Initiating Human-Robot Interactions Using Incremental Speech Adaptation

Kerstin Fischer
Department of Design and Communication
University of Southern Denmark
Sonderborg, Denmark
kerstin@sdu.dk

Lakshadeep Naik
Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
lana@mmmi.sdu.dk

Rosalyn Langedijk
Department of Design and Communication
University of Southern Denmark
Sonderborg, Denmark
rla@sdu.dk

Timo Baumann
Natural Language Systems
University of Hamburg, Germany
baumann@informatik.uni-hamburg.de

Matouš Jelínek
Department of Design and Communication
University of Southern Denmark
Kolding, Denmark
majel18@student.sdu.dk

Oskar Palinko
Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
ospa@mmmi.sdu.dk

ABSTRACT

In this paper, we present a study in which a robot initiates interactions with people passing by in an in-the-wild scenario. The robot adapts the loudness of its voice dynamically to the distance of the respective person approached, thus indicating who it is talking to. It furthermore tracks people based on information on body orientation and eye gaze and adapts the text produced based on people's distance autonomously. Our study shows that the adaptation of the loudness of its voice is perceived as personalization by the participants and that the likelihood that they stop by and interact with the robot increases when the robot incrementally adjusts its behavior.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in interaction design**;

KEYWORDS

initiating interaction, incrementality, voice, gaze, personalization

ACM Reference Format:
Kerstin Fischer, Lakshadeep Naik, Rosalyn Langedijk, Timo Baumann, Matouš Jelínek, and Oskar Palinko. 2021. Initiating Human-Robot Interactions Using Incremental Speech Adaptation. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447205

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447205

1 INTRODUCTION

With robots moving into public spaces, like shopping malls, airports or museums, one task they are likely to fulfill is to address unfamiliar people and initiate interactions, for instance, to offer a service. Establishing an interaction with a stranger is not an easy task, not even for people [7]. For robots, which generally have fewer modalities to initiate interactions and to negotiate whether it is acceptable to approach a person, such a task is even harder (e.g. [20]).

The first problem a robot has to solve in order to address people in public spaces is to get their attention and to indicate that it is talking to them, rather than to anyone else or no one in particular (e.g.[17]). Thus, the robot needs to personalize its behavior in such a way that the person addressed recognizes that he or she is being addressed by the robot. In the current study, we therefore explore the effects of adjusting the robot's speech, gaze and body orientation to the one type of information readily available about a potential addressee, namely his or her distance from the robot.

Humans intuitively adapt their speech behavior based on the distance of the intended communication partner, which their communication partners can exploit to identify to whom the respective utterance is addressed (e.g. [9, 12]). That is, the further away the intended recipient is, the louder the speech (taking context into account; for instance, the chosen intensity level would be higher in a noisy shopping mall than in a quiet library). Thus, the given intensity level informs listeners about the intended addressee.

In the current study, we explore the extent to which this property of human interaction can be employed to select addressees in human-robot interaction and thus to initiate interactions. This is however not completely trivial because for humans, a kind of relaxed baseline intensity level is known, with which an increased intensity level, for instance, to address someone who is further away, can be compared. Concerning the synthesized speech that robots use, such a baseline is not given; that is, an 'extra effort' is not audible from synthesized speech that is played back at a higher volume, as other speech characteristics that characterize loud human speech, such as the fundamental frequency [2], tempo,

Figure 1: The Experimental Set-up

spectral intensity, or vowel-to-consonant ratios [25], are not easily controllable in off-the-shelf speech synthesis systems.

In human-robot interaction, the respective addressee has to understand that changes in intensity are incrementally adjusted to his or her distance in order to identify them as indicators of recipient design [9]. Thus, it can be expected that intensity as such does not function as an attention getting signal, but that the personalization of the intensity to the addressee's distance can be such a cue. Therefore, speech characteristics of the utterance produced by the robot must be adapted online as the utterance unfolds (and the proximity of the human changes).

2 PREVIOUS WORK

Previous work concerns studies that address how robots can initiate interactions in general, and what social cues have been found useful in order to do so, as well as work on incremental speech adaptation.

2.1 Initiating Human-Robot Interactions

Several studies confirm that initiating interactions by robots in the wild is a challenge; for instance, in a study in which a museum robot tries to get visitors involved in hearing more about a particular picture, Pitsch et al. [19] report 63% successful initiations of interactions without a personalized, contingent response by the robot in the dialog opening, compared to 80% success if the robot corresponds contingently.

Gehle et al. [13] and Kato et al. [17] experiment with the right moments and the right robot behaviors to open an interaction by learning from human operators. Similarly, Sidner et al. [27] explore different methods for initiating interactions in an office scenario. In general, speech is found to be more effective than just gaze or body orientation of the robot (e.g. [18, 26]). For instance, in a study by Fischer et al. [10], not a single participant responded to the robot's turn to the participant and gaze in their direction alone, and even a beep sound was largely ignored (see also [1]).

To sum up, previous work on how to initiate human-robot interactions in the wild has documented considerable challenges in getting people to interact with a robot. As the work by Pitsch and colleagues [19] has shown, a response that is contingent on the respective participant's behavior can effectively make people stop and interact, as well as combinations of behaviors in different modalities (e.g. [26, 27]). These findings suggest that multimodal robot behaviors that are tailored to and contingent on the user's behavior should be effective.

2.2 Incremental Speech Adaptation

Incremental processing concerns the piece-meal analysis of an interaction partner's behavior while it is occurring; that is, the computational system does not wait for a user action to be complete before processing it, but starts analyzing the meaning of incomplete, evolving actions [24].

Producing behaviors incrementally can be beneficial in HRI as well: for instance, instead of planning and producing more elaborate behaviors, the robot plans for the delivery of interruptable speech actions that can seemlessly be extended, thus foreseeing self-interruptions [5] and incremental information delivery [6]. Furthermore, adjusting what is spoken depending on the interaction with the context has been found beneficial [3]. In human-robot interaction, this generally leads to greater responsivity because the robot can adapt flexibly to aspects of the user's behavior [14].

Incremental speech production requires incremental speech synthesis [4] so that speech output can be seemlessly extended (or shortened) without audible breaks. As speech is synthesized *online*, with very little latencies, incremental speech synthesis also allows to adapt the way that speech is delivered in an online fashion. For example, speech intensity can be adapted to account for ambient noise, and Rottschäfer et al. [21] find that it is important to adapt more than the pure volume to be as natural as possible.

To sum up, previous work on incrementality in HRI has documented positive effects in terms of responsivity and dialog management. Since incrementality is one method to make a robot's behavior contingent on a user's behavior, one may expect that it is also beneficial to increase people's motivation to interact with the robot; that is, by processing the user's behavior continuously and responding incrementally, the robot can seamlessly adjust its behavior to the respective user. A particular kind of incremental interaction lies in adapting the loudness to react to the user's proximity with low latency and thereby achieving an effect of 'calling out' to the user when far away and creating intimacy with increasing proximity.

3 HYPOTHESES

In the current study, we aim to identify the effects of incremental adaptation of the robot's speech intensity (loudness) based on the addressee's distance, as well as the effects of turning towards a passerby and looking at him or her. Our hypothesis is that such adaptations will be noticed by the person addressed as contingent with their own movement and understood as an attempt to initiate the particular interaction.

4 METHOD

Since we want to study the effects of adjusting a robot's behavior to a particular, unsuspecting person, our experiment was carried out "in the wild", i.e. in a scenario where people did not know that they were going to encounter a robot. Thus, the experiment took place in a corridor of a large university building where participants

were students and staff from various disciplines, as well as cleaning personnel and members of the general public.

4.1 Procedure

The robot's behavior was implemented as shown in Figure 2. The different robot behaviors are the following:

- the robot perceives the person approaching
- the robot visibly turns to the person
- while the person's distance is more than 2m and less than 10m, the robot produces speech
  - the robot chooses the length of its speech based on the person's distance
  - in the adaptive condition, the robot adjusts its loudness based on distance
  - the robot turns to the person while approaching
- when the distance to the person is 2m or less,
  - if the person is slowing down or stopping, the robot abandons the current utterance at a suitable transition point and begins a new utterance (in the adaptive condition, uttered in soft voice)
  - if the person continues to walk past the robot, the robot abandons the current utterance and produces a new utterance (in the adaptive condition, loudness adjusted to distance)
  - if the person returns, the robot begins a new utterance (in the adaptive condition, in a soft voice).

We compare two conditions, one in which the robot adapts the loudness of its voice incrementally to the user's distance, and one in which it does not. All other robot behaviors are the same across conditions.

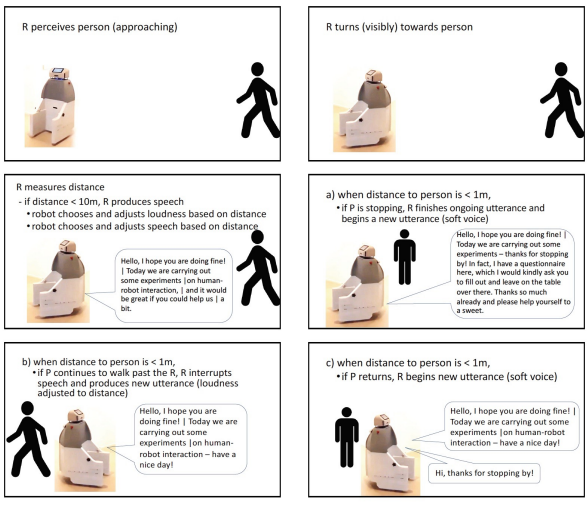

Figure 2: Overview of the interaction

The robot was placed on one side of a corridor with much space for people to pass if they preferred not to interact with the robot. At the entrance to the corridor, we installed a visible sign that a robot was in the corridor and that we were filming the interactions.

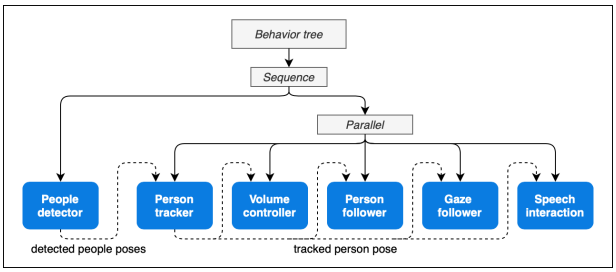

Figure 3: Robots behavior control architecture

People could then simply choose another route if they preferred not to interact with a robot or not to be videotaped. After having interacted with the robot or having passed by, the participants were stopped by a researcher and asked to fill out a short consent form and questionnaire.

4.2 The Robot

We use the Smooth robot [16] for this study. Fig. 3 describes the robot's behavior control architecture to exhibit the behavior described in Fig. 2. The robot has several action nodes, such as a people detector, a person tracker and a volume controller. All of these action nodes are combined using behavior tree [8] control nodes such as *sequence* (for sequential execution), or *parallel* (for parallel execution) to implement the desired behavior. The people detector uses a CNN-based pose estimator along with depth information for estimating the pose of the detected people [15]. Once a person is detected, the behavior tree switches to the parallel node, which executes 5 different actions in parallel. The person tracker tracks the identified person's position using an extended Kalman filter. The volume controller uses this information to change the robot's volume based on the person's distance. The person follower uses the tracked position and the PD controller to control the robot's turn (orientation) to follow the person, while the gaze follower uses the tracked person's position to follow the person using gaze. Since changing the robot's body orientation is slower than changing its gaze, gaze and turn operate independently. Speech interaction is also implemented as a separate state machine that monitors the person's distance and interrupts and plays various speech utterances.

4.3 Data Analysis

The quantitative measure used is people's subjective evaluation of the robot's approaching behavior after their encounter with the robot. Since this was an in-the-wild experiment, we could not keep people too long to fill out questionnaires, and thus we only asked them for their consent to use the video recordings, about their age and gender, and furthermore three self-report questions concerning the extent to which they thought the robot took them into account, responded to their actions and perceived them. Finally, we gave them some opportunity for additional comments.

As for the behavioral measures, we counted how often the robot's approaching behavior led people to stop and comply with the robot's request to fill out a questionnaire. In addition, we carried

Condition	takes into account	responds	perceives
non-adaptive	4.33 (0.70)	3.92 (1.02)	4.04 (1.23)
adaptive	3.96 (0.79)	4.16 (0.75)	4.042 (0.81)

Table 1: Mean (sd) evaluations of the robot's approach behavior in the two conditions

out a micro-analysis to understand how people responded to the different robot behaviors, such as turning, gazing and speaking, and what effect the change in loudness has on the interaction. The micro-analysis is based on ethnomethodological conversation analysis [22, 23], which aims to uncover people's own sense-making mechanisms. In order to make the analysis accessible to the larger HRI community, the presentation of the analysis was simplified.

5 RESULTS

50 people passed by the robot and stopped to fill out the questionnaire, 24 in the non-adaptive condition and 26 in the adaptive condition. Half of the participants identify as female, the other half as male.

A t-test shows that people judge the extent to which the robot has taken them into account significantly higher in the adaptive condition (t=1.74663, p= .043), whereas the differences between the other two judgements do not reach significance.

The qualitative analysis shows that in the adaptive condition 70.6% of the people addressed by the robot stop, compared to 56.3% in the non-adaptive condition. The example analysis in Figure 4 shows that people do notice the change in loudness and respond to it, which indicates that they feel addressed more and more inclined to interact with the robot. In addition to the loudness adaptation, also the fact that the robot acknowledges that they have not stopped by wishing them a good day functions as a strong clue that the robot perceives them and takes them into account.

The comments the post-experimental interview/questionnaire suggest that in the adaptive condition, people noted that the robot was adjusting to them, and they commented positively on the robot's gaze toward them, for instance: "Interaction was nice and the eyes are really nice," "Big friendly eyes. Gaze and turning was amazing," and "The eyes are getting high warm ratings." However, in both conditions, some find the robot also scary that it approaches them at all:"Found it scary that the robot approached me."

6 DISCUSSION

The experiment was carried out in an in-the-wild scenario, which means that people had not chosen to encounter a robot, that many were in a hurry, that they were alone or in groups etc. Since people often check with their partners whether they agree with how to respond to a robot approaching them [11, 28], the interaction dynamics between people in groups has influenced the interactions with the robot. Furthermore, at times, there was a lot of traffic in the hallway, and people were also reluctant to stop by if they were blocking the way for others. All these aspects of the "in-the-wild" situation are likely to have watered down the effect of our interventions.

Furthermore, many people who did stop to interact with the robot were also a bit disappointed that it could not interact with them beyond the dialog initiation. Thus, their evaluation of the degree to which the robot responded to them is not just due to the robot's approach behavior, but concerns its behavior *after* the successful initiation of the interaction.

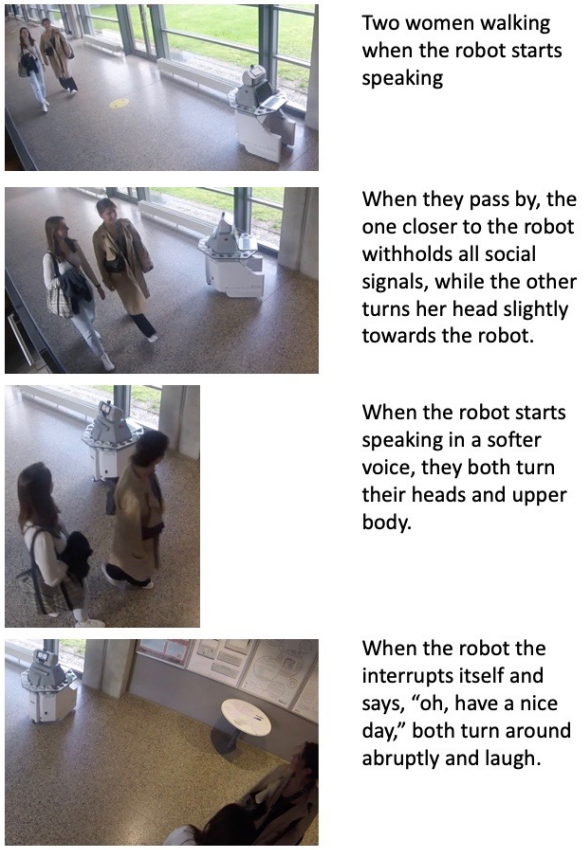

Figure 4: Example Analysis: Two women responding to the robot's adjustments

Given these influencing factors, the video analysis is more revealing in showing that most people respond to the robot's adaptation of the loudness of its speech, and that as many as 70.6% stop to interact with it.

7 CONCLUSIONS

Our analyses have shown that even in an unconstrained "in-the-wild" scenario, adjusting the loudness of a robot's voice to indicate that the robot is adjusting its speech to the particular person it is talking to has the intended effect, such that people feel personally addressed and more inclined to stop to interact with the robot.

For some participants, this strategy was however a bit scary – because it was effective at addressing them personally. We can thus conclude that while incremental loudness adaptation is effective, the personalization achieved may not always be welcome.

ACKNOWLEDGMENTS

This project was funded by the Innovation Fund Denmark.

REFERENCES

[1] Henny Admoni, Caroline Bank, Joshua Tan, Mariya Toneva, and Brian Scassellati. 2011. Robot gaze does not reflexively cue human attention. In *Proceedings of the Cognitive Science Society*, Vol. 33.

[2] Paavo Alku, Juha Vintturi, and Erkki Vilkman. 2002. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication* 38, 3 (2002), 321 – 334. https://doi.org/10.1016/S0167-6393(01)00072-3

[3] Timo Baumann and Felix Lindner. 2015. Incremental Speech Production for Polite and Natural Personal-Space Intrusion. In *Social Robotics, Proceedings of the 7th International Conference, ICSR 2015 (LNAI, Vol. 9388)*. Springer, Paris, France, 72–82. https://doi.org/10.1007/978-3-319-25554-5_8

[4] Timo Baumann and David Schlangen. 2012. Inpro_iSS: A Component for Just-In-Time Incremental Speech Synthesis. In *Procs of ACL 2012 System Demos* (Jeju, Korea). arXiv:urn:nbn:de:0070-pub-25145619

[5] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2016. Look at Me!: Self-Interruptions as Attention Booster?. In *Proceedings of the Fourth International Conference on Human Agent Interaction*. ACM, 221–224.

[6] Monika Chromik, Birte Carlmeyer, and Britta Wrede. 2017. Ready for the Next Step?: Investigating the Effect of Incremental Information Presentation in an Object Fetching Task. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 95–96.

[7] Herbert H. Clark. 1998. Communal Lexicons. In *Context in Language Learning and Language Understanding*, K. Malmkjaer and J. Williams (Eds.). Cambridge: Cambridge University Press, 63–87.

[8] Michele Colledanchise and Petter Ögren. 2016. How behavior trees modularize hybrid control systems and generalize sequential behavior compositions, the subsumption architecture, and decision trees. *IEEE Transactions on robotics* 33, 2 (2016), 372–389.

[9] Kerstin Fischer. 2016. *Designing Speech for a Recipient*. Amsterdam: John Benjamins.

[10] Kerstin Fischer, Bianca Soto, Caroline Pantofaru, and Leila Takayama. 2014. 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*. IEEE, 999–1005.

[11] Kerstin Fischer, Stephen Yang, Brian Mok, Rohan Maheshwari, David Sirkin, and Wendy Ju. 2015. Initiating interactions and negotiating approach: a robotic trash can in the field. In *AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction*. AAAI Press, 10–16.

[12] Maximiliane Frobenius. 2014. Audience design in monologues: How vloggers involve their viewers. *Journal of Pragmatics* 72 (2014), 59–72.

[13] Raphaela Gehle, Karola Pitsch, Timo Dankert, and Sebastian Wrede. 2017. How to open an interaction between robot and museum visitor? Strategies to establish a focused encounter in HRI. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 187–195.

[14] Lars Christian Jensen, Rosalyn Melissa Langedijk, and Kerstin Fischer. 2020. Understanding the Perception of Incremental Robot Response in Human-Robot Interaction. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 41–47.

[15] William Kristian Juel, Frederik Haarslev, Norbert Krüger, and Leon Bodenhagen. 2020. An Integrated Object Detection and Tracking Framework for Mobile Robots. In *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics - Volume 1: ICINCO*. SCITEPRESS Digital Library, 513–520.

[16] William K Juel, Frederik Haarslev, Eduardo R Ramírez, Emanuela Marchetti, Kerstin Fischer, Danish Shaikh, Poramate Manoonpong, Christian Hauch, Leon Bodenhagen, and Norbert Krüger. 2020. SMOOTH Robot: Design for a novel modular welfare robot. *Journal of Intelligent & Robotic Systems* 98, 1 (2020), 19–37.

[17] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I help you?-Design of human-like polite approaching behavior. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 35–42.

[18] Yadong Pan, Haruka Okada, Toshiaki Uchiyama, and Kenji Suzuki. 2015. On the reaction to robot's speech in a hotel public space. *International Journal of Social Robotics* 7, 5 (2015), 911–920.

[19] Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luise Sussenbach, Paul Luff, and Christian Heath. 2009. "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*. IEEE, 985–991.

[20] Md Golam Rashed, Dipankar Das, Yoshinori Kobayashi, and Yoshinori Kuno. 2018. A study on proactive methods for initiating interaction with human by social robots. *Asian Journal For Convergence In Technology (AJCT)* (2018).

[21] Sebastian Rottschäfer, Hendrik Buschmeier, Herwin van Welbergen, and Stefan Kopp. 2015. Online Lombard-adaptation in Incremental Speech Synthesis. In *Proceedings of INTERSPEECH 2015*. Dresden, Germany, 80–84.

[22] Harvey Sacks. 1984. Notes on methodology. *Structures of social action: Studies in conversation analysis* (1984), 21–27.

[23] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* (1974), 696–735.

[24] David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse* 2, 1 (2011).

[25] Richard Schulman. 1989. Articulatory dynamics of loud and normal speech. *The Journal of the Acoustical Society of America* 85, 1 (1989), 295–312. https://doi.org/10.1121/1.397737 arXiv:https://doi.org/10.1121/1.397737

[26] Chao Shi, Satoru Satake, Takayuki Kanda, and Hiroshi Ishiguro. 2018. A robot that distributes flyers to pedestrians in a shopping mall. *International Journal of Social Robotics* 10, 4 (2018), 421–437.

[27] Candace L Sidner, Christopher Lee, Cory Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence, 166(1-2):140-164* (2005).

[28] Stephen Yang, Brian Ka-Jun Mok, David Sirkin, Hillary Page Ive, Rohan Maheshwari, Kerstin Fischer, and Wendy Ju. 2015. Experiences developing socially acceptable interactions for a robotic trash barrel. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 277–284.