Non-Verbal Auditory Aspects of Human-Service Robot Interaction

Jeannie S. Lee[*], Muhamed Fauzi Bin Abbas[*], Chee Kiat Seow[†], Qi Cao[†]
Kar Peo Yar[*], Sye Loong Keoh[†], Ian McLoughlin[*]
[*]*Infocomm Technology Cluster, Singapore Institute of Technology, Singapore*
[†]*School of Computing Science, University of Glasgow, United Kingdom*
{Jeannie.Lee, Fauzi.Abbas, KarPeo.Yar, Ian.McLoughlin}@SingaporeTech.edu.sg
{CheeKiat.Seow, Qi.Cao, SyeLoong.Keoh}@glasgow.ac.uk

***Abstract*—As service robots become ever more pervasive, the number, degree and depth of interaction with humans, particularly fellow workers, is increasing rapidly. Humans are generally shaped alike, respond in predominantly similar ways and are often inherently predictable to other humans. Robots, by contrast, have an exceptional diversity of size, shape, mobility, function, and their intentions or actions are often less predictable.**

Humans working in close proximity have learnt to provide cues to their behaviour, both verbal and non-verbal, and we argue that this is an important aspect of maintaining both safety and comfort in a mixed work or social environment. At present, robots do not provide any such cues to their fellow workers, which can be cause of human discomfort, and indeed contribute to safety issues when working in close proximity to humans.

This paper considers the non-verbal auditory aspects of interaction in a work environment, with particular emphasis on safe and comfortable integration of service robots into such locations. In particular, we propose a classification of interaction levels to inform the construction, programming and operation of robots in the workplace.

***Index Terms*—Robotics, non-verbal communications, human-computer interaction, human-robot interfacing, auditory communications**

I. Introduction

In real-world work environments, it is increasingly common to find robots and humans co-existing in mixed workplaces, rather than situations where robots are functioning in isolation or purely with other robots. Just as socially unaware or impaired humans can be difficult to work with as colleagues (e.g. people who move erratically, or who do not respond to normal cues of social behaviour), so too are socially and context-unaware robots in mixed settings. It is thus imperative, as robots develop in complexity, that they gain the ability to firstly understand, and secondly respond to, the context around them. They would clearly benefit from being able to detect, process and interpret non-verbal cues, as well as the obvious benefits to be gained from verbal communications.

However, this paper is not concerned primarily with the 'feelings' of robots adapting into a mixed work environment, but rather to co-exist in comfort and safety with humans.

In this emerging area of human-robot co-existence in the workplace, robots must be able to move and act in safe, understandable, and appropriate ways. They should ideally take into account social rules, such as social distance [1]–[3]. For example, if social distance is routinely ignored, humans may become uncomfortable, which would reduce robot acceptance. A robot may also not convey expected cues regarding its automated decisions, thus leaving a human to either guess or to exert effort to interpret such decisions – which might in turn introduce collaboration errors.

When encountering other people in a real-world scenario, most humans can capture non-verbal cues and understand the corresponding social context. Humans usually realise when interactions are needed, how to communicate, what information to exchange, and how to wrap up such interactions before continuing with their tasks. As another example, when construction work is found in an operating environment, humans can recognise that situation and try to avoid interrupting or interacting inappropriately.

The way that current mobile robots approach a human can be unpleasant or threatening. For example, a common scenario is where a robot suddenly stops when encountering a human (for path re-planning), as shown in Fig 1. Even when the robot is intending to give the human the right of way, their stop-start motion and long pauses may cause stress to the human involved [4], [5]. The impact of such interactions are important for the deployment of robots in crowded environments. This is true also in non-work or public environments where confrontations have been known to trigger emotional issues in humans, especially in young children.

With the increased adoption of robots 'in the wild', efficient navigation from source to destination, and completion of their tasks, are not the only essential requirements. The *manner* in which they interact with, and are perceived by humans has now become an equally important topic.

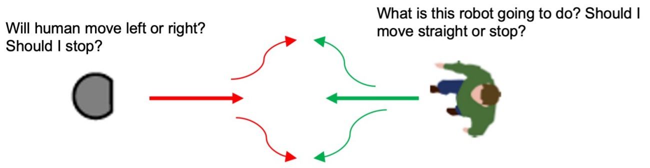
Will human move left or right?
Should I stop?
What is this robot going to do? Should I
move straight or stop?

Fig. 1. An example scenario where a robot encounters a human, both need to plan for appropriate movement and communication.

If there is any deficiency at present in robot-human communications, it is not to be found primarily among the human workers. Most service robots are featureless and – barring their electric motors – soundless. Potential communications methods include verbal or non-verbal means, with non-verbal communications (NVC) spanning both visual and auditory interaction. Non-speech auditory signals are also referred to as non-linguistic utterances (NLU) [6].

Some robots incorporate a visual display in the "face" area, complete with artificial eyes and mouth, moving eyebrows and the ability to frown. However this is generally intended to augment spoken communications in models equipped with automatic speech recognition (ASR), text-to-speech (TTS) and rudimentary natural language processing (NLP) capabilities [7].

Apart from using speech [8], few modern robots are able to communicate their current operation or intention to humans. Speech is clearly one option to do so, but it is unnatural. Humans moving in a crowded environment do not routinely signal their movement intention with speech, for example "I am moving past you, and then I intend to move right... I am now pausing for a short while". Instead they use a range of non-verbal cues. The authors argue that adoption or mimicking of some of those non-verbal cues would better enable service robots to comfortably co-exist with humans in mixed workplaces. Speech, though, is still useful to convey more complex information [8], just as it is when humans need to communicate more complexity than NVC allows.

We note that non-verbal robot communications has been explored through significant prior research. This includes endowing robots with characteristic sounds, much like fictional devices R2-D2 and Wall-E communicate feelings through non-verbal bleeps, whistles and buzzes. Jee *et al.* [9] investigated using precisely these sounds for socially interactive robots, concluding that it was effective at conveying emotion. This demonstrated the potential of sound for communications, although not the ability to communicate intent, or more complex information. Read [6] later investigated robots using NLUs formed with different pitch contours for discrimination and identification, clearly showing that they are able to communicate – although again it was a range of emotions being tested. However results also showed that listeners preferred to interpret those sounds categorically, according to human prototypes.

NLUs in human-*computer* interfacing is, a very well researched subject, for example by Roth *et al.* [10] who included design guidelines for usability. Furthermore, in addition to the naturalness and ease of NVC in certain circumstances, it is very effective at reducing information overload [11]. This is probably also something to consider as an important aspect of scalability. Unlike speech which needs to be linguistic and is lengthy in nature, a key advantage of NLU is the formation of short bursts of sound utterance that can correlate strongly to human understanding [12]. This will resolve the problem that exists in modalities that are not well-received where either the communication channel or sensation is poor.

Applying NLU research into HRI applications is still in a stage of infancy [13], therefore this forms the basis of our analysis, leading to the seven levels of interaction describing NVC for HRI communications as articulated in Section III.

II. How humans do things

Humans have long used sounds to communicate, and this includes semantic meanings ranging from the simplest auditory alert to the deepest conceptual or emotional constructs.

In general, speech (and, by proxy, the written word), is used to convey more complex information between humans, almost always augmented by NLU and prosody (along with other NVC like facial expressions, posture, and gestures) [14].

When communicating to those around us, we actually take a lot of contextual information into account, including who we are talking to, the number of people present, their distance from us (and orientation to us) as well as the subject matter being conveyed – and its perceived urgency or importance. We account for environmental noise without usually being aware of it [7], and make use of repetition, emphasis and phraseology to convey additional information, with acknowledgement (negative of positive) happening via both main and side channels [14].

While speech is generated by lung exhalation [7], NLUs can also be created by other means. This includes tongue clicks, inhalation (as well as exhalation), throat clearing, but also makes use of kinetic sounds (e.g. shuffling or dragging feet noisily – which is sometimes done deliberately to make others aware of movement). In most cases, NLU and NVC is initiated based on the need to convey some information.

Humans are clearly highly responsive to their operating context and to the need to communicate, but are not always conscious that they are doing so. And yet when one person in a shared environment fails to communicate as expected, it can lead to some discomfort for others [14].

III. Proposed robot communications

Service robots are more of a blank canvas for sound design, as they tend to be non-humanoid. As the focus is on service robots, which excludes the specific sub-fields of humanoid or animal-like designs (or shapes like R2-D2 or Wall-E which are already 'characters') for which anthropomorphism is an expected attribute. By contrast, service robots do not usually appear to be human or animal-like, and there is thus much less opportunity for there to be an implicit expectation regarding the *type* of communication used (for example, a humanoid robot may be expected to talk, whereas a dog-like robot might be expected to bark or make similar dog-like noises).

Upon this blank canvas, the audible communications of a service robot can have three main categories:

- Auditory icons or earcons - cues, notifications, informational alerts, feedback.
- Ambient background sound – to indicate that a robot is nearby, or to establish mood and situation.
- Anthropomorphic intent notifiers – specifically to relate to humans in the vicinity.

Earcons are non-verbal audio messages that are used in a machine's user interface to provide information to the user about some object, operation or interaction [15]. They are specific sounds created to communicate a particular function, for example the sound your computer makes to indicate a new email has arrived, or a notification in a vehicle to indicate that a door has been left open. Auditory icons may use naturally occurring sounds as natural methods to convey conceptual objects or metaphors within the computer system [16]. Such sounds can be learned by human operators and users as short-hand to convey information. These auditory icons can be further sub-classed in terms of their importance levels, for example:

1) Emergency notifications
2) Alert or warning sounds
3) Cues, notifications or feedback

Ambient background sound has a similar function to the comfort noise that many electronic devices are designed to make, for example the regulatory sounds that electric vehicles (EV) must produce to warn pedestrians of their presence. Like EV sounds, it is probably sensible for frequency or amplitude to scale with velocity. Where the first and second categories are more process specific, i.e. related to the particular task, design or environment. The third category relates more to how robots co-exist with humans.

Turning to anthropomorphic intent notifiers, this is the class of sounds that should endow a service robot with the ability to comfortably co-exist with humans. These should not require learning – they need to be obvious to an untrained observer, but nevertheless informative. There are many possible notifiers, but a simple set for mobile robots might include:

- About to move
- About to stop
- About to change direction (left/right)
- Acknowledgement of human presence
- A temporary pause to 'think', or if the path is blocked
- Reached destination

With such a list, the service robot is able to convey its intention to nearby humans. This is, however, far from exhaustive. It also takes very little account of the ability of a robot to read and understand human intention, something that humans often automatically acknowledge (think of the situation of passing a colleague in a narrow corridor, which involves a 'negotiation' of passing on the left or right, the potential to pause, slow down, or simply hurry past – all of which can be communicated non-verbally).

Given the wide variation in possibilities, it is useful to categorise NLU human-robot communications into levels of non-verbal interaction, as proposed in Table I.

TABLE I
LEVELS OF NVC IN HRI COMMUNICATION.

Level	**Overall Interaction**		
	Input	**Output**	**Robot communicates:**
Level 0	None	None	Operates without reference to any humans around it, does not communicate.
Level 1	None	Action	Signals current activity (e.g. motion), broadcasts basic intent (e.g. about to move).
Level 1a	Proximity	Action	Signals activity and intent in response to human proximity.
Level 2	Motion	Intent	Signals intent in response to human motion and activity, as well as current action.
Level 3	Activity	Intent + context	Infers human activity, signals intent and action according to context.
Level 4	Intent	Intent + context	Infers human intent, plans and responds accordingly.
Level 5	Intent + context	Intent + context	Infers human intent and its context, plans, responds and signals accordingly.
Level 6	Personality	Personality	Works with humans as part of a team, with similar degree of NVC.

IV. BENEFITS

The use of NLU as a means of endowing service robots with NVC capabilities is clearly important for their ability to co-exist and co-work with humans, particularly in confined areas. Apart from NLU, and the situation of no communications, two current alternatives exist: spoken and visual notification (in future, haptics may be added to this list).

Visual notification requires some kind of display screen (or screens) or gesture, plus the ability for human co-workers to read this. Spoken communications requires auditory playback, and the ability for human co-workers to hear and understand it. Display screens, gesture devices and sound sources may be located on a robot, located externally (e.g. on a wall), or conveyed on a mobile device carried by the humans (e.g. a smartphone). However in any environment with multiple robots, coinciding messages and confusion over message source will dictate that it is preferable for messages to be co-located with their source (i.e. on the robots themselves). In any case, overlapping and coinciding spoken and textual notifications are problematic with multiple devices.

Given such scenarios, there are considerable benefits to be gained from the use of NLU to communicate from service robots to the humans around them:

- NLU can be language and culture independent.
- The use of natural human NLU modalities [14] is easily understood, for example it does not require high levels of cognitive ability (as textual or spoken utterances might).
- Sound alerts reduce the load on the visual senses, they are "eyes free", and also inherently multi-directional (i.e. do not require the human to be facing or looking in a certain direction).
- Sound alerts free up the visual interface bandwidth for something else (and also don't require the subjects proximity to the robot as would be the case when reading text on a robot-mounted screen).
- Suitable for the visually impaired.
- Have been demonstrated to have reduced cognitive load, distraction potential and attention bandwidth [11].

- Shorter processing time – faster to convey an audible NLU message than a spoken or textual alert; could be within milliseconds. This has particular importance in crowded environments where many robots might be operating (imagine a parcel sorting hall containing 100 speaking robot that broadcast their every action and intention through speech).

Beyond the benefits noted above, it is also important to recognise that the computation hardware, and control software, needed to generate audible alert sounds is much simpler than that required to produce speech (which is in turn significantly less complex than that required to understand speech) [17]. There is thus a lower computational load required for conducting non-verbal audible communications from robot to human.

V. Deficiencies and negatives

A. Crowds

While NLUs occupy a lower sensory bandwidth than speech, there is still potential for them to become confusing in crowded environments. Use of directional sounds for playback (and for sensing) and context sensitive volume scaling would help [18], although at the cost of additional complexity. Nevertheless, a level 5 and above interaction would need to scale appropriately with the number of operators (human or robot) that are physically present. This applies to both sensing individuals in a crowd, as well as ensuring NLU signals are noticed in a crowd, and ensuring good directionality or localisation.

B. Environment deficiencies

Although NLUs may work reasonably well in indoor environments, it might face some challenges when the robots are deployed outdoors. One possible challenge is the rapid change in noise level due to sudden weather changes (e.g. thunder) or passing traffic. In such cases, the noise levels would increase exponentially and may impair the NLU information transmitted from robot to human. A human might miss information or even misinterpret an NLU. This is not a trivial situation as increasing the loudness of NLU might result in a different perception of its meaning from the perspective of psychoacoustics [7], [19]. In such situations, visual cues might be more appropriate for the robot to convey an intention to humans.

C. Human deficiencies

Although we have argued based on the fact that humans enjoy more uniformity than robots, there is still considerable diversity within the human population. Audible signals can be designed to work across all cultures and language groups, but this does require careful design and testing, since different cultures make use of a range of NLUs with different meanings (e.g. in-drawn breath in Japanese, ‘tut’ tongue click in English, in contrast to the near-universal ‘huh’ sound [20]).

Beyond language and culture, we note that audible alerts are obviously problematic for the hearing impaired. This includes those with total hearing loss, but also frequency-selective impairment. Accompanying visual indicators are then useful.

D. Robot deficiencies

Just like humans, robots are not all-seeing and all-aware and can be ‘surprised’ too. Not only can humans fail to correctly identify signals from a robot, but robots themselves will occasionally misinterpret activity, motion, proximity and intention. Humans operating in crowded, noisy, or dangerous spaces will learn to use fail-safe mechanisms of communications (e.g. waiting for positive acknowledgement or confirmation, use of bi-directional signalling mechanisms), as well as inherently more robust communications means (for example saying ‘alpha zulu fiver’ instead of ‘AZ5’, an example from the NATO Phonetic Alphabet which is used in high noise environments). Robots likewise need robust NLU methods.

VI. Conclusion

The proliferation and prevalence of service robots has made the coexistence of robots with humans extremely common, particularly with the advent of the Covid-19 pandemic. Two forms of human-robot interaction (HRI) have been discussed regarding their limited viability and effectiveness. While implicit communication translates robot intent into decipherable human actions, explicit communication requires advanced agreement for visual and auditory gestures. This paper proposes a novel robot interaction classification with seven levels of non-verbal communications (NVC) within HRI using non-linguistic utterances (NLU) to overcome these limitations. These levels of incrementally more complex NVC define the generation of understandable auditory and other gestures that align with human emotional responses. A key advantage of auditory NVC from robots to humans is simplicity, making it more natural and faster for human comprehension. This simplicity also leads to significantly lower computational loads and improved responsiveness than existing auditory speech or visual systems. In fact, this classification lends itself to the design of low-overhead NLU-aware service robots to co-exist with humans in a safe and comfortable work environment. Future challenges such as more complex environments and diverse cultures, need to be addressed to tailor these recommendations for broader applications in future.

References

[1] G. R. Collins, “Improving human–robot interactions in hospitality settings,” *International Hospitality Review*, 2020.

[2] J. M. Beer, A. Prakash, T. L. Mitzner, and W. A. Rogers, “Understanding robot acceptance,” Georgia Institute of Technology, Tech. Rep., 2011.

[3] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas, “Safety bounds in human robot interaction: A survey,” *Safety science*, vol. 127, p. 104667, 2020.

[4] Y. Che, A. M. Okamura, and D. Sadigh, “Efficient and trustworthy social navigation via explicit and implicit robot–human communication,” *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 692–707, 2020.

[5] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, “Core challenges of social robot navigation: A survey,” *arXiv preprint arXiv:2103.05668*, 2021.

[6] R. Read and T. Belpaeme, “People interpret robotic non-linguistic utterances categorically,” *International Journal of Social Robotics*, vol. 8, no. 1, pp. 31–50, 2016.

[7] I. V. McLoughlin, *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press, 2016.

[8] P. Stavropoulou, D. Spiliotopoulos, and G. Kouroupetroglou, "Voice user interfaces for service robots: Design principles and methodology," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 489–505.

[9] E.-S. Jee, Y.-J. Jeong, C. H. Kim, and H. Kobayashi, "Sound design for emotion and intention expression of socially interactive robots," *Intelligent Service Robotics*, vol. 3, no. 3, pp. 199–206, 2010.

[10] Z. T. Roth and D. R. Thompson, "Usability of sound-driven user interfaces," 2018.

[11] S. A. Brewster, "Using non-speech sound to overcome information overload," *Displays*, vol. 17, no. 3, pp. 179–189, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141938296010347

[12] R. Read and T. Belpaeme, "How to use non-linguistic utterances to convey emotion in child-robot interaction," *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 219–220, 2012.

[13] H. Wolfe, M. Peljhan, and Y. Visell, "Singing robots: How embodiment affects emotional responses to non-linguistic utterances," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 11, no. 2, pp. 284–295, 2020.

[14] D. Crystal, *The Cambridge Encyclopedia of English Language*. Cambridge University Press, 2003.

[15] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human–Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[16] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-computer interaction*, vol. 2, no. 2, pp. 167–177, 1986.

[17] I. V. McLoughlin and H. R. Sharifzadeh, "Speech recognition for smart homes," *Speech Recognition, Technologies and Applications*, pp. 477–494, 2008.

[18] I. V. McLoughlin and Z.-P. Xie, "Speech playback geometry for smart homes," in *Consumer Electronics (ISCE 2014), The 18th IEEE International Symposium on*. IEEE, 2014, pp. 1–2.

[19] E. Asutay, D. Västfjäll, A. Tajadura-Jiménez, A. Genell, P. Bergman, and M. Kleiner, "Emoacoustics: A study of the psychoacoustical and psychological dimensions of emotional sound design," *Journal of the Audio Engineering Society*, vol. 60, pp. 21–28, 02 2012.

[20] P. O. Staff, "Correction: Is" huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items," *PloS one*, vol. 9, no. 4, p. e94620, 2014.