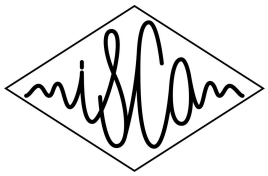


Proceedings of the
1st International Workshop on
Vocal Interactivity in-and-between Humans, Animals and Robots

VIHAR 2017

Skövde, Sweden, 25-26 August 2017

Published by:
Ricard Marxer

ISBN: 978-2-9562029-0-5

Credits:
Editors: Angela Dassow, Ricard Marxer, Roger K. Moore
Cover photo: Jarke, Skövde from city hall, Recolored by Ricard Marxer, CC BY-SA 4.0,
https://commons.wikimedia.org/wiki/File:Skövde_from_city_hall.jpg
Proceedings assembled by: Ricard Marxer

Workshop took place in Skvöde, Sweden — August 25-26, 2017

Published online at http://vihar-2017.vihar.org/ — September 14, 2017

Workshop Organisation

Organising Committee

Roger K. Moore University of Sheffield, UK

Angela Dassow Carthage College, US

Ricard Marxer University of Sheffield, UK

Benjamin Weiss Technical University of Berlin, DE

Serge Thill University of Skövde, SE

Scientific Committee

Andrey Anikin Lund University

Véronique Auberge Lab. d'Informatique de Grenoble

Timo Baumann Universität Hamburg

Tony Belpaeme Plymouth University

Elodie Briefer ETH Zürich

Nick Cambell Trinity College Dublin

Fred Cummins University College Dublin

Angela Dassow Carthage College

Robert Eklund Linköping University

Julie Elie University of California

Sabrina Engesser University of Zurich

Sarah Hawkins Cambridge University

Ricard Marxer University of Sheffield

Roger Moore University of Sheffield

Julie Oswald University of St. Andrews

Bhiksha Raj Carnegie Mellon University

Rita Singh Carnegie Mellon University

Dan Stowell Queen Mary University of London

Zheng-Hua Tan Aalborg University

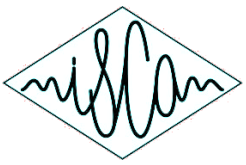
Serge Thill University of Skövde

Petra Wagner Universität Bielefeld

Benjamin Weiss TU Berlin

Sponsors

Attendance of the keynote speakers was supported by a grant from the Swedish Academy of Research.



Conference Program

Proc. 1st Intl. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), Skvöde, Sweden, 25-26 Aug 2017

Interspecies communication: a means of studying the cognitive and communicative abilities of Grey parrots

Irene Pepperberg

Abstract

Pepperberg has been studying the cognitive and communicative abilities of Grey parrots for over 40 years. She will briefly describe the history of research on avian abilities, the training techniques that she has used to establish two-way communication with parrots, and the highlights of her work with Alex that were possible because of this communication system. She will present data on her most recent research on topics such as probabilistic learning and Piagetian tasks that have been carried out with her current subjects, Griffin and Athena, showing how their intelligence compares with that of human children.

Biography

Pepperberg (S.B, MIT, '69; Ph.D., Harvard, '76) is a Research Associate and lecturer at Harvard. She has been a visiting Assistant Professor at Northwestern University, a tenured Associate Professor at the University of Arizona, a visiting Associate Professor at the MIT Media Lab and an adjunct Associate Professor at Brandeis University. She has received John Simon Guggenheim, Whitehall, Harry Frank Guggenheim, and Radcliffe Fellowships, was an alternate for the Cattell Award for Psychology, won the 2000 Selby Fellowship (Australian Academy of Sciences), the 2005 Frank Beach Award for best paper in comparative psychology, was nominated for the 2000 Weizmann, L'Oreal, and Grawemeyer Awards, the Animal Behavior Society's 2001 Quest Award and 2015 Exemplar Award, and was renominated for the 2001 L'Oreal Award and the 2017 and 2018 Grawemeyer Award. She won the 2013 Clavius Award for research from St. Johns University. Her research has been supported by the National Science Foundation (US). Her book, *The Alex Studies*, describing over 20 years of peer-reviewed experiments on Grey parrots, was favorably reviewed in publications as diverse as the New York Times and Science. Her memoir, *Alex & Me*, a New York Times bestseller, won a Christopher Award. She has published over 100 scholarly articles in peer reviewed journals and as book chapters. She is a Fellow of the Animal Behavior Society, the American Psychological Association, the American Psychological Society, the American Ornithologists' Union, AAAS, the Midwestern Psychological Society, and the Eastern Psychological Association. She serves as consulting editor for four journals and as previous associate editor for *The Journal of Comparative Psychology*.

Towards Real-time Coordination in Human-robot Interaction

Gabriel Skantze

Abstract

When humans interact and collaborate with each other, they coordinate their behaviours using verbal and non-verbal signals, expressed in the face and voice. If robots of the future should be able to engage in social interaction with humans, it is essential that they can generate and understand these behaviours. In this talk, I will give an overview of several studies that show how humans in interaction with a human-like robot make use of the same coordination signals typically found in studies on human-human interaction, and that it is possible to automatically detect and combine these cues to facilitate real-time coordination. The studies also show that humans react naturally to such signals when used by a robot, without being given any special instructions. They follow the gaze of the robot to disambiguate referring expressions, they conform when the robot selects the next speaker using gaze, and they respond naturally to subtle cues, such as gaze aversion, breathing, facial gestures and hesitation sounds.

Biography

Gabriel Skantze is an associate professor in speech technology at the Department of Speech Music and Hearing at KTH (Royal Institute of Technology), Stockholm, Sweden. He has a M.Sc. in cognitive science and a Ph.D. in speech technology. His primary research interests are in multi-modal real-time dialogue processing, speech communication, and human-robot interaction, and is currently leading several research projects in these areas. He is also co-founder of the company Furhat Robotics, which develops a social robotics platform to be used in areas such as health care, education and entertainment.

Animals, humans, computers, and aliens. Is there anything in common between all their languages?

Arik Kershenbaum

Abstract

It is often said that one of the greatest unsolved mysteries in biology is the evolution of human language. Somehow, our ancestors made a quantum leap from having no language (like all other animals), to having an infinitely complex communication medium - which no other species displays, even in part. But how can we be sure that animals have no language? What is the fundamental difference between non-human communication, and fully fledged linguistic ability? Is there some kind of "languageness" that can be quantified and measured? Some researchers claim that animal communication is nothing more than an instinctive execution of a set of neural commands. But then, at what point does autonomous computer communication become a language, rather than just a deterministic execution of commands? In the Search for Extra Terrestrial Intelligence, this question becomes crucial - would we recognise an alien language even if we heard it? Is it possible that there are languages so alien that we could never recognise them as such? In this talk, I will explore these ideas using examples from animal communication and human language (but without examples of alien language) and show how the statistical properties of these communication systems may - or may not - help distinguish language from nonsense.

Biography

Arik Kershenbaum is the Herchel Smith research fellow in Zoology at the University of Cambridge, where he researches animal communication from both theoretical and empirical angles, combining field studies with wolves and dolphins, with computer simulations of cooperative behaviour. He received his BA in Natural Sciences from Cambridge, and PhD in behavioral ecology from the University of Haifa, before going on to be a research fellow at the National Institute for Mathematical and Biological Synthesis in the USA. He has also worked developing image processing and artificial intelligence systems for the Israeli aerospace industry.

The socio-affective glue: how to manage with the empathic illusion of human for robot?

Véronique Aubergé

Abstract

Is the social robot the result of the artificial intelligence production or of the natural intelligence perception of human? One main phylogenetic feature of human is to continuously extend his body and environment competences, both cognitively and technologically. The "augmented self" paradigm has the age of human. Technologies to extend the social space of human, "to augment the others", are very old dreams, that can be drawn by the history of speech synthesis premises. However it is only recently that the social robot becomes a societal desire, without any strong hypotheses able to explain how a smart object can become perceptively a subject. In this talk we will propose how some non verbal speech primitives, collected from human sciences explorations, can dynamically manipulate the human relation with the robot by building socio-affective glue, within ethical background challenges and risks. We will explore some ecological experimental methods implying societal people, industrial partners and researchers around the users, in order to co-construct together models, smart data and technologies in the constraints of responsible research and innovations.

Biography

Véronique Aubergé is a CNRS researcher in human sciences at the LIG Lab (Computer Sciences Lab at Grenoble, France) where she heads the Domus Living Lab platform, and she is an associate Professor at the University of Grenoble-Alpes (UGA) where she directs I3L department. She heads the Chair Robo'Ethics at Grenoble National Polytechnics Institute. She has a PhD in Language Sciences and in Computer Sciences. She was a research engineer at the French Company OROS, and a researcher at ICP Lab and then at GIPSA Lab until 2012, where she developed cognitive models, experiments and applications in phonetics, prosody and expressive text-to speech synthesis. At LIG Lab, she focuses on social robotics as instruments to observe and to design models on the human interactional behaviors. She develops co-construction methods for experimenting in Living Lab some real life socio-damaged situations (elderly, children at hospital), for which the robot could be a transitory aid in ethical issues. In particular she is implied in the LIG robotic Social-Touch-RobAir platform developed within the LIG fablab, and in Emox (Awabot Inc) and Diya One (Partnering Robotics Inc.) robots.

Phonetic Characteristics of Domestic Cat Vocalisations

Susanne Schötz[1], *Joost van de Weijer*[1], *Robert Eklund*[2]

[1] Lund University, Sweden
[2] Linköping University, Sweden
susanne.schotz@med.lu.se, vdweijer@ling.lu.se, robert.eklund@liu.se

1. Introduction

The cat (*Felis catus,* Linneaus 1758) has lived around or with humans for at least 10,000 years, and is now one of the most popular pets of the world with more than 600 million individuals [1], [2]. Domestic cats have developed a more extensive, variable and complex vocal repertoire than most other members of the Carnivora, which may be explained by their social organisation, their nocturnal activity and the long period of association between mother and young [3]. Still, we know surprisingly little about the phonetic characteristics of these sounds, and about the interaction between cats and humans.

Members of the research project Melody in human–cat communication (Meowsic) investigate the prosodic characteristics of cat vocalisations as well as the communication between human and cat. The first step includes a categorisation of cat vocalisations. In the next step it will be investigated how humans perceive the vocal signals of domestic cats. This paper presents an outline of the project which has only recently started.

1.1. Previous studies

The phonetic characteristics of domestic cat vocalisations were first described by Moelk [4], and since then a number of acoustic characteristics of cat vocalisations have been described [5]–[12]. Based on previous descriptions as well as analysis of new recordings, an attempt was made to develop a comprehensive phonetic typology of domestic cat vocalisations, with phonetic definitions. Table 1 shows the number of vocalisation types and subtypes identified so far.

Table 1: *The most common domestic cat vocalisation types and subtypes identified in this study.*

Vocalisation type	Subtypes
Meow	Mew, Squeak, Moan, Meow, Trill-meow
Purr	-
Trill	Chirrup, Grunt, Trill-meow
Howl	Howl, Howl-growl
Growl	Growl, Howl-growl
Hiss	Hiss, Spit
Snarl	-
Chirp	Chirp, Chatter

Auditory as well as acoustic analyses have been used to identify and describe the different types. The descriptions include phonetic transcriptions, segmental and prosodic features, as well as typical contexts in which the vocalisations are used. These types are now used in the project for annotating and classifying cat vocalisations (see Figure 1 for an example waveform, spectrogram and fundamental frequency (F_0) contour of a vocalisation, and http://meowsic.info for additional video and audio examples.

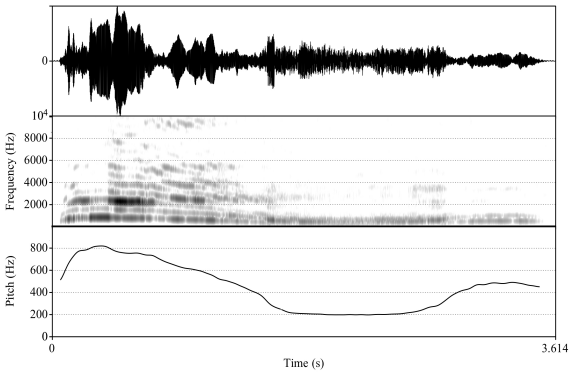


Figure 1: *Waveform (top), spectrogram (mid, bandwidth: 300 Hz) and F_0 contour (bottom) of an example howl-growl.*

2. Vocalisation types

The following list is an overview of the vocalisation types we have identified so far along with their subtypes. Example phonetic transcriptions and typical contexts in which the vocalisation types are used are provided for each type.

1. Sounds produced with the mouth closed
   a. **Purr(ing)**: a low-pitched regular and probably nasalised sound produced during alternating (pulmonic) egressive and ingressive airstream: [↓h:r̃-↑r̃:h-↓h:r̃-↑r̃:h…] or ; used when the cat is content, hungry, stressed, in pain, gives birth or is dying; probably signals ”I do not pose a threat” or ”Keep on doing what you are doing”.
   b. **trill (chirr, chirrup, grunt, murmur)**: a short and often soft, sometimes a bit harsh nasalised sound rolled on the tongue, i.e. a voiced trill: [mhr̃:], [m:r̃:ut], [bʁ̃:h]; used e.g. during friendly approach and greeting, and during play; grunts (murmurs) are usually more low-pitched, while trills or chirr(ups) are more high-pitched; sometimes cats combine a trill with a meow, producing the more complex vocalisation subtype trill-meow
2. Sounds produced with an opening-closing mouth
   a. **meow (miaow) sounds**: Meows can be assertive, plaintive, friendly, bold, welcoming, attention soliciting, demanding, or complaining, sad or even silent. A meow can be varied almost endlessly, and there are several subtypes, including the following:

i. **mew**: a high-pitched meow with [i], [ɪ] or [e] quality: [mi], [wɪ] or [mɪu]; kittens may use it to solicit attention from their mother, and adult cats may use it when they are sad or in distress or when they signal submissiveness

ii. **squeak**: raspy, nasal, high-pitched and often short mew-like call, sometimes with an [ɛ] vowel quality: [wæ], [mɛ] or [ɛʊ], sometimes not ending with a closing mouth; often used in friendly requests

iii. **moan**: with [o] or [u] vowels: [moau] or [mæu]; often used when sad or demanding

iv. **meow (miaow)**: a combination of vowels resulting in the characteristic [iau] sequence: [miau], [ɛau] or [waʊ]; often used in cat-human communication to solicit food or get past an obstacle (e.g. a closed door or window); adult cats mainly meow to humans, and seldom to other cats, so adult meow could be a post-domestication extension of mewing by kittens

b. **trill-meow (murmur-meow)**: combination of a trill (murmur) and a meow: [mrhiau], [mhrŋ-au] or [whrrrau]; used in the same contexts as the meow

c. **howl (yowl, moan, anger wail)**: long and often repeated sequences of extended vocalic sounds – often with [ɪ], [i̞], [j], [ɤ], [aʊ], [ɛɔ], [aw], [ɔɪ], [ɑo] – usually produced by gradually opening the mouth wider and closing it again; used in threatening situations, and often merged or combined with by growls in long sequences with slowly varying F_0 and intensity: [gʀ:awɪjɑoʀ:]

d. **mating cry (mating call)**: long sequences of meow-like sounds, sometimes similar to the cries of human infants; often used in spring during the mating season: [wa͡:uw], [ɪ:ɪ:a͡u:], [mhr:wa:o:u:ɪ:] or [ʀ:w:u:a:u]

3. Sounds produced with an open tense mouth are often associated with either offensive or defensive aggression, but also with prey-directed vocalisations

a. **growl (snarl):** long guttural, harsh, regularly and rapidly pulse-modulated, low-pitched sounds produced during a slow steady exhalation, often with the lip curled up and exposed teeth [gʀ:], with a vocalic [ɹ:] or rhotic [ʌ], occasionally beginning with an [m]; used to signal danger or to warn or scare off an opponent, and often intertwined or merged with howls and hisses

b. **hiss and spit** (the more intense variation): agonistic (aggressive and defensive) sounds produced with the mouth wide open and the teeth exposed, sounding a bit like long exhalations: [h:], [ɦ:], [ç:], [ʃ:] or [ʂ:]; often an involuntary reaction to being surprised by an (apparent) enemy; the cat changes position with a startle and breath is being forced rapidly through the slightly open mouth before stopping suddenly; the spit sounds similar to a hiss, but may sometimes begin with a stop – often a t-like sound: [t͡ʃ:], [t͡ʂ:], [k͡h:]

c. **snarl (scream, cry, pain shriek)**: loud, harsh and high-pitched vocalic sounds, often with [a], [æ], [aʊ] or [ɛo] vowels: [æ:ɋ]; often produced just before or during active fighting, or when in pain

d. **chirp and chatter (prey-directed sounds)**: a hunting instinct where cats copy the calls of their prey, e.g. when a bird or insect catches their attention (by making a sound) and the cat becomes riveted to the prey, and starts to chirp, tweet and chatter:

i. **chatter (teeth chattering)**: unvoiced very quick stuttering or clicking sequences of sounds with the jaws juddering, [k̟̄ k̟̄ k̟̄ k̟̄ k̟̄ k̟̄]

ii. **chirp:** voiced short calls said to be mimicking a bird or rodent chirp, sound similar to a high-pitched phone ring, tone often rises near the end, [ʔə] or reiterated [ʔɛʔɛʔɛ...]

iii. **tweet and tweedle:** tweets are soft weak chirps, often without any clear initial [ʔ] and with varying vowel qualities: [wi] or [ɦɛu]; tweedles are prolonged chirps or tweets with some voice modulation, like tremor or quaver: [ʔəɛəɥə]

Previous pilot studies have revealed that experienced human listeners are fairly good at recognizing the vocal signals of domestic cats [13], [14]. In future studies we intend to investigate this further.

3. Acknowledgements

The authors gratefully acknowledge the Marcus and Amalia Wallenberg Foundation and Lund University Humanities Lab.

4. References

[1] C. A. Driscoll, J. Clutton-Brock, A. C. Kitchener, and S. J. O'Brien, 'The Taming of the cat', *Sci. Am.*, vol. 300, pp. 68–75, 2009.

[2] D. C. Turner and P. P. G. Bateson, Eds., *The domestic cat: the biology of its behaviour*. Cambridge University Press, 2000.

[3] J. Bradshaw, R. A. Casey, and S. L. Brown, *The behaviour of the domestic cat*. CABI, 2012.

[4] M. Moelk, 'Vocalizing in the House-Cat; A Phonetic and Functional Study', *Am. J. Psychol.*, vol. 57, pp. 184–205, 1944.

[5] K. A. Brown, J. S. Buchwald, J. R. Johnson, and D. J. Mikolich, 'Vocalization in the cat and kitten', *Dev. Psychobiol.*, vol. 11, pp. 559–570, 1978.

[6] P. E. McKinley, *Cluster Analysis of the Domestic Cat's Vocal Repertoire*. University of Maryland, 1982.

[7] N. Nicastro, 'Perceptual and Acoustic Evidence for Species-Level Differences in Meow Vocalizations by Domestic Cats (Felis catus) and African Wild Cats (Felis silvestris lybica).', *J. Comp. Psychol.*, vol. 118, no. 3, pp. 287–296, 2004.

[8] S. C. Yeon *et al.*, 'Differences between vocalization evoked by social stimuli in feral cats and house cats', *Behav. Processes*, vol. 87, no. 2, pp. 183–189, 2011.

[9] S. Schötz and R. Eklund, 'A comparative acoustic analysis of purring in four cats', in *Proceedings of Fonetik 2011*, Stockholm, 2011, pp. 9–12.

[10] S. Schötz, 'A phonetic pilot study of vocalisations in three cats', in *Proceedings of Fonetik 2012*, University of Gothenburg, 2012, pp. 45–48.

[11] S. Schötz, 'A phonetic pilot study of chirp, chatter, tweet and tweedle in three domestic cats', in *Proceedings of Fonetik 2013*, Linköping University, 2013, pp. 65–68.

[12] S. Schötz, 'Agonistic Vocalisations in Domestic Cats : A Case Study', in *Working Papers*, 2015, vol. 55, pp. 85–90.

[13] S. Schötz and J. van de Weijer, 'A Study of Human Perception of Intonation in Domestic Cat Meows', in *Proceedings of Speech Prosody*, Dublin, 2014.

[14] S. Schötz, 'A pilot study of human perception of emotions from domestic cat vocalisations', in *Proceedings from Fonetik 2014*, Stockholm University, 2014, pp. 95–100.

Appropriate Voices for Artefacts: Some Key Insights

Roger K. Moore

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

r.k.moore@sheffield.ac.uk

Abstract

The 2011 release of *Siri* hailed the beginning of a sustained period of impressive advances in the capability and availability of spoken language technology. Subsequent years saw the appearance of competitors such as *Google Now*, swiftly followed by consumer products such as Amazon *Echo*. These devices are seen as the first steps towards more advanced 'conversational' artefacts (especially *robots*). However, evidence suggests that the usage of such voice-enabled devices is surprisingly low, perhaps due to noise in the environment, privacy concerns or manual alternatives.. Another possible contributing factor is that the ubiquitous deployment of *inappropriate* humanlike voices for non-living artefacts might deceive users into overestimating their capabilities, thereby creating a conflict of expectations that ultimately leads to a breakdown in communications. This paper highlights the benefits of providing an *appropriate* voice for a given artefact based on three separate studies. Results are presented that demonstrate the positive impact of a non-human voice and illustrate how 'appropriateness' might be measured objectively. Finally, a worked-example is presented of implementing an appropriate voice for the *MiRo* biomimetic robot. It is concluded that these insights could be important for the design of future generations of voice-enabled artefacts.

Index Terms: appropriate voices, robot voices, speaking artefacts

1. Introduction

After more than 40 years of research into spoken language processing, the 2011 release of *Siri* - Apple's voice-based 'personal assistant' for the iPhone - represented a significant milestone in bringing speech technology to the attention of the general public. It also hailed the beginning of a sustained period of impressive advances in the capabilities of the underlying speech technologies with dramatic improvements in the accuracy of 'automatic speech recognition' (ASR) and the quality of 'text-to-speech synthesis' (TTS). Subsequent years saw the appearance of smartphone-based competitors to *Siri* such as *Google Now* and Microsoft's *Cortana*, swiftly followed by voice-enabled consumer products such as Amazon *Echo* and Google *Home*. These latter devices are seen as the first stepping stones towards more advanced 'conversational' artefacts in the future, in particular 'automonous social agents' (such as robots) - see Fig. 1.

Notwithstanding the popularity of contemporary voice-enabled devices, it appears that actual usage is surprisingly low (see Fig. 2) [1]. Indeed, it seems that voice interfaces maintain their notoriety for "*fostering frustration and failure*" [2].

There are a number of potential explanations for this lack of genuine take-up: e.g. noise in the environment, privacy concerns or manual alternatives. However, it is argued here that another contributing factor could be the ubiquitous deployment of humanlike voices for artefacts that are clearly not human. Not only is this true of mainstream speech-based systems such as *Siri* and *Echo*, but it is also typical to find that robot research laboratories have equipped their devices with off-the-shelf humanlike speech synthesis on the basis that it's "*natural*" that people should wish to interact with a robot using 'normal' speech. The reality is that, when faced with such artefacts, users tend to be deceived into overestimating their capabilities, creating a conflict of expectations that ultimately leads to a breakdown in communications (much like the famous 'uncanny valley' in robotics [3, 4, 5]) - the opposite of what was intended.

In practice, it would be relatively easy to manage users' expectations by giving artefacts an appropriate *non-human*, rather than humanlike, voice. In principle, such an approach would avoid the pitfalls of the 'uncanny valley' by aligning an artefact's visual, vocal and behavioural *affordances* [6, 7, 8], and would create a more 'habitable' interface in line with the ideas expressed in Bruce Balentine's seminal book on the usability of

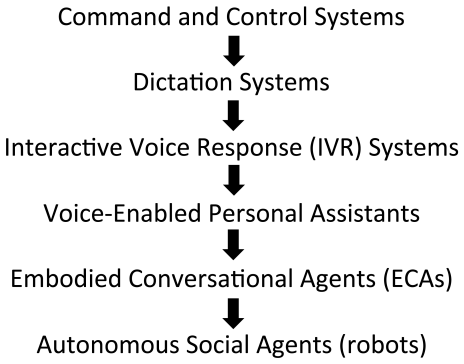


Figure 1: *The evolution of spoken language technology applications from the first 'voice command' systems of the 1970s, through contemporary smartphone-based 'personal assistants' (such as Siri) to future 'autonomous social agents' (i.e. robots).*

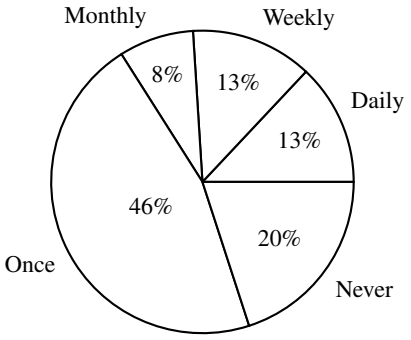


Figure 2: *Speech technology usage on smartphones [1].*

spoken language systems: "*It's Better to be a Good Machine than a Bad Person*" [9]. However, for one reason or another, deploying a robotic voice is still an unpopular idea, and mainstream speech technology R&D continues to strive for voices that as as humanlike as possible [10].

This paper brings together three separate studies which support the general hypothesis that there are benefits to be gained from providing artefacts with *appropriate* voices. Section 2 reprises experiences using a robotic voice in a genuine telephone-based travel planning service, Section 3 describes an experiment that was designed to measure vocal appropriateness, and Section 4 presents a worked-example of implementing an appropriate voice for a biomimetic robot. Finally, Section 5 concludes that this paper has brought together a number of important insights into the potential benefits and practical steps required to create appropriate voices for artefacts.

2. Experiences with a Genuine Telephone-based Travel Planning Service

The first study was conducted some years ago while the author was Head of the UK Government's *Speech Research Unit* (SRU). At the time, there was burgeoning interest in 'spoken language dialogue systems' (SLDS), and there was a need to collect corpora of speech-based transactions for study. Much of the SLDS research during that period was based on *simulated* applications, so a project was established at SRU to attempt to collect *real* conversations in a task-based dialogue - in this case, a telephone-based travel planning service.

2.1. The Setup

As is common in the SLDS research area, a 'Wizard-of-Oz' (WoZ) arrangement was used in which a human operator plays the role of all or part of a supposedly automated system. However, what was special about the SRU study was (i) the service was *genuine* (in that it was advertised with no mention that it was experimental or automated or connected with the SRU), and (ii) callers to the service were handled either by a human operator (in 'normal' mode) or by the same operator with a modified robotic-sounding voice (in 'WoZ' mode).

The enquiry service was configured around a commercially available route planning software package running on a PC. Its main feature was its ability to find the shortest and/or quickest routes between two locations in accordance with a range of user-specified preferences. Such software was not readily available to ordinary members of the public at the time. The call handling system was configured to operate with two incoming telephone lines - one assigned to the human operator's normal voice and one assigned to the robotic voice - and, in order for there to be minimal differences between the operator's behaviour in both conditions, the same operator was used in each case. The WoZ voice was created using a 'voice disguise' unit which changed the talker's pitch and then combined the natural and altered signals to produce a robotic, yet fully intelligible, vocal timbre. On receipt of a call, the operator (in normal or WoZ mode) always used the same introductory announcement: "*Welcome to the route planning service - how can I help you?*".

2.2. Results

The full results were published shortly after the study [11, 12], but the key outcome was the observation that the robotic voice had a dramatic effect on the behaviour of the callers (who, immediately upon hearing the robotic voice, genuinely believed that they had been connected to a fully automated system). The main effect was that callers in WoZ mode did not engage in lengthy social exchanges; they did not feel obliged to explain to the (apparently) automated system *why* they wanted to travel. As a consequence, WoZ-based transactions were considerably more efficient in terms of task completion. In particular, the average number of words spoken by each caller was reduced from 186 in response to the humanlike voice to just 31 for the robotic voice: an 83% reduction. Also, disfluencies were reduced by an order-of-magnitude (see Fig. 3).

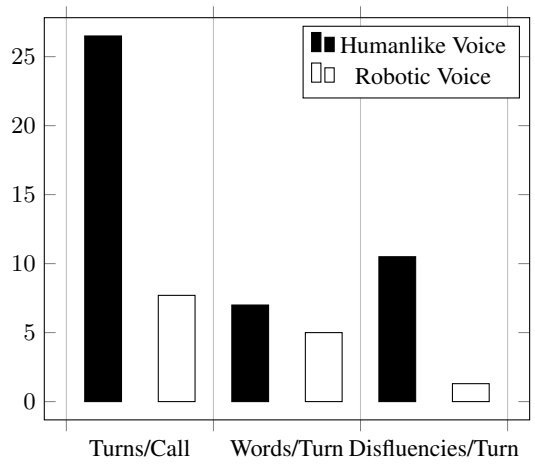


Figure 3: *The effect of the operator's voice on various measures in the telephone-based travel planning service.*

Overall, the results of this study made it clear that merely changing the timbre of a voice can have a dramatic effect on an interlocutor's interactional behaviour. In particular, an *appropriate* robotic voice can successfully reflect the limited social capabilities of an automated system, thereby facilitating efficient and successful voice-based transactions.

3. Measuring Vocal Appropriateness

The second study reported here was conducted as part of the EU-funded project *Social Engagement with Robots and Agents* (SERA). SERA was aimed at investigating the social acceptability of verbally interactive robots and agents, and it conducted long-term field trials in which a *Nabaztag* robot was placed in elderly people's homes to provide advice and encouragement about maintaining an active and healthy lifestyle.

Nabaztag is a 23 cm high WiFi-enabled highly-stylised plastic rabbit with flashing lights on its belly and nose, and rotating ears (see Fig. 4). Subjects described the robot as cute, comical and somewhat like an animation character (particularly *Pokemon*). Feedback from the initial field trials suggested that the agent must be friendly, likeable, polite and submissive, and that *its voice should be consistent with its visual appearance*.

Nabaztag's voice was generated using a state-of-the-art text-to-speech synthesiser (provided by *Loquendo*). Therefore, in order to meet the requirement that the voice should be consistent with the character of the robot, an experiment was conducted to select the most appropriate voice: the default (adult male) voice or one that was more childlike. The aim of the experiment was not simply to ask people's subjective opinions, but to attempt to measure appropriateness *objectively*.

Figure 4: *The Nabaztag robot.*

3.1. The Experiment

3.1.1. Approach distance

The first part of the experiment investigated an established measure based on 'approach distance'. Previous research on 'proxemics' had suggested that the size of the space between humans reflects (and influences) their social relationships and their attitudes to each other [13, 14]. Other studies found that inanimate objects are generally approached closer than other humans [15], and that users do not always respect a robot's interpersonal space (by moving very near to it) [16]. Hence, maintaining a proper social space between a robot and a human had been hypothesised to express the acceptance of the robot as a social actor, and that the distance was influenced by the voice [17].

3.1.2. Dislocation perception

The second part of the experiment investigated a new measure based on 'dislocation perception'. Inspired by the 'ventriloquist effect' [18], it was hypothesised that an appropriate voice for an agent could be physically displaced from an artefact and yet still be perceived as emanating from it: the more appropriate the voice, the larger the displacement. In order to test this for the different synthetic voices, the Nabaztag robot was placed in front of an acoustically transparent screen, and its voice was played through a hidden loudspeaker 29 cm to the side of the robot's 'mouth'. This meant that, at a distance of 120 cm, the voice from the robot was at an angle of approximately 12°, well over the minimum audible angle (MAA) of 1-2° [19].

3.1.3. Subjects

46 normal hearing subjects were recruited for the experiment, all of whom had with little or no prior exposure to agents/robots. Each subject was exposed to one voice only, and met the robot in a specially prepared room, with the agent approximately 3.5 metres from the entrance. Once in the room, the subject was instructed to keep eye contact with the robot, and to wait for it to invite them to come closer. When it was confirmed that they were looking at the robot, it would say: "`Hello, I've been expecting you - please come closer`". The subject then moved towards the robot, and the approach distance was noted.

The robot would then ask the researcher to offer the subject a seat, and a chair was placed directly in front (120 cm from the robot). This ensured that each subject faced the agent at approximately 0° azimuth and elevation. The robot then delivered a short speech explaining its role and purpose, finishing with: "`It was so nice of you to offer to help with this, thank you - now the researcher would like to ask you a few questions`". The researcher informed the subject that the experiment was over and led him/her away from the robot, but then casually asked: "*By the way, where did you think the voice came from - the robot or somewhere else?*", and the response was noted.

3.2. Results

The results of the 'approach distance' experiment are shown in Fig. 5. As can be seen, the majority of subjects chose to occupy the robot's 'personal space' regardless of the selected voice.

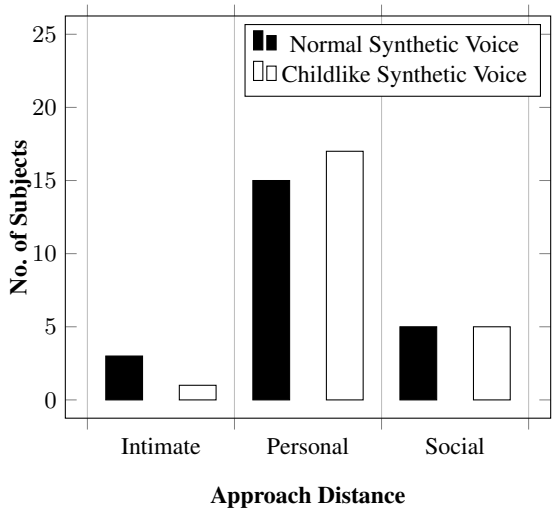


Figure 5: *The number of subjects that entered the Nabaztag robot's 'intimate space' (∼30 cm), 'personal space' (∼80 cm) or 'social space' (∼120 cm). The differences in the responses for the two voices are not statistically significant.*

The results of the 'dislocation perception' experiment are shown in Fig. 6. In this case, there is a clear (and statistically significant) difference between the subjects' responses for the two voices. As expected, the childlike synthetic voice benefitted from the 'ventriloquist effect' and was perceived by the majority of subjects to be emanating from the robot.

Overall, the results of this study suggested that, contrary to expectations, 'approach distance' is not a good objective measure of the appropriateness of a voice to an artefact, whereas 'dislocation perception' appeared to be quite effective [20].

4. A Voice for a Biomimetic Robot

The third study reported here concerns the design of a voice for *MiRo*: a highly featured, low-cost, programmable robot, with a friendly animal-like appearance, six senses, a nodding and rotating head, moveable hearing-ears, large blinking seeing-eyes, and a wagging tail. Designed and built by Consequential Robotics Ltd. in collaboration with the University of Sheffield [21], *MiRo* has been designed to look like a cartoon hybrid of a generic mammal (see Fig. 7) and is targeted at a range of applications such as assistance, companionship, pet therapy and edutainment. A unique brain-based biomimetic control system [22, 23] allows *MiRo* to behave in a life-like way: for example,

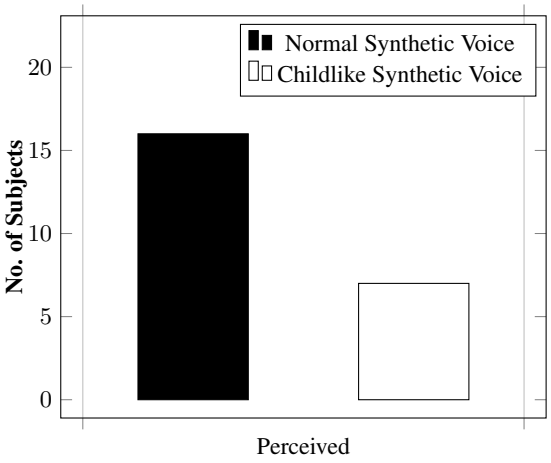


Figure 6: *The number of subjects that perceived the dislocation between the location of the Nabaztag robot and the source of its voice.*

listening for sounds and looking for movement, then approaching and responding to physical and verbal interactions.

Figure 7: *The MiRo biomimetic robot.*

4.1. The Robot

MiRo is constructed around a differential drive base and a neck with three Degrees-of-Freedom (DoF). Additional DoFs include rotation for each ear, tail droop/wag, and eyelid open/close. All DoFs are equipped with proprioceptive sensors, and there is an on-board loudspeaker. The robot is equipped with stereo cameras in the eyes, stereo microphones in the ears and a sonar range-finder in the nose. Four light-level sensors are placed at each corner of the base, and two infrared 'cliff' sensors point down from its front. Eight capacitive sensors are arrayed along the inside of the body shell and over the top and back of the head. Internal sensors include twin accelerometers, a temperature sensor and battery-level monitoring.

MiRo represents its affective state (emotion, mood and temperament) as a point in a two-dimensional space covering valence (unpleasantness-pleasantness) and arousal (calm-excited) [24, 25]. Events arising in *MiRo*'s sensorium are mapped into changes in affective state: for example, stroking *MiRo* drives valence in a positive direction, whilst striking *MiRo* on the head drives valence in a negative direction. *MiRo*'s movements are modulated by its affective state, and it also expresses itself using a set of 'social pattern generators' that drive light displays, movement of the ears, tail, eyelids and *vocalisation*.

4.2. *MiRo*'s Voice

MiRo's ability to vocalise was achieved using a real-time parametric general-purpose mammalian vocal synthesiser [26] tailored to the physical and behavioural characteristics of the robot [27]. The overall structure of the synthesis software is based on a simulation of the flow of energy through a generic mammalian vocal apparatus with an appropriate body mass.

In order to allow the injection of emotion into the vocalisations, key parameters were linked to *MiRo*'s two-dimensional affect map. Arousal modulates the airflow rate and, thereby, the amplitude and tempo of the vocalisations; high arousal leads to high airflow and short vocalisations (and *vice versa*). Valence influences the variance of the fundamental frequency and the voice quality; high valence leads to expressive vocalisation whereas low valence produces more monotonic utterances. For example, stroking *MiRo*'s head increases valence, which leads to 'happier' vocalisations (and a wagging tail).

The outcome of this design approach has been the creation of an 'appropriate' voice for *MiRo* that is perfectly aligned to the physical and behavioural affordances of the robot. It thus successfully avoids the 'uncanny valley' effect mentioned in Section 1 and contributes strongly to the effectiveness of *MiRo* as an attractive interactive vocal agent.

5. Summary and Conclusion

It has been argued that that one reason users fail to engage successfully with speech-enabled devices is the ubiquitous deployment of humanlike voices for artefacts that are clearly not human. Hence, it has been hypothesised that users' expectations could be better managed by giving artefacts an *appropriate* non-human voice, e.g. a voice that is intelligible but robotic.

This paper has brought together three separate studies which support the hypothesis. First, experiences with a genuine telephone-based travel planning service confirmed that an appropriate non-human voice can have a dramatic and beneficial effect on the behaviour of naïve users. Second, results of a study to measure vocal appropriateness *objectively* revealed that 'approach distance' was not a good measure of the appropriateness of a voice to an artefact, whereas 'dislocation perception' proved to be quite effective. Third, a worked-example has been presented of implementing an appropriate voice for a biomimetic robot.

Overall, this paper has highlighted a number of important insights into the potential benefits and practical steps required to create appropriate voices for future generations of voice-enabled artefacts.

6. Acknowledgements

This work was partially supported by the European Commission [grant numbers EU-FP6-507422, EU-FP6-034434, EU-FP7-231868 and EU-FP7-611971], and the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/I013512/1].

7. References

[1] R. K. Moore, H. Li, and S.-H. Liao, "Progress and prospects for spoken language technology: what ordinary people think," in *INTERSPEECH*, San Francisco, CA, 2016, pp. 3007–3011.

[2] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. Cambridge, MA: MIT Press, 2005.

[3] M. Mori, "Bukimi no tani (the uncanny valley)," *Energy*, vol. 7, pp. 33–35, 1970.

[4] W. J. Mitchell, K. A. Szerszen Sr., A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, "A mismatch in the human realism of face and voice produces an uncanny valley," *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[5] R. K. Moore, "A Bayesian explanation of the Uncanny Valley' effect and related psychological phenomena," *Scientific Reports*, vol. 2, no. 864, 2012. [Online]. Available: http://www.nature.com/articles/srep00864

[6] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. Shaw and J. Bransford, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1977, pp. 67–82.

[7] R. K. Moore, "Spoken language processing: Where do we go from here?" in *Your Virtual Butler, LNAI*, R. Trappl, Ed. Heidelberg: Springer, 2013, vol. 7407, pp. 111–125.

[8] ——, "From talking and listening robots to intelligent communicative machines," in *Robots That Talk and Listen*, J. Markowitz, Ed. Boston, MA: De Gruyter, 2015, ch. 12, pp. 317–335.

[9] B. Balentine, *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. Annapolis: ICMI Press, 2007.

[10] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[11] R. K. Moore and A. Morris, "Experiences collecting genuine spoken enquiries using WOZ techniques," in *5th DARPA workshop on Speech and Natural Language*, New York, 1992, pp. 61–63.

[12] R. K. Moore and S. R. Browning, "Results of an exercise to collect 'genuine' spoken enquiries using WoZ techniques," in *Institute of Acoustics Speech and Hearing Conference*, Windermere, 1992.

[13] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold, M. Durbin, M. S. Edmonson, J. L. Fischer, D. Hymes, S. T. Kimball, W. La Barre, J. E. McClellan, D. S. Marshall, G. B. Milner, H. B. Sarles, G. L. Trager, A. P. Vayda, and A. P. Vayda, "Proxemics," *Current Anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968.

[14] R. Sommer, *Personal Space. The Behavioral Basis of Design*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1969.

[15] M. J. Horowitz, D. F. Duff, and L. O. Stratton, "Body-Buffer Zone," *Archives of General Psychiatry*, vol. 11, no. 6, pp. 651–656, 1964.

[16] C. L. Breazeal, *Designing Sociable Robots*. MIT Press, 2004.

[17] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 707–712.

[18] D. Alais and D. Burr, "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology*, vol. 14, no. 3, pp. 257–262, 2004.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 2003.

[20] R. K. Moore and V. Maier, "Visual, vocal and behavioural affordances: some effects of consistency," in *5th International Conference on Cognitive Systems - CogSys 2012*, Vienna, 2012, p. 76.

[21] MiRo: The Biomimetic Robot. [Online]. Available: http://consequentialrobotics.com/miro/

[22] B. Mitchinson and T. J. Prescott, "MiRo: A robot mammal with a biomimetic brain-based control system," in *Biomimetic and Biohybrid Systems. Living Machines 2016. Lecture Notes in Computer Science*, N. Lepora, A. Mura, M. Mangan, P. Verschure, M. Desmulliez, and T. Prescott, Eds. Springer, 2016, vol. 9793, pp. 179–191.

[23] E. C. Collins, T. J. Prescott, B. Mitchinson, and S. Conran, "MiRo: a versatile biomimetic edutainment robot," in *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology - ACE '15*. Iskandar, Malaysia: ACM Press, 2015, pp. 1–4.

[24] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.

[25] E. C. Collins, T. J. Prescott, and B. Mitchinson, "Saying it with light: a pilot study of affective communication using the MiRo robot," in *Proceedings of the 4th International Conference on Biomimetic and Biohybrid Systems - Volume 9222*. Barcelona, Spain: Springer-Verlag New York, Inc., 2015, pp. 243–255.

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

[27] R. K. Moore and B. Mitchinson, "A biomimetic vocalisation system for MiRo," in *Living Machines 2017*, Stanford, CA, 2017. [Online]. Available: http://arxiv.org/abs/1705.05472

Multimodal breathiness in interaction: from breathy voice quality to global breathy "body behavior quality"

Liliya Tsvetanova, Véronique Aubergé, Yuko Sasa

LIG-lab, CNRS UMR 5217, Grenoble Alpes University, France
{Firstname}.{Lastname}@univ-grenoble-alpes.fr

Abstract

The face-to-face interaction is a complex dynamic process in which the interactants are mutually, continuously and reciprocally sharing with each other vocal and non-vocal information. This multimodal information has been defined as relevant to the speakers' interpersonal intimacy degree in the relation process ("socio-affective glue"). Otherwise, the human multimodal intimacy cues dynamics is a crucial aspect of the analysis of the human socio-affective behavior. To address this point, our study is based on the EEE corpus involving spontaneous dialogs between an elderly and a smart home control robot whose vocalizations are primitive pure prosodic expressions. The gradually increasing socio-affective gluing (intimacy) effect of the robot's vocalizations has been shown in previous studies. The utterances produced by the subjects are imposed commands (smart home orders). However, the elderly's vocal and non-vocal behavior changes gradually throughout the experiment by varying in breathiness, commands paraphrasing, and non-vocal cues, which we suppose could be meaningful for the nature of the human-robot relation. In this study, the communication dynamics is observed as an overall behavior. Accordingly, we suggest that the breathiness dynamics could be superposed, in the dialog time, to the dynamics of both the morpho-lexico-syntactical style and the proxemic cues (as postural proximity and gaze direction).

Index Terms: breathiness, socio-affective "glue," multimodal interaction, Human-Robot Interaction (HRI)

1. Introduction

Nowadays, one of the crucial aspects in the field of human-robot interaction is to provide the robot with acceptable and ethical interactional behavior so that it could be a part of the human social environment. In order to address this point, one of the approaches used in the research field is to focus on the robot recognition (detection and analysis) of the human speaker vocal and non-vocal behavioral cues observed during the interaction process (see [1] for an overview).

In terms of vocal behavior, it is well known that an important quantity of information about the speakers' affective state is likely to be discerned throughout the one's vocalizations and this phenomenon is observed in both human and animal species [2]–[4]. Accordingly, one utterance (even as small as the burst "eh") could be vocalized differently to express quite a different meaning [5], and so, quite a different affective state [6]. Acoustically, this kind of subtle affective vocal information is given by the expression style, namely the speech prosody [2]. Moreover, regarding social relations, the prosodic cues have been reported in the literature as an important vocal aspect indicating the speaker's attitude toward his or her interactional partner [5]. Thus, the prosody has been shown as a relevant factor in the establishment of an interpersonal connection between the interactants and this process of connectedness, which is called socio-affective "glue" [7], is based on an altruistic bond built according to the principles of mutual social grooming [8]. Nonetheless, the affective function of the prosodic expressiveness is unlikely to be related to its phonetic parameters [9], [10], but to a 4th prosodic dimension known as voice quality [11]. According to the findings, the voice quality refers to a specific folds' functioning/vibration (more or less folds openness than for modal speech), which seems to be led by the speaker affective state [12] and is reported as having social signaling functions [9]. One particular voice quality (see [13] for an overview) – the breathy voice quality (also known as breathy phonation) – is reported in the literature as the vocal manifestation of the interpersonal *intimacy* and *caring* [12]–[17], namely the highest degrees of altruistic gluing between the interactants.

According to the literature, a variety of non-vocal behaviors are likely to indicate heightened involvement in the interaction process. These behaviors have been studied in the field of proxemics [18], according to which interactional partners in a close relationship tend to unconsciously show their "closeness" physically during the interaction process. The proxemics is related to the notion of physical intimacy (also known as physical closeness or physical distance) [19], and it is observed throughout some non-vocal cues such as the postural proximity [20] and the gaze direction [21]. Moreover, recent studies suggest that the hand gesture could also be related to the relational gluing process [22], [23]. Thereby, according to the proxemic analysis, the whole body appears as an instrument able to express the intimacy level or, as we will refer to it in this paper, the "glue level."

The cited studies state that both vocal and non-vocal behaviors could reflect a kind of intimacy between the interactional partners. Thus, it seems that in the case of intimacy, the interactional partners create together a mental resonation [19], which could merge into a more global process as the interpersonal synchrony [24], [25], during which the speakers create an interactional dynamics resulting in temporal coordination of their behaviors. However, the dynamics (whether individual or interpersonal) of the observed interactional modalities is rarely taken into account in the context of interaction between a human and a robot [1], [26].

The purpose of this preliminary study based on the observation of spontaneous dialogs between elderly and a robot is to show that the dynamics of the vocal breathiness (a tangible intimacy indicator) is related in the dialog time to the dynamics of the glue building. Moreover, as the prosody dynamics could be transmitted and perceived throughout some non-vocal cues such as hands and face gesture [27], [28], we suppose that the

breathiness dynamics could be directly related to the dynamic changes in other vocal and non-vocal intimacy cues such as the linguistic form of the commands addressed to a robot, the postural proximity and the gaze direction. The affective communication could thus be considered as more global behavior characterized by the multimodal breathy dynamics.

2. Background: EEE spontaneous dialogs data corpus

The data used in this study are from the EEE (Elderly Emox Expressions) corpus [29] involving spontaneous dialogs between the non-anthropomorphic robot Emox (Awabot company) and socio-isolated elderly.

1.1. Study contextualization

In order to observe how the altruistic relation emerges in a micro-social system as elderly, a set of primitive sounds supposed to remain a fundamental tool for the mutual building of the socio-affective glue has been implemented in the Emox robot. Those vocalizations are gradually ordered according to their supposed gluing straight (see [29] for more details) in the following range: (1) no speech, (2) pure prosodic mouth noises, (3) interjections and lexicons with supposed gluing prosody, and (4) some subjects' commands imitations with supposed glue prosody, knowing that the imitation, or the so-called chameleon effect [30], has a high potential of establishing relationships. Those vocalizations notified as the lifeblood of the intimacy establishment between interactional partners were observed in a dialog context with socio-isolated elderly for whom the creation of this dynamic relational process seems to be more difficult [31]. In fact, the rate of social isolation increases with aging [32] and with the absence of intimate interactions [19], which seriously affects the elderly's communication skills, which are an essential part of creating and preserving the elderly's social relationships. In other words, the elderly's abilities to create the relational gluing process are damaged [29].

1.2. Collecting ecological data with a Wizard of Oz experimental scenario and "glue level" retrieval

The natural elderly-robot dialogs corpus has been collected using an experimental Wizard of Oz scenario (see [29] for more detailed information), which took place *in situ* in the Living Lab Domus (Computer Sciences Laboratory of Grenoble, France) arranged as a smart-home prototype. In this study, elderly subjects were invited to use a smart home control robot to carry out an imposed list of 31 home automation commands. The experiment was followed by an auto-annotation session, which took place a few weeks later. During this session, each participant was prodded to review, in order to involve his or her autobiographical memory [33], the whole experiment's video recording and to qualify their mental state at every moment of the interaction with the robot. The auto-annotation session aims are multifold, as follows: (1) to define the socio-affective glue value by the participant himself / herself while avoiding a possible incorrect "expert" interpretation of the collected data, (2) to observe the glue global and progressive transformation through the interaction process, and also (3) to detect the breaking points determining the border lines of each glue level.

Regarding the collected data, the EEE corpus comprises a video and audio data captured by the six ceiling cameras and six ceiling microphones in Domus; the participants were also equipped with a headset microphone allowing high-quality speech sounds collection. Concerning the headset microphone use, to avoid every suspiciousness about the fact that speech data are recorded, we let the participants think that the Emox' sound capturing sensor was broken and the only way to give commands to the home control robot was by using the robot's microphone. The auto-annotation session was also audio recorded. All the captured data were temporarily aligned in the ELAN software program, and all the speech was orthographically transcribed.

3. Methods

For this study, we choose from the EEE spontaneous dialogs corpus the multimodal data of five elderly subjects, all women and French native speakers, from 69 to 89 years old and with none or low dependency (corresponding respectively to the 6th and 5th grades according to the French elderly dependency scale AGGIR [34] grouping the elderly from 1 – very dependent to 6 – no dependent). The data represent a total of 226 minutes (approximately 3.8h) of audio and video records with full speech transcription, and a total of 340 commands addressed to the robot.

The audio record from the headset microphone was used to analyze the breathiness level of all vocal commands extracted from the selected corpus. A number of solutions for automatic calculation of the breathiness exist, and their functioning is based on different acoustic properties reported in the literature, such as H1-A3 (difference between the amplitudes of the first harmonic and the third formant) [35], NAQ (Normalized Amplitude Quotient of the glottal waveform and its derivative waveform) [36], HNR (Harmonics-to-Noise Ratio) [37] or the inverse NHR (Noise-to-Harmonics ratio) [38], GNE (Glottal-to-Noise Excitation ratio, which needs inverse filtering to avoid the problem of high-pitched voices) [39], F-aperiodic (boundary frequency between harmonic and aperiodic components) [40] and even F1F3syn (synchronization of the amplitude envelopes of the first and third formant frequency bands) [41]. However, those measures either could not be used for spontaneous speech analysis or are not adapted for elderly high-pitched female voices, which are known as naturally breathier (due to the muscular slackening, which increases with aging [42]). For that reason, in this study, we proceed by an expert labeling of the voice quality (with emphasis on the breathiness level) during the vocal production of the commands.

The elderly's non-vocal behavior has been analyzed on video recordings, focusing on proxemics cues as posture (subject's body position), physical proximity (relative to the Emox' position) and gaze direction (head and eyes cast direction). These modalities have been annotated according to a list of labels (cf. Table 1 below) and only those performed during the command time have been considered in our analysis.

Table 1: *List of labels used to annotate the subject's posture, proximity to the robot and gaze direction.*

Modality	**Labels**
Posture	Standing
	Sitting
	Crouching down
	Laying on the bed
	Leaned forward
Physical Proximity	In other room

	In the same room Close (50cm) Close+ (25cm) Close++ (touch)
Gaze direction	Emox Commands list Object concerned by the command Object that is different from the object concerned by the command Human interlocutor

4. Results

Global analysis of the data from elderly, concerning the voice quality during the commands announcement, showed that the commands are produced in either modal or breathy voice. However, the observation of the breathiness dynamics throughout the experiment revealed that each vocal command could be seen as lying along a continuum of breathiness. On this continuum, the command voice quality varies globally from modal tense (no breathiness) to breathy lax (high breathiness level). Specifically, the modal voice (voice without breathiness) is associated with the lowest glue levels (when the relation between the elderly and the robot is not yet established), and the breathy voice is associated with the highest levels of socio-affective glue (at the end of the experiment when the robot's vocalizations are the most charged in glue). The breathiness scale varies in accordance (or even in response) to the robot's gluing vocalizations. In this way, the breathiness dynamics is following, in a progressive fashion, the socio-affective glue dynamics as it is illustrated in Figure 1 below.

An analysis consolidating the subjects' multimodal behavioral data compared to both the glue level and the previous breathiness observations showed some general tendencies in the variations of the linguistic style and the proxemic modalities. In accordance with the glue potency variation, every separate modality seems to evolve gradually on its continuum as shown in Figure 1 below. Figure 1 also illustrates the moment of emergence of some proxemic cues, which coincided with the first robot's vocalizations and the appearance of the vocal breathiness in the commands. The emergence of the cues revealing lower physical and vocal distance was the beginning of the closeness manifestation in the other modalities. At this moment, the linguistic form started to change from the imposed infinitive form into a "we" form manifesting a kind of togetherness or "we-ness," and then, into an "you" form that seems to be a characterization of the robot like a different entity (which is confirmed by the auto-annotations noting that at this moment, the robot is "like a child," "like another"). This temporal boundary line is also the beginning of the proxemic cues in a manner, which also showed lower distancing. Thus, postural proximity and gaze direction seem to be dynamically interrelated: the increase of the physical distance (decrease of the physical closeness) increases the gaze occurrences.

However, a more detailed data analysis of the behavior of every elderly person showed that the glue obtained from the auto-annotations arranges the subjects in three distinctive groups: (a) those who did not glue (one subject), (b) those who moderately glued (two subjects) and (c) those who glued strongly (two subjects). Moreover, the different gluing type seems to induce different dynamics variations in terms of linguistic behavior (breathiness and morpho-lexico-syntactical style) and proxemic behavior (postural proximity and gaze direction). By taking into account all cited modalities, the three profiles could be summarized as follows:

(a) *Low gluing profile*: the majority of the produced commands are in modal (tense or lax) voice quality (91%), and there are no commands emitted with either breathy or breathy lax voice. The infinitive structure of the command is maintained throughout the whole experiment. In terms of proxemic modalities, the subject maintains a close distance to the robot (labeled as "close"), but she is rarely looking in the direction of the robot; her gaze is more often oriented in the direction of the list of commands and the object concerned by the command.

(b) *Medium gluing profile:* these subjects' voice quality varies from modal tense to breathy lax, with a high percentage of modal lax and breathy tense commands (mean value of 52%). There are very few commands forms modifications. The physical distance with the robot decreased progressively from "close" at the begging through "close/close+" to "close+" at the end of the experiment when the gluing level is the highest. The gaze direction also changes in a similar way: at the experiment's beginning the preferred gaze targets are the list of commands, the object of the command, and Emox, while at the end of the experiment, the preferred gaze targets are the command object and the environment.

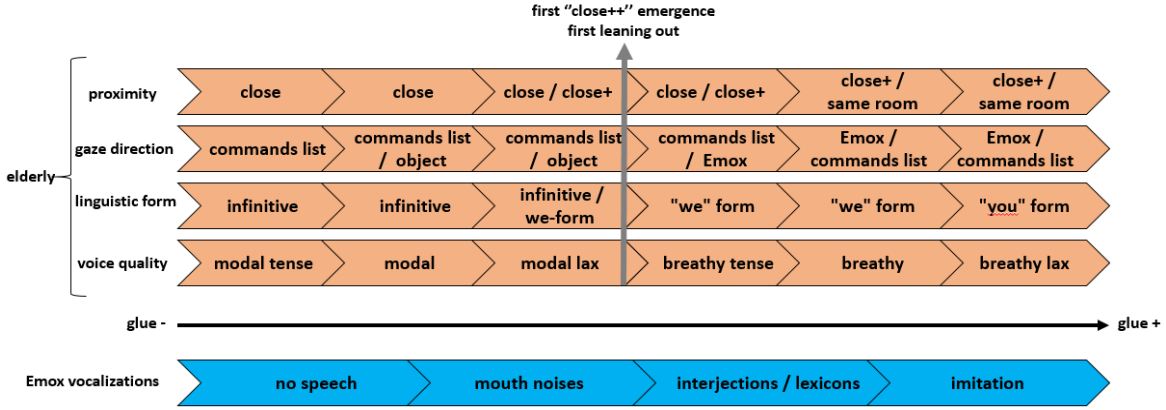


Figure 1: *The subjects' multimodal behavior variation in accordance with the robot's vocalizations and the socio-affective "glue level" (experiment's time)*

(c) *High gluing profile*: the voice quality variation begins from modal to breathy lax, with the highest percentage of breathy and breathy lax commands (on average 60%). The commands form varies accordingly to the perceived breathiness in a progressive manner from the infinitive to a form with the pronoun "we" (e.g. "We turn on the lights" / "On allume la lumière") and then into a form with the pronoun "you" (e.g. "You turn on the lights" / "Tu allumes la lumière"). The proxemic behavior is labeled as "close+" all along the experiment, the first noticeable closeness appears with the "close++" labels in accordance with some body leanings forward in the direction of the robot. The modal voice quality is associated to highly frequent glances in the direction of the list of commands, and the breathy voice quality is associated to glances in the direction of Emox.

5. Discussion

In this study, we tried to investigate the interactional dynamics in an elderly-robot interaction context. First of all, we supposed that as the breathiness is a proved intimacy vector, its dynamics follows the socio-affective glue dynamics. Secondly, we tried to show that the breathiness dynamics could characterize the rhythm of variation (or the dynamics) of the overall elderly's multimodal behavior. In an effort to check those affirmations, we analyzed the elderly's breathiness, linguistic styles and proxemic (postural proximity and gaze direction) behavior within the context of spontaneous interaction between the elderly and a butler robot implemented with socio-affective glue vocalizations.

As shown above, the point in the dialog time of the emergence of the first gluing vocalizations of the robot is corresponding to the beginning of the changes in the elderly's vocal and non-vocal behavior dynamics. On the one hand, the dynamics in the robot's vocalizations seems to influence the human vocal expression in a progressive way. Thus, more the robot's sounds are charged in glue (intimacy), more the elderly voice is breathier and more the level of reported glue is higher. A possible explanation of the dynamic modification of the human vocal behavior according to the robot's vocalizations could be the usual process of synchrony, evolving in the interaction process in whom the speakers share an important degree of intimacy. On the other hand, the analysis of both the breathiness dynamics (as socio-affective glue indicator) and the dynamics of the other modalities (command form, postural proximity, and gaze direction) reveals that there is no a complete temporal correspondence in the modalities changes. So, the granularity of our analysis allows us only to refer to the global tendencies but does not allow us to observe if two (or more) cues tend to change together dynamically. However, we could affirm that all the human modalities change in a manner, which shows vocally and non-vocally a general tendency to go towards more "close" (intimate) behavior.

The reported differences between the groups of subjects indicate that every subject reaches a different glue level, and not all the participants reach the highest glue level. The literature in the field of human-robot interaction often tends to explain the difference in human behavior by the personality traits, but we think that in our study the explanation is quite different. Our experience with the five elderly subjects lets us suppose that this difference in the elderly's gluing behavior could be explained by the "degree" of social isolation. Thus, the profile of elderly who glue less with the robot corresponds to the elderly who are less isolated.

Knowing that the elderly subjects who glued the most in the experiment modify the form of the commands from the imposed list of home automation commands as reported above, we expected to find a decreasing number of glances at the list of commands when the form is modified. Surprisingly, as shown by the results, the occurrences of gazes in the direction of the list remain high, even when the commands are in "we" form and "you" form. Even if the elderly continue to use the list of commands, it seems that this gaze behavior could be explained by a phenomenon of cognitive detachment of the list. This finding suggests that the speech recognition systems (which nowadays are based on lists of commands) have to take into account the state of the established relationship between the interactants. The first works in this direction have been implemented in the robot's dialog system called SARSI (Socio-Affective Robotics Speech Interaction), which is constructed accordingly to the socio-affective glue paradigm.

6. Conclusions

The acoustic breathiness dynamics is in high accordance with both the relation dynamics (the "glue life") and the global multi-dimensional elderly behavior. Thus, we observed the breathiness not only as a vocal quality but overall as a global interactional behavior quality, which could be seen as an indicator of the relation nature between the interactants. Some previous works point out the existence of body prosody as more holistic interactional (verbal and non-verbal) behavior, which is essential for the interaction. In this study, we referred to the intimate prosodic dimension – the breathiness – in order to show that its dynamics could be observed in the dynamic variation of the other intimate cues. This overall dynamics is important, not only for the interaction process but moreover, for the affective, relation process between interactants. In this case, the intimacy (namely the socio-affective glue) appears as a cognitive motor (as stated in the literature) for the discernable global breathiness behavior, which is materialized in what is said, how it is said and how it is shown by the proxemic cues as the postural proximity and the gaze direction.

7. Acknowledgements

This work was partially funded by French grants Interabot, parts of BGLE no2 Investissements d'Avenir and it has been partially supported by the LabEx PERSYVAL-Lab (ANR- 11-LABX-0025-01) and the Major Program for the NSSF of China (13&ZD189). We would like to thank the Awabot Company (robotics), Bien A la Maison Company (caregiving services) and Roger Meffreys elderly housing for their collaboration in this study. Thank you to Romain Magnani, Natacha Borel and Ambre Davat for their support.

8. References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1, pp. 227–256, 2003.

[3] R. M. Seyfarth and D. L. Cheney, "Meaning and emotion in animal vocalizations," *Ann. N. Y. Acad. Sci.*, vol. 1000, no. 1, pp. 32–55, 2003.

[4] P. Pongrácz, C. Molnár, and Á. Miklósi, "Acoustic parameters of dog barks carry emotional information for humans," *Appl. Anim. Behav. Sci.*, vol. 100, no. 3, pp. 228–240, 2006.

[5] N. Campbell, "Perception of affect in speech-towards an automatic processing of paralinguistic information in spoken conversation.," in *INTERSPEECH*, 2004.

[6] M. Schröder, "Experimental study of affect bursts," *Speech Commun.*, vol. 40, no. 1, pp. 99–116, 2003.

[7] V. Auberge, Y. Sasa, T. Robert, N. Bonnefond, and B. Meillon, "Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot," in *WASSS 2013*, Grenoble, France, 2013.

[8] H. Nelson and G. Geher, "Mutual grooming in human dyadic relationships: an ethological perspective," *Curr. Psychol.*, vol. 26, no. 2, pp. 121–140, 2007.

[9] K. R. Scherer, "Vocal affect expression: a review and a model for future research.," *Psychol. Bull.*, vol. 99, no. 2, p. 143, 1986.

[10] T. Johnstone and K. Scherer, "The effects of emotion on voice quality," in *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco: University of California, Berkeley, 1999, pp. 2029–2032.

[11] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, Barcelona, Spain, 2003, pp. 2417–2420.

[12] C. Gobl and A. Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication 40 (1)*, pp. 189–212, 2003.

[13] J. Laver, "The phonetic description of voice quality," *Camb. Stud. Linguist. Lond.*, vol. 31, pp. 1–186, 1980.

[14] C. R. Rogers, *On becoming a person: A therapist's view of psychology*. Boston: Houghton Mifflin, 1961.

[15] A. Wichmann, "Attitudinal intonation and the inferential process," in *Speech Prosody 2002, International Conference*, 2002.

[16] N. Campbell, "On the use of nonverbal speech sounds in human communication," in *Verbal and nonverbal communication behaviours*, Springer, 2007, pp. 117–128.

[17] N. Audibert, V. Aubergé, and A. Rilliard, "When is the emotional information? A gating experiment for gradient and contours cues," in *Proceedings of ICPhS XVI Meeting. Saarbrucken*, 2007, pp. 6–10.

[18] E. T. Hall, "The hidden dimension," 1966.

[19] H. T. Reis and P. Shaver, "Intimacy as an interpersonal process," *Handb. Pers. Relatsh.*, vol. 24, no. 3, pp. 367–389, 1988.

[20] S. E. Scherer and M. R. Schiff, "Perceived intimacy, physical distance and eye contact," *Percept. Mot. Skills*, vol. 36, no. 3, pp. 835–841, 1973.

[21] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.

[22] L. Guillaume *et al.*, "HRI in an ecological dynamic experiment: the GEE corpus based approach for the Emox robot," presented at the IEEE International Workshop on Advanced Robotics and its SOcial impacts (ARSO), Lyon, France, 2015.

[23] M. Girard-Rivier *et al.*, "Ecological Gestures for HRI: the GEE Corpus," 2016. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2016/pdf/1040_Paper.pdf. [Accessed: 25-Sep-2016].

[24] E. Delaherche and M. Chetouani, "Multimodal Coordination: Exploring Relevant Features and Measures," in *Proceedings of the 2nd International Workshop on Social Signal Processing*, 2010, pp. 47–52.

[25] O. Weisman, E. Delaherche, M. Rondeau, M. Chetouani, D. Cohen, and R. Feldman, "Oxytocin shapes parental motion during father-infant interaction," *Biology letters*, 2013.

[26] L.-P. Morency, "Modeling human communication dynamics," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 112–116, 2010.

[27] D. Brentari and L. Crossley, "Prosody on the hands and face: Evidence from American Sign Language," *Sign Lang. Linguist.*, vol. 5, no. 2, pp. 105–130, 2002.

[28] Guellaï, Bahia, Langus Alan, and Nespor Marina, "Prosody in the hands of the speaker," *Frontiers in Psychology*, 2014.

[29] V. Aubergé *et al.*, "The EEE corpus: socio-affective' glue' cues in elderly-robot interactions in a Smart Home with the EmOz platform," in *5th International Workshop on EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA*, Reykjavik, Iceland, 2014.

[30] T. Chartrand and J. Bargh, "The chameleon effect: the perception-behavior link and social interaction," *J. Pers. Soc. Psychol.*, vol. 76, no. 6, pp. 893–910, 1999.

[31] K. A. Bayles, A. W. Kaszniak, and C. K. Tomoeda, *Communication and cognition in normal aging and dementia*. College-Hill Press/Little, Brown & Co, 1987.

[32] J.-L. Pan Ké Shon, "Isolement relationnel et mal-être," 2003.

[33] H. L. Williams, M. A. Conway, and G. Cohen, "Autobiographical memory," *Mem. Real World*, p. 21, 2007.

[34] J. Belmin *et al.*, "Level of dependency: a simple marker associated with mortality during the 2003 heatwave among French dependent elderly people living in the community or in institutions," *Age Ageing*, pp. 298–303, 2007.

[35] H. M. Hanson, "Glottal characteristics of female speakers.," 1995.

[36] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Commun.*, vol. 18, no. 2, pp. 131–138, 1996.

[37] C. T. Ferrand, "Harmonics-to-noise ratio: an index of vocal aging," *J. Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[38] D. Deliyski, "Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications," *Educ. Gerontol.*, vol. 27, no. 2, pp. 159–168, 2001.

[39] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acust. United Acust.*, vol. 83, no. 4, pp. 700–706, 1997.

[40] T. Ohtsuka and H. Kasuya, "Aperiodicity control in ARX-based speech analysis-synthesis method.," in *INTERSPEECH*, 2001, pp. 2267–2270.

[41] C. T. Ishi, "A new acoustic measure for aspiration noise detection.," in *INTERSPEECH*, 2004.

[42] H. Hollien, "'Old voices': What do we really know about them?," *J. Voice*, vol. 1, no. 1, pp. 2–17, 1987.

Bases of Empathic Animism Illusion: audio-visual perception of an object devoted to becoming perceived as a subject for HRI

Romain Magnani, Véronique Aubergé, Clarisse Bayol, Yuko Sasa

LIG, CNRS, UMR 5217, University of Grenoble Alps, France

{name.surname}@univ-grenoble-alpes.fr, Clarisse-bayol@live.fr

Abstract

Ethically, connected objects, in particular, a social robot appearance implies to know what are the effects of its design. Western societies seem to be about embracing robots' advent in the daily living, so the Human Robot Interaction field starts facing the phenomenon we call the "empathic illusion". It supposed to appear while an object is able to switch its socio-cognitive treatment by humans by becoming a subject to them. This phenomenon may change those specific objects' status within the human social sphere. This paper's aim is to start exploring the audio-visual design dimension which influence our first impressions on an object. These dimensions combinations are expected to help better understand the robots design's effects on humans.

Index Terms: social robotic, HRI, socio-affective prosody, perception test, empathic illusion, emotional induction, design

1. Introduction

Before being perceived as animated, the social robot inherit of the audio-visual (emotional, intentional, social and cultural) attributes naturally given to the things [1]. One of the challenges of HRI concerns the models allowing to handle the attachment phenomenon in human-robot interaction which is initiated by these emotional inductions [2][3][4][5]. This attachment builds progressively a relation, which can be related eventually as companionship, the role given to this object perceived as a robot/subject being dependent on the appearance, while a wide range of "companion robots" are actually spreading.

This attachment has been observed along empathic reactions for various robots appearances and different socio-cultural belongings. While humans are asked to be brutal with robots like Pleo, a stuffed animal-like robot, they are not allowing themselves to hurt it [6]. Indeed, they seem to be concerned about the robot, showing empathic reactions with physiological markers [7] as same as observed on EEG of subjects looking at an android hand being cut [8]. This empathic reaction can even bring soldiers to prepare funerals for their destroyed tool-robot [9]. Moreover, while Boston Dynamics emphasised the technical performance of their Atlas robot by hustling and pushing it, in order to challenge its abilities to recover from disturbances, they provoked a lot of indignation on social networks as the public perceived their demonstration as a bullying. Even though that robot is known not to be human or alive it still get perceived with an ability to have feelings, whereas it obviously cannot: it's an object.

A human being involved into an interaction with this kind of technology is led to having an empathic perception of someone else. The impossibility to escape from this animistic perception is what we introduced as the "empathic illusion" [10]. The Atlas anecdote perfectly illustrates this phenomenon, bringing humans to feel a form of pain for the object. The robots' appearances, regularly given by roboticists, are mostly following the human affective sphere [11]. They so can be humanoid (like Nao), "petoid" (like Aibo, Karotz), or inherent from cartoons characters and cuddly toys (like Pleo, Paro). Following up these social robots, its designs seem to focus on the emotional impact they can have on the interaction pleasantness. Roboticists seem to play down the design's effects while it constantly influence the perception guiding our interaction. However, the uncanny valley awareness [12] progressively tends to design robots by trying to pull away instinctively some chosen shapes, for more abstract ones (like Jibo, Diya One), to take them out of the affective sphere. The design's overlayered complexity due to the humanoid form as seen for Atlas, is increasing the risk of anthropomorphic projection which might affect the empathic effects on the humans' perceptions.

This paper is a perceptive inception of the design approach, for French culture, by looking at the appearances dimension, which are making our first impressions on an object that could motivate the premise of animism. Instead, to associate the appearance with emotions, subjects are asked to associate the visual stimuli to acoustic stimuli, which have been referenced in previous studies for their relation effects in HRI. We expect this perceptual rupture shall be at least partially caused by visual primitives. However, we also suppose these visual forms might induce different impressions while they are associated with acoustic stimuli which are reduced only to meaningful sounds but without the lexical content to avoid their semantic influence. These sounds are explored on acoustic impressions by keeping their socio-affective "pure prosodic" information, as they have previously been defined as possible language primitives tools to build a non-dominant and altruistic relation, which dynamics processes' results on a socio-affective glue [13]. This relation building could thus be considered as the beginning of another consideration on which the changes can be firstly influenced on the objects impression itself (with its whole characteristics: colour, size, voice, etc.)

The present study is proposed as a perception test based on the supposed basics properties of visual appearances (shape, colour, size) by associating them with sounds dynamics, by hypothesising the gain of animism through their prosody. This study is thus settled in an impressionistic approach of audio-visual combination influencing the perception of an object.

2. Perception test

The perception test consists in associating a visual object with a sound. The tested parameters and values of the study are summarised in Table 1.

2.1. Audio stimuli characteristics as references

One test is composed of 64 different sounds. These audio stimuli are pure prosodic mouth noises carrying from which the socio-affective information was reproduced by copying the prosody referenced in the French E-Wiz corpus [14], previously labelled (visually and acoustically as the originally collected stimuli were multimodal) and auto-annotated [15][16]. They were also perceptively tested on: a cultural discrimination on linguistic/control degree criteria [17], the informational values perceived intra-culturally [18] and interculturally [19], then gradually tested on the socio-affective hypothesis in HRI [11]. The selected prosodies were the most universal and resistant ones in the previous studies. The sounds were equally produced by two French speakers (male and female). Each speaker produced 4 natures of pre-lexicalised sounds ("ah", "euh", "hum", "waouh"), which prosodies are out of the emotional dominance dimension. These 4 prelexical stimuli carry negative or positive valence different labels, giving 8 distinctive sounds.

Table 1: Test parameters' values and codes

Test parameters	Values
Nature	ah, euh, hum2, waouh
Valence	positive, negative
Gender	male, female
F0	original (O), F0 modification ratio (P)
Amplitude	original, normalized (up, down)
Shape	round (R), round-sharp (N), sharp (S)
Colour	white (w), red (r)
Size	big (Bg), small (Sm)

These test stimuli were produced in an acoustically isolated room, with a portable H6 zoom microphone (mono, .wav file, 16000Hz). The recordings present natural variations of the signal's intensity and amplitude due to the recording settings. In order to control these variations, the amplitude was normalised. The normalisation consisted in amplify or reduce the signals with the Audacity tool, to reach the calculated average amplitude of all the original acoustic cues, giving 16 sounds with original and normalised amplitude.

Finally, the sounds fundamental frequency was also modified to change the voice aesthetics, which is itself a part of an object design possibly changing our perception. An algorithm is applied on the sounds' F0, to change its value, without modifying its prosodic contour, in order to keep the same socio-affective information. The female voices were augmented in pitch, and inversely for the male voice, to avoid the opposite artefacts creation (artefacts as low female voice and high pitched man voice). The modifications were pushed until a perceptive limit of a human vocal tract production. The pitched female voice ratio fluctuate between 1.08 to 1.18, while the pitched male voice fluctuate between 0.76 to 0.80. The 16 previous sounds, into these two aesthetics versions for both genders, give the 64 test sounds.

2.2. Visual stimuli characteristics

The visual cues to associate to the sounds are 12 figures. Each one is determined by three appearance parameters, combined out of 3 shapes (round; round-sharp; sharp), 2 colours (white; red), 2 sizes (big; small). In psychological approaches, moods or emotions were most of the time associated with colour, with for instance the red associated with a high arousal [20], similarly for both women and men [21]. Hence, the objects appearances dimension impressions are also coupled with other modalities as audio inputs, carrying emotional semantic. Thus the chosen red (RGB code: R: 255 G: 0 B: 0 / HSV: 0° 100% 100%) is a very intense one with very high saturation and high value/brightness. It is opposed to white (RGB code: R: 255 G: 255 B: 255 / HSV: 0° 0% 100%) which appears by contrast on a neutralized grey background (RGB code: R: 189 G: 189 B: 189 / HSV: 0° 0% 75%).

The three types of shapes – round, sharp, and "round-sharp" (an intermediary blending the two others) – appear in an impressionist paradigm. The choice of the sharp shapes is motivated by the wide observations of the kiki/bouba experiment [22]. This study has brought to light an ideasthesia effect [23][24][25] between shapes and words. This kind universal sounds impressionism were also depicted for colours [26] with relatively resistant phonemes/colours association, even synaesthesia in music [27], or textures associated with voice quality [28]. But as the colours perception have been shown to mostly depend on socio-cultural backgrounds discussed as in [29] based on the Sapir and Whorf language relativity [30] this dimension is expected to have potential side coupled effects with the sounds gender for instance. In the present study, each shape has a red and a white prototype. The red spiky shape is motivated to be more often matched with negative sounds, and white round shapes with positive sounds which are most culture-independently shared impressions [22]. This might also motivate the design of the animation characters or the "companion robots", which tend to be rounded and white. Likewise, smaller things are expected to emit a higher pitched sound due to the vocal track congruence [31]. Little shapes are so expected to be more often matched with high-pitched sounds, and big shapes with low-pitched sounds.

Complementarily, a short qualitative study has been done in parallel to this perception test, in order to determine how humans qualify the 12 figures without the audio stimuli. The 11 participants (not participating in the perception test) firstly describe each figure with their own words. Secondly, they did an association task on the following propositions: kindness, aggressiveness, valency, potency, dominance, aesthetics and gender. As results, the participants tend to perceive big shapes as more protective, but also more aggressive and dominant than little ones. Red shapes had similar effects over the white ones. Globally, rounded shapes seem to be more likely inoffensive. Only among the little shapes, rounds were sometimes perceived bigger than round-sharps (this effect vanished for the big ones).

3. Experimental settings

3.1. Stimuli presentation

The 64 sounds were presented one by one to be associated each time with one shape among the 12 visual figures presented all at once on a web browser set as in Figure 1. This same interface is kept during a test (no figures position's variation), but the figures configuration and the sounds both changed randomly between judges. There is also a control subset of judges who used one specific set of interface/sound order, which is used to verify the eventual learning effects.

For each stimulus, the judge: 1) click on a "play" button to hear the sound (which can repeatedly played until the validation of the associated figure), 2) choose the best suiting figure to the sound (the figure can be changed until its validation), 3) put a grade on the answer's confidence level using a Likert scale

(from 1: lowest - to 5: highest). Then the interface switched to the next sound. There are no restrictions on the judges' socio-cultural origins criteria.

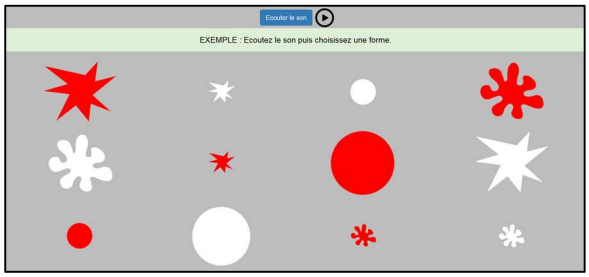


Figure 1: *Screenshot from one of the perception test interface's possible layout.*

The traces collected are: the socio-cultural information; the test start and end times; the response reaction times; the interface actions traces; the chosen values on each stimulus; the interface settings code the interface configuration and sound randomisations; the chosen figure/sound couple confidence degrees.

4. Results

4.1. Collected data samples

A total of 93 judges composed the test panel, representing 5952 pairings. They are mostly French, divided into 62% women and 38% men, with 48% within 25 and 35 years old. The other 52% are within 6 to 75 years old. There are as many participants who are used to manipulate audio-visual material in their job or spare time as the one who does not. Concerning the other cultures participants, the test was completed by 17 subjects within 9 foreign cultures and 3 biculture pairs. As the data samples present a wide variation of cultures but a few participants in each of them, the analysis will focus only on the 77 French subjects. The dataset corresponds to 2668 pairings of 35 subjects in the unrandomised subset and 2240 pairings of 42 subjects in the randomised subset. The subset is firstly analysed separately and is then merged to control the learning effects on the unrandomised subset samples.

4.2. Congruence vs. Incongruence phenomenon

We observed some strong attractions between variables and contrastively several discrepancies. Indeed, it appears that some sounds have global incompatibilities (i.e a specific shape including its colour and size), and some others have specific incompatibilities (i.e. only a colour, or only a size, or only a colour and size, etc.) In contrast, we see global and specific compatibilities. These results could be summarised as follows:

- Idiosyncratic Congruence / Thematic Congruence
- Idiosyncratic Incongruence / Thematic Incongruence

Table 2 illustrates those strong attraction/repulsion phenomena. Each cell presents the association percentage and the codename (cf. Table 1) for the shape appearance.

Table 2: *Strong and weak associations of sounds and appearances for the 77 French judges group*

Normalised amplitude only (to minimize the recording condition effects)		**Shape_Color_Size**			
		Positive		**Negative**	
		Strongest association	**Weakest association**	**Strongest association**	**Weakest association**
Female-P	**ah**	36% / S_r_Bg	0% / *_w_Sm	19% / R_w_Bg	1% S_r_Bg
	euh	29% / N_w_Sm	1% / S_w_Bg	25% / R_w_Sm	0% *_r_Bg
	hum2	55% / N_w_Sm	0-2% / *_r_*	31% / N_r_Sm	0% / R_*_Bg
	waouh	33% / S_w_Bg	~2% / *_*_Sm	22% / R_w_Bg	3% / S_r_Bg
Female-O	**ah**	27% / S_r_Bg	~3% / *_*_Sm	23% / R_w_*	1% / S_r_Bg
	euh	~23%/ N_w_*	~3% / *_r_*	21% / R_w_Sm	~2% / S_*_Bg
	hum2	42% / N_w_Sm	~3% / *_r_*	~25% / S_*_Sm	~3% / *_*_Bg
	waouh	30% / S_w_Bg	~2% / *_*_Sm	~20% / R_w_*	~4% / *_r_*
Male-P	**ah**	21% / N_w_Bg	1% / S_r_*	23% / R_r_Sm	~1% / S_*_Bg
	euh	25%/ N_r_Bg	~2% / S_*_*	~15% / *_*_Sm	~3% / *_*_Bg
	hum2	25% / N_r_Sm	~1% / S_*_Bg	25% / R_r_Sm	~2%/ *_*_Bg
	waouh	34%/ S_r_Bg	~2% / N_*_*	23% / R_r_Bg	~4% / N_*_Sm
Male-O	**ah**	23% / N_w_Bg	~3% / *_*_Sm	19% / R_r_Sm	0% / S_*_Bg
	euh	~19% / N_*_Bg	1% / S_w_Sm	~18% / R_*_Sm	~1% / S_*_Bg
	hum2	22% / N_*_Bg	~2% / S_*_*	25% / S_r_Sm	~1% / R_*_Bg
	waouh	~25% / S_*_Bg	~2% / *_*_Sm	17% / R_w_Sm	0% / N_r_Bg

An asterisk means that one of the appearance dimensions (described as shape_color_size) is not significant. This table also depicts that congruencies are more often idiosyncratic, while incongruences are more often thematic. For example, the positive Female-P hum2 have a solid idiosyncratic congruence with the little white round-sharp (N_w_Sm), and an also strong thematic incongruence with red appearances (*_r_*). One of the present perception test motivation was to explore the covariance/variance of the audio and visual stimuli parameters. As observed in this contingences matrix, this methodology does not show an exact correspondence between one acoustic parameter and one visual parameter as the dynamics of combination could affect the perception differently. In the next section, we applied Manovas in the R program in order to observe which acoustics dimensions coupling might influence the appearances' choice.

4.3. MANOVA applications

In order to explore this potential combination effects, MANOVAs were applied with sounds as factors, in uncombined condition (on 5 acoustics dimensions: nature, valency, gender, amplitude, voice change with F0) then exhaustively combined into 2 to 5 dimensions (corresponding to 27 sounds vectors). Each of these sounds vectors are crossed with the 3 appearance variables (shape, colour, size), also firstly uncombined then combined in pairs (shape-colour, shape-size and colour-size) then both of three altogether. This leads to 198 vectors tested for each of the three subsets' groups (the randomised, the unrandomized and the 77 French subjects group). All the MANOVAs are tested on the 4 Hotteling's-Lawley, Pillai, Roy and Wilks tests and are also analysed by decomposition. Moreover, the four applicability conditions (including independence and normality) were previously verified.

4.3.1. Uncombined acoustic dimensions behaviors

At first, the MANOVAs seem to be robust to analyse the decomposed dimension effects, but only until two combined dimensions, as significant differences between are noticed. But over two dimensions, the effects of each parameter are too strong to see the changes and so all combinations appeared to be significantly linked to the figure choice. Globally, the randomised group and the whole French group are showing the same characteristics. The unrandomised interface group shows the same tendency, but with additional combinations minimising the effects of certain choice. By looking at the parameter through the MANOVAs decomposition, each sound variable (nature, F0, gender, valence and amplitude) presents privileged or avoiding appearance parameters.

Concerning the sounds' nature, there are no noticeable differences, as it always appears as significant in the figure choice ($p<0.001$). We so suppose the information values carried in this parameter to be too rich to isolate focused effects on the forms' choices.

One of the most varying parameters is the F0 giving the voice aesthetics changes. The colour choice is preferentially associated with this F0 variation ($p<0.05$) with a stronger effect if it is associated with a size variation. However, it is the combination (and not only the size) which can explain this choice. This observation concerns all the groups, including the total participants, adding the foreigners. The gender seems to be never significant on the size choice while considering only this dimension. Moreover, the size is neither non-significant for the unrandomised group, only the colour explains the choice from gender ($p<0.001$). Besides, gender and valence have inversed behaviours. While the valence is combined with the two others parameters, the colours are explained significantly the choice ($p<0.001$). Besides, the amplitude seems to be not significant for the shape choice or its association with colours. It is even only justifying the size in the control group. Its effects seem to be minimised while combined, which we can consider is due to the little changes of the amplitude normalisation. Finally, while a choice is highly significant on one appearance basing on one acoustic parameter, its effects seem to appear strongly to change the other parameters' behaviours.

4.3.2. Acoustic dimensions combination effects

While we consider the effects of two sounds parameters combination on the choice of uncombined and combined appearance parameters, the major visible effects concern the pairings with the F0. While this dimension is combined with gender, it is not avoiding the choice of the size alone, and by side effects as seen before, a less significant association with combined or not shape. This effect is amplified for the unrandomised interface group. Moreover, as the effect is only minimised for the control group, it is strengthened for the group with all participants. The amplitude combined with F0 has a minimised effect on the shape and its combination with colours, and even for the shape/size pair for the randomised group (more than 0.05 significant p-values). For the control group, this effect is confirmed with a significant choice only visible on the size. Finally, the valence combined with F0 avoids mostly colours for all groups.

5. Conclusions

In this study, we initiate a research perspective to understand what causes an object to become perceived as a subject, by looking first for the primitives of multimodal appearances which are making our first impressions on an object by modifying its socio-cognitive treatment. The data analysis of the perception test illustrated a congruence vs. incongruence phenomenon between sounds and static shapes: with some pairs strongly chosen by the panel and some others systematically dodged. This could be the first step to enlighten the empathic illusion phenomenon, which is introduced to occur while the object is modifying its socio-cognitive state regarding humans observing it. As one can hypothesise, a consistency of various ontologies including movement (in this case it's sound/shape but it could be other dimensions) is mandatory to induce animism. In a second part, MANOVAs showed some strong effects on some acoustic parameters as F0 for instance. However, the combination between two acoustic features seem to have more complex interactions and could be completed by other approaches as they might not follow a statistical linear regression law as proposed in the MANOVAs. We assume that the social robotic field needs to develop solid knowledge about these emotional inductions to better understand the attachment phenomenon in human-robot interaction and to consider it better for the sake of an essential ethical approach.

6. Acknowledgements

We would like to thank Nicolas Bonnefond (Amiqual4Home, Inria) for his technical assistance, and Ambre Davat (Unniversity of Grenoble Alps, PhD student) for sharing her sounds conversion algorithm.

7. References

[1] N. H. Frijda, « Emotion, cognitive structure, and action tendency », *Cogn. Emot.*, vol. 1, nº 2, p. 115–143, 1987.

[2] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, « Evaluating the Engagement with Social Robots », *Int. J. Soc. Robot.*, p. 1–14, 2015.

[3] J. Hall, T. Tritton, A. Rowe, A. Pipe, C. Melhuish, and U. Leonards, « Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge », *Robot. Auton. Syst.*, vol. 62, nº 3, p. 392–399, 2014.

[4] C. Rich, B. Ponsler, A. Holroyd, et C. L. Sidner, « Recognizing engagement in human-robot interaction », in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference*, p. 375–382, 2010.

[5] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, « Where to look: a study of human-robot engagement », in *Proceedings of the 9th international conference on Intelligent user interfaces*, p. 78–84, 2004.

[6] K. Darling, « Extending Legal Rights to Social Robots », Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2044797, 2012.

[7] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, « An Experimental Study on Emotional Reactions Towards a Robot », *Int. J. Soc. Robot.*, vol. 5, nº 1, p. 17-34, 2012.

[8] Y. Suzuki, L. Galli, A. Ikeda, S. Itakura, and M. Kitazaki, « Measuring empathy for human and robot hand pain using electroencephalography », *Sci. Rep.*, vol. 5, 15924, 2015.

[9] J. Carpenter, « Culture and Human-Robot Interaction in Militarized Spaces: A War Story », *Routledge*, 2016.

[10] V. Aubergé, « Illusion empathique du robot social : réparation de "l'autre" fantôme », *7e rencontres du Pôle Grenoble Cognition*, Grenoble, France, 2016.

[11] V. Aubergé *et al.*, « The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the Emoz platform », *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*, Reykjavik, Iceland, 2014.

[12] M. Mori, « La vallée de l'étrange », *Gradhiva*, nº 1, p. 26–33, 2012.

[13] Y. Sasa and V. Aubergé, « Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue" », *Speech Prosody 7*, Dublin, Ireland, 2014.

[14] V. Aubergé, N. Audibert, and A. Rilliard, « E-Wiz: A trapper protocol for hunting the expressive speech corpora in lab », *4th International Conference on Language Resources and Evaluation*, Lisbonne, Portugal, p. 179-182, 2004.

[15] F. Loyau and V. Aubergé, « Expressions outside the talk turn: ethograms of the feeling of thinking », in *5th LREC*, p. 47–50, 2006.

[16] A. Vanpé and V. Aubergé, « Early meaning before the phonemes concatenation? Prosodic cues for Feeling of Thinking », *GSCP Belo Horiz. Bresil*, 2012.

[17] R. Signorello, V. Aubergé, A. Vanpé, L. Granjon, and N. Audibert, « À la recherche d'indices de culture et/ou de langue dans les micro-événements audio-visuels de l'interaction face à face », in *WACA 2010*, p. 69–76, 2010.

[18] G. D. Biasi, V. Auberge, and L. Granjon, « Perception of social affects from non-lexical sounds », in *GSCP*, 2012.

[19] Y. Sasa, V. Aubergé, and A. Rilliard, « Social micro-expressions within Japanese-French contrast », *WACAI*, Grenoble, France, 2013.

[20] R. Plutchik, « The Nature of Emotions », *Am. Sci.*, vol. 89, nº 4, p. 344, 2001.

[21] A. Mehrabian, « Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament », *Curr. Psychol.*, vol. 14, nº 4, p. 261-292, 1996.

[22] W. Köhler, *Gestalt psychology*. New York: H. Liveright, 1929.

[23] E. Milan, O. Iborra, M.J. de Cordoba, V. Juarez-Ramos, M. A. R. Artacho, and J. L. Rubio, « The Kiki-Bouba Effect A Case of Personification and Ideaesthesia », *J. Conscious. Stud.*, vol. 20, nº 1-2, p. 84-102, 2013.

[24] D. Nikolić, « Is synaesthesia actually ideaestesia? An inquiry into the nature of the phenomenon », in *Proceedings of the Third International Congress on Synaesthesia, Science & Art*, p. 26–29, 2009.

[25] A. Mroczko-Wąsowicz and D. Nikolić, « Semantic mechanisms may be responsible for developing synesthesia », *Front. Hum. Neurosci.* vol. 8, 2014.

[26] K. Watanabe, Y. Greenberg, and Y. Sagisaka, « Sentiment analysis of color attributes derived from vowel sound impression for multimodal expression », in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, p. 1-5.

[27] J. Ward, B. Huckstep, and E. Tsakanikos, « Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? », *Cortex J. Devoted Study Nerv. Syst. Behav.*, vol. 42, nº 2, p. 264-280, February, 2006.

[28] A. Moos, D. Simmons, J. Simner, and R. Smith, « Color and texture associations in voice-induced synesthesia », *Front. Psychol.*, vol. 4, p. 568, 2013.

[29] P. Kay and T. Regier, « Language, thought and color: recent developments », *Trends Cogn. Sci.*, vol. 10, nº 2, p. 51-54, February, 2006.

[30] P. Kay and W. Kempton, « What is the Sapir-Whorf hypothesis? », *Am. Anthropol.*, vol. 86, nº 1, p. 65–79, 1984.

[31] J. J. Ohala, *The frequency code underlies the sound symbolic use of voice pitch. Sound symbolism, ed. by Leanne Hinton, Johanna Nichols, and John J. Ohala, 325–47*. Cambridge: Cambridge University Press, 1994.

Animal–Robot Interaction: The Role of Human Likeness on the Success of Dog–Robot Interactions

Maretta Morovitz[1], Megan Mueller[2], and Matthias Scheutz[1]

[1]$TuftsUniversity, UnitedStates$
[2]Cummings School of Veterinary Medicine, Tufts University, United States
maretta.morovitz@tufts.edu, megan.mueller@tufts.edu, matthias.scheutz@tufts.edu

Abstract

Animals, and specifically dogs, are present throughout our social spaces which nowadays are increasingly populated with technology. Past research has mostly investigated interactions between humans and robots, failing to address possible effects of this technology, on animals, such as canines, in homes and in particular the possible utility of using robots for animal care. However, for dog–robot interactions to be successful and effective, dogs must accept robots and display positive behaviors towards them. Thus, research must determine possible robot characteristics such as particular movements or vocalizations that might be able to facilitate the dog's trust in and acceptance of the robot. The goal of the present exploratory study was to investigate the reaction of dogs to a small humanoid robot under different conditions of vocalization and movement. Our main finding from these dog–robot interaction experiments is that dogs unacquainted with the robot prefer robot vocalizations to robot movement.

Terms: animal–robot interactions, dog–robot interaction experiments, human likeness

1. Introduction

Dogs live in homes, work with our police and military, and are present on commercial farms. Already these three specific groups of dogs have begun to see robotics incorporated into their traditional living spaces. In regard to the first group, the technical world has seen much excitement and innovation in the field of social robotics as our society prepares itself to accommodate the needs and social requirements of an aging population [1]. As a result, robotic research has started to investigate the use of social robots in the home for companionship and help with daily tasks. Dogs in these homes, therefore, have begun to see robots as part of daily life. Additionally, with more adults working longer into life [2], many dogs find themselves alone for significant periods of time. Canine robotic toys, focused particularly on the social needs of dogs, have been introduced to help alleviate the loneliness and agitation expressed by these home-bound dogs [3, 4]. In regard to the second group of dogs, military and police work has begun to rely on robots for tasks previously accomplished by dogs [5]. However, until robotics can utilize the full agility of a dog, both canine and robot will have to work together to accomplish tasks [6]. Finally, as large farms increasingly turn to technology and robotics, farm animals have been increasingly required to interact with this new technology [7]. Thus, by designing robots specifically to interact with these farm dogs, we can create more synergistic partnerships between canine and machine. While these examples of potential animal–robot interaction differ greatly, they share the same root requirements. In each of these situations the dog must accept the robot in order for the interaction to be successful and effective. Thus, research must determine the robot characteristics necessary for facilitating the dog's trust and acceptance. The goal of this research is to begin to guide designers to create robots with features and functionalities most beneficial to establishing and maintaining effective dog–robot interactions, especially as they relate to the human likeness of the robot. For this study we focused on two aspects of human behavior, vocalization and movement. In the **human** condition, the anthropomorphic robotic agent will vocalize and move like a human, while in the **nonhuman** condition the robot will remain silent and stationary. After the interaction, in both conditions, the robot will offer the dog a treat. The culmination of the dogs behaviours throughout the interaction, and its acceptance or rejection of the treat will be used to determine the eventual success or failure of the interaction.

2. Previous Work

As the presence of technology grows in our society, fields such as human–computer interaction (HCI) and human–robot interaction (HRI) have expanded to include animal–robot interaction applications. Past research has suggested that the extension of HCI and HRI into animal–robot interactions could lead to insights in inter-species relationships in the areas of animal cognition, conservation, food production, and even expanding human–computer interaction knowledge [8, 10]. Devices including the FIDO vest [13] and Dog PC [14] represent technology designed specifically for canine users. Additionally, the Canine Assisted Robot Deployment (CARD) robots were designed specifically to work in conjunction with Urban Search and Rescue (USAR) dogs [15]. In field tests, CARD was identified as a "viable technique for delivering a response robot through challenging terrain to a casualty of an urban disaster" [15]. However, while the existence of these technologies lends support to the importance of dog–robot interactions, the device designs fail to examine the aspects of the interaction that contribute to its efficacy and success. This question was addressed in research conducted at Eötvös Loránd University which investigated social dog–robot interactions and examined the effect of social signals displayed by an unfamiliar robot [9]. This study concluded that "the level of sociality shown by the robot was not enough to elicit the same set of social behaviours from the dogs as was possible with humans, although sociality had a positive effect on dog–robot interactions." While not as successful as a human–dog interaction, by utilizing known social cues, the success of the interaction increased. Therefore, the research suggests that the dog is able to recognize and respond to human social cues from a nonhuman agent. Our study will take the next step and determine if the use of a robotic agent that displays humanlike behaviors, namely vocalizations and movement, will increase the success of the interaction compared to a

robotic agent that does not display humanlike behaviors.

3. Experiment

Based on the work outlined previously, we hypothesized that a dog would be more likely to accept a treat from and act positively towards a robot displaying humanlike behaviors than a robot that fails to display humanlike behaviors. To test our hypotheses, we designed a fully between-subjects investigation of the effects of human-likeness on dog–robot interactions. After a brief interaction with a robot, which either acted humanlike or nonhuman-like, the robot offered the dog a treat. Dogs then had a set amount of time to take the treat. The robot then walked directly towards the dog. At the conclusion of the interaction, the human researcher offered the dog a treat. All behaviors during the entirely of the dog–robot interaction were recorded.

3.1. Materials & Methods

3.1.1. Equipment

- **Robot** The robotic agent used for this study was the Nao programmable humanoid robot developed by Aldebaran Robotics. The robot was programmed using Choregraphe to complete a pre-scripted set of vocalizations and movements. Vocalizations were performed by a female human voice. Vocalizations include calling out the dog's name and using phrases such as "Good dog" and "Come here, buddy" and "Do you want a treat?". Movements included, offering the dog a treat by extending and opening the robot's hand, waving the robot's hands, walking side to side, turning the robot's head to follow sounds, and swaying back and forth as part of the Nao's Autonomous Life mode. During the first 10 seconds of the interaction, the robot stands (movement alone), speaks to the dog using its name (vocalization alone), and walks side to side while speaking to the dog using its name (vocalization and movement combined). The remainder of the interaction includes vocalization and movement together.
- **Dog Treat** The dog treat used was the Milk-Bone MaroSnacks Dog Treats for All Sizes Dogs.
- **Study Environment** The study environment was an empty room that contained only the Nao robot. During the interaction with the robot, only the owner and dog were present. The researcher was present in the room at the conclusion of the interaction.

3.1.2. Participants & Procedure

A total of 14 owner/dog teams participated in this study. All dogs were at least 6 months of age. Dogs remained on a leash for the entirety of the interaction, but owners were advised to allow their dog complete freedom to explore the space. Owners were also instructed to avoid all interactions with their dog to avoid influencing behavior. Upon informed consent owner and dog entered the study environment. As the dog entered the room the robot stood. In the **human** condition, the robot then continued to talk and move for the next three minutes. In the **nonhuman** condition, the robot stood on entering, but remained silent and still for the same time period. At the end of the time, the robot raised its arm and opened its hand to offer the dog a treat. In the **human** condition, this action included a vocalization "Would you like a treat?". As soon as the dog took the treat, the robot would walk forward, directly towards the dog. In the event that the dog did not take the treat, the robot would walk forward after 3 minutes. After this action, the researcher enters the room and offered the dog the treat. If the treat was not accepted, the researcher offered the uneaten treat. If the treat was accepted, the researcher offered an identical treat. This action concluded the interaction.

3.1.3. Control

A human researcher offered the dog a treat to ensure that, in the case the dog did not accept the treat when offered by the robot, this rejection was not due to the treat itself.

3.1.4. Independent Variable

We manipulated the robot's human likeness. The **human** condition used vocalizations and movements to mimic normal human–animal interactions. The **nonhuman** condition did not use any such vocalizations or movements.

3.1.5. Dog Behavior Assessment Measures

After analyzing video from each interaction, we used qualitative assessment measures [11] to categorized three subsets of behavior types: positive, negative, and neutral. A positive behavior corresponds to a behaviour that shows affinity for the robot, such as smelling the robot, cocking of the head, and approaching the robot [16]. A negative behavior corresponds to a behavior that shows a disaffinity for the robot, such as backing up, growling, head and tail down [17, 16] (Figure 1). A neutral behavior is one that shows neither affinity nor disaffinity towards the robot, such as smelling the room, laying down or sitting without looking at the robot, or trying to play with the owner.

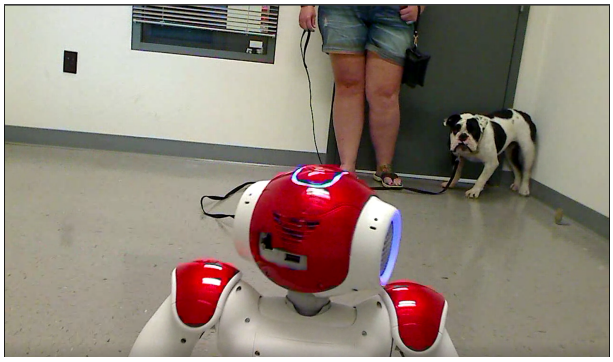


Figure 1: *On average dogs in the **nonhuman** condition displayed negative behaviors more often then they displayed positive behaviors during the interaction*

3.2. Results

Each interaction was analyzed using the following questions:

1. Did the dog accept the treat from the robot? From the human researcher?
2. When did the dog display negative, positive, and neutral responses during the interaction?

3.2.1. Acceptance of Treat

Out of 14 interactions, a total of 4 dogs, all in the **nonhuman** condition, accepted the treat from the robot (Figure 2). This

Figure 2: *All dogs that accepted the treat were part of the **nonhuman** condition*

number is equivalent to 28.6% of all dogs who participated in the study (14 dogs) or 57.1% of dogs who participated in the **nonhuman** condition (7 dogs). All dogs in both conditions accepted the treat from the human researcher.

Table 1: *Breakdown of Dog Behavior Based as Percentage of Entire dog–robot Interaction*

	Negative	**Positive**	**Neutral**
Human Condition			
Mean	38%	14%	48%
SD	0.33129307600	0.09068897300	0.25564644000
SEM	0.12521701288	0.03427720989	0.09662527197
Nonhuman Condition			
Mean	4%	41%	54%
SD	0.06020373600	0.14420790000	0.18573626200
SEM	0.02275487335	0.05450546293	0.07020170839

Table 2: *Unpaired t-test results for means between conditions*

	t	**df**	**standard difference of error**	**p**
Negative Means	267.1533	12	0.127	<.0001
Positive Means	419.3349	12	0.064	<.0001
Neutral Means	50.2365	12	0.119	<.0001

3.2.2. Dog Behavior Breakdown

After marking the videos according to the aforementioned criteria, the resulting breakdown showed that overwhelmingly, the dogs in the **human** condition displayed more negative behaviors than positive behaviors (Figure 3). The mean percentage of time spent displaying each behavior type during the course of an interaction can be seen in Table 1. In the **human** condition, an average of 44% of the interaction was categorized as a negative response, while 17% and 39% were categorized as a positive and neutral response, respectively. In contrast, during the **nonhuman** condition, an average of 4% of the interaction was categorized as a negative response, while 41% and 54% were categorized as a positive and neutral response, respectively. Using an Unpaired t-test (Table 2) to compare the means for each behavior category between the **human** and **nonhuman** conditions, it was determined that all differences between conditions are extremely statistically significant. These mean values show that the **nonhuman** condition elicited a far more positive overall response than the **human** condition. However, while the positive response is higher in the **nonhuman** condition, so is the neutral response.

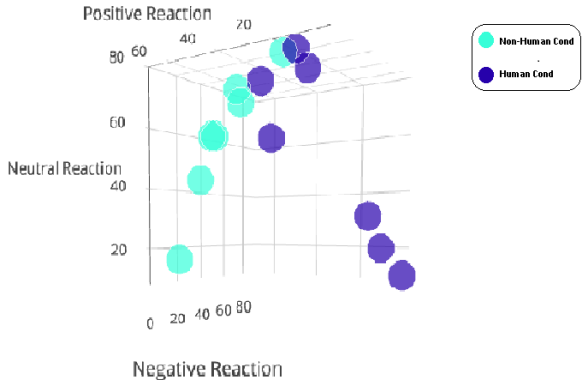


Figure 3: *3D Scatter Plot of Dog Behaviors as a Percentage of Time of dog–robot Interaction.*

3.2.3. Initial Behaviors Towards Robot

Table 3 lists the initial behaviors during the first 10 seconds of the interaction.We see that 29% of the dogs reacted positively to movement only, 100% acted positively to vocalization only, 29% acted positively to a combination of vocalization and movement.

Table 3: *Breakdown of dog behaviors during first 10 seconds of **human** condition*

	Movement Only (2 secs)	**Vocalization Only With and Without Name (5 secs)**	**Movement and Vocalization Combined (3 secs)**
1	negative	positive	negative
2	negative	positive	positive
3	negative	positive	negative
4	positive	positive	negative
5	negative	positive	positive
6	positive	positive	negative
7	negative	positive	negative

4. Discussion

Will a dog react more positively to a robot that acts human-like, compared to robot that does not act human-like? Based on previous research [9] we expected that the dogs would react more positively to a robot that demonstrated humanlike vocalizations and movements than one that did not. However, as can be clearly seen in Table 1, this hypothesis was not supported. Dogs in the **human** condition displayed higher percentages of negative behaviors than positive behaviors, while dogs in the **nonhuman** condition displayed higher percentages of positive behaviors than negative behaviors. In all cases, the dogs displayed positive behaviors to the researcher, by accepting the

treat. Additionally in the **nonhuman** condition, the dogs displayed higher percentages of neutral behaviors. These behaviors represent periods of time where the dog shows no interest in the robot. Robotic designs that will operate in shared spaces with dogs, without the primary goal of facilitate dog–robot interactions, may benefit from incorporating these nonhuman characteristics, as it would allow the robot to operate without interfering with the dog's activity. However, designs that seek to facilitate dog–robot interactions may not be aided by implementing nonhuman characteristics alone, as this may cause a neutral behavior from the dog and impede the interaction. Examining Table 3, we can clearly see the pattern of the dog's initial behaviors to the robot. All of the dogs were comfortable with the initial vocalizations alone. Many of the dogs reacted negatively to either the movement alone, or the combination of vocalization and movement. Interestingly, no dog reacted positively to both the movement alone and the combination of vocalization and movement. These observations suggest that vocalizations could be used to elicit a positive response, especially when the dog's name or a known command is given in a human voice. Thus, vocalization may be successful in creating an initial relationship and allowing for other actions, such as movement, to happen in the interaction after an initial trust between robot and dog is formed. However, since this study combined vocalizations and movement simultaneously after the first 7 seconds, further research is necessary.

4.0.1. Limitations and Future Directions

This study represents one of the preliminary attempts to study the effects of human likeness on a dog–robot interaction where the dog is the primary agent interacting directly with the robot with no cues from a human counterpart. However, there are a number of limitations of this study, which illustrate avenues for additional research in this area.

Population size By and far the largest limiting factor of this study was the number of participants. With a larger sample size, this study could be further expanded and factors such as age, breed, disposition, and whether the dog is treat or toy motivated could have been analyzed to determine if they effect the interaction. As most of the dogs lived in the surrounding area, the population was fairly homogeneous consisting of small to medium sized family dogs. This population is ideal when considering dogs who will be exposed to the continued integration of social robotics in the home. However, as robotics continue to be incorporated into our society, policy/military dogs and farm dogs will be expected to interact closely with robots for working purposes. Thus future work will need to incorporate these dogs, as they are exposed to different stimuli and have differing levels of obedience training than family dogs.

Owner–dog interactions Another limitation of this study was that dogs remained on a leash for the interaction. Had the dogs been allowed off the leash and to enter the room alone, we suspect we would have seen different behaviors. Often, when scared or threatened, the dogs retreated behind their owners. Additionally they were often confused as to why their owners were not interacting with them as they ordinarily would have and displayed more interest in their owner than the robot.

Mode of Evaluation Each of the dog's interactions during the study was categorized as positive, negative, or neutral. This analysis was done using qualitative assessment of behaviour methods [11]. While there is support for this method of assessment [12], the use of instrumental methods of assessment [11] may be necessary if the study is expanded for longer interactions with more participants, as manual behavior coding may no longer be possible.

In terms of future directions, an important next step is to separate vocalizations and movement. This study showed that the combination of vocalizations and movement were not successful in eliciting a positive behavior from the dog. However, future work should investigate if the same behaviors would be found from interacting with a robot that performs only one of these two actions, or uses vocalizations to establish trust before movement. Additionally, specifically looking at vocalizations, further work must be performed to determine if word choice (i.e., use of words known to the dog, such as its name or trained commands) or voice type (i.e., male, female, computer) influence dog behavior. Finally, the dogs' aversion towards the robot's movements may have been due to the fact that the movement, while humanlike, did not match perfectly with normal human behavior. For example, while the act of walking is humanlike, the robot walked using jerky robotic steps. Additionally, while the robot sounded and moved in a humanlike manner, it did not possess other qualities such as human scent which dogs ordinarily use to distinguish humans. These inconsistencies may have resulted in a form of a canine Uncanny Valley, where the inconsistencies between the robot and a human prevented the dog from accepting the robot. However, while these inconsistencies may have resulted in negative responses at first, it may be that, as the dogs get more comfortable with this new robotic stimuli over time, they will accept it. Thus a future study which introduces dogs to robots on multiple occasions would be needed to determine if, after repeated exposure, the dogs would become more comfortable and accepting of the robot and respond more positively to a combination of vocalization and movement than to either action alone.

5. Conclusion

The primary aim of this exploratory research was to determine the effect of human likeness, in the forms of movement and vocalization, on animal–robot, and specifically dog–robot interactions. From the above results it can be concluded that dogs more frequently displayed negative behaviors towards humanlike robots and more frequently displayed neutral and positive behaviors towards nonhumanlike robots. By examining the initial behaviors displayed, we found specifically that the dogs displayed more positive behaviors towards vocalizations than towards movement. Further studies will be required to evaluate whether vocalizations compared to non-vocalizations are preferred, and whether repeated interactions might be able to mitigate the initially negative effects of movements.

6. Acknowledgements

We are grateful for the assistance of Dr. Debbie Linder for her expertise in animal behavior, Willie Wilson for his advice on dog behavior, and Hershey for helping us pick the best treats.

7. References

[1] Broadbent, R. Stafford and B. MacDonald, "Acceptance of Healthcare Robots for the Older Population: Review and Future Directions", *International Journal of Social Robotics*, vol. 1, no. 4, pp. 319–330, 2009.

[2] DeSilver and D. DeSilver, "More Older Americans are Working,

and Working More, Than They Used To", *Pew Research Center*, 2017. [Online]. Available: http://www.pewresearch.org/fact-tank/2016/06/20/more-older-americans-are-working-and-working-more-than-they-used-to/.

[3] T. Mogg, "Pebby the Robotic Toy Means You'll Never Miss Your Pet When You're Out", *Digital Trends*, 2017. [Online]. Available: https://www.digitaltrends.com/home/pebby-robotic-pet-toy/.

[4] "CleverPet", *CleverPet*, 2017. [Online]. Available: https://clever.pet/collections/frontpage/products/cleverpet.

[5] H. Jones, S. Rock, D. Burns and S. Morris, "Autonomous Robots in SWAT Applications: Research, Design, and Operations Challenges", in *Proceedings of the 2002 Symposium for the Association of Unmanned Vehicle Systems International (AUVSI '02)*, Orlando, FL, 2002.

[6] A. Bozkurt, D. Roberts, B. Sherman, R. Brugarolas, S. Mealin, J. Majikes, P. Yang and R. Loftin, "Toward Cyber-Enhanced Working Dogs for Search and Rescue", *IEEE Intelligent Systems*, vol. 29, no. 6, pp. 32-39, 2014.

[7] R. Lenain, B. Thuilot, C. Cariou and P. Martinet, "High Accuracy Path Tracking for Vehicles in Presence of Sliding: Application to Farm Vehicle Automatic Guidance for Agricultural Tasks", *Autonomous Robots*, vol. 21, no. 1, pp. 79-97, 2006.

[8] C. Mancini, "Animal-Computer Interaction", *interactions*, vol. 18, no. 4, p. 69, 2011.

[9] G. Lakatos, M. Janiak, L. Malek, R. Muszynski, V. Konok, K. Tchon and . Miklsi, "Sensing sociality in dogs: what may make an interactive robot social?", *Animal Cognition*, vol. 17, no. 2, pp. 387-397, 2013.

[10] B. Resner, "Rover@Home: Computer Mediated Remote Interaction Between Humans and Dogs", Masters, Massachusetts Institute of Technology, 2001.

[11] A. Miklosi, *Dog behaviour, evolution, and cognition*, 1st ed. .

[12] J. Walker, A. Dale, N. Clarke, M. Farnworth and F. Wemelsfelder, "The Assessment of Emotional Expression in Dogs Using a Free Choice Profiling Methodology", *Animal Welfare*, vol. 19, no. 1, 2017.

[13] M. Jackson, Y. Kshirsagar, T. Starner, C. Zeagler, G. Valentin, A. Martin, V. Martin, A. Delawalla, W. Blount, S. Eiring and R. Hollis, "FIDO - Facilitating Interactions for Dogs with Occupations", *Proceedings of the 17th Annual International Symposium on International Symposium on Wearable Computers - ISWC '13*, 2013.

[14]]B. Heater, "Guess what Dog PC does", TechCrunch, 2017. [Online]. Available: https://techcrunch.com/2016/09/12/guess-what-dog-pc-does/.

[15] Tran, A. Ferworn, M. Gerdzhev and D. Ostrom, "Canine Assisted Robot Deployment for Urban Search and Rescue", *2010 IEEE Safety Security and Rescue Robotics*, 2010.

[16] "Canine Body Language", ASPCA Professional, 2017. [Online]. Available: http://www.aspcapro.org/resource/saving-lives-behavior-enrichment/canine-body-language.

[17] P. Borchelt, "Aggressive Behavior of Dogs Kept as Companion Animals: Classification and Influence of Sex, Reproductive Status and Breed", *Applied Animal Ethology*, vol. 10, no. 1-2, pp. 45-61, 1983.

Cognitive mechanisms underlying speech sound discrimination: a comparative study on humans and zebra finches

Merel A. Burgering[1,2]*,Carel ten Cate*[2,3] *& Jean Vroomen*[1]

[1]Department of Cognitive Neuropsychology, Tilburg University, Warandelaan 2, 5000 LE Tilburg, the Netherlands
[2] Institute Biology Leiden (IBL) Leiden University, Sylviusweg 72, 2300 RA Leiden, the Netherlands
[3] Leiden Institute for Brain and Cognition (LIBC), Leiden University, Leiden, the Netherlands
m.a.burgering@tilburguniversity.edu, c.j.tencate@biology.leidenuniv.nl, j.vroomenen@tilburguniversity.edu

Abstract

Speech sound discrimination in different species seems in many ways comparable to that of humans. Yet it is unclear what type of cognitive mechanisms are involved and whether these are shared among species.

To examine this, we trained human adults and birds (zebra finches) to discriminate two pairs of synthetic speech sounds that varied either along one dimension (vowel or sex of the speaker) or along two dimensions (vowel and speaker information needed to be integrated or combined). Subjects were assigned to one of the four stimulus-response mappings. Once training was completed, we tested generalization to new speech sounds that were either more extreme or more ambiguous than the trained sounds. Generalization to new sounds would reflect if they apply a rule or rely on an exemplar-based memory.

Humans learned the one-dimensional mappings faster than the two-dimensional mappings. Zebra finches learned all mappings equally fast, but showed the same tendency as humans. During the test, zebra finches performed in general higher on the trained sounds than on the extreme and ambiguous test-sounds, whereas humans performed higher on the extreme and trained test-sounds than on the ambiguous sounds. Humans had great difficulty with the task that required combining dimensions to form categories. These results demonstrate that birds rely on exemplar-based memory with some evidence for rule learning, whereas humans use a rule if possible.

Index Terms: categorization – information-integration – speech perception – comparative cognition – songbirds – zebra finches – human - XOR

1. Introduction

A variety of animal species can be trained to discriminate human speech sounds and form speech sound categories [1]. A recent study showed that zebra finches maintain discrimination between vowels when words were spoken by new speakers from the same sex or the other sex, which reveals the capability to generalize [2].

However, what type of cognitive mechanisms underlie this discrimination and generalization and whether animals and humans share these mechanisms is yet unclear. Learning to categorize sounds can be achieved via different mechanisms, such as exemplar-based memorization, prototype learning, rule-based learning or information-integration (II) [3].

To examine the cognitive mechanisms underlying auditory categorization, we developed a rule-based stimulus-response (SR) mapping, wherein the subject either had to discriminate the sounds based on the vowel (/i/ vs. /e/) or on the sex (male vs. female) of the speaker (hereafter: speaker). In addition, we developed two-dimensional SR-mappings: an II task and an exclusive-or (XOR) task that required the use of both dimensions to classify the stimuli.

Via a two-alternative forced-choice task with corrective feedback, we first trained birds and Dutch adults to categorize four sounds based on one or two dimension(s). Once training was completed, we tested generalization to new speech sounds from a matrix of sounds based on male-female and /e/-/i/ continua. These sounds were either more extreme, more ambiguous or intermediate between the trained sounds. For rule-based memory, we expected faster learning speed on one-dimensional mappings and generalization to new extreme and intermediate sounds. For exemplar-based memory, we expected no significant differences in learning speed between the various mappings, and similar generalization on ambiguous and extreme test-sounds.

2. Methods

2.1. Subjects & apparatus

Thirty-six adult zebra finches from the Leiden University breeding colony were individually housed in an operant conditioning chamber in a sound-attenuated room. Three horizontally aligned pecking sensors in the back wall of the cage, a fluorescent lamp, a food hatch, and a speaker were connected to an operant conditioning controller that registered all sensor pecks. Pecking the middle sensor elicited a sound. Depending on the sound, the bird had to peck the left or right sensor. A correct response resulted in temporary food access and an incorrect response led to a short period of darkness.

For humans, sixty students from Tilburg University were individually tested in a dimly lit sound-attenuated room. After a sound was presented through headphones, the participant responded by pressing one of two buttons on a response box after which they received immediate corrective feedback.

2.2 Stimulus material

Three stimulus matrices of morphed speech sounds were constructed with Tandem-STRAIGHT, each based on four different natural speech recordings from an earlier study [2] *wet* and *wit* spoken by a male and a female speaker. Sounds were decomposed into f0 trajectory, a time-frequency and an aperiodicity spectrogram, and next female-male continua for

wet and *wit* were created by manually mapping time-frequency anchors of matching features in the spectrograms of the two sounds. Next, the female-male continua were matched in similar way to create *wet-wit* morphs. Four training-stimuli and twelve test-stimuli, including more extreme, ambiguous and intermediate sounds were used for all experiments.

2.3 Design & procedure

The subjects were randomly assigned to one of the four SR-mappings: based on vowel, speaker, XOR or II. Every task was completed by 15 humans and nine birds.

All subjects were trained to sort four training sounds into two categories (see figures 1 and 2). After performing at >0.75 for three days (birds) or one training-block of 32 trials (humans), the subject was tested on the trained and non-reinforced test-sounds.

2.5 Analyses

Learning speed was defined as the number of training trials (birds) or trainingblocks (humans) required to reach criterion of >0.75 correct.

For the test, the proportions 'correct' for different sound-groups were calculated by taking the average scores of the proportion of responses to a particular sound group on each side of the midline between the differentially reinforced stimuli (e.g. taking the average of the proportion of pecks to 'extreme *wit*' and 'extreme *wet*' for the vowel test). The proportions correct for the trained sounds included non-reinforced trials only.

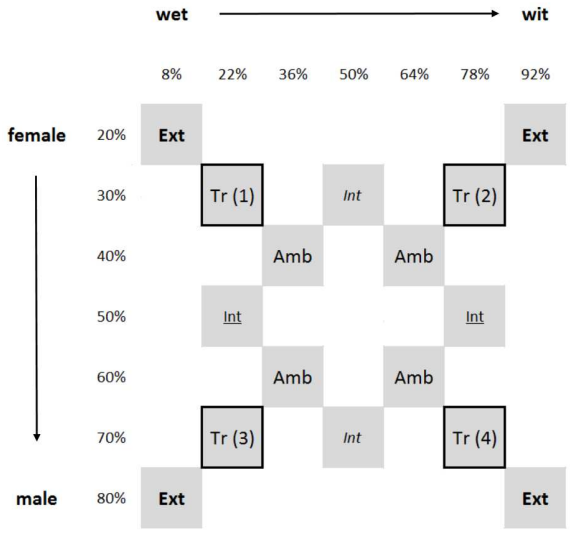


Figure 1: *Subjects were trained to sort four training sounds (Tr1, Tr2, Tr3, Tr4 for the vowel-, speaker- or XOR-task) into two categories. Upon reaching criterion they were tested on the trained and non-reinforced sounds, including intermediate sounds for the vowel (Int) and speaker task (Int). In the vowel task, Tr1 and Tr3 were assigned to one category and Tr2 and Tr4 to the other category. In the speaker task, Tr1 and Tr2 were assigned to one category and Tr3 and Tr4 were assigned to the other category. In the XOR training, Tr1 and Tr4 were assigned to one category and Tr2 and Tr3 to the other category.*

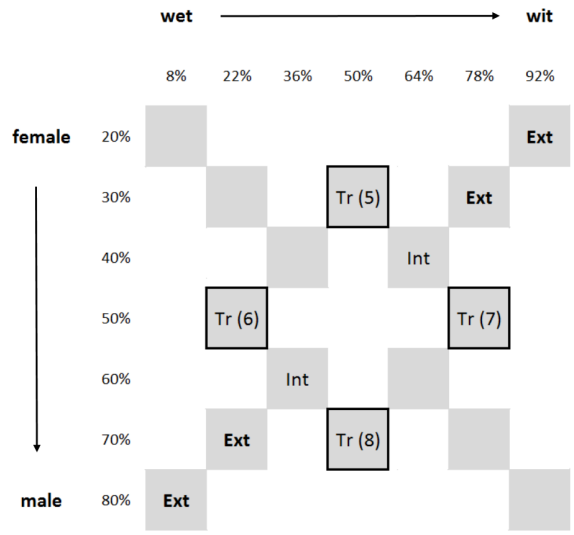


Figure 2: S*ubjects were trained to categorize four training sounds (Tr5, Tr6, Tr7 and Tr8 for the II-task) into two categories. Upon reaching criterion they were tested on trained and non-reinforced sounds. Here, Tr5 and Tr7, were assigned to one category and Tr6 and Tr8, were assigned to the other category.*

3. Results & conclusion

Humans learned the one-dimensional SR-mappings (categorization based on vowel or speaker) faster than the two-dimensional mappings (the II and XOR task). Zebra finches learned all mappings equally fast but showed the same tendency as humans. During the test phase, birds usually performed higher on the trained exemplars than on the extreme and ambiguous test-sounds whereas humans mostly performed higher on the extreme and trained test-sounds than on the ambiguous ones. These results reflect that birds rely more on exemplar-based memory than humans. In the rule-based task based on speaker, birds also show generalization for more extreme and intermediate sounds. Compared to birds, humans showed more generalization in both rule-based tasks. Humans had great difficulty with the XOR task, presumably because they confused the SR-mapping. These results demonstrate that birds rely on exemplar-based memory with weak evidence for rule learning, whereas humans prefer rule-based learning if possible.

4. References

[1] Dooling, R. J., & Brown, S. D. (1990). Speech-perception by budgerigars (Melopsittacus undulatus) - Spoken vowels *Perception & Psychophysics*, *47*(6), pp 568-574.

[2] Ohms, V. R., Gill, A., Van Heijningen, C. A. A., Beckers, G. J. L., & ten Cate, C. (2010). Zebra finches exhibit speaker-independent phonetic perception of human speech. *Proceedings of the Royal Society B-Biological Sciences*, *277*(1684), pp. 1003-1009

[3] Smith, J. D., Zakrzewski, A. C., Johnson, J. M., Valleau, J. C., & Church, B. A. (2016). Categorization: The View from Animal Cognition. *Behavioral Sciences*, *6*(2), 24.

Recording Vocal Interactivity among Turtles using AUVs

Nick Campbell [1], *Angela Dassow*[2]

[1]Trinity College Dublin, Ireland
[2]Carthage College, United States
nick@tcd.ie, adassow@carthage.edu

Abstract

This paper is the reflection of a brainstorming session at the Dagstuhl Seminar 16442 VIHAR in which potential costs and practical building constraints were made secondary to consideration of emerging technologies that might combine robotics and animal research. We have identified a practical use-case for an emerging technology and propose modifications to the devices that would enable their use in our case. The paper describes how turtles may be tracked by autonomous devices that (a) provide a corpus of their behaviour, and (b) eventually help to protect young turtles by identifying key habitats used by hatchlings.

Index Terms: Autonomous Underwater Vehicles, Turtle behaviour, Vocalisations, Image processing, RCNN, Submersible devices, Tracking-robots, Conservation

1. Introduction

Vocal interactivity is assumed to take place between turtles and may provide a source of information regarding food locations, environmental activity regarding predators moving in the area, location of refuges that the turtles may be using, or changes in water current, among other possibilities. To date there has been little research into turtle vocal interaction because of scarcity of data and difficulties of recording vocal behaviour in situ.

In this paper we propose a method whereby such information may be gathered for analysis of the relation between vocal interaction and animal behaviour. In particular we propose techniques for tracking and recording hatchlings and young turtles after they have been tagged by a human and when they start to travel further afield, beyond the range of human observers. Particularly interesting is the depth information associated with their dispersal localities since this information cannot be obtained using drone technology.

Previous work has reported success in using Autonomous Underwater Vehicles (AUVs) for tracking acoustically tagged fish near Juneau, Alaska [1] . The use of AUVs and a low energy sensor for tracking jellyfish has also been proposed [2], suggesting there may be a broader range of applications for further developing this technology for use in wildlife research. We propose a similar device suitable for tracking young turtles after hatching by following them amongst the dense vegetation which forms their habitat.

If the turtles were confined to a two-dimensional surface then the tracking problem could be easily solved by the use of drone technology, but because the turtles move in a three-dimensional space, a submersible device is required. As mentioned earlier, depth information cannot be obtained using drones, but it is important information to obtain for learning about hatchling turtle movement.

Currently many scientists and practitioners in the field (wildlife biologists and volunteers, citizen scientists etc) are devoting much of their time to protecting nesting beaches and the adult females during egg laying, or rearing hatchlings in head-starting facilities [3]. These are costly, labor-intensive efforts and involve much waiting around for eggs to hatch. In spite of these efforts, populations of most species are continuing to decline and are at risk of extinction. Table 1 (from [4]) shows that more than a third of species are endangered, and one of the major issues yet to be addressed is protecting the hatchling turtles in their first years of life in a natural environment. This has been a problem due to our inability to track the location of the hatchlings after they leave their nests. Key questions regarding hatchling dispersal are a) how far they travel from their nest and b) at what depths. Additionally, budgetary and time constraints are restricting our ability to provide further protection to the turtles, but AUVs could aid our conservation efforts by directing us to the optimal habitats used by dispersing hatchlings.

Table 1: *Worldwide IUCN chelonian population assessment*

Category	Species Count	Percentage
Extinct	7	3.1%
Critically Endangered	32	14.0%
Endangered	44	19.3%
Vulnerable	58	25.4%
Near Threatened	10	4.4%
Lower Risk	66	29.0%
Data Deficient	11	4.8%

2. Capturing Turtle Talk

It is known that turtles use vocalisations to communicate [5,6]. High quality ultrasound hydraphones are currently available for recording these calls and regularly used in cetacean research. Identifying the turtle vocalisations from among the various noises found in their environment may be problematic; however this is a problem that has already been addressed by previous research with *Chelodina oblonga* turtles [5]. Individuals can be recorded in the laboratory under a training phase to determine the vocal repertoire and known vocalisations can be compared to samples acquired from the natural environment. These techniques can be used for studying any species of turtle.

There may be little need to identify which individuals are making the particular sounds in order to associate the sounds with their subsequent behaviour, since the group may respond to any call from any individual in a consistent way. Our research therefore principally concerns identifying the sounds that trigger the movements. The body of the work involves collecting representative data in the wild, recording both acoustic events and associated movements, and then training statistical classifiers to map between features of the recordings and particular behavioural patterns.

3. Incorporating a Mother-Node

To track hatchlings in the first month of their life, small submersible devices (that we refer to as turtle-tracking robots) can be adapted for remote use, recording data continuously. Information storage on the tracking robot is not a problem but there is a foreseeable difficulty regarding battery life in extended deployment. A solution can be found by using a mother-node as a local charging station as well as for data aggregation. Once a tracking robot can sense that its turtle is inactive, it can 'ping' its GPS location, store its depth, and briefly visit the mother-node for battery replacement, data/information delivery, and debris removal before returning to the pinged location, with some compensation for potential drift. In this way, a long-term record of locations and navigation behaviour can be learnt in addition to the vocal behaviour characteristics.

At hatching, each turtle in the nest is tagged with an RFID chip using standard procedures [7]. Each turtle-tracking robot will only 'follow' one specific individual from each clutch of eggs marked but because all will have been RFID tagged, it will have the ability to record the location of any nearby siblings at the same time. Multiple devices will be needed for tracking individuals in the group. In addition to the hydraphone, the tracking robot will be fitted with 360-degree image capture. By knowing the location of its target, the device will use a form of image processing to maintain minimum distance from its source, based on perspective. The captured images can later be used in the laboratory to determine the identification of other species in the hatchlings' environment.

In addition to long battery life, an optimal device would require: navigation ability, image processing facility, GPS and depth location, etc., in order to associate vocal and bodily activity with coordinates in three-dimensional space. By including the 'mother-node' in the swarm, and assuming that the pack doesn't disperse completely but stays in a relatively closely delimited area, the task of sending data back to the researcher can be performed by the larger coordinating device.

4. A Pied-Piper Robot for Turtle Protection

There are pros and cons to utilising AUVs in wildlife research. In this section we highlight what we believe are the main concerns and the main benefits that will arise from integrating fields of robotics and fields of ethology.

There may be concerns regarding invasion of privacy in this work as there has been debate about the interaction of robotic devices with species in the wild [8]. For this reason, the proposed AUV would need to maintain a minimum distance from its target specimen. There is little justifiable concern about planting the RFID chip in the hatchling as it is already established practice [7]. There is a potential for misuse of the technology for capturing young turtles for the illegal pet trade, but collecting the eggs would be easier. There may also be a potential danger to other wildlife, for example diving birds which come into contact with the submerged devices, but the chance of this actually happening is minimal.

The technology may be put to good use at a later stage of the research when it can be used as an underwater 'sheepdog' to guide hatchlings to a protected zone or safer location provided by concerned researchers in the field. In addition to tracking native species, another application of the turtle-tracking robot is to identify the locations of invasive turtle species. The technology may enable the removal of invasive species thereby protecting the natural habitats for local species.

5. Discussion and Conclusion

In this paper we proposed a system for studying turtle behaviour underwater. The system is composed of one or more small submersible devices equipped with sensors for detecting acoustic events over a range of frequencies, and 360-degree image processing around each device. Manually swappable mother-nodes provide in-situ supplies and collect local data. We detailed the scope and limitations regarding the movement of the tracker robots and their 'relationship' with their turtle subjects, and we described how the small AUVs will be able to follow the young hatchlings and serve a practical use in providing additional information about the turtle locomotion and location of refuges.

The purpose of the research is to better understand how turtle vocalisations relate to group behaviour and movement within the environment. The method involves recording a large corpus of hatchling vocalisations along with data related to their position in the environment and in the group. The collected data will form part of a larger study employing statistical procedures (deep nets, recursive convolutional neural networks etc.,) to find mappings between the observations. There has been relatively little understanding of hatchling behaviour because of the remoteness of the location in which it typically occurs, but when this corpus becomes available ethologists will be able to observe finer details of the behaviour. This is an important tool for the future of conservation.

6. Acknowledgements

The first author is supported by Science Foundation Ireland under Grant No. 13/RC/2016, through the ADAPT Centre for Digital Content Technology at Trinity College, Dublin (www.adaptcentre.ie). The coauthor is supported by the Faculty Research and Development Grant through Carthage College.

7. References

[1] Eiler, J.H., Grothues, T.M., Dobarro, J.A., and Masuda, M.M., "Comparing autonomous underwater vehicle (AUV) and vessel-based tracking performance for locating acoustically tagged fish." Marine Fisheries Review 75.4 (2013): 27-42.

[2] Rife, J., and Rock, S., "A low energy sensor for AUV-based jellyfish tracking." Proc. 12th International Symposium on Unmanned Untethered Submersible Technology. (2001).

[3] Heppell, S.S., Crowder, L.B., and Crouse, D.T., "Models to evaluate headstarting as a management tool for long-lived turtles: Ecological Applications" Ecol Appl, vol. 6, no. 2, pp. 556-565, (1996).

[4] Rhodin, A. G. J., Pritchard, P.C.H., van Dijk, P.P., Saumure, R.A., Buhlmann, K.A., Iverson, J.B., and Mittermeier, R.A., "TURTLE TAXONOMY WORKING GROUP." (2010).

[5] Giles, J.C., Davis, J.A., McCauley, R.D., and Kuchling, G., "Voice of the turtle: The underwater acoustic repertoire of the long-necked freshwater turtle, Chelodina oblonga." JASA 126.1 (2009): 434-443.

[6] Ferrara, C.R., Vogt, R.C., and Sousa-Lima, R.S., "Turtle vocalizations as the first evidence of posthatching parental care in chelonians." Journal of Comparative Psychology 127.1 (2013): 24.

[7] Gibbons, W.J., and Andrews, K.M., "PIT tagging: simple technology at its best." Bioscience 54.5 (2004): 447-454.

[8]Bendel, O., "Considerations about the relationship between animal and machine ethics", AI & Soc (2016) 31: 103. doi:10.1007/s00146-013-0526-3

A proposal to use distributional models to analyse dolphin vocalisation

Mats Amundin[2], Henrik Hållsten, Robert Eklund[3], Jussi Karlgren[1,4], Lars Molinder[5]

[1]Gavagai, Sweden;
[2]Kolmården Wildlife Park, Sweden;
[3]Linköping University, Sweden;
[4]KTH, Sweden;
[5]Carnegie Investment Bank, Sweden

Abstract

This paper gives a brief introduction to the starting points of an experimental project to study dolphin communicative behaviour using distributional semantics, with methods implemented for the large scale study of human language.

1. Dolphin communication

Dolphins vocalise and communicate using complex signals of at least two kinds, whistles and clicks, produced in separate systems. Most dolphin species produce whistles. Whistles can last between tens of a millisecond to several seconds and consist of continuous, narrow-band, frequency-modulated signals. Most whistles can be found in the range of 2 to 20 kHz [1, 2]. Notable among whistles are the "signature whistles" which appear to be individually specific for each dolphin [3]. All dolphins also produce pulsed signals, or "clicks". These sounds are presumably used both for communication and sensing the environment. Typically the clicks come in trains with inter-click intervals ranging from few ms to several hundred ms and have most of its sound energy above the human hearing range [4, 2, 5].

2. More data and better tools

Both clicks and whistles have been studied in detail with respect to their acoustics, their relation to dolphin behaviour, and their occurrence patterns. Recent analyses have been able to describe dolphin whistle patterns using formalisms similar to those used to describe the morphological patterns of human language in terms of regularities in the way constituent elements form patterns [6, 7]. How the constituent elements of those patterns relate to each other, has not been formally described. Doing this will require much larger data sets than before: for example the most recent pattern mining experiments are performed on no more than 25 audio files.

Recent advances in computational hardware make possible the capture, storage, and analysis of analogue signals on a scale which was unthinkable even only a few years ago. Simultaneous advances in the in-memory analysis of streaming data make new processing models technically attainable. The wide availability of human linguistic data in speech and text form has made use of the technical possibilities to build unsupervised learning and dynamic on-line analysis models for inferring emerging semantic patterns in streaming data.

3. An opportunity for distributional models

Distributional analysis was first formulated by Zellig Harris [8] and such methods have gained tremendous interest due to the proliferation of large text streams and new data-oriented learning computational paradigms. Distributional semantic models collect observations of items from linguistic data and infer semantic similarity between linguistic items based on them. If linguistic items – e.g. the words $grid$ and $distributed$ – tend to cooccur – say, in the vicinity of the word $computation$ – then we can assume that their meanings are related. The primary relations of interest are *replaceability* and *combinability* of items [9]. Distributional analysis allows us to infer similarities between fundamental units, based on their observed occurrences in various patterns through the computation of second order cooccurrence relations: not only that a precedes x with some regularity, but that a and b both frequently occur with x, even if they never occur together.

4. Aims: a thesaurus of dolphin signals

While we in the current prestudy use methodology originally developed for the analysis of human language, we refrain from claiming that dolphins communicate in ways which are human-like. The task of our project is to find a representation of the signals vocalised by dolphins which allows us to infer usage similarities between identified recurring communicative tokens in dolphin communication. This aim involves a cascade of interconnected challenges.

The general task of making sense of continuous signals, assuming that they are of a sequential nature, involves three tasks: segmenting the signal into chunks of suitable level of abstraction; identifying similarities between such chunks across situations to recognise *fundamental units* of interest, corresponding to words or morphemes in human language; and then to identify patterns of occurrence among those fundamental units, corresponding to phrases or utterances in human language to be able to establish similarity of usage of such items. The result of such a procedure is a library of patterns and a thesaurus of items.

5. Challenges: the hermeneutic circle

individuation Dolphins vocalise without visible articulation [10, 11, 12]. Separating signals from a a number of vocalising individuals at the same time without knowing where the speech from one dolphin ends and another starts will be necessary, but is a known challenge in the field: *"...Identifying the vocalizers still remains one of the greatest challenges to the study of dolphin communication signals today"* [1, 13].

feature palette Humanly obvious acoustic features such as frequency and amplitude spectrograms become more complex as the interplay between the two communicative mechanisms of whistles and clicks are taken into account. Prosodic features such as pitch, quantity, stress or overlay between whistles and click bursts can be expected to communicatively relevant as well. The features of interest to identify segments from a continuous signal are manifold and involve temporal analysis of pauses and bursts, observable changes in dynamics or amplitude of frequency and harmonics, or observation of other contiguous action on the part of the vocaliser and potentially of its peers. Previous studies, have e.g. used a categorisation of context into play, for-

aging, aggression, and mother–calf interaction.

segmentation and phonetic similarity Most discovery algorithms in previous work on analysis of dolphin vocalisation have used distance-based approaches to segment signals into communicative tokens by firstly manual inspection of a transposed acoustic signal or a graphical rendition of its contours and later by computationally more convenient elastic matching of the same explicit surface signal.

directionality Directionality of sounds, especially the click sounds, is used by dolphins when they address social signals to specific conspecifics. [4] Directionality is difficult to establish, and cannot be captured at all using fixed hydrophones: it will require acoustic recordings devices that can be attached to the animals; this is not to be included in this study.

distributional similarity Once a signal has been segmented into communicative tokens and a cross-situational and cross-individual similarity measure has been defined, a distributional analysis will allow for models of similarity between tokens: "token A is used much like token B. This is the key to creating a thesaurus of communicative tokens, and the main challenge of our project.

situational factors Distributional semantic models are normally constrained to the analysis of occurrences and cooccurrences of linguistic items, but there is no conceptual need to limit the analysis to words or constructions: other contextual factors are quite reasonable candidates for inclusion in the computation. In this proposed project, factors such as the presence of stimuli of interest (e.g. food, play, humans, peers, threats) might well be used as distributional features. Enriching the model to handle context is a theoretical challenge for any distributional model.

signal and grounding Our basic assumptions are that dolphins emit and perceive sequences of fundamental items in their communicative patterns, that some of the vocalisation is intended for communication between individuals, and that dolphins are able to individuate the sounds they make to each other. Our assumption is that the communicative signal is largely sequential. This may be a risky assumption in view of the two communicative mechanisms and their interaction. Our somewhat daring assumption is also that there are segmentable communicative tokens in the signal and that those tokens are composed by combinations of separable features, much as phonemes are combined into syllables and words.

meaning Going to the heart of the entire effort, the question is what dolphins communicate about. While it is likely that some referential expressions can have shareable semantics across species, it is possible or even likely that much of dolphin–dolphin communication concerns states and aspects of dolphin life which are difficult to observe and may be near impossible for humans to conceptualise. Variation in the communicative signal may encode such content, similarly to how prosodic features are used in human–human communication. Our model will start from concrete events, observable by dolphins and humans alike, there is a risk of missing such salient variation from the signal that might refer to abstractions only accessible to dolphins. Studying the communicative behaviour of another species ranges between two theoretical extremes: On the one hand we can have a overly broad notion of what constitutes a language *everything is language*. We will then interpret every observed behavioural pattern of the studied species as a negotiation or dialog between the individual and its surroundings, including other individuals. On the other hand, if we hold to the narrowest notion of language *Only human-like communication behaviour is language* then we run the risk of finding nothing or only finding crude versions of human language. As an example, should the cheetah agonistic sound sequence moaning-growling-hissing-spitting, with "pawhit" [14] be interpreted as four distinct signals, signalling four distinct and identifiable mental states, or simply as four different "modes" of one and the same escalating mental state?

Addressing these challenges must be iterated over in turn, since the results from one will inform the processing models in both preceding and subsequent ones. After signal segmentation, we will study both similarities between those tokens as well as differences between specific individuals' uses of those tokens. The results of these studies may well force us to revisit the way we segmented the signal. It is therefore important that we capture the signals in their entire frequency spectrum with a minimum of pre-study notions as to what the relevant range of frequencies are: if the dolphins can hear it, we intend to capture it.

6. Current state of the prestudy

We are currently recording dolphins at Kolmården with a fixed hydrophone set-up, and expect to start processing the data during this year. Results will be released both as data sets and as methods and algorithms for further application in other projects. Several of the results we expect are potentially extensible to other species as well; some of the results are contributions not only to our understanding of dolphins but to our general understanding of the capacity and limits of distributional modelling.

7. References

[1] D. L. Herzing, "Making sense of it all: Multimodal dolphin communication," *Dolphin Communication and Cognition: Past, Present, and Future*, 2015.

[2] M. O. Lammers and J. N. Oswald, "Analyzing the acoustic communication of dolphins," *Dolphin Communication and Cognition: Past, Present, and Future*, vol. 107, 2015.

[3] M. C. Caldwell, D. K. Caldwell, and P. L. Tyack, "Review of the signature-whistle hypothesis for the Atlantic bottlenose dolphin," *The bottlenose dolphin*, pp. 199–234, 1990.

[4] C. Blomqvist and M. Amundin, "High-frequency burst-pulse sounds in agonistic/aggressive interactions in bottlenose dolphins, Tursiops Truncatus," in *Echolocation in bats and dolphins*, Thomas, Moss, and Vater, Eds. University of Chicago, 2004.

[5] W. Au, *The sonar of dolphins*. Springer, New York, 1993.

[6] D. Kohlsdorf, C. Mason, D. Herzing, and T. Starner, "Probabilistic extraction and discovery of fundamental units in dolphin whistles," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 8242–8246.

[7] D. Kohlsdorf, D. Herzing, and T. Starner, "Feature learning and automatic segmentation for dolphin communication analysis," *Interspeech 2016*, pp. 2621–2625, 2016.

[8] Z. Harris, *Mathematical structures of language*. Interscience Publishers, 1968.

[9] M. Sahlgren, "The distributional hypothesis," *Italian Journal of Linguistics*, vol. 20, pp. 33–54, 2008.

[10] M. Amundin and S. Andersen, "Bony nares air pressure and nasal plug muscle activity during click production in the harbour porpoise, phocoena phocoena, and the bottlenosed dolphin, tursiops truncatus," *Journal of Experimental Biology*, vol. 105, no. 1, pp. 275–282, 1983.

[11] S. Ridgway, D. Carder, R. Green, A. Gaunt, S. Gaunt, and W. Evans, "Electromyographic and pressure events in the nasolaryngeal system of dolphins during sound production," in *Animal sonar systems*. Springer, 1980, pp. 239–249.

[12] T. W. Cranford, M. Amundin, and K. S. Norris, "Functional morphology and homology in the odontocete nasal complex: implications for sound generation," *Journal of Morphology*, vol. 228, no. 3, pp. 223–285, 1996.

[13] M. Hoffmann-Kuhnt, D. Herzing, A. Ho, and M. Chitre, "Whose line sound is it anyway? identifying the vocalizer on underwater video by localizing with a hydrophone array," *Animal Behavior and Cognition*, vol. 3, no. 4, pp. 288–298, 2016.

[14] R. Eklund, G. Peters, F. Weise, and F. Munro, "An acoustic analysis of agonistic sounds in wild cheetahs," in *FONETIK 2012. Gothenburg, Sweden, May 30–June 1, 2012*. University of Gothenburg, 2012, pp. 37–40.

Extended Abstract:
Development of vocal cord mechanism for a robot capable of infant-like speech and reproducing the pitch of a babbling and a shout

Tomoki Kojima[1], *Nobutsuna Endo*[2], *Minoru Asada*[1]

[1]Adaptive Machine Systems, Osaka University
[2]Graduate School of Science and Technology for Future Life, Tokyo Denki University
asada@ams.eng.osaka-u.ac.jp

Abstract

There are many mysteries related to the speech development process of a baby; the influence that the constraint imposed by the structure of the vocal organ such as the larynx / throat has on the speech development has not been elucidated. Therefore, we have developed a speech robot, specifically, the Lingua series, which has an infant's articulatory ability and can produce voices similar to infants [1, 2]. The purpose of development of Lingua is to develop a vocal robot platform that can be utilized in behavioral studies regarding the experiments of infant-caregiver interactions as a substitute for a real infant (Figure 1). In this article, we focus on an utterance platform that can precisely reproduce various utterances of real infants.

Index Terms: Lingua-R, pitch, babbling, shout, arytenoid cartilage

1. Lingua-R: a vocal robot platform

The Lingua can reproduce a pitch of 410–2000 Hz when the vocal cords are deformed manually [3]. However, several issues remain unsolved. The vocal cord folds could not satisfy the lower pitch of the babbling (300–400 Hz), the drive mechanism is not implemented, and the pitch control program is not coded. In this study, we propose the following (Lingua-R):

(i) a new vocal cord which satisfies the pitch range of the babbling/shout at the same time (300–1000 [Hz])

(ii) a drive mechanism to control the pitch

(iii) a pitch control code with auditory feedback

2. New vocal cords

First, regarding the vocal cord folds, the optimum values of shape parameters were determined in terms of five parameters related to the thickness and hardness of the fold [3] (Figures 2, 3). In this version, we improved the performance by adding two more parameters (angle of the fold and vertical thickness of the surface layer) in an attempt to keep the minimum and maximum pitch as low and high, respectively, as possible. As a result, the lowest pitch decreased from the conventional 410 Hz to 325 Hz, but at the same time, the maximum pitch dropped from 2000 Hz to 1028 Hz. However, these values cover 96 % of the pitch range of 300 (babbling)–1000 Hz (shout). Although it does not seem perfect, it seems to achieve the initial purpose.

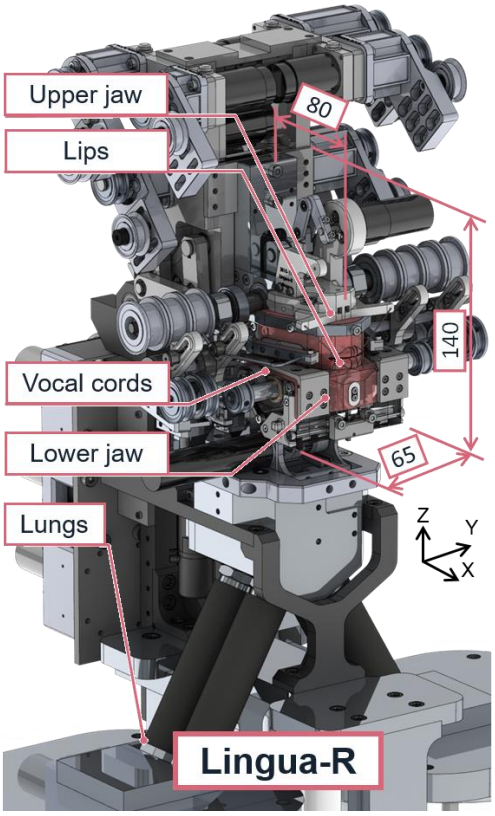


Figure 1: *Overview of Lingua-R.*

3. Arytenoid cartilage mechanism

To control the pitch of the voice, we implemented an arytenoid cartilage mechanism. The mechanism has two degrees-of-freedom (DOFs): open/close and stretch/relaxation of the vocal folds (Figure 4). Each DOF is driven by a DC motor. A motor driver controls the motor speed and an encoder measures the motor angle displacement. A Windows PC with I/O board controls the motor angle. The period of the position control of the motor is 2 ms. The control PC also controls the airflow into the vocal cords with the mechanism of the lungs. This system configuration enables simultaneous control between the vibration of the vocal cords by the airflow and stretching of the fold. Consequently, Lingua-R can produce voice and control its pitch.

4. Pitch control program

We also developed a pitch control program with auditory feedback. A PC measures Lingua-R's voice by a microphone, calculates the pitch, and sends the pitch value to the control PC. The control program changes the two-dimensional position of the distal end of the arytenoid cartilage mechanism according to a preliminarily calibrated pitch-position mapping. If there is an error between the desired and measured pitch, the program fine-tunes the position according to the gradient of the mapping.

5. Experiments and conclusion

As a result of the speech experiment, it was confirmed that the robot exhibited sufficient responsiveness, pitch followability, and stability to reproduce an infant's utterance (Figure 5). Although the pitch performance was numerically approved, the shout voice did not appear as realistic. Further improvement is required in addition to better coding for pitch control.

6. Acknowledgements

This research was supported by the JSPS Grants-in-Aid for Scientific Research (Research Project Number: 24000012).

7. References

[1] N. Endo, T. Kojima, Y. Sasamoto, H. Ishihara, T. Horii, and M. Asada, “Design of an articulation mechanism for an infant-like vocal robot ”lingua”,” in the 3rd Conference on Biomimetic and Biohybrid Systems (Living Machines 2014), pp. 389–391, 2014.

[2] N. Endo, T. Kojima, H. Ishihara, T. Horii, and M. Asada, “Design and preliminary evaluation of the vocal cords and articulator of an infant-like vocal robot ”lingua”,” in the IEEE-RAS International Conference on Humanoid Robots, pp. ThuI2–3.8, 2014.

[3] T. Kojima, N. Endo, T. Kojima, and M. Asada, "Development of Vocal Cords of an Infant-like Vocal Robot based on Anatomical Structure", Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2015), 2015.

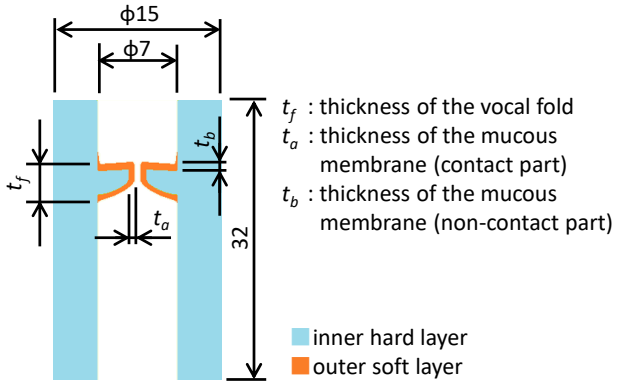


Figure 2: *Two-layered vocal cords for Lingua.*

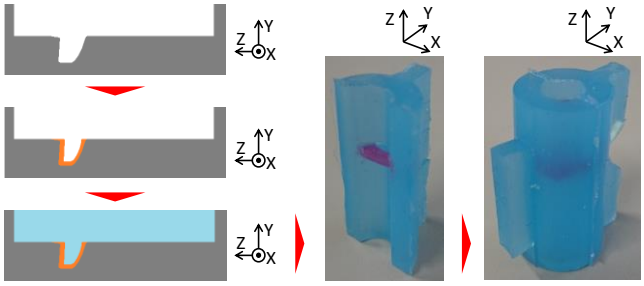


Figure 3: *Fabrication process of the vocal cords.*

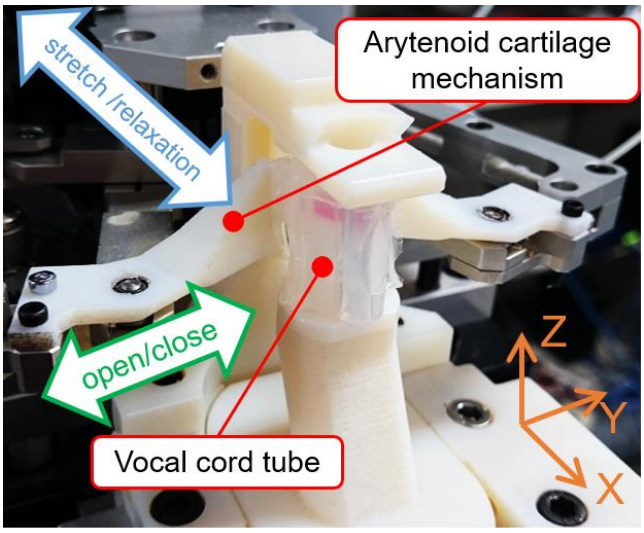


Figure 4: *Arytenoid cartilage mechanism.*

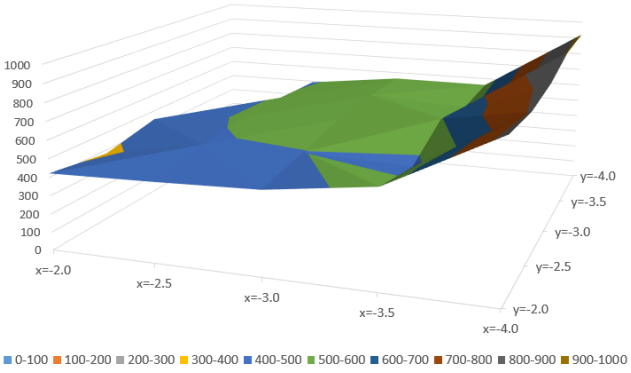


Figure 5: *Pitch performance in terms of two (x,y) parameters in driving the arytenoid cartilage arm.*

Perceptual and acoustic correlates of spontaneous vs. social laughter

Takaaki Shochi[1],[2], *Marine Guerry*[1], *Jean-luc Rouas*[2], *Marie Chaumont*[3], *Toyoaki Nishida*[4], *Yoshimasa Ohmoto*[4]

[1]LaBRI, Univ. Bordeaux, CNRS UMR 5800, 33400 Talence, France
[2]CLLE-ERRSaB UMR5263, Bordeaux, France
[3]Bordeaux Montaigne University, France
[4]Kyoto University, Japan

takaaki.shochi@labri.fr, marine.guerry@etu.u-bordeaux-montaigne.fr,
jean-luc.rouas@labri.fr, marie.chaumont@etu.u-bordeaux-montaigne.fr,
nishida@i.kyotou.-ac.jp, ohmoto@i.kyoto-u.ac.jp

Abstract

The current paper focuses on the various types of laughter recorded during real social interactions in a virtual immersive environment. In this experiment, we investigate whether human beings are able to discriminate perceptually determined volitional social laughs from spontaneous involuntary laughs using only audio information without any contextual cues. Towards this aim, we designed a perceptual experiment taken by 82 French and 20 Japanese subjects. Each subject listened to 162 laughs and chose one response among three possibilities : social, spontaneous or unknown (I don't know). The results show that all listeners are able to discriminate these two types of laughter with quite good confidence without contextual information : the correct identification rate for spontaneous laughter is about 70% with a similar amount for social laughter. We then extracted acoustic characteristics for each laughter in order to investigate possible differences between the two types of laughter. Moreover, multiple factor analysis shows that perceptual behaviours and some acoustic features (F0 and duration) are correlated. Especially, we observe a significant difference between social and spontaneous laughter through the features of F0 and total duration.

Index Terms : human laughter, recognition, acoustics, perception, culture

1. Introduction

Comprehensive knowledge about the vocal characteristics of social affective interaction has been neglected for a long time because of the lack of sufficient understanding about cognitive processing of various affective meanings as well as technical realization of such expressions. However, automatic recognition and synthetic realization of such affective meaning became one of the important issues for researchers of various scientific research fields like social robotics, medical hearing tools or language learning tools etc. [1, 2].

Such human social interaction is an exchange of social information conveyed by voice, eye contact, gestures, facial expressions, sighs or laughter [2, 3, 4]. Among these modalities, laughter must be one of the most important behaviours in the development of speech and in human and animal communication [5, 6, 7]. Laughter is often considered as a physical reaction to external stimuli which are often linked to positive valence (i.e. joyful reaction). Although laughter is deeply rooted in human biology, it also serves very strong social roles to bring about positive, mutually beneficial relationships among people and communities [8, 9]. In [10], the authors reported that laughter is usually provoked by external stimuli, and organized on three different axes : neuro-hormonal involving periaqueductal gray, the reticular formation with inputs from cortex basal ganglia and the hypothalamus [11], including muscular inputs and the respiratory axis.

In [9], the authors suggest the existence of two different types of laughter : spontaneous and volitional (or social) by neuro-physiological differences. Spontaneous laughter is considered an involuntary reaction to external stimuli. It is supposed to be innate because it occurs even before the first words. Physiological changes during such involuntary laughter are quite different from what occurs during a voluntary one. For instance, involuntary laughter is characterized by a higher activation of hypothalamus than for the voluntary one, and the chest expansion and amplitude of sound waves show more regular cycle patterns than the voluntary one which exhibits a speech-like pattern. On the other hand, social laughter is supposed to be an intentional communicative act in order to set up a positive relationship or to tone down the conflictive tension.

Concerning the acoustic realization of these various types of laughter, [12] suggests three levels of description : "bouts", "calls" and "segments". With regard to segmentation, [13] made a distinction between "spontaneous" and "social" laughter. According to recent work ([14], [12], [15], [16]), the spontaneous is higher and has a more variable F0, as well as higher variability in acoustic parameters in general. In addition, the spontaneous laughter is also characterized by longer duration with a shorter burst duration, ingressive and chuckle sounds ([13], [17]). However, there is no significant difference for both types of laughter regarding the breathiness and the mouth aperture. According to our assumption, (1) human beings are able to discriminate between social (voluntary) laughs and spontaneous (involuntary) ones using only audio information without any context. (2) Perceptually determined spontaneous laughter may have common acoustic cues among different cultures. On the contrary, (3) volitional social laughter may be perceived differently from one culture to another following cultural conventional manners.

Following these hypotheses, the current research investigates (1) whether French and Japanese subjects can discriminate between social volitional laughter and spontaneous involuntary laughter using only auditory laughs extracted from an immersive virtual interaction without any context. Independently we aim at investigating (2) acoustic characteristics for each type of laughter in French and Japanese.

2. Corpus

The stimuli were recorded in an immersive virtual environment at Kyoto University, Japan. This database consists of spontaneous affective speech recorded during a virtual reality game played by three participants. The game was designed to study communications between people in virtual environments and was made using Unity. Each player was alone in his own individual immersive virtual environment (completely surrounded by displays or in an immersive dome), but could communicate with the others using cameras and microphones. They were required to communicate in order to solve various tasks instructed by three different virtual characters. One of the main interests of this approach is that each participant can be recorded individually. A total of 12 spontaneous affective speech data files 9 Japanese (2F/7M) and 3 French (1F/2M) were recorded. A total of 254 sequences containing only laughter were manually segmented using PRAAT [18]. A first pilot test was conducted in order to investigate what acoustic features distinguish spontaneous emotional vocalizations of laughs from volitional forms which are considered as social laughs ([17], [14], [19]). 7 experimenters (3 Japanese males and 4 French (3F/1M) are instructed to annotate each sample using two labels spontaneous and social. According to a selection threshold criterion based on more than 70% of identification of the stimulus perceived as "spontaneous", a set of 27 spontaneous laughs was chosen. Another pilot test designed to choose the 27 volitional social laughs was done under the same criterion as for the spontaneous one by the 4 French experimenters who participated in the first pilot test.

3. Perceptual experiment

3.1. Paradigm

82 French native listeners (48F/34M, Mean age = 22.39 years) and 20 Japanese native listeners (9F/11M, Mean age = 24.55 years) were recruited in both countries. The stimuli were displayed 3 times each in audio alone condition in a randomized order (54 laughs (27 spontaneous / 27 social) x3 (repetitions) = 162 stimuli).

Before the test, subjects were informed about the definition of each type of laughter and the procedure of the experiment. The test was conducted individually using a GUI based interface developed under the "OpenSesame" software [20]. The total duration of the session took about 25 minutes. The subjects were required to listen to each stimulus at least once but could listen to the stimulus a second time maximum. Then, they had to select one choice among three possible answers : "spontaneous", "social", "I dont know". In the cases when spontaneous or social were selected, the subjects had to select a degree of certainty on a scale from 1 (not sure) to 7 (very sure). Definitions of the type of laughter provided in the instruction were :

- Spontaneous : it seems to you that the person is laughing in a spontaneous manner to an external event (e.g. a funny clip)
- Social : it seems to you that the person is laughing to maintain the communication with the other (e.g. embarrassed laughter, polite laughter, cynical laughter...)

3.2. Results

First of all, the χ^2 test was computed to investigate whether the distributions of listeners responses (Social, Spontaneous or Unknown) are independent or correlated. According to the result, a significant difference of the distribution of answers was observed (χ^2= 5284.7, ddl :2, p $<$ 0.001). According to the Table 1 (stimuli are in rows and the responses given by the subjects are in columns), the two types of laughter are well recognized : French subjects identified 69.24% for spontaneous laughs and 69.41% for social laughs ; Japanese listeners recognised 70.49% for spontaneous laughs and 74.63% for social laughs. These results confirmed that the listeners of both groups were able to recognize 2 types of laughter without visual indices or context.

TABLE 1: *Results for the perceptual test for French and Japanese listeners. Raw results are presented with their frequency for each row*

FRENCH	**Spontaneous**	**Social**	**Unknown**
Spontaneous	4599 (69.24%)	1683 (25.34%)	360 (5.42%)
Social	909 (13.69%)	4444 (69.41%)	1289 (19.41%)
Total result	5508 (41.46%)	6127 (46.12%)	1649 (12.41%)
JAPANESE	**Spontaneous**	**Social**	**Unknown**
Spontaneous	1142 (70.49%)	289 (17.84%)	189 (11.67%)
Social	370 (22.84%)	1209 (74.63%)	41 (2.53%)
Total result	1512 (46.67%)	1498 (46.23%)	230 (7.10%)

3.3. Correspondence analysis

In order to observe the perceptual distance of all responses based on the classification made by the listeners (spontaneous, social, I don't know) for 54 stimuli, we computed a Correspondence Analysis (CA) using FactoMineR package ([21]) under R software. According to the CA, the perceptual behaviour for 26 stimuli in the French group and 22 stimuli in the Japanese one, listeners showed an important contribution (i.e. above the expected average contribution for both 1st and 2nd dimensions).

Figure 1 and Figure 2 describe the distribution of 26 perceptual points for French and 22 for Japanese subjects on two psychometrical dimensions. The blue points on the figures represent the distribution of the perceptual behaviour and the three triangles represent the concept subjects have of the three answers. These two figures indicate that both French and Japanese listeners discriminate clearly the two types of laughter. It is also important to note that social and unknown categories are close together on the 1st dimension and far from spontaneous, which represents a well discriminated category. It indicates that volitional social laughs are more difficult to perceive than spontaneous ones. French subjects felt more difficulty to identify 5 laughs (located in the category of "unknown") rather than Japanese who had only two laughs in this category).

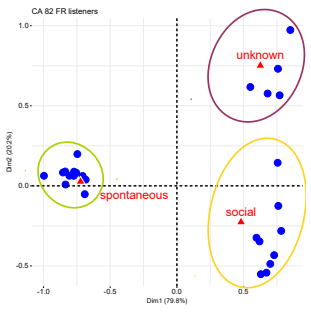


FIGURE 1: *Distribution of the perceptual behaviour of the French listeners for 26 stimuli*

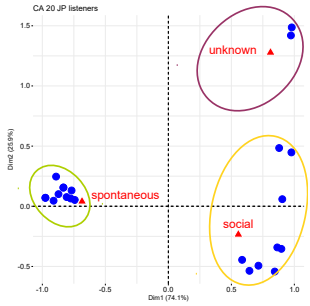


FIGURE 2: *Distribution of the perceptual behaviour of the Japanese listeners for 22 stimuli*

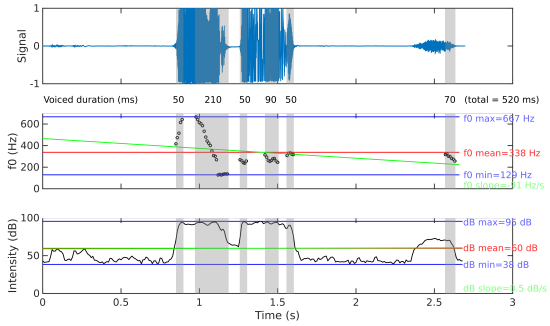


FIGURE 3: *Extraction of acoustic features on a spontaneous laughter excerpt*

4. Acoustic analysis

For the purpose of the acoustic analysis, we measured several acoustic features that were previously reported to predict affective ratings and categorization for laughter as well as for more general affective voice analysis [17].

4.1. Features extraction

For the acoustic analysis, fundamental frequency (F0) and intensity are computed every 10 ms. They were extracted using a customized version of the Snack toolkit [22]. Most analyses are carried out on the voiced parts of the laughter as detected by the F0 extraction algorithm, thus ignoring non-voiced segments.

We extracted a set of 14 features in four main categories : F0 values for assessing the variability of the fundamental frequency (we expect, for instance, to have higher frequencies as well as more variability for spontaneous laughs), Intensity values - where higher levels and variability are also expected for spontaneous laughs, Duration values - social laughs are expected to be shorter and less voiced, Harmonics-to-noise ratios which were not explored in previous laughter studies but are expected to measure to some extent the breathiness level.

- F0mean (Hz) : the mean value of F0 extracted on voiced parts of the laughs
- F0SD : the standard deviation of F0 values on a laughter excerpt (voiced parts)
- F0slope (Hz/s) : the approximated slope of F0 (voiced parts only)
- NRJmean (dB) : mean of intensity values (whole file)
- NRJsd (dB) : standard deviation of intensity values during a laughter
- NRJslope (dB/s) : approximated slope of intensity during a laughter
- total.duration : duration of a manually annotated laughter
- voiced.duration : duration of all the voiced parts of a laughter
- NBvoiced : number of voiced segments
- HNR05 : harmonic to noise ratio in the frequency band between 0 and 0.5 kHz
- HNR15 : harmonic to noise ratio in the frequency band between 0 and 1.5 kHz
- HNR25 : harmonic to noise ratio in the frequency band between 0 and 2.5 kHz
- HNR35 : harmonic to noise ratio in the frequency band between 0 and 3.5 kHz

An example of basic features extracted on a spontaneous laughter from our corpora is displayed on Figure 3.

4.2. Multiple Factor Analysis

To explore the global correlation between the acoustic features of F0 (mean, slope, standard deviation), intensity (mean, slope, standard deviation), total duration and voiced segment duration and the perceptual values (responses provided by the subjects) of both French and Japanese groups (abbreviated as Res_FR and Res_JP), a Multiple Factor Analysis (MFA) was carried out. Before computing the MFA, all acoustic and perceptual values were converted into z-scores setting average value as reference value for each parameter. The result showed that the distribution of the responses of French as well as for Japanese listeners were correlated with F0 features (mean and standard deviation) and the total duration of the laughter segments and of the voiced segments. However, the intensity (mean, slope, standard deviation) and F0 slope were less correlated with the perceptual responses of the two groups (Figure 4). Table 2 shows the values for F0 mean, F0 sd, mean duration and voiced segment mean duration. Significant differences were found between spontaneous and social laughter for F0 sd ($t(52)=5.669$, $p.=0.05$), for duration mean ($t(52)=2.696$, $p.=0.05$) and the voiced segment duration mean ($t(52)=2.595$, $p.=0.05$) between spontaneous and social volitional laughs. The variations of F0 values are higher, total duration and voiced segment duration is longer for spontaneous laughs than for social ones.

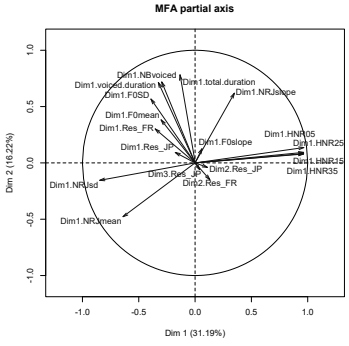


FIGURE 4: *Correlation between acoustic and perceptual values described by Multiple Factor Analysis*

TABLE 2: *Mean F0, F0 SD, total duration mean and voiced duration mean for the spontaneous and the social laughs*

	Spontaneous	Social	t-test
F0 mean (Hz)	203.59	160.80	ns
F0 SD	54.75	25.69	2.696*
Total duration mean (s)	1.81	0,65	5.669**
Voiced duration mean (s)	0.25	0.13	2.595*
*$p<.05$	**$p<.01$		

4.3. Principal Component Analysis

Previous MFA analysis showed only the global correlation between all responses and all acoustic parameters. Therefore, a Principal Component Analysis (PCA) was applied to all acoustic parameters by the two types of stimuli that were categorized by all of the listeners (French and Japanese groups). We first analysed all types of laughs for the intensity (mean, slope, standard deviation) and total duration. Ellipses indicate a normal probability (=68%) for each group of laughter. Correlations are found between the intensity standard deviation and the intensity mean vectors. The direction of the vector corresponding to the total duration on the component 2 (vertical axis) reveals that these acoustic features help differentiate spontaneous laughs from the social ones (Figure 5).

A second PCA was applied on the voiced laughs only (6 completely unvoiced laughs were removed from the set) in order to add the acoustic features related to voicing to the analysis : F0 mean, F0 slope, F0 standard deviation, voicing duration, number of voiced segments. The result (Figure 6) shows that the voiced duration and the number of voiced segments are correlated. F0 standard deviation and total duration are closely correlated. Then, F0 mean and intensity slope are correlated as well. According to the distribution of the type of laughs related to the direction of each vector on the component 1, it was found that the acoustic features concerning the voiced segment duration, the number of voiced segments, the total duration and the F0 standard deviation help differentiate spontaneous and social laughs.

Figure 7 represents the variations in duration of each laughter (normalised values). Spontaneous laughs show greater variability than social laughs.

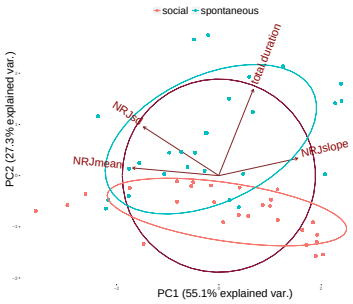


FIGURE 5: *Correlation between the acoustic values (intensity and total duration) and the 54 laughs*

5. Conclusion

The current paper investigates whether human beings can perceptually discriminate between social volitional laughter and spontaneous involuntary laughter from a corpus of spontaneous

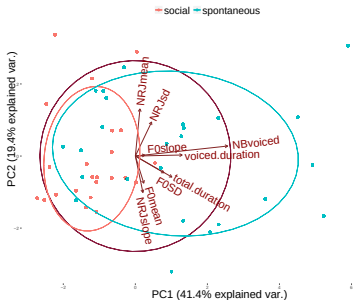


FIGURE 6: *Correlation between the acoustic values (F0, intensity, total duration, voicing duration) and the 48 laughs*

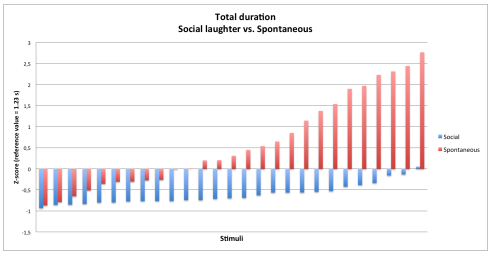


FIGURE 7: *Total duration of the two types of laughs*

laughs recorded in a virtual immersive environment using only sound information without any context or any foreign language skill. According to the perceptual discrimination experiment with native French and Japanese subjects, participants are able to discriminate these two types of laughter indicated by more than twice the chance level of recognition rate without context. This result confirms the existence of two types of laughter on the voluntary-involuntary control dimension as mentioned in previous research [9, 17].

Fourteen acoustic features including F0, harmonic to noise ratio, intensity and duration for each type of laughter are also investigated. Multiple factor Analysis was conducted to explore the global correlation between the acoustic characteristics and the participants' perceptual behaviour. Results showed that the perceptual behaviours of both French and Japanese groups were correlated with F0 features (mean and standard deviation), the total duration and the voiced segment duration. After this global result, we further investigated the important acoustic factors associated to each type of laughter (spontaneous or social). The results showed that the total duration helps to differentiate spontaneous laughs from the social ones. Moreover, we found that the voiced duration, the number of voiced segments and the F0 standard deviation also contribute to the differentiation between spontaneous and social laughs.

For future work, we will implement an additional perceptual experiment with social laughs to explore sub-categories of social laughter (embarrassment, politeness or mirthful) among two different cultures/languages (i.e. Japanese and French).

6. Acknowledgements

This study has been carried out with financial support from the French State in the frame for the "Investments for the future" program IdEx Bordeaux (ANR-10-IDEX-03-02), the Bordeaux PEPS IDEX/CNRS project "Virtual Laughter", the Bordeaux International Support in collaboration with Tsukuba University

"Cultural difference of Social Laughter". We also wish to thank all the speakers and listeners from Kyoto University and Bordeaux University for their kind participation.

7. References

[1] H. Ishiguro, "Communication robots," in *International Congress of Psychology*, Yokohama, Japan, 2016.

[2] T. Shochi and A. Rilliard, "La prononciation des apprenants de fle et la multimodalite expressive," in *La prononciation du français dans le monde*. CLE International, 2016, pp. 257–263.

[3] A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion," in *International Speech Communication Association Tutorial and Research Workshop on Speech and Emotion*, 2000.

[4] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter." in *Interspeech*, Lisbon, Portugal, 2005.

[5] R. R. Provine, "Laughter," *American Scientist*, vol. 84, pp. 38–45, 1996.

[6] ——, *Laughter : A scientific investigation*. Penguin, 2001.

[7] M. Owren and J. Bachorowski, "Acoustic assessment of vocal expression of emotion," *Handbook of emotion elicitation and assessment. Oxford University*, pp. 239–266, 2007.

[8] D. Erickson, C. Menezes, and K.-i. Sakakibara, "Are you laughing, smiling or crying ?" in *Asia-Pacific Signal and Information Processing Association Annual Summit Conference*, Sapporo, Japan, 2009.

[9] S. K. Scott, N. Lavan, S. Chen, and C. McGettigan, "The social life of laughter," *Trends in cognitive sciences*, vol. 18, no. 12, pp. 618–620, 2014.

[10] T. Jacykiewicz and F. Ringeval, "Automatic recognition of laughter using verbal and non-verbal acoustic features," Ph.D. dissertation, Masters thesis, Department of Informatics, University of Fribourg, Switzerland, 2014.

[11] B. Wild, F. A. Rodden, W. Grodd, and W. Ruch, "Neural correlates of laughter and humour," *Brain*, vol. 126, no. 10, pp. 2121–2138, 2003.

[12] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[13] H. Tanaka and N. Campbell, "Acoustic features of four types of laughter in natural conversational speech," *International Congress of Phonetic Sciences*, 2011.

[14] A. Anikin and C. F. Lima, "Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations," *The Quarterly Journal of Experimental Psychology*, pp. 1–21, 2017.

[15] S. Kipper and D. Todt, "Variation of sound parameters affects the evaluation of human laughter," *Behaviour*, vol. 138, no. 9, pp. 1161–1178, 2001.

[16] J. Vettin and D. Todt, "Human laughter, social play, and play vocalizations of non-human primates : an evolutionary approach," *Behaviour*, vol. 142, no. 2, pp. 217–240, 2005.

[17] N. Lavan, S. K. Scott, and C. McGettigan, "Laugh like you mean it : Authenticity modulates acoustic, physiological and perceptual properties of laughter," *Journal of Nonverbal Behavior*, vol. 40, no. 2, pp. 133–149, 2016.

[18] P. Boersma and W. David, *Praat, a system for doing phonetics by computer [Computer program], Version 5.3.51*, 2013.

[19] R. Jürgens, K. Hammerschmidt, and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Frontiers in psychology*, vol. 2, 2011.

[20] S. Mathôt, D. Schreij, and J. Theeuwes, "Opensesame : An open-source, graphical experiment builder for the social sciences," *Behavior research methods*, vol. 44, no. 2, pp. 314–324, 2012.

[21] R Core Team, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available : https ://www.R-project.org

[22] K. Sjolander, *The Snack Sound Toolkit [computer program]*, 2000.

Proc. 1st Intl. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), Skvöde, Sweden, 25-26 Aug 2017

Robot, Alien and Cartoon Voices: Implications for Speech-Enabled Systems

Sarah Wilson[1], *Roger K. Moore*[2]

[1]Scribestar Ltd., Central Point, 45 Beech Street, London, UK
[2]Speech & Hearing Research Group, Dept. Computer Science, University of Sheffield, UK
sarahwilson608@gmail.com, r.k.moore@sheffield.ac.uk

Abstract

Since the early days of cinema and television, fictional characters such as 'robots' and 'aliens' have almost always been portrayed with correspondingly robotic or alien voices. Likewise, animated cartoon characters are often given quirky or unusual vocal characteristics. A wide variety of different techniques are used to create these imaginary voices, and the precise properties of each are usually carefully selected to fit the narrative context. In marked contrast, the voices of speech-enabled artefacts in the non-fictional world (such as Apple's *Siri* or Amazon Echo's *Alexa*) invariably sound humanlike, despite the risk that users might be misled about the capabilities of the underlying technology. The research reported here attempts to bridge the gap by collating and analysing a large corpus of robot, alien and cartoon voices with a view to understanding the relationship between particular vocal characteristics and the perceived 'persona' of the different characters portrayed. The results show that voice quality, delay/echo/reverberation and voice breaks are major factors, and it is concluded that a more in-depth understanding could lead to guidelines and tools that would allow designers of speech synthesis systems to create more *appropriate* voices in line with the 'affordances' of the target persona.

Index Terms: robot voice, alien voice, cartoon voice, vocal affordances, speech synthesis

1. Introduction

Since the early days of cinema and television, fictional characters such as 'robots' and 'aliens' have almost always been portrayed with correspondingly robotic or alien voices. Perhaps one of the most famous examples (certainly in the UK) is the harsh metallic (and terrifying) voice of the 'Daleks' - a race of hostile alien machine-like organisms which appeared in the BBC television science-fiction series *Doctor Who* in 1963. The Dalek's voice was produced using a technique known as 'ring modulation', and the catchphrase "*Exterminate!*" in a suitably monotonic tone has subsequently become an icon of evil.

In a similar manner, animated characters are often given quirky or unusual voices. For example, cartoon series made by Warner Brothers such as *Looney Tunes* and *Merrie Melodies* featured the popular character 'Daffy Duck' - an anthropomorphic black duck who spoke with a heavily exaggerated (and much imitated) lisp. Daffy was given this particular speech impediment specifically in order to reflect the possible consequences of a duck having an extended mandible.

A wide variety of different approaches are used to create these imaginary voices, from skilled voice actors to technology-based vocal manipulation. In each case, the aim is to select the vocal characteristics that fit the narrative context. In other words, such voices are specifically tailored to be *appropriate* to the character being portrayed, and this is regarded as a part-objective part-subjective highly-skilled activity.

> "*I usually first think, if these objects, places, robots or machines really existed what would they sound like? How would they be powered? What would be the actual physics of how they work? But if I find a sound isn't working within a scene, I'll abandon the science and go with what works emotionally.*"
>
> Ben Burtt [1]
> (sound designer for *R2-D2*, *ET* and *Wall-E*)

In marked contrast, the voices of speech-enabled artefacts in the non-fictional world (such as Apple's *Siri* or Amazon Echo's *Alexa*) are invariably designed to be as humanlike as possible using the latest technology for 'text-to-speech' synthesis [2]. For such devices, it is taken for granted that users prefer 'natural' voices over artificial or robotic voices. However, a human-sounding voice encourages users to overestimate the capabilities of the underlying technology, with negative consequences for subsequent interaction [3, 4, 5]. Nevertheless, consumer resistance and the lack of a suitable design methodology mitigate against the deployment of non-humanlike voices.

Based on research carried out by the first author as part of her MSc Dissertation [6], this paper attempts to bridge this gap by collating and analysing a large corpus of robot, alien and cartoon voices. The aim has been to gain some understanding of the relationship between particular vocal characteristics and the perceived 'persona' of the different characters portrayed [7]. It was hoped that this information could be used to better inform the design of future artificial voices in line with the principles espoused in [8]: "*It's better to be a good machine than a bad person*". Not only could this lead to the design of more appropriate voices for speech-enabled artefacts, but could also avoid entering the 'uncanny valley' [9] in which mismatched perceptual cues give rise to confusion and feelings of repulsion [10].

The paper is structured as follows: Section 2 reviews the ways in which voices may be manipulated, Section 3 describes the corpus of collected vocal samples, Section 4 presents an analysis of the data, and Section 5 summarises the results and concludes with suggestions for further work.

2. Robot, Alien and Cartoon Voices

The voices that were of interest in this study were not strictly limited to robots, aliens and cartoon characters; we were also concerned with talking machines, talking animals and indeed any real or imaginary artefact that might vocalise. In practice, there are three possible approaches to creating a desired vocal characterisation: (i) employ a skilled voice actor to adopt an unusual range, skill or voice quality, (ii) use a suitably configured speech synthesiser, or (iii) modify a voice in post-production by analogue manipulation or by digital signal processing [11]. The latter may be applied to natural or synthetic speech, hence it was of special interest to the study reported here.

2.1. Vocal Manipulation Techniques

There are many ways in which a voice (real or synthetic) may be manipulated in order to change some aspect of its characteristics, and several commercial products are available - particularly for use in professional music recording studios. One of the earliest devices was *Sonovox* (invented in 1939) which fed a sound source into a performer's throat so that they could use their tongue to shape the emitted sound. This arrangement enabled artefacts and musical instruments to articulate, and *Sonovox* was famously used in 1947 to make a piano talk in *Sparky's Magic Piano*. A more modern example is *Auto-Tune* [12] which was used by Cher in 1998 to create a unique pitch-jumping effect in her song "Believe". *Auto-Tune* was also used to create the voice of 'GLaDOS' (in *Portal 2*) and 'Brian' (for *Confused.Com*). In addition, there are many 'voice-changers' available on the internet: e.g., *Voxal* [13].

Techniques for vocal manipulation operate in either the time-domain or the frequency-domain [11]. Not only are these non-exclusive, but multiple techniques may be applied in any order. As a result, the number of potential effects is huge. Examples of specific manipulations are listed in Table 1.

Table 1: *Examples of vocal manipulation techniques (roughly in order of increasing complexity).*

Technique	**Method**
Time reversal	*delay line*
Speed change	*delay line*
Tremolo	*modulated amplitude*
Vibrato	*modulated pitch*
Ring modulation	*multiplication of two signals*
Comb filter	*short delayed version added to the original*
Echo	*long delayed version added to the original*
Flanger	*delay-modulated version added to the original*
Chorus	*multiple flangers with different delays*
Phaser	*phase-modulated version added to the original*
Reverberation	*convolution with room acoustic*
Pitch shift	*homomorphic filtering*
Harmony	*pitch-shifted version added to the original*
Filtering	*frequency shaping*
Formant shift	*altered vocal tract length*
Vocoding	*linear prediction analysis-synthesis*

Many manipulations involve a 'low frequency oscillator' (LFO) that gives a time-varying character to the modified output. For example, vibrato and tremolo are achieved using an LFO to control amplitude or frequency respectively, and a "*wah-wah*" effect can be created by using an LFO to control the characteristics of a low-pass filter.

The consequence of each of these manipulations is to alter the tone and timbre of a voice in various ways. Of course, the initial voice could be natural or synthetic and could already be imbued with a particular characterisation. For example, the voice actors for the 'Daleks' (from *Doctor Who*) speak in a stilted monotone prior to their voice being subjected to modification by ring modulation using a 30Hz LFO.

2.2. Example Voices

In general, there are some fairly standardised ways that have been found to produce acceptable imaginary voices. For example, an effective robot voice can be achieved by a small increase in pitch, followed by adding back the original (c.f. 'harmony') and introducing some echo. On the other hand, a reasonable alien sound may be created by decreasing the pitch and applying a chorus effect. Finally, a cartoon-like voice may be produced by applying a large pitch increase followed by a chorus effect and added tremolo. These, and many others, are often available as 'presets' in voice-changing products such as *Voxal* [13]. Specific examples of characters with voices created through the application of the techniques mentioned in Section 2.1 are listed in Table 2.

Table 2: *A selection of characters with manipulated voices.*

Character	**Production**	**Technique**
Aliens	*Toy Story*	chorus
Celestria	*Power Rangers*	phaser
Dalek	*Doctor Who*	ring modulation
Jinx	*Spacecamp*	pitch increase
King Laufey	*Thor*	pitch decrease
Klutzy	*Robot Holocaust*	comb filter
Marvin	*Hitchhikers Guide*	vibrato
Max	*Flight of the Navigator*	reverberation
Mechanoids	*Doctor Who*	tremolo
Proteus	*Demon Seed*	flanger
Tassadar	*Starcraft*	reverse reverb.
Ultron	*Ultimate Alliance*	echo

3. The 'RAC' Corpus

A corpus of relevant voices was collected by searching the internet for films and TV series with robot, alien and cartoon characters, online reviews, forums and YouTube's recommender side bar. Further suggestions were obtained by uploading a publicly editable document and providing anonymous social media users an opportunity to contribute suggestions. Voices were not limited to any accent, ethnicity or age range, nor were they required to be speaking a known human language. However, it was decided that there must be some human element to each voice, so voices made from animal sounds or beeps and whistles (such as 'Chewbacca' or 'R2-D2' from *Star Wars*) were excluded.

All characters were labelled as being either 'robot' or 'alien', as well as given an estimate of their size, gender, material (metal or organic) and good, evil or neutral 'persona'. Voices were also labelled with subjective impressions of delay, harmony, modulation or speed change, as well as objective vocal measurements such as pitch (mean and standard deviation), jitter, shimmer, harmonic-to-noise ratio (HNR) and number of voice breaks. The latter were computed using *Praat*, a standard open-source speech analysis tool [14]. Vocal features such as breathy, creaky or whispery voice quality were also labelled.

Cartoon voices were assigned as 'robots' or 'aliens' on the basis that the latter category includes anything that does not exist in the real world. So a talking chipmunk is an alien in the same way that a 'Dalek' (from *Doctor Who*) is an alien because, although chipmunks exist, they do not speak. So, for example, the cartoon character 'Stitch' (from *Lilo and Stitch*) was classed as an alien, whereas the 'Iron Giant' (from a cartoon series of the same name) was classed as a robot. In addition, the robot category was more specific; not only could it include characters that were made of metal, but it could also be subdivided into 'cyborgs' (human-robot combinations), computers (such as 'HAL 9000' from *2001: A Space Odyssey*) and automobile robots (such as 'Optimus Prime' from *Transformers*, 'KITT' (from *Knight Rider* and 'Crimebuster' from *Heart Beeps*).

In total, 93 voices were collected and annotated, with samples spanning a period from 1939 (*The Wizard of Oz*) to 2015 (*Chappie*) - see Table 3. We are not permitted to share the data.

Table 3: *List of the 93 robot, alien and cartoon voices in the 'RAC' corpus.*

AlvinChipmunk	BigHero-Baymax	BicentMan-Galatea	BSG-Cylon	CaptainScarlet-mysterons
Chappie	Confused.com-Brian	Cyborgcop	DarkStar-Bomb20	DemonSeed-Proteus
DrWho-Icewar	DrWho-Cybermen	DrWho-Dalek	DrWho-Davros	DrWho-GreatIntelligence
DrWho-K9	DrWho-Mechaniod	DrWho-Silence	DrWho-Silence2	Dumbo-Casey
ET	Evolver	FlightoftheNav-Max	ForbiddenPlanet-Robby	GhostITShell-Proj2501
GIJ-CobraCommander	GOTG-Groot	GuyverDarkHero-Guyver	HarryPotter-Dobby	Heartbeeps-Crimecar
Heartbeeps-Val	HGTTG-Penguin	HGTTG-Vogon	HGTTG-Marvin	Hulk-Abomination
InspGadget-DrClaw	Intersteller-TARS1	Intersteller-TARS2	IronGiant-Giant	IronMan-Jarvis
JudgeDredd-ABCWar	KnightRider-Kitt	Lilo&Stitch-Stitch	LostInSpace-B9	LOTR-Gollum
LOTR-MouthSauron	LOTR-Treebeard	Marv-AlutAlianc-Ultron	Marv-SHSquad-Ultron	MenInBlack2-Zarthan
MichWeb-Cheesoid	Moon-Gerty	Portal2-GlaDos	PowerRangers-Alpha	PowerRang-Cestria
PowerRangers-Goldar	PowerRangers-Zordon	QuantumQuest-Fear	ReturnToOz-Ticktok	Robocop
RobotHolocaust-Klutzy	Rocky-Sico	ShortCircuit-Johnny5	SmashRobots	Spacecamp-Jinx
SpaceOdyssey-HAL	SparkyPiano	Starcraft-Tassadar	StarTrek-Borg	StarWars-C3PO
StarWars-DarthVador	StarWars-EmperorP	StarWars-EV-9D9	StarWars-JabbaTheHutt	StarWars-JarJarBinks
StarWars-Yoda	TheBlackCauldron-HornedKing	Tekken-Yoshi	TheBlackHole-Vincent	TheHobit-Smaug
Thor-KingLaufey	TMNT-Shredder	ToyStory-Aliens	Transformers-Deceptacon	Transformers-OptimusPrime2
Transformers-OptimusPrime-low	Tron1982-MCP	TronLegacy-Gem	Walle-Eve	Walle-Auto
Walle-Walle	WhatHappenedToRJ-RobotJones	WizardOfOz-Witch		

4. Data Analysis

4.1. General Observations

Of the 93 voices in the 'RAC' corpus, 50 were classed as 'robot' and 43 were classed as 'alien'. 64 were single recordings, 29 were concatenated samples and a large number (81) had audible background noise. Interestingly, 87 were categorised as 'male', but only 6 as 'female'. The most common effect in the corpus was echo or delay (66), followed by harmony (45), some form of modulation (40), slowing down (15) and speeding up (4). One of the more interesting effects was reverse reverberation in a character called 'Tassadar' (from *Starcraft*) which created an unusual inhalation sound prior to the speech. Pitch-changing effects were also found, such as quantised pitch shifts in 'Brian' (from *Confused.Com*) and a monotone in the 'Cylons' (from *Battlestar Galactica*). In terms of phonetic voice quality, 8 voices were creaky, 6 were whispery, 6 hoarse, 3 breathy and 3 tense/glottal.

In order to determine the relationship between the character voices in the 'RAC' corpus and normal unaltered human voices, 89 male and 42 female voices were selected from the TIMIT corpus [15] as 'controls' for comparison. The natural human voices were subjected to the same analysis techniques as the character voices, and various statistics were calculated across both sets.

4.2. Summary Statistics

Correlations were computed between the various parameters and simple 'persona' characteristics (such as characters *vs.* controls, 'robots' *vs.* 'aliens', and 'good' *vs.* 'evil') - see Table 4. As might be expected, the results indicate that character voices differ from normal (control) voices on most of the measures, reflecting the manipulations that have taken place (especially in delay, voice quality and breaks). The difference between 'robot' voices and 'alien' voices not only shows up (to a modest extent) in the voice quality measures, but also in the mean pitch. It seems that the 'aliens' in the corpus had somewhat higher pitched voices than the 'robots' (unlike the *Voxal* pre-set mentioned in Section 2.2), but both have a much larger range than controls - see Fig. 1.

As mentioned, Table 4 suggests that voice quality plays a role in distinguishing the various 'personae'. For example, Fig. 2 shows that 'alien' voices have a slightly more unusual voice quality than 'robot' voices, both of which are quite different from unmanipulated control voices. Table 4 also indicates that voice quality plays a role in distinguishing the voices of 'good' characters from 'evil' characters - see Fig. 3.

Table 4: *Correlations between measured vocal parameters and various simple 'personae'.*

	Character-Control	**Robot-Alien**	**Good-Evil**
Pitch (μ)	-0.1470	0.2225	0.0290
Pitch (σ)	-0.4732	0.1954	0.1439
Jitter	-0.5535	0.1865	0.3514
Shimmer	-0.6857	0.2470	0.3968
HNR	0.6646	-0.1868	-0.3568
Delay	-0.6905	0.0705	0.1871
Harmony	-0.5494	-0.1095	-0.0965
Breaks	-0.6550	-0.1133	-0.0307

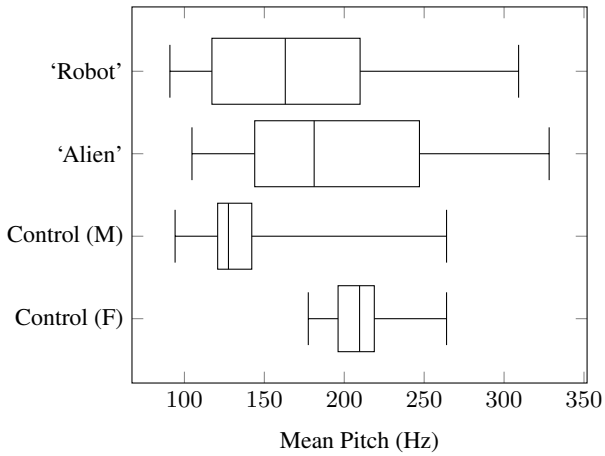


Figure 1: *Distribution of mean pitch for 'robot' and 'alien' character voices compared to unmanipulated male and female control voices.*

It can also be seen from Table 4 that an important difference between the character voices and the controls is that the characters often contain an unusually large number of breaks (often caused by the use of a low frequency modulation effect). Fig. 4 illustrates an almost complete lack of overlap between the two groups for this parameter, with one character in particular - the 'Mechanoids' (from *Doctor Who*) - showing up as the most extreme example.

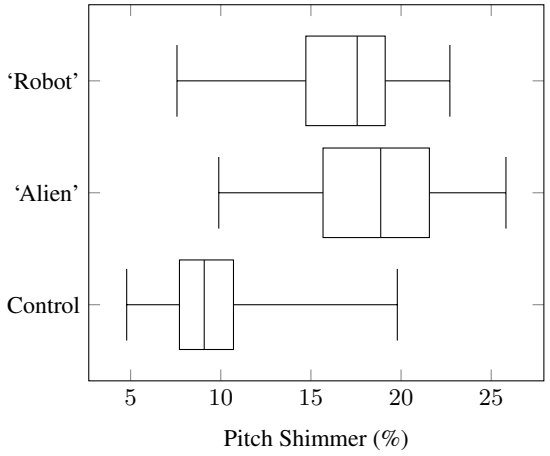


Figure 2: *Distribution of pitch shimmer for 'robot' and 'alien' character voices compared to unmanipulated control voices.*

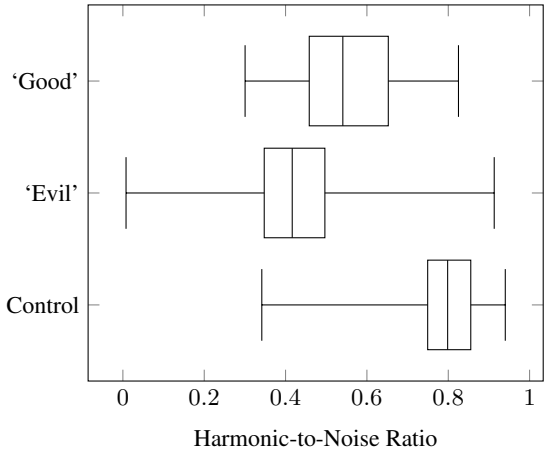


Figure 3: *Distribution of voice qualities (based on measured HNR) for 'good' and 'evil' character voices compared to unmanipulated control voices.*

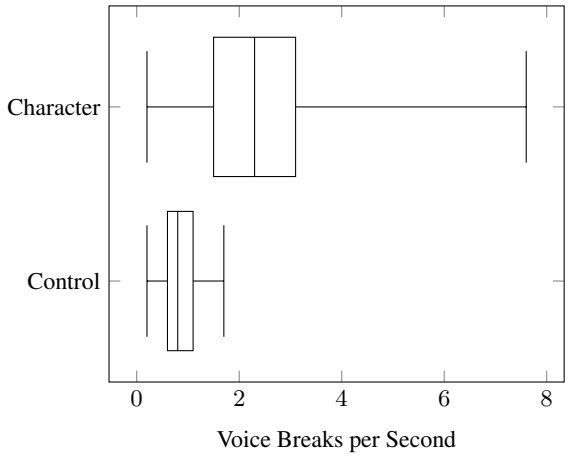


Figure 4: *Distribution of the number of voice breaks for character voices compared to unmanipulated control voices.*

4.3. Principal Component Analysis

In addition to computing the statistical correlations between various vocal parameters, the 'RCA' corpus was also analysed using principal component analysis (PCA) [16]. It was found that four components accounted for 67.4% of the total variance. The first component appeared to correspond to aspects of voice quality, serving to distinguish 'good' personae from 'evil' personae. The second component was linked to pitch which, with more female character data, could have related to gender. The third correlated to size, voice breaks and material, which could be regarded as aspects of 'appearance'. The fourth component related to echo, delay and reverberation, which seemed to distinguish fictional from non-fictional characters.

As examples, the extreme characters for the first principal component were 'GLaDOS' (from *Portal 2*), 'Galatea' (from *Bicentennial Man*) and 'Jar Jar Binks' (from *Star Wars*) as the most 'good' characters, and 'The Silence' (from *Doctor Who*), 'Abomination' (from *The Incredible Hulk*) and the 'Daleks' (from *Doctor Who*) as the most 'evil' characters. Perceptually, the first three (the 'goodies') have near normal voice quality, whereas the final three (the 'baddies') are heavily manipulated. Extreme characters for the second principal component were 'Gerty' (from *Moon*) at the low-pitch end and 'Robot Jones' (from *Whatever Happened to Robot Jones?*) at the high-pitch end.

Overall, it is interesting to note that the PCA revealed that most of the variance in the data arises as a result of personality rather than appearance, thereby confirming the importance of a character's voice as a key indicator of 'persona'.

5. Summary and Conclusion

The research reported in this paper has attempted to bridge the gap between voice-enabled artefacts in the fictional and non-fictional worlds by collating a large corpus of robot, alien and cartoon voices and comparing them with normal control voices. The aim has been to gain some understanding of the relationship between particular vocal characteristics and the perceived 'persona' of the different characters portrayed. It was hoped that this information could be used to better inform the design of future artificial voices in line with the principle that "*It's better to be a good machine than a bad person*" [8].

The study has confirmed that the majority of robot, alien and cartoon voices are manipulated to fit the narrative context, and that such manipulations are correlated with different 'personae' in predictable ways. In particular, it has been shown that voice quality, delay/echo/reverberation and voice breaks are major factors that influence the perceived character. These results, coupled with existing evidence that it is possible to infer a speaker's physical attributes such as age, weight and height from their voice [17], lend support to the view that future voice-enabled artefacts should not be designed to be as humanlike as possible, but should adopt vocal characteristics that are *appropriate* to their physical makeup and cognitive capabilities.

Ultimately, what is required is a set of guidelines (and associated tools) that would allow the designers of voice-enabled artefacts to 'dial-up' appropriate vocal characteristics in line with the visual and behavioural affordances of the target 'persona'. In order to achieve this, a more in-depth understanding of the relevant dependencies is required than the preliminary results reported here, and this is the subject of ongoing research.

6. References

[1] J. Ludwig, "Animation Sound Design: Ben Burtt Creates the Sounds for Wall-E (Part 2 of 2)," 2009. [Online]. Available: https://www.youtube.com/watch?v=eySh8FOUphM

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[3] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. Cambridge, MA: MIT Press, 2005.

[4] R. K. Moore, "Spoken language processing: Where do we go from here?" in *Your Virtual Butler, LNAI*, R. Trappl, Ed. Heidelberg: Springer, 2013, vol. 7407, pp. 111–125.

[5] ——, "From talking and listening robots to intelligent communicative machines," in *Robots That Talk and Listen*, J. Markowitz, Ed. Boston, MA: De Gruyter, 2015, ch. 12, pp. 317–335.

[6] S. Wilson, "Characteristics of Robot, Alien and Cartoon Voices," MSc in Computer Science with Speech and Language Processing, University of Sheffield, 2015.

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

[8] B. Balentine, *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. Annapolis: ICMI Press, 2007.

[9] M. Mori, "Bukimi no tani (the uncanny valley)," *Energy*, vol. 7, pp. 33–35, 1970.

[10] R. K. Moore, "A Bayesian explanation of the Uncanny Valley' effect and related psychological phenomena," *Scientific Reports*, vol. 2, no. 864, 2012.

[11] J. Rose, *Audio Postproduction for Film and Video*. Taylor & Francis, 2012.

[12] *Auto-Tune (by Antares Audio Technologies)*. [Online]. Available: http://www.antarestech.com

[13] *Voxal Voice Changer (by NCH Software)*. [Online]. Available: http://www.nchsoftware.com/voicechanger/

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: http://www.praat.org/

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Darpa TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, pp. 385–403, 1993.

[16] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[17] R. M. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, no. 6, pp. 618–625, 2002.

Index of Authors

VIHAR 2017

http://vihar-2017.vihar.org/