

Perception of Mechanical Sounds Inherent to Expressive Gestures of a NAO Robot - Implications for Movement Sonification of Humanoids

Emma Frid

Department of Electrical Engineering
and Computer Science (EECS)
emmafrid@kth.se

Roberto Bresin

Department of Electrical Engineering
and Computer Science (EECS)
roberto@kth.se

Simon Alexanderson

Department of Electrical Engineering
and Computer Science (EECS)
simonal@kth.se

ABSTRACT

In this paper we present a pilot study carried out within the project *SONAO*. The *SONAO* project aims to compensate for limitations in robot communicative channels with an increased clarity of Non-Verbal Communication (NVC) through expressive gestures and non-verbal sounds. More specifically, the purpose of the project is to use movement sonification of expressive robot gestures to improve Human-Robot Interaction (HRI). The pilot study described in this paper focuses on mechanical robot sounds, i.e. sounds that have not been specifically designed for HRI but are inherent to robot movement. Results indicated a low correspondence between perceptual ratings of mechanical robot sounds and emotions communicated through gestures. In general, the mechanical sounds themselves appeared not to carry much emotional information compared to video stimuli of expressive gestures. However, some mechanical sounds did communicate certain emotions, e.g. frustration. In general, the sounds appeared to communicate arousal more effectively than valence. We discuss potential issues and possibilities for the sonification of expressive robot gestures and the role of mechanical sounds in such a context. Emphasis is put on the need to mask or alter sounds inherent to robot movement, using for example blended sonification.

1. INTRODUCTION

The work presented in this paper was carried out in the context of the *SONAO* (Robust non-verbal expression in artificial agents: Identification and modeling of stylized gesture and sound cues) research project. The *SONAO* project aims to improve the comprehensibility of robot non-verbal communication (NVC) using data-driven methods and physical acting styles. In other words, the purpose is to compensate for limitations in robot communicative channels with an increased clarity of NVC through expressive gestures and non-verbal sounds.

In the current paper we briefly introduce the *SONAO* project by outlining the background and problem domain of transferring expressive human gestures to a humanoid

robot. We present a pilot study focusing on sounds inherent to movement of the humanoid robot NAO¹ (see Figure 1) and discuss implications for the sonification of these movements.

2. BACKGROUND

Robots are not quiet; they often rely on for example servo motors in order to move, and thus their movements produce sounds. Sounds generated by robot movement could potentially have an effect on Human-Robot Interaction (HRI), as certain sounds may alter interpretation of the message that the robot is trying to convey. For example, mechanical sounds could influence interpretation of the robot's emotional reactions. Despite the fact that sounds inherent to robot movement could implicitly convey meaning and affect social interaction, sound design is often an overlooked aspect in the field of HRI. As for all consumer products, sound plays a role in our aesthetic, quality, and emotional experience [1]. Of course, this also applies for robots that we interact with. There are several examples of projects in which the sound design has been neglected, resulting in significant effects for the HRI. For example, motor sounds of the pet robot *Pari* negatively interfered with interactions [2] and the Boston Dynamics LS3 pack-mule robot was found to be "too loud" to be integrated in military patrols [3].

In [1], Langeveld et al. make a distinction between sounds that are generated by the operating of the product itself (*consequential sounds*), and sounds that are intentionally added to a product (*intentional sounds*). In the context of HRI, we have to consider both the sounds that have been specifically designed for communication of a robot's functions and emotional reactions (intentional sounds) and the sounds that are produced by the robot's movements (consequential sounds). Unfortunately, little work in the field of HRI has focused on consequential sounds. The fact that robot's active motion makes motor noise has been discussed mainly in research focusing on robots with audition, since motor noises makes auditory processing more difficult (see e.g. [4]). Few studies have focused specifically on sounds inherent to robot movement. In [5], authors investigated perception of sounds generated by a robotic arm, concluding that the presence and quality of the sound shaped subjective perception. In another study by Moore

Copyright: © 2018 Emma Frid et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

¹ <https://www.ald.softbankrobotics.com/en/robots/nao>



Figure 1: Two humanoid NAO robots.

et al. [6], aural impressions associated with servo motors commonly used in robotic motion were investigated. Participants made subjective ratings of motor sounds. Results suggested both anthropomorphic associations with sounds and negative impressions of the sounds overall.

Interestingly, numerous previous studies have focused on affective communication through expressive movements of the humanoid robot NAO (e.g. [7–9]). Apart from defining expressive and affective gestures, research in social HRI has also focused on how to achieve effective communication through intentional sounds. An extensive review of semantic-free utterances in social HRI was carried out by Yilmazyildiz et al. [10]. Several studies have focused on affective sounds for humanoid robots [11, 12], some of which have specifically dealt with affective sounds for the NAO robot [13, 14]. However, little work has been done on sonification in the context of social HRI (see e.g. [15]). In particular, little research has focused on augmenting expressive robotic movement with sound (see e.g. [16, 17]). The SONAO project aims to fill this gap by incorporating movement sonification in the robot’s non-verbal communication. The following section describes the project in detail.

3. THE SONAO PROJECT

The aim of the SONAO project is to establish new methods for achieving robust interaction between users and humanoid robots and virtual agents, based on sonification of expressive gestures. This is done by combining competences of research team members in the fields of social robotics, sound and music computing, affective computing and body motion analysis. Focus is put on *Non-Verbal Communication (NVC)*, i.e. communication that does not involve semantics in natural spoken language but can still facilitate rich communication and expression. The aim is to move from a discrete identification scheme, based merely on detecting a particular emotion (i.e. *the robot is sad/happy*), to a continuous scheme where a set of emotions could be viewed on a continuum. The main objectives of the SONAO project are to:

- Develop representations of a humanoid robot’s or virtual agent’s internal states based on gestures and sounds.

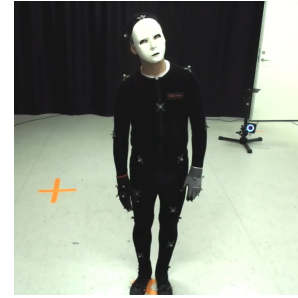


Figure 2: Motion capture setup with mime actor.

- Develop methods for clear multisensory communication of intentions, including emotions, with regard to limitations in facial, verbal and bodily expression of a humanoid robot or virtual agent.
- Contribute with new knowledge and methods for supporting feedback to, and understanding of, body motion qualities by means of sound.

This will be achieved through:

1. The creation of a database of expressive (stylized) gestures.
2. Mapping of (human) gestures to humanoid robots and virtual agents.
3. Sonification of gestures of humanoid robots and virtual agents.

An important aspect of the SONAO project is to use expressive gestures and sonifications to compensate for the reduced degrees of freedom² in which a robot can express her/himself, resulting in increased clarity of robot NVC. The work presented in the current paper is based on previous research conducted by Alexanderson et al. [18], focusing on mapping human motion to virtual agents. The study by Alexanderson et al. connects mainly to point 1 and 2 of the list presented above. It focused on full body motion capture of a mime actor (see Figure 2) interacting in short dialogues, varying the intensity level (from low, i.e. level 1, to high, i.e. level 5) of the non-verbal expression along five different dimensions (e.g. frustration, attention, joy). Findings suggested that the mapping of body movements from human movements to a virtual agent tended to preserve the recognition of emotional renderings originally performed by the mime actor. Other work within the SONAO project includes developing re-targeting techniques for the NAO robot grounded on virtual character animation research [19]. In [19], focus was put on motion sequences that addressed constraints such as joint angle and angular velocities. The work connects mainly to item 2 of the above presented list. Moreover, a vocal sketching experiment (see [20] for definition of vocal sketching methodology) with the mime actor participating in [18] was carried out. In this experiment, the actor was shown

² Robots have fewer degrees of freedom than humans, as they do not have the same number of movable joints, etc.



Figure 3: Screenshot of video stimulus of the NAO robot performing a joyful gesture.

mutated videos of his own movements, and asked to vocalize sounds that expressed particular emotions. The work was done as a prestudy intended to inform sonification design in the *SONAO* project, and connects to item 3 of the above presented list.

4. PILOT STUDY

The purpose of the pilot study described in this paper was to perceptually evaluate a set of sounds produced by mechanical movement of a NAO robot performing expressive gestures. The term *mechanical sounds* in this context refers not only to sounds produced by motors but also to friction sounds and contact sounds produced by robot movement. The main aim of the pilot study was to investigate if these sounds could communicate affective states or induce emotions. Depending on how these sounds are perceived, future implementations, as well as sonifications of expressive robot movements, could focus on either enhancing or masking mechanical robot sounds, using *blended sonification* techniques [21]. Blended sonification describes “the process of manipulating physical interaction sounds or environmental sounds in such a way that the resulting sound signal carries additional information of interest while the formed auditory gestalt is still perceived as coherent auditory event” [21].

4.1 Method

We carried out an online perceptual rating experiment³ in which participants were asked to rate emotions in the interactions presented in different stimuli on a set of five-step scales (sad, joyful, frustrated, relaxed), ranging from not at all (0) to very much (4), with an annotated step size of 1. The survey was distributed to students and colleagues at KTH Royal Institute of Technology, Stockholm, and Uppsala University. In addition, the survey was shared on music and audio-related mailing lists in the research community. Participants were also allowed to re-distribute the survey to anyone they believed would be interested in participating in the study.

³The entire survey can be accessed at <http://www.surveymzmo.com/s3/3852959/DM2350-Robot-Motor-Sounds>.

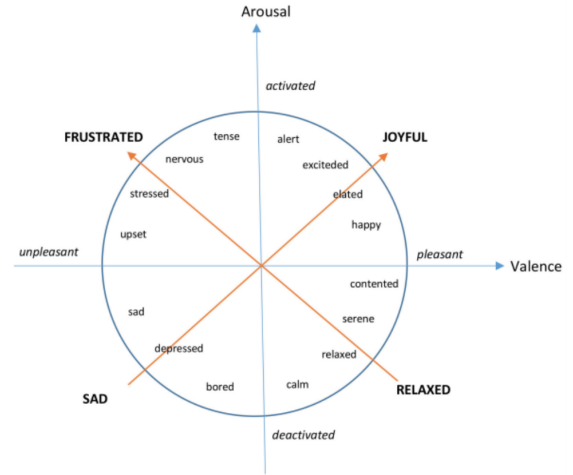


Figure 4: The Circumplex Model of Affect.

The survey was based on three types of stimuli: audio, audio-video and video, with 4 stimuli per respective stimuli category, giving a total of 12 stimuli⁴. Presentation order was randomized for each participant. The stimuli were expressive gestures performed by a NAO robot and/or sounds produced thereof. The gestures originated from human movements performed by the mime actor in [18]. These movements were translated into a robot gestural representation, as described in [19]. The sounds produced by the robot’s movements were recorded using a Brüel Kjær BK 4003 microphone connected to a Brüel Kjær 2812 pre-amplifier, which in turn was connected to a RME Baby-Face soundcard. Sound files were normalized and synchronized to video recordings. A screen shot of a video stimulus is seen in Figure 3.

The gestures used to generate stimuli expressed the following affective states: 1) frustration, 2) relaxation and 3), 4) joy (two different gestures). These particular affective states were chosen as they are opposites on the cross-diagonal of the circumplex model of affect [22], see Figure 4. The same categories have previously been used by robot researchers in [23]. Sadness was not included as stimuli in the current study as no such movement was available from the re-targeted gestures presented in [19]. Due to the difference in dialogue for respective affective state, the stimuli differed somewhat in length (from 4.0 to 7.6 seconds). The two joyful gestures also differed in terms of variation of the nonverbal expression along the joyful axis, as described in [18] (one gesture was more joyful than the other). These joyful stimuli are here referred to as joyful clip 1 (level 5, i.e. the most joyful of the two clips) and joyful 2 (level 4, i.e. joyful, one level less joyful than the other clip).

As previously mentioned, participants rated emotions for each stimulus along four different scales (sad, joyful, frustrated, relaxed). By adding additional ratings along the non-intended dimensions (all scales were present for all stimuli) we could investigate the intensity of the emotion displayed not only along the intended dimension, but also

⁴Stimuli can be accessed from <https://kth.box.com/v/sonao2018>.

Clip	Word Frequency
Frustrated	fear (5), nervousness (3), panic (2), frustration (2), hesitance (2), annoyance (2), anxiousness (2)
Relaxed	frustrated (5), annoyed (4), confused (4), apathy (3), carelessness (2), exasperation (2), fed up (2)
Joyful 1	annoyance (3), acceptance (2), calm (2), care-free (2), friendly (2), forgiveness (2), helpfulness (2), indifference (2)

Table 1: Most frequently used words to describe the emotional reaction in respective gesture of original videos of the mime actor. Number of occurrences is given in parenthesis.

the perceived emotion towards complementary axes in the circumplex model.

In order to discuss potential acoustic differences between respective sound stimuli, we analyzed all audio clips using MIR Toolbox⁵ and the function `iosr.dsp.ltas` from IoSR Matlab Toolbox⁶. A detailed description of MIR features for respective sound clip is presented in Section 4.2.

Prior to the pilot study described in this paper, the original motion capture recordings (which were translated into expressive gestures of the NAO robot) were evaluated in a free-labeling experiment with 23 participants (10 M, 13 F) (see [18]). Results are summarized in Table 1⁷. These labels were for full body movements of the mime actor wearing a mask. Subtitles of respective dialogue were also shown (e.g. mime actor: “*Sorry I broke this glass*”, interlocutor: “*No problem I’ll fix it*”).

4.2 RESULTS

4.2.1 Acoustic Properties

Acoustic features for respective audio clip are presented in Table 2. A waveform (time domain) and spectrogram (frequency domain) of the first joyful sound stimulus is depicted in Figure 5. Long Time Average Spectrums (LTAS) of the sounds are displayed in Figure 6. The LTAS was calculated from the average power spectral density (PSD) obtained from a series of overlapping FFTs, with an FFT length of 4096 and hop size is 2048, with Hann-windowed segments. The average PSD was Gaussian-smoothed to 1/3-octave resolution.

In terms of acoustic features, the sounds did not differ much in terms of Root-Mean-Square Energy (RMS). The lowest value was obtained for the relaxed clip. The Zero-Crossing Rate (ZCR), which is an indicator of noisiness, was highest for the joyful 1 clip, followed by the relaxed clip, the joyful 2 clip, and lastly the frustrated clip, suggesting that the joyful 1 clip was the noisiest one, while the frustrated clip was the least noisy. For spectral roll-

⁵ <https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox>

⁶ <https://github.com/IoSR-Surrey/MatlabToolbox>

⁷ For the relaxed stimulus, there was a one step difference between the stimulus used in the free-labeling experiment and the stimuli used in our pilot study; the free labeling experiment used level 1, i.e. very relaxed, whereas the stimuli used in our current study used level 2, i.e. relaxed.

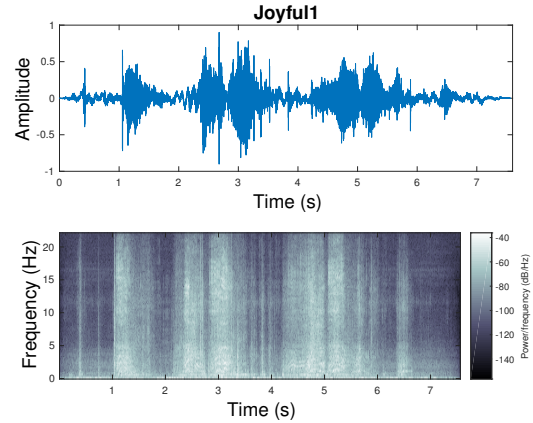


Figure 5: Waveform and spectrogram of the first joyful sound stimulus. For the spectrogram, window size was set to 256 sample points, 50% overlap.

off, which estimates the amount of high frequencies in the signal by finding the frequency that the majority of the total energy is contained below, the highest frequency was also observed for the joyful 1 clip, and the lowest for the frustrated clip. For the flatness feature, which indicates whether the distribution is smooth or spiky (this provides a way of quantifying how noise-like or tone-like a sound is), lowest value was obtained for the frustrated clip (more tone-like) and highest for the joyful 2 clip (more noisy). The brightness feature is calculated based on measuring the amount of energy above the frequency of 1500 Hz. Results indicated highest brightness for the joyful clips, and lowest for the frustrated. The spectral spread, accounting for the standard deviation of the spectrum, was highest for the joyful 1 clip, and lowest for the frustrated clip. For spectral flux, calculated as the distance between the spectrum of each successive frame, thereby being a measure of the amount of spectral change in a signal, highest rate of change was observed for the joyful 1 clip, and lowest for the relaxed. Moreover, for the roughness feature, which is an estimation of the sensory dissonance, values were notably higher for the two joyful clips than for the relaxed and frustrated clips.

Clip	Frustrated	Relaxed	Joyful 1	Joyful 2
RMS	0.07	0.06	0.08	0.09
ZCR	1051.95	1785.14	1972.65	1712.61
RollOff	12836.65	13082.09	13893.88	13785.54
Flatness	0.51	0.54	0.58	0.59
Brightness	0.71	0.77	0.80	0.80
Spread	5522.7	5531.96	5643.86	5618.22
Flux	58.57	48.86	66.78	73.40
Roughness	356.29	422.99	843.05	812.23

Table 2: MIR Features. For Zero-Crossing-Rate (ZCR), spectral roll-off and spectral spread, results are reported in Hertz (Hz).

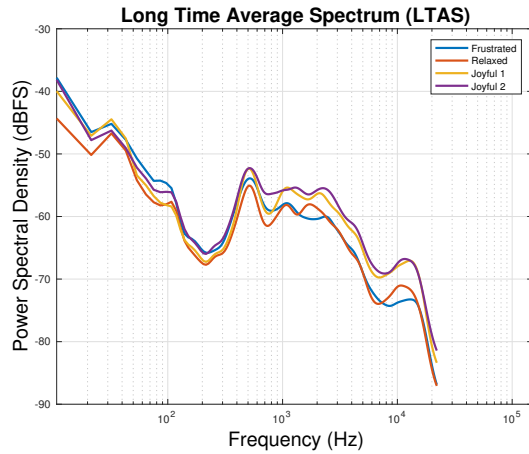


Figure 6: Long Time Average Spectrum for sound stimuli, separated by color.

4.2.2 Perceptual Ratings

A total of 40 participants took part in the experiment ($M=24$, $F=16$, average age=33.9 years). The collected data is readily available online⁸. In total, 15 participants (37.5%) were students at KTH. Most of the participants stated that they had no (17.5%), little (40%) or some (35%) experience from dance or similar motor activities. More than half of the participants (65%) had some musical experience, mainly in terms of playing a musical instrument, singing and/or composing music.

Descriptive statistics for each stimulus category (frustrated, relaxed, joyful 1 and 2), separated by condition (audio, audio-video or video) for respective emotional scale (sad, joyful, frustrated, relaxed) are shown in Figures 7a - 7d. As the data was collected on scales that displayed numeric values of equal distance, we proceeded with statistical analysis using parametric methods. With more than 30 observations per category, data should be approximately normally distributed according to the central limit theorem.

Initially, we conducted a Three-Way Repeated Measures ANOVA with the within-subjects factors stimuli category (frustrated, relaxed, joyful 1 respective 2), condition (audio, audio-video, video) and emotional scale (frustrated, joyful, relaxed, sad) to compare main effects of stimuli type, condition and scale (as well as interactions between these parameters) on perceptual ratings. Results indicated a significant main effect of emotional scale, $F(2.16, 84.39) = 31.62, p < 0.001$. There was also a significant interaction between emotional scale and stimuli category, $F(9, 351) = 16.76, p < 0.001$, and emotional scale and condition, $F(3.23, 125.83) = 3.32, p = 0.019$. Furthermore, a significant interaction between emotional scale, stimuli category and condition was found $F(9.48, 369.72) = 6.42, p < 0.001$. Mauchly's test of sphericity was significant for all of the above-discussed within-subjects effect but the interaction between emotional scale and condition, hence degrees of freedom were

corrected using Greenhouse-Geisser estimates of sphericity. Results suggest that perceptual ratings across stimuli categories were different for different scales and conditions.

For the purpose of our current project, analyzing perceptual ratings separately for different stimuli categories makes more sense for interpretation, as behavior may differ between different expressive gestures. In the sections below, we present figures based on a detailed analysis of ratings within each stimulus category (frustrated, relaxed, joyful 1 and 2), through separate Two-Way Repeated Measures ANOVAs with the following within-subjects factors: emotional scale (frustrated, joyful, relaxed, sad) and condition (audio, audio-video, video). The purpose of these tests was to investigate if there was an interaction effect between emotional scale and condition.

Frustrated Stimuli

For the frustrated stimuli, a Two-Way Repeated Measures ANOVA indicated a significant main effect of emotional scale, $F(2.17, 84.49) = 39.21, p < 0.001$. There was also a significant interaction between emotional scale and condition, $F(3.70, 144.32) = 7.49, p < 0.001$. Mauchly's test of sphericity was significant for both measures; therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. The significant interaction effect indicates that ratings across emotional scales were different for the audio, audio-video respective video conditions. Pairwise comparisons based on estimated marginal means (with applied Bonferroni corrections) indicated that the frustrated scale was rated as significantly higher than all other scales ($p < 0.001$ for all comparisons). Since we are mainly interested in differences between conditions for the frustrated scale in this context, post-hoc paired t-tests comparing these ratings were also conducted (with Bonferroni corrections accounting for 3 comparisons). There was a significant difference in ratings for the audio ($M = 1.30, SD = 1.34$) and video ($M = 2.60, SD = 1.41$) conditions; $t(39) = -4.24, p < 0.001$. There was also a significant difference in ratings for the audio ($M = 1.30, SD = 1.34$) and audio-video ($M = 2.45, SD = 1.40$) conditions; $t(39) = -3.88, p = 0.001$. In other words, higher ratings were obtained for the frustrated scale in conditions involving a video representation.

Relaxed Stimuli

For the relaxed stimuli, the Repeated Measures ANOVA indicated a significant main effect of condition $F(2, 78) = 4.89, p = 0.010$. There was also a significant interaction between emotional scale and condition, $F(5.33, 207.85) = 7.24, p < 0.001$ (degrees of corrected using Huynh-Feldt estimates of sphericity). Pairwise comparisons based on estimated marginal means (with Bonferroni corrections) indicated a significant difference in ratings for audio ($M = 1.00, SD = 0.09$) and video ($M = 1.21, SD = 0.08$) conditions, $p = 0.013$. Post-hoc paired t-tests comparing ratings specifically for the relaxed emotional scale (with Bonferroni corrections accounting for 3 comparisons) indicated a significant difference

⁸ Data can be accessed from <https://kth.box.com/v/sonao2018>

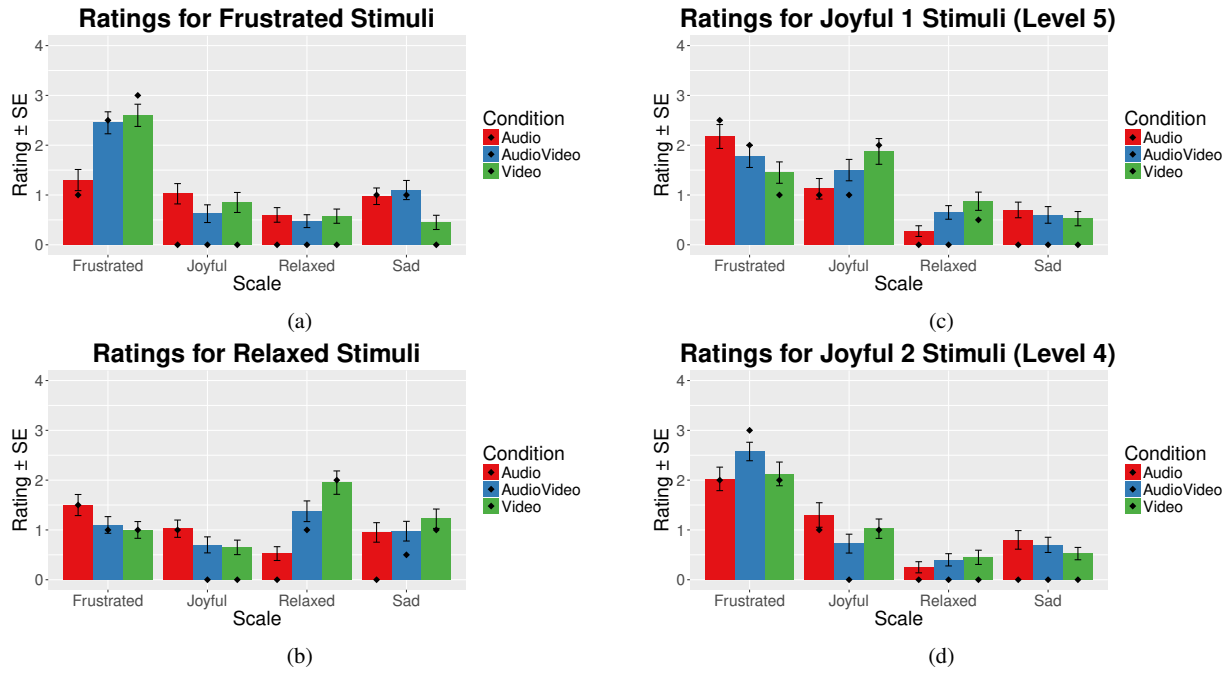


Figure 7: Mean perceptual ratings for respective stimuli category. Error bars represent Standard Error of the Mean (SEM). Medians represented with diamonds.

in ratings for the audio ($M = 0.53, SD = 0.88$) and video ($M = 1.95, SD = 1.48$) conditions; $t(39) = -6.01, p < 0.001$. There was also a significant difference in ratings for the audio ($M = 0.53, SD = 0.88$) and audio-video ($M = 1.36, SD = 1.31$) conditions; $t(39) = -3.98, p = 0.001$. To conclude, relaxed ratings were significantly higher when video was presented.

Joyful Stimuli

For the first joyful stimuli (level 5, i.e. the most joyful of the two clips), the Repeated Measures ANOVA indicated a significant main effect of emotional scale $F(2.20, 85.89) = 20.22, p < 0.001$ (degrees of corrected using Greenhouse-Geisser estimates of sphericity). A significant interaction between emotional scale and condition was also observed, $F(3.77, 147.68) = 3.97, p = 0.005$ (df corrected using Greenhouse-Geisser correction). Pairwise comparisons based on estimated marginal means (with Bonferroni corrections) between emotional scales indicated significantly higher values for the frustrated scale, compared to relaxed and sad scales (all with $p < 0.001$). Moreover, significantly higher values were obtained for the joyful scale than for the relaxed and sad scale ($p < 0.001$). Post-hoc paired t-tests comparing ratings specifically for the joyful emotional scale (with Bonferroni corrections accounting for 3 comparisons) indicated significantly lower ratings for the audio ($M = 1.13, SD = 1.30$) than for the video ($M = 1.86, SD = 1.64$) condition; $t(39) = -2.64, p = 0.035$. Interestingly, rather high ratings were obtained for the frustrated scale for these stimuli (as well as for the joyful 2 stimuli), despite the fact that the originally intended emotion was joy. However, no significant dif-

ferences were observed between conditions for frustrated ratings.

Also for the second joyful stimuli category (level 4, i.e. joyful, one level less joyful than the other joyful clip), the repeated measures ANOVA indicated a significant main effect of emotional scale $F(2.49, 90.99) = 39.76, p < 0.001$ (degrees of corrected using Huynh-Feldt estimates of sphericity). Pairwise comparisons based on estimated marginal means (with Bonferroni corrections) indicated significantly higher values for frustrated ratings, compared to the other emotional scales (all had $p < 0.001$). The joyful ratings were significantly higher than for the sad ratings ($p < 0.001$), but significantly lower than the frustrated ratings ($p < 0.001$).

5. DISCUSSION

It is evident that a lot of information is lost when the mechanical robot sounds were presented without a video representation of respective expressive gesture. Pairwise comparisons for three out of the four stimuli categories (frustrated, relaxed, joyful 1 and 2) indicated significantly lower values for the intended emotional scale (e.g. frustrated rating for the frustrated category) in audio conditions than in video conditions. The mechanical sounds themselves appeared to contribute to emotional coloring only to a small extent. An emotional coloring effect, indicated by high perceptual ratings of emotional scales, was mainly present for the joyful stimuli group. A detailed discussion on perceptual ratings of e.g. frustration for the joyful stimuli is presented below.

Since significant interaction effects were observed for all

stimuli categories but the joyful 2 stimuli, interpretation of main effects should be done with caution, as there may be carryover effects. Generally, the frustrated stimuli appeared to be the most successful category in terms of communication of emotions, at least for the video and audio-video conditions. There was a significant interaction between emotional scale and condition, suggesting that ratings across emotional scales were different for the audio, audio-video respective video conditions. Even if ratings of video and audio-video stimuli received rather high ratings, the mechanical robot sound in isolation did not appear to communicate much frustration ($M = 1.30$, $SD = 1.34$).

Results for the relaxed stimuli category were not as strong as for the frustrated one. Also for this category a significant interaction between emotional scale and condition was observed, however, as seen in Figure 7b, ratings for the relaxed scale were generally rather low. Interestingly, the audio condition had significantly lower ratings than the audio-video and video conditions. In fact, the sound stimulus was not rated as relaxed at all, with a mean value of 0.53 ($SD = 0.88$). From these results we can conclude that the presence of sound appeared to counteract the perception of the video in this case; ratings were lower when the relaxed sound was added.

Perhaps the most interesting results were obtained for the joyful stimuli categories. For these stimuli, there appear to have been some confusion between the frustrated and joyful scales. As can be seen in Figure 7c and 7d, frustrated ratings were rather high. Interestingly, there were some differences between the two joyful clips. For example, for the first joyful stimuli (joyful 1) there was a significant interaction between emotional scale and condition, but no significant interaction was observed for the second joyful stimuli (joyful 2, one level less joyful than joyful 1). For the second stimuli, pairwise tests based on estimated marginal means indicated significantly higher values for the frustrated scale than for the joy scale. In general, it appears as though the communication of valence (positive versus negative) was somewhat ambiguous for the joyful stimuli. The arousal dimension (excited versus calm), on the other hand, appeared to be easier to interpret. Participants appear to have identified that communicated emotion was not relaxation or sadness for these stimuli (emotions characterized by low valence, as indicated in the circumplex model of affect, see Figure 4). Instead, rather high ratings were obtained for the frustrated and joyful scales (emotions characterized by high arousal). The confusion between frustration and joy suggests that the distinction between positive and negative valence is not clear in this context.

Possibly, the reason why the joyful sounds were generally rated as frustrated might be explained by their spectral properties and overall amplitude envelope characteristics. As mentioned in Section 4.2.1, the joyful stimuli had high spectral roll-off, i.e. a large amount of high frequency energy. They also had higher values on spectral flatness and somewhat higher brightness values, compared to other audio clips. Moreover, both joyful clips had somewhat higher spectral spread and spectral flux than other au-

dio clips. Finally, both of the joyful audio clips had notably higher roughness. In the context of emotion recognition in speech, spectral and F0 features have been found to produce the most accurate predictions of valence (i.e. positive versus negative) [24]. Interestingly, angry and happy sentences have been found to share similar acoustic patterns [25]. In other words, these emotions are similar in the activation domain but different in the valence domain, at least for speech sounds.

The premise of the pilot study discussed in this paper was that perception of mechanical sounds produced by the NAO robot can affect the overall perception of the robot's affective communication. Depending on how the mechanical sounds are perceived, they could be processed and blended with movement sonification, in order to further emphasize or attenuate emotional coloring. In other words, the consequential sounds could be used as input for other intentional sounds, e.g. as sound source in carefully designed movement sonification. Our results suggest that there was a low correspondence between perceptual ratings of sounds and the originally intended emotion of respective stimuli category (compare e.g. frustrated ratings for the frustrated stimuli category or relaxed ratings for relaxed stimuli categories). The mechanical sounds inherent to the NAO's movements appear not to be clearly linked to the emotional scales used in the current study (sad, joyful, frustrated and relaxed). However, the sounds appear to have some emotional content. For example, the joyful sounds were rated as being somewhat frustrated.

Considering that certain mechanical sounds produced by the NAO robot can communicate other affective states than those originally intended to be expressed through stylized gestures, one should perhaps consider whether the mechanical sounds should be completely masked, or altered to a rather large extent, when sonifying movements. Possibly, sounds produced by the NAO's movements could be processed in order to minimize particular audio features associated with certain affective states. In particular, the valence dimension could be emphasized and clarified through the use of certain sounds. Perhaps the confusion between the frustrated and joyful scales for the joyful stimuli might be explained by the fact that the joyful sound did not express any evident positive valence. This is something that could easily be corrected through the use of e.g. a major mode. In any case, it is important to take consequential sounds into consideration when designing movement sonifications for robots, so that mechanical sounds do not interfere with the interpretation of the sonification.

After our study, some subjects gave verbal accounts of their experience of participating in the experiment. Several subjects emphasized that it was difficult to interpret the NAO robot, as it did not show any facial expressions. Some participants stated that the motor sounds were rather disturbing to listen to and did in fact not communicate much emotion at all. Our results can be compared to previous findings presented in [6], indicating that participants had overall negative impressions of motor sounds. However, some of our participants stated that the sounds indeed communicated certain emotional properties and did not mind

listening to the mechanical sounds. The authors suspect that the overall experience of listening to these sounds might be affected by level of musical experience. Some participants may be very accustomed to listening to noise and abstract sounds as they are familiar with e.g. contemporary music, whereas others have never before been exposed to such sounds in an active-listening context.

The movement sequences that were used in the study described in this paper were restricted by the fact that they had to be optimized to subset of videos selected for [19], taken from the original dataset presented in [18]. The conclusions that can be drawn from the free-labeling results shown in Table 1 are that communication through merely body movements (without facial expressions) could be ambiguous and very context specific. A lot of information is lost when no facial expression is shown. Moreover, it should be noted that labeling of emotions, or ratings of such, is easier if you first present a reference to what is actually a “neutral” emotion in the particular context (as in e.g. [26]). Not presenting such an example stimulus may result in large inter-subject variability.

One may suggest that robotic consequential sounds produced by humanoids should ideally function similar to sounds produced by humans; the audible sounds should primarily be frictional sounds produced by interacting with external objects, rather than sounds produced by gears in servo motors. Technologies such as non-g geared brushless motors that operate silently⁹ may enable development of more quiet robots, thereby reducing the need to compensate for undesired motor sounds in future HRI research.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the project SONAO and a pilot study performed within this project. The study focused on perceptual ratings of sounds produced by expressive movement of the NAO robot. Participants rated emotions communicated through NAO’s movements and/or sounds on a set of scales (frustrated, relaxed, sad, joyful) in three different conditions (audio, audio-video and video). The following conclusions could be drawn:

- The mechanical sounds inherent to the robot NAO’s movement appear not to be clearly linked to the emotional scales used in the current study (sad, joyful, frustrated and relaxed). For example, the sound of a relaxed movement did not necessarily sound very relaxed.
- Ratings were generally lower for audio-only conditions than for video conditions, suggesting that a lot of information is lost when transitioning to the auditory domain.
- Certain mechanical sounds might communicate emotional properties. For example, findings suggest that audio clips originally intended to express joy were perceived as frustrated.
- The mechanical sounds appear to communicate arousal more effectively than valence.

We suggest that future work on sonifications of expressive robot gestures should consider that sounds inherent to robot movement can communicate other affective states than those originally intended to be expressed through respective gesture. Such mechanical sounds should therefore be either masked or altered, before being incorporated into the sound design and movement sonification of a humanoid robot.

Future experiments within the SONAO project should involve more detailed evaluations of sounds inherent to robot movement, using on a free-labeling descriptor methodology. Evaluations of larger datasets, including also some sad stimuli, should also be done. Moreover, large-scale studies focusing on correlations between perceptual ratings and acoustic features could be performed.

Acknowledgments

This project was funded by the SONAO Visionary Project funded by the School for Computer Science and Communication at the KTH Royal Institute of Technology, Stockholm, Sweden, and Grant 2017-03979 from the Swedish Research Council.

The authors would like to thank Aravind Elanjimattathil Vijayan and Iolanda Leite for their valuable contributions to the project.

7. REFERENCES

- [1] L. Langeveld, R. van Egmond, R. Jansen, and E. Ozcan, “Product sound design: Intentional and consequential sounds,” in *Advances in industrial design engineering*. InTech, 2013.
- [2] K. Inoue, K. Wada, and Y. Ito, “Effective application of paro: Seal type robots for disabled people in according to ideas of occupational therapists,” in *International Conference on Computers for Handicapped Persons*. Springer, 2008, pp. 1321–1324.
- [3] H. H. Seck, “Marine corps shelves futuristic robo-mule due to noise concerns,” *Military.com*, 2015.
- [4] K. Nakadai, H. G. Okuno, and H. Kitano, “Real-time sound source localization and separation for robot audition,” in *Seventh International Conference on Spoken Language Processing*, 2002.
- [5] H. Tennent, D. Moore, M. Jung, and W. Ju, “Good vibrations: How consequential sounds affect perception of robotic arms,” *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 928–935, 2017.
- [6] D. Moore, H. Tennent, N. Martelaro, and W. Ju, “Making noise intentional: A study of servo sound perception,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 12–21.

⁹ See e.g. <https://odriverobotics.com/>.

- [7] A. Beck, L. Cañamero, and K. A. Bard, "Towards an affect space for robots to display emotional body language," in *Ro-Man, 2010 IEEE*. IEEE, 2010, pp. 464–469.
- [8] M. S. Erden, "Emotional postures for the humanoid-robot NAO," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 441–456, 2013.
- [9] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 407–414.
- [10] S. Yilmazyildiz, R. Read, T. Belpeame, and W. Verhelst, "Review of semantic-free utterances in social human-robot interaction," *International Journal of Human-Computer Interaction*, vol. 32, no. 1, pp. 63–85, 2016.
- [11] M. Häring, N. Bee, and E. André, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *Ro-Man, 2011 IEEE*. IEEE, 2011, pp. 204–209.
- [12] C. Becker-Asano, T. Kanda, C. Ishi, and H. Ishiguro, "How about laughter? Perceived naturalness of two laughing humanoid robots," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.
- [13] J. Monceaux, J. Becker, C. Boudier, and A. Mazel, "First steps in emotional expression of the humanoid robot NAO," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*. ACM, 2009, pp. 235–236.
- [14] R. Read and T. Belpaeme, "How to use non-linguistic utterances to convey emotion in child-robot interaction," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2012, pp. 219–220.
- [15] R. Zhang, M. Jeon, C. H. Park, and A. Howard, "Robotic sonification for promoting emotional and social interactions of children with ASD," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 2015, pp. 111–112.
- [16] J. Bellona, L. Bai, L. Dahl, and A. LaViers, "Empirically informed sound synthesis application for enhancing the perception of expressive robotic movement." Georgia Institute of Technology, 2017.
- [17] L. Dahl, J. Bellona, L. Bai, and A. LaViers, "Data-driven design of sound for enhancing the perception of expressive robotic movement," in *Proceedings of the 4th International Conference on Movement Computing*. ACM, 2017, p. 16.
- [18] S. Alexanderson, C. Osullivan, M. Neff, and J. Beskow, "Mimebot - Investigating the expressibility of non-verbal communication across agent embodiments," *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 4, p. 24, 2017.
- [19] J. B. A. Elanjimattathil Vijayan, S. Alexanderson and I. Leite, "Using constrained optimization for real-time synchronization of verbal and nonverbal robot behavior," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017.
- [20] I. Ekman and M. Rinott, "Using vocal sketching for designing sonic interactions," in *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. ACM, 2010, pp. 123–131.
- [21] R. Tünnermann, J. Hammerschmidt, and T. Hermann, "Blended sonification: Sonification for casual interaction," in *ICAD 2013-Proceedings of the International Conference on Auditory Display*, 2013.
- [22] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.
- [23] S. Song and S. Yamada, "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 2–11.
- [24] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [25] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso, "An acoustic study of emotions expressed in speech," in *Eighth International Conference on Spoken Language Processing*, 2004.
- [26] R. Bresin and A. Friberg, "Emotion rendering in music: Range and characteristic values of seven musical variables," *Cortex*, vol. 47, no. 9, pp. 1068–1081, 2011.