Report from Dagstuhl Seminar 16442

Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)

Edited by

Roger K. Moore[1], Serge Thill[2], and Ricard Marxer[3]

1 **University of Sheffield, GB,** r.k.moore@sheffield.ac.uk
2 **University of Skövde, SE,** serge.thill@his.se
3 **University of Sheffield, GB,** r.marxer@sheffield.ac.uk

Abstract

This seminar was held in late 2016 and brought together, for the first time, researchers studying vocal interaction in a variety of different domains covering communications between all possible combinations of humans, animals, and robots. While each of these sub-domains has extensive histories of research progress, there is much potential for cross-fertilisation that currently remains underexplored. This seminar aimed at bridging this gap. In this report, we present the nascent research field of VIHAR and the major outputs from our seminar in the form of prioritised open research questions, abstracts from stimulus talks given by prominent researchers in their respective fields, and open problem statements by all participants.

Seminar October 30–4, 2016 – http://www.dagstuhl.de/16442
1998 ACM Subject Classification I.2.7 Natural Language Processing, I.2.9 Robotics
Keywords and phrases animal calls, human-robot interaction, language evolution, language universals, speech technology, spoken language, vocal expression, vocal interaction, vocal learning
Digital Object Identifier 10.4230/DagRep.6.10.154

1 Executive Summary

Serge Thill
Ricard Marxer
Roger K. Moore

Almost all animals exploit vocal signals for a range of ecologically-motivated purposes. For example, predators may use vocal cues to detect their prey (and vice versa), and a variety of animals (such as birds, frogs, dogs, wolves, foxes, jackals, coyotes, etc.) use vocalisation to mark or defend their territory. Social animals (including human beings) also use vocalisation to express emotions, to establish social relations and to share information, and humans beings have extended this behaviour to a very high level of sophistication through the evolution of speech and language – a phenomenon that appears to be unique in the animal kingdom, but which shares many characteristics with the communication systems of other animals.

Also, recent years have seen important developments in a range of technologies relating to vocalisation. For example, systems have been created to analyse and playback animals calls, to investigate how vocal signalling might evolve in communicative agents, and to interact with users of spoken language technology (voice-based human-computer interaction using speech technologies such as automatic speech recognition and text-to-speech synthesis). Indeed, the

Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), *Dagstuhl Reports*, Vol. 6, Issue 10, pp. 154–194
Editors: Roger K. Moore, Serge Thill, and Ricard Marxer
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

latter has witnessed huge commercial success in the past 10-20 years, particularly since the release of *Naturally Speaking* (Dragon's continuous speech dictation software for a PC) in 1997 and Siri (Apple's voice-operated personal assistant and knowledge navigator for the iPhone) in 2011. Research interest in this area is now beginning to focus on voice-enabling autonomous social agents (such as robots).

Therefore, whether it is a bird raising an alarm, a whale calling to potential partners, a dog responding to human commands, a parent reading a story with a child, or a businessperson accessing stock prices using an automated voice service on their mobile phone, vocalisation provides a valuable communications channel through which behaviour may be coordinated and controlled, and information may be distributed and acquired.

Indeed, the ubiquity of vocal interaction has given rise to a wealth of research across an extremely diverse array of fields from the behavioural and language sciences to engineering, technology and robotics. This means that there is huge potential for crossfertilisation between the different disciplines involved in the study and exploitation of vocal interactivity. For example, it might be possible to use contemporary advances in machine learning to analyse animal activity in different habitats, or to use robots to investigate contemporary theories of language grounding. Likewise, an understanding of animal vocal behaviour might inform how vocal expressivity might be integrated into the next generation of autonomous social agents. Some of these issues have already been addressed by relevant sub-sections of the research community. However, many opportunities remain unexplored, not least due to the lack of a suitable forum to bring the relevant people together.

Our Dagstuhl seminar on the topic of "Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)" provided the unique and timely opportunity to bring together scientists and engineers from a number of different fields to appraise our current level of knowledge. Our broad aim was to focus discussion on the general principles of vocal interactivity as well as evaluating the state-of-the-art in our understanding of vocal interaction within-and-between humans, animals and robots. Some of these sub-topics, such as human spoken language or vocal interactivity between animals, have a long history of scientific research. Others, such as vocal interaction between robots or between robots and animals, are less well studied – mainly due to the relatively recent appearance of the relevant technology. What is interesting is that, independent of whether the sub-topics are well established fields or relatively new research domains, there is an abundance of open research questions which may benefit from a comparative interdisciplinary analysis of the type addressed in this seminar.

2 Table of Contents

Executive Summary

Seminar Organization

Overview of Talks

Open Problems

3 Seminar Organization

3.1 Participants

Participants at the seminar spanned the entire range of academic career stages and provided broad global coverage. The mix of attendants was also particularly interdisciplinary, covering animal vocalisation, human language, machine language production and understanding, as well as various intersections of these topics. The success of the seminar can largely be attributed to this presence of multi- and pluridisciplinary interests and the discussions across boundaries that this generated

3.2 Overall organisation

We intentionally kept the structure of the seminar rather loose and self-organising. We started the seminar with brief presentations from all participants in which they were asked to summarise their background, and what they felt the most pressing issues were. This was then followed by invited "stimulus" talks by selected participants from different home disciplines. The intention with these talks was to foster an initial interdisciplinary interest between the different communities present.

The remainder of the seminar was spent in groups that focussed on specific issues thus identified, which spanned a broad range of topics, as can be seen in the remainder of this report.

3.3 Prioritisation of open research questions

The seminar was guided by a recent review paper by the organisers, which identified a number of open research questions in the field of VIHAR. During the seminar, we asked the participants to identify the three questions they considered most crucial / relevant. This allows us to here present the resulting order of questions that received at least one vote:

1. *(16 votes)*
   - What is the relationship (if any) between language and the different signalling systems employed by non-human animals?
2. *(11 votes)*
   - What tools might be needed in the future to study vocalisation in the wild?
3. *(10 votes)*
   - What are the similarities/differences between the vocal systems (including brain organisation) in different animals?
4. *(8 votes)*
   - How does one evolve the complexity of voice-based interfaces from simple structured dialogues to more flexible conversational designs without confusing the user?
   - Are there any mathematical modelling principles that may be applied to all forms of vocal interactivity and is it possible to derive a common architecture or framework for describing vocal interactivity?
5. *(7 votes)*
   - What are the limitations (if any) of vocal interaction between non-conspecifics?
6. *(6 votes)*
   - How can vocal interactivity as an emergent phenomenon be modelled computationally?

7. *(5 votes)*
   - What are the common features of vocal learning that species capable of it share, and why is it restricted to only a few species?
8. *(4 votes)*
   - What are the common features of vocal learning that species capable of it share, and why is it restricted to only a few species?
   - How are vocal mechanisms constrained or facilitated by the morphology of the individual agents involved?
   - How is information distributed across the different modes and what is the relationship between vocal and non-vocal (sign) language?
   - To what degree can affective states be interpreted and expressed, and should they be treated as superficial or more deeply rooted aspects of behaviour?
   - Do the characteristics of vocalisations carry information about the social relationship connecting the interactants (for example, how is group membership or social status signalled vocally)?
9. *(3 votes)*
   - Do artificial agents need ToM in order to interact effectively with human beings vocally?
   - How are multi-modal behaviours orchestrated, especially in multi-agent situations?
   - Who should adapt to whom in order to establish an effective channel?
   - How would one model the relevant dynamics (whether to study natural interactivity or to facilitate human-machine interaction)?
   - How can insights from such questions inform the design of vocally interactive artificial agents beyond Siri?
10. *(2 votes)*
    - How are vocalisations manipulated to achieve the desired results and is such behaviour reactive or proactive?
    - Is ToM crucial for language-based interaction?
    - To what degree is there a phonemic structure to animal communications, and how would one experimentally measure the complexity of vocal interactions (beyond information-theoretic analyses)?
    - What is it about the human-dog relationship that makes the one-sidedness of this relation sufficient, and conversely, what can biases in communication balancing say about social relationships?
    - To what extent are vocal signals teleological, and is it possible to distinguish between intentional and unintentional vocalisations?
    - Given the crucial nature of synchrony and timing in interactivity between natural agents, to what extent does this importance carry over to human-machine dialogue?
    - Is it necessary to create new standards in order to facilitate more efficient sharing of research resources?
11. *(1 vote)*
    - What is the role of vocal affect in coordinating cooperative or competitive behaviour?
    - How does a young animal (such as a human child) solve the correspondence problem between the vocalisations that they hear and the sounds that they can produce?
    - Does the existence (or absence) of prior relationships between agents impact on subsequent vocal activity?
    - How is vocalisation used to sustain long-term social relations?
    - What can be learned from attempts to teach animals human language (and vice versa)?

3.4 Conclusions and next steps

Overall, the seminar proved very successful in the fostering of a new community targetting the interdisciplinary challenges in the field of VIHAR. we now intend to keep the momentum going and, as a next step, are organising a workshop on VIHAR as a satellite event of Interspeech 2017 (see http://vihar-2017.vihar.org/ for details). This workshop is a direct consequence of the Dagstuhl seminar; indeed its organising committee consists of participants in the seminar. In conclusion, we hope that, with this seminar, we have laid the foundations for a new and vibrant research community that will remain active and meet regularly for years to come.

4 Overview of Talks

4.1 (Vocal) interaction with the artificial

Tony Belpaeme (University of Plymouth, GB)

Artificial Intelligent systems, from digital assistants to humanoid robots, are already part and parcel of our daily lives or are expected to be in the not too distant future. When interacting with artificial systems, we now use channels different to those we use to interact with others. We type commands and search terms, but do not yet engage in dynamic social interactions with machines. It is however interesting to see how anthropomorphisation colours our interactions with machines: we cannot help interpret behaviours of machines (even the simplest Braitenberg vehicles) as having intention and personality, and this can be exploited by engineers when building socially interactive robots. I will present two studies, the first study looks into how people interpret robotic clicks and beeps (such as the sounds made by R2D2). These non-linguistic utterances are readily interpreted as containing emotion, and adults interpret these sounds categorically. The second study shows how a robot can leverage the human propensity to tutor: using a social robot we study how people teach it, and show that the they form a mental model of the robot used to tailor the teaching experiences for the robot. Both studies not only show how human social interaction spills over to interacting with machines, but also demonstrate the promise of using robots as research tools, achieving a level of control and repeatability unachievable by current research methods in human and animal interaction.

4.2 Acoustic communication in animals: a window to their inner state

Elodie Briefer (ETH Zürich, CH)

My presentation focuses on three main questions, relevant for VIHAR (Moore et al. 2016 Front. Robot. AI 3:61); 1) To what degree is there a phonemic structure to animal communications, and how would one experimentally measure the complexity of vocal interactions (beyond information–theoretic analyses)?; 2) To what degree can affective states be interpreted and expressed, and should they be treated as superficial or more deeply rooted aspects of behavior?; 3) What are the limitations (if any) of vocal interaction between non-conspecifics?

1. To what degree is there a phonemic structure to animal communications, and how would one experimentally measure the complexity of vocal interactions (beyond information-theoretic analyses). I discuss the first question through my PhD research on skylarks. Skylark songs are among the most complex acoustic signals compared to other songbird species, both in terms of the number of different acoustic units produced by each male, but also in how they are arranged within songs (high diversity of transitions). Markov chain analyses show that skylark's song are best modelled by a 1st order Markov chain governed by a finite-state grammar. However, as each acoustic unit bears no "meaning" per se (different units do not have different referential meaning), this structure could thus be described as "phonetic patterning". In addition, geographical variation exists at the sequential level; sequences of syllables are shared by neighbouring birds. Playback experiments revealed that the dialect constitutes a group signature used by birds to discriminate neighbours (birds from the same group) from strangers (birds from a different group), and that the order of acoustic units within these particular sequences is an important feature for the neighbourhood identity coding.
2. To what degree can affective states be interpreted and expressed, and should they be treated as superficial or more deeply rooted aspects of behavior? I discuss the second question through my current research on vocal expression of emotions. Expression of emotions plays an important role in social species, including humans, because it regulates social interactions. Indicators of emotions in human voice have been studied in detail. However, similar studies testing a direct link between emotions and vocal parameters in non-human animals are rare. In particular, little is known about how animals encode in their vocalisations, information about the valence (positive/negative) of the emotion they are experiencing. I combined new frameworks recently adapted from humans to animals to analyse vocalisations (source-filter theory), and emotions (dimensional approach), in order to decipher vocal expression of both arousal (bodily activation) and valence in domestic ungulates. I present my results on horses, which are part of a large project aimed at investigating the evolution of vocal expression of emotions in ungulates (goats, horses, pigs and cattle) and the effect of domestication on human-animal communication. I measured physiological, behavioural and vocal responses of the animals to several situations characterised by different emotional arousal and valence. Physiological and behavioural measures collected during the tests confirmed the presence of different underlying emotions. My results showed that horse whinnies are composed of two fundamental frequencies (two "voices"), suggesting biphonation, a rare case among mammals. Interestingly, one of these fundamental frequency and the energy spectrum indicates emotional arousal, while the other and the duration indicates the emotional valence of the producer. These findings show that cues to emotional arousal and valence are segregated in different, relatively independent parameters of horse whinnies. Most of the emotion-related changes to vocalisations that I observed are similar to those observed in humans and other species, suggesting that vocal expression of emotions has been conserved throughout evolution.
3. What are the limitations (if any) of vocal interaction between non-conspecifics? To discuss the third question, I am showing current experiments that I am currently running, testing if domestic and wild horses perceive indicators of emotions in conspecific vocalisations, vocalisations of closely related species (wild horses to domestic horses and vice versa), as well as in human voice. I am also testing if humans can perceive emotions in the

vocalisations of domestic (goats, horses, pigs and cattle) and wild (Przewalski's horses and wild boars) ungulates using an online questionnaire. These experiments will shed light on the evolution of vocal expression of emotions, and on the impact of domestication on human-animal communication of emotions. From this work, more questions arise, and an inter-disciplinary approach joining research on vocal communication within and between animals, humans and robots would be greatly beneficial to share tools and skills, in order to lead to further advances in these fields of research.

4.3 Lessons about Vocal Interaction from Joint Speech, or How I learned to love languaging

Fred Cummins (University College Dublin, IE)

The theme of this seminar encourages us to look beyond the well-studied terrain of inter-human communication towards some novel and unexplored challenges. This attitude seems to suggest that inter-human communication is sufficiently well-studied to support generalisation, but I will argue that there is plenty of reason to believe that we have failed to identify language as an object of study. The shortcomings of received approaches to language are graphically illustrated by the wholly Christian and literate foundation of orthodox accounts of language that ignore such essential parts of the fabric of communication as (1) gesture, (2) gaze, (3) posture, (4) prosody and my favourite topic, (5) joint speech. Joint speech is speech produced by multiple speakers all saying the same thing at the same time, as found in practices of prayer, ritual, protest, sports traditions, and beyond. The study of such common and important practices throws up some conundrums and sensitises us to some themes notably absent from most of contemporary linguistics: Common Ground emerges as an essential concept for understanding the dialogical push and pull of many interactions. This may be relevant when we come to consider human-animal communication, where it may find expression in the Bayesian language of shared priors. Co-presence is an important theme that gets lost when we approach languaging from a representationalist perspective. The importance of context, rather than lexis and text, jumps out at us, and in place of an exegesis based on the analysis of encoded messages, we are encouraged to look instead at the role of real-time reciprocal interaction. Thus joint speech may alert us to some blind spots we have in accounts of inter-human communication that will gain significance as our study extends to communication with other types of beings.

4.4 Universals in Vertebrate Vocal Communication

Angela Dassow (Carthage College – Kenosha, US)

What is the difference between communication and language? What properties are shared between animal vocalizations and human speech? What approach should be taken when comparing vocalizations produced by different species and how can we employ signal processing techniques to address these questions?

My stimulus talk began by addressing common approaches to studying animal-animal interactions and the attempts made to date to find linguistic properties in animal vocal communication systems. Many such approaches have relied on a single field of interest, either ethology or linguistics. This has resulted in over generalization of capabilities or under appreciation for the complexity of vocal systems. Neither have done a sufficient job to address core evolutionary structures or commonalities across taxa. As an example of this issue, we explored the relevance of searching for vowel harmony in other species. This discussion was based on a study which suggests cotton-top tamarins do not perceive vowel harmony in playback experiments. While there is experimental evidence to support this conclusion, the fundamental question of why would we expect to find vowel harmony in tamarins, when it doesn't exist in all human languages remains. I proposed that this issue and others similar to it, could be avoided by employing an interdisciplinary approach to searching for phonological properties in animal communication systems.

To address the question of what approach should be taken when comparing vocalizations of various species, I proposed a movement from focusing on vocal repertoires, which can be misleading with respect to vocal complexity, to analyzing sequences of acoustic units. As an example of this issue, we discussed the problem of categorizing acoustic units in species with graded vocalizations versus animals with discrete vocalizations. For animals with discrete vocalizations, there is a mismatch between the number of measureable acoustic categories in a given species and their evolutionary relationships. For example, the Irrawaddy dolphin produces fewer categories of sound than the Madagascar treefrog or the Australian long-neck turtle. If we were to use the number of acoustic categories as our main measure of complexity in vocal communication systems, we may erroneously conclude that a dolphin has a more primitive form of communication than a frog or a turtle. To avoid this issue, it is important to examine how these units are being combined. Sequence analysis may offer insight into potential meaning in a vocal communication system which would provide a basis for comparisons made between closely related taxa, such as the sixteen species of extant gibbons.

4.5 Temporal models for machine listening in mixed audio scenes

Dan Stowell (Queen Mary University of London, GB)

We wish to be able to analyse soundscapes with multiple vocalising individuals, where those individuals might be human, animal, or otherwise, and might or might not be interacting. Many current models for analysing vocalisation sequences are surprisingly limited for this purpose: they assume there is a single symbol sequence with a strict chain of causality (this might be one individual, or a turn-taking exchange between individuals); they neglect important aspects such as the timing of vocalisations; they assume each vocal unit is a single quantum of meaning.

Simplified models enable efficient inference and can be applied to many species – we do gain a lot from the abstraction of the Markov model, for example. Generic models are essential for handling outdoor sound scenes with dozens of potential species present. But in order to apply machine listening methods to multi-party sound scenes, we need models designed for multiple parties acting in parallel. These models have two complementary purposes: we fit them to data to measure animal behaviour, and we use fitted models to make inferences in new sound recordings.

I give two specific examples of multi-party models:

1. Multiple Markov renewal processes running in parallel. With this, we can segregate concurrent streams of events. [1]
2. A point-process model in which calls from individuals influence each others' probability of calling. This deals well with multiple parallel influences converging on an individual. With it, we characterise the communication network in a group. [2]

These two paradigms hint at ways forward. Future methods will need richer underlying structure – but what? Sequence modelling? Affective state? Physiological state? Theory of mind? The key question for feasible analysis is, how little can we get away with?

As a separate issue I also discuss "active spaces" and our ideas of signal content. We often treat a vocal unit as having a single purpose and a single audience. In animal communication the concept of an "active space" is the physical space in which a receiver can hear enough of the sound to decode the message conveyed. But it is well known to students of human language that a single utterance can simultaneously have multiple meanings targeted at different audiences. Birdsong contains structural features such as chirp sounds (rapid frequency modulation) which offer a mechanism for multivalent utterances with different spatial extents [3]. We shouldn't accidentally overlook that animals might make use of such possibilities.

References

1 D. Stowell and M. D. Plumbley, *Segregating event streams and noise with a Markov renewal process model.* Journal of Machine Learning Research 14, 1891–1916, 2013.

2 D. Stowell, L. F. Gill, and D. Clayton. *Detailed temporal structure of communication networks in groups of songbirds.* Journal of the Royal Society Interface, 13(119), 2016.

3 Mathevon, N. and Aubin, T. and Vielliard, J. and da Silva, M.-L. and Sebe, F. and Boscolo, D., *Singing in the Rain Forest: How a Tropical Bird Song Transfers Information.* Plos One 3(2), 2008.

5 Open Problems

5.1 Statement by Andrey Anikin

Andrey Anikin (Lund University, SE)

The central assumption, for me, is that humans possess a number of species-specific (innate, as opposed to culturally learned) vocalizations, at least some of which are shared with other primates. If we can pinpoint these vocalizations through phylogenetic reconstruction and cross-cultural research, we will have a better understanding of the developmental constraints under which humans acquire their vocal repertoire, including both non-speech sounds and prosodic features of spoken language. This, in turn, will improve human-machine interaction through better recognition and production of vocalizations and prosodically natural speech by machines.

This formulation is broad enough to make it natural to include humans, animals, and robots in the same framework, but still specific enough to lead to testable predictions for empirical research and to have specific practical implications for affective computing. To unpack, this view of VIHAR involves coordinated efforts and contributions from three fields, as follows:

1. Animal communication.
   a. Data. To know which sounds humans share with other primates, it is essential to have good descriptions of the vocal repertoire of species closely related to humans, especially the great apes: the acoustic form of vocalizations and typical contexts of their production.
   b. Method. Researchers studying vocal communication in animals cannot simply ask their subjects what a sound "means". This has led to a search for stringent methods of classifying sounds into acoustic types (unsupervised classification using some form of cluster analysis, etc), without assuming a priori that each vocalization is specific to one particular context. In my opinion, this methodology is superior compared to the tendency in human research to map sound onto meaning directly, bypassing the level of vocalization.
2. Psychology.
   a. Large cross-cultural corpora. The existing corpora of non-linguistic vocalizations are relatively small (e.g., compared to the size of speech corpora and collections of animal vocalizations) and limited to a few Western cultures. To find acoustic universals, larger and more diverse corpora have to become available.
   b. Sound-to-meaning mapping. The relative contribution of within-call and between-call variation needs to be addressed. What range of emotions can a scream indicate? Are the acoustic differences between a scream of anger vs. fear the same as those between an aggressive "evil" laugh and a friendly laugh? Do Morton's structural-motivational rules apply to human vocalizations? Are (some) vocalizations and/or emotions perceived categorically? These and other questions can be approached via perceptual studies of both natural recorded vocalizations and synthetic sounds (hybrids of natural vocalizations and/or sounds generated "from scratch").
3. Affective computing.
   a. Machine learning for sound recognition. This buzzing field is developing very rapidly, but arguably suffers from a piecemeal approach with each team using different training corpora and categories. In my opinion, a more systematic approach with standardized corpora and a more theoretically justified architecture could improve the generalizability of results. In particular, it may be fruitful to introduce a priori constraints on classifiers (e.g., specify dedicated detectors for innate vocalizations, such as laughs and screams) and an intermediate level of acoustic categories distinct from meaning.
   b. Sound production. There is already considerable interest in producing emotionally charged computer speech. Non-speech vocalizations are a natural extension of this project, and again, just as with recognition, their production can benefit from a more theoretically sound framework. I can conclude by stating, in all humbleness, that I've been trying to peck at the problem from all of the perspectives described above. To do more than scratch the surface, however, collaborative efforts are a vital necessity, which is why I believe that VIHAR as a cross-disciplinary framework is the answer.

5.2 Statement by Timo Baumann

Timo Baumann (Universität Hamburg, DE)

Highly Responsive Vocal Interaction through Incremental Processing

Vocal interaction (and this is not limited to vocal interaction but also extends to gesture, mimickry and interactive behaviours) is like an intricate dance: what one interlocutor does is potentially immediately analyzed and interpreted by the other and likely incorporated in that interlocutors response behaviour (e.g. backchannelling while listening to speech). While taking turns (a coarse-grained differentiation of sending/receiving in an ongoing interaction) is the *modus operandi* of most human-machine interaction, the more responsive behaviours like backchannelling, blinking, and timing contributions are probably similarly important to achieve good and natural interaction performance.

I work in the area of system architectures for very low-latency reactions and controllable reflexive behaviours, based on incremental processing [1] which allows the concurrent and modular processing of information *as it happens*, including the extrapolation/prediction into the future. Challenges in incremental processing are plentiful and my systems focus on novel interactive behaviours rather than on accomplishing well what existing systems already do (activities like booking a train ticket). Thus, the topic of VIHAR interests me primarily for two reasons:

- Human-animal interaction is often less task-driven and more interaction-driven than human-human interaction (and spoken human-machine interaction which focuses on solving particular problems). Thus, it's a domain in which meaningful interactions are first becoming feasible for incremental systems and I want to learn from researchers on animal interaction about the underlying patterns. Similarly, I believe that contact with roboticists will help to improve interaction capabilities of robots.
- Secondly, I am interested in optimality of the complex interaction system (between and among humans, animals and robots). The wide variety of decision making that is possible at any moment during an interaction and may just look like a small cause may have large effects on the overall outcome. Yet, it is unclear which causes have which effects and to correctly anticipate their magnitudes. I believe VIHAR as a testbed of interaction research is highly valuable to sketch out the possible design spaces of various (natural) interaction systems. What is more, I believe that ultimate human-machine interaction need not necessarily mimic human-human interaction patterns but that better spots in the interaction design space may exist. Inspiration across species-specific research will be very helpful to find better ways of interacting.

References

1 Baumann, Timo, *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components.* PhD Thesis, Universität Bielefeld, Germany, 2013.

5.3 Statement by Tony Belpaeme

Tony Belpaeme (University of Plymouth, GB)

How can robots tap into interactivity?

What fascinates me is the point where vocal interactivity becomes verbal interactivity. The point where vocal utterances are no longer mere grunts, whistles or calls, but where the vocal signal is a package containing distinct chunks. These chunks seem to be the solution (one of many) which animals and humans adopted to communicate symbolic semantics. In animal communication, these chunks seem to break the boundaries of mating, alarm and territorial calls, and carry more complex meaning: they still might be alarm calls, but will now –for example– distinguish the type of threat, as some lemurs and monkeys do. In human language, vocal chunks are now words or grammatical markers, and when strung together they can carry complex, recursive semantic content.

Human cognition relies heavily on intelligent others to evolve and develop, and vocal/verbal communication plays a central role here. When trying to build intelligent machines, such as robots, these not only require the ability to interact with people, but might need a process which helps them tap into human interactions for to mere purpose of bootstrapping and developing their artificial cognition. Concepts, for example, are only to a certain extent acquired through perceiving the physical environment (the so-called "physical grounding" of concepts), but are predominantly subject to a cultural process which relies on interaction. We learn to demarcate the concept of RED not just through experiencing red, but through communicating with others using the word "red" in an appropriate context [1].

Can machines – computers, cloud-based systems, robots – have access to these concepts? To some extent it seems that it is possible to let machines tap into human communication and extract semantic structure: big data approaches show that the simple processes of co-occurrence and correlation can extract semantic relations from mere text. But are there limits to our current methods? Big data and Deep Learning are very much en vogue, and it would seem that the performance of their applications, such as speech recognition, keeps improving with ever more data. But while they are connectionist methods, and therefore have some natural plausibility, they also require huge amounts of annotated data and are therefore fundamentally different to natural learning processes. So a question we need to ask is: are there skills that are fundamentally outside the grasp of these new AI techniques?

One aspect that sets these machine learning methods apart from human learning is the fact that they are batch learners: they feed on huge datasets without the need to interact while learning. Human learning and social learning in animals rely on a tightly coupled interactions between learner and tutor. The tutor, often an adult, will spend considerable resources teaching the learner and will shape the interaction to meet the learner's needs. From motherese to acquire speech sounds to demonstrations of skills, people seems to have a propensity to structure their interactions to allow the transmission of knowledge and skills. Can machines leverage this to move away from the need for large training sets? Would people be willing to teach machines? What would human-robot interaction look like if machines would learn through interaction? And while building such robots, we not only build novel learning methods, but also develop new methods with which we might study interaction and cognition. We have build a set-up to explore these questions and results indicate that indeed people build a mental model of the robot and tailor the interaction to fit the robot's learning needs [2].

A different, but still related issue is the relation people have with machines, and with robots in specific. We know that robots are seen as having agency and that much of what a robot does is interpreted is being meaningful. When developing robots to interact with people we need to be aware of how the robot verbal and non-verbal behaviour will be interpreted. With respect to vocal communication, we are quite used to hearing robots utter clicks and beeps, which we call non-linguistic utterances [3]. These utterances are readily interpreted as meaningful by people and seem to be subject to categorical perception, showing how neural mechanisms which evolved for natural communication seems to be sensitive to artificial communicative acts as well [4].

References

1 Luc Steels and Tony Belpaeme, *Coordinating perceptually grounded categories through language: a case study for colour.* Behavioral and brain sciences 28(4), 469–488, 2005.

2 Joachim de Greeff and Tony Belpaeme, *Why Robots Should Be Social: Enhancing Machine Learning through Social Human-Robot Interaction.* PLOS ONE 10(9): e0138061, 2015

3 Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst, *Review of Semantic-Free Utterances in Social Human-Robot Interaction.* International Journal Of Human-Computer Interaction, 32(1), 2016

4 Robin Read and Tony Belpaeme, *People Interpret Robotic Non-linguistic Utterances Categorically.* International Journal of Social Robotics, 8(1), 31–50, 2016

5.4 Statement by Elodie Briefer

Elodie Briefer (ETH Zürich, CH)

Since my PhD, I have been investigating the acoustic communication of several species, including skylarks, fallow deer, goats, horses, Przewalski's horses, pigs, wild boars and cattle. All these species differ widely in the form and complexity of the sounds they produce and raise different questions/challenges for VIHAR. I will here only focus only on my main current project, the study of vocal expression and contagion of emotions in ungulates.

Emotions play an important role in social species, because they guide behavioural decisions in response to events or stimuli of importance for the organism and hence regulate social interactions (e.g. approach or avoidance). Indicators of emotions in human voice have been studied in detail. However, similar studies testing a direct link between emotions and vocal structure in non-human animals are rare. In particular, little is known about how animals encode in their vocalisations, information about the valence (positive/negative) of the emotion they are experiencing. Furthermore, the potential for emotions to be transmitted to conspecifics and hetero-specifics through vocalisations (vocal contagion of emotions) has been poorly studied. A comparative approach between humans and other animals would give us a better understanding of how the expression of emotions evolved.

My current project aims at combining methods to study emotions and vocalisations in order to investigate the evolution of vocal expression of emotions and the impact of domestication on humananimal communication of emotions. My project focusses on (1) vocal expression of emotions in domestic and wild ungulates; (2) perception and contagion of emotions between conspecifics; (3) perception and contagion of emotions between closely related domestic and wild ungulates; (4) perception and contagion of emotions between domestic and wild ungulates and humans. It includes goats, horses, Przewalski's horses, pigs, wild boars and cattle.

I am listing below some of the challenges for VIHAR that my project raises:

Fundamental challenges

- How can we best compare vocal expression of emotions in animals and humans? My research focusses on "subtle" acoustic variation occurring within call types (e.g. within horse whinnies) as a function of emotional valence and arousal; Is it correct to compare emotion-related changes in vocalisation types (e.g. bark → growl) to human nonverbal emotion expressions (e.g. laughter → screams), while variation within vocalisation types is closer to affective prosody?
- Can we really differentiate between "emotional" and "intentional" signals in animals?

The main application of my research resides in the assessment and improvement of animal welfare.

Applications

- Emotion expression: Development of automated tools that would recognize animal's emotions from their vocalisations. Can these tools be trained on the calls of each individual? Such tools could allow animal keepers to be informed when a certain threshold of vocalisations indicating negative emotions are produced and could thus take action to improve welfare.
- Emotional contagion: Development of acoustic tools that would decrease negative arousal (e.g. during stressful husbandry procedure) and promote positive emotions. These tools could take the form of synthetic vocalisations based on our knowledge of parameters that trigger emotions in receivers.

5.5 Statement by Nick Campbell

Nick Campbell (Trinity College Dublin, IE)

Pragmatism, Context-sensitivity, and the Robot-Dialogue Interface

I come to this meeting from a background of speech processing for human-human translation machines with a specific emphasis on speech synthesis, particularly concerning utterance generation and timing. Now working on autonomous robot dialogue interfaces for human-robot interaction, my prime interest is in the style and content of utterances delivered by the device: for a natural-seeming spoken interaction, the speech must be relaxed, apparently spontaneous, and contextually appropriate.

Previous research with the "Herme" conversational robot [1] has shown that even without an understanding module (or even functioning speech recognition) a machine (robot) can maintain a natural-seeming conversation with a human for between three and five minutes. However, going beyond this simple time limit will require an element of understanding on the part of the robot in order to continue the conversation and contribute satisfactorily.

The goal of our work in the Speech Communication Lab at the University of Dublin, Trinity College, is not to create yet another chatbot but to understand how to improve the delivery of predetermined utterances in the context of engaging the interlocutor and assessing the cognitive effect of each message. For me, an issue to be addressed at this VIHAR meeting is the extent of understanding required by the robot for an efficient situated dialogue; whether full Theory-of-Mind is required for linguistic grounding or whether simpler pragmatic/functional constraints of the dialogue context can sufficiently restrict the interaction for the robot to respond from a limited list of pre-prepared default utterances or utterance-types. I bring to the meeting a small study of dog barks in the context of human engagement and show from that work how rather than relying on an underlying 'language of barks' that each dog/human pair has to learn, there is a situational context dependency from which an interpretation can be gained.

In the barking study [2], we did not find that bark type generalised widely between different dogs of the same species but infer that each dog had developed its own similar-sounding bark type in response to a common set of everyday situations under common articulatory constraints. In the context of human-robot interaction, it may prove to be the case that rather than share a common (human) language, sufficient sounds may trigger an appropriate reaction in a given context when the constraints of that context are understood by both participants. In implementing such a model, we focus first on determining the degree of engagement of the human (i.e., where his or her attention is directed when the robot is speaking or about to speak) and on maintaining sufficient contact throughout a speech interlude so that the desired message may be delivered and the interaction satisfactorily completed.

Here the example of a receptionist robot dialogue interface collaboratively built during the recent eNTERFACE workshop becomes relevant [3]; the situation is extremely constrained but practical, and the receptionist simply has to direct each customer/patient to the desired room as they arrive. In our test case, there are only two rooms and only two humans in the robot's universe. We built an exhaustive model of how to deal with each customer (including the utterances required for each move) and how to manage the queueing of customers when more than one was present. The robot also had a set of idling behaviours to return to when each customer was served. Rather than program this behaviour deterministically, we had access to the Flipper dialogue management engine [4] that continually tests the envorinment for a set of given conditions and then acts (and resets the environmental state) accordingly. The set of condition-behaviour-response tokens is large but finite. The success of this model depends on the responses also being finite, but we claim that this might be the case for a large number of real-world situations and that full 'understanding', particularly 'linguistic understanding' on the part of the robot, might not be necessary. The use of other sensors, however, is mandatory, and our robot is able to see the environment and to recognise simple gestures such as pointing.

It will be interesting to hear whether colleagues from the animal sciences have any contributions to make to this model from their observations of animal behaviour and of the very restricted use of 'language' that animals appear to make.

References

1 Han, J., Gilmartin, E., De Looze, C., Vaughan, B., and Campbell, N., *The Herme Database of Spontaneous Multimodal Human–Robot Dialogues.*, Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21-27 May 2012, edited by ELRA, 2012

2 Nick Campbell, *An acoustic analysis of 7395 dog barks.* in editor(s) Rudiger Hoffman, Festschrift, Book Chapter, 2013

3 Daniel Davison, Binnur Gorer, Jan Kolkmeier, Jeroen Linssen, Bob Schadenberg, Bob van de Vijver, Nick Campbell, Edwin Dertien, Dennis Reidsma *Things that Make Robots*

Go HMMM: Heterogeneous Multilevel Multimodal Mixing to Realise Fluent, Multiparty, Human-Robot Interaction. Proc eNTERFACE, Enschede, Netherlands, 2016 (in press)

4 MarkMaat, T., and Heylen, D., *Flipper: An Information State Component for Spoken Dialogue Systems.* in Intelligent Virtual Agents. Reykjavik: Springer Verlag, 2011, pp. 470–472.

5.6 Statement by Fred Cummins

Fred Cummins (University College Dublin, IE)

My work thematises *Joint Speech*, defined as speech produced by multiple people at the same time. It is a familiar form of speech, serving to empirically pick out highly valued domains of human practice including practices of prayer, ritual, protest, and the enactment of collective identity among sports fans. The absence of any empirical scientific work in this domain is revealing. It demonstrates a fixation within the human sciences on a solipsistic, Cartesian, and, yes, very obviously Christian, approach to the person, that has treated langauge as a form of modality-neutral passing of encoded messages containing information. This obsession of the science of "language" is necessarily blind to the manner in which we bring a shared world into being through our coordinated practices, including our vocalisations. In my view, linguistics has failed to identify language in the first place.

The study of joint speech serves to bring some neglected themes to the fore in place of the concerns of academic linguistics. Joint speech is clearly a highly central example of language use, as old as humanity, and instrumental in bringing many kinds of human collectivities into being. Yet in joint speech some familiar distinctions vanish. The opposition of speaker and listener is no longer relevant, as everybody is both and the texts uttered are authored elsewhere. Likewise, there is no principled manner to distinguish between speech and music any more, as we find all possible points on a continuum from the amusical recitation of an oath on a singular occasion, through the rhythmically and melodically exaggerated chants of repeated prayers or protest calls, to the unison singing of plainsong or the familiar Happy Birthday. Joint speech cannot be replaced by written texts. It is performative at its core: taking part in joint speech practices is not a neutral activity conducted in an intellectual tone. It is an act of commitment, through which many of the structural elements of any human society are made manifest and are maintained by doing. Studying joint speech changes the principal themes we might pursue as we study vocal behaviour, and these concerns, once recognised, may be extended far beyond the rather narrow specification of the definition of joint speech itself. They appear to me to shed light on all human vocal communication and to extend naturally to human-animal interactions as well. To the extent that they generate novel ways of conceiving of the role of the voice in interaction, they may prove relevant to human-robot interactions as well. I have not pursued that particular thread yet.

Once the message-passing metaphor is no longer relevant, we uncover instead a realisation that the *real time recurrent interaction* among participants is an essential element to any joint speech event. Participants are in contact with each other in a very important manner. This recognition also serves to shift the focus from notions of representation and reference, to an awareness of the importance of *co-presence* among participants. Making *vocal interactivity* a research theme seems to me to offer a more promising starting point for understanding what people are doing in such situations than anything on offer within an anaemic and abstract "linguistics".

Occasions in which joint speech is important are inevitably embedded in rich and highly charged suties of practices that are themselves highly informative about the values and lifeworlds of the participants. Any single instance of joint speech needs to be interpreted with a keen sense of the context in which it is embedded. Transcription is irrelevant. What is needed instead is the notion of *thick description*, providing as much supporting material to reveal the specific context-bound manner in which one or other instance of joint speech is integrated into meaning-making activities. I suspect we have a lot more observation to do before we can redress the shortcomings of the received syntax-first approach to language that seems to be good mainly for bible translation.

References

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

2 Cummins, F. (2014) *The remarkable unremarkableness of joint speech* in Proceedings of the 10th International Seminar on Speech Production, pages 73–77, Cologne, DE.

5.7 Statement by Angela Dassow

Angela Dassow (Carthage College – Kenosha, US)

Understanding evolutionary relationships through examining vocal interactivity

In non-human animals, communication is widely viewed as a behavior; it is a reflexive activity designed to produce behavioral responses in conspecifics or across species. In contrast, language is a human affair. It transfers conceptual knowledge from speaker to listener and has extraordinarily generalizable descriptive powers. While spoken language may lead to behavior, this is unnecessary.

Longstanding debates regarding the use of language in non-human animals have focused on nested and recursive syntactic structures as proxies for the core competencies associated with human cognition. However, current understanding of non-human cognition precludes the need for complex conceptual representations that require the deep mathematical structure of human language. Put plainly, what precisely do these animals have to think about, let alone communicate? Current paucity in understanding cognition in other species renders this question simply provocative.

That aside, evaluating the existence of language in non-humans solely via analogies to syntax begs the question: how would vocalizations hypothetically possessing syntax be constructed? While insistence of parity with the structural complexity of human language are widespread, the precursor questions of morphology and phonology in non-humans have largely been ignored. The constituent pieces that enable formation of more elaborate structures must be examined before comparisons of structural complexity can be met.

The goal of my research is to characterize linguistic commonalities in different vertebrate species that communicate vocally. Specifically, my interests lie within the following problems:

- **Commonalities derived thru evolution:** Much like Darwin noticed similarities of physical features such as wings, we are noticing similarities in sound categories across several diverse species. As a component of evolution, selection can only occur on existing structures. This stands to reason then for vocal species, that there is some connection to how vocalizations are made as well as, why and what meaning the vocalizations have. Part of my research agenda is to better understand these connections.

- **Developmental differences within clades:** Within monophyletic groups, there is variation between different species vocal development patterns. While vocal production and comprehension is innate in some species, other species require a sensorimotor learning style and others require something in between. I am developing methods to make inferences of vocal complexity based upon a pre-existing understanding of how various species learn to meaningfully vocalize. My goal in this pursuit is to determine what environmental and genetic factors are important for developing a more complex way of communicating vocally.
- **Potential for linguistic structure:** Cognitive studies of animals have provided some insight into what certain species are capable of. My work strives to further this understanding by viewing this problem from a cross-disciplinary approach. Instead of focusing on making a direct connection to human language, I first examine what meaningful connections individuals within single species are making with conspecifics. Once these connections are explored, I then expand my view of the interactions to include individuals from other species that may come into regular contact with my focal species. My goal in this approach is to first understand what meaningful communication may be going on within a community with which the species has coevolved in before making a larger leap towards how that may relate to us.

5.8 Statement by Robert Eklund

Robert Eklund (Linköping University, SE)

Personal statement

Given a background in Speech Technology (I worked on the first concatenative speech synthesizer for Swedish, the first commercial ASR system for Swedish (now Nuance) and the first open prompt human–computer support system in Scandinavia (Telia 90 200) it has, for a long time been ”natural” for me to think in terms of interaction, and concepts like agents, avatars, Theory of Mind and interface design (auditory and visual) have all been part of parcel of my work activities during the period 1994 to (roughly) 2012.

For completely unrelated reasons I started to expand my research interests into animal vocalizations in the year 2009 when I made a recording of a cheetah purring [1] and these activities did then snowball into a five-year-long project where me and colleagues will study human–cat interaction with focus on prosody/melodic aspects [2].

My Stimulus Talk during the Dagstuhl conference did not focus on or describe my previous research on the topic (cheetah, lion and domestic cat vocalizations) but instead raised some “larger issues” concerning “cross-species” (with a wide definition of ‘species’, including robots) communication.

These will shortly be described below:

[1] http://www.youtube.com/watch?v=ZFvULxbN3NM
[2] http://meowsic.info

Personality issues

The literature is replete with studies of personality (and was crucial in e.g. how to put together submarine crews during WWII). However, such studies are not constrained to human but several studies of personality in different species of felids are also to be found (see Bibliography), partly for husbandry reasons. My issue-to-raise here is to what extent individual personalities play a role when humans interact with other species.

New form of "uncanny valley"?

In 1970 Masahiro Mori published a paper title "The Uncanny Valley" (in Japanese translation) [1] where he described a dip in the easiness with which we approach and regard humanoids. If these are completely not like us (like 1930s teddy bears or cartoon characters) we have no problem, which is also the case if there are very similar to us. However, if something is "eerily" similar to us – not completely not like us, but not completely not like us, either – we get a spooky feeling around them. My question here is whether this can occur in the auditory domain, too. If computers sound very much like machines, or whether animals respond to or signal to us, in ways that are definitely not human-like, we (obviously) have no problem. But what happens when either robots or animals start communicate with us in very human-like manners – both voice-quality and content-wise: will this created another/a new form of more abstract uncanny valley?

Was Wittgenstein right?

Wittgenstein famously stated that "if a lion could speak we would not understand him". This obviously played on the idea that the lion world is so basically different from the human world that there is no way that we could understand the lion's worldview. (Note that this argument has also been forwarded within anthropology when studying other – most often non-Western cultures.) But is this necessarily true? Although undeniably true that a lot has happened since humans lived "on the savannah", we still most likely share the same basic emotions, and are governed by them. This should, in my view, provide some solid common ground for mutual understanding.

Health effects?

To spend time with a pet, or even robots, is beneficial from a health perspective. Will this effect be enhanced by improved communication with animals or robots? Or will a potential new uncanny valley effect reverse this?

Symbol mapping?

The cheetah is particularly famous for its agonistic moan–growl–hiss–spit+paw hit sequence[3] (most felids exhibit this, minus the paw hit). How to interpret this sequence? As one agonistic sequence that qualitatively changes character as it escalates, or as four different "symbols", all with their own intrinsic meaning? The basic question is: to what extent can we use the standard linguistic toolbox when we describe animal vocalizations?

[3] http://www.youtube.com/watch?v=bBIf5g2Fp1U&amp;feature=youtu.be

Language learning?

That several species of animals are capable of language learning – and consequently also dialectal variation – has been known since Aristotle [3]. What can we learn about our own acquisition of language, phylogenetically, from the study of language learning in animals?

Role of hearing?

Animals vary a lot when it comes to hearing abilities, both frequency-wise and from a source location point of view (see Bibliography below). To what extent do we need to take other species' hearing abilities into account when trying to communicate across species? Case in point: the Beluga whale described in [2] who deliberately made an effort to vocalize outside its comfort zone when addressing humans.

Motherese?

It is well-known that humans – at least in the western world – make use of what is sometimes called "motherese" when they address infants (or small children). This speech style is characterized by en exaggerated prosody and simplified phone and word repertories. It is also known that humans use the same "trick" when addressing their pet animals. Does this have any benefits on the animal side of things, or is it simply something that we do semi-automatically for our own benefit?

Summing it all up

There are, obviously, loads of things to consider when expanding our knowledge on how animals communicate, and on how we as humans can improve our communication with those animals. Although not exactly the same, there is considerable overlap in our communication with robots (and animated agents and/or avatars) and there is no doubt in my mind that there will be vast cross-fertilization between all those fields in the future. And I hope to be part of this!

Web resources

- http://roberteklund.info
- http://ingressivespeech.info
- http://purring.info
- http://meowsic.info

References

1 Mori, Masahiro, *The Uncanny Valley.* Energy vol 7, no 4, 33–35, 1970.

2 Ridgway, Sam, Donald Carders, Michelle Jeffries & Mark Todds, *Spontaneous human speech mimicry by a cetacean.* Current Biology, vol 22, no 20, R860–R861, 2012.

3 Zirin, Ronald A., *Aristotle's Biology of Language.* Transactions of the American Philological Association, vol 110, 325–347, 1980.

5.9 Statement by Julie E. Elie

Julie E. Elie (University of California – Berkeley, US)

As humans, spoken language is central in our everyday life. We use it to exchange information, to express our emotions and to form social bonds with other human beings. The auditory system plays a fundamental role in the perception and interpretation of these communication sounds. Both in humans and animals, the auditory system parses the auditory stream coming to the ear and extracts the behaviorally relevant acoustic features of sounds, leading to the percept of meaning for communication signals. Auditory neuroscientists have obtained a relatively good model of how complex sounds are represented in the primary auditory cortex primarily in terms of their spectro-temporal features. We also know that a network of higher-level auditory and associative cortical areas is involved in processing speech in humans and communication calls in animals. However, the neural circuits and the corresponding non-linear transformations that occur between primary auditory cortical areas and cortical regions that categorize communication sounds in terms of their meaning remains unknown. One first area of knowledge that I think needs a research effort is to identify the computational steps leading from the perception of communicative sounds to the invariant representation of meaning in the brain.

Furthermore, as young humans, we don't only learn to produce speech but also learn to understand the meaning of words and other non-verbal vocal communication signals. The correct interpretation of communication signals is necessary not only for eliciting the appropriate behavioral response but also for learning the appropriate usage of the vocalization. This ability to learn the meaning of vocalizations is not restricted to humans. Young vervet monkeys, for instance, progressively refine their reaction to alarm calls, adopting progressively the right behavioral response to the nature of the predator (e.g. terrestrial or aerial) signaled in the alarm call. While the neural basis of vocal learning and plasticity has been well studied in animal models, mostly in songbirds, the role of learning and its neural underpinnings in the correct interpretation of communication signals has yet to be investigated. Previous research has demonstrated that exposition to particular sound statistics or reinforcement learning with sounds does enhance the neural representations of these behaviorally relevant sounds. However, the role of plasticity in auditory cortex during development for the correct categorization of communication signal is unknown. While the auditory extraction of some relevant behavioral information could be innately implemented in the wiring of the brain (such as the basic response to alarm vocalizations in vervet monkeys), the extraction of other informative features (such as the type of predator encoded in the alarm call) is likely learned by experience and likely rely on the maturation of the auditory cortex. As such, another area that is likely of interest is to explore the extent of innate processing for social cues in vocalizations and to describe changes in neural processing of social information as the brain matures. Knowing the maturation/learning steps in animals might help us better calibrate machines/robots that should also be able to mature/learn with the environment they are navigating in.

Finally, besides the meaning conveyed by single sound elements or sequences of signals, the rhythm with which individuals exchange information might also be informative about the vocalizer internal/emotional state, or about the urgency of the situation. Brief alarm calls repeated several times might for instance be more effective in achieving the desired behavior of rising others attention and fleeing for cover and could gradually indicate the imminence of the danger. In another line of studies, the synchrony of vocal exchanges between close related individuals seem also to serve social relationship by maintaining or straightening the bonds. Duets between paired individuals in the context of territory defense is as such most likely advertising the strength of the alliance between the mates to potential intruders. When such duets are performed in a more intimate context then the hypothesis of the pair-bond reinforcement has been proposed. However, we still don't know in terms of both physiology and information content, what are the consequences of the precise synchrony between individuals during these vocal interactions, and we are even further from understanding how this precise timing is achieved.

5.10 Statement by Sabrina Engesser

Sabrina Engesser (Universität Zürich, CH)

Vocal combinations in non-human animals

Research over the last five decades has indicated that numerous aspects of human language also exist in non-human communication systems [1]. Reference and intentionality represent two key components of language, with meaning being assigned to vocal structures, and information being voluntarily communicated [2]. Analogue forms of these components are found in various forms in non-human species. Animal vocalisations can, for example, refer to current external events or objects [3], and signals can be flexibly used by animals to inform or manipulate receivers, or equally, information can be withheld in the presence or absence of certain individuals [4]. Such strategic, flexible use of vocalisations indicates that vocalisations and the decision to call are not necessarily hardwired in animals, but individuals might have a certain degree of control over their vocal production [5]. Whilst these findings have been argued to provide insights into understanding the evolution of linguistic abilities central to language, there remains a problem with regard to language's generative nature, particularly its evolutionary origin and the selective conditions promoting its emergence [1, 6]. Theoretical work hypothesises that language's combinatorial layers evolved in order to overcome productional and perceptional limitations [7]. Specifically, stringing meaningless sounds (phonemes) together can enhance the discriminability between otherwise similar sounding signals, and hence decrease perception mistakes [7]. Once the number of messages to be encoded exceeds the number of discrete signals present in a communicative system, and in order to offset memory limitations, meaningful signals can then be assembled in a systematic way into higher order meaningful structures [7].

Empirical data on animal communication systems can help to test such hypotheses, and a broad comparative approach can provide insights into the evolutionary progression of human language's combinatorial components. In line with the comparative approach, my research investigates the prevalence and diversity of vocal combinations in two highly social passerine birds which do not sing, but instead possess an array of discrete vocalisations: southern pied babblers (Turdoides bicolor) and chestnut-crowned babblers (Pomatostomus ruficeps). Given the extensive array of behaviours that require coordination, there has likely been a significant selective pressure on both species to evolve new and diverse call types. However, like most animal species, babblers appear to be anatomically constrained in the number of different calls they can produce. Combining existing sounds and calls may therefore represent a potential mechanism applied by both species to increase the amount of information that can be encoded, facilitating the smooth management of a plethora of behaviours upon which the stability of these species' social and breeding system depend (for more information see [8, 9]).

VIHAR related statements/thoughts

While "vocal learning is thought to be a key precursor of [...] language" [10] a fundamental question arises concerning why – of the few known animal taxon possessing the ability to generate novel sounds – this capacity is primarily allocated to the creation of sound combinations devoid of conventional meaning [11], with complex structures/songs being primarily driven by female preferences for elaborate male songs, or by selection for individually recognisable signals functioning, for example, in bonding behaviour [12, 13]. Potentially the loose association between signal structure and conventional meaning has enabled the creation of ever-more complex vocal sequences. But how crucial is vocal learning for the evolution of meaningful generative capacities (i.e. rudimentary phonemic and syntactic structures)?

Whilst "the physical apparatus for articulation and audition differs from species to species" [10], it is crucial to also consider to what degree the environment a species inhabits shapes the spectral features of its vocalisations and their perception. Are species with 'fixed' vocal repertoires actually constrained in their vocal production (i.e. did a species adapt) or do they simply 'adjust' to an 'acoustic/environmental niche' with underlying vocal plasticity? How and to what degree do anatomical and environmental constraints affect the structure of vocal signals, and how does this in turn shape the emergence and the forms of combinatorial structures in non-human animals?

"Vocal interactivity is likely often teleological and is thus conditioned on underlying intentions. [...] To what extent are vocal signals teleological, and is it possible to distinguish between intentional and unintentional vocalisations?" [10]. Besides asking whether a signal is intentional or unintentional, from a receiver's perspective a signal may not necessarily have to be teleological to serve a communicative purpose. Some calls may simply encode emotional states of the caller and still transfer information triggering an evolutionary adaptive response in receivers. Concerning vocal sequences, is intentionality a prerequisite for combinatoriality? Do signals have to be purposefully combined to encode information in a compositional fashion and to be meaningful for receivers?

References

1 Hauser M. D., Chomsky N., Fitch W. T. *The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?* Science 298, 2002.
2 Hockett C. F.. *The Origin of Speech* Sci. Am. 203, 1960.
3 Townsend S. W., Manser M. B.. *Functionally referential communication in mammals: the past, present and the future* Ethology 119, 2013.
4 Tomasello M. *Origins of Human Communication* Cambridge, MA: MIT Press, 2008.
5 Marler P., Dufty A., Pickert R.. *Vocal communication in the domestic chicken: II. Is a sender sensitive to the presence and nature of a receiver?* Anim. Behav. 34, 1986.
6 Bolhuis J. J., Tattersall I., Chomsky N., Berwick R. C. *How Could Language Have Evolved?* PLoS Biol. 12, 2014.
7 Nowak M. A., Krakauer D. C. *The evolution of language* Proc. Natl. Acad. Sci. USA 96, 1999.
8 Engesser S., Crane J. M., Savage J. L., Russell A. F., Townsend S. W.. *Experimental Evidence for Phonemic Contrasts in a Nonhuman Vocal System* PLoS Biol. 13, 2015.

9 Engesser S., Ridley A. R., Townsend SW. *Meaningful call combinations and compositional processing in the southern pied babbler* Proc. Natl. Acad. Sci. USA 113, 2016.

10 Moore R. K., Marxer R., Thill S. *Vocal Interactivity in-and-between Humans, Animals, and Robots* Front. Robot. AI 3, 2016.

11 Rendall D. *Q&A: Cognitive ethology – inside the minds of other species* BMC Biol. 11, 2013.

12 Catchpole C. K. *Bird song, sexual selection and female choice* Trends Ecol. Evol. 2, 1987.

13 Janik V., Slater P. *Vocal Learning in Mammals* Adv. Stud. Behav. 26, 1997.

5.11 Statement by Sarah Hawkins

Sarah Hawkins (University of Cambridge, GB)

Sound and meaning. Researchers into human's speech perception typically rely heavily, and often entirely, on the units of formal linguistic theory as the elements that 'need identifying', unless the focus is emotion. I believe that privileging such atomistic, non-redundant units that neglect communicative function provides a distorted view of how people understand spoken language: non-redundancy is biologically implausible; information in spectrotemporal properties of the spoken signal is ignored unless it contributes directly to lexical identification and narrow sentence meaning, which leads to unlikely models of perceptual processes; and important aspects of human communication are neglected. An utterance's meaning can be quite different from the meaning of its individual words and grammar, being modulated by voice quality, facial expression, and the situation itself – cf. the range of responses that hold on or take a break/brake or even the cat's over there! invite, given different renditions and situations. Rich, subtle meaning can also be conveyed without words yet phonetically reflect the implied words (Hawkins, 2003, Table 2 erratum). So prioritizing linguistic unit identification provides an incomplete 'sterile world' analysis – it neither uses all information inherent in the multi-modal physical signal, nor guarantees a full description of the talker's meaning, nor allows for that intended meaning to be filtered through the listener's preconceptions. Instead, we need to prioritize the input, rich interpretation, and their interaction. This entails using stimuli that are recorded and responded to in contexts that demand attention to broad meaning, as well as refocussing effort onto the input signal ('below' linguistic units), and on how to represent meanings without being forced to depend on identification of intermediary linguistic units. This alternative way to conceptualize speech perception processes may connect more straightforwardly with both animal work and robotics, for when communicative function is clear, meanings can be clearly conveyed without phoneticallysegmentable units in the physical signal, and I speculate that the processes that make this possible are those that are fundamental to communication within and between species, and also to engaging with inanimate events.

I work with Polysp (POLYsystemic Speech Perception), which centres on mapping perceived properties of the physical signal to metrical and rhythmic structures appropriate for situated communication in the specific language. Details of structures must be species- and language-specific, but in general, sound chunks and associated information (visual, situational) activate competing structures to differing degrees. An attribute of a sound chunk can signify several types of structural element, and different attributes can activate one element. Strongly-activated metrical structures influence mapping by changing weights on specific sound chunks, depending on the likelihood of one meaning over another and prior knowledge of expected sound patterns in the context. E.g. English /s z m n/ vary acoustically less than /t d/ and 'th' as in this, but all are affected by grammatical status. So prediction influences how physical features are attended to, interpreted and hence mapped. When meaning is reached without identifying less-certain elements, these are 'filled in' afterwards by pattern completion processes. But if a filled-in candidate structure does not match the perceived rhythm, that structure is discarded. Distinctive (re psycholinguistics) aspects of Polysp include that relative timing is fundamental, no unit can be described independently of its context, identifying units between sound and meaning is not essential, and the distinction between 'top-down' and 'bottom-up' processes has limited value.

Though its principles have been used for text-to-speech, Polysp has not been implemented as a recognition system, and drawbacks include that it lacks well-specified high-level functional/intentional information capable of dealing efficiently with the detail, and it needs extending to account for interaction. Several perception-action robotics systems have such high-level control, but their speech models do not exploit the rich communicative information available from phonetic detail. I hope the two approaches can inform each other and hence come together.

Interaction & Generality. The claim that rhythm is fundamental to understanding an utterance leads naturally to examining interaction, with the literature suggesting temporal entrainment between musicians, and phase-locked neural oscillations between talkers. We have recent evidence of entrainment across turn boundaries in well-formed Question-Answer pairs, of seamless transfer of pulse between conversational speech and jointly-improvised music, especially when the musical rhythmic pulse is less variable, and new data tentatively suggesting that experience in predicting turn-taking during improvisational music-making and language games enhances empathy amongst teenagers, compared with just playing and (e.g.) rapping together. The questions I ask on slide 3 about interaction and generality seem to me ideally answered in a cross-disciplinary and cross-species forum. I value most highly models that are not just biologically plausible but are also as biologically general as possible. I welcome work that is cautious about making speech and language too special: true, many language attributes seem to be largely specific to humans, but each species' communication has unique aspects, and for me there is much interest in finding commonalities.

References

1 Hawkins, S. (2003) *Roles and representations of systematic fine phonetic detail in speech understanding.* Journal of Phonetics 31(3–4), 373–405. http://dx.doi.org/10.1016/j.wocn.2003.09.006. Erratum in J. Phonetics 32(2), 289.

2 Ogden, R., & Hawkins, S. (2015) *Entrainment as a basis for co-ordinated actions in speech.* The Scottish Consortium for ICPhS 2015 (Ed.), 18th International Congress of Phonetic Sciences. Univ. Glasgow; ISBN 978-0-85261-941-4. Paper number 0599.

3 Hawkins, S. (2014) *Situational influences on rhythmicity in speech, music, and their interaction.* In R. Smith, T. Rathcke, F. Cummins, K. Overy, S. Scott (eds.) Communicative Rhythms in Brain and Behaviour. London: Philosophical Transactions of the Royal Society B 369: 20130398. http://dx.doi.org/10.1098/rstb.2013.0398

5.12 Statement by Ricard Marxer

Ricard Marxer (University of Sheffield, GB)

Vocal interaction plays a fundamental role in our day-to-day relations to our environment and to others. We are capable of explaining complex ideas to others and recognising the emotional state of someone from the tone of their voice. Animals make extensive usage of vocalisations, whether to establish territory, sound an alarm or establish social bonding. Vocal signals are also central in the study and design of autonomous agents, nowadays we can perform Internet searches with spoken commands and maintain short conversations with virtual personal assistants.

The research I have conducted has mainly revolved around humans' perception and production of acoustic signals. From the study of music perception and singing voice to the modelling of speech intelligibility, the work I have done investigated several aspects of human listening [1, 2] and multimodal vocal interaction [3].

My interest in organising the VIHAR seminar comes from wanting to understand the underlying principles, commonalities and differences governing vocal interactivity in humans in animals. I'm also interested in seeing how these principles can be used to modify our interactive experiences with autonomous agents, and how these agents could be used to further explore animal behaviour.

I also think understanding the underlying principles of vocal interactivity across species could have an important impact on well-established fields. Knowledge about the formation of perceptual acoustic units throughout species could have an impact on speech recognition systems for under-resourced languages. Better understanding of the role of top-down and bottom-up processes in the perception of acoustic categories could significantly improve human speech intelligibility modelling. Insights on the role of expectations in vocal interaction in both animals and humans could change the way in which we build dialog models or interactive music systems.

In particular some of the initial questions that spark my interest are:

- Are there common processes involved in the categorical perception of acoustic units in humans and animals? What computational models could be used to reproduce such processes?
- What top-down processes are involved in the perceptions of vocal signals in animals? How are these related to those involved in human speech perception?
- What's the role of expectations in the perception of vocal signals (and sequences of them) in both humans and animals? And how may these be exploited when establishing interaction with autonomous agents?
- What principles underlie the perception/production of sequences of acoustic units in animals? How do these relate to the principles governing human vocal interaction?
- What modelling and machine learning techniques can help us understand the general principles (if any) of vocal interactivity across multiple species?
- How do multiple modalities influence the process of acoustic perception in animals? Is there an analogue to the McGurk effect in animal vocalisations?

References

1 Marxer R. & Purwins H., *Unsupervised Incremental Online Learning and Prediction of Musical Audio Signals*, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 863–874, May 2016.

2 Marxer R., Barker J., Cooke M. & Garcia Lecumberri M. L., *A corpus of noise-induced word misperceptions for English*, in J. Acoust. Soc. Am. (in press), 2016.

3 Abel A., Marxer R., Barker J., Watt R., Whitmer B., Derleth P., & Hussain A. *A. A Data Driven Approach to Audiovisual Speech Mapping*, in Advances in Proc. of BICS 2016, Beijing, China, November 28-30, 2016.

5.13 Statement by Roger K. Moore

Roger K. Moore (University of Sheffield, GB)

Since joining the Speech and Hearing Research (SPandH) group at Sheffield in 2004, I've developed (and published) a unified theory of spoken language processing called PRESENCE (PREdictive SENsorimotor Control and Emulation) which weaves together accounts from a wide variety of different disciplines concerned with the behaviour of living systems in general – many of them outside the normal realms of spoken language – and compiles them into a novel framework that is breathing life into a new generation of research into spoken language processing.

PRESENCE (first published in 2007) presents a number of practical implications with regard to new models for automatic speech recognition and generation. However, it also poses some fundamental questions about the nature of vocal interactivity – not just about speech communication between one human being and another, or between human beings and machines, but questions concerning the foundations on which all such interactive behaviours are based – questions such as:

- How do (living) systems coordinate their activities by vocal signalling?
- What is the role of prosody and emotion in establishing relations between equal/unequal social partners?
- How does mimicry and imitation facilitate learning (e.g. in development)?
- How are interaction skills acquired?
- What evolutionary constraints are implicit in such behaviours?

Contemporary approaches to spoken language interaction and dialogue (e.g. Siri – Apple's voice-enabled personal assistant) are understandably based on rather naïve models of message passing and strict turn-taking. PRESENCE, on the other hand, points to a more fluid model of interactivity based on continuous interaction between coupled dynamical systems. PRESENCE also shows how behaviours such as emotion serve to drive an organism's intentions and that 'empathy' between interacting agents facilitates signalling efficiencies. What I'm trying to do now is to go back and address some of these fundamental scientific and technical questions, and what seems to be needed is a comparative approach that is not limited to human behaviour but which encompasses computational models of vocal interactivity in and between humans, animals and robots. It is my view that progress in this interdisciplinary area will unlock key behaviours for interactive systems, and will pave the way for much more effective human-machine interfaces – especially if they involve spoken language.

I've managed to conduct some preliminary research in the area. For example, I've performed experiments using e-Puck robots interacting and vocalising using a novel general-purpose mammalian vocal synthesiser configured to generate rat-squeaks (see http://www.youtube.com/watch?v=E4aMHK7AH5M). Likewise, I've been working with Zeno (the

RoboKind humanoid robot) to investigate synchronous agent-to-agent behaviour within a PRESENCE framework. I've also been modelling the consequences of category misalignment between different modes of interactivity (visual, vocal and behavioural), which led to my 2013 Nature paper presenting the first quantitative model of the well-known 'uncanny valley' effect.

My overall aim is to demonstrate that many of the little-understood paralinguistic features exhibited in human speech (including prosody and emotion) are derived from characteristics that are shared by living systems in general. Modelling such behaviours in this wider (situated and embodied) context, using robots as an experimental platform, should eventually enable us to implement usable and effective interaction between human beings and artificial intentional agents. The research aims to address fundamental interactive behaviours such as mimicry, imitation, adaptation, learning, speaker-listener coupling, acquisition, evolution and cooperative/competitive social interaction.

References

1 Moore, R. K. *Spoken language processing: piecing together the puzzle.* Speech Communication, 49(5), 418–435, 2007

2 Moore, R. K. *PRESENCE: A human-inspired architecture for speech-based human-machine interaction.* IEEE Trans. Computers, 56(9), 1176–1188, 2007

3 Moore, R. K. *A Bayesian explanation of the "Uncanny Valley" effect and related psychological phenomena.* Nature Scientific Reports, 2(864), doi:10.1038/srep00864, 2012

5.14 Statement by Julie Oswald

Julie Oswald (University of St Andrews, GB)

Acoustic species recognition in delphinids

Dolphin species tend to be acoustically active and produce a variety of sounds. However, many acoustic recordings of dolphins do not have associated visual observations and it can be difficult to identify species in the recordings. Whistles produced by dolphins are narrowband, tonal sounds that are believed to function as social signals and carry information related to individual identity, arousal state and possibly other information such as species identity. As such, much research in recent years has focused on developing tools for classifying whistles to species. Whistle contour shapes are highly variable within species and exhibit a great deal of overlap in time-frequency characteristics when compared between species, which makes classification of these sounds challenging. There are several facets of this topic that are especially relevant in a comparative VIHAR context:

- Computational methods for classification: Many statistical and machine learning techniques have been employed, with varying levels of success, to create classifiers, including random forest analysis, Hidden Markov Models, clustering algorithms, neural networks, and others. Collaborations between computer scientists, bioacousticians, signal processors, etc. are crucial for developing the best classifiers.
- Big data: Large datasets that encompass the variability in vocal repertoires are necessary for training successful classifiers. These data can be difficult to compile, organize, share and store. What are the best practices for dealing with these big datasets?

- Between-species communication: Acoustic species identification is important for researchers, but how important is it for dolphins? Do dolphins use whistles to communicate species identity? If so, what features are they attending to for this information? These questions require collaborations among scientists from fields such as cognition, animal behaviour, bioacoustics, signal processing, and others.
- Applications from research on human speech and communication in other taxa: Lessons learned from research on communication in other taxa may give insight to inter-species communication in dolphins and acoustic species recognition in these species.

5.15 Statement by Bhiksha Raj

Bhiksha Raj (Carnegie Mellon University – Pittsburgh, US)

When to interrupt: a comparative analysis of interruption timings within collaborative communication tasks

This study seeks to determine if it is necessary for the software agent to monitor the communication channel to effectively detect appropriate times to convey information or "interrupt" the operator in a collaborative communication task between a human operator and human collaborators. There is empirical research dedicated to manipulating time on the delivery [Bailey and Konstan 2006; Czerwinkski et al. 2000b; Monk et al. 2002] of system-mediated interruptions [McCrickard et al. 2003] in multi-task environments [McFarlane and Latorella 2002]. There is also literature that explores immediate interruption or notification dissemination [Czerwinski et al. 2000a; Dabbish and Kraut 2004; Latorella, 1996] within dual-task scenarios. Studies have shown that delivering interruptions at random times can result in a decline in performance on primary tasks [Bailey & Konstan 2006; Czerwinski et al. 2000a; Kreifiedt and McCarthy 1981; Latorella, 1996; Rubinstein et al. 2001]. Additionally, studies have illustrated that interrupting users engaged in tasks has a considerable negative impact on task completion time [Cutrell et al. 2001; Czerwinski et al. 2000a, 2000b; Kreifeldt and McCarthy 1981; McFarlane 1999; and Monk et al. 2002]. Much of the current literature is focused on one user engaged in a primary task interrupted by a peripheral task.

This study differs from previous studies in that the primary task is collaboration between two or more users and the secondary task is presented to one of the collaborating users. This study explores the outcome of overall task performance and time of completion (TOC) of a task at various delivery times of periphery task interruptions. The study attempts to determine if there is a need for a system to monitor a collaborative communication channel prior to disseminating interruptions that improves efficient communication and prevents information overload within a human exchange. The study uses a simulated collaborative, goaloriented task via a dual-task where an operator participates in the primary collaborative communication task and a secondary monitoring task. User performance at various interruption timings: random, fixed, and human-determined (HD) are evaluated to determine whether an intelligent form of interrupting users is less disruptive and benefits usersóverall interaction.

There is a significant difference in task performance when HD interruptions are delivered in comparison with random and fixed timed interruption. There is a 54% overall accuracy for task performance using HD interruptions compared to 33% for fixed interruptions and

38% for random interruptions. Additionally when the TOC for the dual-task is compared across interruption types, the TOC for HD interruptions is lower than fixed and randomly timed interruptions. Although on average users complete the dual-task in less time when the communication channel is monitored, the TOC averages are close and there is no significant difference in the completion times. Results show that the use of HD interruptions results in improved task performance in comparison to fixed and randomly timed interruptions. These results are promising and provide some indication that monitoring a communication channel or adding intelligence to the interaction can be useful for the exchange.

5.16 Statement by Rita Singh

Rita Singh (Carnegie Mellon University – Pittsburgh, US)

Thoughts on human-human, human-animal and human-robot interactions

Currently, at the time of writing this report, shortly after the completion of VIHAR at Schloss Dagstuhl, my primary focus is on the application of Artificial intelligence techniques to voice forensics. Specifically, I work on profiling humans from their voice. Profiling in this context refers to the generation of a complete description of the speaker's persona from their voices. This includes the deduction of the physical appearance, medical status, demographic, sociological and other parameters of a person, and also the person's surroundings, entirely from their voice. In my recent work with the US Coast Guard Investigative Services, I have analyzed scores of hoax distress calls transmitted over national distress channels, and have provided physical descriptions of the perpetrators, of their location and their equipment sufficiently accurately to enable significant success in the investigative process. The ability to track and describe humans through their voice is useful in several disciplines of intelligence, where voice is part of the intelligence information gathered.

The relevance of my work to VIHAR is founded on the methodology I use for this work. My work builds on the fact that humans make numerous judgments about other people from their voices, such as their gender, emotional state, state of health, intelligence etc. There have been hundreds of studies on the ability of humans to make a surprisingly diverse range of judgments about other people entirely from their voices. My approach involves the utilization of AI techniques to achieve super-human capabilities that enable machines to make faster, more accurate, more abundant and deeper assessments of people from their voices. The methodology that I have developed for this is called micro-articulometry. It involves using state-of-art automatic speech recognition and audio processing technologies, to fragment voice recordings into pattern-consistent segments of very short durations, with high precision in high-noise environments. Scores of different "micro-features" are then extracted from these (processed) fragments. These are characteristics of the signal that are usually not observable or measurable by humans manually, and carry signatures of the speaker's persona, upbringing, medical conditions etc. in a manner similar in concept to DNA-biomarker encoding. Amongst other things, the list includes signatures of the physical environment in which the voice was produced and the devices and mechanisms that were used to transmit it. This derived information then feeds into relevant AI techniques designed for learning and discovery from ensembles of data. Suitable machine learning algorithms are then used to "derive" or "predict" the speaker's persona from these micro features. I hope to be able to build physically accurate holograms of humans from their voices in the future.

This methodology has direct relevance to the goals of VIHAR: especially that of enhancing the effectiveness of interactions. Auditory judgments are an aspect of human (and perhaps animal) sensory intelligence that have not been tapped into as a resource for interaction-enhancement until recently. In my interactions with the diverse community of human-human, human-robot and human-animal interactivity researchers at VIHAR, I explored the viability of using the micro-characteristics of human voice – which I also refer to as infra-sensory information, since it is often neither under voluntary control of the speaker, nor consciously perceivable by the human – to enable both robots and animals to understand humans better.

VIHAR was a tremendously enriching experience for me in some ways. From my colleagues who work on human-robot interactions, I learned about their techniques for simulating emotional intelligence in robots. I was able to easily see how my approaches in forensics could enhance those simulations by allowing for subtle reactions in robots in response to changes in the voice (and by association the physical and mental status) of the humans they interact with. My colleagues who work on the analysis and understanding of animal vocalizations in different settings, and on human-animal interactions changed my perspective of the field of interactivity in general. I now believe that while animals may not have the capability of understanding human language as humans do, they may nevertheless be able to discern changes in voice (and speech) patterns at multiple levels, and may be taught to react appropriately to them. I was able to make this hypothesis after listening to presentations, and participating in discussions about animal vocalizations with my colleagues. The vice-versa may also be possible, where humans may be able to interpret the nuances in animal behavior more meaningfully by utilizing the computational techniques that I now use in forensic profiling to enhance their ability to interpret animal sounds. I now firmly believe that AI systems may be able to revolutionize the field of human-animal interaction.

5.17 Statement by Dan Stowell

Dan Stowell (Queen Mary University of London, GB)

How do we model vocalisations in general, across hundreds of species?

Speech research has had the luxury of focusing on a single species, tailoring models for that one species' communication system. Furthermore the models developed are often tailored to a single task (e.g. speech recognition vs speaker identification). Many animal communications researchers also focus on specific species or taxa, concentrating on the aspects that are particularly salient for their questions.

We wish to model vocalisations for general-purpose multi-species machine listening. To do this – especially for sequences of vocalisations (whether intra- or inter-individual) – we need models that can capture enough of the relevant details, yet which are generic enough to be reused across widely different species.

A classic approach has been to transliterate animal sounds as sequences of symbols (ABBBABBBABABC), and to study the resulting sequences using n-grams or Markovian models. The reader might miss the discretisation issues swept under the carpet: how was the continuous sound stream segmented into units, and how were those units assigned one of a small set of labels? In a few studies, correspondence with categorical perception in the target species has been measured, but more commonly we trust a human listener or a clustering algorithm. Even where the discretisation does match up with perceptual categories, it obscures qualitative modulation (how it was said) and the fact that one signal can carry multiple meanings simultaneously. It also usually discards all timing information, while it is clear that the timing of the intervals between units is structured (even if not meaningful) in many species.

In my own work with bird sounds I have recently focused on models which properly integrate the timing of vocalisations. These are to be integrated with analyses of the content of vocal units. Even within the songbirds we have a massive variety of communication styles (tonal vs. non-tonal; simple vs. richly-structured; solos, duets).

Various basic paradigms are available. Markovian models such as HMMs and their extensions. Point processes. Deep learning such as RNNs. Is any of these paradigms appropriate, sufficient?

Modelling 'state' in a sound scene: how little can we get away with?

If we are to develop machines that can make sense of vocalisations in multi-species sound scenes, we need models that can reflect all the important aspects of a sound scene, which presumably includes some information relating to the actors in the scene. Yet the true 'internal state' of those actors may be quite different depending on their species – birds, humans, animals, robots – and we do not want to be forced to specify a highly-complex model for every species we might encounter. Do we wish to maintain a model of each 'agent' detected in a sound scene – or can we get away without it, leaving it implicit in the network of individual vocalisations affecting each other and the observer?

Assuming that we do want a model of agents, is there a minimal 'internal state' model that can be applied in many situations? Theory-of-mind approaches tend to imply a rich model of agents with beliefs, motivations, affordances. At the opposite end, a basic HMM-like model of actors could contain nothing more than a small set of unlabelled states. Is there something useful between these, e.g. information access combined with the handy circumplex model of affect?

Not all state is captured in the agents: contextual variables come in too, such as the temperature or the noise background.

Machine audition – how can it ever be as robust and flexible as human/animal audition?

However impressive recent results have been, there remains quite a gulf between the performance of any given machine and human/animal audition. Whether through inherited or learnt structure, we are able to cope with a vast array of: interfering noises (weather, traffic, television); modifications that the enviroment makes to a sound (reflections in a forest, reflections from a wall, frequencydependent attenuation, turbulence); novel sounds.

Can such robustness and flexibility be represented in computational methods economically? (e.g. without having to model general learning)

Reaching across disciplines

A practical challenge: relevant disciplines here include bioacoustics, ecology, animal behaviour, signal processing, machine learning, robotics, etc. – and to address these issues we need sustained cross-disciplinary engagement. The various disciplines often have different ideas about what can be taken for granted, which conferences to go to – and how much a conference should cost...

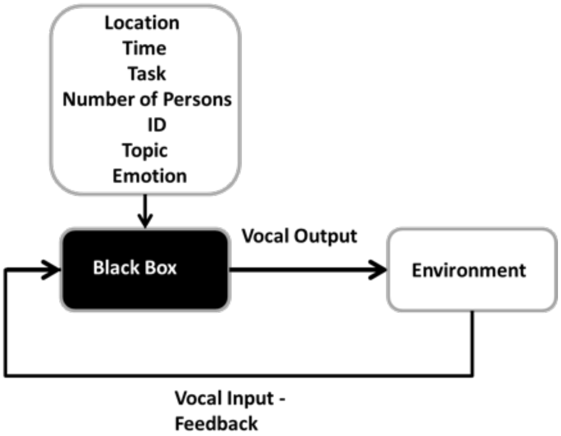
5.18 Statement by Zheng-Hua Tan

Zheng-Hua Tan (Aalborg University, DK)

Durable vocal interactive system for socially intelligent robots

A social robot should be able to interact in a meaningful way with its users and to maintain a long-term relationship with them. To meet this requirement, a durable vocal interactive system is very important, as speech interaction is perhaps the most important aspect of interpersonal communication. A vocal interactive system for robots should not only operate as a simple answer machine, but should also understand and respond accordingly to different users and different situations. For example, the robot should have memory of its users and past conversations with them in order to be able to sustain a long-term interaction. Furthermore, it should extract the scene information and also the information of the users during the conversation in order to have natural vocal interaction just like in human communication.

We aim to develop a durable vocal interactive system as described in the figure below, where we have a black box (machine learning methods) which can utilize the conversation information, take the vocal input and sensor the environment as feedback in order to give natural vocal output. This black box is capable of life-long learning, which means it can model its users, extract and remember information from their past conversations and selfadapt based on this information.



To develop such a system, several challenges arise:

1. We need to figure out how to model the vocal interaction between humans and robots, which can take all the related input into account to generate the output and also save and manage extracted information.
2. What input aspects are important for human robot interaction (HRI)? E.g., location, time, task, number of persons, identity of persons, topic of discussion, communicated emotion and etc.
3. What kind of explicit/implicit methods to use for acquiring the input aspects? Similar to reinforcement leaning, could we fuse the input aspects using reinforcement fusion?
4. Can reinforcement learning be applied to a vocal interaction system? If the answer is yes, what language, or code phrases, or key terms should be used as a reward (positive/negative feedback) for the robotic system if we use reinforcement learning?

5. How to realise life-long learning for a vocal interactive social robot, which can enable the robot being self-adapting to the environment and users?

Project iSocioBot (Durable Interaction with Socially Intelligent Robots): http://socialrobot.dk/

5.19 Statement by Serge Thill

Serge Thill (University of Skövde, SE)

Concept grounding has received much attention over the past few decades. It is arguably one of the defining aspects of embodied cognitive science as a breach with good old-fashioned computationalism (as opposed to embodied cognitive science as a continuation of American naturalism and ecological psychology as championed, for example, by Chemero). The exact degree, if any, to which a concept needs to be grounded in an agent's own experience however remains a matter of much debate. On one hand, there is plenty of evidence for the involvement of sensorimotor cortices in the processing of language (particularly words that directly relate to the sensorimotor aspect in question, see [5], for a discussion). On the other hand, one cannot deny that computational linguistics has made significant progress in machine understanding of language over the past decades, based on statistical information alone [4].

My interest is in trying to characterise what concepts are "made up from" internally. This can include direct grounding in sensorimotor experience (where I defend the idea that sensorimotor experience must be considered to encompass more than just interaction with the external world – interoception matters just as much and it is not sufficient to merely consider the perception of a given sound attached to some other (e.g. visual) input), but also statistical information, and information conveyed by others (via, for example metaphors in the Lakoff & Johnson sense). An initial stab at formulating such a characterisation – using Chris Eliasmith's *semantic pointer architecture* for a number of reasons [3] – is presented in [1].

My main reason for investigating such questions is because I am interested in social artificial cognitive systems, and what it takes to create some that are natural to interact with. In this context, there are two points that need to be made. The first is that artificial agents need to be able to understand human concepts (rather than the other way around) to be intuitive to interact with. The second is that, if theories of embodiment are right, then the inevitable differences between robot and human bodies are bound to limit this understanding since there will be limitations to the degree to which a human concept can be grounded in a robotic body (this is particularly true when we consider interoceptive aspects of concept grounding [1]).

To create social artificial agents, we must therefore find a way to overcome these differences, which implies that we need to understand how, precisely, the body may matter in concept formation. It is here that studying vocal interaction between all types of living beings – with all the commonalities and variation in embodied and sensorimotor experience this implies – as well as explicitly trying to build machines that can offer the same types of interaction despite having a different embodied experience of the concepts used will help to further the state of the art. Of particular interest from the perspective of creating social robots are the possibilities and limitations in interaction between non-conspecifics [4].

All that said, there is still much for me to learn about animal vocalisations. While it seems relatively uncontroversial to me to state that studying vocalisations across a wide range of embodied experiences will shed light onto the role that such an experience plays, the exact approach by which this potential can be unlocked remains to be explored in the context of VIHAR. Similarly, all of the above conflates vocalisation and meaningful communication. While I won't attempt to separate these here in the hope to retain a somewhat succinct statement (but see [2], for a somewhat more detailed discussion), it is very clear that this is going to substantially shape research in this direction.

References

1 Thill S., Twomey K. *What's on the inside counts: A grounded account of concept acquisition and development* in Frontiers in Psychology: Cognition, 7(402), 2016.

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

3 Thill S. *Embodied neuro-cognitive integration* in Proceedings of the Workshop on "Neural-Cognitive Integration" (NCI@KI 2015), 2015.

4 Thill S., Padó S., Ziemke T. *On the importance of a rich embodiment in the grounding of concepts: perspectives from embodied cognitive science and computational linguistics* in Topics in Cognitive Science, 6(3), p. 545–558, 2014.

5 Chersi F., Thill S., Ziemke T., Borghi A. M. *Sentence processing: linking language to motor chains* in Frontiers in Neurorobotics, 4(4), 2010.

5.20 Statement by Petra Wagner

Petra Wagner (Universität Bielefeld, DE)

What do we minimally need to communicate and how do we assess that communication is working?

As the concept of language is meaningless without assuming its being shared and used interactively by a linguistic community (Wittgenstein‘s private language argument), the investigation of communicative vocalizations does not make sense from a solipsistic, monadic perspective. Unfortunately, much work in linguistics has done exactly this and has focused on aspects of either production, perception or grammatical intuition. Among many other things, we are therefore surprisingly ignorant about the prerequisites of felicitous interactions. This deficit makes it all the more difficult to determine which aspects of linguistic or proto-linguistic communicative skills should be necessarily realized in artificial systems.

In any new communicative encounter with con-species (e.g. speaking a different language) or other interlocutors (e.g. artificial systems, animals, aliens), we need to negotiate how a process of informational grounding can be successfully implemented. To achieve this, we seem to rely on an a priori set of communicative "customs´´ which ultimately pave the way for communicative interaction, or, rather a common ground concerning how communication works. This common ground needs to be explored further and HRI provides a very useful platform for this endeavor.

I hypothesize that at least from a human perspective, this a priori common ground would have to include the following:

1. Some fundamental mechanism(s) organizing the sequentiality or simultaneity of interlocutors‘ vocalizations, e.g. by anticipating upcoming speech onsets and terminations (e.g. by respiratory cues, slowing down, falling intonation) together with some general customs for organizing the floor exchange.
2. The expression of general responsiveness and some agreement on what signals this responsiveness (feedback, attention, entrainment?)
3. An initially very coarse pool of shared signs, e.g. related to universal concepts (e.g. the frequency code, where f0 expresses size or movement direction or the effort code, where more articulatory effort expresses relevance, e.g. danger or surprise).
4. The presupposition that shared signs are flexible in the sense that they can be transferred into other system architectures (e.g. those equipped with different sound production mechanisms) or modalities (e.g. gestures).

We have worked to some degree on all these issues, trying to understand better the communicative function of subtle phonetic cues such as inhalations [2], disfluencies in (synthetic) speech [1], the success of an entrainment-based multimodal feedback mechanism in an artificial agent [3], the multimodal expression of attention [4], the usage of iconic prosody and speech-movement synchronization in a coaching scenario [5] and the flexibility of cues, extending to the domain of co-speech gestures [7, 5]. While we are still only beginning to comprehend these various complex communicative factors, we furthermore believe that we need to find novel methodological paradigms to investigate interactions both “in the wild“ and under more controlled laboratory conditions [6].

Naturally, I believe that most of my assumptions sketched above are false or at least need a lot more thinking, extension and exploration. But in order to assess the general applicability of these ideas, we need to come up with working methodological paradigms on how to properly assess whether an interaction is perceived as felicitous by the interlocutors. Unfortunately, in my opinion, we currently lack suitable online (!) approaches to evaluate HRI or ongoing human-human interactions.

References

1 Betz, S., Wagner, P., & Schlangen, D. *Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis.* Proceedings of Interspeech 2015, Dresden, 2015.

2 Cwiek, A., Neueder, S., & Wagner, P. *Investigating the communicative function of breathing and non-breathing “silent” pauses.* PundP 12 – Phonetik und Phonologie im deutschsprachigen Raum. München, 2016.

3 Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. *Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent.* In J. Epps, F. Chen, S. Oviatt, K. Mase, A. Sears, K. Jokinen, & B. Schuller (Eds.), Proceedings of the 15th International Conference on Multimodal Interaction, ICMI’13 - Sydney New York: ACM, 2013.

4 Malisz, Z., Wlodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., & Wagner, P. *The ALICO Corpus: Analysing the Active Listener.* Language Resources and Evaluation, 50(2), 2016.

5 Skutella, L. V., Süssenbach, L., Pitsch, K., & Wagner, P. *The prosody of motivation. First results from an indoor cycling scenario.* In R. Hoffmann (Ed.), Studientexte zur Sprachkommunikation: Vol. 71. Elektronische Sprachsignalverarbeitung 2014 (pp. 209–215). TUD Press, 2014.

6 Wagner, P., Trouvain, J., & Zimmerer, F. *In defense of stylistic diversity in speech research.* Journal of Phonetics, 48, 1–12, 2015.

7 Wagner, P., Malisz, Z., & Kopp, S. *Gesture and Speech in Interaction: An Overview.* Speech Communication, 57(Special Iss.), 209–232, 2014.

5.21 Statement by Benjamin Weiss

Benjamin Weiss (TU Berlin, DE)

Interpersonal Perception and Evaluation

The first (acoustic) impression results in immediate person attributions. However, which impression is dominant in the listener is still hard to grasp, as it depends on individual expectations and preferences. Several acoustic correlates of such person attributions have already been identified, but we still lack a model of relevant general (physiologically grounded) and individual, i.e. interpersonal, attributions that incorporates non-linear relationships with acoustic and/or articulatory features as well as listener properties (e.g. personality, background, voice). Here, insights from animal vocalizations might be fruitful to consider.

A subsequent evaluation of the dialog partner (e.g. on competence/benevolence and likeability), or even of the dialog (e.g. satisfaction), can only be successfully studied, if the most salient attributions of a dialog partner are known, and the situational context is taken into account. This challenge is even higher when moving from simple listening situations to interactive ones, as the attributions and attitudes towards the speaker will be reflected in conversational behavior.

When applying results from HHI to HRI, e.g. backchannel or turn-taking signals, there arise several methodological issues during evaluation.

- Whereas human observed behavior can mostly be considered as situationally adequate and congruent on multiple linguistic levels (semantic, pragmatic, para-linguistic, nonverbal), current implementations can, of course, only consider one a or few of such levels and features, which might limit validity of evaluations results. Trying to identify different communication strategies might help to reduce complexity without simplifying communication behavior too much (e.g. should a certain degree of acoustic-prosodic entrainment not also be reflected on other linguistic and gestural levels and on back-channeling behavior in a robot?).
- However, even basic communication signals in robots have been found to affect humans in real social situations. Such approaches (even popular in design and arts) seem to be very promising to study social aspects in HRI, or even HHI, than trying to build complex AI. In order to better understand und interpret such results, new methods to assess relevant interaction events are necessary, with the aim to address relevance of vocal signals outside the laboratory.
- Does making a robot more human-like by implementing vocal communication skills really results in a better user experience? The limits of such aims have not yet been explored to a satisfying degree.

References

1 Weiss, B., F. Burkhardt & M. Geier *Towards perceptual dimensions of speakers' voices: Eliciting individual descriptions"*. Proc. Workshop on Affective Social Speech Signals, Grenoble, 2013.

2 Weiss, B. & K.Schoenenberg *Conversational structures affecting auditory likeability.* Proc. Interspeech, 1791–1795l, 2014.

3 Weiss, B. & S Hillmann *Feedback Matters: Applying Dialog Act Annotation to Study Social Attractiveness in Three-Party Conversations.* 12. Joint ACL – ISO Workshop on Interoperable Semantic Annotation, Portorož, pp. 55–58, 2016.

4 Fernandez Gallardo, L. & B. Weiss *Speech Likability and Personality-based Social Relations: A Round-Robin Analysis over Communication Channels.* Proc. Interspeech. pp. 903–907, 2016.

Participants

- Andrey Anikin
Lund University, SE
- Timo Baumann
Universität Hamburg, DE
- Tony Belpaeme
University of Plymouth, GB
- Elodie Briefer
ETH Zürich, CH
- Nick Campbell
Trinity College Dublin, IE
- Fred Cummins
University College Dublin, IE
- Angela Dassow
Carthage College – Kenosha, US
- Robert Eklund
Linköping University, SE
- Julie E. Elie
University of California – Berkeley, US
- Sabrina Engesser
Universität Zürich, CH
- Sarah Hawkins
University of Cambridge, GB
- Ricard Marxer
University of Sheffield, GB
- Roger K. Moore
University of Sheffield, GB
- Julie Oswald
University of St Andrews, GB
- Bhiksha Raj
Carnegie Mellon University – Pittsburgh, US
- Rita Singh
Carnegie Mellon University – Pittsburgh, US
- Dan Stowell
Queen Mary University of London, GB
- Zheng-Hua Tan
Aalborg University, DK
- Serge Thill
University of Skövde, SE
- Petra Wagner
Universität Bielefeld, DE
- Benjamin Weiss
TU Berlin, DE