2016-01

People Interpret Robotic Non-linguistic Utterances Categorically

Read, R

http://hdl.handle.net/10026.1/4407

10.1007/s12369-015-0304-0
International Journal of Social Robotics
Springer Science and Business Media LLC

People Interpret Robotic Non-linguistic Utterances Categorically

Robin Read & Tony Belpaeme

International Journal of Social Robotics

ISSN 1875-4791

Int J of Soc Robotics
DOI 10.1007/s12369-015-0304-0

Springer

Int J of Soc Robotics
DOI 10.1007/s12369-015-0304-0

People Interpret Robotic Non-linguistic Utterances Categorically

Robin Read[1] · **Tony Belpaeme**[1]

Accepted: 5 May 2015

Abstract We present results of an experiment probing whether adults exhibit categorical perception when affectively rating robot-like sounds (Non-linguistic Utterances). The experimental design followed the traditional methodology from the psychology domain for measuring categorical perception: stimulus continua for robot sounds were presented to subjects, who were asked to complete a discrimination and an identification task. In the former subjects were asked to rate whether stimulus pairs were affectively different, while in the latter they were asked to rate single stimuli affectively. The experiment confirms that Non-linguistic Utterances can convey affect and that they are drawn towards prototypical emotions, confirming that people show categorical perception at a level of inferred affective meaning when hearing robot-like sounds. We speculate on how these insights can be used to automatically design and generate affect-laden robot-like utterances.

Keywords Non-linguistic Utterances · Social human–robot interaction · Categorical perception · Affective displays · Multi-modal human–robot interaction

Electronic supplementary material The online version of this article (doi:10.1007/s12369-015-0304-0) contains supplementary material, which is available to authorized users.

✉ Robin Read
robin.read@plymouth.ac.uk

Tony Belpaeme
tony.belpaeme@plymouth.ac.uk

[1] Centre for Robotics and Neural Systems, Plymouth University, Drake Circus, Plymouth PL4 8AA, Devon, UK

1 Introduction

Utterances made by machines—and in this article specifically—by *social* robots, need to be congruent with the expectations of users. Although robots come in all shapes and sizes, not all embodiments lend themselves well to the use of natural language interaction (NLI), nor do they demand that NLI be the sole means of vocal expression [36,55]. While it is well established that NLI is an important element of social machine interfaces, the current state-of-the-art in natural language processing means that NLI is limited to closed domains and interactions [46,48], making its application to social robotics limited. So too does the violation of expectations that may have been formed of a robot due to the use of NLI [51]. If a design decision is taken for a robot not to have NLI, this leaves some very open questions as to what kind of audible behaviour a robot should have instead, if any at all.

In this article, we turn to a relatively untapped source of inspiration to tackle this problem for social robots: the world of film and animation. This industry has shown us that social robots (on the "big screen") do not need to utilise NLI in order for them to be interactive, expressive and engaging for people. Rather, robots like R2D2 and many of the characters from the film WALL-E readily use robotic sounds as a means of rich expression, rather than relying on natural language. Furthermore, the audience is seemingly un-phased by this, tending to show few signs of confusion or negative reactions to the robot's use of these sounds as social cues [31]. Moreover, these robots utilise a wide variety of alternative robot-specific modalities and social cues that are arguably equally rich and expressive in their own right and allow for robot-specific methods of communication of affect and inner states. What is interesting about these different types of cues is that they tend to be overlooked when compared with the

Published online: 14 October 2015

study of other, more conventional models of communication, such as facial and body body gestures, gaze and natural language.

Through observations of the world of animation and many of the real robots that we see today, it is clear that the design space for creating robots with different morphologies is large and under the complete control of the robot designer, and as such, so too is the design space for creating robot-specific modalities and social cues [20]. Robot modalities can be inspired by nature, for instance, in the use of expressive ears [7] or tails [68]. Alternatively, these modalities can be derived from technology and thus are more robot-specific. These include the exploitation of lights and colours [24,71], shapes and robot specific body parts [20], movement [61], or certain sounds [38,56,65,72]. However, large areas of these different design spaces are uncharted, and as a result, many of the possible ways in which robots may utilise robot-specific social cues have not been explored and their utility is not well understood, nor is the understanding of how people respond to these in turn. Through studying and understanding these rich types of social cues, important design insights and principles will likely emerge (such as those presented in this article), which will help the future design of a host of different social and service robots.

The work presented here explores how robotic sounds (which we term Nonlinguistic Utterances) may be used by social robots as a means of rich affective expression during social human–robot interaction, rather than placing an unnecessarily emphasis or requirement on the need for NLI or a human-like voice. In essence, we argue that not all robots need to use natural language as a means of audible communication and that there are other alternatives that may be suitable also. The world of animation has shown us that such sounds have great effect when used on the big screen, however, one cannot assume that what works on the big-screen directly translates to autonomous robots. The work presented here is part of a larger effort to explore and understand to what extent Non-linguistic Utterances can be utilized for social HRI, where the application in animation and HRI differ, and what the limitations may be for HRI from a practical perspective.

Specifically, this article is concerned with the manner in which people infer affect from non-linguistic Utterances. Do people see a wide spectrum of rich emotions when they head robots make expressive robotic sounds, or are their interpretations more coarse? We consider these aspects as they have direct implications on the design of the acoustic behaviour of robots. Our line of reasoning is that one cannot reliably develop and use (robot-specific) social behaviours—in this case, robotic sounds—without understanding how people interpret these first.

1.1 Non-linguistic Utterances

In this section we take a more detailed look at what non-linguistic Utterances are, and are not. Following this, we outline main motivations for using such robotic sounds, and what they afford for social robots, as well as caveats that sound be considered also. Finally, we outline related work, and draw out the scientific query that we seek to investigate with this article.

1.1.1 What are Non-linguistic Utterances?

We define non-linguistic Utterances (NLUs) as sounds consisting of chirps, beeps, squeaks and whirrs, which are used as social cues during HRI. Rather than being designed to resemble spoken language or artificial languages, they tend to sound "robot-like" (for example the sounds made by R2D2 from Star Wars). However, is it still possible to include many similarities between natural language and NLUs. For example, general prosodic features from the human voice may be mapped to NLUs in order to make them sound more natural or human-like. Research also suggests that high pitch "voices" are more suited to smaller robot platforms and lower pitch "voices" [73]. While it is possible to include these similarities, it is not clear as to whether subjects respond in the same way to NLUs that follow similar naturalistic patterns found in the human voice, as they would to the human voice alone. This is in itself a strong motivation for conducting research into NLUs and other robot-specific methods of expression—to increase the depth of scientific understanding.

As social robots also produce other audible sounds also (such as motor noise, pneumatics, ect), we draw a distinction between these mechanical sounds and NLUs based upon their use as a social cue. NLUs are designed to be distinct from these mechanical sounds in order to avoid confusion and are typically used in a manner where they are meaningful auditory cues. While motor sounds may co-occur with gestural cues in a robot, we argue that through the consistent design and use of NLUs, users are are unlikely to confuse these mechanical sound with NLUs.

While NLUs are the focus of this work, it is worth noting that there is a complementary approach in the form of *Gibberish Speech* (GS) [7,50,74,76]. GS consists of utterances designed to *resemble* natural language, but deliberately have no linguistic semantic content. As with NLUs, this lack of linguistic semantic content is a deliberate design choice, founded on the notion that it is not necessity for robots to behave or interact socially. A prominent example of this can be found in the robot Kismet [7], where GS was used in place of natural language, and natural, fluid and engaging interactions were observed between subjects and the robot during user evaluations.

An important issue to address early on when presenting the subject of NLUs and GS is their relation to *language*. More precisely, whether these types of utterances constitute a real language. The standpoint that we take in this article is that they do not. Using fundamental properties[1] of a language as proposed by Hackett [28] as a reference, namely *Semanticity, Displacement, Arbitrariness, Productivity, Discreteness, Duality and Cultural Transmission*, we argue that both NLUs and GS indeed have the capacity to accommodate all of these. However, there are three vital elements missing: vocabulary (lexicon), syntax and grammar.

As such, we argue that NLUs and GS do not currently constitute a real language and urge caution at the notion of thinking about these sounds in this way due to the unnecessary confusion and misunderstanding that may arise as a result. At best, we feel that they may be considered a *protolangauge*. However, we do highlight that the step toward a true language is small, as illustrated by the robot-interaction language (ROILA) [48,69]. ROILA is presented as new artificial language that has been optimized for communication between machines and humans, having a fully specified vocabulary, syntax and grammar that must be learnt by users.

We argue that NLUs are different from this. NLUs are non-*linguistic*[2] and are unable to communicate complex ideas (e.g. "that ball is red") when compared to natural language. This is why we urge caution at thinking about NLUs in the sense of a language. Moreover, we argue that people readily and naturally attribute meaning to novel NLUs as suggested by our previous work [55,56]. Furthermore, we have found that peoples' interpretations of NLUs are guided and biased by the context/scenario in which they are used with the acoustic properties of the utterances having little influence with respect to biasing interpretation [57].

While we do not consider NLUs to be a language in their current format, it is still possible for NLUs to contain well established semantic meanings. This is done through introducing strong parallels and similarities with auditory/iconic sounds [25] and "ear-cons" [5], both of which have very well established meanings across a broad range of different cultures and are commonly part of user interfaces in personal electronic products such as smart-phones and personal computers. An analogy with gestural cues would be that NLUs can in some cases be considered as *emblematic* gestures [11]. Given the complexity of NLUs, they are able to extend far beyond such iconic sounds. In this light, it is our opinion that the amount of effort required to decode and understand utterances is minimal. We feel that NLUs are not a language that need to be learnt. Rather, they are very comparable to the auditory signals made by many modern technologies such as smart-phones, computers and many household appliances and their interpretation can be very intuitive as a result.

1.1.2 Motivations. Affordances, and Caveats of Using NLUs

NLUs have been used to great effect within the world of animation as a means to help bring inanimate objects, such as robots, to life and allowing them to be portrayed as social agents who can interact with social peers with ease, and without the need to use spoken language (robots such as R2D2 and WALL-E provide vivid examples of this). There is also a growing number of examples of commercial systems that employ NLUs as a means of expressive displays (e.g. Aldebaran's Nao, BeatBot's My Keepon, Wow Wee's RoboQuad and Sony's Aibo robot dog), showing the commercial application of NLUs. This in itself shows that there is indeed application and value in endowing robots with NLUs and GS. However, there is currently little to no principled understanding as to how NLUs are perceived by robot users, how they can be automatically generated and how they can be used effectively in real-world robotic systems.

The use of NLUs and GS does also have potential pitfalls however, as the type of utterance needs to be aligned with the physical morphology of the robot in which they are embodied [55,72], hence why some styles of utterance appear to intuitively match one type of robot more than another. It is however, not clear as to what exactly drives this. Thus, as robot and HRI designs can vary considerably, care should be taken when making the decision on what audible behaviour to align with a particular robot morphology in order to avoid violating what naive users deem as appropriate.

As a general guideline [55], it appears that GS tends to be better matched with humanoid robots well due to the close resemblance to natural language,[3] while it is deemed less appropriate for zoomorphic robots. On the other hand, NLUs seem to be appropriate for a broader range of robot morphologies, with the caveat being that the aesthetic should not be easily confused with a real biological entity. For example this is why an android robot using NLUs may be deemed as inappropriate.

[1] Hackett [28] proposes in total 13 properties that are universal to language, however the remaining properties (the vocal-auditory channel, broadcast transmission and directional reception, rapid fading, specialisation and total feedback were the listener can reproduce what they hear) relate specifically language through vocal/acoustic expression, and in the light of artificial languages such as sign language or programming languages, their value with respect to the broader concept of language is deemed as limited.

[2] We argue that NLUs do not contain *linguistic* semantic content. They do however contain semantic content in the same way that the audible sounds made by computers, smart-phones, etc, contain semantic content.

[3] Given the close resemblance between Gibberish Speech and Natural Language, it may be argued that Gibberish Speech could be perceived as a foreign language rather than meaningless nonsense to the naive observer.

While the shortcomings in comparison to natural language are obvious, NLUs do have qualities that hold promise for HRI. For example, utterances are not bound to a particular spoken dialect, thus their use in multi-lingual and multi-cultural settings may be advantageous. Also, given that NLUs hold little semantic content, there is generally less necessity to process semantic information from input user speech, thus situational settings that pose challenges for sensory equipment and technologies such as embedded microphones and NLP can be considered less problematic[4]. Furthermore, as NLUs are generally considered to hold less semantic content (with less need for a robotic system to consider semantic content), the burden of interpretation lies with the user, the *intelligent other*, with their inherent understanding of situational context and natural tendency to anthropomorphize inanimate objects such as robots [17], and treat them as socially competent [58]. Given this, the presence of an intelligent other may also be exploited to allow utterances to be used in far less restricted scenarios, widening the range of potential application areas.

NLUs also have another potential affordance—the ability to allow robotic designers to subtly manage end user expectations. It is a common observation in HRI that as the sophistication of a robotic system increases, so too do the user's expectations of that system, and thus the greater the risk that they discover the system's limitations and disengage [60]. This however can be circumvented through expectation setting where both information about a robot's capabilities (e.g. vision, tactile sensing, speech recognition, etc.) and observable behaviour (e.g. reactive behaviour to input stimulus, and expressive displays, etc.) can be used as a tool to set user expectations [51]. In theory, by employing NLUs rather than Natural Language, the robot designer is able to help keep the bar of expectation at an appropriate level.

Perhaps one obstacle in this respect are the expectations that may have been distilled in users through exposure to popular media. Both film and animation have long presented robots in a light that far exceeds the true technological state-of-the-art and can lead to naive users having drastically incorrect expectations of the real technology. While it is almost impossible to prevent this from happening, we take solace in the fact that there are many ideas and concepts that stem from popular culture that can potentially have very positive applications in real social HRI. As such we strive to focus on these rather than the examples and expectations that negatively impact our field.

1.1.3 Related Work on NLUs

Previous work on NLUs within the context of *social* HRI has focused on a variety of issues. For example, early work explored how auditory icons could be used by a mobile service robot to communicate intended directional motion trajectories to nearby humans [33]. It has also been shown that different agent embodiments affect how subjects interpret NLUs [37,55]. The type of utterance made by a robot should be aligned with the users' expectations of the robot: an animal-like robot should make animal-like sounds, not human-like GS. Jee *et al.* have investigated how the design of NLUs can be aided by insights from the world of musicology [31], and have also found that the intensity of the interpretation is increased when the utterances are presented along-side an affective facial expression [32]. It has also been found that NLUs can be used to convey "positive" and "negative" attitudes as well as differing degrees of confidence when a robot provides information to a user [35]. There have also been explorations into the effectiveness of NLUs as a means of expressing affect during child–robot interaction [56]. More recent work has also explored the use of NLUs as a means of proxemic feedback [65] (i.e. using sensory input relating to proxemics as a driver for the type of NLU displayed).

It is generally agreed that natural language has two distinct but interleaved components that require to be encoded/decoded by interactants: *what* is said (i.e. the meaning), and *how* it is said (i.e. cues relating to the affective state of the speaker) [52]. In the case of NLUs, it is only the expression of affect that can be addressed due to the lack of semantic content. Thus, an important aspect to understand is how subjects perceive and interpret NLUs on an affective level as insights here can be used to better inform the production and use of NLUs in real settings and scenarios, as well as helping provoke the illusion of life and anthropomorphization [17].

A common trend in the previous NLU and GS research has been the method for capturing subject interpretations of stimuli. In the vast majority of cases, the experimental setups have involved subjects listening to an utterance and then providing a self-reported affective interpretation (c.f. [7,31–33,35,37,50,55,74,76]). Specifically, this has been done by presenting subjects with a small list of affective labels (most commonly the "basic six" emotions has suggested by Ekman and Friesen [19]) rather than using affective measures that are based upon continuous dimensional representations of affect [13]. While this approach allows for the construction of *confusion matrices*, which indicate which utterances may cause confusion between interpretations, it is not possible to investigate the *transition* between these affective interpretations as the affective meansument tools based on categories do not provide the resolution to do so.

The relevance for this with respect to NLUs is that very little is learnt about the dynamics of the relationship between

[4] Such settings tend to be in dynamic and unpredictable *real world* environments that are far from the protected and controlled, "safe" laboratory environments.

the properties of utterances and their affective interpretations. This makes the interpolation between two different affective portrayals difficult as assumptions have to be made regarding the mappings between the utterance parameters corresponding to the different affective prototypes (e.g. should the interpolation in utterance parameters between "happy"' and "angry" be linear?). It is better to conduct evaluations that directly investigate these underlying transitions in order to uncover these dynamics in the mappings.

Results from a previous experiment [56] in which young children provided affective interpretations of a wide variety of NLUs indicated that while children readily attribute different emotions to NLUs, they do not do so in a consistent or predictable manner. What was observed however was that affective interpretations appeared to be clustered around certain "basic" affective prototypes (e.g. happy, angry, sad, scared and surprised) rather than evenly distributed across all the possible interpretations. This led to the hypothesis that affective inferences of NLUs may be subject to some degree of Categorical Perception. This article presents the results of an experiment which seeks to test this hypothesis.

1.2 Categorical Perception

Categorical perception (CP) [27,44] is the phenomenon whereby sensory stimulation along a continuum is not seen as gradual, but as instances of discrete categories. In essence, when perceiving a stimulus continuum with equal, linear physical differences (for example, the *hue* of the colour spectrum), people perceive the continuum consisting of discrete categories (for example containing red, blue, green, yellow bands). The perceptual stimuli are drawn towards perceptual categories. The hallmark of CP is people exhibiting greater sensitivity to a physical change that occurs over a perceptual *boundary* than when the same physical change occurs within a perceptual *region* [29,42]. As such, stimuli that are near such a boundary are commonly subject to a "magnet effect" [39]. This is a mechanism whereby the perception of the stimulus is *pulled* toward a particular well established category, resulting in a non-linear relationship between the stimulus continuum and the perceived class membership of each stimulus within the continuum.

Figure 1 sketches a theoretical illustration of how values along a perceptual continuum are drawn towards one of two classes. The magnet effect introduces non-linear perception across the continuum. The figure also shows the *discrimination profile* (ratings of whether neighboring stimuli in the continuum are perceived as being of the same class) of the two classes—as the stimulus steps approach the *categorical boundary*, people exhibit greater sensitivity to the differences between the stimuli and thus assign different class memberships to each.

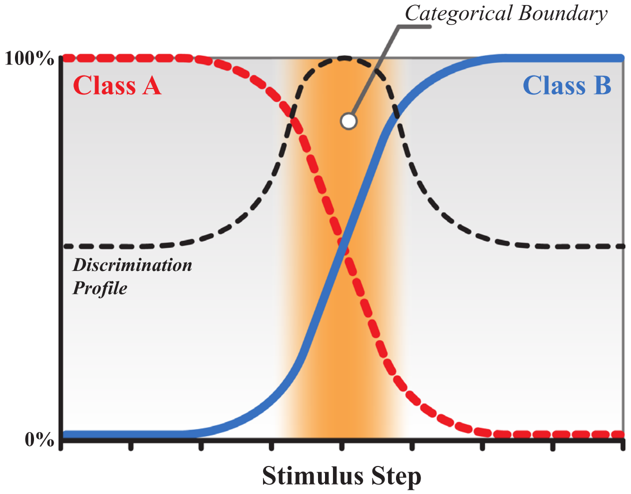

Fig. 1 Illustrative example of the dynamics of class membership associated with categorical perception

The phenomenon of CP has been shown to take place during the processing of a broad variety of sensory information in both adults and children. Examples are the processing of phonetic sounds [39,44], colour [6,23,77], acoustic pitch [43,66], facial expressions [4,12,21] and affect in synthesised speech [42]. As such, CP has been proposed as a fundamental foundation upon which human cognition is built [29]. For the interested audience, Repp [59] provides a good overview of issues, experimental methods and findings surrounding the scientific study of CP.

The issue of CP also holds relevance for areas closer to HRI. Moore [45] has proposed that the Uncanny Valley [47] effect may be a particular manifestation of CP, where, in the presence of multi-modal perceptual cues feeding into a category membership, conflicts in these cues could lead to the feelings of *discomfort* akin to those as described by Mori [47]. While this example specifically pertains to the relation between the physical appearance of an agent, and its physical behaviour, it may also be possible that similar effects occur between an agent's physical appearance and the acoustic behaviour it exhibits, for example, voice quality, or in the case of this article, the quality of NLUs. This latter notion does however presuppose that NLUs are indeed subject to CP, the validity of which is investigated in this article.

2 Experimental Setup

In psychological experiments, the typical methodology for testing CP involves presenting subjects with a stimulus continuum in which there are at least two *prototype* perceptual categories represented, with all other members of the continuum providing equal, linear transitions between these prototypes. This presupposes that CP is occurring and that at least one readily established and recognisable categorical boundary exists at some point along the continuum. Subjects are then asked to complete two tasks: a *discrimination* task and an *identification* task.

The purpose of the discrimination task is to determine whether subjects exhibit a perceptual *difference* between two stimuli. Are the two stimuli perceived as members of the same category? This is done by presenting stimulus pairs (usually neighbouring stimuli) from the continuum and asking subjects whether they judge them as *similar* or *different* without explicitly stating class membership.

The identification task entails subjects explicitly assigning category membership to each stimulus individually. This is done by presenting a single stimulus and asking subjects to rate it in some way, where the rating metric relates to the underlying representative categories (e.g. labels or sliders for dimensions, in the case of affect). The results for these two tasks together are then used to assess whether CP is occurring.

From the discrimination task, indications of CP are that subjects rate neighbouring stimuli that cross a categorical boundary as different, while they rate neighbouring stimuli that sit within a categorical region as similar. Thus, one would expect to see the frequency of “different” ratings increase as the neighbouring stimulus pairs approach a categorical boundary, providing an inverted “V” discrimination profile (see the discrimination profile line in Fig. 1).

In the identification task, indications of CP are when there are at least two clusters of neighbouring stimuli each situated near prototype stimuli that are closely rated, with clusters being separated by a sharp change (or step) in the identification rating. This represents the crossing of a categorical boundary and is characterised by a clear step in the class membership profiles (see the class membership profiles in Fig. 1).

Given these characteristics (and referring to Fig. 1), it is possible to formalise the following conditions under which CP may be said to occur:

- C_1 The two ends of the stimulus continuum receive class membership ratings during the Identification Task that are significantly different.
- C_2 Subjects rate neighbouring stimuli in the continuum that are near a prototype stimulus as “different” to a degree that is not statistically above chance (this indicates the presence of a categorical region).
- C_3 Subjects rate neighbouring stimuli that lay in the middle of the continuum as “different” to a degree that is above statistical chance, forming an inverted “V” shape in the discrimination profile (this indicates the presence of a categorical boundary).

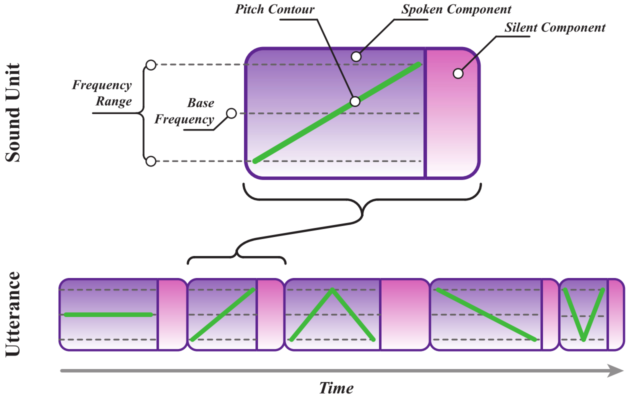

Fig. 2 Illustrative example of the anatomy of an NLU as used in this article

- C_4 Stimuli that are near a particular prototype stimulus have the same class membership rating as the prototype and a significantly different rating to the *other* prototype stimulus in the continuum, forming a *step* function.
- C_5 The peak in discrimination ratings during the Discrimination Task coincides with the step in the class membership rating in the Identification Task.

This experiment adopts the same basic CP methodology with some minor alterations to serve the focus on HRI, namely that utterances were embodied in an Aldebaran Nao robotic platform and a facial gesture tool was used for capturing affective ratings from subjects (see Sect. 2.3). The remainder of this section details the experimental set-up, covering the stimuli, the three different tasks that were completed, and the overall experimental procedure.

2.1 Utterance Stimuli

The stimuli, or utterances, used in this experiment consist of a collection of concatenated *Sound Units* and have four parameters used to characterise the utterances. Referring to Fig. 2, each sound unit has an *auditory* component and a *silent* component, the ratio of which is defined as the *Pause Ratio* (a small pause ratio value results in a long audible component and a short silent component). The auditory component consists of a single sinusoidal wave form whose frequency is modulated over time and is defined as the *Pitch Contour*, while the silent component provides a means of splitting the different sounds units affording rhythmic modulations at the global utterance level. All sound units in an utterance have a *Base Frequency* (Hz) which the Pitch Contour is centered around, and a *Frequency Range* (Hz) which defines the maximum and minimum frequencies of each Pitch Contour. Both the Base and Frequency Range remain constant for all sound units within an utterance. Finally, the duration of the utterances can be modulated, and this is defined as the *Speech Rate*

Table 1 Utterance parameter configurations for each utterance in both stimulus Sets 1 and 2

Utter	Parameter configuration			
	Base freq	Freq rang	S. rate	P. ratio
0	1500	1500	6	0.05
1	1333.33	1333.33	5.5	0.166
2	1166.67	1166.67	5	0.2833
3	1000	1000	4.5	0.4
4	833.33	833.33	4	0.5166
5	666.67	666.67	3.5	0.633

Table 2 Pitch Contour specifications for the utterances in stimulus Set 1 and Set 2

Sound uit	Stimulus set	
	Set 1	Set 2
1	Flat	Rising–falling
2	Falling	Flat
3	Rising	Falling–rising
4	Falling–rising	Rising
5	Rising–falling	falling

(sounds units played per second). These acoustic parameters have been informed and inspired by findings and trends from the domains of speech synthesis [64] and psychology [1,2,62], and are intended to cater for, and promote overlap with these domains where possible.

For this experiment, a stimuli set of a total of 12 utterances was produced, comprised of two continua (Set 1 and Set 2) each consisting of six utterances (Utter-0 to Utter-5), each with a different Utterance Parameter configuration[5]. Within each continuum there were two prototype utterances (Utter-0 and Utter-5) separated by four utterances with linear transitions in the four utterance parameters (Table 1). Each utterance was comprised of five sound units, and across the two continua only the Pitch Contour specifications were different (see Table 2), whilst within each continuum these Pitch Contour specifications remained the same for each utterance.

The parameter configurations of the two *prototype* utterances came from a previous experiment [56] where subjects were able to distinguish between the two different parameter configurations, which results in significantly different ratings along the Dominance dimension of the AffectButton. As such, these two parameter configurations were used to represent the two extremes of each stimulus continuum. Spectrograms of the utterances are illustrated in figure 3. Note the differences in the Pitch Contour specifications as outlined in Table 2.

2.2 Embodying Utterances in a Robot

Utterances were embodied in an Aldebaran Nao humanoid platform. The Nao is a small humanoid robot, standing 58cm tall and boasting a broad variety of different sensors. This platform has also become an attractive platform to use in HRI research due to its relative low cost, and human-friendly and aesthetically pleasing design. The motivations for this choice of platform were threefold. Firstly, our previous research [55] has suggested that a combination of NLUs and a small humanoid morphology is deemed as more appropriate by users than NLUs combined with a zoological morphology such as Sony's Aibo robot.

Secondly, it is a common observation that robots with a anthropomorphic humanoid morphology evoke strong responses from people both young and old [8]. Coupling this with the general observation that people tend to treat computer based technologies as socially competent [58], it was felt that the notion of attributing emotional states to a robot based upon the sounds that it made would be plausible to subjects.

Thirdly, we opted to use a standard platform that is now widespread in research labs around the world rather than a unique, custom built platform. Not only does this make replication efforts easier to facilitate (as one does not have to *build* a robot specifically for replication purposes), but the Nao represents the current state-of-the-art in social robotics platforms from a commercial perspective also, with increasing uptake in real world applications. We feel that this is a valuable consideration as this the Nao is a likely candidate when it comes to real-world users encountering a social robot outside a research/lab setting in the near future.

With respect to the proxemic arrangement, the robot was located on the table and stood behind the subjects laptop such that the eye level of both the human and the robot was approximately the same as it has been shown that differences in hight and distance can impact peoples' perception of and behaviour toward robots [49,54]. The human and robot were separated by approximately 60cm. While subjects were free to touch the robot, at no point through the experiment was this a necessity (only when the subject *asked* to touch the robot was this allowed). Only the experimenter was required to physically touch the robot.

2.3 Measuring Affect

To measure affect, a facial gesture tool—the *AffectButton*—was used to capture affective ratings from subjects, as opposed to the more typical approach of using categorical

[5] To listen to the utterances, please refer to the Online Resources. Resources 1–6 are the utterances in Set 1, and resources 7–12 are for Set 2.

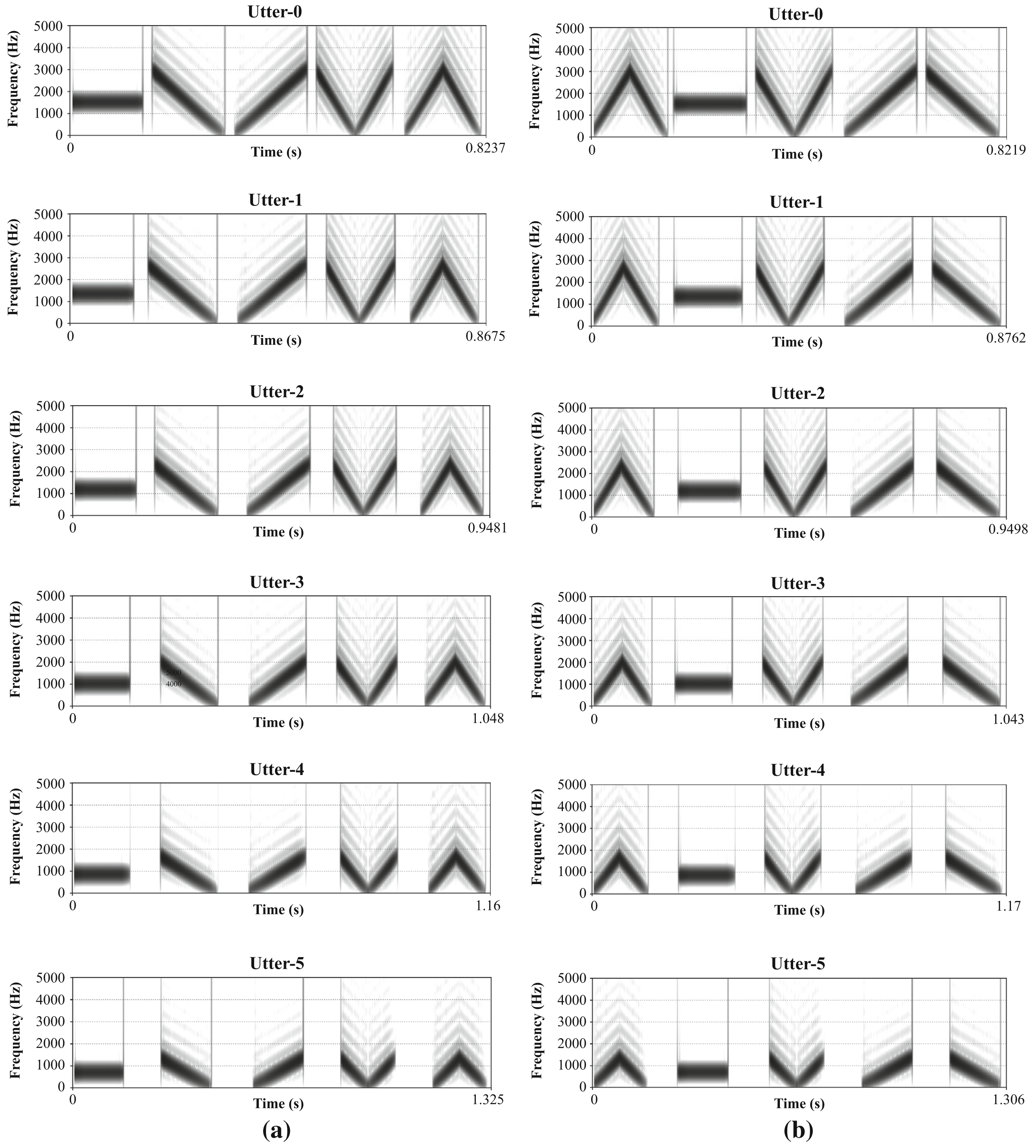

Fig. 3 Spectrograms of the six Utterance Stimuli in each of the two Sound Sets. *Top* Utter-0. *Bottom* Utter-5. **a** Stimulus Set 1. **b** Stimulus Set 2

labels/adjectives/check lists [53, Chapter 5], or sliders to represent affective dimensions such as FEELTRACE [14] or the Self Assessment Manikin (SAM) [41]. The AffectButton [9] is an open source[6], dynamic facial expression tool designed for obtaining explicit affective ratings from subjects in a real-time, simple and intuitive manner.

As the user moves the mouse cursor on screen, the tool dynamically interpolates between nine prototypical facial expressions (see Fig. 4 and Online Resource 13), encoding the facial expression into a coordinate within a 3D *affect space* where the three axis correspond to *Pleasure*, *Arousal*

[6] Python and Java source code for the AffectButton can be downloaded at http://www.joostbroekens.com/

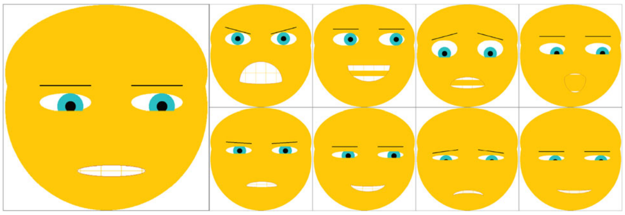

Fig. 4 AffectButton prototype facial expressions with PAD values. From left, clockwise: Neutral (0,−1,0), Angry (−1,1,1), Excited (1,1,1), Scared (−1,1,−1), Surprised (1,1,−1), Annoyed (−0.5,−1,0.5), Happy (0.5,−1,0.5), Sad (−0.5,−1,−0.5), Content (0.5,−1,−0.5). Adapted from Broekens et al. [10]

and *Dominance* (PAD) each with the range [−1, 1]. In practice however, arousal is calculated as the hypotenuse of the pleasure and dominance dimensions, thus providing a 2D to 3D mapping. As a result of this, pleasure corresponds to the horizontal movement of the mouse cursor, while Dominance corresponds to vertical movement[7].

This tool has been employed primarily for its intuitive nature and the fact that the underlying affective dimensions are essentially hidden from the user through being encoded into facial gestures. This is useful as classically it is difficult to explain the nature of affective dimensions, and their use in affective measuring tools, to subjects [9]. This in turn can lead to reduced inter-rater reliability. Moreover, as the Nao robot does not have an expressive or animated face, using facial gestures is a intuitive manner of having subjects assigning affective states to the robot. Furthermore, less time is required to explain to subjects how to use the tool, when compared with other metrics such as affective dimensions.

2.4 Labelling Task

In order to use the AffectButton as an effective tool for recording affective ratings, it is important to ascertain how coherent subjects are in their use of the tool. Do they all use the tool in the same way? Thus, this task was aimed at forming an impression of the overall coherence between subjects in their use of the AffectButton and whether they assigned similar facial gestures to a particular emotional interpretation.

Subjects were given some time to familiarise themselves with the tool and explore the range of facial gestures that that may be generated, and the associated mouse cursor locations onscreen. This process consisted of the experimenter explaining that by moving the mouse to different locations onscreen, the facial gesture of the onscreen face would change. Subjects were encouraged to verbalise their interpretations of the facial expressions that they saw as they moved the mouse onscreen. At no point were subjects told that a certain mouse location would yield a particular labelled facial expression (e.g. “Moving the mouse to the top-left of the box makes an ‘angry’ face”). Rather, the associations between different facial gesture interpretations and their mouse locations were left to be decided by the subject. This process took no longer than 5 min.

Once subjects felt that they were familiarised, an affective label was then displayed on the subject’s laptop screen above the button. Subjects were then asked to assign a facial gesture to the label by moving the mouse to a location yielding the desired facial gesture (Fig. 5a). The affective labels that were used were: *Happy*, *Excited*, *Angry*, *Annoyed*, *Surprised*, *Scared*, *Sad, Calm* and *Relaxed* and were presented in a random order (the respective affective coordinates for each label are show in Table 3). This choice of labels was motivated by the prototype facial expressions that are hard coded in the AffectButton and the overlap with the theory of *basic emotions*[8] [18,53]. As such, it was considered that a wide audience of subjects (e.g. children, adults, individuals from different cultural backgrounds and mother tongues, etc.) would also be familiar with these affective labels and facial gestures that would represent them.

2.5 Discrimination Task

The discrimination task was performed using an AX discrimination paradigm [12,26], where two stimuli were presented in pairs, sequentially (but randomly ordered), and subjects were asked to report whether they thought the robot *felt* different or the similar between the stimuli. On their laptop screen, subjects could choose from either “same”, “different” or “don’t know” options (Fig. 5b). The experimenter presented stimulus pairs by tapping the Nao on the head on the touch sensor. At each head touch the robot played the two utterances in succession, after which the three response options were then displayed on screen with the cursor reset to the centre of the screen. Centering the mouse cursor was done in order to avoid subjects hovering the cursor over a particular response button.

In total, each subject rated 13 pairs of utterances in this task. The first three pairs were test pairs whose order remained constant across all subjects (and are not used in the results analysis). They consisted of one stimulus pair with two extreme prototype utterances, one pair with identical utterances and one pair with neighbouring utterances. All of these stimuli were different from the actual experiment stimuli as they had a different *pitch contour*. In each Stimulus Set, neighbouring utterances were paired (e.g. Utter-0 vs. Utter-

[7] Broekens et al. [9] provide a detailed description of the AffectButton functionality and so this will not be described here.

[8] The *basic emotion* theory as proposed by Ekman and Friesen [19] states that there are certain facial behaviours which are universally associated with particular emotions, namely *anger*, *happiness*, *sadness*, *surprise*, *fear* and *disgust*.

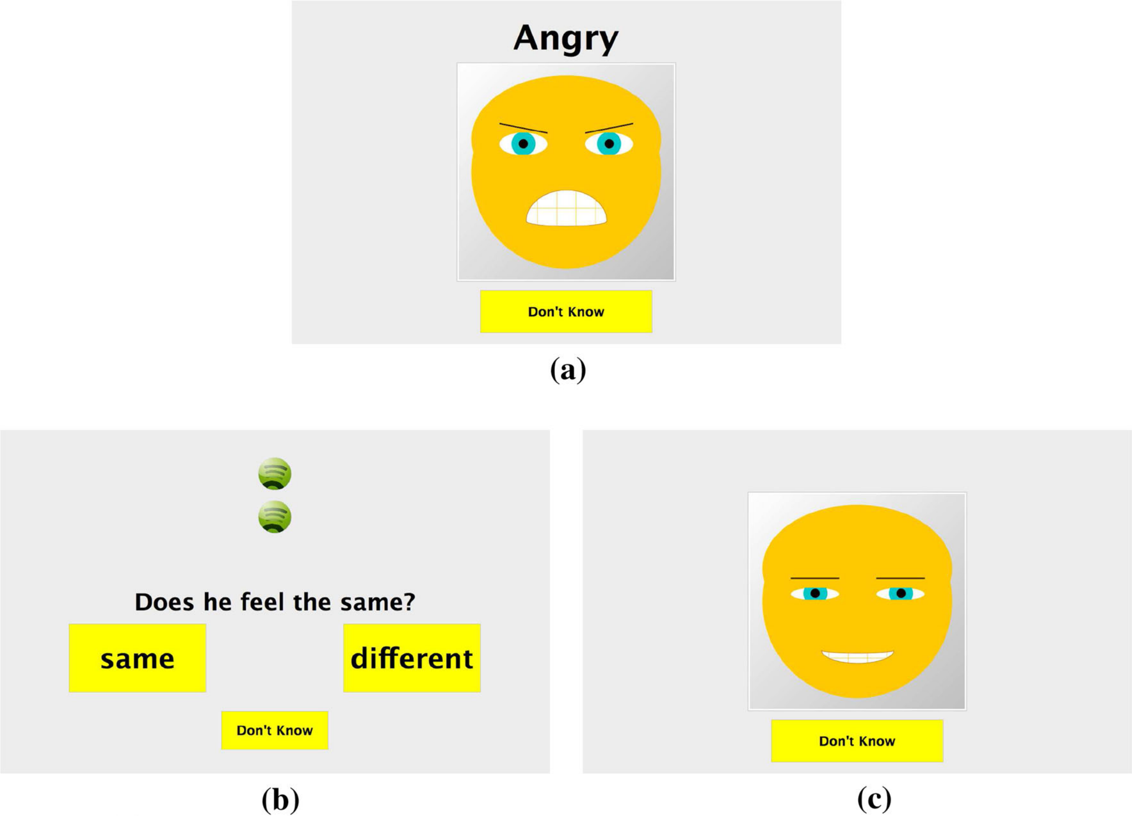

Fig. 5 Images of the subjects' laptop screen during each of the three tasks in the experiment. **a** Labelling task. **b** Discrimination task. **c** Identification task

Table 3 Affective co-ordinates in the AffectButton affect space of the labels (and associated prototypical facial gestures) used during the labelling task

Label	Affect space coordinate		
	Pleasure	Arousal	Dominance
Angry	-1	1	1
Annoyed	-0.5	-1	0.5
Happy	0.5	1	0.5
Excited	1	1	1
Sad	-0.5	-1	-0.5
Scared	-1	1	-1
Surprised	1	1	-1
Calm	–	–	–
Relaxed	–	–	–

Note that calm and relaxed are not prototypes used in the AffectButton

1, Utter-1 vs. Utter-2, ect) accounting for the remaining ten utterance pairs.

2.6 Identification Task

This task involved presenting subjects with a single utterance stimulus and asking them to provide an affective interpretation by assigning a facial expression on the AffectButton to their affective interpretation of the utterance (Fig. 5c). For this task, a simplified version of the AffectButton was used (see Online Resource 14), where the facial gestures were limited to interpolate between only the *sad*, *neutral*, *happy* and *excited* prototypes as the mouse cursor was moved horizontally (vertical movement had no effect). In doing this, the Pleasure value was modulated via the horizontal cursor movement, and Dominance (d) was then set equal to Pleasure (p) with both values falling in the range of $-0.5 \leq p \leq 1.0$ and $p = d$.

This Pleasure/Dominance mouse position mapping is shown in Fig. 6, with the corresponding prototypical facial expressions at these PAD coordinates in the AffectButton. By placing the extreme facial expressions at the upper and lower quartiles of the range of mouse movement, expression selection became a more cognitive task, avoiding subjects swinging to the extreme locations of the AffectButton (see Fig. 6).

During typical psychological CP identification tasks it is common for subjects to assign category membership by selecting from a small set of category labels (e.g. happy, sad, angry), however doing this explicitly promotes the notion of splitting the stimulus continuum into two or more discrete

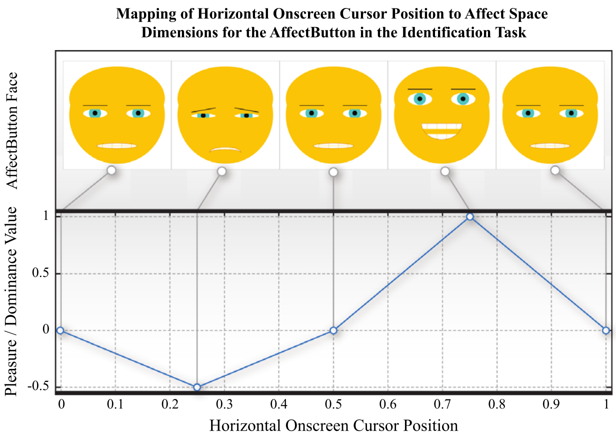

Fig. 6 Plot of the pleasure/dominance values as a function of the horizontal onscreen cursor position and the resulting AffectButton prototype expressions associated with the PAD values (from *left* to *right*): *neutral* (0, −1, 0), *sad* (−0.5, −1, −0.5), *neutral* (0, −1, 0), *excited* (1, 1, 1), *neutral* (0, −1, 0)

categories. The use of the AffectButton overcomes this by presenting subjects with a continuous scale of measurement. By using a continuum of possible facial expressions subjects are not forced to make an explicit categorical distinction, rather it leaves room for any CP to present itself in a more unrestricted manner.

2.7 Experimental Procedure

Subjects were recruited through advertisements located around the university campus, and the experiment took place within a lab setting. Subjects were rewarded with £5 in cash at the end of the experiment.

All utterance stimuli were played through the Nao's built in speakers, embodying the utterances in the robot. To further the notion of embodiment and agency, the Nao was also programmed to exhibit some neutral behaviours (random gazing, shifting weight from foot to foot and subtly moving the arms and fingers) in order to provoke the "illusion of life" and agency in the subjects.

A laptop was placed in front of the subject and was used to capture their responses for each task. The touch sensors on the robot's head were used to play/repeat utterances and were used only by the experimenter. A second laptop was operated by the experimenter and was used to orchestrate and monitor the overall experiment from a global perspective. The experimental arrangement is shown in Fig. 7.

Subjects were instructed to provide affective interpretations of the utterances made by the Nao, and that there were no correct or incorrect answers. Furthermore, they were told that the sounds were pre-recorded and were not influenced by the physical interactions with the robot or the motion behaviour displayed by the robot, nor the use of the AffectButton in any way. They were also told that they should try and respond as quickly as possible and use their "gut feeling" so as to avoid over thinking the problem. The Labelling Task was completed first as this task was intended to enable the subjects to become familiarised with the AffectButton tool. It is common practice in CP studies for the Discrimination Task to be completed before the Identification Task in order to avoid the process of assigning categories to stimuli biasing the process of discriminating between two stimulus pairs, a practice that was followed here. The total duration of the experiment was 20 minutes and once completed, subjects were free to ask any questions, and then presented with the £5 reward.

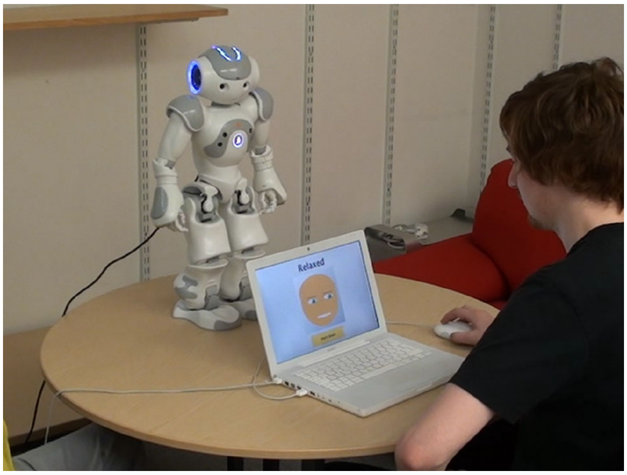

Fig. 7 Image of the experimental setup

3 Results

In total, 29 adult subjects took part: 17 women (mean age = 32.2, std = 10.6) and 12 men (mean age = 28.8, std = 6.8). All subjects were fluent english speakers and had mixed backgrounds with respect to their profession. These included, students (both post and undergrad), members of staff (e.g. cleaners, administrative staff, technical support, post-docs) and even members of the public who happened to spot the on-campus advertisements. All subjects were recruited from outside the Robotics Department. As a result familiarly with computer technology was varied.

This section presents the results for each of the three tasks, beginning with the labelling task, then the discrimination task, and finally the identification task.

3.1 Labelling Task Results

Figure 8 shows plots of the mean values and standard deviations for the ratings of each affective label. An initial visual inspection of the results revealed that the subjects provided

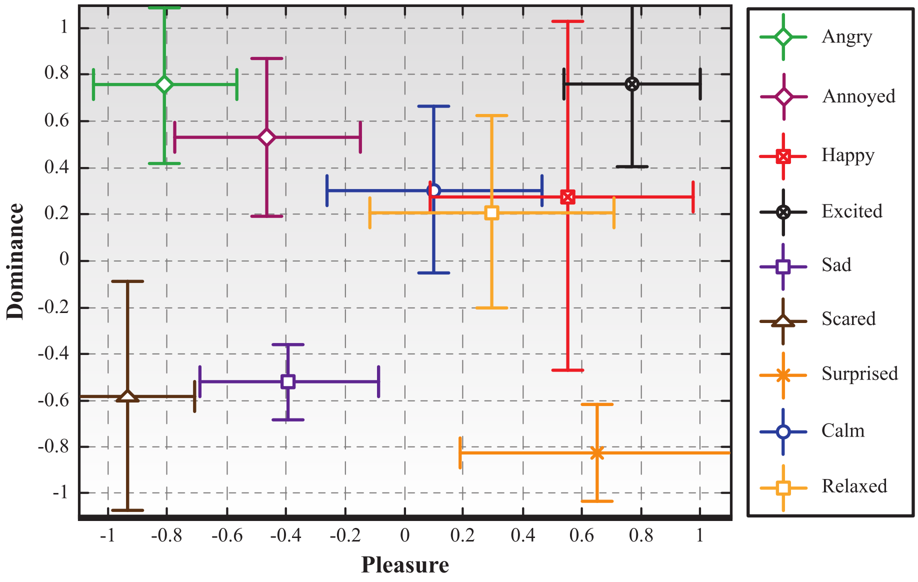

Fig. 8 Plots of the mean values and standard deviations of the ratings for each affective label in the labelling task

ratings that covered the majority of the AffectButton affect space.

The ratings for some of the labels were found to *not* follow a normal distribution, and as such, non-parametric statistical tests were employed to perform the analysis of the results in this task. As such, Friedman tests used to compare the ratings for two individual labels at a time. All of these tests were performed for each affective dimension individually.

These tests found that there were significant differences in how the affective labels were rated along the Pleasure ($\chi^2(8) = 184.35$, $p < 0.001$), Arousal ($\chi^2(8) = 143.55$, $p < 0.001$) and Dominance ($\chi^2(8) = 153.6$, $p < 0.001$) dimensions. Table 4 shows the χ^2 values calculated through the Friedman tests for pair-wise comparison. The values indicate that the subjects provided significantly different ratings between the majority of the labels overall, along at least one of the three dimensions.

3.2 Discrimination Task Results

Referring to Fig. 1, and specifically the *differential* profile, the indication of CP is that as the AX pairs of neighbouring stimuli approach the categorical boundary, these are rated as "different" to a degree that is above chance levels (i.e. tending toward a 100 % "different" rating) while utterance pairs that lay within a categorical region received discrimination ratings that remain at chance level (i.e. tending toward a 50 % "different" rating). As such, the discrimination profile follows an inverted "V" shape.

Figure 9 shows bar graphs of the percentage of "different" ratings for each of the neighbouring utterance pairs in Stimulus Set 1 and 2. Upon visual inspection, it can be seen that the ratings for both the sets appear to follow a general inverted "V" profile, with this being more prominent for the utterances in Set 1 than in Set 2. It is also notable that there is a general skew in the ratings, with the highest ratings occurring for the comparison of Utter-3 versus Utter-4 overall.

χ^2 goodness-of-fit tests were performed to identify which of the ratings were *above* chance level and which were at chance level, comparing each AX utterance pair against a uniform distribution (50 % "same" and 50 % "different"). These tests were performed for the results for results for Stimulus Sets 1 and 2 independently. Bars marked with a star in Fig. 9 represent overall rating percentages that are above chance.

With respect to Stimulus Set 1, AX pairs comparing Utter-2 vs. Utter-3 ($\chi^2(1, N = 28) = 3.572$, $p = 0.05$), and Utter-3 vs. Utter-4 ($\chi^2(1, N = 28) = 3.572$, $p = 0.05$) were found have ratings that were above chance levels. For the utterances in Stimulus Set 2, the table shows that overall, only the ratings for Utter-0 vs. Utter-1 ($\chi^2(1, N = 28) = 1.690$, $p = 0.194$) were *not* above chance. These results are summarized in Table 5.

3.3 Identification Task Results

A two-way (6 × 2), within-subjects, repeated measures ANOVA was performed for the Pleasure ratings using the six Utterances Parameter Configuration values and the Stimulus Set as the two factors. The ANOVA was followed up by post-hoc multi-comparison tests with Bonferoni Corrections for both the main effects and interaction.

The ANOVA found that there were statistically significant main effects due to both the Utterance Parameter Configuration ($F(1, 130) = 64.732$, $MSE = 6.386$, $p < 0.0001$) and the Stimulus Set ($F(1, 26) = 10.051$, $MSE = 0.988$, $p = 0.004$). There was no interaction effect found between the two factors ($F(5, 130) = 1.169$, $MSE = 0.101$, $p =$

Table 4 Results of the Friedman pairwise comparisons for the affective ratings for the affective labels in the Labelling Task. The table shows the $\chi^2(1)$ results and indicate the associated p-value for each dimension of the AffectButton affect space independently

Affective label	Dimension	Ang	Ann	Hap	Exc	Sad	Scar	Sur	Calm	Rel
Angry	P	–								
	A									
	D									
Annoyed	P	$14.29^{\ddagger}$	–							
	A	$17.64^{\ddagger}$								
	D	$12.45^{\ddagger}$								
Happy	P	$29^{\ddagger}$	$29^{\ddagger}$	–						
	A	1.47	$17.64^{\ddagger}$							
	D	0.31	5.83∗							
Excited	P	$25.14^{\ddagger}$	$29^{\ddagger}$	0.03	–					
	A	0.25	$18.62^{\ddagger}$	1.47						
	D	0.03	1.69	0.86						
Sad	P	$12.45^{\ddagger}$	0.31	$29^{\ddagger}$	$25.14^{\ddagger}$	-				
	A	$22.15^{\ddagger}$	2.13	$27^{\ddagger}$	$24.14^{\ddagger}$					
	D	$25.14^{\ddagger}$	$29^{\ddagger}$	$25.14^{\ddagger}$	$15.21^{\ddagger}$					
Scared	P	1.69	$18.24^{\ddagger}$	$29^{\ddagger}$	$29^{\ddagger}$	$25.14^{\ddagger}$	–			
	A	0.89	$14.44^{\ddagger}$	0.05	4.26∗	$20.57^{\ddagger}$				
	D	$25.14^{\ddagger}$	$25.14^{\ddagger}$	$25.14^{\ddagger}$	$15.21^{\ddagger}$	2.79				
Surprised	P	$25.14^{\ddagger}$	$25.14^{\ddagger}$	0.03	1.69	$25.14^{\ddagger}$	$29^{\ddagger}$	–		
	A	0.6	$18.62^{\ddagger}$	2.88	0	$24.14^{\ddagger}$	2.25			
	D	$29^{\ddagger}$	$29^{\ddagger}$	$29^{\ddagger}$	$21.55^{\ddagger}$	$15.21^{\ddagger}$	7.76∗			
Calm	P	$25.14^{\ddagger}$	$18.24^{\ddagger}$	$25.14^{\ddagger}$	$21.55^{\ddagger}$	$15.21^{\ddagger}$	$25.14^{\ddagger}$	$15.21^{\ddagger}$	–	
	A	$24.14^{\ddagger}$	6∗	$29^{\ddagger}$	$29^{\ddagger}$	0	$21.55^{\ddagger}$	$29^{\ddagger}$		
	D	$15.21^{\ddagger}$	2.79	$18.24^{\ddagger}$	2.79	$25.14^{\ddagger}$	$25.14^{\ddagger}$	$29^{\ddagger}$		
Relaxed	P	$25.14^{\ddagger}$	$21.55^{\ddagger}$	$11.57^{\ddagger}$	$9.97^{\dagger}$	$21.55^{\ddagger}$	$29^{\ddagger}$	$12.45^{\ddagger}$	7∗	–
	A	$19.59^{\ddagger}$	1.64	$23.15^{\ddagger}$	$24.14^{\ddagger}$	2.91	$15.38^{\ddagger}$	$19.59^{\ddagger}$	1.09	
	D	$12.45^{\ddagger}$	$9.97^{\dagger}$	$21.55^{\ddagger}$	0.86	$21.55^{\ddagger}$	$17.29^{\ddagger}$	$25.14^{\ddagger}$	1.69	

∗$p < 0.05$
$^{\dagger}p < 0.005$
$^{\ddagger}p < 0.001$

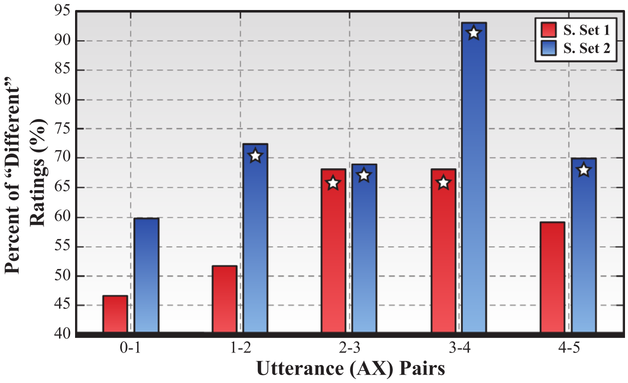

Fig. 9 *Bar graphs* showing the percentage of "different" ratings given by the adults for the neighbouring utterance AX pairs for both stimulus sets. *Bars* marked with a star are ratings found to be significantly above chance at the 0.05 level. The ratings shown in this figure are summarised in Table 5

0.328). Figure 10 show a plot of the ratings for the six utterances across the two stimulus sets, while Table 6 shows the exact mean values, standard error and 95 % confidence intervals respectively.

For the main effect due to the Stimulus Sets, the post-hoc tests revealed that the ratings for the utterances in Set 2 (mean $= 0.402$, 95 % CI [0.318 0.485]) received overall higher ratings than the utterances in Set 1 (mean $= 0.291$, 95 % CI [0.214 0.368]), $p = 0.004$.

With respect to the main effect due to the Utterance Parameter Configuration the post-hoc tests revealed that Utter-0 received the highest rating (mean $= 0.784$, 95 % CI [0.639 0.858]) and Utter-5 received the lowest rating (mean $= -0.095$, 95 % CI [-0.213 0.023]). All the other Utterances presented a negative slope of ratings (see Table 6). Utter-0 and Utter-1 were found not to be significantly different ($p = 1.0$), with both were found to be significantly different from all other utterances ($p < 0.05$). Similarly,

Table 5 χ^2 Goodness of fit tests for the adult subjects' comparison of neighbouring utterances in each of the stimulus sets during the discrimination task

Set	Utterance		Rating (%)	$\chi^2(1)$
	A	X		
1	Utter-0	Utter-1	46.667	0.310
	Utter-1	Utter-2	51.724	0.143
	Utter-2	Utter-3	67.517	3.572*
	Utter-3	Utter-4	67.857	3.572*
	Utter-4	Utter-5	59.259	0.926
2	Utter-0	Utter-1	60.000	1.690
	Utter-1	Utter-2	72.414	7.000**
	Utter-2	Utter-3	68.966	5.143*
	Utter-3	Utter-4	93.103	$20.571^{\dagger}$
	Utter-4	Utter-5	70.000	5.828*

* $p < 0.05$

** $p < 0.01$

† $p < 0.005$

Table 6 Mean values and 95 % Confidence Intervals of the ratings for the different utterances parameter configurations across the two sound sets as part of the identification task (see Fig. 10)

Set	Utter	Mean	SE	95 % Conf inter	
				Lower	Upper
Both	0	0.748	0.053	0.639	0.858
	1	0.680	0.050	0.577	0.784
	2	0.492	0.045	0.398	0.585
	3	0.177	0.053	0.067	0.286
	4	0.075	0.054	−0.036	0.187
	5	−0.095	0.057	−0.213	0.023
1	0	0.715	0.061	0.589	0.841
	1	0.633	0.073	0.483	0.783
	2	0.473	0.053	0.364	0.582
	3	0.049	0.071	−0.097	0.195
	4	−0.009	0.066	−0.145	0.126
	5	−0.113	0.075	−0.267	0.040
2	0	0.782	0.057	0.665	0.900
	1	0.728	0.053	0.619	0.837
	2	0.510	0.066	0.375	0.645
	3	0.304	0.066	0.169	0.440
	4	0.160	0.083	−0.011	0.332
	5	−0.076	0.064	−0.208	0.055

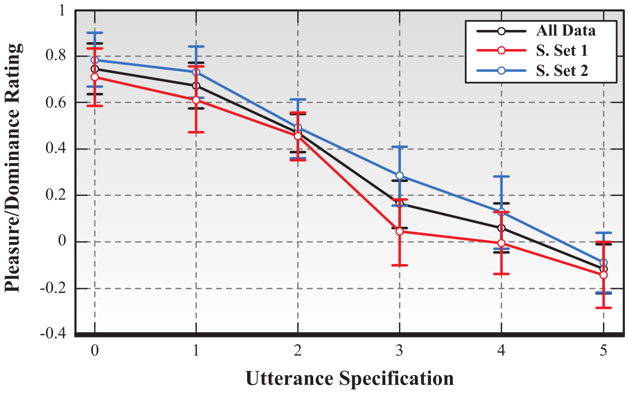

Fig. 10 Plot showing the mean values and 95 % confidence intervals of the ratings of each utterance parameter configuration across the stimulus sets. See Table 6 for the exact values

Utter-3 and Utter-4 were not found to be significantly different ($p = 1.0$) but too were significantly different from all the other utterances ($p < 0.05$). Finally, Utter-2 and Utter-5 were significantly different from all the other utterances ($p < 0.05$).

The post-hoc tests also revealed that there were important differences in the ratings for utterances across the two Stimulus Sets. When isolating the ratings for utterances in Set 1, it was found that Utter-0,1 and 2 received ratings that were not different to a statistically significant degree ($p > 0.132$), essentially forming one *cluster* of ratings. Similarly, Utter-3, 4 and 5 received ratings that were lower and were not significantly different ($p = 1.0$), forming a second cluster of ratings. Concretely, all utterance ratings *within* a cluster were not significantly different from each other, while all utterances ratings *between* the clusters were significantly different.

When isolating the ratings pertaining to utterances only in Stimulus Set 2. The post-hoc tests showed that no such cluster formation existed. Rather, the utterance ratings followed a linear negative slope. Utter-0 and Utter-1 received the highest rating and were not significantly different ($p = 1.0$) and Utter-6 received the lowest rating, with this rating being significantly different from all others ($p < 0.004$). All the other utterances received ratings that were not significantly different from their direct neighbours ($p > 0.126$).

4 Discussion

This section provides a discussion of the results obtained during this experiment, and the relevance of CP in when interpreting NLUs.

The findings of this experiment are discussed in a slightly broader perspective regarding HRI in Sects. 4.3 and 4.4.

4.1 Results and Discussion

The results of the labelling task show that the subjects were able to reliably associate different facial gestures with affective labels using the AffectButton. Figure 8 shows that indeed the *extreme* affective labels (Angry, Excited, Scared and Surprised) were rated differently; this is evidenced by the lack of overlap in the confidence intervals of the ratings. This validates the AffectButton as an appropriate tool for reporting affect.

With regard to the Discrimination and Identification Tasks and their relation to CP, these two tasks should be considered as a collective as neither can be used to confirm CP alone. During the Discrimination Task both the results for Stimulus Set 1 and Stimulus Set 2 tended to follow an inverted "V" shape, with the ratings located at the top of the V being statistically above chance while those at the bottom of the "V" were not significantly above chance levels, though this was more prominent for the ratings for Stimulus Set 1. This supports condition C_2 (that neighbouring stimuli near a prototype stimulus are rated as "different" to a degree that is not above chance) and C_3 (the neighbouring stimuli in the middle of the continuum are rated as "different" to a degree that is statistically above chance and forms an inverted "V" shape). This already suggests the presence of a categorical boundary in the stimulus continua.

It is interesting to see that in the case of Stimulus Set 1 the profile of the affective ratings in the Identification Task followed a step function (between Utter-2 and Utter-3), which supports C_4 (that stimuli near a prototype stimulus have the same class membership and have a significantly different rating to the stimuli near the other prototype stimulus) while this was not the case of Stimulus Set 2. Furthermore, this step coincides roughly with the peak in the corresponding results of the Discrimination Task (Utter-2 vs. Utter-3, and Utter-3 v.s Utter-4), which supports C_5 (that the peak of the inverted "V" in the differentiation profile and the step in the category membership profile occur at the same location in the continuum). When marrying the results of these two tasks, there is strong evidence for the presence of a categorical boundary along the Pleasure/Dominance dimension of the AffectButton affect space.

The Identification Task found that there was a significant main effect due to the difference in the Stimulus Sets, (which were differentiated by their Pitch Contour specification). This suggests that the Pitch Contour does appear to play a role in how subjects affectively interpret an NLU, though the magnitude of this effect remains unclear. The results of the Discrimination Task lend support to this as there was a notable visual difference in the distribution of results between Stimulus Sets 1 and 2 (see Fig. 10).

During the Identification Task, the subjects' affective ratings of the two prototype utterances (Utter-0 corresponding to positive, and Utter-5 corresponding to negative) that were significantly different. This supports C_1, confirming that the two prototype utterances did indeed represent two different categories (and by category we mean two different regions of the affect space).

Considering the results of all three tasks we conclude that there is strong evidence suggesting that prototype utterances of each of the two stimulus continuums did represent two different categorical regions with a categorical boundary in the middle. Moreover, we conclude that in the case of Stimulus Set 1 subjects exhibited behaviour that is consistent with the presence of CP.

4.2 Methodological Remarks

While the results of the experiment indicate that CP is occurring, this experimental set up is not without methodological provisos. Gerrits and Schouten [26] have argued that CP findings depend upon the type of Discrimination task that has been employed. Here the AX paradigm is used, where subjects are presented with two stimuli and have to say whether they feel that they are the "same" or not. While this task has low cognitive load, there is a bias toward subjects providing more "different" ratings as there are not other pairs (presenting identical stimulus pairs) in the trail with which comparison may be made [26,63]. As such, where neighbouring stimuli might be expected to fall within the same categorical region, these have a higher chance of being deemed as "different". The result is a reduced inverted "V" differential profile.

Other paradigms commonly used are the ABX and 4IAX comparison tasks. In the ABX task, subjects are presented with three stimuli, two of which are the same, and subjects must identify which of the first two stimuli (A or B) is the same as the last (X). The 4IAX task is a far more cognitively demanding task than either the AX or ABX tasks, where subjects are presented with two pairs per trail (e.g. AA-BA, AB-BB, etc.) and subjects must identify which of the two pairs contains the odd one out (for example, AA or BA).

In the case of the ABX task, it is common to find a bias due to presentation order where the B and X stimulus are more likely to be identified as the same than A and X. This is theorised to be linked (in cases which use auditory stimuli) to the loading on auditory memory. The 4IAX task is a method that holds less overall bias, however has high cognitive loading, and requires that subjects listen to a total of four stimuli rather than 2 per trial. This study has employed the AX as it was deemed to be the least demanding paradigm to use, which was an appealing factor when using auditory rather than visual stimuli.

During experiment and explicitly during the discrimination task, the robot was assigned with the male gender (see Fig. 5b). It can be argued that this could have skewed the results considerably. However, this was a deliberate choice with respect to the methodological design. The rationale for this was that by assigning an explicit gender to the robot that was constant for all subjects, we remove a potentially confounding factor where the gender of the robot would be left to the subjects discretion. Furthermore, evidence from HRI research (e.g. [22,67,70]) suggests that subjects respond differently to robots that that have different attributed genders, reinforcing the need to explicitly declare the robots gender rather than letting subjects decide for themselves.

Given that the Nao was programmed to exhibit some random neutral behaviours to help bring it to life and that the utterances were played by touching the robot on the head, it can be argued that multi-modal HRI was taking place and that the touch interaction should be accounted for within the results analysis a should the differences in motions of the robot. However, this is not the case as both the random motion of the robots "neural behaviour" and the fact that the touch was only instigated by the experimenter were held constant for all subjects. Thus, there should be no confounding factor introduced via either of these two aspects of the experiment.

There is also an issue that revolves around the notion that by placing the Labelling Task before the Identification Task, one might be introducing a bias where subjects would exhibit CP more readily (due to the associations made between certain regions of the affect space, and discrete affective labels). Whether or not such as bias is occurring is difficult to determine, however, we propose a line of reasoning that suggests that even if such a bias is occurring, there is little difference between how subjects would use the AffectButton, and if they were using affective labels to assign category membership during the identification Task. The latter of which has been common practice in many CP studies over the years [59]. It was considered more important to ensure that subjects were indeed familiar with how to use the AffectButton before they performed the Identification Task, and thus this was facilitated by having subjects perform the Labelling Task first.

Finally, the actual use of the AffectButton as a means for capturing affective ratings in this experiment may be subject to criticism. Given that it has been well established that humans exhibit CP of facial expressions, it may be argued that this experimental setup has an inherent bias that would promote evidence supporting the presence of CP.

It is difficult to asses whether such a bias is indeed taking place, as well the magnitude that it may have. One would need to use a completely different measurement tool to gauge this, which in itself can lead to more criticisms—for instance whether a different tool does indeed provide a robust representation of an affective interpretation (it is argued here that using facial expressions does this). However, setting this aside, in Sect. 4.4 we argue that the findings of CP in this study are still relevant to the field of HRI.

4.3 Categorical Perception and NLUs

Let us begin with a discussion regarding whether the confirmation that NLUs indeed are subject to CP is actually *surprising*. Initially it may seem that subtle changes in the acoustics of NLUs do not equate to subtly different affective inferences, is indeed a novel and unexpected result. However, when one considers the vast other domains in which CP is observed (see Sect. 1.2) it is clear that CP is common place in sensory processing and perception [29]. As such, here we do not wish to purely celebrate the fact that NLUs too are subject to some form of CP. We use this section to outline why this finding has importance for the design of HRI systems.

Firslty, we wish to draw attention to *level* at which it has been found to be occurring: the level of inferred affective meaning. In comparison emotional facial gestures (which are also subject to CP of inferred affective meaning), NLUs are a novel social display to observe and decode. As such, we draw attention to the fact that people have exhibited CP of these displays made by a novel agent, and that this has occurred within a brief time period. The suggestion is that when engaging with new interactive robotic technologies that utilise novel social cues, people are quick to decode and perceive these as having meaningful and familiar affect-laden content, having had little prior experience. This may be seen as a affirmation of the work by Heider and Simmel [30] who observed that people naturally attribute human-like characteristics to animated geometric shapes displayed on a screen.

The fact that peoples' affective interpretations are categorical impacts the generation of NLUs as well as the general use of NLUs during social HRI. With respect to the use of NLUs in HRI, the finding of CP suggests that subtle differences in the acoustic features of utterances do not necessarily translate to subtle differences in how these utterances are interpreted with respect to their affective meaning. Rather, the results show that utterances can be subject to a "magnet effect" whereby they are drawn to prototypical affective interpretations (e.g. happy, angry, sad, act). This is an important insight when it comes to attempting to predict how a given utterance may be interpreted by a subject/listener and highlights that the mapping between the parameters of an NLU and a users affective interpretation are non-linear and complex.

This insight impacts the design of system to automate NLU generation. The goal of this is move away from handcrafting NLUs for specific experiments and develop a system where a desired affective interpretation on the users part may be specified and an appropriate NLU be generated and synthesized. Fundamental to this is the need to uncover the mapping between the acoustic parameters of an utterance (and any external factors) and how that utterance as a whole is affectively interpreted. Given that peoples' interpretations are subject to CP, this impacts the techniques that may be used to uncover this mapping. For example, linear regression is likely unsuited to this task. Currently, nearly all of the previous NLU research has used hand-crafted and prerecorded utterances as stimuli. The only exception is the work of Schwent and Arras [65]. As such, this aspect of NLUs is in it's infancy and required considerably more attention.

In the experiment, utterances were presented in a scenario with minimal situational context[9], and yet found evidence showing CP of NLUs. However, real-world HRI is rich in context: all HRI contains implicit situational context. As such, a valuable extension to the experiment presented in this chapter would be to investigate how the use of NLUs within a scenario with a more defined situational context may may differ from the use in context-free settings, and whether the perceptual magnet effect may be more prominent in such situations. This is something that has been addressed in part in previous work [57], where it was found that people's affective ratings of NLUs are primarily driven by the context within which they are used. Though it was found that people are still sensitive to the changes in acoustic properties of NLUs, something that is also evidenced in this work.

4.4 Categorical Perception and Multi-modal Human–Robot Interaction

The notion of CP holds relevance for other modalities (other than NLUs) employed for affective displays during social HRI also. For example, recently, there have been developments in the methods of allowing robots to make affective displays through bodily gestures (c.f. [3,34,68]), and in retro-projected technologies that provide robots with animated faces that can express realistic emotions (cf. [15,16,40]). There is also a keen interest in developing speech synthesis engines that are capable of conveying affect in a realistic and convincing manner to improve natural language interfaces (c.f. [64]).

Real-world HRI is also not uni-modal, but rather multi-modal, which provides many more cues which can be drawn upon to gain understanding of the interaction. With respect to multi-modal HRI and utterances of a non-linguistic nature, previous work has found that for both GS and NLUs, combining utterances with facial gestures significantly strengthens the interpretation that people have of the robot's behavior [32,75]. In short, we predict that multi-modal HRI will amplify the magnet effect significantly making it more pronounced. This is something that is on our agenda for future work. However we have started with investigating NLUs in a uni-modal format as this will help us disentangle the effects of each single modality when studying NLUS as part of multi-modal HRI.

5 Conclusion

We presented an experiment aimed at probing whether adults exhibit Categorical Perception when affectively interpreting non-linguistic Utterances made by a social robot. The methodology employed closely followed traditional methodologies matured in experimental psychology, with some minor adaptations – stimuli were embodied in a humanoid robotic platform, and a novel facial expression tool was used as a means of capturing affective ratings from subjects.

The experiments provide two important insights. They confirm that NLUs are not just beeps and clicks, instead they can convey affect. In addition, when interpreting NLUs there is a magnet effect: the interpretation of NLUs is drawn towards prototypical emotions. However, categorical perception seems to be utterance specific and not all gradual variations in NLUs are interpreted categorically.

The previous literature on both NLUs and GS has primarily focused on experiments in which stimuli have been created to represent discrete affective states (e.g. happy, angry, sad, scared, disgust and surprise), measuring a person's ability to correctly infer the desired affective state from the stimuli. Such experiments yield little insight as to the dynamics of affective inference across the different affective states portrayed by the stimuli (e.g. [7,31,32,36,38,50,74,76]), something that is equally as important as understanding what drives the correct recognition of utterances. With respect to NLUs, the experiment presented here addresses this grey area directly using a well established methodology, finding that the dynamics of affective interpretation are variable: transitions between affective interpretations can be subtle and gradual, or sudden and abrupt. Moreover, this appears to be utterance specific, complicating the matter further. What governs this with respect to the properties of utterances remains unclear, highlighting that this is an area that requires further investigation.

These insights can help inform the automated generation of affective non-linguistic Utterances. It is clear that the mapping between an affect space and the parameter space of NLUs is complex, but the fact that NLUs are readily recognised as prototypical emotions is a strong advantage. Automating this mapping in an "on the fly" manner holds promise, as it does not rely on pre-recorded samples, but it remains a challenging and multi-faceted problem. It is for example not completely clear as to what extent the situational context, or multi-modal displays may override or even drive the affective interpretations of NLUs. These are issues that require investigations in future work in order to fully understand the potential of this modality.

Acknowledgments This work was (partially) funded by the EU FP7 ALIZ-E project (Grant 248116).

[9] By "minimal situational context" we refer to the fact that the robot did not engage in vocal interaction, nor did the robot and subject engage in a complex interaction (e.g. a game of chess). Subjects were simply asked to rate sounds made by the robot, with the knowledge that the sounds were pre-recorded, and touching the robot on the head would play the next sound. In this scenario, there are no other cues that subjects can turn to in order to aid in the interpretation of the sounds.

References

1. Banse R, Scherer K (1996) Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70(3):614–636
2. Banziger T, Scherer K (2005) The role of intonation in emotional expressions. Speech Commun 46(3–4):252–267
3. Beck A, Stevens B, Bard KA, Cañamero L (2012) Emotional body language displayed by artificial agents. Trans Interact Intell Syst 2(1):1–29
4. Bimler D, Kirkland J (2001) Categorical perception of facial expressions of emotion: evidence from multidimensional scaling. Cogn Emot 15(5):633–658
5. Blattner M, Sumikawam D, Greenberg R (1989) Earcons and icons: their structure and common design principles. Hum Comput Interact 4:11–44
6. Bornstein MH, Kessen W, Weiskopf S (1976) Color vision and Hue categorization in young human infants. Human perception and performance. J Exp Psychol 2(1):115–129
7. Breazeal C (2002) Designing sociable robots. The MIT Press, Cambridge
8. Breazeal C (2003) Emotion and sociable humanoid robots. Int J Hum Comput Stud 59(1–2):119–155
9. Broekens J, Brinkman WP (2013) Affectbutton: a method for reliable and valid affective self-report. Int J Hum Comput Stud 71(6):641–667
10. Broekens J, Pronker A, Neuteboom M (2010) Real time labelling of affect in music using the affect button. In: Proceedings of the 3rd international workshop on affective interaction in natural environments (AFFINE 2010) at ACM multimedia 2010. ACM, Firenze, pp 21–26
11. Cassell J (1998) A framework for gesture generation and interpretation. In: Cipolla R, Pentland A (eds) Computer vision for human–machine interaction. Cambridge University Press, Cambridge, pp 191–216
12. Cheal JL, Rutherford MD (2011) Categorical perception of emotional facial expressions in preschoolers. J Exp Child Psychol 110(3):434–443
13. Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Commun 40(1–2):5–32
14. Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M (2000) 'FEELTRACE': An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA tutorial and research workshop (ITRW) on speech and emotion. Newcastle, pp 19–24
15. Delaunay F, de Greeff J, Belpaeme T (2009) Towards retro-projected robot faces: An alternative to mechatronic and android faces. In: Proceedings of the 18th international symposium on robot and human interactive communication (ROMAN 2009). Toyama, pp 306–311
16. Delaunay F, de Greeff J, Belpaeme T (2010) A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In: Proceedings of the 5th international conference on human–robot interaction (HRI'10). ACM/IEEE, Osaka, pp 39–44
17. Duffy BR (2003) Anthropomorphism and the social robot. Robot Autonom Syst 42(3–4):177–190
18. Ekman P (2005) Basic emotions. In: Dalgleish T, Power M (eds) Handbook of cognition and emotion. Wiley, Chichester, pp 45–60
19. Ekman P, Friesen W (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124–129
20. Embgen S, Luber M, Becker-Asano C, Ragni M, Evers V, Arras K (2012) Robot-specific social cues in emotional body language. In: Proceedings of the 21st international symposium on robot and human interactive communication (RO-MAN 2012). IEEE, Paris, pp 1019–1025
21. Etcoff N, Magee J (1992) Categorical perception of facial expressions. Cognition 44:227–240
22. Eyssel F, Hegel F (2012) (S)he's got the look: gender stereotyping of robots. J Appl Soc Psychol 42(9):2213–2230
23. Franklin A, Davies IR (2004) New evidence for infant colour categories. Br J Dev Psychol 22(3):349–377
24. Funakoshi K, Kobayashi K, Nakano M, Yamada S, Kitamura Y, Tsujino H (2008) Smoothing human-robot speech interactions by using a blinking-light as subtle expression. In: Proceedings of the 10th international conference on multimodal interfaces (ICMI'08). ACM, Chania, pp 293–296
25. Gaver W (1986) Auditory icons: using sound in computer interfaces. Hum Comput Interact 2(2):167–177
26. Gerrits E, Schouten M (2004) Categorical perception depends on the discrimination task. Percept Psychophys 66(3):363–376
27. Goldstone RL, Hendrickson AT (2009) Categorical perception. Wiley Interdiscip Rev 1(1):69–78
28. Hackett C (1960) The origin of speech. Sci Am 203:88–96
29. Harnad S (ed) (1987) Categorical perception: the groundwork of cognition. Cambridge University Press, Cambridge
30. Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57:243–259
31. Jee E, Jeong Y, Kim C, Kobayashi H (2010) Sound design for emotion and intention expression of socially interactive robots. Intel Serv Robot 3:199–206
32. Jee ES, Kim CH, Park SY, Lee KW (2007) Composition of musical sound expressing an emotion of robot based on musical factors. In: Proceedings of the 16th international symposium on robot and human interactive communication (RO-MAN 2007). IEEE, Jeju Island, pp 637–641
33. Johannsen G (2004) Auditory displays in human–machine interfaces. Proc IEEE 92(4):742–758
34. Karg M, Samadani Aa, Gorbet R, Kuhnlenz K (2013) Body movements for affective expression: a survey of automatic recognition and generation. Trans Affect Comput 4(4):341–359
35. Komatsu T, Kobayashi K (2012) Can users live with overconfident or unconfident systems?: A comparison of artificial subtle expressions with human-like expression. In: Proceedings of conference on human factors in computing systems (CHI 2012). Austin, pp 1595–1600
36. Komatsu T, Yamada S (2007) How appearance of robotic agents affects how people interpret the agents' attitudes. In: Proceedings of the international conference on Advances in computer entertainment technology: ACE '07
37. Komatsu T, Yamada S (2011) How does the agents' appearance affect users' interpretation of the agents' attitudes: experimental investigation on expressing the same artificial sounds from agents with different appearances. Int J Hum Comput Interact 27(3):260–279
38. Komatsu T, Yamada S, Kobayashi K, Funakoshi K, Nakano M (2010) Artificial subtle expressions: intuitive notification methodology of artifacts. In: Proceedings of the 28th international conference on human factors in computing systems (CHI'10). ACM, New York, pp 1941–1944
39. Kuhl PK (1991) Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Percept Psychophys 50(2):93–107
40. Kuratate T, Matsusaka Y, Pierce B, Cheng G (2011) "Mask-bot": A life-size robot head using talking head animation for human–robot communication. In: Proceedings of the 11th IEEE-RAS international conference on humanoid robots (Humanoids 2011). IEEE, Bled, pp 99–104
41. Lang P, Bradley M (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Therapy Exp psychiatry 25(1):49–59

42. Laukka P (2005) Categorical perception of vocal emotion expressions. Emotion 5(3):277–295
43. Levitin DJ, Rogers SE (2005) Absolute pitch: perception, coding, and controversies. Trends Cognit Sci 9(1):26–33
44. Liberman A, Harris K, Hoffman H (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54(5):358–368
45. Moore RK (2012) A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Sci Rep 2:864
46. Moore RK (2013) Spoken language processing: where do we go from here? In: Trappl R (ed) Your virtual butler. Springer, Berlin, pp 119–133
47. Mori M (1970) The Uncanny Valley. Energy 7:33–35
48. Mubin O, Bartneck C, Leijs L, Hooft van Huysduynen H, Hu J, Muelver J (2012) Improving speech recognition with the robot interaction language. Disrupt Sci Technol 1(2):79–88
49. Mumm J, Mutlu B (2011) Human–robot proxemics: physical and psychological distancing in human–robot interaction. In: Proceedings of the 6th international conference on human–robot interaction (HRI'11), Lausanne
50. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Stud 59(1–2):157–183
51. Paepcke S, Takayama L (2010) Judging a bot by its cover: an experiment on expectation setting for personal robots. In: Proceedings of the 5th international conference on human–robot interaction (HRI'10). ACM/IEEE, Osaka, pp 45–52
52. Picard RW (1997) Affective computing. MIT Press, Cambridge
53. Plutchik R (1994) The psychology and biology of emotion. HarperCollins College Publishers, New York
54. Rae I, Takayama L, Mutlu B (2013) The influence of height in robot-mediated communication. In: Proceedings of the 8th international conference on human–robot interaction (HRI'13). IEEE, Tokyo, pp 1–8
55. Read R, Belpaeme T (2010) Interpreting non-linguistic utterances by robots : studying the influence of physical appearance. In: Proceedings of the 3rd international workshop on affective interaction in natural environments (AFFINE 2010) at ACM multimedia 2010. ACM, Firenze, pp 65–70
56. Read R, Belpaeme T (2012) How to use non-linguistic utterances to convey emotion in child–robot interaction. In: Proceedings of the 7th international conference on human–robot interaction (HRI'12). ACM/IEEE, Boston, pp 219–220
57. Read R, Belpaeme T (2014) Situational context directs how people affectively interpret robotic non-linguistic utterances. In: Proceedings of the 9th international conference on human–robot interaction (HRI'14). ACM/IEEE, Bielefeld
58. Reeves B, Nass C (1996) The media equation: how people treat computers, television, and new media like real people and places. CSLI Publications, Stanford
59. Repp B (1984) Categorical perception: issues, methods, findings. Speech Lang 10:243–335
60. Ros Espinoza R, Nalin M, Wood R, Baxter P, Looije R, Demiris Y, Belpaeme T (2011) Child-robot interaction in the wild: Advice to the aspiring experimenter. In: Proceedings of the 13th international conference on multimodal interfaces (ICMI'11). ACM, Valencia, pp 335–342
61. Saerbeck M, Bartneck C (2010) Perception of affect elicited by robot motion. In: Proceedings of the 5th international conference on human–robot interaction (HRI'10). ACM/IEEE, Osaka, pp 53–60
62. Scherer K (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40(1–2):227–256
63. Schouten B, Gerrits E, van Hessen A (2003) The end of categorical perception as we know it. Speech Commun 41(1):71–80
64. Schröder M, Burkhardt F, Krstulovic S (2010) Synthesis of emotional speech. In: Scherer KR, Bänziger T, Roesch E (eds) Blueprint for affective computing. Oxford University Press, Oxford, pp 222–231
65. Schwent M, Arras K (2014) R2–d2 reloaded: a flexible sound synthesis system for sonic human–robot interaction design. In: Proceedings of the 23rd international symposium on robot and human interaction communiation (RO-MAN 2014), Edinburgh
66. Siegel J, Siegel W (1977) Categorical perception of tonal intervals: musicians can't tell sharp from flat. Percept Psychophys 21(5):399–407
67. Siegel M, Breazeal C, Norton M (2009) Persuasive robotics: the influence of robot gender on human behavior. In: International conference on intelligent robots and systems (IROS 2009). IEEE, St. Louis, pp 2563–2568
68. Singh A, Young J (2012) Animal-inspired human–robot interaction: a robotic tail for communicating state. In: Proceedings of the 7th international conference on human–robot interaction (HRI'12), Boston, pp 237–238
69. Stedeman A, Sutherland D, Bartneck C (2011) Learning ROILA. CreateSpace, Charleston
70. Tay B, Jung Y, Park T (2014) When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. Comput Hum Behav 38:75–84
71. Terada K, Yamauchi A, Ito A (2012) Artificial emotion expression for a robot by dynamic coluor change. In: Proceedings of the 21st international symposium on robot and human interactive communication (RO-MAN 2012). IEEE, Paris, pp 314–321
72. Walters ML, Syrdal DS, Dautenhahn K, te Boekhorst R, Koay KL (2007) Avoiding the uncanny valley: robot appearance, personality and consistency of behaviour in an attention-seeking home scenario for a robot companion. Auton Robots 24(2):159–178
73. Yilmazyildiz S, Athanasopoulos G, Patsis G, Wang W, Oveneke MC, Latacz L, Verhelst W, Sahli H, Henderickx D, Vanderborght B, Soetens E, Lefeber D (2013) Voice modification for wizard-of-OZ experiments in robot–child interaction. In: Proceedings of the workshop on affective social speech signals, Grenoble
74. Yilmazyildiz S, Henderickx D, Vanderborght B, Verhelst W, Soetens E, Lefeber D (2011) EMOGIB: emotional gibberish speech database for affective human–robot interaction. In: Proceedings of the international conference on affective computing and intelligent interaction (ACII'11). Springer, Memphis, pp 163–172
75. Yilmazyildiz S, Henderickx D, Vanderborght B, Verhelst W, Soetens E, Lefeber D (2013) Multi-modal emotion expression for affective human–robot interaction. In: Proceedings of the workshop on affective social speech signals (WASSS 2013), Grenoble
76. Yilmazyildiz S, Latacz L, Mattheyses W, Verhelst W (2010) Expressive Gibberish speech synthesis for affective human–computer interaction. In: Proceedings of the 13th international conference on text., speech and dialogue (TSD'10). Springer, Brno, pp 584–590
77. Zhou K, Mo L, Kay P, Kwok VPY, Ip TNM, Tan LH (2010) Newly trained lexical categories produce lateralized categorical perception of color. Proc Natl Acad Sci USA 107(22):9974–9978

Robin Read received his Ph.D. in human–robot interaction from Plymouth University in 2014 where he worked on the FP7 ALIZ-E project under the supervision of Prof. Tony Belpaeme. He currently works at Dyson Technology Ltd., as a Robotics Research Engineer.

Tony Belpaeme received his Ph.D. in Computer Science from the Vrije Universiteit Brussel in 2002 and is currently Professor in Robotics and Cognitive Systems at Plymouth University (United Kingdom) where he leads a research lab in the Centre for Robotics and Neural Systems. Starting from the premise that cognition is rooted in social interaction, Belpaeme and team try to further the science and technology behind artificial intelligence and social robots. This results in a spectrum of findings, from theoretical insights to practical applications. He coordinated the FP7 ALIZ-E project, and collaborated on the ROBOT-ERA, DREAM and ITALK projects.