Abstract—Robots are often envisaged as embodied agents that might be able to intelligently and expressively communicate with humans. This could be due to their physical embodiment, their animated nature, or to other factors, such as cultural associations. In this study, we investigated emotional responses of humans to affective non-linguistic utterances produced by an embodied agent, with special attention to the way that these responses depended on the nature of the embodiment and the extent to which the robot actively moved in proximity to the human. To this end, we developed a new singing robot platform, ROVER, that could interact with humans in its surroundings. We used affective sound design methods to endow ROVER with the ability to communicate through song, via musical, non-linguistic utterances that could, as we demonstrate, evoke emotional responses in humans. We indeed found that the embodiment of the computational agent had an affect on emotional responses. However, contrary to our expectations, we found singing computers to be more emotionally arousing than singing robots. Whether the robot moved or not did not affect arousal. The results may have implications for the design of affective non-speech audio displays for human-computer or human-robot interaction.

Index Terms—Human Robot Interaction, Non-linguistic Utterances, Emotive Sound, Lab Study, Iterative Prototyping Design

I. INTRODUCTION

The notion that humans can respond emotionally to embodied robots in their surroundings has long been a staple of science fiction. Even the way that a robot moves affects the emotional responses of people observing it [1]. The presence of movement is often a valuable clue as to whether an object is animate or not. For example, the two main cues children use in distinguishing between animate and inanimate objects are movement and visual features (such as the presence of a face or dominant texture) [2]. Because movement can be indicative of animate embodiment, it can affect the associations that people make to an object in their environment, and the way that they might interact with it, as we review below. Consequently, we expected that people might respond emotionally to a mobile singing robot more readily than to a non-mobile singing robot.

From Asimov’s first and second laws [3], and other popular associations such as C3PO, one could conclude that robots should be easily usable, trainable, accessible and pleasant. Indeed, when technologies conform to social expectations, people find their interactions to be enjoyable, and empowering [4]. While it is not easy to support natural conversations and emotions through an embodied agent, such qualities can make robots more relatable and predictable [5]. In order to better afford interactions with broad arrays of untrained users, it is valuable for a robot to be able to communicate without relying on constrained language or command vocabularies. Non-linguistic utterances (NLUs), which consist of sounds that are not associated to words or other parts of language, have been widely explored in robotic systems, both in research, and in popular media, as we review below. NLUs can help to facilitate emotional communication while lowering expectations for speech, through warning sounds, calming, or questioning sounds, among many other possibilities, that might streamline communication between robots and people. Auditory communication is a natural vehicle for emotional content. People react ten times faster to acoustic than to visual cues [6], and one of the fastest triggers for emotion is sound [7]. Visual appearance plays a complex and multifaceted role in affective robotics, as has been demonstrated in prior literature [8], [9]. Despite prior research (which we review below), less is known about how the sounds produced by a robot affect the emotional responses it engenders, even though the value of multimodal interaction in ensuring accessible, usable, and satisfying interactions has been well established [10], [11], [12].

It is often tempting to design robotic agents that respond like humans, whether through sound (i.e., speech) or visual appearance, but this can be a perilous undertaking, due to the presence of the “uncanny valley”. The latter refers to the discomfort that is felt when something seems to be mostly, but not quite human [13]. The quality of synthesized speech is worse than synthesized facial expressions at expressing emotion [14], often exacerbating “uncanniness” effects. As such, it may be preferable, where appropriate, to convey emotion through non-linguistic auditory cues, paralanguage, or prosody.

In this study, we explored effects of embodiment and interaction on emotional responses to synthetic, non-linguistic utterances of a computational agent. Embodiment can be interpreted broadly, as perception, action and introspection are grounded in bodily sensation and states [15]. In our specific case embodiment refers to the level of presence that a subject experiences when interacting with a robot compared to interacting with a computer [16]. Although one could investigate the difference between a singing robot and a non-singing...
robot, we are focusing the effect of embodiment, the difference between a singing robot compared to a singing computer. Following a review of the literature on robot communication, sound and emotion (reproduced below), we compared emotional responses to NLUs elicited in three conditions: One in which participants approached a stationary robot that produced the NLUs, one in which the robot approached the participant, and a third in which a participant sat at a computer that produced the NLUs. We designed a singing robot, ROVER, as a platform for studying human robot interaction. In order to enable ROVER to produce sounds (NLUs) that could elicit specified emotional responses, we used perceptually-informed algorithms to synthesize musical phrases that varied in base frequency, frequency range, speech rate and pause rate, envelope, musical mode and intonation. Our results are dependent on how the NLUs are perceived. Participants reported high levels of valence when hearing the most excited NLUs. Although we expected a robotic embodiment to elicit a stronger emotional response, an analysis of the results indicated that the computer yielded significantly greater levels of arousal. The following sections review the relevant literature on robotics, sound, and emotion, introduce the ROVER platform, present the experiments, analysis, and results, and discuss implications for the design of NLU-based auditory displays for human-computer or human-robot interaction.

II. BACKGROUND

A. Robots, Embodiment, and Communication

Prior research indicates that people empathize and engage more with embodied robots than avatars [17], [18], [19], [20]. Physical embodiment can affect a social agent’s capabilities and a user’s enjoyment of a task [21]. In a study which compared a robotic and virtual medical assistant, participants felt a greater sense of presence, felt it was more lifelike, and disclosed less private information with the embodied robot versus an avatar [20]. Based on a survey on experiments with embodied robots, telepresent robots and avatars, people preferred an embodied robot. People felt a higher level of arousal, responded more favorably, had a stronger response, and found physically present robots more persuasive [22].

B. Robots and Auditory Communication

People prefer to communicate with robots via voice, and they prefer voice to be human-like [23], [24], [25]. In film, robots are typically voiced in one of two ways: by the voice of actors, whose speech is subsequently filtered (e.g. C3PO), or by non-speech computer generated sounds (e.g. R2D2) [26]. Despite prior research on the parametric synthesis of emotional speech, currently predominant approaches to speech synthesis use fragments of pre-recorded human speech.

1) Robots and Speech: Communicating emotion is important. A learning robot may receive more and better training data if it expresses emotion through the pre-recorded speech that consists of entire utterances of an actor [27]. However, using pre-generated speech can limit the range of possible interactions, motivating the use of synthesized speech. Cahn et al. implemented a system, using the DECTalk3 speech synthesizer, where 19 parameters controlled emotive speech. They found that the correct emotion was chosen 53% of the time out of six distinct emotions [28].

Nourbakhsht et al. created a robotic tour guide that expressed emotion through voice [29], but expressivity and communication were only briefly and informally discussed by the authors. Likewise, Roehling and Xingyan propose systems for producing emotional natural language speech by robot, but these systems were not systematically evaluated [30], [31]. Niculescu et al. found that modifying the pitch of speech could elicit differences in perception, and that a robot dressed as a woman with a higher pitched voice was rated as having a more attractive voice, as being more aesthetically appealing, outgoing, and as having better social skills [32]. This indicates that people can project human qualities onto robots in ways that depend on the audio qualities of speech audio.

2) Gibberish as Non-verbal Communication: A simpler approach to emotional communication by a robot is to use gibberish or non-linguistic utterances. Such an approach might better avoid the uncanny valley of robot speech. Cynthia Breazeal’s robot, Kismet, uses child-like utterances to reinforce emotions [33]. Oudeyer used child-like babble was used convey emotions that could be interpreted by people from different countries [34], by means of the MBROLA synthesiser. The latter could produce 30 sounds, expressing Happiness, Sadness, Anger, Comfort and Calmness, using different input strings. The participants had a high accuracy in categorizing the sounds, though they confused the sounds representing comfort and calmness.

Yilmazlyidiz et al. proposed a method for generating gibberish using a database of sounds and a prosody template, but this algorithm was systematically evaluated [35]. In 2010, the same group presented a different approach, based on presenting vowel sounds from sentences with different emotional content, generated via a text to speech synthesizer [36]. They found that it is important to match the input language with the language of the text to speech synthesizer, and that it was easier to determine whether the statement is positive or negative if it was presented in the correct semantic context. They combined both two methods in 2011 and found that users were able to accurately recognize seven emotions that were presented [37].

We are aware of no prior studies investigating the effect of embodiment and interactivity on emotional responses to gibberish, and few studies have utilized generative systems for producing non-linguistic utterances to convey emotion, as we do.

3) Non-Linguistic Utterances: Non-linguistic utterances (NLUs) are computer generated phrases that have audio qualities similar to speech but do not use vocal sounds [38]. Non-linguistic utterances are inexpensive computationally [38], and can be cross-culturally effective [34]. NLUs are an example of non-speech audio, similar to the related earcons, which are nonverbal sounds that are used to convey information. Earcons have been extensively investigated in the human factors literature, for applications ranging from conveying weather information [39] to warning drivers[40].
Read et al. found that children will assign emotional meaning to non-linguistic utterances produced by a Nao robot, though they could disagree on the emotion [38]. In contrast, adults more consistently categorized emotion in non-linguistic utterances [41]. Non-linguistic utterances can create the appearance of a stronger emotional reaction in robots when produced in response to an action [26].

Other researchers have investigated combinations of visual and auditory cues to emotion. Sparky, a small (50 cm) robot with an expressive face, a movable head on a long neck, and wheels was created by Mark Scheeff at Interval Research in order to investigate human-robot interaction [42]. The robot could make chirping sounds, but these were found to be confusing. Bartneck found that an avatar could more effectively elicit emotional responses using visual cues or audio and visual cues than with audio cues alone [43].

Robot toys like those from Keepon and WowWee use small sound databases to communicate emotion. Keepon uses them as a means of attracting attention and in response to sensory input [44]. WowWee robots also use sounds in response sensory input [26]. These preprogrammed sound sets are analogous to pre-recorded voice acted sound sets, and are not as flexible as a generative system, like we present here.

Jee et al. designed non-linguistic utterances for robots using musical structures, including tempo, key, pitch, melody, harmony, and rhythm to represent happiness, sadness, fear and dislike. Results showed that composed music was able to express emotion, and worked best when paired with visual cues [45]. Jee et al. later proposed an algorithmically generated musical system for robots, using tempo, pitch and volume to express joy, distress, shyness, irritation, expectation, dislike, pride and anger [46], but the effectiveness received limited evaluation. The same authors analyzed sounds from film robots R2D2 and Wall-E, and found that intonation, pitch and timbre were used to express emotions. Using this information, they designed five sounds and found that about half (55%) of people felt that the sounds reflected the intended emotion, and that most (80%) felt the sounds carried emotional expression [47]. While successful, these sounds could not be algorithmically produced to yield a specified emotional response, as our system is able to do. Other systems have been proposed but not tested [48]. Wallis et al. designed computer generated emotional music which was shown effective for expressing emotion, but has not been tested using robots [49].

Komatsu et al. found that sounds produced by a robot could affect whether a suggested action was performed [50], and that people preferred robots to communicate confidence using non-linguistic utterances (earcons) rather than through language or paralanguage [51]. They also asked participants to select attitudes that matched sounds produced by a robot or a personal computer. The results showed that the participants were more able to interpret computer sounds than sounds from the robots [52]. The design of their study was prone to bias because the NLUs were originally categorized by users listening at a computer [51]. This may explain why they found that NLUs were interpreted correctly more often on a computer in their latter study. We remove this bias by measuring how the participants report their emotional response, not how accurately they can categorize them. Unlike Komatsu et al., our sounds are more varied, more reminiscent of speech, and can work for any duration.

C. Emotional Spaces and Assessment

Prior research in robotics, like that in other fields, has examined emotion through several distinct models. Although discrete categorizations of emotion are often employed for their simplicity, many theories of emotion adopt continuous dimensional models, which have been found to better reflect the range of emotions that people report [53], [54], including intermediate or mixed emotions. They also allow contributing factors to be examined. Most models use pleasure and arousal as major axes in a two-dimensional emotion space [54]. In three-dimensional models, like that employed in our study, a dominance/submission axis is commonly used, creating what is known as the PAD (Pleasure, Arousal, Dominance) emotion space [55].

The Self Assessment Manikin (SAM) is one common tool that is used to gather emotional response using the Pleasure and Arousal Scale, or the full Pleasure-Arousal-Dominance spectrum. In SAM, users select images that represent ratings on a 5-point Likert Scale [80]. An alternative to SAM is the Positive and Negative Affect Schedule (PANAS), which involves rating 20 emotions on a 5-point Likert Scale [81]. (PANAS-X [82] expands the survey to 60 emotional items, and I-PANAS-SF [83] shortens the survey to 10 internationally recognized terms.) We chose to use SAM for the present study because of its simplicity (it only requires participants to answer three questions), because it is easily correlated with PAD, and because it is well validated, having been used extensively for the past 20 years.

D. Emotion and Prosody

The relation between emotion and prosody, the pattern of stress and intonation, has been extensively investigated in both music perception and speech research. Most studies of prosody and emotion through the audio signal have focused on four

TABLE I

<table>
<thead>
<tr>
<th>Audio Feature</th>
<th>Prior Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>[56] [57] [58] [59] [60]</td>
</tr>
<tr>
<td>F0 mean</td>
<td>[61] [62] [63] [50] [64]</td>
</tr>
<tr>
<td>F0 perturbation/ range</td>
<td>[64] [65]</td>
</tr>
<tr>
<td>F0 variability</td>
<td>[62] [63] [56]</td>
</tr>
<tr>
<td>F0 contour</td>
<td>[63] [56] [65]</td>
</tr>
<tr>
<td>High frequency-energy</td>
<td>[56] [63]</td>
</tr>
<tr>
<td>Pitch</td>
<td>[66] [67] [68] [69] [70] [71] [65] [72]</td>
</tr>
<tr>
<td>Pitch average</td>
<td>[73]</td>
</tr>
<tr>
<td>Pitch range</td>
<td>[73] [74] [65]</td>
</tr>
<tr>
<td>Pitch Variation</td>
<td>[68] [73] [75]</td>
</tr>
<tr>
<td>Pitch Maximum</td>
<td>[16]</td>
</tr>
<tr>
<td>Major/ Minor Mode</td>
<td>[77] [76] [78] [79] [74]</td>
</tr>
</tbody>
</table>
attributes: Fundamental frequency (F0) or Pitch, Amplitude or Intensity, Speech Rate or Tempo, and Articulation or Timbre (Table I - IV) Emotional valence has been correlated with pitch, intensity and rate in speech, with increased valence correlated with a higher pitch, higher deviation, larger range, higher mean intensity, larger intensity deviation, faster speech rate, shorter syllable duration and shorter/less frequent pausing. Decreased valence is correlated with the opposite effects of increased valence [64].

There has been conflicting results in research on how tonality and timbre affects emotion. In researching an instrument’s ability to express sadness, Huron found that instruments that were more percussive were perceived to be unable to express sadness by musicians [70]. In a pilot study, Le Groux found that valence was not correlated with any emotive sound, but the study only used percussive sounds [86]. Musical major/minor mode has been found to be capable of eliciting emotions with different valences [77]. Jee et al. used musical key in composing hand crafted emotion-carrying non-linguistic utterances for robots [45], but an algorithm proposed by the same authors did not use this strategy [46]. Our generative algorithm uses percussive and steady state envelopes, as well as major or minor mode as parameters.

III. ROVER: A SINGING ROBOT

To investigate how the emotional responses to a robot are affected by embodiment, movement, and sound, for this study, we designed and fabricated a singing mobile robot, ROVER: the Reactive Observant Vacuous Emotive Robot (Figure 1). ROVER includes a base unit (iRobot CREATE), embedded computer (Raspberry Pi) with digital camera, a far infrared sensor array (Melexis 90620), a microcontroller (Arduino Mega), proximity sensors and speakers. ROVER is 6 ft tall, with a thin white frame that was laser cut from acrylic.

Motivated by the aforementioned results in music and speech perception, ROVER communicates via non-linguistic utterances in order to convey emotion. To this end, it modulates audio qualities, including timbre, fundamental frequency, contour, mode, and tempo.

ROVER is also able to navigate its surroundings, mapping obstacles and finding people, while navigating using a combination of bump, proximity and far infrared sensors. ROVER has two basic interaction modes, both of which are used in our study: an active mode where it moves about the space looking for people to interact with, and a passive mode where it is stationary and will only interact when approached. ROVER’s Raspberry Pi’s software had 3 main parts, playing the NLUs, analyzing the video and sending/receiving serial data (Figure 3).

IV. ALGORITHM FOR NLU GENERATION

ROVER produces sounds in the form of non-linguistic utterances in order to communicate with users. NLUs are designed as musical phrases, which are produced algorithmically in order to elicit desired emotional responses. In order to lend ROVER the largest possible repertoire, and to ensure engaging interactions, we designed an algorithm that could produce a very large variety of NLUs that are designed to elicit specified emotions, greatly increasing the vocabulary of the robot over other approaches, such as pre-recorded speech.

Melodies are produced by a generative algorithm, which varies the fundamental frequency (F0), amplitude, timbre (temporal envelope), and motive according to the specified emotion, based on the sound perception and emotion results cited above. Each of these aspects is controlled by two to five descriptors that specify a given musical phrase (Table VI).

Our study focused on two families of emotion in non-linguistic utterances, corresponding to “excited” or “sad” sounds. The emotion “sad” was defined as low valence, low arousal and submissive in the PAD emotion space. In contrast, “excited” was defined as high valence, high arousal and dominant. The parameters it used were base fundamental frequency, frequency range, speech rate, pause ratio, envelope, pitch contour and mode. For illustration, a sad emotional state is often expressed in speech by low base frequency, small frequency range, low speech rate and high pause rate, while an excited emotional state is often expressed in speech by high base frequency, large frequency range, high speech rate and low pause rate. The emotional arousal was manipulated through the temporal envelope (“percussive envelope” and “steady state envelope”) of the sound. Two musical modes were used in order to affect emotional valence: major and
ROVER Schematic (left and center), and ROVER interacting with a user (right). ROVER is built from a base unit (iRobot CREATE), embedded computer (Raspberry Pi) with digital camera, a far infrared sensor array (Melexis 90620), a microcontroller (Arduino Mega), proximity sensors and speakers. ROVER is 6 ft tall, with a thin white frame that was laser cut from acrylic.

**TABLE V**

| Parameters for Two Emotions: This Study Focused on Two Emotions “Sad” and “Excited”. The Values to Control Those Non-Linguistic Utterances Were Based on Prior Research. |

<table>
<thead>
<tr>
<th>Emotion</th>
<th>“Sad”</th>
<th>“Excited”</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base F0 (tonic note)</td>
<td>G (1567.98)</td>
<td>G (1567.98)</td>
<td>[26]</td>
</tr>
<tr>
<td>F0 range (variance)</td>
<td>3 notes</td>
<td>5 notes</td>
<td>[26]</td>
</tr>
<tr>
<td>Pause Length</td>
<td>Long (avg. .4s, std. .2s)</td>
<td>Short (avg. .2s, std. .1s)</td>
<td>[64]</td>
</tr>
<tr>
<td>Pause Ratio</td>
<td>High (avg. 3:1, std. 2)</td>
<td>Low (avg. 6:1 std. 2)</td>
<td>[64]</td>
</tr>
<tr>
<td>Envelope</td>
<td>Steady State (.2s attack, .6s sustain, .2s decay)</td>
<td>Percussive (.05s attack, .2s decay)</td>
<td>[70]</td>
</tr>
<tr>
<td>Pitch Contour</td>
<td>Rising (contour end difference = 5)</td>
<td>Flat (contour end difference = 0)</td>
<td>[87]</td>
</tr>
<tr>
<td>Mode</td>
<td>Minor</td>
<td>Major</td>
<td>[77]</td>
</tr>
</tbody>
</table>

A. NLU Melody Generation

NLUs comprised short randomly generated melodies (described here) composed of synthesized notes (discussed below). The frequency of each note is dependent on five values:

- **F0 Frequency**
  - tonic note of key
  - major or minor key
  - the contour of the phrase (start)
  - the contour of the phrase (end)
  - variation around the contour

- **Amplitude**
  - attack and sustain amplitude difference
  - steady state amplitude

- **Timbre**
  - length of the attack
  - decay
  - sustain mean
  - sustain variance
  - release

- **Motive**
  - pausing length mean
  - pausing length variance
  - motive length center
  - motive length range
Fig. 2. Technology design for ROVER: The study required ROVER, an iPad, a computer and a server. The server ran a Django website with a MySQL database to collect data and manage the state of the system. When interacting with the computer, it was used by the participant to listen to the NLUs and enter data while being recorded by a webcam. The iPad and ROVER were used when the participant interacted with ROVER. The iPad was used for rating each NLU that ROVER played. ROVER uses both a Raspberry Pi (single board computer) and an Arduino Mega (micro-controller) to control the system. The Arduino Mega’s software navigates the iRobot CREATE, using data hierarchically from the iRobot CREATE’s bump sensors, proximity sensors, and Melexis thermophile. The Arduino also sends the sensor readings over serial to the Raspberry Pi, and listens for a signal that the Raspberry Pi has detected a face.

Fig. 3. ROVER software design: The software controlled the movement of ROVER, the video recording, and audio playback.

Fig. 4. The waveform for an “excited” NLU (A) and a “sad” NLU (B).

Fig. 5. Melody of phrase algorithm: The melody is dependent on the tonic note of the key, whether the NLU is in a major or minor key, the contour of the phrase (start and end values) and variation around the contour. In this chart the randomly generated melody, comprised of grey dots representing notes, is shifted to fit the contour. The creation of the melody is described in the Section IV-A.

tonic note of the key, whether the NLU is in a major or minor key, the contour of the phrase (start and end values) and variation around the contour (Figure 5). The tonic note, and whether it is a major or minor key, determines the scale. The contour, the rise and fall of the melodic line, is defined by two contour values. It is created by concatenating two vectors, each which have equally spaced values from first contour value to zero, and from zero to the last contour value. This allows the pitch to ascend, descend and plateau at the beginning or end of the melody. We added the contour to a random vector of equal length with a mean of the tonic note and the variance defined in Table V. We varied the degree of excursion (or variation) above and below the contour in order to generate notes for the melody that comprises the NLU. The major/minor mode of a melody was not forced, and particularly with shorter melodies may be imperceptible.

B. Note Timbre and Amplitude

The timbre of the notes in the melody that comprise the robot’s non-linguistic utterances was controlled through an attack decay sustain release (ADSR) envelope that was applied to the sound (Figure 6). Different timbres were produced by manipulating the attack, decay, sustain mean, sustain variance and release of the envelope. A sustain length was randomly generated using these values, and an ADSR envelope was generated for each note. Two types of ADSR envelopes were applied to produce different timbres: a (short) percussive envelope and a (longer) steady state envelope. The percussive envelope had a sustain mean and variance of zero. The steady state envelope difference in attack and sustain
Fig. 6. ADSR Envelope: The timbre is created using an attack decay sustain release (ADSR) envelope. In this study we specifically used a steady state and a percussive envelope.

Fig. 7. Motive length and range algorithm: The motive is described by the motive length center and range, as well as the pausing length mean and variance. An example of how a series of notes can be divided into motives is denoted by the purple lines, with the average motive length 3 notes long and the motive range 1 note.

amplitude was zero and had a decay time of zero. While the algorithm to generate emotive NLUs allowed for more complexity in defining the amplitude, we decided for this study to keep amplitude consistent between the two emotive utterances because the amplitude of the utterance differed between the computer and the robot playing it.

C. Motive

The phrasing, or prosody, of the note sequences was algorithmically generated by controlling parameters that consisted of motive length, center, and range, and pause length, mean, and variance (Figure 7). The motive length describes after how many notes there is a pause. The motive length is calculated by choosing a random value within the motive range, until the sum of the motive lengths was at least as large as the length of the phrase. The length of each pause is sampled from a normal distribution parametrized by pause length mean and variance. The length was specified with millisecond resolution.

V. STUDY: EMOTIONS ELICITED BY A SINGING EMBODIED AGENT

In two large exhibitions for public audiences (Fig. 1), we observed dozens of individuals and groups of people interact with ROVER over the course of several days, and discussed their responses with them informally. We noted that the NLUs produced by the robot could elicit a range of responses, and that people enjoyed the interactions. We also noticed differences in whether people approached the robot or waited for it to approach before listening, and wondered if this difference would affect how they interpreted the utterances produced by the robot.

Inspired by this, we designed an experimental campaign to investigate whether a robot could elicit different emotional responses using NLUs, and to assess how the elicited emotions depended on the embodiment and actions of the robot. The study assessed emotional responses of participants to NLUs designed using the methods described above, and produced from three different embodied agents: a computer, an immobile robot and a mobile robot.

A. Participants

A total of 20 subjects completed the study, 12 men and 8 women. 95% of participants were between the ages of 18-24, recruited using the Psychology Study Participant Pool of the authors’ institution. Participants were compensated with a five dollar gift certificate, and were all university students at the authors’ institution. Participants were naïve with respect to the content of this study. All participants gave their informed consent. The experiments were approved by the Institutional Ethics Review Board of the authors’ institution.

B. NLUs

Auditory stimuli consisted of the non-linguistic utterances (sound signals) described above, and were generated according to one of the tested emotions (Table 5). In order to avoid accidental effects that could arise if particular NLUs elicited unusual responses (for example, due to familiarity effects or other associations), new sets of NLU stimuli were randomly generated for each trial and for each participant in the study. All presented NLUs were recorded during the experiment. The NLU sound files were generated in advance for playback by the system (Figure 2).

C. Procedure

During the experiment that in total lasted less than 30 minutes, participants listened to NLUs produced by a robot or computer and rated their feelings using the SAM scale. At the beginning of the experiment, there was a 5 minute instructional period, participants were introduced to the apparatus, the procedure, and the response method. The Self Assessment Manikin scale was explained with 3 descriptors for each extreme of the scale (Table VII). After any questions, the participant filled out a short demographic survey.

Before each interaction block the investigator explained to the participant how to interact with the sound source. Each
TABLE VII
EMOTIONAL RESPONSE SURVEY DESCRIPTORS AND MAPPING

<table>
<thead>
<tr>
<th>Scale</th>
<th>-2</th>
<th>0</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleasure</td>
<td>Unhappy</td>
<td>Neutral</td>
<td>Happy</td>
</tr>
<tr>
<td></td>
<td>Unsatisfied</td>
<td></td>
<td>Satisfied</td>
</tr>
<tr>
<td></td>
<td>Annoyed</td>
<td></td>
<td>Pleased</td>
</tr>
<tr>
<td>Arousal</td>
<td>Relaxed</td>
<td>Neutral</td>
<td>Stimulated</td>
</tr>
<tr>
<td></td>
<td>Sleepy</td>
<td></td>
<td>Awake</td>
</tr>
<tr>
<td></td>
<td>Calm</td>
<td></td>
<td>Excited</td>
</tr>
<tr>
<td>Dominance</td>
<td>Controlled</td>
<td>Neutral</td>
<td>Controlling</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td></td>
<td>Big</td>
</tr>
<tr>
<td></td>
<td>Influenced</td>
<td></td>
<td>Influential</td>
</tr>
</tbody>
</table>

interaction block, including instruction, lasted up to 7 minutes. The participant would start the interaction whenever they were ready. After the participant had completed the three blocks the participant was debriefed, received compensation and the investigator answered questions the participant had about the study. The debriefing took up to 5 minutes.

Participants listened to the NLUs and provided responses in three different interaction conditions, which were presented sequentially in randomized order. Participants heard 8 NLUs for each of the three interaction types, for a total of 24 unique NLUs. They heard 4 “excited” and 4 “sad” NLUs in a random order for each interaction. Each NLU that they heard was a randomly generated unrepeated melody with the average length of 6.39 seconds (std. 0.65s, min. 4.95s, max. 9.22s). The “excited” songs were 17-19 notes long and the “sad” songs were 3-5 notes long. The rest intervals between each NLU was the time it took for the participant to fill out the survey plus 5 seconds.

In different interaction conditions, participants listened to 8 different non-linguistic utterances produced by a computer, by a stationary robot (ROVER), or by a mobile robot (also ROVER). During each trial in the stationary robot condition, participants began a short distance away (13 feet) and walked toward the robot, and stood in front of it (2 feet), at which time the robot played the NLU. During each trial in the computer condition, the robot was replaced by a laptop computer, which played the sound as participants sat in front of it. During each trial in the mobile robot condition, participants stood a short distance away (13 feet) and were approached by the robot. The robot used a digital camera and proximity sensors in order to determine when it had reached the same distance from the participant as in the stationary conditions, then stopped and played the NLU. The process was repeated, in each condition, for all 24 utterances, of which 12 corresponded to each emotion. After each trial, participants completed a Self Assessment Manikin (SAM) survey on the computer or on an iPad. Each question was explained to the participants with 6 descriptors for each scale (Table VII) [88]. A custom internet application (using Django and MySQL running on an Ubuntu server) collected the participants’ responses. Participant interactions also filmed on video during the experiment, and all other data were anonymized for the analysis.

D. Analysis

Each question from the SAM survey responses was analyzed using two-way repeated measures ANOVA using SPSS Statistics. The resulting p-values (reported below and in accompanying tables) were Bonferroni corrected (BC) by multiplying them by the relevant number of tested hypotheses. The scale for each question was a -2 to 2 scale, with 0 as neutral (Table VII).

Fig. 8. Emotional (valence, arousal, dominance) response on a scale of -2 to 2 to “sad” and “excited” NLUs: The difference in valence between the two NLU types was very statistically significant F(1, 13) = 32.126, p=0.0002, np2=.71. On average the participant rated the “excited” NLUs as causing them a higher level of pleasure (mean 0.923, standard deviation 0.154), while the “sad” NLUs were rated causing neutral or negative level of pleasure (mean -0.208, standard deviation 0.173).

Fig. 9. Emotional (valence, arousal, dominance) response on a scale of -2 to 2 in different interaction conditions: The difference in arousal between interaction conditions was statistically significant with F(2, 26) = 8.829, p=0.003, np2=.41. We combined the sad and excited tones because the NLU type and Interaction Condition relationship was non-significant, see Table VIII, IX, and X. There was no significant difference between ROVER stationary and moving (p=3.0), or the computer and ROVER moving (p=0.09) but there was a significant difference between interacting with the computer and ROVER stationary (p=0.009). The computer was rated as causing a higher rate of arousal (mean -0.304, standard deviation 0.186) compared to ROVER stationary (mean -0.804, standard deviation 0.154) and ROVER moving (mean -0.732,SD 0.233).
VI. RESULTS

The analyzed factors were the NLU type (i.e., sad or excited NLUs) and interaction condition (i.e., stationary robot, mobile robot, or PC). The difference in emotional valence between the two NLU types across all interaction conditions was very statistically significant $F(1, 13) = 32.126, p = 0.0002, np^2 = .71$ (Figure 8 and Table VIII). On average the participant rated the “excited” NLUs as causing them a higher level of pleasure (mean 0.923, standard deviation 0.154), while the “sad” NLUs were rated as causing neutral or negative level of pleasure (mean -0.208, standard deviation 0.172).

The difference in arousal between interaction conditions was statistically significant with $F(2, 26) = 8.829, p = 0.003, np^2 = .41$. (Figure 9 and Table IX) There was a significant difference between the PC condition and the stationary robot condition ($p = 0.009$). There was no significant difference between the mobile and stationary robot conditions ($p = 0.3$), or the PC and mobile robot conditions ($p = 0.09$). The PC was rated as causing a higher rate of arousal (mean 0.304, standard deviation 0.186) compared to the stationary robot condition (mean 0.208, standard deviation 0.172) and moving robot condition (mean -0.732, standard deviation 0.233), indicating that participants felt more aroused when interacting with the computer than with the stationary robot. There was not a significant difference between interacting with a stationary or moving robot, or with a moving robot or a computer.

The difference in dominance was not significant for either interaction type or non-linguistic utterances. Although it is a standard measure, participants seemed to have the greatest difficulty in interpreting the concept of submission/dominance, which may have contributed to these results. The relationship between interaction conditions and NLUs was not statistically significant (Figure 10).

VII. DISCUSSION

The results showed that the NLUs could elicit different feelings, and that the feelings differed in valence in ways that matched our expectations. The “excited” NLUs were rated to cause a higher level of valence than the “sad” NLUs, consistent with prior work in music psychology. (We note that the literature indicates that reported perceived and felt emotions in response to music agree, and that the level of felt emotion was higher than perceived emotion in connection with positive valence [89].)

There were differences in arousal between interaction conditions could be explained by differences in participants interacted with the computer and robot. The difference in the response to the robot and laptop might be attributable to how participants interpreted this interaction as affecting their personal space, as the robot may have been interpreted as affecting what Hall described as their personal reaction bubbles: People can feel anxiety, anger or discomfort when their
personal space is encroached upon [90]. In our experiment, this could cause greater anxiety or heightened arousal if the robot were interpreted as entering their space to a greater degree.

Among the differences between interaction conditions, several are inherent to the differences between robots, modes of interacting with robots, and interacting with computers. In the PC condition, in order to mirror typical interactions with computers, participants sat in front of the computer. This could explain the difference in arousal that we observed, from which it might be concluded that users may be more susceptible to being aroused by non-linguistic utterances when seated in front of a computer, than when standing or moving in proximity to an embodied robot.

VIII. CONCLUSION

This study investigated emotional responses of humans to affective non-linguistic utterances that were designed to convey emotion, when these utterances were produced by an embodied agent. Our study paid special attention to the way that these responses depended on the nature of the embodiment, including whether a moving or stationary robot was involved, or the sounds were produced by a personal computer. We developed algorithms for generating affective non-linguistic utterances to endow these agents with the ability to communicate via musical phrases that could, as we showed, evoke different feelings in humans.

As robots are becoming more common in home environments, and are increasingly employed in the service sector, the value of ensuring usability through effective communication is paramount. With the help of ROVER, a singing robot introduced here, and a novel algorithm for generating a variety of emotion-carrying non-linguistic utterances, we investigated whether feelings could be induced by these utterances, how the feelings they elicited sounds depend on what a robot is doing, or how it is embodied. The NLU generation algorithm is advantageous, because it is capable of producing a very large variety of sounds that correspond to any emotional template, greatly increasing the repertoire of the robot over other approaches, such as pre-recorded speech, and can do so in ways that preserve the emotional content - every "excited" utterance in our study was different, but these utterances were nonetheless rated by participants as causing a higher level of valence than the "sad" utterances.

There are several advantages to our NLU design over prior methods. The NLU design method we introduce is more strongly motivated by the sound and emotion literature, based on variation around the pitch contour. Our NLUs are highly flexible with the length of the utterance independent of emotional content. Our approach also takes advantage of more emotion-related parameters than previously proposed methods, providing more variety and control over the synthesized NLUs. Due to the flexibility of the input parameters, there is room for future research on extending the system to express other emotions outside of those tested in this study. Compared to pre-recorded or composed NLUs, our NLUs can be dynamically generated to any length and each one can be unique which supports repeated interaction. Designers should use this type of algorithm when they need sounds of any length and are expecting repeat interactions where a user might find the same utterance monotonous.

A surprising outcome of this study was that participants felt more aroused when interacting with the personal computer than with a stationary robot, despite the charged associations that are sometimes linked to robotic embodiments. The listening interactions with the computer mirrored those used in many prior studies on emotion and music, and by sound designers, who often work at a computer. As such, designers may wish to consider that the degree of arousal that is elicited by an embodied robot may be lower than expected. This could be due to postural differences or to the familiarity of interacting with a computer.

This investigation, the foundational subjects reviewed in the foregoing sections, and the system and methods introduced here, could help to advance the field of human robot interaction, by contributing models and guidelines for the design of auditory communication capacities for robots that could enliven and improve a wide range of future activities in which embodied agents will play a role.

IX. ACKNOWLEDGMENTS

The authors thank the UCSB Systems Lab and the Mosher Foundation for financial support, and JoAnn Kuchera-Morin, Michael Mangus, Tobias Hollerer, Marcos Novak, and Pamela Wolfe for valuable input. Photo credit of Figure 1 to Mohit Hingorani. This work was also supported by NSF awards No. 1446752 and 1527709.

REFERENCES

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2774815, IEEE Transactions on Affective Computing.
C. Breitenstein, D. V. Lancker, and I. Daum, “The contribution of speech..."  
J. R. Murray and J. L. Arnott, “Toward the simulation of emotion in..."  
S. Dalla Bella, I. Peretz, L. Rousseau, and N. Gosselin, “A development..."  

**Hannah Wolfe** is currently PhD student in Media Arts and Technology and a Masters student in Computer Science at University of California Santa Barbara. She received her B.A. in Art from Bennington College in 2009 and her M.S. in Media Arts and Technology from University of California Santa Barbara in 2016. Her research interests include human robot interaction, affective computing, virtual reality, and computational creativity.

**Marko Peljhan** is a Professor of interdisciplinary studies, University of California, Santa Barbara, where he serves as the director of the Systemics lab, devoted to systems research in immersive environments, aerospace paradigms, sensor networks, robotics and communications at the intersection of arts, engineering and science. He is the recipient of several prizes for his work, including the Golden Nica for Interactive art at Ars Electronica and the UNESCO prize for Art and Technology. In 1992 he founded the arts and technology organization Projekt Nica for Interactive art at Ars Electronica and the ECE Department at Drexel University. Post-Doctoral Fellow (2011-2012) at the Inst. of Intelligent Systems and Robotics, Universite Pierre et Marie Curie. He has been employed in industrial R&D for sonar, speech recognition, and music DSP at several high-technology companies. His research interests include haptic science and engineering and soft robotics.

**Yon Visell** received the PhD Degree in Electrical and Computer Engineering from McGill University (2011), and MA and BA degrees in Physics (Univ. Texas-Austin, Wesleyan Univ.). Since 2015, he is Assistant Professor of Media Arts and Technology, Electrical and Computer Engineering, and Mechanical Engineering (by courtesy) at the University of California, Santa Barbara, where he directs the RE Touch Lab. Assistant Professor (2012-2015) in the ECE Department at Drexel University. Post-Doctoral Fellow (2011-2012) at the Inst. of Intelligent Systems and Robotics, Universite Pierre et Marie Curie. He has been employed in industrial R&D for sonar, speech recognition, and music DSP at several high-technology companies. His research interests include haptic science and engineering and soft robotics.