Perceptual Evaluation of Blended Sonification of Mechanical Robot Sounds Produced by Emotionally Expressive Gestures: Augmenting Consequential Sounds to Improve Non-verbal Robot Communication

Emma Frid, Roberto Bresin

International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00788-4

Perceptual Evaluation of Blended Sonification of Mechanical Robot Sounds Produced by Emotionally Expressive Gestures: Augmenting Consequential Sounds to Improve Non-verbal Robot Communication

Emma Frid[1] · **Roberto Bresin**[1]

Accepted: 17 April 2021

Abstract
This paper presents two experiments focusing on perception of mechanical sounds produced by expressive robot movement and blended sonifications thereof. In the first experiment, 31 participants evaluated emotions conveyed by robot sounds through free-form text descriptions. The sounds were inherently produced by the movements of a NAO robot and were not specifically designed for communicative purposes. Results suggested no strong coupling between the emotional expression of gestures and how sounds inherent to these movements were perceived by listeners; joyful gestures did not necessarily result in joyful sounds. A word that reoccurred in text descriptions of all sounds, regardless of the nature of the expressive gesture, was "stress". In the second experiment, blended sonification was used to enhance and further clarify the emotional expression of the robot sounds evaluated in the first experiment. Analysis of quantitative ratings of 30 participants revealed that the blended sonification successfully contributed to enhancement of the emotional message for sound models designed to convey frustration and joy. Our findings suggest that blended sonification guided by perceptual research on emotion in speech and music can successfully improve communication of emotions through robot sounds in auditory-only conditions.

Keywords Sonification · Non-verbal sounds · Expressive gestures · Emotions in robotics · Affective computing · Sonic Interaction Design

1 Introduction

Non-verbal sound plays an important role in communication between humans. As robots are gradually becoming an integral part of modern society, it also becomes increasingly important that these agents can express their internal states through non-verbal communication. The work described in the current paper focuses on sounds of humanoid robots. Research on sonic interactions that allow robots to express intention and emotion through sounds is an emerging field in Human–Robot Interaction (HRI). There are numerous examples of studies focusing on sounds in social HRI (see e.g. the extensive review of semantic free utterances presented in [54]). Some research has also focused on how Foley artists, the people who generate sounds in movies using their own body motions (e.g. footsteps) [36], use sounds to express and enhance emotional reactions of robots in movies [8,24,28]. As opposed to Foley artists, who produce sounds for films, sound designers who create sounds for non-virtual robots have to consider that robots in real life produce sounds as a result of their mechanical movements. Despite novel technologies such as non-geared brush-less motors which may operate silently, most modern robots are far from silent. Robots often rely on servo motors in order to move, and thus their movements produce sounds. These sounds could potentially affect the HRI in the sense that they may alter the interpretation of the message conveyed by the robot. For instance, mechanical sounds could influence interpretation of emotional reactions of a robot, especially if it is out of sight. Despite the fact that sounds inherent to robot movement could implicitly convey meaning and affect social interac-

This project was supported by the SONAO Visionary Project funded by the KTH Royal Institute of Technology, Stockholm, Sweden, and Grant 2017-03979 from the Swedish Research Council.

✉ Emma Frid
emmafrid@kth.se

Roberto Bresin
roberto@kth.se

[1] KTH Royal Institute of Technology, Stockholm, Sweden

Published online: 22 May 2021

tion, sound design is still often an overlooked aspect in the field of HRI.

The relationship between sounds and the gestures generating them has been extensively researched in the fields of Sonic Interaction Design (SID) [14] and musical gestures [19]. It has been found that sound conveys information about the nature of the gesture that has produced it if we can recognize the sound's source. Persons listening to recorded sounds have successfully been able to mimic the gestures corresponding to the creation of respective sounds, reproducing for example the gesture of crushing a metallic can [9] or a pianist performing a piece of music [18]. Sound also plays a role in our aesthetic, quality, and emotional experience of consumer products [27]. Several sound quality analysis tools have been developed to evaluate how consumer preference relates to a product sound and to quantify this preference based on objective measurements [37]. Interestingly, noise has been found to have a negative influence on overall pleasantness of products [13] and auditory cues have been found to influence product perception at a semantic level [45]. Langeveld et al. [27] make a distinction between sounds that are generated by the operation of the product itself (*consequential sounds*), and sounds that are intentionally added to a product (*intentional sounds*). In the context of HRI, we both have to consider sounds that are specifically designed for the communication of a robot's functions and emotional reactions (intentional sounds) and sounds produced by its movements (consequential sounds).

Although sound has been stressed to have an implicit influence on human-robot interactions [33], relatively little work in the field of HRI has focused on consequential sounds and perception of sounds inherent to robot movement. 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. [35]). Interestingly, there are several examples in which the sound design of a robot has been neglected, resulting in significant effects on the HRI. For example, motor sounds of the pet robot Paro negatively interfered with interactions [23], and the Boston Dynamic's LS3 pack-mule robot was found to be "too loud" to be integrated in military patrols, as it could endanger troops by giving up their position [44].

Up to this point, little work has been done on sonification in the context of HRI (see e.g. [40,56]). In particular, little research has focused on augmenting expressive robotic movement with sounds (see e.g. [2,10]). The work presented in this paper aims to contribute to the field of sound design in social robotics (see e.g. [29,30,54]) by designing and evaluating non-speech sounds for enhancing and supporting robot movements and their emotional expression when communicating with humans.

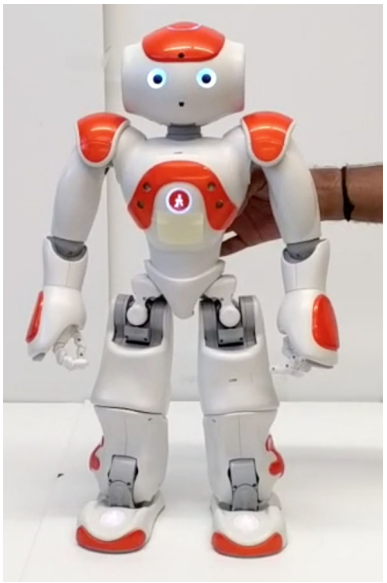

Fig. 1 The NAO social robot

In this study we used a humanoid social robot, NAO[1] (see Fig. 1). The NAO robot has been used for example in education, entertainment, and health care applications. It provides 25 degrees of freedom, is 58 cm tall, and is equipped with tactile sensors, an inertial unit, 2D cameras, sonar, omnidirectional microphones, and two loudspeakers[2]. Previous research on creating affective sounds for the NAO robot have mainly adopted other sound synthesis paradigms [31,39,41]. The current work is novel in the sense that it aims to fill this gap by incorporating movement sonification in non-verbal robot communication.

2 Background

The work described 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. The purpose is to compensate for limitations in robot communicative channels with an increased clarity of NVC through expressive gestures and non-verbal sounds. For more details about the *SONAO* project, please see our previous paper [15]. In the current study, we present two experiments focusing on evaluation of sounds inherent to movements of a NAO robot and discuss how these sounds could be used in blended sonification (see definition in [51], which is also presented in the section below).

[1] https://www.softbankrobotics.com/emea/en/nao.

[2] https://www.softbankrobotics.com/emea/sites/default/files/press-kit/NAO-press-kit-EN.pdf.

Sonification is a rather young discipline focusing on translating data into non-speech sound in a systematic and reproducible way. Multiple definitions of sonification have been proposed. The most commonly agreed upon one is given in the *NSF Sonification Report* [26], where the term is defined as *"(...) the use of nonspeech audio to convey information. More specifically, sonification is the transformation of data relations into perceived relations in an acoustic signal for the purpose of facilitating communication or interpretation."* Thomas Hermann later expanded on this definition of sonification to: *"(...) data-dependent generation of sound, if the transformation is systematic, objective and reproducible, so that it can be used as a scientific method"*[3]. *Interactive Sonification* focuses on the interactive representation of data by means of sound, and it is therefore suitable when real-time feedback is required [6,22]. The method can be considered the acoustic counterpart of interactive visualization [46]. Interactive Sonification makes use of sound for exploring data in a fast and meaningful way, and it is especially suitable for data that change over time, such as those collected from body movements [7,52].

In the current study, we use blended sonification, a sonification technique that involves *"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"* [51]. In other words, the sound resulting from the manipulation and blending of both the original sound (in our case produced by the robot) and new added sounds is perceived as a whole, i.e. as a coherent sound (not as two separate sound sources).

Previous research on consequential sounds in HRI includes work by Tennent et al. [49], who investigated perception of sounds generated by robotic arms in a 2 × 3 experimental design, with 2 contexts (social versus functional) and 3 different sound conditions (no sound, sound from a high-end robotic arm vs sound from a low-end robotic arm). The authors found that robot motor sounds negatively color visual perception of interactions presented in videos and that the sounds of the robotic arm significantly reduced how positively people perceive a robotic arm. Interestingly, sounds from a high-end robot increased ratings of perceived competence when performing a social task (in this case, the robotic arm placed a block in a person's hand), but decreased the ratings of perceived competence when performing a functional task (the robotic arm placed a block on a preexisting tower of blocks, i.e. there was no human interaction). Overall, the social conditions had higher ratings nearly across the board, highlighting that context appears to play a significant role in how sound is interpreted.

[3] https://sonification.de/son/definition/.

A purely acoustic-driven study on aural impressions of servo motors commonly used to prototype robotic motion was presented in [34]. The authors constructed a framework to objectively and subjectively characterize sound using acoustic analyses and novice evaluators, checking for correlations between objective measures and subjective preference. Participants evaluated the sounds through pairwise comparison, in which they made subjective ratings of two servo motor sounds. They also left qualitative commentary. Overall, subjective measures of sound correlated weakly with objective acoustic measures. Moreover, qualitative commentary suggested negative impressions of the sounds overall.

The extent to which robot noise affects proxemics in HRI was explored in a study described in [50]. The authors also investigated how noise can be eliminated in order to be more tolerable in a real world setting based on masking roughness. Noise was masked by addition of a signal inspired by natural sounds and music. Results partially confirmed that the masked sound succeeded in nullifying the negative effects of the added noise. Another recent online study on motor sounds was presented in [32]. In this work, participants evaluated servomotor sounds using two methods from sensory science, *Check All That Apply (CATA)* questions and *Polarized Sensory Positioning (PSP)*. CATA involves checking all words that you associate with a particular sound, whereas PSP involves comparing a sound to references by moving a slider to indicate how similar or different sounds are. The authors discuss benefits and limitations of applying these methods to study subtle phenomena (in this case, subtle differences in robot sounds) within the Human-Computer Interaction (HCI) community.

3 Perceptual evaluation of robot sounds

The research presented in this paper builds on our previous work on perception of sounds produced by expressive movements of a humanoid NAO robot [15]. Results from this work suggested that mechanical sounds inherent to expressive movements of the NAO robot were not clearly coupled to the emotional reactions associated with respective movements. For example, sounds produced by a joyful gesture conveyed a sensation of frustration. We also observed that certain mechanical sounds did convey emotional characteristics when presented in an auditory-only condition. Moreover, sounds generally communicated arousal more effectively than valence.

In the current work, we expand on the previous study, taking a mixed-methods approach combining quantitative and qualitative methods. While the previous study focused on quantitative ratings of emotions for robot sounds, the first experiment of the study presented in the current paper focuses on descriptions of sounds in free-form text. The purpose of

this approach was to explore and characterize robot sounds going beyond a set of predefined scales. Following up on this experiment, we also conducted a second experiment to explore how blended sonification can be used to enhance certain acoustic characteristics of robot sounds in order to improve clarity of non-verbal robot communication. In other words, the study consisted of two separate parts, Experiment 1, focusing on descriptions of the original sounds produced by expressive movements of a NAO robot, and Experiment 2, focusing on perceptual ratings of these sounds and blended sonifications thereof. A schematic representation of the procedure is presented in Fig. 2.

The experiments were organized as follows: participants were first welcomed by the instructor (author 1) and then received instructions on the first page of an online form designed for data collection. They were informed that the *SONAO* project focused on HRI, but did not know about the origin of the sounds used in the experiments, i.e. that the sounds were produced by expressive robot movements. After the initial instructions, participants were asked to fill out a demographics and musical experience form. They then proceeded with the experiments. Experiment 1 focused on labelling sounds using free-form text annotations. Experiment 2 focused on rating emotions conveyed by sonifications on a set of predefined emotional scales. The participants took part in both experiments and performed them after each other (the same participants took part in Experiment 1 and 2). The experiments were carried out in a lab setting at KTH Royal Institute of Technology and KMH Royal College of Music in Stockholm. Participants listened to the sounds in an online web interface and the experiment was purely acoustic-driven: no video representation of the robot's gestures were shown.

3.1 Participants

A total of 31 participants (14F, avg age = 36.26 yrs) took part in the experiments. However, one participant did not complete the second experiment, reducing the number of participants included in the data analysis to 30 (14F, avg age = 36.23) for Experiment 2. In our previous study [15], we observed that some participants found it hard to put into words the sonic qualities of the mechanical sounds produced by the NAO robot. We hypothesized that the overall experience of listening to these sounds might be affected by level of musical experience; musicians might be more accustomed to listening to (as well as describing) abstract sounds since they are more familiar with e.g. contemporary music. Since our current experiment made use of the same original recordings of robot sounds that were used in our previous study [15], a prerequisite for participants to take part in the current study was to have some musical experience. In other words, the decision to recruit participants with a certain level of musical expertise was guided by the hypothesis that this would result in more rich and informed free-form text descriptions of the evaluated sounds. Moreover, it has been shown in several investigations (see for example [3,43,47]) that people with musical skills have shown to acquire an auditory expertise that is not found in laypeople. This expertise includes for example more precise detection of pitch height, discrimination of specific audio streams in noisy situations, and high sensitivity in the identification of frequency deviations (such as in miss-tuned sounds). Therefore, musical experts can provide more stable and reliable data for our investigation, and as a consequence help us in the design of reliable and robust sonic feedback from a robot.

Participants were recruited from the staff and students from the KTH Royal Institute of Technology, staff from the KMH Royal College of Music, as well as amateur musicians from the KTH Symphony Orchestra (KTHs Akademiska Kapell, KTHAK) and the KTH brass band (Promenadorquestern, PQ). Some students from the Interactive Media Technology master programme which had previously completed courses in Sound and Music Computing also took part. In total, 17 of the participants were students at KTH, and 14 participants were not. The level of musical expertise ranged from expert/full-professional activity as musician or singer (9 participants) to little experience as amateur musician/singer (7 participants). A total of 7 participants reported semi-professional activity as a musician or singer with several years of practice, and 7 participants reported being advanced amateur musicians/singers with some years of practice. A mixed-model analysis revealed no significant between-subjects effect of musical expertise on ratings, nor any significant interactions with musical expertise.

3.2 Compliance with Ethics Standards

The authors declare that this paper complies with the ethical standards of this journal. All subjects gave informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki. At the time that the experiments were conducted, no ethics approval was required from KTH for behavioral studies such as the one reported in this paper. For the management of participants' personal data, we followed regulations according to the KTH Royal Institute of Technology's Ethics Officer. Participants did not receive any monetary compensation, however, subjects from the Symphony Orchestra and the KTH brass band received a cinema ticket for their participation.

3.3 Technical Setup

The two experiments were conducted after each other at the KTH Royal Institute of Technology lab, using a laptop connected to a pair of Genelec 8030B speakers. For the partic-

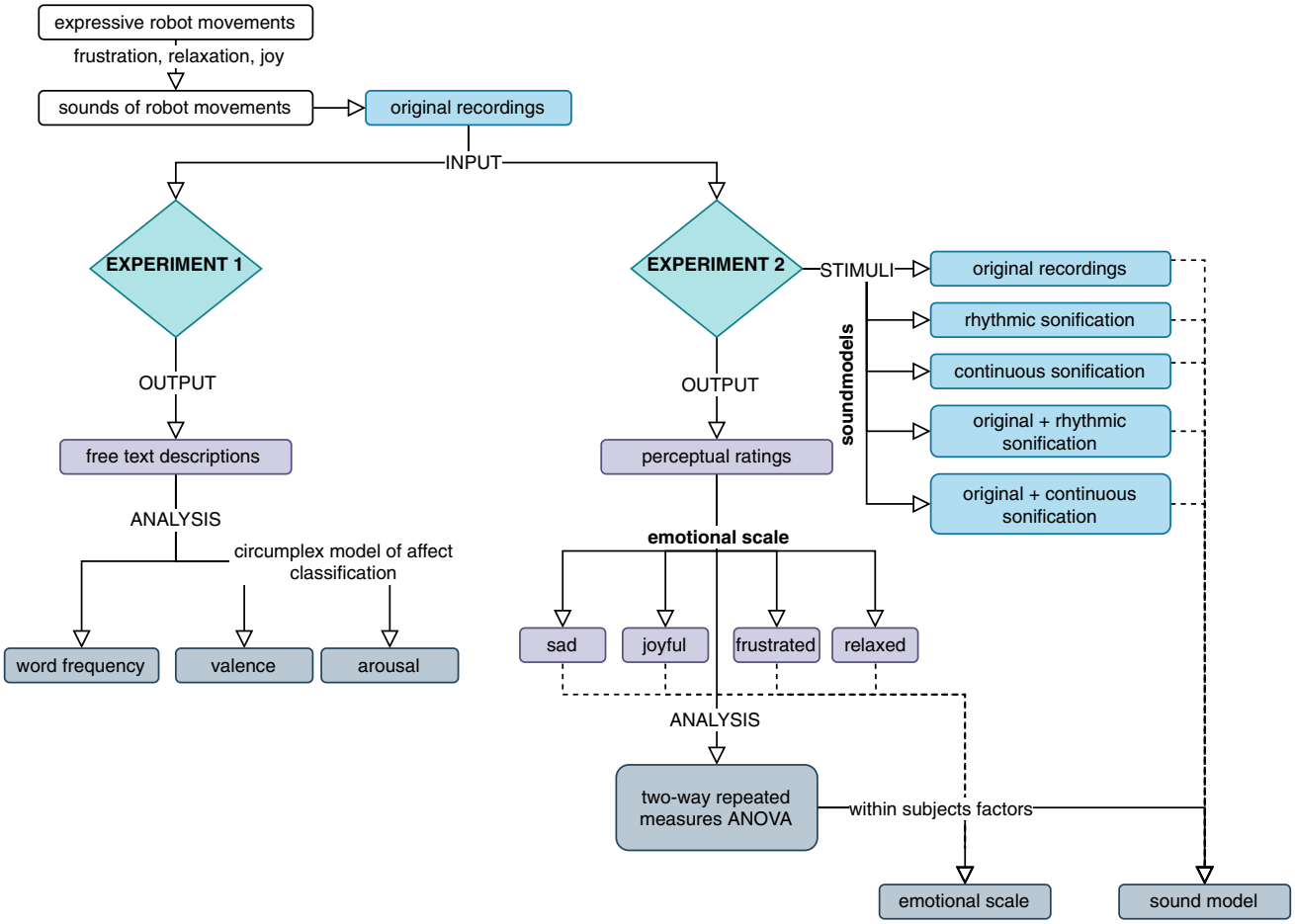

Fig. 2 Schematic representation of the methodology

ipants from KMH Royal College of Music, the experiments were conducted in their studios, with Genelec 8250A speakers.

3.4 Experiment 1

In our previous work we observed that some participants described the mechanical sounds produced by the NAO robot as unpleasant and disturbing [15]. We wanted to know if participants had positive or negative associations to these mechanical robot sounds overall. Moreover, we wanted to investigate how participants would describe these sounds, as well as the emotions that they conveyed, when being allowed to answer in free-form text, i.e. when participants were not restricted to predefined scales. One advantage of using free-form text descriptions is that coherence in listener's responses could be of more significance than if they were merely items checked from a pre-defined list [16].

3.4.1 Stimuli

The same audio recordings that were used in [15] were used in the current study[4]. These recordings were sounds of a NAO robot performing expressive gestures (frustration, relaxation and joy), i.e. sounds produced by the mechanical movement and engines of the NAO robot. There was a total of 4 stimuli, for the following emotions: frustration, relaxation and joy. The frustrated sound file was 9 seconds long, the relaxed sound file was 6.5 seconds and the two joyful sound files were 10 versus 9 seconds. The two joyful sounds were produced by two gestures that differed in terms of variation of the non-verbal expression along a joyful axis, as described in [1]. Two versions were included for comparative purposes. The gesture that produced the first sound file was rated as more joyful than the gesture producing the second sound (see [1]).

[4] Sound files are provided as supplementary material.

3.4.2 Procedure

The presentation order of the stimuli was randomized for each participant. The participants could listen to the sounds as many times as they wanted. For each stimulus, they were asked the following questions:

1. Please describe the emotion(s) that you think that this sound conveys.
2. Do you have any other comments about this particular sound?

Participants had to answer question 1 but were not required to answer question 2. There was no limitation in terms of how many words that could be used for respective question.

3.4.3 Analysis

It has been shown in previous research on emotions in film music that a two-dimensional model (valence, arousal) of emotions gives comparable results to that of a three-dimensional model (valence, arousal, tension) [11]. Therefore, for the sake of simplicity, we adopted a two-dimensional model for the classification of emotional descriptions in our experiment. Since the majority of the text entries were in the form of a single word or few synonyms, we decided to take a high-level approach focusing on aspects related to motion (activity) and emotion (valence) expressed by these words, based on a categorization into the two-dimentional circumplex model of affect [38]. The model provides a dimensional approach, in which all affective states arise from two fundamental neurophysiological systems, one related to valence (a pleasure-displeasure continuum) and the other to arousal, or alertness [42]. As such, we categorized the free-form text words and sentences into two dimensions, based on their arousal (activity) and valence. A schematic representation of the two-dimensional circumplex model of affect is displayed in Fig. 3. The categorization based on the circumplex model enabled evaluation of the descriptions for respective sound file along a limited set of emotional dimensions.

In the first step of this analysis, lists of words for respective sound stimulus were given to the authors. The authors were not aware of which list that originated from which sound file. Separate analyses were then conducted for respective word list. In the first step of the analysis, both authors independently classified keywords based on the activity dimension, categorizing words into the following categories: *activation* (high arousal), *deactivation* (low arousal) or *in between*. In the next step, the same words were further categorized into positive or negative valence categories. Finally, the encoding of the two authors' lists were compared. An inclusion criteria was defined so that words were included in the results if both authors agreed in their categorization along both dimensions.

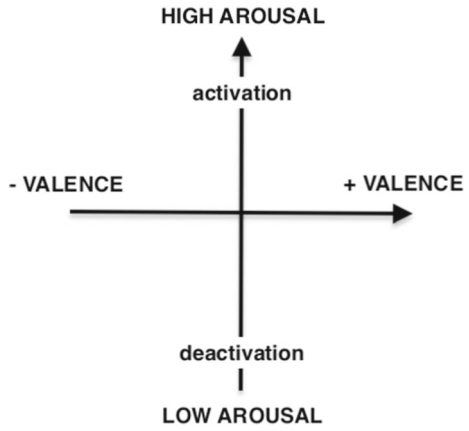

Fig. 3 Two-dimensional circumplex model of affect. The x-axis depicts the valence dimension, which goes from negative to positive. The y-axis depicts the arousal or activity dimension, which goes from deactivation to activation

For example, if a word was described in terms of activation and positive valence by both authors, it was included in the final results. This approach was used to remove words with ambiguous meanings. The schematic representation of analysis procedure is shown in Fig. 4.

3.5 Experiment 2

One of the conclusions presented in our previous work was that the emotional expression of a NAO robot's gestures is not necessarily tightly clearly coupled with how sounds produced by these gestures are perceived [15]. In other words, sounds inherent to robot movements can influence how a robot's gesture is perceived. This can be problematic, especially if the sound communicates something that contradicts the gesture. To solve this issue, we wanted to investigate if mechanical robot sounds could be processed in a way that enhances, rather than disturbs, the emotion conveyed through robot gestures. This blended sonification strategy is described in detail below.

3.5.1 Stimuli

Two different sonification models were implemented: one "rhythmic sonification", producing shorter sounds with regular or irregular Inter-Onset-Intervals (IOI), and one "continuous sonification", producing a continuous stream of sounds, without interruptions. The sonification models were designed based on previous research on emotions in speech and music. A detailed description of the sonification for respective emotion (i.e. expressive gesture) is presented below. The sound design is also summarized in Table 1. Sound files for respective model are available as supplementary material.

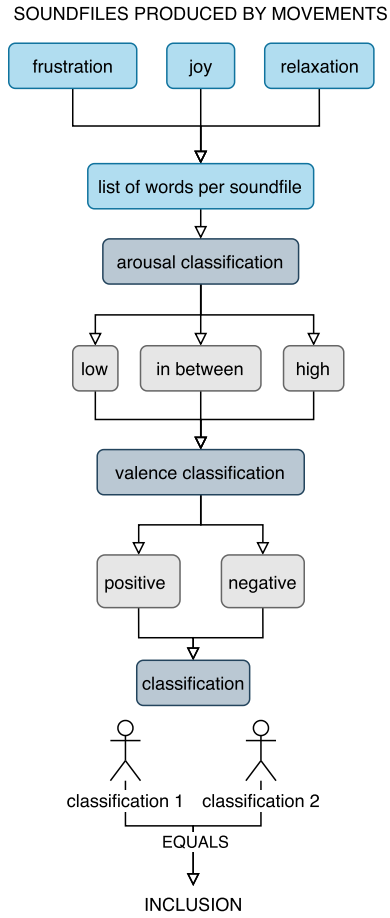

Fig. 4 Schematic representation of the qualitative analysis

Table 1 Sonifications

Sonification	Emotion	Sound synthesis
Rhythmic	Frustrated	Sample-based (granular) synthesis using grains from the original frustrated recording, with random length (range 200–300 ms) and random IOI (range 80–112 ms). The grains were pitch-shifted 1-3 times original pitch (depending on magnitude of the original signal). Also included a distortion effect
Rhythmic	Relaxed	Pulses of bandpass-filtered noise (center frequency 800–1000 Hz, Q-value 0.4–0.3) fed through a reverb, IOI 700 ms
Rhythmic	Joyful	Rectangular (pulse) oscillator (C major scale) with an envelope duration of 150 ms, IOI 180 ms. Pitch was selected based on magnitude of the input signal
Continuous	Frustrated	Sample-based (granular) synthesis, with 200–220 ms grains taken from the original frustrated recording, pitch-shifted 1–4 times original pitch (depending on magnitude of the original signal)
Continuous	Relaxed	Resonant bandpass-filtered noise (center frequency 300–500 Hz, Q-value 0.2–1.0) with a reverb effect
Continuous	Joyful	FM synthesizer (C major scale, tonic + major third). Pitch was selected based on magnitude of the input signal

Since the mechanical sounds of the NAO robot are always present, unless they are masked by other sounds, we decided to investigate how the sonifications would be perceived when presented in combination with the original recordings of the expressive gestures. Sonifications were therefore mixed with original recordings. Since we also wanted to investigate how the sonifications were perceived when presented alone, the final stimuli collection consisted of three emotions,[5] presented in five different soundmodel conditions:

- Original sound file
- Rhythmic sonification
- Continuous sonification
- Original sound file + rhythmic sonification
- Original sound file + continuous sonification

Since there were three emotions and five conditions, a total of 15 stimuli was obtained. The output level of all rhythmic sonifications was obtained by scaling the magnitude of the original input sound file using an exponential scale. For the continuous sonifications, output level of the sound was mapped to peak amplitude of the original signal.

3.5.2 Rhythmic Sonification

The speed at which sonic events are produced (events/second) is one of the most important cues for the communication of emotions in both speech (speech rate) and music (tempo). The speed can be perceived if there is a rhythmic regularity in the display of the sonic events, otherwise it is difficult to perceptually identify it [12]. In a previous study we have found that experienced musicians chose clearly defined speed values for communicating different emotional intentions in music

[5] The sound produced by the most joyful gesture was used for the joyful stimulus. This was the first joyful stimulus used in Experiment 1.

performances [5]. Therefore, we decided to test if introducing a rhythm in the sonification of the robot movements, for communicating their speed, could help the perception of its emotional intentions.

While little work has focused specifically on the term *frustration* in the context of emotion in music and speech research, substantial work has been conducted on the term *anger*. In the review on emotion in vocal expression and music performance conducted by Juslin and Laukka [25], anger is said to be associated with fast speech rate/tempo, high voice intensity/sound level, much voice intensity/sound level variability, much high-frequency energy, high F0/pitch level, much F0/pitch variability, rising F0/pitch contour, fast voice onsets/tone attacks, and microstructural irregularity. In a study focusing on modeling the communication of emotions by means of interactive manipulation of continuous musical features, changes in loudness and tempo were associated positively with changes in arousal, but loudness was dominant [5].

Based on above described findings, the frustrated sonification was designed to be characterized by fast and loud sounds, with a lot of intensity variability, and irregular rhythms. This was achieved by triggering a simple sample-based (granular) synthesis engine at the first peak of the original sound file. The synth was then triggered based on the magnitude of the original frustrated audio recording. The synthesis engine produced grains of random size between 200-300 ms.[6] The grains were randomly sampled from the original frustrated sound recording. The pitch of each grain was shifted based on the magnitude of the original audio signal. Finally, distortion was added to the outputted sound.

In contrast to frustration, which is characterized by negative valence and high arousal, *relaxation* is characterized by positive valence and low arousal. In general, positive emotions appear to be more regular than negative emotions; irregularities in frequency, intensity and duration seem to be a sign of negative emotion [25]. In music, differences in arousal are mainly associated with differences between fast and slow tempi [16]. Relaxed speech has been found to more whispery and breathy than stressed speech [17]. We designed the relaxed sonification model so that it presented soft sounds with little frequency and amplitude variability. These sounds were presented with a regular tempo. This was achieved by generating pulses of noise, with 700 ms time difference, that were filtered through a resonant band-pass filter, with variable center frequency (800–1000 Hz) and Q-value (0.3–0.4), depending on absolute magnitude of the input sound. The output was then feed through a reverb.

In Western music, differences in valence are mainly associated with major versus minor mode [16]. For the purpose of our study, the emotion *joy* could be considered to be similar to the sensation of *happiness*. According to Juslin and Laukka [25], happiness is associated with the following cues: fast speech rate/tempo, medium to high voice intensity/sound level, medium high-frequency energy, high F0/pitch level, much F0/pitch variability, rising F0/pitch contour, fast voice onsets/tone attacks, and very little microstructural irregularity. In [4,5], authors conclude that happy music performances are characterized by a relatively fast tempo and loud sound with staccato articulation and clear phrasing (i.e. changes in tempo and sound level organized in accelerando/crescendo and rallentando/decrescendo couples for the communication of a musical phrase). Moreover, happiness is said to best be expressed with major mode, high pitch and high tempo, flowing rhythm and simple harmony [5,20]. For the joyful sonification, we used a rectangular (pulse) oscillator with a rather short envelope duration, mapping magnitude of the input signal to pitches in a C major scale, with an Inter-Onset-Interval (IOI) of 180 ms between notes (about 5.5 notes/second).

3.5.3 Continuous Sonification

The continuous sonifications were, as the name suggests, not characterized by any particular rhythm. Similar to the rhythmic case, the continuous sonification of *frustration* made use sample-based (granular) synthesis, but without a rhythmic component and added distortion. Grains were also generated based on peak amplitude. The sonification model for *relaxation* was similar to the one described for the rhythmic relaxed condition, with the difference that the sound was continuous, and that the resonant band-pass filter was set to a variable range of 300-500 Hz and the Q value to 0.2-1.0. Scaling between absolute magnitude was also slightly different. For the *joyful* sonification, a simple FM synthesizer was used to generate a tonic and major third in a C major scale, with increasing pitch depending on magnitude of the original sound file.

3.5.4 Procedure

After completing Experiment 1, participants proceeded with Experiment 2. They were presented with sound stimuli that they could listen to as many times as they wanted. Participants were then asked to rate perceived emotions 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. These are the same scales that were used in our previous study [15]. The reason why the sad scale was included, despite the fact that no "sad" stimuli was used, was to obtain results

[6] In the context of this paper we refer to these sonic events as "grains", even if they are exceeding the 1-100 ms duration commonly used in granular synthesis

comparable to those presented in [15]. The participants were given the following instructions: *"This sound represents an emotional reaction. Rate how much of the following emotions you perceive in the sound."* The presentation order of the stimuli was randomized for each participant.

3.5.5 Analysis

Since the data was collected on scales that displayed numeric values of equal distance, we proceeded with statistical analysis using parametric methods. With 30 observations per category, data could be assumed to be normally distributed according to the Central Limit Theorem. For the purpose of this study, we performed analysis of ratings within each stimulus category (frustrated, relaxed and joyful), through separate Two-Way Repeated Measures ANOVAs with the following within-subjects factors: emotional scale (frustrated, joyful, relaxed and sad) and condition, i.e. sound model (original sound, rhythmic sonification, continuous sonification, original + rhythmic sonification and original + continuous sonification). The purpose of these tests was to investigate if there was an interaction effect between emotional scale and condition (sound model). Greenhouse-Geisser estimates of spehericity were used when the assumptions of sphericity were not met. When the omnibus test for the interaction was significant, it was followed by the application of a post hoc procedure to explore which pairs of cell means that were significantly different, i.e. which condition that resulted in significantly different ratings for respective scale compared to the original sound. Paired t-tests with Bonferroni corrections were used to account for multiple comparisons.

4 Results

4.1 Experiment 1

4.1.1 Frustrated

A total of 53 different phrases were identified for the frustrated sound. A summary of the analysis, in which both authors divided words into categories based on the circumplex model of affect, is presented in Table 2. For arousal, no larger difference between the number of activation versus deactivation terms could be observed. Regarding valence, there was a slight tendency towards more negative terms than positive ones. The most frequently used words for describing the frustrated sound were *sadness* (3) and *stress* (3). However, a range of positive words were also identified.

Table 2 Keywords used for describing frustrated sounds categorized by valence (positive versus negative) and arousal (D = deactivation A = activation, I = in between)

D	A	I
Positive		
Careful	Surprise	None (2)
Relaxed	Playfulness	Not much
Unconcerned	Less stressful	
	Energy to relaxed	
Negative		
Sad(ness) (3)	Stress(-ed/ful) (3)	Confusion
Lowkey	In a hurry	
Hopelessness	Rushing	
Surrender	Angry	
Dejection	Nervousness	
Melancholy	Backing off	

Numbers in brackets are provided when more than one participant used a specific term

4.1.2 Relaxed

A total of 57 different terms were observed for the relaxed sound. A summary of the analysis is presented in Table 3. Interestingly, there was a clear tendency towards more active words. For valence, most terms were positive. However, negative terms such as depression, hopelessness and loneliness were also used, and the most commonly used word overall was *stress* (3). This was followed by the words *progress* (2) and *calm* (2).

4.1.3 Joyful

A total of 52 different phrases were observed for the first joyful sound, i.e. the sound produced by the most joyful gesture. A summary of the analysis is presented in Table 4. For arousal, there were more terms in the activation category, compared to the deactivation one. For valence, no larger difference could be observed between positive and negative category counts, although there was a tendency towards more terms for the negative category. The most frequently used words were *stress* (4), *annoyance* (2), *frustration* (2), *happy* (2) and *playful* (2). In general, many high-energy words with both positive and negative valence were identified.

In total, 54 different words were observed for the second joyful clip, which was produced by the less joyful gesture. A summary of the analysis is presented in Table 5. Similarly to the first joyful stimulus, there were more activation than deactivation terms. However, this stimulus had a clearer tendency towards negative associations, with few positive phrases; the most commonly used terms were *stress* (4), *anger* (3), *frustration* (2) and *annoyed* (2).

Table 3 Keywords used for describing relaxed sounds categorized by valence (positive versus negative) and arousal (D = deactivation, A = activation, I = in between)

D	A	I
Positive		
Calm (2)	(Happy) progress (2)	None
Relaxed	Energetic	Common emotions
	Movement	More positive
	Action	Exploration
	Applied force	Neither happy nor sad
	Anticipation	
	An opening	
	Playfulness	
	Arrival	
	Happiness	
	Gleeful	
	Getting going again	
	Moving forward	
Negative		
Depression	Stress(-ed/ful) (3)	I don't know
Hopelessness	Yelling	
Loneliness	Friction	
	Combative/heated	

Numbers in brackets are provided when more than one participant used a specific term

Table 4 Keywords used for describing joyful sounds (stimulus 1) categorized by valence (positive versus negative) and arousal (D = deactivation, A = activation, I = in between)

D	A	I
Positive		
Calm	Happy (2)	Likeable
Peaceful	Playing/playful (2)	Opposite of other stimuli
	Dynamic	
	Determined	
	Joyful	
	Starting	
	Excited	
	Need to speed up	
	Like a child exploring	
Negative		
Annoyance (2)	Stress(ful) (4)	Flimsy
Sadness	Annoy(-ance/ing) (2)	
	Frustration (2)	
	Disturbance	
	Not working correctly	
	Aggressive	
	Angry	
	Strong objection	
	Agony	
	Anxiety	
	Irritation	

Numbers in brackets are provided when more than one participant used a specific term

Table 5 Keywords used for describing joyful sounds (stimulus 2) categorized by valence (positive versus negative) and arousal (D = deactivation, A = activation, I = in between)

D	A	I
Positive		
Peaceful	Playful	
	Happier	
	Curious	
Negative		
Annoy (-ed/ing) (2)	Stress(-ful) (4)	Ambivalence
Sad	Anger/angr(-y/iness) (3)	Similar to other stimuli
Tired	Frustrat(-ed/ion) (2)	
Passive aggressive	Forcing a robot to work	
	Violence	
	Aggressiveness	
	Rage	
	Being dragged	

Numbers in brackets are provided when more than one participant used a specific term

4.2 Experiment 2

4.2.1 Frustrated

Plots displaying ratings for the frustrated stimuli are displayed in Fig. 5. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of emotional scale, $\chi^2(5) = 16.12$, $p < 0.01$, and the interaction term between emotional scale and sound model, $\chi^2(77) = 169.56$, $p < 0.001$. Degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity ($\epsilon = 0.71$ for the main effect of emotional scale and 0.450 for the interaction effect). All effects are reported as significant at $p < 0.05$; there was a significant main effect of emotion scale, $F(2.12, 61.54) = 32.84$, and condition, $F(4, 116) = 3.89$, on ratings. There was also a significant interaction effect between emotional scale and condition, $F(5.4, 156.59) = 10.16$, $p < 0.001$. This indicates that the sound model had different effects on people's ratings depending on the type of emotional scale that was used. An interaction graph for the frustrated stimuli is depicted in Fig. 6.

To break down the significant interaction, post hoc tests were performed. We were interested in whether condition, i.e. choice of sound model, significantly increased frustrated ratings and if there was a significant difference between the mix of the original sound and the sonification, versus the sonification only. All sound models significantly increased frustrated ratings compared to the original sound ($p < 0.001$). There was a significant increase in ratings of frustration for the continuous sonification ($M = 2.27, SD = 1.31, t(29) = -6.29$), the mix of the continuous sonification and the original sound ($M = 2.13, SD = 1.20, t(29) = -7.24$), the rhythmic sonification ($M = 2.53, SD = 1.33, t(29) =$

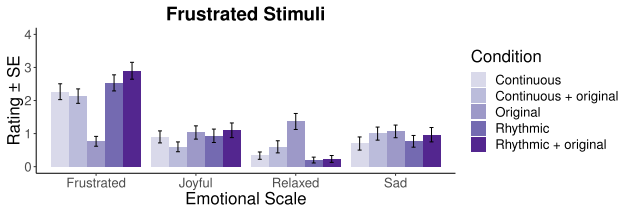

Fig. 5 Mean ratings for frustrated stimuli, with standard errors (SE) for respective sound model

−6.65), and the mix of the rhythmic sonification and the original sound ($M = 2.9, SD = 1.40, t(29) = -7.55$), compared to ratings of the original sound file ($M = 0.77, SD = 0.82$). However, there was no significant difference in frustrated ratings between the mix of sonifications and the original sound (i.e. original sound file + continuous/rhythmic sonification), versus the sonification only. For the relaxed scale, all ratings were significantly lower than for the original sound file ($p < 0.01$).

4.2.2 Relaxed

Plots displaying ratings for the relaxed stimuli can be seen in Fig. 7. Analysis revealed that there was a significant main effect of emotional scale and condition as well as a significant interaction between the two variables. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of emotional scale, $\chi^2(5) = 23.37$, $p < 0.001$, and the interaction term between emotional scale and condition, $\chi^2(77) = 157.50$, $p < 0.001$. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.69$ for the main effect of emotional scale, and 0.51 for the interaction effect). All effects are reported as significant at $p < 0.05$; there was a significant main effect of emotional scale, $F(2.07, 59.99) = 25.57$, and condition, $F(4, 116) = 2.90$, on ratings. There was also a significant interaction effect between emotional scale and condition, $F(6.10, 176.99) = 8.46$, $p < 0.001$. The interaction graph for the relaxed clip is depicted in Fig. 8.

Post hoc tests revealed three significant differences between sound models for the relaxed scale. There was a significant lowering in ratings of relaxation for the mix of the continuous sonification and the original sound ($M = 0.40$, $SD = 0.62$) and the mix of the rhythmic sonification and original sound ($M = 0.43$, $SD = 0.77$), compared to the original sound ($M = 1.1$, $SD = 1.37$, $t(29) = 2.91$, $p = 0.04$). Moreover, there was a significant difference between the mix of the continuous sonification and the original sound ($M = 0.40$, $SD = 0.62$) and the continuous sonification only $M = 1.13$, $SD = 1.22$, $t(29) = -3.5204$, $p = 0.009$). There was no significant difference in relaxed ratings between the mix of sonifications and the original sound, versus the sonification only. Interestingly, all sound models were rated as significantly more sad than the original sound ($p < 0.01$) and all models apart from the continuous sonification were rated as significantly more frustrated than the original ($p < 0.001$).

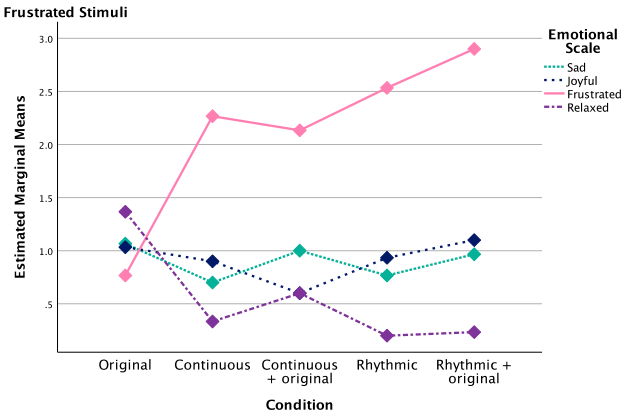

Fig. 6 Interaction graph for frustrated stimuli. Type of emotional scale is represented by the four lines with different colors. Different sound models are displayed as different points on the x-axis

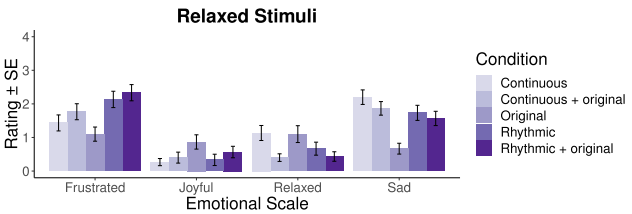

Fig. 7 Mean ratings for relaxed stimuli, with standard errors (SE) for respective sound model

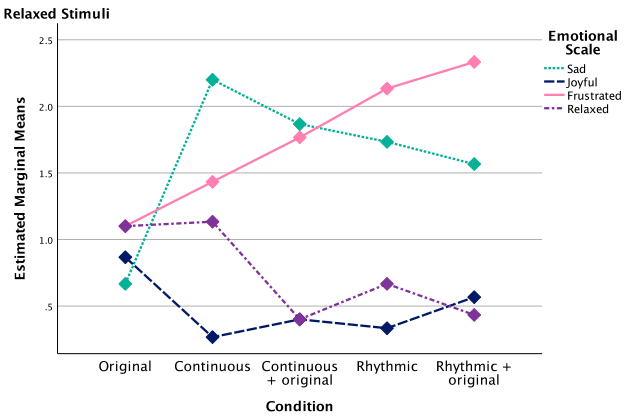

Fig. 8 Interaction graph for relaxed stimuli. Type of emotional scale is represented by the four lines with different colors. Different sound models are displayed as different points on the x-axis

4.2.3 Joyful

Plots displaying ratings for the joyful stimuli can be seen in Fig. 9. Mauchly's test indicated that the assumption of sphericity had been violated for the main effects of emotional scale $\chi^2(5) = 21.772$, $p = 0.001$ and the

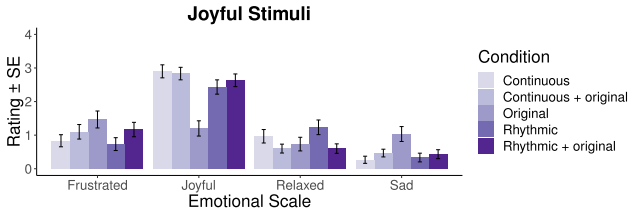

Fig. 9 Mean ratings for the joyful stimuli, with standard errors (SE) for respective sound model

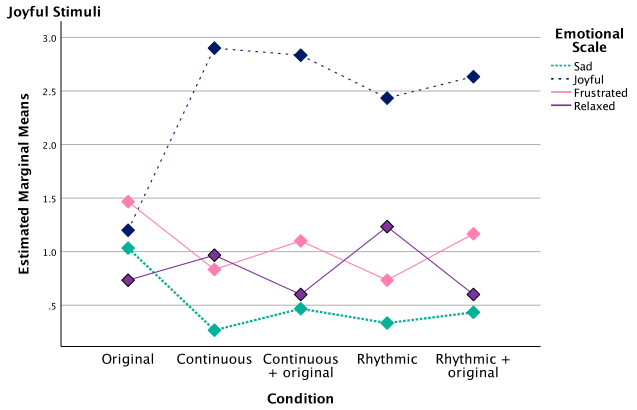

Fig. 10 Interaction graph for joyful stimuli. Type of emotional scale is represented by the four lines with different colors. Different sound models are displayed as different points on the x-axis

interaction term between emotional scale and condition, $\chi^2(77) = 135.639$, $p < 0.001$. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.665$ for the main effects of emotional scale, and 0.549 for the interaction effect). There was a significant main effect of emotion scale on ratings, $F(2.00, 57.85) = 49.09$. There was also a significant interaction effect between the type of emotional scale and type of soundmodel used, $F(6.586, 191.00) = 8.99$, $p < 0.001$. An interaction graph for the joyful stimuli is depicted in Fig. 10.

Post hoc tests revealed that all sound models significantly increased joyful ratings compared to ratings of the original sound file ($p <= 0.001$); there was a significant increase in ratings for the continuous sonification ($M = 2.90, SD = 1.06, t(29) = -8.10$), the mix of the continuous sonification and the original sound ($M = 2.83, SD = 1.02, t(29) = -7.53$), the rhythmic sonification ($M = 2.43, SD = 1.17, t(29) = -4.31$), and the mix of the rhythmic sonification together with the original sound ($M = 2.63, SD = 1.03, t(29) = -6.14$), compared to ratings of the original sound file ($M = 1.20, SD = 1.24$). For the sad scale, only the continuous sonification had significantly lower ratings than the original sound ($t(29) = 3.10$, $p < 0.02$). No significant difference between the combination of the original sound and sonifications versus sonification only was observed for the joyful scale.

5 Discussion

5.1 Experiment 1

The purpose of this experiment was to build upon our previous work in which participants rated the same sound files on a set of quantitative scales [15]. By allowing participants to describe these sounds in free-form text in a new experiment, we aimed to get a better understanding of the emotional content of consequential sounds generated by expressive robot movements. Analysis of free-form text answers suggested that participants had both positive and negative associations to these sounds; the stimuli were different in terms of activity/arousal (activation versus deactivation) as well as valence (positive/negative) dimensions. A word that reoccurred for all sounds and was mentioned by several participants was *stress*. This is not surprising since the sound generated by the mechanical robot movements is a form of broadband noise, and noise has been found to elicit both emotional and physiological stress responses [53], and unpleasantness [13]. We present the results of the free-form text experiment for respective sound stimuli (frustrated, relaxed and joyful) below.

5.1.1 Frustrated

For the frustrated stimulus, there were slightly more negative terms than positive ones. No larger differences were observed between the words categorized in the activation versus deactivation category. Interestingly, this sound appears to have conveyed stress and sadness rather than frustration.

5.1.2 Relaxed

One of the findings from our previous study was that the sound of a relaxed gesture was not necessarily perceived as relaxed by listeners, and that the sound alone was not enough to convey a sensation of relaxation. Interestingly, the current study shows that both positive and negative words were used to describe the relaxed sound. However, there was no clear thematic tendency towards words related to relaxation. An interesting aspect of the results presented in Table 3 was that there were many terms in the activation category. Usually, a relaxed emotion should be placed in the deactivation part of the two-dimensional circumplex model of affect.

5.1.3 Joyful

For the joyful stimuli, it is clear that more positive terms were identified for the most joyful gesture (stimulus 1), compared to the slightly less joyful one (stimulus 2). However, what stands out for both of these stimuli is that there is a considerable number of terms in the activation column that are characterized by negative valence. This further supports our

observations from previous work [15]; the sounds produced by joyful gestures appear to also convey emotions related to frustration and anger.

5.2 Experiment 2

Based on the quantitative ratings, we can conclude that blended sonification successfully contributed to enhance the communication of intended emotions conveyed by the sounds generated by frustrated and joyful robot movements. We discuss these findings in light of previous research on emotions in speech and music in the sections below.

5.2.1 Frustrated

As demonstrated in Fig. 5, and also confirmed in our previous work, the sound produced by a frustrated gesture alone did not communicate very much frustration. Interestingly, all sonification models proposed in this work significantly increased frustrated ratings, suggesting that the blended sonification techniques successfully contributed to conveying frustration more clearly. Highest mean rating was observed for the rhythmic sonification presented together with the original sound. The design of this sound model was guided by previous research on emotions with negative valence and high arousal in speech and music [5,25]. The frustrated property was successfully represented using fast and loud sounds with large intensity variability and irregular rhythms, as well as distortion.

5.2.2 Relaxed

Similarly to the sound produced by the frustrated gesture, the sound generated by a relaxed movement did not convey relaxation to any considerable extent, see Fig. 7. However, as opposed to the sonification models proposed for frustration, the models used in blended sonification of relaxation were not successful in terms of enhancing a relaxed sensation. A significant decrease in ratings of relaxation was observed for both sonification models when presented together with the original relaxed sound, suggesting that the sounds do not blend well together, thus contributing to a disunite message.

The proposed sound design for relaxation was unfortunately not successful in terms of clarifying the emotional expression. It was difficult to modify the mechanical robot sound in order to communicate relaxation, without making the audio sound more sad or frustrated. Interestingly, sonifications significantly increased ratings of sadness and frustration compared to the original sound (apart from the continuous sonification for frustration). Nevertheless, the fact that the blended sonifications (presented both together with the original sound and alone) were classified as more sad is somewhat expected considering that both listeners and automated systems often have difficulty distinguishing between low-arousal categories such as "calm" and "sad" [21], and that listeners have a tendency to mutually confuse sadness with tenderness in the classification of expressive music [48].

5.2.3 Joyful

In our previous work, we observed that consequential sounds produced by joyful movements were not necessarily perceived as joyful. As demonstrated in Experiment 2, the joyful recording was also often described using high activation words with negative valence, such as stress, anger and frustration. In the current work, we have shown that this effect can be counteracted through blended sonification. As results presented in Fig. 9 suggest, all proposed sonification models significantly increased joyful ratings compared to the original sound. A tendency towards higher mean ratings for the continuous model could be observed. This particular sound model was designed based on previous work presented in [5,20] suggesting that happiness in music is best expressed with a major mode, high pitch, flowing rhythm and simple harmony. For this purpose, a simple FM synthesizer was used to generate a tonic and major third in a major scale, with increasing pitch depending on magnitude of the input signal.

5.3 Methodological Concerns

This study was performed in a controlled experiment setting in which participants judged the emotional content in sounds as presented in a web interface, without a visual robot representation. The conclusions presented in this paper should thus be viewed in the light of the current setting; i.e. that participants were not actually in the room together with the robot when listening to the sounds. Additional research on how these sonic aspects are perceived in an interactive context with an actual robot should be explored in future work.

It should be noted that the results presented in this work are based on audio-only evaluations of sounds produced by robot movement and blended sonifications thereof. To fully understand the impact of consequential robot sounds on non-verbal communication, studies involving multimodal stimuli and audiovisual conditions should of course also be studied. However, since the findings presented in the current study are largely supported by the audiovisual evaluations of the same stimuli presented in [15], we believe that the results reported in this paper are reliable.

Moreover, the qualitative analysis of Experiment 1 was based on coding performed by the two authors of this paper. Of course, it would have been better if this coding procedure had included a larger group of independent researchers. However, since both coders are experts in music and emotion research, we judge the terms presented in Tables 2, 3, 4 and 5 to be an accurate representation of unambiguous terms used

by participants to describe respective sounds, rather than a measure of coding consistency between the two researchers.

6 Conclusions

In this paper we have presented two experiments focusing on perception of consequential sounds produced by expressive robot movements. In the first experiment, 31 participants evaluated emotional content of robot sounds in free-form text. In the second experiment, blended sonification was used in an attempt to improve clarity of the emotional message conveyed through the consequential robot sounds evaluated in the first experiment. The sonifications were evaluated by 30 participants through quantitative ratings along a set of emotional scales.

Results obtained from both experiments suggest that there was no strong coupling between the emotional expression of the gestures producing the original sounds of the robot and how they were perceived. For example, sounds produced by a frustrated gesture were described using a combination of positive and negative words and not rated as being very frustrated. Sounds produced by a relaxed gesture included many positive active words and were not rated as relaxed at all. Finally, sounds produced by joyful gestures included more negative than positive words, and were not rated as joyful. Interestingly, the word *stress* reoccurred for all sounds produced by expressive robot movements.

Blended sonification, i.e. the sonification of robot movements combined with the original sound of the robot, helped the participants to recognize the intended emotional expressions. More specifically, analysis of quantitative ratings revealed that blended sonification guided by previous research on emotion recognition in speech and music could clarify emotional messages such as frustration and joy. In other words, blended sonification techniques can counteract the ambiguity of the mechanical sounds produced by the NAO robot. In the current study, consequential robot sounds were successfully blended with sonifications characterized by fast and loud sounds, intensity variability, irregular rhythms and distortion, to enhance the communication of frustration. In contrast, for the communication of joy, consequential robot sounds were successfully enhanced with blended sonifications based on simple harmonies in a major scale with increasing pitch depending on magnitude of the input signal. We also identified challenges when it comes to expressing less active emotions such as relaxation using a blended sonification technique. This low arousal and positive valence emotion was particularly difficult to portray using blended sonification in the presence of robot motor sounds.

We can conclude that blended sonification models guided by perceptual research on emotion in music and speech successfully can be used to improve communication of emotions in auditory-only conditions. It should be noted that the communication of emotional expressions through sounds can be further enhanced and made clearer to participants if presented together with the robot movements (both live or in a video), as it has been shown in a previous study [55]. The findings on blended sonification presented in our work can serve as guidelines for future work on sonification of expressive robot gestures.

Acknowledgements The authors would like to thank all participants for their essential contributions to the study.

Funding Open access funding provided by Royal Institute of Technology. This study was funded by KTH Royal Institute of Technology, Stockholm, Sweden, and the Swedish Research Council (Grant 2017-03979).

Availability of Data and Materials The data from the two experiments as well as sound files of robot sounds, and blended sonification thereof, are available as supplementary material.

Code availability The code used for sound synthesis is available upon request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Alexanderson S, Osullivan C, Neff M, Beskow J (2017) Mimebot—investigating the expressibility of non-verbal communication across agent embodiments. ACM Trans Appl Percept (TAP) 14(4):1–13
2. Bellona J, Bai L, Dahl L, LaViers A (2017) Empirically informed sound synthesis application for enhancing the perception of expressive robotic movement. In: Proceedings of the 23rd International Conference on Auditory Display (ICAD). Georgia Institute of Technology
3. Besson M, Schön D, Moreno S, Santos A, Magne C (2007) Influence of musical expertise and musical training on pitch processing in music and language. Restor Neurol Neurosci 25(3–4):399–410
4. Bresin R, Friberg A (2000) Emotional coloring of computer-controlled music performances. Comput Music J 24(4):44–63

5. Bresin R, Friberg A (2011) Emotion rendering in music: range and characteristic values of seven musical variables. Cortex 47(9):1068–1081
6. Bresin R, Hermann T, Hunt A (2012) Interactive sonification. J Multimodal User Interfaces 5(3–4):85–86
7. Bresin R, de Witt A, Papetti S, Civolani M, Fontana F (2010) Expressive sonification of footstep sounds. In: Proceedings of ISon 2010: 3rd Interactive Sonification Workshop, pp 51–54
8. Burtt B (2001) Galactic Phrase Book & Travel Guide: Beeps, Bleats, Boskas, and Other Common Intergalactic Verbiage (Star Wars), Chap. Part II-Behind the Sounds. Del Rey
9. Caramiaux B, Bevilacqua F, Bianco T, Schnell N, Houix O, Susini P (2014) The role of sound source perception in gestural sound description. ACM Trans Appl Percept (TAP) 11(1):1–19
10. Dahl L, Bellona J, Bai L, LaViers A (2017) Data-driven design of sound for enhancing the perception of expressive robotic movement. In: Proceedings of the 4th International Conference on Movement Computing. ACM, p 16
11. Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. Psychol Music 39(1):18–49
12. Elowsson A, Friberg A (2015) Modeling the perception of tempo. J Acoust Soc Am 137(6):3163–3177
13. Fenko A, Schifferstein HN, Hekkert P (2011) Noisy products: does appearance matter? Int J Des 5(3):77–87
14. Franinović K, Serafin S (2013) Sonic Interaction Design. MIT Press
15. Frid E, Alexanderson S, Bresin R (2018) Perception of mechanical sounds inherent to expressive gestures of a NAO robot—implications for movement sonification of humanoids. In: Proceedings of the Sound and Music Computing Conference (SMC) 2018. pp 53–59
16. Gabrielsson A, Lindström E (2011) The role of structure in the musical expression of emotions. In: Handbook of Music and Emotion: Theory, Research, Applications. Oxford University Press, p 371
17. Gobl C, Chasaide AN (2000) Testing affective correlates of voice quality through analysis and resynthesis. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion
18. Godøy RI, Haga E, Jensenius AR (2005) Playing "air instruments": mimicry of sound-producing gestures by novices and experts. In: International Gesture Workshop. Springer, pp 256–267
19. Godøy RI, Leman M (2010) Musical gestures: sound, movement, and meaning. Routledge
20. Hevner K (1937) The affective value of pitch and tempo in music. Am J Psychol 49(4):621–630
21. Hong Y, Chau CJ, Horner A (2017) A study of what makes calm and sad music so difficult to distinguish in music emotion recognition. In: Proceedings of the International Computer Music Conference (ICMC). Michigan Publishing, University of Michigan Library, Ann Arbor
22. Hunt A, Hermann T (2011) Interactive sonification. In: T. Hermann, A. Hunt, J.G. Neuhoff (eds.) The Sonification Handbook, Chap. 11. Logos Verlag, Berlin, pp 273–298
23. Inoue K, Wada K, Ito Y (2008) 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, pp 1321–1324
24. Jee ES, Jeong YJ, Kim CH, Kobayashi H (2010) Sound design for emotion and intention expression of socially interactive robots. Intell Serv Robot 3(3):199–206
25. Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull 129(5):770
26. Kramer G, Walker B, Bonebright T, Cook P, Flowers J, Miner N, Neuhoff J, Bargar R, Barrass S, Berger J et al (1999) The sonification report: status of the field and research agenda. Report prepared for the national science foundation by members of the international community for auditory display. International Community for Auditory Display (ICAD), Santa Fe
27. Langeveld L, van Egmond R, Jansen R, Özcan E (2013) Product sound design: intentional and consequential sounds. In: Advances in industrial design engineering. InTech
28. Latupeirissa AB, Frid E, Bresin R (2019) Sonic characteristics of robots in films. In: Proceedings of the Sound and Music Computing Conference (SMC)
29. Löffler D, Schmidt N, Tscharn R (2018) Multimodal expression of artificial emotion in social robots using color, motion and sound. In: Proceedings of the 2018 ACM/IEEE International Conference on Human–Robot Interaction, pp 334–343
30. Mavridis N (2015) A review of verbal and non-verbal human-robot interactive communication. Robot Auton Syst 63:22–35. https://doi.org/10.1016/j.robot.2014.09.031
31. Monceaux J, Becker J, Boudier C, Mazel A (2009) First steps in emotional expression of the humanoid robot NAO. In: Proceedings of the 2009 International Conference on Multimodal Interfaces. ACM, pp 235–236
32. Moore D, Dahl T, Varela P, Ju W, Næs T, Berget I (2019) Unintended consonances: methods to understand robot motor sound perception. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–12
33. Moore D, Ju W (2018) Sound as implicit influence on Human-Robot Interactions. In: Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction, pp 311–312
34. Moore D, Tennent H, Martelaro N, Ju W (2017) Making noise intentional: A study of servo sound perception. In: Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction. ACM, pp 12–21
35. Nakadai K, Okuno HG, Kitano H (2002) Real-time sound source localization and separation for robot audition. In: 7th International Conference on Spoken Language Processing
36. Pauletto S (2019) Invisible seams: the role of Foley and voice post-production recordings in the design of cinematic performances. In: Filimowicz M (ed) Foundations in Sound Design for Linear Media: a Multidisciplinary Approach. Routledge
37. Pietila G, Lim TC (2012) Intelligent systems approaches to product sound quality evaluations—a review. Appl Acoust 73(10):987–1002
38. Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17(3):715–734
39. Read R, Belpaeme T (2012) How to use non-linguistic utterances to convey emotion in child-robot interaction. In: Proceedings of the 2012 ACM/IEEE International Conference on Human–Robot Interaction. ACM, pp 219–220
40. Robinson FA (2020) Audio cells: a spatial audio prototyping environment for Human-Robot Interaction. In: Proceedings of the 14th International Conference on Tangible, Embedded, and Embodied Interactions, pp 955–960
41. Rossi S, Dell'Aquila E, Bucci B (2019) Evaluating the emotional valence of affective sounds for child-robot interaction. In: International Conference on Social Robotics. Springer, pp 505–514
42. Russell JA (1980) A circumplex model of affect. J Personal Soc Psychol 39(6):1161
43. Schinkel-Bielefeld N, Lotze N, Nagel F (2013) Audio quality evaluation by experienced and inexperienced listeners. J Acoust Soc Am 133(5):3246. https://doi.org/10.1121/1.4805210
44. Seck HH (2015) Marine corps shelves futuristic robo-mule due to noise concerns. https://www.military.com/dailynews/2015/12/22/marine-corps-shelvesfuturistic-robo-mule-due-to-noiseconcerns.html

45. Spence C, Zampini M (2006) Auditory contributions to multisensory product perception. Acta Acust United Acust 92(6):1009–1025
46. Spence R (2006) Information visualization: design for interaction. Prentice Hall
47. Strait DL, Kraus N (2014) Biological impact of auditory expertise across the life span: musicians as a model of auditory learning. Hear Res 308:109–121
48. Taruffi L, Allen R, Downing J, Heaton P (2017) Individual differences in music-perceived emotions: the influence of externally oriented thinking. Music Percept: Interdiscipl J 34(3):253–266
49. Tennent H, Moore D, Jung M, Ju W (2017) Good vibrations: how consequential sounds affect perception of robotic arms. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp 928–935
50. Trovato G, Paredes R, Balvin J, Cuellar F, Thomsen NB, Bech S, Tan ZH (2018) The sound or silence: investigating the influence of robot noise on proxemics. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, pp 713–718
51. Tünnermann R, Hammerschmidt J, Hermann T (2013) Blended sonification: sonification for casual interaction. In: Proceedings of the of the 19th International Conference on Auditory Display (ICAD)
52. Turchet L, Bresin R (2015) Effects of interactive sonification on emotionally expressive walking styles. IEEE Trans Affect Comput 6(2):152–164
53. Westman JC, Walters JR (1981) Noise and stress: a comprehensive approach. Environ Health Perspect 41:291–309
54. Yilmazyildiz S, Read R, Belpeame T, Verhelst W (2016) Review of semantic-free utterances in social Human–Robot Interaction. Int J Hum-Comput Interact 32(1):63–85
55. Zahray L, Savery R, Syrkett L, Weinberg G (2020) Robot gesture sonification to enhance awareness of robot status and enjoyment of interaction. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp 978–985. https://doi.org/10.1109/RO-MAN47096.2020.9223452
56. Zhang R, Jeon M, Park CH, Howard A (2015) Robotic sonification for promoting emotional and social interactions of children with ASD. In: Proceedings of the 10th annual ACM/IEEE International Conference on Human–Robot Interaction Extended Abstracts. ACM, pp 111–112

Emma Frid is a postdoctoral researcher at Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Paris, France and KTH Royal Institute of Technology, Stockholm, Sweden, where she works on sound and music interfaces designed to promote health and inclusion. Her research interests focus predominantly on Accessible Digital Musical Instruments (ADMIs). She holds a PhD in Sound and Music Computing from KTH Royal Institute of Technology, Stockholm, Sweden.

Roberto Bresin is Professor in Media Technology at KTH Royal Institute of Technology, Stockholm, Sweden. Bresin is the head of the KTH Sound and Music Computing group. His expertise is in design and analysis of expressive control of sound models used in music performance, sound in interaction, interactive sonification in performing arts, sports and robotics.