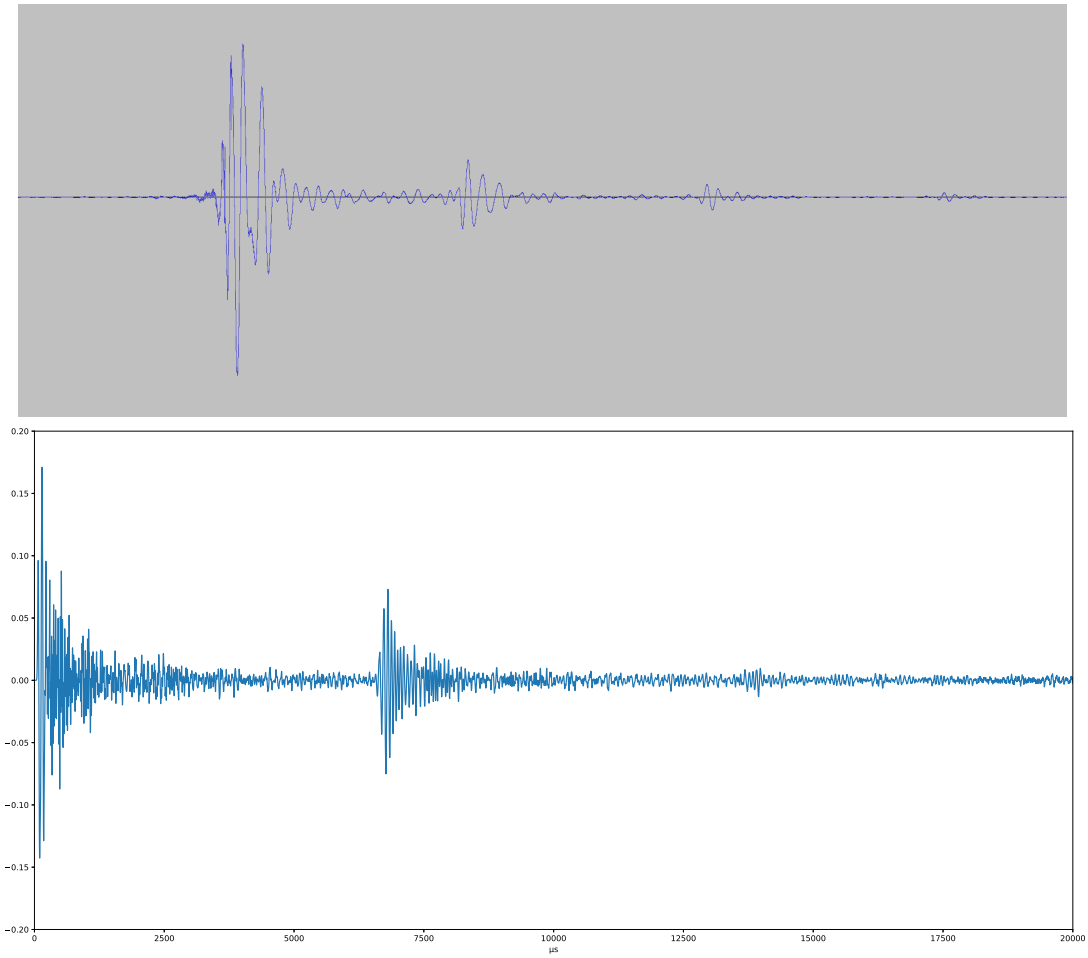

Figure 3. Top: Recording of sperm whale. Bottom: Simulated pressure at the excitation point.

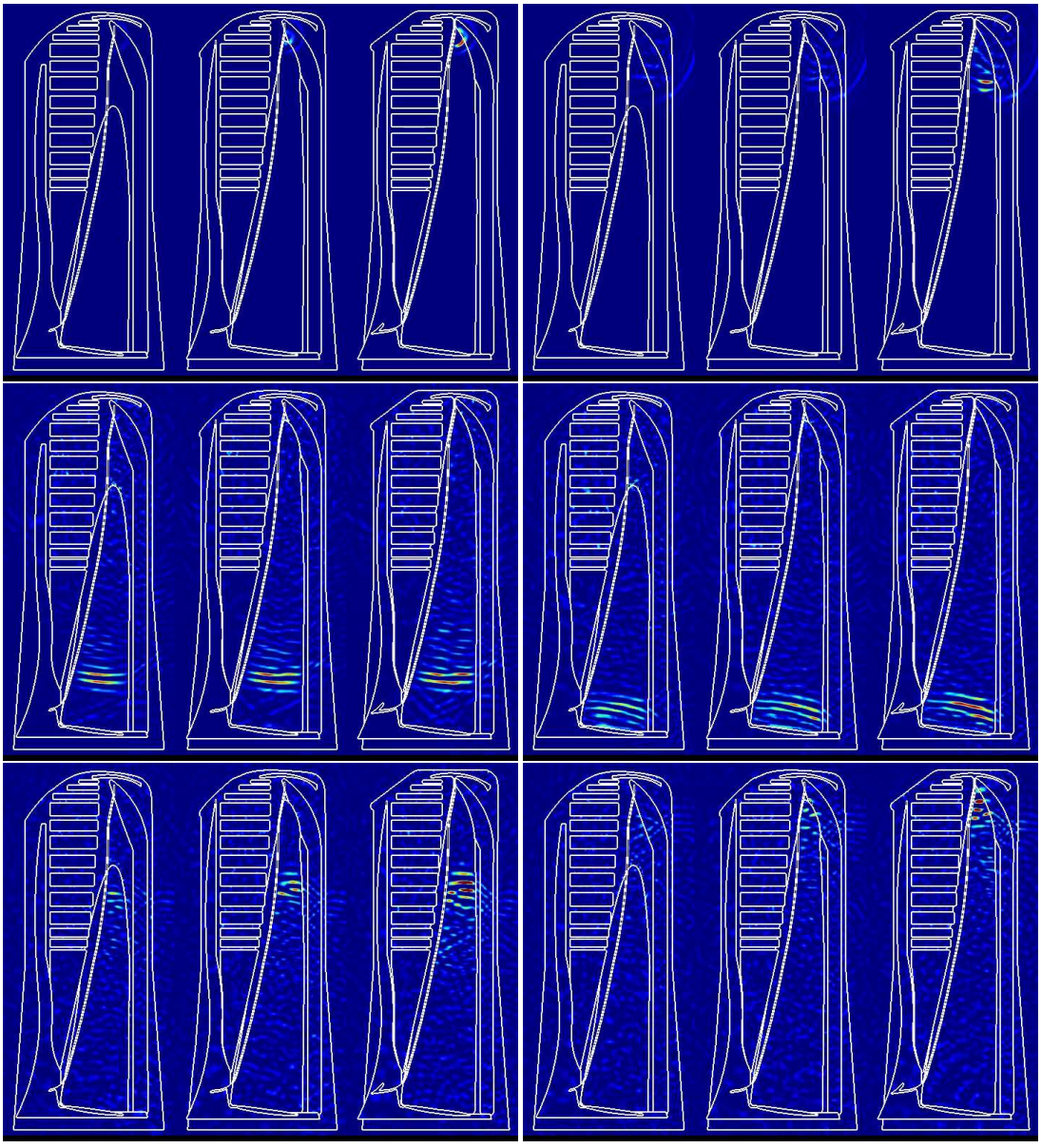

Figure 4. Multiple frames of the simulation, with the stress component (normalized) being plotted. Each picture is made of three slices of the 3D volume. The right one is the sagittal plane, the middle one is a plane 10 cm on the left of the sagittal plane, and the left one has an offset of 20 cm regarding the sagittal plane.

REFERENCES

Aroyan, J. L., Cranford, T. W., Kent, J., and Norris, K. S. (1992). Computer modeling of acoustic beam formation in d elphinusdelphis. *The Journal of the Acoustical Society of America*, 92(5):2539–2545.

Aroyan, J. L., McDonald, M. A., Webb, S. C., Hildebrand, J. A., Clark, D., Laitman, J. T., and Reidenberg, J. S. (2000). Acoustic models of sound production and propagation. In *Hearing by whales and dolphins*, pages 409–469. Springer.

Clarke, M. R. (1978). Structure and proportions of the spermaceti organ in the sperm whale. *Journal of the Marine Biological Association of the United Kingdom*, 58(1):1–17.

Cranford, T. W. (1999). The sperm whale's nose: Sexual selection on a grand scale? 1. *Marine mammal science*, 15(4):1133–1157.

Cranford, T. W., Krysl, P., and Hildebrand, J. A. (2008). Acoustic pathways revealed: Simulated sound transmission and reception in cuvier's beaked whale (ziphius cavirostris). *Bioinspiration & Biomimetics*, 3(1):016001.

Cranford, T. W., Trijoulet, V., Smith, C. R., and Krysl, P. (2014). Validation of a vibroacoustic finite element model using bottlenose dolphin simulations: the dolphin biosonar beam is focused in stages. *Bioacoustics*, 23(2):161–194.

Goold, J. C., Bennell, J. D., and Jones, S. E. (1996). Sound velocity measurements in spermaceti oil under the combined influences of temperature and pressure. *Deep Sea Research Part I: Oceanographic Research Papers*, 43(7):961–969.

Gordon, J. C. (1991). Evaluation of a method for determining the length of sperm whales (physeter catodon) from their vocalizations. *Journal of Zoology*, 224(2):301–314.

Growcott, A., Miller, B., Sirguey, P., Slooten, E., and Dawson, S. (2011). Measuring body length of male sperm whales from their clicks: the relationship between inter-pulse intervals and photogrammetrically measured lengths. *The Journal of the Acoustical Society of America*, 130(1):568–573.

Higdon, R. L. (1986). Absorbing boundary conditions for difference approximations to the multidimensional wave equation. *Mathematics of computation*, 47(176):437–459.

Huggenberger, S., Andre, M., and Oelschläger, H. H. (2016). The nose of the sperm whale: overviews of functional design, structural homologies and evolution. *Journal of the Marine Biological Association of the United Kingdom*, 96(4):783–806.

Møhl, B., Wahlberg, M., Madsen, P. T., Heerfordt, A., and Lund, A. (2003). The monopulsed nature of sperm whale clicks. *The Journal of the Acoustical Society of America*, 114(2):1143–1154.

Norris, K. S. and Harvey, G. W. (1972). A theory for the function of the spermaceti organ of the sperm whale (physeter catodon l.).

Soldevilla, M. S., McKenna, M. F., Wiggins, S. M., Shadwick, R. E., Cranford, T. W., and Hildebrand, J. A. (2005). Cuvier's beaked whale (ziphius cavirostris) head tissues: physical properties and ct imaging. *Journal of experimental biology*, 208(12):2319–2332.

Song, Z., Xu, X., Dong, J., Xing, L., Zhang, M., Liu, X., Zhang, Y., Li, S., and Berggren, P. (2015). Acoustic property reconstruction of a pygmy sperm whale (kogia breviceps) forehead based on computed tomography imaging. *The Journal of the Acoustical Society of America*, 138(5):3129–3137.

Toyoda, M., Takahashi, D., and Kawai, Y. (2012). Averaged material parameters and boundary conditions for the vibroacoustic finite-difference time-domain method with a nonuniform mesh. *Acoustical Science and Technology*, 33(4):273–276.

Wei, C., Zhang, Y., and Au, W. W. (2014). Simulation of ultrasound beam formation of baiji (lipotes vexillifer) with a finite element model. *The Journal of the Acoustical Society of America*, 136(1):423–429.

Vocal Interactivity in Crowds, Flocks and Swarms: Implications for Voice User Interfaces

Roger K. Moore

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

ABSTRACT

Recent years have seen an explosion in the availability of Voice User Interfaces. However, user surveys suggest that there are issues with respect to usability, and it has been hypothesised that contemporary voice-enabled systems are missing crucial behaviours relating to user engagement and vocal interactivity. However, it is well established that such *ostensive* behaviours are ubiquitous in the animal kingdom, and that vocalisation provides a means through which interaction may be coordinated and managed between individuals and within groups. Hence, this paper reports results from a study aimed at identifying generic mechanisms that might underpin coordinated collective vocal behaviour with a particular focus on closed-loop negative-feedback control as a powerful regulatory process. A computer-based real-time simulation of vocal interactivity is described which has provided a number of insights, including the enumeration of a number of key control variables that may be worthy of further investigation.

INTRODUCTION

Background

Recent years have seen an explosion in the availability of 'voice user interfaces' (VUIs), initially stimulated by the 2011 launch of *Siri* - Apple's smartphone-based voice assistant - followed in 2015 by Amazon's release of the first 'smart speaker' - *Alexa*. Since then, such smartphone and smart speaker based voice assistants have become almost ubiquitous. For example, *Siri* has had over 40 million monthly active users in the U.S. since July 2017, and smart speaker shipments reached 78 million units worldwide in 2018[1,2]. In the U.K., the number of people who own a smart speaker doubled from one-in-twenty to one-in-ten over a period of just six-months from autumn 2017 to spring 2018[3].

However, setting aside the impressive sales figures, a more critical aspect of such voice assistants is the extent to which they are actually used. For example, a survey conducted in 2015 (i.e. prior to the appearance of the first smart speaker) found that only 26% of the respondents used a voice assistant regularly and the majority of voice assistant users preferred typing to talking (Moore et al., 2016a). A more recent study by Kim (2019) investigating the usage of voice assistants on both smartphones and smart speakers found that over half of the smart speaker owners used their voice assistant several times a day. In contrast, only one-third of smartphone owners used their voice assistants on a daily basis, and half hardly used their voice assistants at all.

These studies also reveal that the majority of users employ quite stylised language, e.g. using simple voice commands to access music playlists, to perform searches using spoken queries, or to set alerts and reminders. Such shallow linguistic interaction is somewhat predictable given the nature of the problems users encounter with contemporary voice-enabled devices. For example, nearly half of the users surveyed reported difficulties with not being understood or simply not being able to do very much.

[1] https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

[2] https://www.canalys.com/newsroom/smart-speaker-market-booms-in-2018-driven-by-google-alibaba-and-xiaomi

[3] https://yougov.co.uk/topics/politics/articles-reports/2018/04/19/smart-speaker-ownership-doubles-six-months

Key challenges

The usage statistics reported above confirm that contemporary VUIs are still a long way from being able to provide the "*conversational interface*" often promoted in the marketing literature for such systems. Indeed, the fact that users are effectively resorting to 'voice button-pressing' suggests that there is a fundamental difference between the richness of everyday human-human spoken language and the simplicity of the voice-based interaction that takes place between humans and machines. It has been argued elsewhere that a 'mismatch' between interlocutors is not only an important obstacle that needs to be explored in a human-machine context (Moore, 2015), but that it may even be an unsurmountable problem (Moore, 2016). In particular, if spoken language interaction is viewed as being based on the co-evolution of two key traits – *ostensive-inferential* communication and *recursive mind-reading* (Scott-Phillips, 2015), then contemporary voice-based systems are essentially only dealing with one aspect – inference. Some of the high-level issues relating to recursive mind-reading have been addressed in Moore and Nicolao (2017), but low-level concerns relating to ostensive vocal behaviour remain an open question. Hence, there seems to be something essential missing from contemporary voice-enabled system in the area of user engagement and interaction – not just what to say, but when to say it and, in a group context, to whom (Moore, 2015).

Potential solutions

Of course, interactive ostensive behaviours are ubiquitous in the animal kingdom, and vocalisation provides a means through which such activities may be coordinated and managed between individuals and within groups (Moore et al., 2016b). Vocalisations are often carefully timed in relation to each other (and other events taking place in an environment), and this may take the form of *synchronised* ritualistic behaviour (such as rhythmic chanting, chorusing or singing) or *antisynchronous* turn-taking (which can be seen as a form of dialogue) (Cummins, 2014; Fusaroli et al., 2014; Ravignani et al., 2014).

Of particular interest here are the *mechanisms* that support the emergence of synchronised vocal interactivity in crowds, flocks and swarms, and the implications of those mechanisms for future voice user interfaces. Hence, this paper presents results from an investigation into such mechanisms using a real-time simulation (i.e. a computational model) of interactive vocal dynamics and synchrony.

COLLECTIVE BEHAVIOUR

The coordinated behaviour of large numbers of independent living organisms has been the subject of scientific enquiry for many years. For example, studies have been conducted into the flocking of birds (Reynolds, 1987), the synchronised flashing of fireflies (Ermentrout, 1991), the dynamics of human crowd movement (Still, 2000), waves of coordinated clapping by audiences (Néda et al., 2000), and spatial sorting in shoals of fish (Couzin et al., 2002). Of particular interest are the transitions from one type of collective behaviour to another, especially in the context of attraction and repulsion between individuals (Katz et al., 2011), and predator-prey interactions (Handegard et al., 2012). Much of the research has involved computational simulation (perhaps the most famous being '*Boids*'[4]), as well as physical implementations in the field of swarm robotics, e.g. Bo et al. (2005).

One important aspect of synchronous behaviours is that they involve parallel coupled simultaneous action, as opposed to sequential action-reaction (Cummins, 2011). Such collective behaviours can thus be viewed as rhythmic entrainment, and thereby constitute a form of accommodation between individuals in a population (De Looze et al., 2014). It has also been posited that such behaviours underpin the links between different modalities, such as between vocalisation and physical movement (Cummins, 2009).

Vocal Synchrony

Whilst there have been a number of studies of vocal synchrony in animals, e.g. male zebra finches (Benichov et al., 2016) and monkeys (Takahashi et al., 2013), what is important here is the synergy with human vocal behaviours. Much of this work has involved 'joint speech', i.e. people speaking in unison (Cummins, 2014), and a more sophisticated view of 'turn-taking' in human dialogue (Heldner and Edlund, 2010). Of particular relevance is evidence that verbal synchrony in large groups of people produces affiliation (von Zimmermann and Richardson, 2016), and that some conversational partners tend to converge their vocal behaviours (Edlund et al., 2009), while others do not (Assaneo et al., 2019).

[4] https://www.red3d.com/cwr/boids/

Mechanisms

With regard to the mechanisms underpinning coordinated collective behaviour, by far the most popular approach is based on *coupled oscillators* (Kuramoto, 1975; Strogatz and Stewart, 1993; Strogatz, 2012), particularly through 'pulsatile coupling' (Mirollo and Strogatz, 1990). Not only has this been a very productive modelling paradigm with real-world implications (such as the simulation of clustered synchrony in electricity distribution networks (Pecora et al., 2014)), but new results are continuing to emerge (Matheny et al., 2019). The coupled-oscillator paradigm is also attractive because of its potential compatibility with known neural mechanisms (Ermentrout, 1991; Matell and Meck, 2000). However, it is only one way of formulating a complex non-linear attractor space, and it overlooks a number of potentially important conditioning variables – e.g. *energetics* (Moore, 2012).

As a consequence, the work reported here departs from the standard coupled-oscillator approach. In particular, attention is given to an alternative paradigm for creating a space of behavioural attractors – 'closed-loop negative-feedback control' – a powerful *regulatory* mechanism with roots in 'cybernetics' (Wiener, 1965) and commonly deployed for stabilising engineering systems (DiStefano III et al., 1990) as well as providing a powerful *non-behaviourist* paradigm for modelling the behaviour of living systems (Powers, 1973). The main differences between this approach and coupled oscillators is that the convergence criteria can be made more explicit, thereby offering the potential to gain a deeper understanding of the implications of particular parameters/settings on the emergent collective behaviours. It also offers the advantage that it can, in principle, be generalised to the synchronisation of more complex metrical structures, e.g. as discussed by Fitch (2013).

SIMULATION FRAMEWORK

Basic principles

The basic operation of classic closed-loop negative-feedback control is as follows: (i) a reference signal specifies the *desired* consequences of a system's actions, (ii) the *actual* consequences are sensed/interpreted by the system and compared with the reference target, (iii) the resulting *error* generates a control signal that drives the system in a direction such that the error is minimised. The process continues around the loop causing the system to not only converge to the desired behaviour but, more significantly, to maintain that behaviour in the face of arbitrary disturbances *without* having to sense such disturbances directly.

The tracking behaviour of such a negative-feedback control system is a function of the 'loop gain' of the controller. If the loop gain is too low, then stabilisation may take a long time – an 'overdamped' system. On the other hand, if the loop gain is too high, then the system may overshoot and even oscillate – an 'underdamped' system. The point here is that the loop gain effectively corresponds to the degree of *effort* (energy) applied to a regulatory task, i.e. from a psychological standpoint, it is analogous to *motivation*. An agent that 'cares' about controlling a particular variable would have a high loop gain, whereas a loop gain of zero implies the agent doesn't care at all (i.e. it gives up control). These are important individual differences that are not explicit in the coupled-oscillator approach.

Implementation

The simulations described herein have been implemented in Pure Data[5] – known as "*Pd*" – an open-source object-oriented dataflow programming language that is designed for real-time audio processing (Farnell, 2008). An environment has been constructed in which any number of vocalising (and listening) 'agents' may be connected to each other in arbitrary network topologies. Each agent comprises two feedback-control loops: one to regulate the interaction with other agents and another to regulate the agent's own behaviour. The first of these control loops aims to maintain synchrony between an agent's own vocalisations and those from agents that it can 'hear' (i.e. those to which it is connected). The second control loop attempts to maintain the agent's own preferred vocal rhythm. This arrangement means that each agent has two control parameters that influence the priority given to 'self' versus 'other'.

In addition, each agent has settings for the amplitude and duration of its vocalisations, their phase relation with the rhythmic 'beat', and the agent's preferred rhythm. In principle, these parameters could also be the subject of optimisation using feedback-control, but this was not implemented in the experiments reported here.

[5] `http://puredata.info/`

The sound output from each agent was produced using real-time synthesis of human, bird or insect vocalisations (as selected by the user). Other vocal characteristics for each agent (e.g. pitch frequency) were initialised randomly in order to provide a moderate level of 'individuality'.

Figure 1 illustrates a particular configuration of the simulation environment.

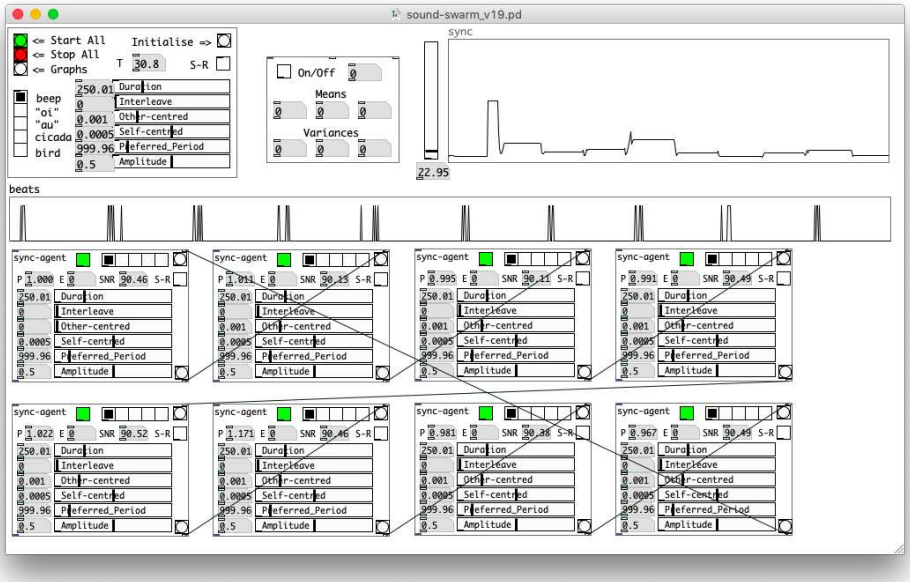

Figure 1. *Screenshot of the user interface for the* Pd*-based vocal simulation environment. The main control panel is shown in the top-left corner; buttons and sliders allow a user to specify global parameters for the population of agents, such as the type of vocalisation (human, insect, bird), duration, loudness etc. The lower half of the user interface facilitates the creation of an arbitrary number of agents, and the specification of which agent is listening to which other agent(s). In the example shown, eight agents have been configured in a loop topology (agent-2 is listening to agent-1, agent-3 to agent-2, agent-4 to agent-3, . . . , agent-1 to agent-8). As can be seen, sliders on each agent allow a user to set parameters individually if required. The graph shown at the top-right of the interface displays a timeline of the overall vocal synchrony between the agents, and the graph shown across the middle displays the individual rhythmic 'beats' from each agent (bunching indicates a degree of synchrony).*

Experiments

A range of experiments has been conducted based on varying numbers of interacting agents, different interconnection topologies, and alternative parameter settings. There is not space here to report all the findings. So what follows is a selected highlight.

One of the overarching research questions is concerned with the relationship between the topological connections between agents (i.e. the 'ostensive' relationships) and the emergent collective behaviour of the agents. In this context, one particular configuration is a *chain* with a 'master' (pacemaker) agent and a sequence of 'slave' agents. Figure 2 illustrates the outcome of simulating such a configuration with a chain of eight agents. As can be seen, on average, all of the agents in the feedback-control configuration maintained synchrony, but the agents further down the chain exhibited less stable rhythms. In contrast, agents in an action-reaction configuration maintained the rhythm, but the agents further down the chain were increasingly out of sync.

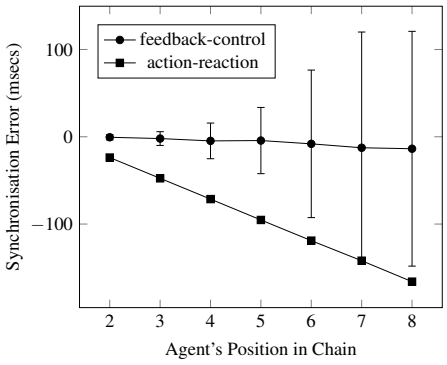

Figure 2. *Relationship between a 'slave' agent's position in a chain and its vocal synchronisation error with respect to the 'master' agent at the head of the chain.*

DISCUSSION & CONCLUSION

As a result of this research, it is possible to draw some conclusions about the control variables that are worthy of investigation with respect to vocal interactivity. These are summarised in Table 1.

Table 1. *Dimensions of vocal interactivity identified in this study. The left-hand column specifies the relevant control variables, and the right-hand column gives an indication of the expected range of values.*

VOCAL AGENTS	
Individuality (e.g. style of vocalisation)	average ⇔ extreme
Ostention (i.e. stance towards other agents)	connected ⇔ disconnected
Intentionality (i.e. goals wrt other agents)	same ⇔ different
Motivation/effort expended on pursuing others' goals	high ⇔ low
Motivation/effort expended on pursuing own goals	high ⇔ low
VOCALISATIONS	
Intensity (e.g. volume)	low ⇔ high
Clarity (e.g. intelligibility/SNR)	low ⇔ high
Period (i.e. timing)	short ⇔ long
Mark-to-space ratio (i.e. duration)	0% ⇔ 100%
Sentiment (i.e. valence)	-ve ⇔ +ve
Meaning (e.g. category)	named-entity-1 ⇔ named-entity-2
VOCAL INTERACTVITY	
Synchrony (i.e. engagement)	low ⇔ high
Simultaneity (i.e. overlap/interleaving)	0% ⇔ 100%
Dependency (i.e. between vocalisations)	dependent ⇔ independent

In conclusion, this paper has outlined some of the key issues facing contemporary voice-user interfaces, with a special focus on emergent properties of collective vocal behaviour, especially ostensive interaction and timing. The focus has been on closed-loop negative-feedback control as a regulatory mechanism, which implements a coincidence detection scheme that is compatible with known neural mechanisms (Matell and Meck, 2000). The simulation of real-time interacting vocal agents has already provided a number of insights into such behaviour, and more are expected as the full parameter space is investigated. In particular, it should be possible to show (i) how dialogue emerges as a compensatory response to the automatic regulation of intelligibility, not as a trivial action-reaction behaviour (Benichov et al., 2016), (ii) how cooperative *vs.* competitive interaction conditions vocalisations, and (iii) how communicative behaviour emerges from vocal interaction (Rosenthal et al., 2015).

REFERENCES

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., and Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience*, 22:627–632.

Benichov, J. I., Benezra, S. E., Vallentin, D., Globerson, E., Long, M. A., and Tchernichovski, O. (2016). The forebrain song system mediates predictive call timing in female and male Zebra finches. *Current Biology*, 26(3):309–18.

Bo, L., Tian-Guang, C., Long, W., and Zhan-Feng, W. (2005). Swarm dynamics of a group of mobile autonomous agents. *Chinese Physics Letters*, 22(1).

Couzin, I., Krause, J., James, R., Ruxton, G., and Franks, N. (2002). Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 218:1–11.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica*, 66(1-2):15–28.

Cummins, F. (2011). Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience*, 5(170):1–9.

Cummins, F. (2014). Voice, (inter-)subjectivity, and real time recurrent interaction. *Frontiers in Psychology*, 5:760.

De Looze, C., Scherer, S., Vaughan, B., and Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

DiStefano III, J. J., Stubberud, A. R., and Williams, I. J. (1990). *Feedback and Control Systems*. McGraw-Hill, New York, 2nd edition.

Edlund, J., Heldner, M., and Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *INTERSPEECH*, Brighton, UK.

Ermentrout, B. (1991). An adaptive model for synchrony in the firefly Pteroptyx malaccae. *Journal of Mathematical Biology*, 29(6):571–585.

Farnell, A. (2008). *Designing Sound*. Applied Scientific Press Limited, London.

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. *Frontiers in systems neuroscience*, 7:68.

Fusaroli, R., Rączaszek-Leonardi, J., and Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157.

Handegard, N. O., Boswell, K. M., Ioannou, C. C., Leblanc, S. P., Tjøstheim, D. B., Couzin, I. D., Walczak, A., Parisi, G., Procaccini, A., Viale, M., and Al., E. (2012). The dynamics of coordinated group hunting and collective information transfer among schooling prey. *Current Biology*, 22(13):1213–7.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Katz, Y., Tunstrøm, K., Ioannou, C. C., Huepe, C., and Couzin, I. D. (2011). Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):18720–5.

Kim, Y. (2019). *Usage of Speech Technology Systems*. Final-year project dissertation, Dept. Computer Science, University of Sheffield.

Kuramoto, Y. (1975). Self-entrainment of a population of coupled non-linear oscillators. In Araki, H., editor, *International Symposium on Mathematical Problems in Theoretical Physics*, pages 420–422.

Matell, M. S. and Meck, W. H. (2000). Neuropsychological mechanisms of interval timing behavior. *Bioessays*, 22(1):94–103.

Matheny, M. H., Emenheiser, J., Fon, W., Chapman, A., Salova, A., Rohden, M., Li, J., de Badyn, M. H., Pósfai, M., Duenas-Osorio, L., Mesbahi, M., Crutchfield, J. P., Cross, M. C., D'Souza, R. M., and Roukes, M. L. (2019). Exotic states in a simple network of nanoelectromechanical oscillators. *Science*, 363(1057).

Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662.

Moore, R. K. (2012). Finding rhythm in speech: a response to Cummins. *Empirical Musicology Review*, 7(1-2):36–44.

Moore, R. K. (2015). From talking and listening robots to intelligent communicative machines. In Markowitz, J., editor, *Robots That Talk and Listen*, chapter 12, pages 317–335. De Gruyter, Boston, MA.

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

Moore, R. K., Li, H., and Liao, S.-H. (2016a). Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH*, pages 3007–3011, San Francisco, CA. ISCA.

Moore, R. K., Marxer, R., and Thill, S. (2016b). Vocal interactivity in-and-between humans, animals and robots. *Frontiers in Robotics and AI*, 3(61).

Moore, R. K. and Nicolao, M. (2017). Towards a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention. *Frontiers in Robotics and AI*, 4(66).

Néda, Z., Ravasz, E., Brechet, Y., Vicsek, T., and Barabási, A.-L. (2000). Self-organizing processes: The sound of many hands clapping. *Nature*, 403:849–850.

Pecora, L. M., Sorrentino, F., Hagerstrom, A. M., Murphy, T. E., and Roy, R. (2014). Cluster synchronization and isolated desynchronization in complex networks with symmetries. *Nature communications*, 5(4079).

Powers, W. T. (1973). *Behavior: The Control of Perception*. Hawthorne, NY: Aldine.

Ravignani, A., Bowling, D. L., and Fitch, W. T. (2014). Chorusing, synchrony, and the evolutionary functions of rhythm. *Frontiers in psychology*, 5:1118.

Reynolds, C. (1987). Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87: Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 25–34, Anaheim, USA.

Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S., and Couzin, I. D. (2015). Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15):4690–4695.

Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

Still, G. K. (2000). *Crowd Dynamics*. Phd, University of Warwick.

Strogatz, S. H. (2012). *Sync: How Order Emerges from Chaos In the Universe, Nature, and Daily Life*. Hatchette Book Group.

Strogatz, S. H. and Stewart, I. (1993). Coupled oscillators and biological synchronization. *Scientific American*, 269(6):68–75.

Takahashi, D. Y., Narayanan, D. Z., and Ghazanfar, A. A. (2013). Coupled oscillator dynamics of vocal turn-taking in monkeys. *Current biology : CB*, 23(21):2162–8.

von Zimmermann, J. and Richardson, D. C. (2016). Verbal synchrony and action dynamics in large groups. *Frontiers in Psychology*, 7:2034.

Wiener, N. (1965). *Cybernetics: or Control and Communication in the Animal and the Machine*. The MIT Press, Cambridge, Mass., 2nd edition.

Index of Authors

VIHAR 2019

http://vihar-2019.vihar.org/

9 782956 202912