3

Learning from Humans: How Research on Vocalizations can Inform the Conceptualization of Robot Sound

Hannah Pelikan and Leelo Keevallik

3.1 Introduction 33
3.2 Robot Sound Design 35
3.3 Vocalization in Human Interaction 36
3.4 Method 37
3.5 Applying Insights on Human Vocalizations to Robots 38
3.5.1 Sounds are Semantically Underspecified 39
3.5.2 Sound Production is Embodied 41
3.5.3 Sound can be Adapted for Complex Participation 45
3.6 Discussion and Implications 49
3.6.1 Meaning as Potentials 49
3.6.2 Sound and Multimodality 50
3.6.3 Variable Form and Reflexive Adaptation to Multiple Participants 50
3.6.4 Designing Sound for Interactional Sequences 51
3.7 Conclusion 52
Acknowledgment 52
Bibliography 52

3.1 Introduction

Beeps and whirrs are just some examples of sounds that robots produce. Such sounds are not exclusive to robots: non-lexical vocalizations such as *ouch*, *wohoo*, and *tadaa* have recently been shown to be an important and effective element of human-human communication: people consistently make sense of

DOI: 10.1201/9781003320470-3

these sounds when interacting [36]. Taking an interactional perspective, this chapter provides examples of how insights on human vocalizations and prosody can inform the analysis and design of robot sound.

Human vocalizations feature special prosodies and convey important information on the state of its producer [34] and some of them are understood universally [9]. They function at the margins of human language, provide affordances for cross-cultural understanding, and crucially facilitate interaction. In our chapter we discuss how vocalizations and prosodies that index emotional states such as sighs and moans indicating disappointment [23, 25] or sensorial and proprioceptive experiences such as the vocalization of gustatory pleasure with a *mmm* [84] or the sonification of body movement [32] are intuitively understood and acted on by humans in concrete activity contexts. Even though semantically less specific than lexical words, vocalizations are an important interactional resource precisely because they are sufficiently flexible to be adaptable to a variety of contexts. In this chapter we argue that their interactional function and special properties can inspire design of robot sounds for interaction with humans.

In robots, similar sounds have been glossed as "semantic-free utterances" [86], including gibberish speech, paralinguistic aspects such as backchannels, voice quality or pitch, and "non-linguistic" sonifications such as beeps. The majority of studies have evaluated them through questionnaires, whereas knowledge of how people interpret them in real-time interaction is still lacking. While semantic-free utterances may stand out as deliberately designed for communicational purposes, even "consequential sounds", originating from the physical embodiment of the robot contribute to how a robot is perceived [46] and may be interpreted in interaction. Both of these categories feature in our work, as we are interested in designing recognizable robot behavior that displays interactional affordances in an intuitive and implicit way [30, 48]. Our aim is to enrich the sounding opportunities for robots by taking inspiration from recent findings regarding the use of vocalizations in human-to-human interaction.

We will discuss how insights on various aspects of human vocal behavior can be used as a resource for designing recognizable robot behavior that makes robots more expressive and natural to interact with. First, we demonstrate how one and the same vocalization or sound can be interpreted differently in different contexts: it gains situated meanings depending on the exact interactional context that it occurs in. This problematizes the design of sound all too rigidly for specific goals but also points to opportunities for sounds to be used for more flexible outcomes depending on when exactly they are produced, i.e. their sequential context in a particular activity. Second, we discuss how sounds are tightly connected to embodiment and movement, thus not to be designed as an independent mode. Third, we highlight some qualitative differences between sounds, such as single versus repeated sounds, and how they can achieve coordination between several agents in interaction. We end the chapter by briefly presenting a method for designing and prototyping the timing of

non-lexical sound based on video recordings of concrete interactions. Drawing on three transcribed video recordings from human interaction and three from human–robot interaction, we contribute lessons that could inform the methods of sound design for robots.

3.2 Robot Sound Design

Evaluating users' interpretations is a central concern for robot sound design, usually with the goal of ensuring that they are in line with the designer's intentions [5, 13, 46, 64, 69, 73, 78]. This is typically done by playing audio or video recordings to study subjects and asking them to rate the sounds along different scales. Few studies have taken a different approach, testing sounds in interactional contexts. Using different video scenarios, such as a robot being hit or kissed just before a sound was played, Read and Belpaeme [65] demonstrated that the situational context influences how sounds are interpreted. Others have explored how users interpret sound during live interactions with a robot in the lab [54]. In real world encounters the evaluation of robot sound has to deal with practical issues such as whether sound can even be heard in a particular environment, such as on a busy road [45, 59].

While some studies focus on audio as a separate modality, sound is often tightly intertwined with a robot's material presence. Whether sound is suitable or not may depend on the specific embodiment of the robot [41, 66] and people may prefer different robot voices depending on the context in which they are used [80]. Consequential sound and musical sonifications are naturally paired with movement [13, 29, 69, 73, 78]. These can even accomplish interactionally relevant actions such as managing delay, for instance through a combination of cog-inspired sounds and turning away from the human interlocutor [57]. Work on backchannels and affect bursts in robots combines facial expressions with vocalizations [54]. The interplay of different modalities has been given particular attention in the design of emotion displays [79].

Concerning how a sound can be varied, studies have compared the use of beeps versus words [12] and explored how variations in intonation, pitch, and rhythm influence the interaction [11, 63]. While the majority of studies focus on evaluating specific robot sounds, more recent work has started to formulate general design principles, reflecting among others on how sound could be varied and modified throughout longer interactions [68].

In short, audio is typically not designed as a standalone resource but is intertwined with another resource such as a movement or a facial expression – but rarely involves a range of modalities at the same time. While much work has focused on designing a set of particular animations, some studies are exploring how sound can be varied throughout an interaction. Importantly, studies of robot sound in real-time interaction remain rare, a gap which our work tries to address.

3.3 Vocalization in Human Interaction

While human interaction can be markedly centered around language, non-lexical vocalizations provide different affordances from lexical items, as they are, a) underspecified (vague in meaning), b) part of complex multimodal conjectures that may reveal information about the mental or physical state of the body producing them, and c) relatively variable in their form. In addition, prosodic aspects of all vocal delivery contribute specific meanings. All of these aspects may be useful for robot sound design.

While lexical items have meanings that are traditionally captured in dictionaries, vocalizations such as pain cries (*uuuw*), strain grunts, or displays of shivering (*brrr*) do not generally figure there. They are semantically underspecified, and we understand them in the context of someone hitting their head on a microwave door, lifting a spade of manure, or wading into cold water. Indeed, they lack propositional meaning but humans make sense of them in concrete activity contexts where they can also take specific action: parental lipsmacks encourage infants to eat [85], a strain grunt recruits others to rush to help [33], and a pain cry shows that students have understood a self-defence technique [82]. Vocalizations such as clicks may be used to hearably not say anything, thus leaving assessments implicit [50]. This vagueness makes the vocalizations usable in a broad variety of functions, while it is also true that the meaning of any word is determined by its context of use.

As is clear from the brief examples above, vocalizations are necessarily embedded in multimodal trajectories of action. The accomplishment of social action is intimately tied to people's ability to behave in a comprehensible manner and to competently interpret these very behaviors in entirety, not merely separating out a single stream of information, such as contained in the vocal tract sound. If someone sniffs and gazes away, it may make evident that the person will not speak at this point in conversation [24]. When a glass is simultaneously lifted to the sniffer's nose, they may be publicly demonstrating access to a source of a smell, such as in beer tasting sessions [44]. When someone gasps there is reason to check where their gaze is for a spill or potentially broken glasses [2]. A *mmm* with a specific rise-fall intonation after taking a spoonful of food is typically interpreted as expressing gustatory pleasure [84]. In short, from a human interaction perspective, sound is not a standalone resource but gets interpreted in combination with other aspects such as movement, facial expressions and so forth. Many vocalizations also express sensory immediacy, such as just having smelled, tasted, or dropped something. A reaction to pain or discomfort needs to be immediate in order to be deemed visceral [83]. A Finnish *huh huh* (a double heavy outbreath) is uttered at transitions from strenuous activities that have just come to an end [55]. Notably, emotion displays such as "surprise" or "appreciation" feature distinct embodied aspects, as was shown very early by Goodwin and Goodwin [19]. A display of "disappointment" may

be performed through a particular interjection in combination with a distinct pitch movement such as the English "oh" [7] or by a visible deflation of the body [6]. In co-present activities in particular, human actions are performed as, and interpreted through, multimodal displays.

Vocalizations can furthermore be adjusted in many ways: repetition, loudness, lengthening, or sound quality. A moan expressing disappointment at a boardgame move can include and combine any back vowel, i.e. *a*,*o*, and *u* and feature variable lengths [25]. Rhythmicity (and arrhythmicity) is a way to exhibit affiliative (or disaffiliative) relations between turns. Vocalized celebrations can be performed in chorus [77]. All of this means that vocalizations can be adjusted to their sequential and action environment and interpreted flexibly [36]. Work in robotics has often glossed parts of this variability under the category of prosody [86], including elements such as loudness and pitch curves. For designing robot sound, the variability and flexibility can be of particular interest, since it means that they need not emulate a very specific human vocalization to be understood as meaningful, and can be adapted to a variety of contexts.

To summarize, research on vocalizations in interaction has resulted in a better understanding of the contextualized methods and resources participants use for making sense of each other. These methods include attention to not only the position of the item in an action sequence but also its indexical aspects, exact timing in relation to current bodily action, the articulatory and prosodic features of the utterance, as well as material, spatial, and other contextual matters in the local environment. In this chapter, we will proceed to use the same method to target robot sounds in interactional settings and show that it can inform new ways of thinking about those.

3.4 Method

We take an Ethnomethodology and Conversation Analysis (EMCA) approach to studying interaction. EMCA originated from sociology, with an initial focus on human spoken interaction in phone calls [72]. Fairly soon it made some of its most impactful advances within anthropology, self-evidently using video rather than merely a tape-recorder to capture human interaction holistically [16,18] and in linguistics, pioneering a new branch that came to be called interactional linguistics [8,49]. In these areas, close studies of video-recorded interaction in naturally occurring situations have provided a solid ground for revealing the underlying organization of human collaboration. EMCA has been successfully applied to study human-computer interaction [3,38,62,76] and interaction with robots [58,61,81].

In this chapter we draw on video recordings from a variety of settings. Participants in all recordings have given their consent to be videorecorded and

we only show data from the persons who agreed to share their videos. The examples from human-human interaction have been previously published [31, 34, 85]. The examples from human–robot interaction stem from two previously collected corpora: A Cozmo toy robot in Swedish and German family homes [57, 60], and autonomous shuttle buses on public roads, on which we also tested own sound designs [59]. Please see the original publications for details on the video corpora.

EMCA treats video recordings as data, beginning the analysis with a multimodal transcription. Transcription is done manually because it is at a level of detail that cannot (yet) be handled by automated transcription and image analysis software, crucially because all the locally relevant details cannot be predicted. Several transcription conventions are available, and in this article we follow the most established and readability-focused Jeffersonian transcription system for verbal utterances [21] and sounds [56], while we use Mondada's transcription system [43] for tracing embodiment and movements at tenth-of-a-second intervals. A close transcription enables the analyst to unpack how interactions evolve in real time.

EMCA methodology is specialized to find "order at all points" [71, p.22], revealing how people systematically calibrate their behavior to each other and to machines, even though it may look disorderly at first sight. Analytic questions include: How do others respond to what someone (or a robot) just did? What understanding of the prior action do they display in their response? And what opportunities and expectations do they create through their own subsequent action? Drawing on detailed transcripts, the researcher typically looks at each turn in an interaction, trying to identify what is accomplished by it. Our particular study objective is to use this method for both human-to-human and human–robot interaction in order to locate similarities and identify possible sources of inspiration.

3.5 Applying Insights on Human Vocalizations to Robots

In this section we present examples of video recorded interactions to highlight three main aspects of how humans make sense of vocalizations among themselves and how they interpret robot sounds: meaning potentials, the multimodal embeddedness of sounds, and their flexible production. In each section, we highlight how the findings can be leveraged by roboticists to create sounds designs that are more in line with human expectations.

3.5.1 Sounds are Semantically Underspecified

When robot sound is evaluated in user studies, the aim is usually to validate that a specific sound can convey the meaning intended by design, for instance a particular emotion [5, 28]. We would like to highlight that this is important for narrowing down the overall range of potential meanings, but this can never nail down a sound to one fixed interpretation. In everyday interaction, sound is always interpreted in the specific local context that it occurs in, which can be an asset for robot sound design: Sequencing (the way in which behavior by different participants follows one after another) and context are essential for how humans make sense of each other's actions and moves. Importantly, all meaning is negotiated in interaction. Utterances get their meaning specified in a local context: Consider the word "nice", which according to the Merriam Webster Dictionary can mean different things, ranging from "polite, kind", to "pleasing, agreeable", to "great or excessive precision". While a dictionary can provide a number of example phrases, it is impossible to list what exactly it would mean at any moment in interaction [22, p. 143]. Similar arguments can be made for many words, Norén and Linell [47] for instance provide a parallel discussion on the Swedish word *ny* "new". Thus, even though lexical items are seemingly more fixed in their meaning, it is ultimately only a question of degree. It is only possible to establish meaning potentials both for vocalizations and words [39].

Consider the extract in Figure 3.1, where an infant produces a short vocalization *mmh* [85, p.248-250]. It is easy to imagine that a *mmh* by an infant could mean anything from satisfaction to dawning unhappiness, depending on the circumstances, while in this case, taken from a mealtime, it basically emerges as a request. The infant produces a *mmh* sound with rising pitch and attempts to grab a bite of food on a tray (Figure 3.1, line 1). Mum gazes at the infant and interprets it as "wanting" that specific bit (l.3). Making sense of all kinds of sounds is an inherent aspect of human interaction and happens with regard to the activity context, gesture, prosody, etc. A "mmh" in a different context, such as after a question, can easily be interpreted as a positive answer [15].

Similarly, sound designed for robots gets assigned a specific situated meaning when a robot plays the sound at a particular moment in interaction. Consider an animation by the Cozmo robot which the designers intended to mean "happy to meet you" (see Figure 3.2) highlighting success after a relatively lengthy face learning activity: When played at the dinner table in a context when Cozmo has just been offered a sip of beer, the animation may get understood as "Cozmo likes beer". Cozmo is programmed to play this animation at the end of every successful face learning sequence. After saying the user's name twice, the robot launches a sequence of quick sounds, while showing smiley eyes on its display and waving its forklift arms. In our corpus the animation typically gets responded to by smiles and petting the robot. We could thus observe that the sound gets treated as closing the face learning sequence, as designed

```
1 INF   mmh?#
  inf   >>gaze on a piece of food
  img       #img1
2       (0.5)^(0.4)
  mum         ^reaches over to the tray-->>
3 MUM   want that bit over there,
4       (0.3)&(0.9)
  inf         &lifts a handful
```

Img 1. Infant utters 'mmh' and reaches for a bit of food.

FIGURE 3.1
Finger-licks with and without lip-smacks (Lewis002_0515_LSS30 & 31).

for. However, we also recorded an example in which the sound-animation gets interpreted in a quite different way, with participants in a family home formulating their understanding of the animation as "full agreement" to a question whether Cozmo likes beer [60].

The extract in Figure 3.3 provides a transcript of the interaction in which a couple, Ulrich and his wife, meet the robot. The robot has been scanning Ulrich's face when he gets impatient and proposes a different thing. Ulrich proceeds to ask "do you like Giesinger beer?" (l. 01), while grabbing his glass. Cozmo says *Ulrich* (l. 03) while Ulrich is speaking, but it gets ignored by everybody present. In this moment Cozmo plays a *oaaaaow* sound that resembles a "wow" (l. 06), which Ulrich responds to with the German change-of-state token *ah* (l. 07) (this resembles an "oh" in English). Cozmo's second formulation of Ulrich's name drowns in laughter (l. 08-11). When Cozmo then finally plays the happy animation (l. 16), Ulrich interprets it as "oh yes, full approval" (l. 17-19). This example highlights very vividly that designers cannot ultimately define how a robot sound gets interpreted in the specific local context of a dynamic family life. While a verbal utterance such as "happy to meet you" would simply not fit as a response to the question "do you like beer", the sound (as part of a display of happiness) gets interpreted as a fitted, positive response about the beer.

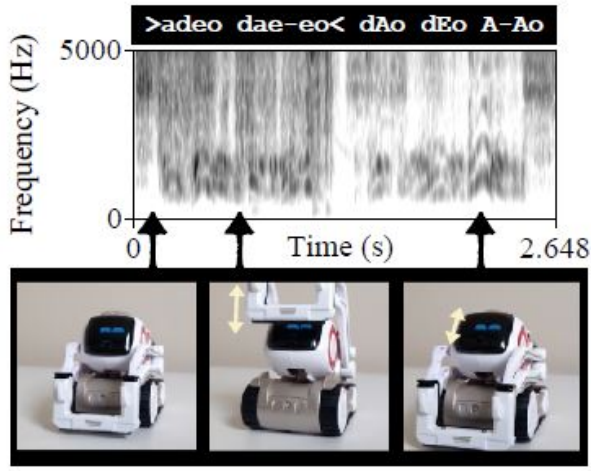

FIGURE 3.2
Cozmo's happy animation at the end of a face learning sequence (adapted from [60]).

The above extracts highlight how humans make sense of vocalizations and sounds, treating them not as pre-defined but instead as meaning different things depending on the specific local context that they are uttered in, and the multimodal aspects that go with them. The examples highlight the importance of repeatedly deploying and testing prototypes in real-world contexts, in order to learn about the range of possible associations by humans. This can also help to identify at what moments narrow meaning potentials are necessary and when interaction would be eased by more interpretative flexibility. Featuring broader meaning potentials, sounds invite broader interaction possibilities and can function efficiently in different interaction contexts.

3.5.2 Sound Production is Embodied

Some of the studies that explore the interplay of robot sound with other modalities focus on identifying which modality is most effective for conveying a specific message or robot state, such as happiness [79]. In contrast, work in sonification and perception generally tends to treat sound as a multisensorial and multimodal phenomenon [4, 14], in which these modalities are not ranked but contribute to an impression that may be more than the sum of its parts. In interactional sense-making, modalities are necessarily intertwined. As we highlighted in the extracts in Figure 3.1 and 3.3, neither the infant's nor Cozmo's sounds stand isolated in defining their meaning. They are interpreted alongside body movements and facial expressions.

In human interaction, vocalizations generally acquire their meaning from their context and ongoing bodily activities. Being essentially embodied, they

```
01 HUS  ma[gst du k- magst du giesin]ger bier?=
           do you like do you like Giesinger beer?
                      ((beer from local brewery))
02 COZ    [u::lrich:::              ]
03 WIF  =[(h)]
04 JON    [k(h)]a
05 RES  e(h)[h]
06 COZ      [o]aaaaow
07 HUS  A[H    ]
        oh
08 RES    [ha↑ha]↓ha[haha]
09 HUS              [haha][ha[haha hahaha]ha  ha ha he]
10 WIF                    [haha   (h)(h)(h)  (h)(h)(h)]
11 COZ                       [°°   rich°°]
12 HUS  &[.hhh [hehe [ha]heha .h] (h)] (h)   ]
13 WIF   [(h)  (h)(h)(h) (h) (h)]
14 RES   [(h)(h)  (h)(h)      (h) (h)]
15 HOS         [↑hi: ↓ha]
16 COZ                [>adeo dae-eo< dAo deo A]-Ao=
17 HUS  =ahJA#
         oh yes
   img       #img1
18      (1.6) ((more joint laughter))
19 HUS  volle zustimmung#
        full approval
   img       #img2
  ((joint laughter))
```

Image 1. Cozmo just played the "happy" anmation.

Image 2. Ulrich translates it as 'full approval'.

FIGURE 3.3
Full approval (A1 [00:38-00:50]).

can instruct movement. An example is shown in the extract in Figure 3.4, which is taken from a Pilates class. The teacher has just demonstrated a new exercise called helicopter that involves moving legs and hands in a circular motion to opposite sides while balancing on the buttocks. When the students try it out, the teacher produces a long nasal sound in combination with a large gesture to accompany them [34]. The pitch trace is shown above the transcript line.

To start with, the teacher times the beginning of the exercise by uttering a slightly lengthened *ja* "and" while simultaneously launching a two-hand gesture that shows the required circular movement of the legs (Figure 3.4, Images 1-2). As she dips into the shape, she produces the strained vocalization that ends in a very high pitch (Figure 3.4, l. 1), marking the limbs' arrival

FIGURE 3.4
Pilates I.

back at the top position (Figure 3.4, Image 3). The teacher's bodily-vocal performance makes the exercise visually and audibly available, highlighting the shape, trajectory, as well as the ostensible proprioceptive experience of the exercise: the strain and its temporal extension are illustrated through her voice. The students toiling each at their own pace show that they attend to the teacher's production as an instruction.

This instance illustrates how one person's vocalizing in combination with their embodiment can have an immediate impact on other people moving. Sounding practices are usable for scaffolding others' activities (as above) as well as coordinating bodies together, such as a collaborative lift of a heavy object [33], making others move in a certain manner [32], and achieving synchrony in movement [26]. Humans may raise their voice [67] or increase repetition tempo, in order to persuade others to comply [52]. There is thus a wide range of ways how sounds embedded in multimodal action can be used for coordination across participants. Needless to say, a *qnnn* sound might mean very different things in other contexts.

The lack of comparable expressivity for robots is often attributed to the difficulty of designing a convincing coordination of different robot resources, such as movements, facial expressions, and sound. Depending on their

morphology, robots may have different resources available than humans, including for instance colored lights or vibration [40, 75]. Robots have different bodies than humans and it is unclear whether they could meaningfully instruct fine-grained physical movement such as in the Pilates example, but there is nevertheless a general point to be learnt here. Sounds are not heard and made sense of in isolation, but humans are used to interpreting them in combination with movement, gestures and other nonverbal, embodied behavior. With Cozmo, an *uuuuuu* sound only becomes meaningful as a greeting when paired with a lifted head and large eyes on its screen-face, as we illustrate in the extract in Figure 3.5.

```
01 COZ  quack +chrrrrrtrr((motor sound))
   coz         +drives forward-->
   coz                 ±lifts head-->

02 COZ  u#u±udeo+:
   coz  -->±
   coz       -->+
   img   #img1

03      (0.6)

04 MOM  °j↑a↓a°
        yes

05 COZ  ±ch+r[r +]# ((motor sound))
06 SON        [(h)]
   coz  ±lifts head-->>
   coz     +turn+
   img             #img2

07      (0.6)

08 COZ  +chr + ((motor sound))
   coz  +turn+

09      (0.4)

10 MOM  hej
        hello

11      +(0.3)     +
   coz  +lifts fork+

12 COZ  ±↑uuuuuu#
   coz  ±large eyes-->
   img           #img3

13 MOM  H↑E±:J
        hello
   coz  -->±
```

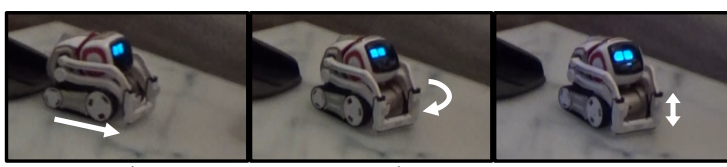

FIGURE 3.5
Cozmo's sound animation is interpreted as a greeting (FAM6 Day 3 [01:28-01:36]).

Cozmo has just been switched on, has left the charger and played an animation that resembles "waking up". It drives forward and raises its head, playing a gibberish *uuudeo* sound (l. 02), which is acknowledged by mother and son with a *ja* "yes" in Swedish (l. 04) and a laughter particle (l. 06). The robot turns and lifts its head, now facing the mother, who greets the robot by saying *hej* "hello" (l. 10). The robot briefly moves its forklift arms and displays large, "cute" eyes on its display while playing an *uuuuuu* sound with rising intonation (l. 12). The mother treats this as a response to her greeting and greets the robot once more, now with increased prosodic marking (l. 13). It is the activity context of Cozmo just waking up that this precise multimodal production is interpreted as a greeting by a human and responded to affectively. Paired with a backwards driving movement and a raised forklift, *uuuuuu* could indicate something else, for instance surprise.

In short, sound should be considered as a multimodal and multisensory phenomenon. Having demonstrated how humans tightly intertwine body and voice, we want to encourage designers to pay particular attention to this intertwinement. Studies that focus on how various resource combinations are interpreted in environments of everyday life can be particularly informative for robot design, helping to evaluate for instance how sound draws attention to particular movements.

3.5.3 Sound can be Adapted for Complex Participation

In addition to the sequential interpretation, pitch movement, and embodied nature of human sounding practices, we would like to highlight a further aspect: sounds and words can be repeated with particular prosodies and thereby convey persuasion to follow a contextualized request for coordination. Vocal devices can suggest actions, addressing not only one person at a time but also dealing with complex participation frameworks involving several people [17, 20].

We may consider another example from Pilates training in the extract in Figure 3.6, where the teacher is asking the students to roll up and balance on their buttocks [31]. She first provides the instruction to stay up (l. 1-2) and then repeats the word "hold (it)" seven times (l. 3-4), while the students are rolling up each at their own tempo (l. 1-4). The words are uttered at a high speed (indicated with >< in the transcript), with sounds floating together almost to the point of the words not being recognizable. The pitch trace is marked above the transcription line. With her level prosody on the first five items (l. 3) she is indicating that she will continue to provide this instruction, while on the last two repetitions her pitch rise already projects an end (l. 4), which is informative for all the students in the class. Furthermore, she coordinates her final repetition minutely with the last student (marked in a circle) arriving in the balanced position.

This example shows how humans mutually coordinate actions across multiple participants: the teacher instructs while she is also accommodating to the students moves, the students comply with the teacher's instruction and the ones who have arrived early wait for the others as well as the teacher who is to

1 Teacher **ni*i, jääge üles^se, leidke tasakaal,**
okay, stay up, find (your) balance,
stuA,B *stay up
stuC,D ^stay up

2 **jääge ülesse#**
stay up
img #1

3 **>püsi&ge* püsige püsige püsige püsige**
hold it hold it hold it hold it hold it
stuE &stays up
stuF *stays up

4 **püsige< püsig#e?**^
hold it hold it
stuG ^stays up
img #2

Image 1. Students at various stages of the exercise.

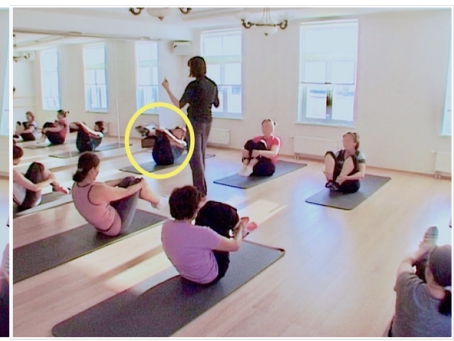

Image 2. Last student arriving.

FIGURE 3.6
Pilates II.

produce a next instruction. Different participants need to do different things to coordinate with others: while some need to stay in position, others have to deploy abs to get into alignment with the class (comp. Images 1 and 2 in Figure 3.6), and the teacher is merely using her voice and gestures to coordinate with students. The teacher's repetition indicates that the required action by others or its quality has not yet been reached [42, 51, 74]. It also highlights that the message is less about the semantic meaning of the word itself than about the action that is supposed to be accomplished and coordinated with others.

Adapting these findings to robots, it is important to notice this mutual reflexivity and continuous adjustment in human-human interaction. While robots may not reach the same level of fine coordination, they can likewise repeat sounds and preferably do so in a manner that is minutely adjusted to other participants' actions. In the following extract we look at an EasyMile EZ10 autonomous shuttle bus for public transport, on which we tested a range of sounds in live traffic. The bus drives on invisible tracks and is often stopped by cyclists and pedestrians that are getting too close, and we were interested in exploring ways that could help the bus to maintain traffic flow by asking other road users to keep a distance [59]. The researcher acts as a Wizard of Oz, playing pre-recorded samples on top of the buse's own soundscape through a Bluetooth speaker. In the extract in Figure 3.7 she is triggering three saxophone rolls with rising intonation while the bus is moving forward.

The bus leaves its designated stop (Figure 3.7, l. 01) and approaches a crossing where it often triggers unnecessary emergency braking when people get too close. Two cyclists are approaching the bus on its left in the same lane, and a pedestrian walks toward the crossing from the right. The wizard triggers a first sound (l. 02), a saxophone riff with rising pitch, inspired by question intonation. Both the cyclists and the pedestrian are still relatively far away but seem to be aware of the approaching bus evident in their gaze and head orientation (l. 03). As the groups are moving closer toward the approaching bus, the wizard triggers another sound (l. 04). While the sound is playing, the right cyclist clearly starts gazing at the bus (l. 04), displaying an orientation to the sound. The left cyclist immediately moves further toward the left, onto a sidewalk lane (l. 04, Image 1). Soon after, the right cyclist also steers away (l. 05, Image 2). Meanwhile, the pedestrian has been slowing down their steps, but keeps approaching the intersection (l. 05). The wizard triggers a third sound (l. 06), drawing the gaze of the right cyclist once more (l. 06), who then moves even further toward the left, onto the sidewalk lane (l. 07). The pedestrian who has slowed down further (Image 3) now also gazes at the bus (l. 07), and finally stops completely, until the bus has passed.

We showed how different parties in interaction mutually adjust to each other, even though they may maintain asymmetrical roles, such as a bus being on invisible tracks or when one person is officially instructing the others. In both examples, repetition of a sound fitted to the current movement worked as a tool for continuous responsivity to others, and it seems to achieve the wanted outcome: mutual attention and adjustment. In the example with humans exercising, each person was completing the exercise at their own speed. In the example with the autonomous bus, everyone needed to act slightly differently, depending on their activity trajectory (walking, cycling, or driving), which results in the whole situation being mutually coordinated. Notably, this reflexive adjustment was in the current instance achieved by a human being and not an automated machine, but it provides an example of where robot sound design could be headed, dynamically adjusting sound to an evolving situation.

```
1       |(1.5)
  bus   |leaves bus stop->>

2 WIZ   da↑deep

3       (2.3)*(1.7)     *(1.5)+(0.3)*(0.4)  +(0.4)
  cycR      *gazes fwd*gaze cycL--*gazes fwd-->
  ped                       +glance to bus+

4 WIZ   da*↑deep%
  cycR -->*gazes at bus-->
  cycL          %turns onto sidewalk-->

5     (0.1)†(0.3)#(0.7)•(0.5)#(0.3)*(0.1)•(0.7)%
 ped       †slows down steps-->
 cycR                  •steers away------•
 cycR                             -->*
 cycL                                      -->%
 img              #1         #2

6 WIZ   da↑deep*
  cycR        *gazes at bus-->

7       (0.3)+•(0.2)#(0.1)*†(1.2) •
  ped       +gazes at bus->>
  cycR       •moves onto sidewalk•
  cycR                -->*
  ped                  -->†stops->>
  img              #3
```

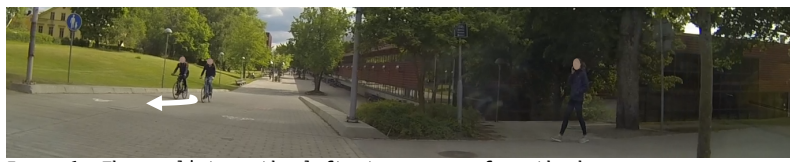

Image 1. The cyclist on the left steers away from the bus.

Image 2. The cyclist on the right steers away from the approaching bus.

Image 3. The pedestrian stops.

FIGURE 3.7
Prototyping sounds for an autonomous bus (EM-f sax round 3 cyclists at Blå Havet).

3.6 Discussion and Implications

By comparing human-human to human–robot interaction, we have demonstrated different ways in which sound can be a constitutive part of performing social actions, showing instances where coordination in real time is relevant. We have highlighted that it is never the sound alone that creates meaning, the message is not exclusively encoded into the sound. Instead, a sound entails a meaning potential that achieves significance in relation to the embodiment of its producer and other details of the local context. In the following, we highlight three main lessons that sound designers and researchers in robotics can take away from this work.

3.6.1 Meaning as Potentials

Reviewing literature and examples on non-lexical vocalizations, prosodies, and sense-making in humans, we highlighted that the term "semantic-free" [86] for robot sound is not entirely accurate. These sounds, like human vocalizations, do not necessarily carry a fixed meaning but they certainly feature potentials to be interpreted in particular ways, depending on the local interactional contexts. More generally, an important lesson for robotics is that human language does not function like math, with pre-defined symbols that always mean the same thing and lead to inevitable outcomes. Instead, meaning is partially flexible and negotiated, even for regular words. One may conceptualize robot audio as a continuum of fixedness, words like for instance "Hello" have relatively narrow meaning potentials, implying mostly a greeting at the beginning of encounters, while sound can have broader potential meanings, such as Cozmo's happy animation, which may be treated as a greeting or as accepting an offer for beer and a range of other things, depending on what exactly has just happened.

Our work highlights how studying human non-lexical vocalizations can inform the goals and questions relevant for robot sound. While it is important to ensure that potential meanings are going in the right direction (such as happy or sad valence in emotion displays), the design should not strive for setting an absolute meaning to them. Rather, we demonstrated that they gain rather specialized meanings once placed in concrete interactional contexts: asking for help, answering a question, instructing a move, greeting, coordinating, etc. For robotics, this kind of flexibility could be an advantage in settings where lexical expressions – especially in a specific language – are inappropriate or inefficient. The above explorations also highlight that it is worth considering and designing meaning potentials carefully: a sound that evokes associations with warnings such as a horn may not be appropriate for inviting people to come closer to a robot (the rolling bus). At the same time, what users and other people who encounter robots ultimately make of these sounds cannot be entirely pre-planned by the designers, leaving space also for creativity on

behalf of the users. Accordingly, field studies are essential in working out the meaning of each particular sound in situated contexts.

3.6.2 Sound and Multimodality

We demonstrated how humans skillfully design combinations of vocalizations and gestures and that Cozmo's visual behavior contributes to how its sounds get interpreted. Previous work has explored the interplay between sound and facial displays in experimental settings [40, 79], and we want to highlight the importance of studying sound in combination with embodied behavior like movement, gestures and facial expressions even in interaction. Our work demonstrates how various resource combinations are interpreted in the "messy" environments of everyday life; and we are showing the orderliness and logic in those. In real world interaction the amplifying function of sound [79] may be what draws people's attention to a movement or change in facial expression that would otherwise go unnoticed. Beyond communicating specific things through sound quality, the presence of sound may sometimes be particularly useful in marking a behavior as intentional or in highlighting the character of a movement, such as the Pilates teacher in the extract in Figure 3.4 was doing when accompanying an exercise.

Robot sound design can take inspiration from the growing body of work on human vocalizations, particularly during physical activities [1, 32, 33, 67]. This can be informative for settings in which humans and robots collaborate on physical tasks, such as to minutely coordinate lifts and handovers, or when working alongside large industrial robots. Sonifying parts of movements that are especially difficult for a robot could contribute to making the robot's behavior explainable for humans, rendering the collaboration more rewarding. Such sounds can also be a natural, implicit [30] tool for robots to ask for help (cf. Figure 3.1, the infant example). We do not necessarily envision that robots should copy humans but suggest looking more closely at the interplay of sound and visual behavior, also in the designers' own bodies when brainstorming robot sounds. Are we making a specific facial expression, movement or gesture while vocalizing a suggestion? How can this be translated to robot behavior in a meaningful way? We believe that paying attention to the interplay of these modalities, as argued for instance also in somaesthetic design approaches [27], can be beneficial even for robot sound design.

3.6.3 Variable Form and Reflexive Adaptation to Multiple Participants

Finally, we pointed out how sound can be variably produced in regard to pitch height, length, and repetition in order for it to be reflexively adjusted to ongoing action and the locally emerging context, such as for the moving bus. Our research extends prior work on robot sound that has focused on different intonation curves [11], highlighting that modifying other prosodic

and rhythmic elements may be a promising direction for further research. Such variation becomes especially relevant when robots interact in naturalistic settings, where interaction is not limited to one-user-one-robot, but where multiple participants need to coordinate.

When designing sound that can function in a range of contexts [68], repetition is a particularly promising resource. Varying the rhythm and number of repetitions, a single sample such as the saxophone roll used on the autonomous bus in the extract in Figure 3.7, can accomplish a range of different actions. Rather than designing a wide range of complex variations, design of so-called "semantic-free" utterances may benefit from inspecting closely how one sound can accomplish multiple actions in different real-life contexts. With modifications such as lengthening, repetition, and loudness variations, an agent can signal urgency, extension of the activity, as well as the fact that attempts at coordination (at least by some participants) are not yet sufficient. Even though the highest level of reflexivity between the sounding and those moving around a robot was only achieved by a human wizard in our study, the success highlights that there are opportunities for future design.

3.6.4 Designing Sound for Interactional Sequences

When aiming to design robot sound for interaction, we argue that it is crucial to consider its interpretability in the precise context where it will be used. Field methods such as observations and recordings of actual situations of sound use are essential, as is the close documentation of participant action. We hope to have demonstrated how a video-based study of sound in specific settings, through transcription can yield detailed insights of the relative timings of participant behavior and provide an empirical ground for discussion.

Most importantly, robot sound should not be designed in isolation, but prototyped in interactional sequences, in which timing, embodiment, and multimodal aspects of the local context play a crucial role. We are specifically interested in developing methods for adopting an interactional perspective that do not require specialist training in transcribing video. Close observation and video recordings in the setting in which the robot is used, and repeated sound design interventions in concrete interactional contexts, captured on video are key to such an approach. Specifically, we developed a video voice-over technique [59] that extends vocal sketching techniques [53, 70] by sketching over recordings of human-human or human–robot interaction: A short video snippet (about 30 seconds) is played on loop and sound is prototyped by performing voice-overs of how a robot could sound in this situation, either with one's own voice or by playing a sample. Repeatedly testing sounds on top of video recordings of actual interaction with the moving robot helps to get a sense of how they fit the particular embodiment of the robot. More importantly, it enables designers to intuitively produce the most interactionally relevant timing and duration, and potentially this kind of "annotated" data could be further used for teaching robots at what moments in interaction

they should produce sound [54]. Further exploring this interactional sound design approach, we also tested sound with Wizard-of-Oz setups in real world environments, as exemplified in the extract in Figure 3.7. This provides a sense of the specific soundscape and the interplay of sound and other multimodal aspects. The setup can reveal different meaning potentials in a range of real-life consequential situations and is particularly well-suited for gaining insights on variations through repetition, pitch modulation, and duration.

3.7 Conclusion

We set out to highlight lessons for robot sound design that can be learnt from studying how humans use non-lexical vocalizations and prosody to accomplish social actions. We demonstrated how a semantically underspecified sound gains meaning locally in concrete interactions, making sound particularly useful for contexts in which verbal utterances risk being too specific. We then provided insights on the multimodal nature of interaction, showing how sound and visual behavior are intertwined, arguing that robot sound designers can gain from scrutiny of their own bodily moves when brainstorming sounds and from paying close attention to how a robot moves in a concrete space and context while playing the designed sounds. Finally, we looked at how sound can be adapted to multiple addressees at the same time by prosody and repetition and provided an example from our own sound design with a Wizard-of-Oz setup on a public road. Overall, we argued that "non-semantic" or "semantic-free" sound is indeed semantically underspecified but not meaningless, and provided examples of how sound can be flexibly adjusted to coordinate actions with multiple participants. Robot sounds should preferably be designed as multimodal displays for interactional sequences.

Acknowledgment

We would like to thank Sally Wiggins, Adrian Kerrison and Emily Hofstetter for their invaluable comments on a previous draft of this chapter. This work is funded by the Swedish Research Council grant "Vocal practices for coordinating human action" (VR2016-00827).

Bibliography

[1] Albert, S., and vom Lehn, D. Non-lexical vocalizations help novices learn joint embodied actions. *Language & Communication 88* (Jan 2023), 1–13.

[2] Ben Moshe, Y. Hebrew stance-taking gasps: From bodily response to social communicative resource. *Language & Communication* (in press).

[3] Brown, B., and Laurier, E. The trouble with autopilots: Assisted and autonomous driving on the social road. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2017, pp. 416–429.

[4] Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., and Susini, P. The role of sound source perception in gestural sound description. *ACM Transactions on Applied Perception 11*, 1 (Apr 2014), 1–19.

[5] Chan, L., Zhang, B. J., and Fitter, N. T. Designing and validating expressive cozmo behaviors for accurately conveying emotions. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (Aug 2021), IEEE, pp. 1037–1044.

[6] Clift, R. Visible deflation: Embodiment and emotion in interaction. *Research on Language and Social Interaction 47*, 4 (Oct 2014), 380–403.

[7] Couper-Kuhlen, E. A sequential approach to affect: The case of disappointment. In *Talk in Interaction: Comparative Dimensions*, M. L. Markku Haakana and J. Lindström, Eds. Finnish Literature Society (SKS), Helsinki, 2009, pp. 94–123.

[8] Couper-Kuhlen, E., and Selting, M. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.

[9] Dingemanse, M., Torreira, F., and Enfield, N. J. Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE 8*, 11 (2013), 1–10.

[10] Ekman, I., and Rinott, M. Using Vocal Sketching for Designing Sonic Interactions. *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, (2010) 123–131. https://doi.org/10.1145/1858171.1858195

[11] Fischer, K., Jensen, L. C., and Bodenhagen, L. To beep or not to beep is not the whole question. In *Social Robotics. ICSR 2014. Lecture Notes in Computer Science, vol 8755*, M. Beetz, B. Johnston, and M. Williams, Eds. Springer, Cham, 2014, pp. 156–165.

[12] Fischer, K., Soto, B., Pantofaru, C., and Takayama, L. Initiating interactions in order to get help: Effects of social framing on people's responses to robots' requests for assistance. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (Aug 2014), pp. 999–1005.

[13] Frid, E., and Bresin, R. Perceptual evaluation of blended sonification of mechanical robot sounds produced by emotionally expressive gestures: Augmenting consequential sounds to improve non-verbal robot communication. *International Journal of Social Robotics 14*, 2 (2022), 357–372.

[14] Frid, E., Lindetorp, H., Hansen, K. F., Elblaus, L., and Bresin, R. Sound forest: Evaluation of an accessible multisensory music installation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland UK, May 2019), ACM, pp. 1–12.

[15] Gardner, R. *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins, Amsterdam, 2001.

[16] Goodwin, C. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, London, 1981.

[17] Goodwin, C. Participation, stance and affect in the organization of activities. *Discourse & Society 18*, 1 (2007), 53–73.

[18] Goodwin, C. *Co-Operative Action*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press, Cambridge, Nov 2018.

[19] Goodwin, C., and Goodwin, M. H. Concurrent operations on talk: Notes on the interactive organization of assessments. *IPrA Papers in Pragmatics 1*, 1 (1987), 1–54.

[20] Goodwin, C., and Goodwin, M. H. Participation. In *A Companion to Linguistic Anthropology*, A. Duranti, Ed. Blackwell, Oxford, 2004, pp. 222–244.

[21] Hepburn, A., and Bolden, G. B. *Transcribing for Social Research*. SAGE, London, 2017.

[22] Heritage, J. *Garfinkel and Ethnomethodology*. Polity Press, 1984.

[23] Hoey, E. M. Sighing in interaction: Somatic, semiotic, and social. *Research on Language and Social Interaction 47*, 2 (Apr 2014), 175–200.

[24] Hoey, E. M. Waiting to inhale: On sniffing in conversation. *Research on Language and Social Interaction 53*, 1 (Jan 2020), 118–139.

[25] Hofstetter, E. Nonlexical "moans": Response cries in board game interactions. *Research on Language and Social Interaction 53*, 1 (Jan 2020), 42–65.

[26] Hofstetter, E., and Keevallik, L. Prosody is used for real-time exercising of other bodies. *Language & Communication 88* (Jan 2023), 52–72.

[27] Höök, K. Designing with the Body: Somaesthetic Interaction Design. (2018) MIT Press.

[28] Jee, E.-S., Jeong, Y.-J., Kim, C. H., and Kobayashi, H. Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics 3*, 3 (2010), 199–206.

[29] Johannsen, G. Auditory displays in human–machine interfaces of mobile robots for non-speech communication with humans. *Journal of Intelligent and Robotic Systems 32*, 2 (2001), 161–169.

[30] Ju, W. *The Design of Implicit Interactions.* Springer International Publishing, 2015.

[31] Keevallik, L. Linguistic structures emerging in the synchronization of a pilates class. In *Mobilizing Others: Grammar and Lexis within Larger Activities*, C. Taleghani-Nikazm, E. Betz, and P. Golato, Eds. John Benjamins, Amsterdam/Philadelphia, 2020, pp. 147–173.

[32] Keevallik, L. Vocalizations in dance classes teach body knowledge. *Linguistics Vanguard 7*, s4 (2021), 20200098.

[33] Keevallik, L. Strain grunts and the organization of participation. In *Body, Participation, and the Self: New Perspectives on Goffman in Language and Interaction*, L. Mondada and A. Peräkylä, Eds. Routledge, London, 2023.

[34] Keevallik, L., Hofstetter, E., Weatherall, A., and Wiggins, S. Sounding others' sensations in interaction. *Discourse Processes* (Jan 2023), 1–19.

[35] Kerstin Norén, Per Linell "Meaning potentials and the interaction between lexis and contexts: an empirical substantiation", *Pragmatics*, 17 (2007), pp. 387–416.

[36] Keevallik, L., and Ogden, R. Sounds on the margins of language at the heart of interaction. *Research on Language and Social Interaction 53*, 1 (Jan 2020), 1–18.

[37] Kerstin Norén and Per Linell, "Meaning potentials and the interaction between lexis and contexts: an empirical substantiation", *Pragmatics*, vol. 17, no. 3 (2007), pp. 387–416.

[38] Licoppe, C., Luff, P. K., Heath, C., Kuzuoka, H., Yamashita, N., and Tuncer, S. Showing objects: Holding and manipulating artefacts in video-mediated collaborative settings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, Association for Computing Machinery, pp. 5295–5306.

[39] Linell, P. *Rethinking Language, Mind, and World Dialogically.* Advances in Cultural Psychology: Constructing Human Development. Information Age Publishing, Charlotte, NC, 2009.

[40] Löffler, D., Schmidt, N., and Tscharn, R. Multimodal Expression of Artificial Emotion in Social Robots Using Color, Motion and Sound. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 334–343 (2018). https://doi.org/10.1145/3171221.3171261

[41] Merat, N., Louw, T., Madigan, R., Wilbrink, M., and Schieben, A. What externally presented information do vrus require when interacting with fully automated road transport systems in shared space? *Accident Analysis & Prevention 118* (2018), 244–252.

[42] Mondada, L. Precision timing and timed embeddedness of imperatives in embodied courses of action: Examples from french. In *Imperative Turns at Talk: The Design of Directives in Action*, M.-L. Sorjonen, L. Raevaara, and E. Couper-Kuhlen, Eds., Studies in language and social interaction. John Benjamins Publishing Company, Amsterdam, Philadelphia, 2017, pp. 65–101.

[43] Mondada, L. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics 145* (May 2019), 47–62.

[44] Mondada, L. Audible Sniffs: Smelling-in-Interaction. *Research on Language and Social Interaction 53*, 1 (Jan 2020), 140–163.

[45] Moore, D., Currano, R., and Sirkin, D. Sound decisions: How synthetic motor sounds improve autonomous vehicle-pedestrian interactions. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (New York, NY, USA, 2020), AutomotiveUI '20, Association for Computing Machinery, pp. 94–103.

[46] Moore, D., Tennent, H., Martelaro, N., and Ju, W. Making noise intentional: A study of servo sound perception. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2017), HRI '17, Association for Computing Machinery, pp. 12–21.

[47] Norén, K. and Linell, P. "Meaning potentials and the interaction between lexis and contexts: an empirical substantiation", *Pragmatics*, vol. 17, no. 3 (2007), pp. 387–416.

[48] Norman, D. *The Design of Everyday Things: Revised and Expanded Edition.* Basic Books, 2013.

[49] Ochs, E., Schegloff, E. A., and Thompson, S. A., Eds. *Interaction and Grammar*. Cambridge University Press, Cambridge, 1996.

[50] Ogden, R. Audibly not saying something with clicks. *Research on Language and Social Interaction 53*, 1 (2020), 66–89.

[51] Okada, M. Imperative actions in boxing sparring sessions. *Research on Language and Social Interaction 51*, 1 (Jan 2018), 67–84.

[52] Okada, M. Lexical repetitions during time critical moments in boxing. *Language & Communication* (in press).

[53] Panariello, C., Sköld, M., Frid, E., and Bresin, R. From vocal sketching to sound models by means of a sound-based musical transcription system. *Proceedings of the SMC Conference 2019. Sound and Music Computing*. (2019). https://www.smc2019.uma.es/articles/S2/S2_05_SMC2019_paper.pdf

[54] Park, H. W., Gelsomini, M., Lee, J. J., and Breazeal, C. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2017), IEEE, pp. 100–108.

[55] Pehkonen, S. Response Cries Inviting an Alignment: Finnish huh huh. *Research on Language and Social Interaction 53*, 1 (Jan 2020), 19–41.

[56] Pelikan, H. Transcribing human–robot interaction. In P. Haddington, T. Eilittä, A. Kamunen, L. Kohonen-Aho, T. Oittinen, I. Rautiainen, & A. Vatanen (Eds.), Ethnomethodological Conversation Analysis in Motion: Emerging Methods and New Technologies (1st ed.). Routledge. (2023). https://doi.org/10.4324/9781003424888

[57] Pelikan, H., and Hofstetter, E. Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction* (2022).

[58] Pelikan, H. R., and Broth, M. Why that nao?: How humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, May 2016), ACM, pp. 4921–4932.

[59] Pelikan, H. R., and Jung, M. Designing robot sound-in-interaction: The case of autonomous public transport shuttle buses. In *ACM/IEEE International Conference on Human-Robot Interaction* (2023).

[60] Pelikan, H. R. M., Broth, M., and Keevallik, L. "Are you sad, cozmo?": How humans make sense of a home robot's emotion displays. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (2020), Association for Computing Machinery, pp. 461–470.

[61] Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., and Heath, C. The first five seconds: Contingent stepwise entry into an interaction as a means to secure sustained engagement in hri. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication* (Sep 2009), IEEE, pp. 985–991.

[62] Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 640:1–640:12.

[63] Read, R., and Belpaeme, T. How to use non-linguistic utterances to convey emotion in child-robot interaction. In *Proceedings of the Seventh annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12* (New York, New York, USA, 2012), HRI '12, ACM Press, p. 219.

[64] Read, R., and Belpaeme, T. People interpret robotic non-linguistic utterances categorically. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (Piscataway, NJ, USA, 2013), HRI '13, IEEE Press, pp. 209–210.

[65] Read, R., and Belpaeme, T. Situational context directs how people affectively interpret robotic non-linguistic utterances. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2014), HRI '14, Association for Computing Machinery, pp. 41–48.

[66] Read, R. G., and Belpaeme, T. Interpreting non-linguistic utterances by robots. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments - AFFINE '10* (New York, New York, USA, 2010), AFFINE '10, ACM Press, pp. 65.

[67] Reynolds, E. Emotional intensity as a resource for moral assessments: The action of "incitement" in sports settings. In *How Emotions Are Made in Talk*, J. S. Robles and A. Weatherall, Eds. John Benjamins Publishing Company, 2021, pp. 27–50.

[68] Robinson, F. A., Bown, O., and Velonaki, M. Designing sound for social robots: Candidate design principles. *International Journal of Social Robotics 14*, 6 (2022), 1507–1525.

[69] Robinson, F. A., Velonaki, M., and Bown, O. Smooth operator: Tuning robot perception through artificial movement sound. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2021), HRI '21, Association for Computing Machinery, pp. 53–62.

[70] Rocchesso, D., Lemaitre, G., Susini, P., Ternström, S., and Boussard, P. Sketching sound with voice and gesture. Interactions, 22(1) (2015), 38–41. https://doi.org/10.1145/2685501

[71] Sacks, H. Notes on methodology. In *Structures of Social Action: Studies in Conversation Analysis*, J. Heritage and J. M. Atkinson, Eds. Cambridge University Press, Cambridge, UK, 1984, pp. 21–27.

[72] Sacks, H., Schegloff, E. A., and Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language 50*, 4 (1974), 696.

[73] Savery, R., Rose, R., and Weinberg, G. Establishing human-robot trust through music-driven robotic emotion prosody and gesture. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2019), IEEE Press, pp. 1–7.

[74] Simone, M., and Galatolo, R. Timing and prosody of lexical repetition: How repeated instructions assist visually impaired athletes' navigation in sport climbing. *Research on Language and Social Interaction 54*, 4 (Oct 2021), 397–419.

[75] Song, S., and Yamada, S. Expressing Emotions Through Color, Sound, and Vibration with an Appearance-Constrained Social Robot. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2–11 (2017). https://doi.org/10.1145/2909824.3020239

[76] Suchman, L. A. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge, 1987.

[77] Tekin, B. S. Cheering together: The interactional organization of choral vocalizations. *Language & Communication* (in press).

[78] Tennent, H., Moore, D., Jung, M., and Ju, W. Good vibrations: How consequential sounds affect perception of robotic arms. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Aug 2017), IEEE, pp. 928–935.

[79] Torre, I., Holk, S., Carrigan, E., Leite, I., McDonnell, R., and Harte, N. Dimensional perception of a 'smiling McGurk effect'. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Nara, Japan, Sep 2021), IEEE, pp. 1–8.

[80] Torre, I., Latupeirissa, A. B., and McGinn, C. How context shapes the appropriateness of a robot's voice. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (Naples, Italy, Aug 2020), IEEE, pp. 215–222.

[81] Tuncer, S., Licoppe, C., Luff, P., and Heath, C. Recipient design in human–robot interaction: The emergent assessment of a robot's competence. *AI & SOCIETY* (Jan 2023).

[82] Weatherall, A. "Oh my god that would hurt": Pain cries in feminist self-defence classes. *Language & Communication* (in press).

[83] Weatherall, A., Keevallik, L., La, J., Stubbe, M. and Dowell, T. The multimodality and temporality of pain displays. *Language and Communication*, 80: 2021, 56–70.

[84] Wiggins, S. Talking with your mouth full: Gustatory mmms and the embodiment of pleasure. *Research on Language and Social Interaction 35*, 3 (Jul 2002), 311–336.

[85] Wiggins, S., and Keevallik, L. Parental lip-smacks during infant mealtimes: Multimodal features and social functions. *Interactional Linguistics 1*, 2 (2021), 241–272.

[86] Yilmazyildiz, S., Read, R., Belpeame, T., and Verhelst, W. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction 32*, 1 (Jan 2016), 63–85.