Sounding Robots: Design and Evaluation of Auditory Displays for Unintentional Human-robot Interaction

Sounding Robots: Design and Evaluation of Auditory Displays for Unintentional Human-robot Interaction

BASTIAN ORTHMANN, IOLANDA LEITE, and ROBERTO BRESIN, KTH Royal Institute of Technology, Sweden
ILARIA TORRE, Chalmers University of Technology, Sweden and KTH Royal Institute of Technology, Sweden

Non-verbal communication is important in HRI, particularly when humans and robots do not need to actively engage in a task together, but rather they co-exist in a shared space. Robots might still need to communicate states such as urgency or availability, and where they intend to go, to avoid collisions and disruptions. Sounds could be used to communicate such states and intentions in an intuitive and non-disruptive way. Here, we propose a multi-layer classification system for displaying various robot information simultaneously via sound. We first conceptualise which robot features could be displayed (robot size, speed, availability for interaction, urgency, and directionality); we then map them to a set of audio parameters. The designed sounds were then evaluated in five online studies, where people listened to the sounds and were asked to identify the associated robot features. The sounds were generally understood as intended by participants, especially when they were evaluated one feature at a time, and partially when they were evaluated two features simultaneously. The results of these evaluations suggest that sounds can be successfully used to communicate robot states and intended actions implicitly and intuitively.

CCS Concepts: • **Human-centered computing** → **Auditory feedback**; *Empirical studies in HCI*; *Sound-based input/output*; • **Computer systems organization** → *External interfaces for robotics;*

Additional Key Words and Phrases: Sonification, unintentional Human-Robot Interaction, Non-verbal communication, Auditory Display, Design Evaluation

ACM Reference format:
Bastian Orthmann, Iolanda Leite, Roberto Bresin, and Ilaria Torre. 2023. Sounding Robots: Design and Evaluation of Auditory Displays for Unintentional Human-robot Interaction. *ACM Trans. Hum.-Robot Interact.* 12, 4, Article 49 (December 2023), 26 pages.
https://doi.org/10.1145/3611655

This work was partially funded by grants from the Swedish Research Council (2017-05189), the Swedish Foundation for Strategic Research (SSF FFL18-0199), the S-FACTOR project from NordForsk, the Digital Futures research Center, the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.
Authors' addresses: B. Orthmann, I. Leite, and R. Bresin, KTH Royal Institute of Technology, Stockholm, Sweden; emails: {orthmann, iolanda, roberto}@kth.se; I. Torre, Chalmers University of Technology, Gothenburg, Sweden and KTH Royal Institute of Technology, Stockholm, Sweden; email: ilariat@kchalmers.se.

2573-9522/2023/12-ART49 $15.00
https://doi.org/10.1145/3611655

1 INTRODUCTION

Much research in **Human-Robot Interaction (HRI)** is dedicated to studying intentional interactions between humans and robots, that is to say, situations where the two agents need to actively perform a task together. Situations such as a person interacting with a receptionist robot or children learning from a robot tutor or a team comprised by humans and robots in Search and Rescue operations, all involve a task that needs to be carried out together. However, when it comes to everyday interactions with robots, arguably, most people in the near future will only interact with a robot "unintentionally," i.e., they will have to share a common space together, for example, in the case of cleaning or delivery robots. Communication plays an important role in both types of situations: While intentional interactions might need some kind of language to work, unintentional interaction does not necessarily need explicit back-and-forth communication. If we take the example of pedestrians walking in a crowded space, they (mostly) succeed in reaching their destination without crashing into each other and without otherwise disrupting the crowd's flow. This even works when people are looking at their phone! To achieve this, people do not use explicit language to communicate things such as their intended direction or whether they are not in a hurry and will move away from other people's path. Instead, they "read" the environment using a variety of multimodal signals that are intuitively and implicitly understood, such as walking speed, facial expressions, and sound. Sound is particularly interesting, because it can reach us from farther away than nuanced visual information such as facial expressions, and thus it can potentially convey information earlier. Additionally, our brain is always "listening," our ears do not have lids, and sound reaches us even when we have our eyes closed (or if we are visually impaired). With sound, we can tell if someone walking behind is walking faster than us and wants to overtake us; we are also constantly paying attention to the sound of technology: We use the sound of cars, traffic lights (accessible pedestrian systems), coffee machines and other technology. Therefore, as displayed in Figure 1, sound could be used to convey information about a robot in cases where explicit communication might be disruptive and unnecessary. Additionally, sounds are made of combinations of audio parameters that contribute to creating the final sound that we hear (e.g., loudness, pitch, rhythm, timbre, brightness). The focus of our investigation is whether these parameters can each be mapped to an individual robot-related information and whether this can be recognised as such, even in combination with other parameters/information.

1.1 Contribution

In this article, we present a comprehensive overview of the process of sonification applied to human-robot interaction. Following best practices from Design Thinking and Participatory Design, we first identify iconic features of robots that could benefit from sonification, via a focus group with experts in robotics. Then, we conceptualise the results of this focus group in a novel framework for displaying a robot's identity and internal state via sound, whereby one sound contains several "layers" of information (Figure 2). We then apply this framework to our scenario of interest—unintentional human-robot interaction, i.e., a situation where humans and robots have to navigate and co-exist within a shared space. We select five robot features that we transform into sound, and we give a detailed description of why we designed each sound in a certain way. Finally, we use rigorous quantitative methods to evaluate the individual sonifications as well as combinations of them.

To the best of our knowledge, this is the first work to describe the conceptualisation, design, and evaluation of sounds for robots in detail. Furthermore, we propose and evaluate the novel idea that one sound could convey several different pieces of information at the same time. We hope that other HRI researchers can benefit from our work by:

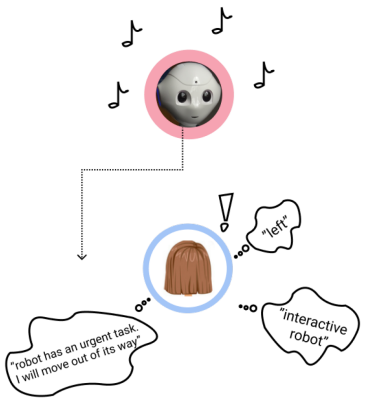

Fig. 1. Conceptualisation of the usage of auditory displays for unintentional human-robot interactions.

— applying our layer-based framework to other HRI scenarios;
— applying our sound design process to the design of new sounds that can be used to convey information from robots;
— taking inspiration from the sounds we designed (which we make available to the community) to augment existing human-robot interactions with the addition of a sound modality.

The article is structured as follows: First, we situate our work with respect to previous literature in Section 2; then, we describe our proposed layer-based framework in Section 3 and the sound designs in Section 4; we present several evaluation studies and their results in Section 5; finally, we discuss the relevance of our work and provide design recommendations in Sections 6–7.

2 RELATED WORK

2.1 Robot Sounds

The sounds that robots make have received limited attention in the past. Here, it is worth distinguishing between robot *voices*—by which we mean the medium with which robots communicate using spoken language, akin to human voices—and robot *sounds*—i.e., the noises or other non-linguistic sounds that robots emit [63]. Speaking of robot voices, recently, McGinn and Torre [37] highlighted how many HRI researchers choose the voice of their speaking robotic platforms out of convenience, without considering the effect that the choice of voice might have on users' behaviour and interaction experience. However, just like human voices affect the formation of first impressions of a newly met individual [36], robot voices have been shown to influence people's first impressions and, consequently, expectations, of robots [31, 37, 56, 61]. Having a voice at all, whether machine-like or human-like, makes people assume that the robot will have spoken language capabilities [50], and natural-sounding voices in particular can produce expectations of human-likeness in other robot capabilities as well [37]. However, Natural Language Processing and spoken dialogue systems are advancing at a much slower rate than synthetic voices [9, 33, 35, 63], so we are currently at risk of deploying high-end, human-sounding voices on robots that cannot live up to the expectations that these voices afford. While some researchers claimed that robot voices should be human-like, to help understandability [22, 53], recently there have been some calls to design "appropriate" voices for robots [3, 39]. This appropriateness refers to making sure that the voices are congruent with, e.g., the robot's physical embodiment [37, 39, 40] and the context in which the robot needs to carry out its tasks [10, 57, 58]. Arguably, for tasks where spoken

verbal interaction is not necessary, voices might also not be necessary, and instead communication could be achieved with nonverbal signals, such as sounds [63].

Robotic nonverbal sounds can be divided into consequential sounds that a robot emits (e.g., motor sounds) [38, 54] and intentional sounds that fulfil a certain function (e.g., beeps and alarms) [14, 23]. A recent review of these kinds of sounds can be found in Reference [46]. Here, we will briefly mention works that pertain specifically to the communicative functions of these sounds and related user experience and perception. For example, Cha et al. investigated how intentional robot sounds affect auditory localisation in a human-robot collaborative task, and found that adding either a tonal or a broadband sound signal to the robot increased localisation performance, compared to having no added sound [11]. Further, they found that the tonal sound was the most noticeable, but also the most annoying, of the three conditions. This was expected, as tonal sounds are often used for alarms, which are not meant to be used as a continuous information stream. Trovato et al. also found that participants walked closer by a robot whose consequential sound had been turned off or masked, and they rated the noisy robots more discomforting [59]. In general, people seem to dislike servo robot motor sounds [24, 38]. Robots displaying sounds that were masking their consequential sounds were rated as warmer, more positive, and less discomforting [65]. Consequential sounds of robotic arms negatively influenced people's perceptions of the arms, although this was mediated by how "high-end" the arm looked and the task it was carrying out (functional or social) [54]. Cha et al. [11] conclude with a call for "creating intelligent auditory signaling policies that use iconic sounds to convey robot state information in an intuitive manner." Since many people only have knowledge about robots based on their portraits in media, we could take inspiration from films and TV to give robots sounds that people are already familiar with [27, 32, 62]. Thus, intentional nonverbal sounds seem promising for human-robot communication. Previous studies also showed that these sounds can be used to communicate robot trajectories [29].

A concept for synthesising sound from different output parameters of a robot can be found in Reference [48], which also compares the generated audio output to a human voice. The necessary mappings between physical dimensions and sound parameters can partly be based on findings from Dubus and Bresin [16], who analysed existing publications on the sonification of physical properties. In this work, they created a database of sonification mappings by associating physical and auditory dimensions, and they identified, e.g., the predominance of pitch compared to other auditory dimensions used for sonifications. However, to the best of our knowledge, there is no previous work trying to use sounds to communicate different information simultaneously and systematically, but rather most studies focused on evaluating the effect of a specific sound or sound characteristic (e.g., pitch, tone). In the current article, we want to address this gap in the literature by describing a novel framework for robot auditory displays, followed by its conceptual evaluation.

2.2 Auditory Display

An extensive overview of the field of Auditory Display has been given by Hermann et al. [25]. The authors describe it as the practice of utilising the human auditory system as main interface for acquiring information or, in other words, as the "process of transforming acoustic waves into meaning and response behaviour." The authors motivate the use of Auditory Display with the "complex and powerful listening system" that humans have, which makes the decoding of audio information appear effortless.

Sonification is the art and science of transforming data into acoustically perceivable information, and it belongs to the field of Auditory Display. Similar to visualization, it can be done in a broad variety of approaches but is less categorized in terms of psychological perception [60].

Hermann et al. [25] describe how data and interaction contribute to the audio output and mention the main functions of auditory displays, beside alarms, alerts and warnings, to be status and monitoring reports, data exploration, or art and entertainment. Sonification methods can differ, e.g., in how literally they translate data into audio, their learning requirements, and their flexibility. As can be seen with the design of alarm sounds, which always carry some kind of information about the severity (intensity) and surroundings (type of sounds), sonification can bring additional information to multiple people at the same time, in a subtle way.

Due to the complex nature of the auditory displays that we envision, we chose methods based on **Parameter Mapping Sonification (PMSon)**, which associates audio parameters with multivariate data and therefore might be capable of displaying a more complex data bundle [25]. Potential parameters to be included in PMSon are acoustic parameters such as frequency, sound pressure level or timbre, and multidimensional features such as time and space. Different frequencies will result in different perceptual sound heights, i.e., pitches. Pitch is particularly malleable, as it provides intuitive connections with spatial dimensions. For example, pitch polarity (perceived as a rising or falling pitch) can be mapped to increasing or decreasing data values [25]. Pitch has also been used to represent vertical movements [28, 41, 47], and even horizontal movements, with low pitch being associated with left movements and high pitch with right movements [47].

The sound designs presented in this work draw inspiration from Sonification theories, such as PMSon [25], and from findings from the aforementioned previous studies.

3 LAYER-BASED SONIFICATION

To tackle the challenges that arise from the broad variety of robots being developed and deployed, together with the wide range of use cases [10], we propose a layer-based sonification framework that allows to map robot features to audio parameters combined in one single audio stream. The basic idea of such a framework is to decide on several layers of information pertaining to a robot's characteristics, actions, and intentions—akin to a manifesto for the entity that the robot is acting as, combined with a description of how it is behaving and what it is doing. Each layer is made of multiple information that are then mapped to an audio parameter, creating an ideally unique sound that could be understood intuitively. Such a manifesto could be easily adaptable to the particular robotic platform and deployment context of interest, thus decreasing the risk brought by "one-size-fits-all" voices [10].

3.1 Focus Group with Roboticists

The layers, and the robot features to sonify contained in each layer, were decided after conducting a participatory design workshop with roboticists, who discussed what features should be prioritised in the context of mapping them onto a sound. We recruited 10 experts in robotics (5 women and 5 men) who had been in the field for an average of 7.5 years (from PhD students to full professors). The one-hour focus group took place online, via a video-conferencing tool, and was moderated by the first and last author. The full results from this focus group are beyond the scope of the current article and are in preparation for publication elsewhere. For the purposes of the current article, we will briefly describe the setup and outcomes.

The focus group was structured as follows: First, participants introduced themselves by describing iconic features of typical actions from the robots they were the most familiar with. This first step was used to identify use cases of robot features that could potentially be sonified (e.g., an unintentional interaction between a robot and a group of pedestrians). Then, participants were divided into three smaller groups to discuss a single use case each. Each group cooperatively designed a robot concept to address the assigned scenario; robot concepts included the main features and components of a theoretical robot, such as size, appearance, and movement. After completing

Table 1. Focus Group Response Codes Related to Robot Sound's Potential Applications for HRI that Occurred Most Frequently

Category	Concept	"Sound affects this concept by:..."
Motion	Autonomy	Adding to motion to support a sense of autonomy
	Discomfort	Reducing disruption caused by quick changes in consequential sound
	Feedback	Conveying spatial and motion information
	Predicability	Forewarning users of motions
State	Intent	Communicating the robot's purpose
	Presence	Notifying a user of the robot's proximity
	Interactivity	Indicating readiness for interaction
	Feedback	Communicating successes or errors
	Emotion	Conveying a robot's "emotional" state
Interaction	Multimodal	Leading attention and communicating alongside other interaction modalities
	Transparency	Showcasing their abilities or limitations
	Qualities	Corresponding with physical attributes
Design	Affordances	Using established iconic sounds, such as from media, that users will recognize
	Personalization	Allowing different sound profiles or configurable sound profiles

their designs, the participants were asked to imagine and discuss how their familiar robot actions, which were mentioned earlier, as well as the designed concepts, might be sonified. The results created in the smaller groups were then shared with the whole group and discussed briefly.

The focus group was designed to follow an inductive reasoning approach, going from concepts that would be most familiar to roboticists (e.g., talking about features and capabilities of the robots they use the most in their work) to concepts that we assumed might prove more difficult for them (e.g., thinking about how a robot's movement or internal state could be expressed with sound). Transcripts from the focus group were then coded using thematic analysis, resulting in the categories listed in Table 1.

3.2 Conceptual Layers

Based on the categories identified by the Robotics experts, we conceptualised three main "layers" of robot-related information that could be displayed via sound (a schematic representation of the layers is also shown in Figure 2):

Layer 1: Physical base. The first layer contains information about the nature of the robot, such as its model, size, degrees of freedom, and so on. These core rules give the robot an individual sound aesthetic, comparable to the tonal features of human voices, by driving sound parameters such as, e.g., timbre or reverberation. This layer can be seen as the basic manifesto of each robotic entity, intending to communicate the general concept of the robot.

Layer 2: Internal state. This layer contains information about how an action is being performed by the robot and how demanding or urgent the task is. The robot's cognitive load could be parametrised to convey information on how much a robot can deviate from its current action and interact with humans nearby. Further keywords that describe the robot's performance could be the current acceleration or its alertness towards unpredicted events.

Layer 3: Action. The top layer contains information about the action that the robot is presently carrying out. As this information can be most likely also obtained from visual stimuli, most of the encoded clues might be redundant. Nevertheless, sonifying actions such as walking, turning, waiting, grabbing, or listening might contribute to preventing collisions caused by misunderstandings or lack of robot visibility.

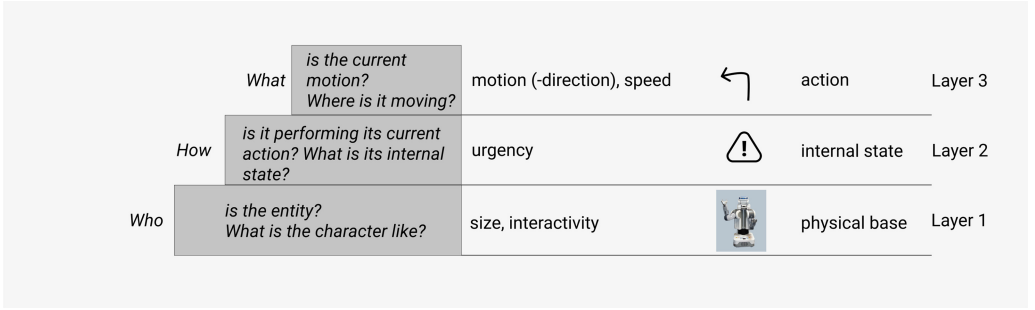

Fig. 2. Layer-based sonification framework.

3.3 Information Sonification

While we propose that the individual features to be sonified can differ based on the envisaged scenario, for our case study (robot and people conducting their individual tasks in the same environment), we selected the following robot-related information to sonify: size and interactivity (i.e., whether the robot is currently available for interaction with humans), belonging to Layer 1; urgency of the robot's task, belonging to Layer 2; speed and directionality, belonging to Layer 3.

In the next section, we describe in detail the design process that led us to communicate these features with a certain sound. Please note that for our approach, we only take into account intentional robot sounds. For simplicity, and to allow a more detailed exploration of intentional sounds, consequential sounds are not taken into account. For this work, we argue that consequential sounds are likely to be disregarded when intentional sounds are added, similarly to how people are able to block out roadwork sounds coming from the window but pay attention to alarms. However, this assumption could be investigated in future work (see, e.g., Reference [45]).

4 SOUND DESIGN PROCESS

4.1 Audio Terminology

Sound design processes are not often described in HRI literature. To aid the reader in this regard, we provide here a brief description of audio terms that we will use later in this section. Further information on audio theory can be found in, e.g., References [30] and [13].

— **Low Frequency Oscillator (LFO)**: A low rate oscillator used for automating audio parameters;
— Harmonics: Integer multiples of the fundamental frequency;
— Sine wave: Basic waveform that derives from a classic periodic oscillation (consists of a single fundamental frequency);
— Sawtooth wave: Waveform that is rich in harmonics; consists of an infinite set of even and odd harmonics of the fundamental frequency [13];
— Additive synthesis: Method of combining individual harmonics; allows precise manipulation of the frequency domain [13];
— Subtractive synthesis: Synthesis method where the main tonal modification technique is filtering; a simple way of generating signals with very characteristic tonal qualities by reducing or amplifying the amplitudes of many frequencies simultaneously [13];
— Bell filter: Audio effect that increases (gain) or attenuates the signal at a specific center frequency;
— Filter slope: Steepness of the filter curve; measured in dB per octave;

— Q-factor: "Quality Factor" that defines the bandwidth of a filter effect; lower Q-factor means broader bandwidth, higher Q-factor smaller bandwidth;
— Panning: Panoramic controls for the amount of left- and right-audio channel mixed to the output signal; used for simulating spatial position of an audio source;
— Spatial Audio (also known as 3D Audio): 3-dimensional simulation of an audio environment as how it is perceived by humans;
— Dry mix: The unprocessed audio signal;
— Wet mix: The processed audio signal that is the output of manipulating the audio material with audio effects;
— Filter & **equaliser (EQ)**: Audio effect for processing sound in the frequency domain; modifying the timbre of the audio material; filters reduce the amplitude for given frequencies, while EQs can either enhance or reduce the amplitude [44];
— Binaural stereo (binaural): Stereo rendering of a virtual audio environment that simulates the listener's head-rotation; used for displaying Spatial Audio scenes, e.g., on headphones;
— Flanger: Modulation audio effect that manipulates the frequency domain by duplicating the audio signal and shifting the (wave-) phase of one copy.

4.2 Sound Design

Mapping audio parameters to the identified robot features is a complex task. To achieve it, we relied both on previous research demonstrating the association between certain audio parameters and meaning (see below) and on our own sound aesthetics. Thus, we do not claim that our final designed sounds are the most appropriate and intuitive sounds to communicate the information we want to communicate; rather, we suggest that the individual audio parameters that comprise a sound can each communicate a piece of the full set of robot-related information. The sounds we designed are just an example of many such sounds that could communicate the same information.

Our five robot features of interest (size, interactivity, urgency, speed, directionality) were individually mapped to an audio parameter, under the main design consideration of providing immediate and simple feedback. Given the flexibility to generate sounds in real-time afforded by sound synthesis, the designs were based on synthesised sounds created from filtered white noise signal (subtractive synthesis).

This way, information can be encoded in an audio signal by modulating the filter parameters with LFO, such as the range and amplitude of modified frequencies, and by modulating the LFO parameters. In the case of directionality, one design additionally encodes spatial information by applying spatial panning, a conventional way of simulating spatial information by balancing the audio signal between the output channels. The automation of parameter modulation allows to create complex patterns of information; while more complex concepts with continuous values are desirable for future implementations, we reduced complexity by creating binary value-states for each of the targeted features. We designed two different sounds for each of the five features. We will refer to them as "design 1" and "design 2" and also give intuitive names for each of them. A detailed technical description of each designed sound can be found in Tables 1–5 in the supplementary materials.

The designed sounds can be found online www.bastianorthmann.eu/lts/hriaudio/0323.zip here.

4.2.1 Size. Size is naturally related to the pitch and loudness dimension of emitted sounds. For example, people and animals with bigger vocal tracts tend to have lower-pitched voices than people and animals with smaller vocal tracts [42]. Also, a heavier person tends to be associated with footsteps with a lower spectral mode (lower frequency), while a lighter person walks with footsteps that have a higher spectral mode (higher frequency) [34, 52]. Based on these findings, we decided

to display large size through more energy in the lower frequency domain and small size through more energy in the higher frequency domain. This results in modifications of the overall timbre of the sound, its *tonal character* [13], leaving the pitch domain available for other feature mappings.

For design 1 (*multiband frequency*) the frequency bandwidth is broader for large size display, containing low-, mid-, and high-frequencies; and it is narrower for small size, with reduced energy in the mid-frequency range (see Figures 1 and 2 in the supplementary materials). The resulting sound should be perceived as *broader*, *heavier*, and *solid* for large size and as *thinner*, *lighter*, and *airy* for small size. Thus, the bandwidth of the signal spectrum is generally wider and more balanced for large size display than for small size, where the high spectrum from 5 kHz to 15 kHz predominates. The resulting distribution of energies among frequencies can be seen in Figure 3 of the supplementary materials.

Similarly, for design 2 (*single-band frequency*) large size sounds have a broader frequency spectrum than small size sounds, with the difference that small size is also reduced in the high frequency spectrum and has a generally more moderate filtering, with no enhanced frequencies. The result is a more *intense*, *broad* sound for large size and a more *subtle* and *softer* sound for small size (see Figure 4 in the supplementary materials). Overall, for single-band frequency the noise signal is modified less than for multiband frequency, and small and large size are distinguished mainly by altering the total frequency bandwidth, while for multiband frequency the frequency spectrum is split into multiple bands, with thinner bandwidth on all frequency bands for small size compared to large size sounds.

4.2.2 Speed. Speed falls into the category of kinematics, and it is often displayed with pitch-related and temporal auditory dimensions and at times also through loudness [16]. We therefore decided to sonify the robot's speed using the auditory dimensions of tempo and loudness. Specifically, we modulated the amplitude of the signal in patterns: slow speed was displayed with gradually increasing and decreasing amplitude slopes, whereas fast speed was shown through quick amplitude modulations. The results for design 1 (*flow-modulation*) are two distinct motion patterns: a gentle, ocean-like wave motion represents slow speed, with a constant audio signal lying under the modulations, with a pace of 0.82 cycle during a 2-second sequence. The fast speed pattern instead resembles the choppy sound of helicopter blades cutting through air, giving it a staccato articulation and a pace of 21.6 cycles during the sequence (see Figure 5 in the supplementary materials).

The fact that the modulation rates are not synced to the white noise sequence duration makes the patterns less repetitive and gives a constantly changing sound structure—an important design consideration to avoid listener fatigue.

The second design *(constant-modulation)* for slow speed sounds more "restless" (see Figure 6 in the supplementary materials), suggesting a steady build-up of intensity rather than ocean waves. The constant-modulation design for fast speed oscillates at a faster rate than in flow-modulation design, but with less harshness (see Figure 7 in the supplementary materials).

4.2.3 Interactivity. Displaying availability for interaction requires more abstraction, since we cannot automatically be inspired by the audio cues surrounding us daily. We started with the assumption that, for our imagined scenario, non-interactive robots would be the default, and that we would only indicate when a robot can be interacted with. Therefore, we only modify the audio signal when we want to signal that a robot can be interacted with. We took inspiration from emotion theories, suggesting that people tend to approach people who show signs of positive emotions, such as happiness [17, 55]. This approach invitation could be translated into robots that are available for interaction with "happy" sounds. These could be created by adding brightness or

positive tonality, thus tweaking the timbre or pitch of an audio signal, which has been shown to invoke positive affect [20, 21].

To achieve a perceptually *happy*, *friendly*, and *sociable* sound, *additive synthesis* was used in both designs. The resulting audio from design 1 *(basic-buzz)* has a buzzing quality to it and stands out compared to the non-interactive white noise base signal.

Design 2 *(harmonic-buzz)* included enhanced frequencies corresponding to the main frequencies of a C-major chord (C5-E5-G5), a very common chord from a major scale, which are generally attributed as having a positive emotional connotation [4]. The chord is intentionally not played by the sawtooth-wave synthesiser, but instead indicated through enhancing the frequencies of the audio signal, which is expected to be more subtle and seem less than an obvious music chord. Therefore, harmonic-buzz does not differ too much from basic-buzz in its total sound colour, but still augments the buzzing sound with a harmonic impression.

4.2.4 Urgency. When conveying urgency, we want to naturally catch people's attention and make them aware of an urgent situation. For this, we can borrow concepts from alarm sound design (e.g., References [18, 19]). In our current HRI scenario however, we do not want to warn the whole surrounding area of a dangerous situation, but we only want to alert the people around the robot that the situation requires additional attention; for example, this could be an implicit request from the robot to move aside and let it pass, in cases, e.g., where the robot has been summoned urgently. A more appropriate term for this type of sound would be *alert* sound rather than *alarm* sound, although we could consider alert and notification sounds as sub-categories of alarm sounds. Typical features that are employed in the design of alarm sounds are the use of frequencies that stand out in our hearing system, and the creation of rhythms, where fast patterns signal higher alertness, as can be seen, for example, in ambulances, police cars, fire engines, or medical devices [18].

The most prominent frequency range in the human hearing system lies between 1 kHz and 4 kHz [8] and is also referred to as equal-loudness contour. As for non-interactive sounds, we decided on non-urgent display to be the default parameter and to only modify the audio to display urgency. Design 1 *(single tone alert)* shows a higher amount of high-frequencies and has the character of a single tone, while design 2 *(multi-tone alert)* shows a smaller bandwidth and is less distinct.

4.2.5 Directionality. For sonifying directionality, we took inspiration from Reference [16], which lists spatialization and pitch as the main audio parameters to convey this information. Spatialization can either be realised by distributing the signal between two stereo channels—*panning*—or by utilising Spatial Audio—a way of simulating a 3-dimensional audio environment. An intuitive approach for design 1 *(spatial-direction)* was therefore to pan the audio signal from the center to the left and right side of the audio system. The panning automation was synchronised to the trigger cycle of the audio signal and took 2 seconds.

This sonification implies that the sounding robot will be equipped with a stereo speaker setup; however, there might be cases where it is not possible to attach multiple speakers to the robot or to place them far enough away from each other, which will result effectively in monophonic audio playback. Therefore, the stereo panning approach could be problematic in some cases. An alternative concept, that we used for design 2 *(pitch-direction)* is to display directionality through pitch modulation. While most real-world examples map pitch to a vertical axis, there are some examples of pitch being mapped to a horizontal axis, too. One such example is the piano, where the lower pitch tones are placed on the left side of the instrument and the higher ones on the right side. Thus, we implemented this concept by modulating filter frequencies to create a perceived change of pitch.

4.3 Stimuli

In summary, we designed two versions of each sonified robotic feature; these were then combined into sounds comprising all possible combinations of the sonifications, resulting in 40 final audio files to be used for evaluation (see Table 1 in the supplementary materials for details. The sound files can be downloaded here). These stimuli were between 6 and 15 seconds long.

5 EVALUATION

We conducted a series of online studies (using the platform Prolific) to evaluate the designed sounds. The evaluation was split into two sets of studies: The first set of studies was used to evaluate each feature (speed, size, interactivity, urgency, directionality) individually (*individual feature evaluation*), while in the second set of studies, participants evaluated two different features simultaneously (*combined feature evaluation*). We would like to point out that the main aim of the evaluation was not to prove that our implementations of the sound designs were the most suitable ones; rather, we wanted to see whether certain information can be conveyed via sound at all (*individual feature evaluation*) and whether the same sound can convey more than one source of information simultaneously (*combined feature evaluation*). If people are able to perceive the information, as we intended, based on the audio material, then that would indicate that mapping different robot characteristics or internal states to audio features could be implemented with the same technique in future designs of robot auditory displays. The designs therefore mainly serve as a way to demonstrate this concept.

All the studies were conducted according to KTH ethical guidelines; the data was anonymised and participants were paid an average amount of £8.4 per hour. During the studies that took longer than approximately 15 minutes, participants were actively asked to take a break after they had completed 50% of the trials.

5.1 Individual Feature Evaluation: Size, Speed, Urgency, Interactivity

The intent of the individual feature evaluation was to find out whether our sound designs could theoretically convey one type of information at a time, proving that our auditory display for robot characteristics conveys the intended information and that nonverbal, generative audio can be used for such display. To be able to compare the results of this study, the audio signals already contained all our layers of information, but the participants only had to rate one feature at a time.

5.1.1 Method. As described in Section 4.3, we created a total of 40 sound stimuli, encoding different levels and different designs of our robot features of interest (size, speed, urgency, interactivity, and directionality; for the evaluation of directionality, see Section 5.2). Since asking participants to evaluate these sounds for all our features of interest would have resulted in a very long study, where participants' attention and response quality could have easily deteriorated over time, we divided the questions in three sub-studies. Each of these sub-studies included 32 sounds (that is, all the stimuli excluding the 8 additional sounds created to encode directionality). We recruited 50 participants for each study, on Prolific. The 32 sounds that were used for each study consisted of combinations of small or large size, slow or fast speed, non-urgency or urgency, and non-interactivity or interactivity. Participants were pre-screened to be at least fluent in the English language. Furthermore, participants were given the option to leave a free-text comment after each trial and at the end of the study.

In the first study, we focused on the evaluation of size and speed. The 32 sounds were played to participants in random order, and for each they were asked to provide a rating either on the size or the speed of a robot producing these sounds. For size, participants were asked to rate, based on the sound they were hearing, what size they expected the robot making that sound to be. The

evaluation was based on a scale from 1 to 7, with 1 labelled as *very small*, 4 as *average human size,* and 7 as *very big*. For speed, the same participants were asked to rate, based on the sound they were hearing, at what speed they expected the robot making that sound to be moving. The same scale as for size was used, with 1 labelled as *very slow*, 4 as *average walking speed,* and 7 as *very fast.* After participants completed the evaluation of all the 32 sounds, they were asked to fill out a short demographics survey asking about their age, gender, country of residence, and previous experience with robots.

The participants' countries of residence were South Africa (23), Portugal (6), USA (3), France (2), Spain (2), UK (1), Canada (1), Israel (1), Italy (1), Mexico (1), Morocco (1), Nigeria (1), Pakistan (1), Poland (1), UK-Israel-Turkey (1), and 3 people did not disclose this information. There were 27 males and 22 females; their age ranged from 18 to 59 years old (median = 25). Regarding previous experience with robots, 26 people reported only seeing a robot in TV or other media, 14 interacted with a robot before, 4 interact with a robot on a regular basis, and 5 had never seen a robot before. This study lasted, on average, 34 minutes.

The second study evaluated the same sounds for urgency in a randomised order. The participants were asked to rate, based on the sound they were hearing, whether they expected the robot making that sound to move out of their way or whether it would expect them to move out of its way. The evaluation was based on a scale from 1 to 7, with 1 labelled as *will move out of my way*, 4 as *will either move out of my way or expects me to move out of its way,* and 7 as *expects me to move out of its way.* Participants filled out the same demographics questionnaire after the sound evaluation was complete. The participants' countries of residence were Poland (12), Portugal (8), Mexico (6), South Africa (6), Spain (6), UK (3), USA (1), Latvia (1), Greece (1), and 5 did not disclose this information. There were 25 males, 22 females, 1 non-binary person, and 1 person who self-described as "other"; their age ranged from 18 to 56 years old (median = 24). Regarding previous experience with robots, 27 people reported only seeing a robot in TV or other media, and 22 interacted with a robot before. This study lasted, on average, 18 minutes.

The third study evaluated the same sounds for interactivity in a randomised order. The participants were asked to rate, based on the sound they were hearing, whether they expected the robot making that sound to be interactive or not. The evaluation was based on a scale from 1 to 7, with 1 labelled as *not available for interaction*, 4 as *either not available for interaction or is available for interaction,* and 7 as *available for interaction.* Participants filled out the same demographics questionnaire after the sound evaluation was complete. The participants' countries of residence were South Africa (16), Portugal (6), Spain (6), Poland (5), Hungary (3), Greece (2), Italy (2), UK (2), Latvia (1), Mexico (1), Namibia (1), Pakistan (1), Slovenia (1), Zimbabwe (1), Estonia (1), and 1 person did not disclose this information. Their age ranged from 18 to 53 years old (median age was 27). There were 24 males, 25 females, and 1 non-binary person; their age ranged from 18 to 53 years old (median = 27). Regarding previous experience with robots, 32 people reported only seeing a robot in TV or other media, 17 interacted with a robot before, and 1 interacts with robots on a regular basis. This study lasted, on average, 17 minutes.

5.1.2 Results. Analyses were conducted in R version 4.2.0.

For studies 1–3, we performed a series of ANOVA tests to see whether the average rating was different based on the feature of interest (size, speed, interactivity, urgency) and the specific design variation (design 1 or 2).

For *size*, there were significant differences in the rating of small and large sounds ($F(1, 1541) = 29.99$, $MSE = 2.63$, $p < .001$, $\hat{\eta}^2_G = .019$); as can be seen in Figure 3(A), sounds designed as *large* were perceived as belonging to larger robots, and sounds designed as *small* were perceived as belonging to smaller robots. There was no statistically significant difference between the two

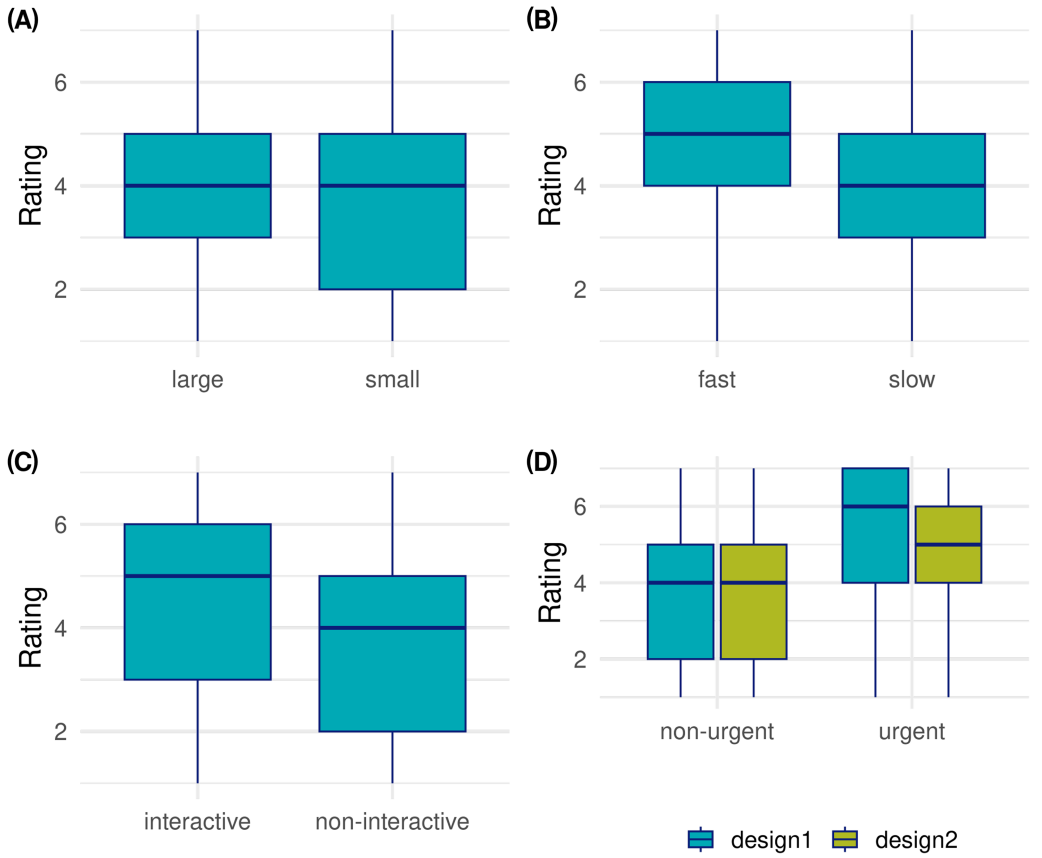

Fig. 3. A: Average ratings of sounds designed to convey the information "large robot" and "small robot"; higher ratings correspond to bigger sizes. B: Average ratings of sounds designed to convey the information "fast robot" and "slow robot"; higher ratings correspond to faster speeds. C: Average ratings of sounds designed to convey the information "interactive robot" and "non-interactive robot"; higher ratings correspond to higher interactivity. D: Average ratings of sounds designed to convey the information "robot performing an urgent task" and "robot performing a non-urgent task"; higher ratings correspond to higher urgency.

design variations ($F(1, 1541) = 0.07$, $MSE = 2.63$, $p = .796$, $\hat{\eta}^2_G = .000$), indicating that both designs conveyed the information equally well.

For speed, overall, people rated the sounds designed as *fast* higher in terms of speed than the sounds designed as *slow* ($F(1, 1547) = 133.29$, $MSE = 2.71$, $p < .001$, $\hat{\eta}^2_G = .079$), with no significant differences between the two designs ($F(1, 1547) = 0.16$, $MSE = 2.71$, $p = .687$, $\hat{\eta}^2_G = .000$), as can be seen in Figure 3(B).

For interactivity, overall, people rated the sounds designed as *interactive* higher in terms of interactivity than the sounds designed as *non-interactive* ($F(1, 1568) = 39.26$, $MSE = 3.67$, $p < .001$, $\hat{\eta}^2_G = .024$), as shown in Figure 3(C). There was no effect of design variation ($F(1, 1568) = 0.06$, $MSE = 3.67$, $p = .808$, $\hat{\eta}^2_G = .000$).

For urgency, people rated the sounds designed as *urgent* higher in terms of urgency than the sounds designed as *non-urgent* ($F(1, 1543) = 199.54$, $MSE = 3.07$, $p < .001$, $\hat{\eta}^2_G = .115$), and the two sound designs were rated significantly differently from each other ($F(1, 1543) = 12.55$,

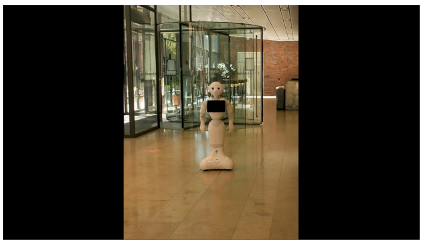

Fig. 4. Screenshot taken from the videos used in the directionality study (Pepper robot).

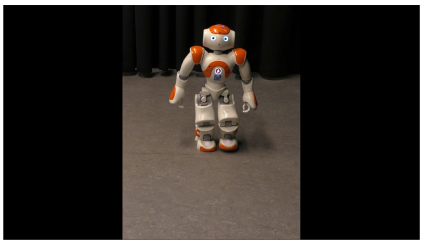

Fig. 5. Screenshot taken from the videos used in the directionality study (Nao robot).

$MSE = 3.07$, $p < .001$, $\hat{\eta}^2_G = .008$), with single tone alert conveying more urgency than multi-tone alert (see Figure 3(D)).

5.2 Individual Feature Evaluation: Directionality

5.2.1 Method. We reasoned that it would have been unintuitive for participants to rate how much a robot was about to turn left or right; therefore, the directionality feature was evaluated in a slightly different way than the other features. First, we ran an online study (similar to those presented above), where participants were shown a video of a robot walking towards them in a straight line, and the video stopped just before the robot reached them; they were then asked to state whether the robot would continue left or right next. Half of the participants were shown eight trials with a video featuring the robot Pepper (Figure 4); the other half were shown eight trials with a video featuring the robot Nao (Figure 5). Both robots are commonly used in HRI research and development, and we decided to show more than one robot to ensure that the meanings were understood regardless of the robotic platform. In both cases, the videos were shown to participants in random order. Additionally, since the Nao robot is much smaller than Pepper, half of the participants heard the sounds with the *small size* feature and the other half with the *large size* feature (the robot shown and size feature were counterbalanced between participants). Figure 6 shows how directionality was rated. The eight different sounds used for evaluating directionality had consistent base values of small or large size, slow speed, non-urgency and non-interactivity, combined with left or right directionality, as shown in Table 2. For this study, we recruited 100 participants on Prolific, who were pre-screened to be at least fluent in the English language. As in the other studies, participants rated all the videos and then filled out a short demographics questionnaire.

Seven participants were removed from the results due to technical issues with video loading. Of the remaining 93 participants, reported countries of residence were Portugal (21), South Africa (21), Poland (18), Italy (5), Greece (4), Hungary (4), UK (4), Spain (3), USA (2), Belgium (1), France (1), Mexico (1), Slovenia (1), Turkey (1), and 5 did not disclose this information. Their age ranged from 18 to 56 years old (median age was 23). There were 45 males and 48 females. Regarding previous experience with robots, 49 people reported only seeing a robot in TV or other media, 42 interacted with a robot before, and 2 had never seen a robot before. The study lasted, on average, 9 minutes.

5.2.2 Results. For the online study, we performed a binary logistic regression to see whether people were more likely to indicate that the robot was about to turn right or left based on the designed sounds, the robot type (Pepper or Nao), the *size* feature (small or large), and the specific design variations (design 1 or 2).

We found that people thought that the robots playing the sounds designed as *right-turning* were about to turn right, and that the robots playing the sounds designed as *left-turning* were about to turn left ($b = 1.23$, 95% CI $[0.92, 1.55]$, $z = 7.59$, $p < .001$), as shown in Figure 7. There

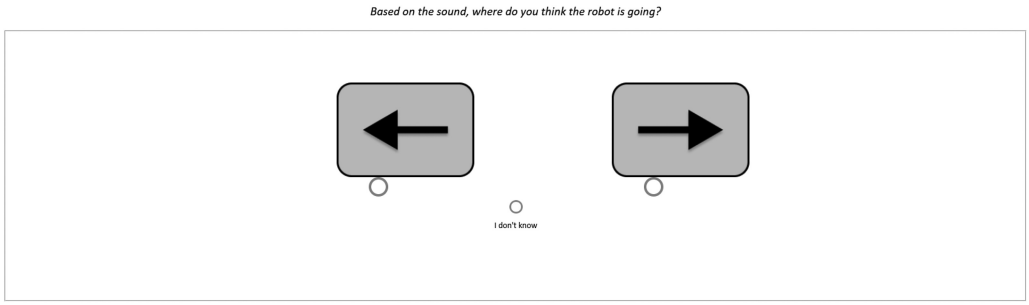

Fig. 6. Screenshot taken from the directionality study, showing the question presented to participants after each video.

Table 2. Value Combinations for the Stimuli Used in the Online Study on Robot Directionality

Large	Slow	Non-interactive	Non-urgent	Left (spatial-direction)	Pepper
Large	Slow	Non-interactive	Non-urgent	Right (spatial-direction)	Pepper
Large	Slow	Non-interactive	Non-urgent	Left (pitch-direction)	Pepper
Large	Slow	Non-interactive	Non-urgent	Right (pitch-direction)	Pepper
Small	Slow	Non-interactive	Non-urgent	Left (spatial-direction)	Pepper
Small	Slow	Non-interactive	Non-urgent	Right (spatial-direction)	Pepper
Small	Slow	Non-interactive	Non-urgent	Left (pitch-direction)	Pepper
Small	Slow	Non-interactive	Non-urgent	Right (pitch-direction)	Pepper
Large	Slow	Non-interactive	Non-urgent	Left (spatial-direction)	Nao
Large	Slow	Non-interactive	Non-urgent	Right (spatial-direction)	Nao
Large	Slow	Non-interactive	Non-urgent	Left (pitch-direction)	Nao
Large	Slow	Non-interactive	Non-urgent	Right (pitch-direction)	Nao
Small	Slow	Non-interactive	Non-urgent	Left (spatial-direction)	Nao
Small	Slow	Non-interactive	Non-urgent	Right (spatial-direction)	Nao
Small	Slow	Non-interactive	Non-urgent	Left (pitch-direction)	Nao
Small	Slow	Non-interactive	Non-urgent	Right (pitch-direction)	Nao

were no main effects of design variation, size, or robot, meaning that all the sounds we designed managed to convey the information right-motion and left-motion regardless of the robot they were associated with.

5.3 Combined Feature Evaluation

Finally, we conducted a study to see whether different pieces of information could be perceived simultaneously within one sound, thus lending support to our proposed layer-based system.

5.3.1 Method. We recruited 50 participants on Prolific. Participants were played the 32 sounds containing the speed, size, interactivity, and urgency parameters (in random order) and were asked to evaluate them in terms of two features at the same time: either speed and size, interactivity and size, or urgency and size. Participants were pre-screened to be at least fluent in the English language. One participant had to be excluded due to technical reasons; of the remaining 49, countries of residence were: South Africa (14), Poland (6), Italy (3), Portugal (3), USA (2), UK (2), Israel (2), Hungary (2), Belgium (2), Canada (1), Greece (1), Latvia (1), Nigeria (1), Pakistan (1), Spain (1), Uganda (1); there were 26 males, 22 females, and 1 non-binary persons; their age ranged from 18 to 38 years old (median = 24). Regarding their previous experience with robots, 27 people reported

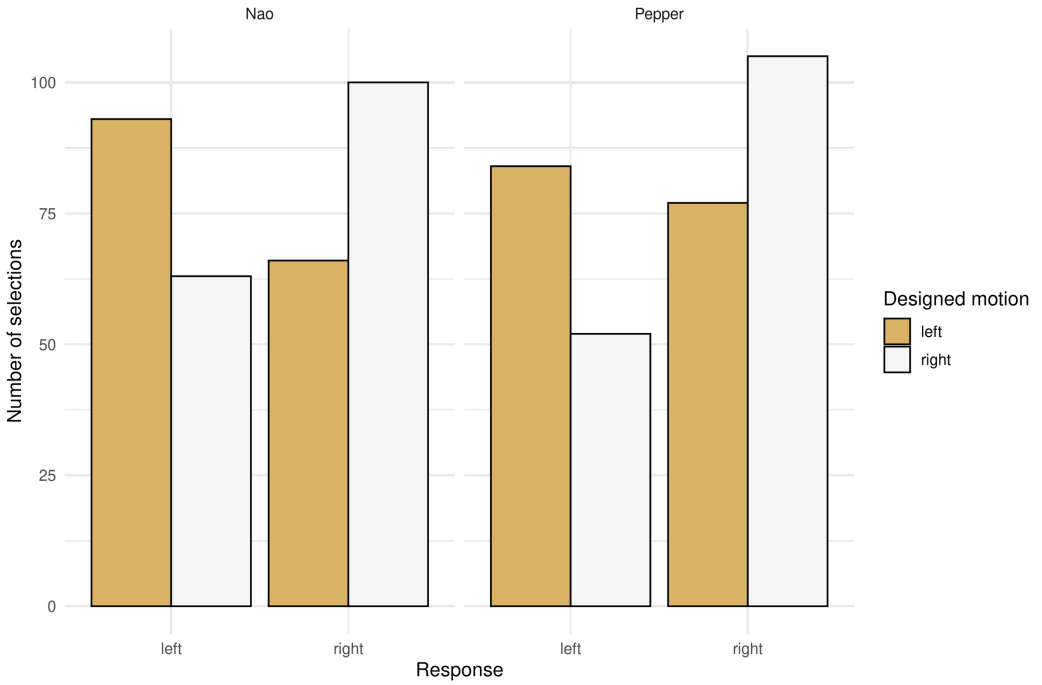

Fig. 7. Number of times participants in the online study indicated that the robot in the videos was about to turn left or right.

Table 3. Number of Times a Certain Combination of Size and Speed Values Was Selected upon Hearing a Sound Showing the Designed Values

	Response values			
	Large + Fast	Large + Slow	Small + Fast	Small + Slow
Large + Fast	109 (*)	41 (*)	181 (*)	58
Large + Slow	89	168 (*)	59 (*)	69
Small + Fast	80	35 (*)	207 (*)	62
Small + Slow	47 (*)	155 (*)	67 (*)	111 (*)

"*" Indicates Statistical Significance (adjusted $\alpha = .003$).

only seeing a robot in TV or other media, 15 interacted with a robot before, 4 interact with a robot on a regular basis, and 3 had never seen a robot before.

5.3.2 Results. We conducted chi-square tests for independence on each of the robot features of interest—size, speed, urgency, interactivity—to see whether there was a causal relationship between the robot features being evaluated and the designed sounds. Post hoc analyses—to see whether a feature was selected more often than the others for each sound—were conducted by testing the adjusted residuals of each cell of the contingency table a critical z value and adjusting the alpha level for multiple comparisons (Bonferroni correction). This procedure for post hoc testing of a chi-square test is described in Reference [5].

For the size + speed combination, a chi-square test of independence was performed on the whole contingency table (see Table 3) and found a significant association between the designed sounds and evaluations of size and speed ($\chi^2(9) = 339.09, p < .001$, Cramer's V = 0.27). As shown in

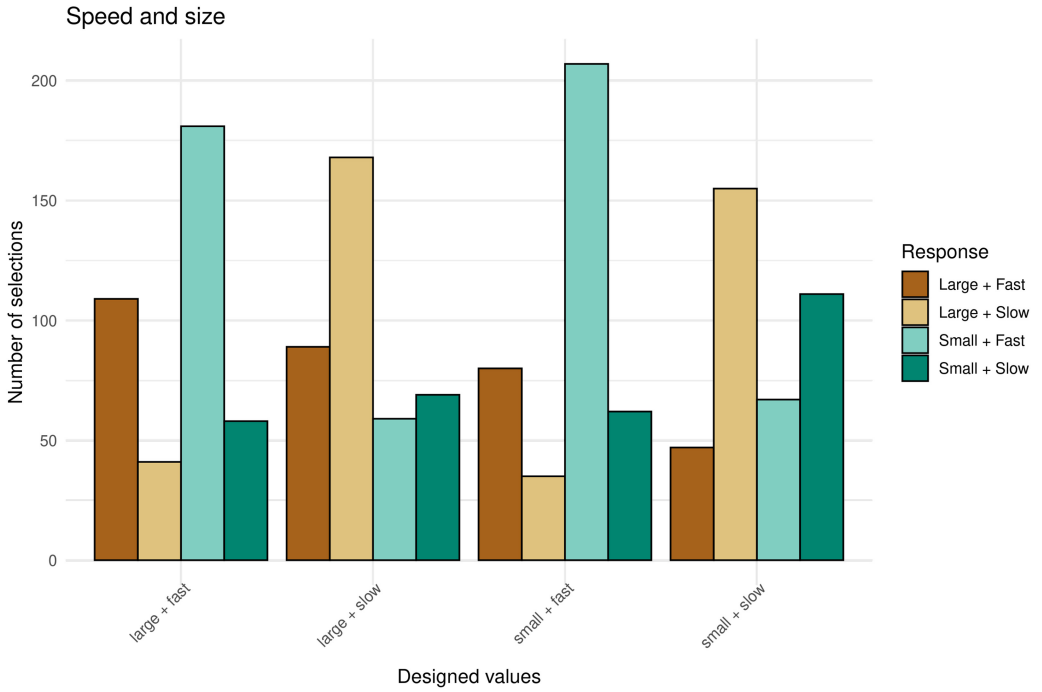

Fig. 8. Number of times a certain combination of size and speed values was selected upon hearing a sound showing the designed values.

Table 4. Number of Times a Certain Combination of Size and Interactivity Values Was Selected upon Hearing a Sound Showing the Designed Values

	Response values			
	Large + Interactive	Large + Non-interactive	Small + Interactive	Small + Non-interactive
Large + Interactive	26	34	18	19
Large + Non-interactive	33	49 (*)	6 (*)	7 (*)
Small + Interactive	17	27	26	27
Small + Non-interactive	28	28	17	20

"(*)" indicates statistical significance (adjusted $\alpha = .003$).

Figure 8, people mostly associated the sounds they were hearing with the combined information they were intended to convey.

For the size + interactivity combination, a chi-square test of independence was performed on the whole contingency table (see Table 4) and found a significant association between the designed sounds and evaluations of size and interactivity ($\chi^2(9) = 37.365, p < .001$, Cramer's V = 0.18). As shown in Figure 9, only the sound designed to convey the information "large robot" + "robot not available for interaction" were interpreted as intended by participants.

For the size + urgency combination, a chi-square test of independence was performed on the whole contingency table (see Table 5) and found a significant association between the designed sounds and evaluations of size and urgency ($\chi^2(9) = 42.911, p < .001$, Cramer's V = 0.19). As shown in Figure 10, it seems that people only understood partial information from the sounds they heard: They indicated a sound to be non-urgent in combination with both "small robot" and "large robot," and they indicated a sound to be urgent in combination with both "small robot" and "large robot."

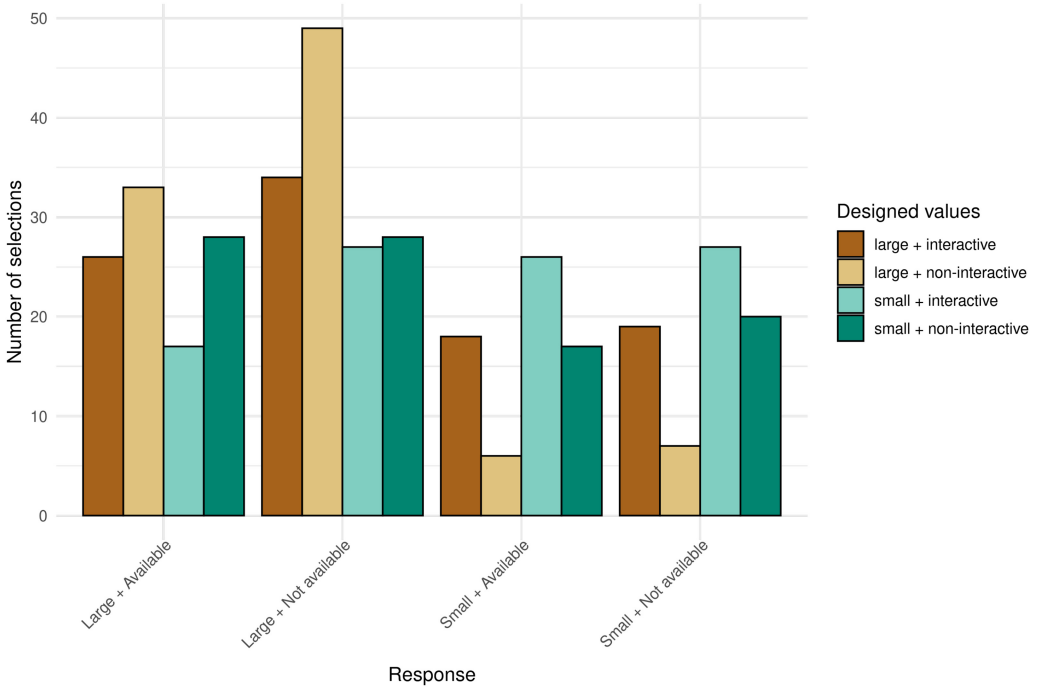

Fig. 9. Number of times a certain combination of size and interactivity values was selected upon hearing a sound showing the designed values.

Table 5. Number of Times a Certain Combination of Size and Urgency Values Was Selected upon Hearing a Sound Showing the Designed Values

	Response values			
	Large + Urgent	Large + Non-urgent	Small + Urgent	Small + Non-urgent
Large + Non-urgent	42	36 (*)	8 (*)	10
Large + Urgent	32	15	31 (*)	18
Small + Non-urgent	36	28	10	21
Small + Urgent	35	16	31 (*)	14

"(*)" indicates statistical significance (adjusted $\alpha = .003$).

6 DISCUSSION

In this work, we detailed the process of designing sounds for human-robot interaction, going from a conceptualisation of the problem and the scenario at hand to the sound creation informed by formal research as well as designer's aesthetics to a series of evaluations of the created sounds.

We suggested stratifying the robot features to be sonified into multiple layers, corresponding to the robot's identity, internal state/goal, and current action (see Figure 2). Then, we designed different examples of sounds that could represent each feature. These designs were then evaluated in a set of studies to investigate whether:

— each different state of a feature (small vs. large, slow vs. fast, non-interactive vs. interactive, non-urgent vs. urgent, turning left vs. right) is conveyed through the corresponding sound designs when focusing on one feature at a time;

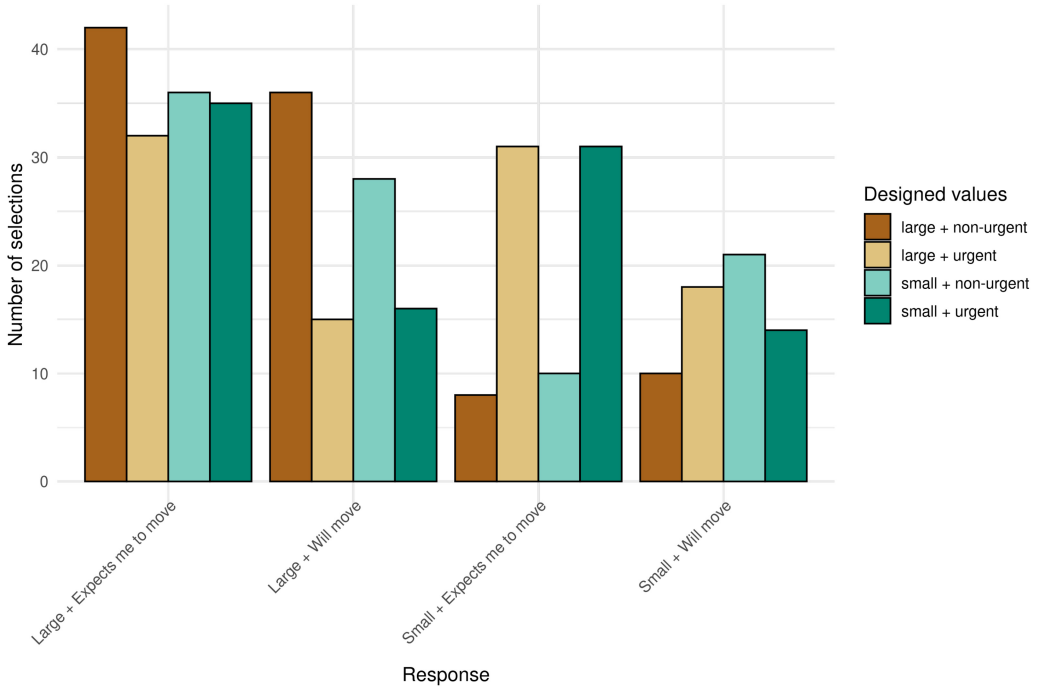

Fig. 10. Number of times a certain combination of size and urgency values was selected upon hearing a sound showing the designed values.

— the same different states can be identified when focusing on two features simultaneously while listening to the same sounds.

Overall, all our sound designs successfully conveyed the intended information when participants focused on one feature at a time, implying that sounds can convey information about relative differences between states in robot's internal processes and actions. The audio parameters that we chose to display size (low-cut filter, high-cut filter, and bell filter) were successful for conveying relative differences in size information and correspond to findings from Reference [16]. The parameters we used to display speed (amplitude modulation) successfully conveyed slow and fast speed—also lending support to the list of most prominent mappings in Reference [16]. The audio parameters that we chose to display interactivity (changing the tonal character through Additive Synthesis) resulted in higher ratings of interactivity than the default *non-interactive* designs. The parameters chosen to display urgency (bell filter in combination with frequency modulation and compression) successfully conveyed urgency over non-urgency, with better results for the single tone alert design, with the central frequency modulating in a broader range, between 470 Hz and 3.2 kHz, compared to the multi-tone alert design. Finally, both designs intended to convey the robot's directionality gave positive results. This was interesting, because, while it was expected that the spatialised sounds used for design 1 (spatial-direction) would be able to convey the robot's direction, the positive results from the pitch-direction design were more surprising. Therefore, we can conclude that pitch-based sonifications can also be used to convey directionality. This supports previous findings by Reference [47], which found a correlation between low pitch and left side and high pitch and right side. This is an interesting finding, since it means that directionality can be easily displayed on a robot with sound without the need for stereo output. However, since we did

not collect data on participants' handedness, it remains to be seen whether this sonification will be intuitive for both right- and left-handed individuals [15]. All in all, while there exists previous research evaluating whether one sound can convey one meaning (e.g., urgency [12] or directionality [7, 49]), we are not aware of any study designing and evaluating whether sounds meant to convey several different robot features.

The results for combined feature evaluation were less successful. While the combinations of size + speed conveyed the intended feature states in most cases, the combinations of size + interactivity and size + urgency were not conclusively identified. The results from Table 4 indicate that in most cases the sounds were found to convey non-interactive behaviour, with large + non-interactive display being the only combination significantly identified as intended. For size + urgency, relatively more sounds were rated to convey an urgent than non-urgent state. Both feature combinations indicate to have also confused the perception of size, since, e.g., in addition to most sounds associated to a non-interactive state, sounds designed as small size were wrongly assigned as displaying large size. Even though both urgency and interactivity were successfully understood when people were asked to focus on one feature at a time, this understanding was lost when they were asked to identify a double feature. This suggests that, when designing sounds that need to convey more than one information at the same time, the information to be conveyed should not only be evaluated in isolation, but also in combination, as evidently this combination does not necessarily ensure understanding in additive terms. This is in line with previous work showing that nonverbal communication that is more nuanced than a simple binary alternative is difficult to interpret [12]. However, the results from size + speed are promising and warrant further investigation. It is possible that the designs for sound and speed were particularly iconic and intuitive for participants, while the others were not. Given these results, one way forward would be to keep designing and evaluating new sounds until they can consistently convey all these combinations of meanings. Another way might be to have a multi-modal layer-based framework, for example, by having Layer 1 (the *who* in Figure 2) represented by a sound, Layer 2 (the *how*) represented by a visual display such as a colour or flashing light, and Layer 3 (the *what*) represented by a sound again.

The successful results for most designs might have to do with the fact that, instead of absolute values, only the display of two relatively different feature states was investigated. While, e.g., the non-urgent display was rated, on average, with a neutral value (4; see Figure 3(D)) in the individual feature evaluation, the urgency display as a whole was still successful, since the urgent display was rated with significantly higher values. This shows that focusing on displaying relative differences between feature states might be easier to achieve, which allows for the use of simpler sound designs that can be generated in real-time on various hardware devices.

All in all, amplitude modulation appears to be a suitable way for displaying the robot's speed, with low modulation rates representing slow speed and high rates indicating fast speed. Further, indicating opposite sizes with different frequency bands seemed to convey the intended information, with significant differences between increasing the low-frequency spectrum for sounds designed as large size and the high spectrum for sounds designed as small size or simply decreasing the bandwidth for small size and increasing it for large size.

6.1 Limitations and Future Work

The initial evaluations presented here relied on explicit feedback from participants (e.g., "how fast is the robot?"). While this kind of approach allowed us to collect a large amount of data and to draw conclusions based on statistical inference, these evaluations should in the future be complemented with qualitative, open-ended feedback aimed at collecting people's unbiased opinions on what meaning the sounds conveyed to them. Similarly, whether people can distinguish the intended meanings of our sounds played on a computer screen is a different matter than actually

understanding them while immersed in the intended context. The next step in our long design process will be to evaluate the sounds in an unconstrained environment, such as a shopping centre or museum, and observe the behaviour of passers-by. We plan to run qualitative analyses of video-recorded interactions and conduct interviews with a random sample of passers-by. We also plan to include all possible combinations of designs for parameter mappings and how learnable the designs are.

For evaluating the sounds conveying directionality, we chose to show videos of two commonly used robots, Nao and Pepper. Since these robots are considered to be anthropomorphic [43], it is possible that people might not understand that nonverbal sounds such as the ones we designed could be emitted by these robots and would rather expect more "human-like" vocalisations. When thinking of our scenario of interest, we had excluded speech as a way for robots to communicate their actions and intentions, because this would disrupt the navigation flow (we do not go around exclaiming "I am in a hurry!" or "I'm about to turn right"; see also Reference [26]). However, future work should also research the appropriateness of these nonverbal sounds for the specific robotic platform being used (see also References [37, 39]).

Furthermore, it is possible that individual differences influence how people interpret the meaning of a sound. Specifically, cultural differences can play a role in the perception of audio features such as pitch, rhythm, and timbre—although these seem to be more likely to occur in verbal rather than non-verbal communication [25]. Also, the perception of sounds varies for different age groups, as, e.g., with increasing age the perception of concurrent sounds declines [1], peripheral sensory problems increase, and cognitive aging affects auditory perception, such as speech processing [2]. Therefore, future work should include different cultures and age groups for evaluating the introduced concepts, as issues due to intercultural and inter-generational differences might arise and influence the efficacy of the proposed sonification methods.

For a more universal information display, more work is needed also for defining action spaces and internal states. A modular framework that can be adapted for different use cases would be desirable. A promising implementation of the presented layer-based approach could include detecting when to display information through a sound and when to remain silent. This would require a complete system that allows the robot to create a self-model of its role in the current environment and how it is expected to navigate within. A complete auditory feedback system for the robot could include this flexible decision making about when to display its internal processes, how it reacts to environmental processes (valence feedback, different degrees of feedback such as, e.g., notification, alert, alarm) and generate the sounds in real-time. Thus, the next step in our research is to build a synthesiser that, given inputs on a robot's internal state, next action, and base identity, could automatically generate the required information via sound in real-time. We plan to make this synthesiser modular and portable, so different robots could be equipped with it.

Finally, while in the current work we focused exclusively on sounds, in the future, we plan to integrate other communicative signals, such as gestures and lights, to further augment understanding and increase accessibility. For example, we could leverage emotion theories to display colours to trigger approach or avoidance behaviours together with interactivity and urgency sonifications, and we could investigate adding a pointing gesture or flashing light to the left- and right-turning sounds. Previous works focusing on lights and gestures have shown their communicative potentials in HRI [6, 12, 51, 64].

6.2 Implications for the Field

Based on our experience from this work's process, we illustrate in Figure 11 how an iterative conceptualisation, design, and evaluation process can be followed when creating auditory displays for unintentional HRI. In combination with particular design recommendations drawn from the

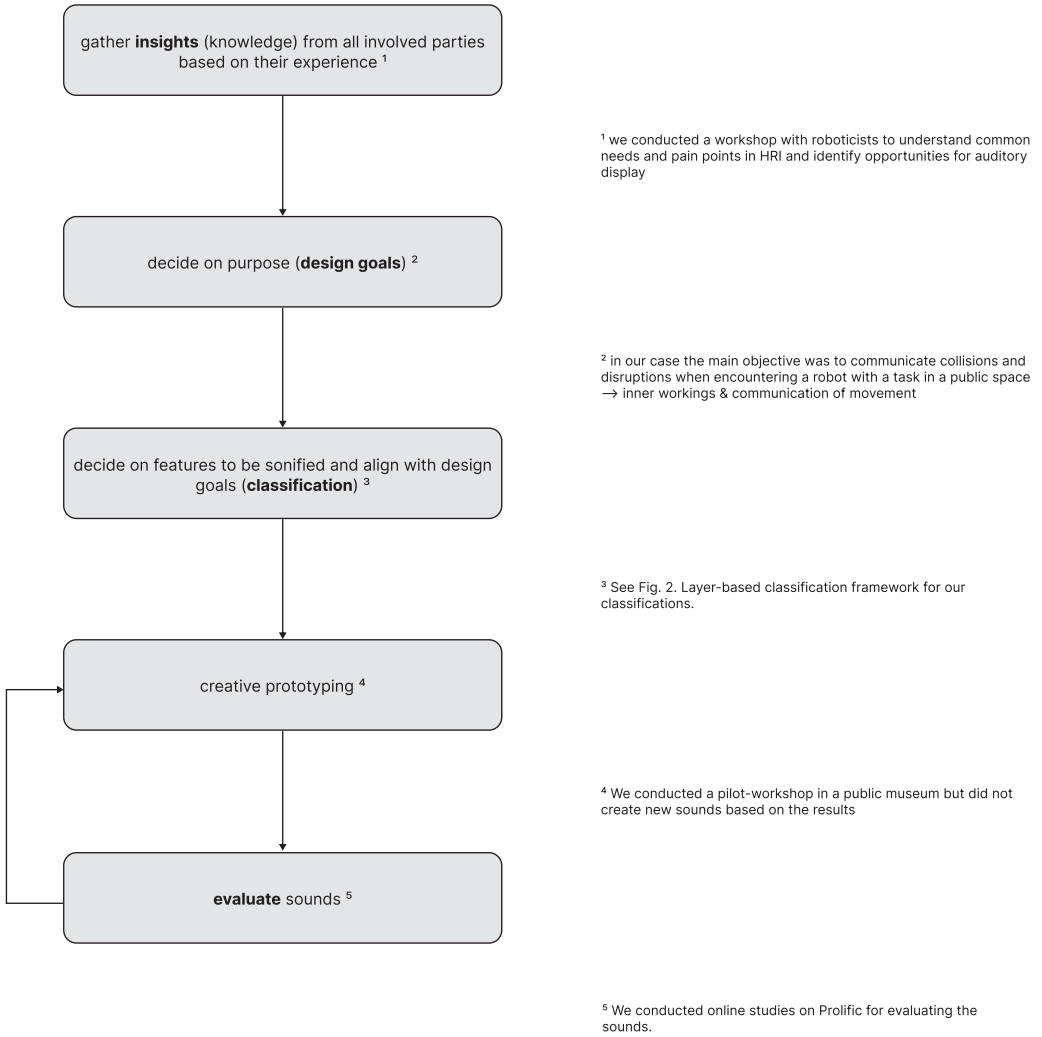

Fig. 11. Iterative conceptualisation, design, and evaluation process.

evaluation process, we suggest to consider some additional points when starting the process of developing an auditory display system:

- A clearly defined conceptual sonification framework supports learning and identification of auditory displays;
- Including experts in the conceptualisation phase for defining design goals is crucial for generating clear auditory displays:
  - Gather as many insights, such as requirements, pain points, requests, and visions as possible from roboticists and interaction designers.
- Decide on primary and secondary design goals together with stakeholders:
  - E.g., is the main objective to communicate internal processes (inner workings), upcoming or occurring actions (movement), limitations (abilities), or emotion? Map out *design journeys* that anticipate the HRI experience.

- Decide on which features to be sonified and align them with the design goals:
  - We suggest classifying features into groups based on the purpose of their communication (see *action, internal state, physical base* in Figure 2).
- Iterative creative prototyping processes improve the design quality and comprehensibility:
  - Include audio experts in the design process and provide tools for creative prototyping (e.g., custom synthesisers, vocal sketching methods);
  - Evaluate the results and restart creative prototyping based on evaluation.
- Mapping information about robots' internal states and intentions to audio parameters can be done and understood by non-experts:
  - Borrow from established and learned sound concepts;
  - Try to create either intuitive, learnable, or simple designs.
- Layering multiple information by chaining various audio parameters enables simultaneous multi-information display:
  - People are able to extract multiple information that is encoded into a single sound.
- White noise can be used as audio material base;
- Subtractive synthesis is a suitable method for generating different signals with characteristic tonal qualities, which can be layered; however, the layering of different information needs to be carefully evaluated, as our current results were not completely successful in this regard;
- Amplitude modulation is a comprehensible parameter mapping for displaying robot speed;
- Manipulations of the audio's frequency domain (filtering) can help display robot size;
- Enhancing frequencies that resonate within the human ear [8] can be used to convey urgency (and to draw attention);
- While the spatial dimension is very suitable for displaying directionality, pitch can be used either redundantly or complementarily when stereo audio playback is not perceived or generally possible;
- Multiple evaluation cycles should be planned and conducted:
  - Wizard-of-Oz studies can help map the *user journey* for people interacting with robots unintentionally (qualitative evaluation);
  - Online studies provide a way for recruiting large amounts of participants from diverse backgrounds (quantitative and possibly qualitative evaluation);
  - In-person evaluation can provide thorough user experience evaluation in lab environment; e.g., technology-based studies with virtual simulations on computer or Virtual Reality devices (qualitative and possibly quantitative evaluation);
  - Real-world studies are necessary before deploying robots with auditory displays to test different reactions of participants on robots based on designs.

7 CONCLUSION

In this work, we have shown that intentional nonverbal sounds can convey information about multiple features of a robot, both individually and simultaneously. The sound modality can thus be employed specifically for those scenarios where visual feedback is limited and where speech output is reserved for other interaction or simply not desired, such as people and robots co-existing and navigating in the same space. In these interaction contexts, the use of nonverbal auditory display can provide a useful additional modality for informing about robots' actions and intentions.

REFERENCES

[1] Claude Alain and Kelly L. McDonald. 2007. Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. *J. Neurosci.* 27, 6 (2007), 1308–1314.

[2] Jennifer Aydelott, Robert Leech, and Jennifer Crinion. 2010. Normal adult aging and the contextual influences affecting speech and meaningful sound perception. *Trends Amplif.* 14, 4 (2010), 218–232.
[3] Matthew P. Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The right kind of unnatural: Designing a robot voice. In *1st International Conference on Conversational User Interfaces*. 1–2.
[4] David Radford Bakker and Frances Heritage Martin. 2015. Musical chords and emotion: Major and minor triads are processed for emotion. *Cognit., Affect. Behav. Neurosci.* 15, 1 (2015), 15–31.
[5] T. Mark Beasley and Randall E. Schumacker. 1995. Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *J. Experim. Educ.* 64, 1 (1995), 79–93.
[6] Alisha Bevins and Brittany A. Duncan. 2021. Aerial flight paths for communication. *Front. Robot. AI* 8 (2021).
[7] Gabriele Bolano, Arne Roennau, and Ruediger Dillmann. 2018. Transparent robot behavior by adding intuitive visual and acoustic feedback to motion replanning. In *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. IEEE, 1075–1080.
[8] Editors of Encyclopaedia Britannica. 2020. *The Physiology of Hearing*. Retrieved from https://www.britannica.com/science/ear/The-physiology-of-hearing
[9] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *CHI Conference on Human Factors in Computing Systems*. 1–13.
[10] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proc. ACM Hum.-comput. Interact.* 3, CSCW (2019), 1–19.
[11] Elizabeth Cha, Naomi T. Fitter, Yunkyung Kim, Terrence Fong, and Maja J. Matarić. 2018. Effects of robot sound on auditory localization in human-robot collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction*. 434–442.
[12] Elizabeth Cha and Maja Matarić. 2016. Using nonverbal signals to request help during human-robot collaboration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16)*. IEEE, 5070–5076.
[13] David Creasey. 2016. *Audio Processes: Musical Analysis, Modification, Synthesis, and Control*. Routledge.
[14] Luke Dahl, Jon Bellona, Lin Bai, and Amy LaViers. 2017. Data-driven design of sound for enhancing the perception of expressive robotic movement. In *4th International Conference on Movement Computing*. 1–8.
[15] Diana Deutsch. 1995. *Musical Illusions and Paradoxes*. Philomel.
[16] Gaël Dubus and Roberto Bresin. 2013. A systematic review of mapping strategies for the sonification of physical quantities. *PloS One* 8, 12 (2013), e82491.
[17] Stefanie Duijndam, Nina Kupper, Johan Denollet, and Annemiek Karreman. 2020. Social inhibition and approach-avoidance tendencies towards facial expressions. *Acta Psychol.* 209 (2020), 103141.
[18] Judy Edworthy. 2011. Designing effective alarm sounds. *Biomed. Instrum. Technol.* 45, 4 (2011), 290–294.
[19] Judy Edworthy. 2013. Medical audible alarms: A review. *J. Am. Med. Inform. Assoc.* 20, 3 (2013), 584–589.
[20] Tuomas Eerola, Rafael Ferrer, and Vinoo Alluri. 2012. Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Percept.: Interdisc. J.* 30, 1 (2012), 49–70.
[21] Tuomas Eerola, Anders Friberg, and Roberto Bresin. 2013. Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. *Front. Psychol.* 4 (2013), 487.
[22] Alexander L. Francis and Howard C. Nusbaum. 2009. Effects of intelligibility on working memory demand for speech perception. *Attent., Percept. Psychophys.* 71, 6 (2009), 1360–1374.
[23] Emma Frid and Roberto Bresin. 2022. Perceptual evaluation of blended sonification of mechanical robot sounds produced by emotionally expressive gestures: Augmenting consequential sounds to improve non-verbal robot communication. *Int. J. Soc. Robot.* 14, 2 (2022), 357–372.
[24] Emma Frid, Roberto Bresin, and Simon Alexanderson. 2018. Perception Of mechanical sounds inherent to expressive gestures Of A nao robot-implications for movement soniication Of humanoids. In *Proceedings of the Sound and Music Computing Conference (SMC'18)*. Sound and Music Computing Network, Limassol, Cyprus, 43–51.
[25] Thomas Hermann, Andy Hunt, and John G. Neuhoff. 2011. *The Sonification Handbook*. Logos Verlag, Berlin.
[26] Guy Hoffman and Cynthia Breazeal. 2007. Cost-based anticipatory action selection for human–robot fluency. *IEEE Trans. Robot.* 23, 5 (2007), 952–961.
[27] Eun-Sook Jee, Yong-Jeon Jeong, Chong Hui Kim, and Hisato Kobayashi. 2010. Sound design for emotion and intention expression of socially interactive robots. *Intell. Serv. Robot.* 3, 3 (2010), 199–206.
[28] Qi Jiang and Atsunori Ariga. 2020. The sound-free SMARC effect: The spatial-musical association of response codes using only sound imagery. *Psychonom. Bull. Rev.* 27, 5 (2020), 974–980.
[29] Gunnar Johannsen. 2001. Auditory displays in human–machine interfaces of mobile robots for non-speech communication with humans. *J. Intell. Robot. Syst.* 32, 2 (2001), 161–169.
[30] R. A. Katz. 2015. *Mastering Audio: The Art and the Science*. Focal Press. Retrieved from https://books.google.se/books?id=P8QwMQEACAAJ

[31] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Front. Neurorobot.* 14 (2020), 105.

[32] Adrian Benigno Latupeirissa, Emma Frid, and Roberto Bresin. 2019. Sonic characteristics of robots in films. In *Sound and Music Computing Conference*. 1–6.

[33] Sébastien Le Maguer and Benjamin R. Cowan. 2021. Synthesizing a human-like voice is the easy way. In *3rd Conference on Conversational User Interfaces*. 1–3.

[34] Xiaofeng Li, Robert J. Logan, and Richard E. Pastore. 1991. Perception of acoustic source characteristics: Walking sounds. *J. Acoust. Soc. Am.* 90, 6 (1991), 3036–3049.

[35] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *CHI Conference on Human Factors in Computing Systems*. 5286–5297.

[36] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say hello? Personality impressions from brief novel voices. *PLoS ONE* 9, 3 (2014), e90779.

[37] Conor McGinn and Ilaria Torre. 2019. Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. IEEE, 211–221.

[38] Dylan Moore, Nikolas Martelaro, Wendy Ju, and Hamish Tennent. 2017. Making noise intentional: A study of servo sound perception. In *12th ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. IEEE, 12–21.

[39] Roger K. Moore. 2017. Appropriate voices for artefacts: Some key insights. In *1st International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots*.

[40] Roger K. Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer, 281–291.

[41] Uran Oh, Shaun K. Kane, and Leah Findlater. 2013. Follow that sound: Using sonification and corrective verbal feedback to teach touchscreen gestures. In *15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.

[42] John J. Ohala. 1983. Cross-language use of pitch: An ethological view. *Phonetica* 40 (1983), 1–18.

[43] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is human-like? Decomposing robots' human-like appearance using the Anthropomorphic roBOT (ABOT) database. In *ACM/IEEE International Conference on Human-robot Interaction*. 105–113.

[44] Cambridge Street Publishing. 2020. *Tutorial for the Handbook for Acoustic Ecology*. Retrieved from http://www.sfu.ca/sonic-studio-webdav/cmns/Handbook%20Tutorial/Filters.html

[45] Frederic Anthony Robinson, Oliver Bown, and Mari Velonaki. 2020. Implicit communication through distributed sound design: Exploring a new modality in human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction*. 597–599.

[46] Frederic Anthony Robinson, Mari Velonaki, and Oliver Bown. 2021. Smooth operator: Tuning robot perception through artificial movement sound. In *ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[47] Elena Rusconi, Bonnie Kwan, Bruno L. Giordano, Carlo Umilt, and Brian Butterworth. 2006. Spatial representation of pitch height: The SMARC effect. *Cognition* 99, 2 (2006), 113–129. DOI: https://doi.org/10.1016/j.cognition.2005.01.004

[48] Markus Schwenk and Kai O. Arras. 2014. R2-D2 reloaded: A flexible sound synthesis system for sonic human-robot interaction design. In *23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'14)*. IEEE, 161–167.

[49] Moondeep C. Shrestha, Ayano Kobayashi, Tomoya Onishi, Hayato Yanagawa, Yuta Yokoyama, Erika Uno, Alexander Schmitz, Mitsuhiro Kamezaki, and Shigeki Sugano. 2016. Exploring the use of light and display indicators for communicating directional intent. In *IEEE International Conference on Advanced Intelligent Mechatronics (AIM'16)*. IEEE, 1651–1656.

[50] Valerie K. Sims, Matthew G. Chin, Heather C. Lum, Linda Upham-Ellis, Tatiana Ballion, and Nicholas C. Lagattuta. 2009. Robots' auditory cues are subject to anthropomorphism. In *Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. SAGE Publications Sage CA, Los Angeles, CA, 1418–1421.

[51] Sichao Song and Seiji Yamada. 2018. Bioluminescence-inspired human-robot interaction: Designing expressive lights that affect human's willingness to interact with a robot. In *13th ACM/IEEE International Conference on Human-Robot Interaction (HRI'18)*. IEEE, 224–232.

[52] Ana Tajadura-Jiménez, Maria Basia, Ophelia Deroy, Merle Fairhurst, Nicolai Marquardt, and Nadia Bianchi-Berthouze. 2015. As light as your footsteps: Altering walking sounds to change perceived body weight, emotional state and gait. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. 2943–2952.

[53] Rie Tamagawa, Catherine I. Watson, I. Han Kuo, Bruce A. MacDonald, and Elizabeth Broadbent. 2011. The effects of synthesized voice accents on user perceptions of robots. *Int. J. Soc. Robot.* 3, 3 (2011), 253–262.

[54] Hamish Tennent, Dylan Moore, Malte Jung, and Wendy Ju. 2017. Good vibrations: How consequential sounds affect perception of robotic arms. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'17)*. IEEE, 928–935.

[55] Alexander Todorov. 2008. Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Ann. New York Acad. Sci.* 1124, 1 (2008), 208–224.
[56] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *ACM/APA Technology, Mind, and Society Conference*. 1–6.
[57] Ilaria Torre, Adrian Benigno Latupeirissa, and Conor McGinn. 2020. How context shapes the appropriateness of a robot's voice. In *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'20)*. IEEE, 215–222. DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223449
[58] Ilaria Torre and Sébastien Le Maguer. 2020. Should robots have accents? In *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'20)*. IEEE, 208–214.
[59] Gabriele Trovato, Renato Paredes, Javier Balvin, Francisco Cuellar, Nicolai Bæk Thomsen, Soren Bech, and Zheng-Hua Tan. 2018. The sound or silence: Investigating the influence of robot noise on proxemics. In *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. IEEE, 713–718.
[60] Bruce N. Walker and Michael A. Nees. 2011. Theory of sonification. *Sonif. Handb.* 1 (2011), 9–39.
[61] Michael L. Walters, Dag Sverre Syrdal, Kheng Lee Koay, Kerstin Dautenhahn, and René Te Boekhorst. 2008. Human approach distances to a mechanical-looking robot with different robot voice styles. In *17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'08)*. IEEE, 707–712.
[62] Sarah Wilson and Roger K. Moore. 2017. Robot, alien and cartoon voices: Implications for speech-enabled systems. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR'17)*. 40–44.
[63] Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst. 2016. Review of semantic-free utterances in social human–robot interaction. *Int. J. Hum.-Comput. Interact.* 32, 1 (2016), 63–85.
[64] Lisa Zahray, Richard Savery, Liana Syrkett, and Gil Weinberg. 2020. Robot gesture sonification to enhance awareness of robot status and enjoyment of interaction. In *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'20)*. IEEE, 978–985.
[65] Brian J. Zhang, Nick Stargu, Samuel Brimhall, Lilian Chan, Jason Fick, and Naomi T. Fitter. 2021. Bringing WALL-E out of the silver screen: Understanding how transformative robot sound affects human perception. In *IEEE International Conference on Robotics and Automation (ICRA'21)*. IEEE, 3801–3807.

Received 25 May 2022; revised 4 March 2023; accepted 22 June 2023