Findings from human-android dialogue research with ERICA

Divesh Lala, **Koji Inoue**, **Kenta Yamamoto** and **Tatsuya Kawahara**
Kyoto University Graduate School of Informatics
{lala,inoue,yamamoto}@sap.ist.i.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

Human-android interaction is a domain where dialogue management is combined with realistic humanoids. In this work we provide a summary of our dialogue research with the android ERICA. We provide an outline of what we have accomplished until now, with discussions of ERICA's dialogue management in several scenarios, both linguistic and non-linguistic. From formal experiments and informal commentary from users, we draw upon several findings during the project that should be considered with human-android dialogue research but can also be applied to other robots and agents.

1 Introduction

One of the ultimate goals of human-robot dialogue research is to create a robot which can autonomously hold a conversation at the level of a human being. Additionally, we may treat this robot as a complete social entity if it physically resembles one, such as an android. In this paper we review a long-running project on the development of an android named ERICA [Glas *et al.*, 2016]. The motivation behind ERICA is to produce a social robot which can be used for a number of different social roles which require differing types of dialogue.

ERICA's physical appearance is that of a woman in her 20s, as seen in Figure 1. We expect that ERICA's physical appearance shapes user expectations of her social behavior and dialogue [Kontogiorgos *et al.*, 2019; Haring *et al.*, 2013; Fong *et al.*, 2003] and that this project will add to the growing body of android research [Ishiguro, 2016]. ERICA is designed to physically resemble humans more than other android prototypes [Ramanathan *et al.*, 2019; Nishio *et al.*, 2007; Pioggia *et al.*, 2007; Oh *et al.*, 2006], and so our challenge in this project in terms of dialogue is to create autonomous dialogue management systems which support this physical realism.

In this short paper we present an overview of ERICA, the research that we have accomplished so far, and the findings that we have learned over time for human-android interaction. We present these findings for other researchers addressing situated dialogue with robots, in particular for more unstructured conversation. We note that our findings aren't based solely on what we have learned from formal experimentation. Informal demonstrations of ERICA have been performed in our laboratory and also in public, but there is always important feedback after these demonstrations about what users found impressive and what can be improved.

Figure 1: Situated interaction with ERICA.

2 System architecture

ERICA's full architecture is quite complex, so we will provide just a brief overview of the components related to dialogue and conversation. ERICA is controlled by a centralized system which is used to govern all aspects of the interaction. Figure 2 shows how this is managed and further details can be found in previous papers [Lala *et al.*, 2016].

The speaker is identified by isolating the direction of the audio source and combining this information with the visual feedback from the Kinect sensor. It is possible for ERICA to interact with multiple users, but for now we focus on one-to-one conversations. Audio, speech recognition results [Lee and Kawahara, 2009] and visual data are continuously streamed to the centralized controller, where decision making about ERICA's dialogue is made. Turn taking and backchannels models run independently from the controller but are integrated into the system.

As an android, users should talk with ERICA as if she were

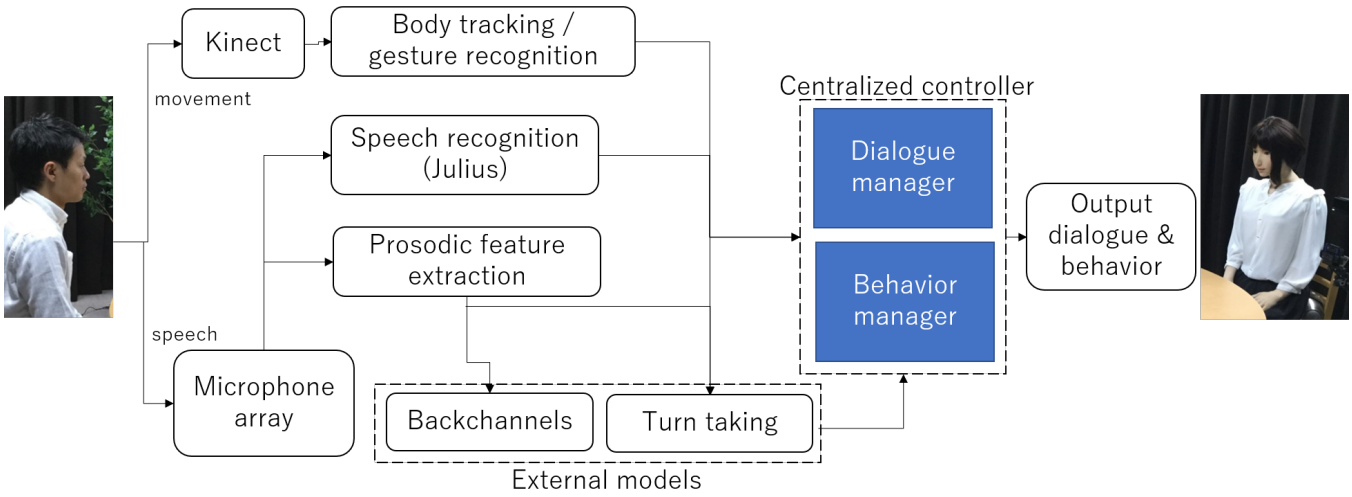

Figure 2: Overview of ERICA's dialogue behavior system architecture.

a human. There are no external devices or magic words required for speaking. The user also does not need to wear any microphones or devices to speak with ERICA as a microphone array will pick up their voice. The microphone array itself is disguised as a vase so that the user does not feel they have to speak towards it, ensuring that the environment of the conversation is as natural and unobtrusive as possible.

ERICA's text-to-speech is provided through a service which uses both synthesized and pre-recorded utterances which are based on a real voice actress chosen to match ERICA's appearance. The pre-recorded utterances contain a large variety of non-linguistic utterances such as fillers and backchannels. Most interactions with ERICA are done in Japanese. It is possible to use English with ERICA, however English-speaking ERICA does not have the variety of non-linguistic utterances which are present in the Japanese system.

3 Data collection for android interaction

Given the complex nature of ERICA's interaction environment, we discovered early on that the process used to design the models was crucial. We receive raw speech recognition results from Julius as an inter-pausal unit (IPU) and the 40 mel filterbank features associated with them. We also only have access to a particular set of streamed data, namely prosodic features of pitch and power and Kinect body information. Using pre-trained models was a possibility, but we found that often these were trained for different languages or required data streams (e.g. infra-red camera data) which we could not integrate easily with ERICA. This required us to develop our own conversation models given the constraints on our data inputs.

Towards this, we have created a large corpus containing over 200 sessions of human-android dialogues with ERICA. These dialogues consist of several scenarios (described in Section 5) and level of autonomy (fully autonomous or Wizard of Oz). Audio data is captured from a 16 channel microphone array, shotgun and pin microphones. Video data from numerous angles is recorded including recording the face of the user. The user's body data from Kinect is also recorded so that gesture recognition can be conducted.

Another intent of the data collection is to measure the user's behavior during conversation. We have created models to recognize laughter, nodding, backchannels and the direction of eye gaze [Inoue *et al.*, 2018]. These models have been used to predict the engagement of the user [Inoue *et al.*, 2019b]. One concern of training mutimodal models using corpora is that the reported performance differs in a live system due to different input methods or measurement devices. We were particularly careful in how much the corpus data could be replicated in ERICA's live system by only using features we knew we had access to in real time.

4 Non-linguistic behaviors

For ERICA, we have spent much time researching several aspects of conversation which are not related to the dialogue itself - backchanneling, turn-taking, gaze and shared laughter. Early on in the project we focused on creating models for these non-linguistic phenomena as we felt they were necessary to enhance the naturalness of the conversation. We now discuss some of these in more detail.

Backchannels are used as ERICA's primary listening behavior and was one of the earliest model we implemented in the system. The backchannel model is based only on prosodic features [Lala *et al.*, 2017] and so can be used in real-time. We found that backchannel timing was more important than the actual form used. In fact, we only use three backchannel forms which are selected at random.

Surprisingly, even this fairly simple model left a favorable impression on users. We found that listener behavior is necessary for embodied agents, at the very least it lessens silences in dialogue and ideally shows the robot is engaged and attentive. We are investigating backchannel models which can better predict the appropriate form, perhaps by using the utterances spoken by the user. Since we found timing to be

most important, we are aware of the need to create continuous models which have little delay.

Turn-taking is another major aspect which we found was needed for ERICA's system. In early demonstrations with ERICA we used a naive form of turn-taking using threshold of user silence, but we found this to be inadequate as ERICA either took unnaturally long pauses or had many false interruptions. Eventually we developed a model which uses acoustic and linguistic features to modify the silence time between turns [Lala *et al.*, 2018].

We could successfully quantify how much better this model was compared to our previous naive implementation and could use it without constant tuning. Furthermore, we also experimented with an extension of this model which is able to do overlapping turn-taking, through the use of fillers (e.g. "'Ah"') to grab the turn just as the user finishes speaking [Lala *et al.*, 2019a]. The intention was to make the turn-taking speed more natural, as overlapping turns often happen in real conversation [Stivers *et al.*, 2009]. We are aware of even more powerful models which can also do this [Roddy and Harte, 2020] so we may be able to improve this further.

Gaze is also highly related to turn-taking, particularly in the extended model. When we first developed ERICA, her gaze was always directed towards the user. Although we believed that this showed her interest in the conversation, for some users it was a little uncomfortable and unnatural. This is to be expected, as in real dialogues speaker and listener do not always engage in mutual gaze. We have recently been experimenting with other gaze models for attentive listening where ERICA limits mutual gaze until turn-switching occurs, similar to the patterns found in other research [Jokinen *et al.*, 2013; Andrist *et al.*, 2013].

Currently we are developing methods to have ERICA engage in shared laughter with the user to increase naturalness and engagement. Since we had previously developed a laughter recognition model [Inoue *et al.*, 2018], it was feasible to create a system that could respond to a laugh with a laugh of its own. We implemented such a system for ERICA and intend to conduct experiments related to this. From preliminary work we identified two main issues. The first issue is that shared laughter often overlaps so a continuous recognition model may be required. Secondly, the type of laugh (polite or mirthful) appears to be important [Tanaka and Campbell, 2014] and so we need a robust laughter selection method.

5 Dialogue scenarios and experiments

The goal of our project is for ERICA to act in several social roles that are sufficiently diverse in terms of the nature of the dialogue between ERICA and the user. Figure 3 shows roles which have been developed for ERICA so far.

Much of our research has been conducted on ERICA as an attentive listener and job interviewer. We have yet to implement sophisticated dialogue generation techniques such as BERT [Devlin *et al.*, 2018] because we have found success using relatively simpler models based on patterns we found by analyzing the corpus, rather than training a black-box dialogue generation model.

In the attentive listening system, the user talks at length

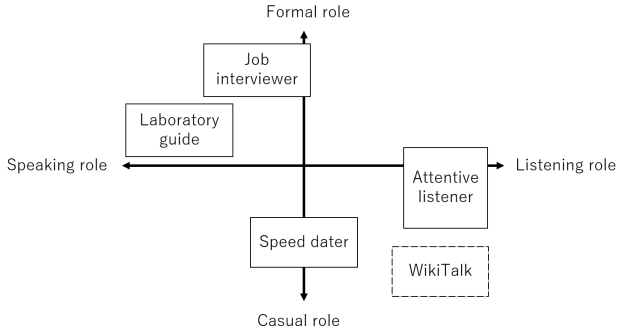

Figure 3: Roles that have been developed for ERICA. Axes are the formality of the robot's role (vertical) and what the role mostly entails (horizontal).

about any topic. ERICA's role is to listen to the user and interject with short questions or react in an empathetic manner. ERICA also uses backchannels when listening to show interest in the conversation. The motivation for this scenario is to provide social interaction with elderly people who may live alone. In theory ERICA can listen to any topic for any length of time as long as the user continues talking. Therefore attentive listening has a wide breadth of topics, but a shallow depth of individual topics.

The model we use is relatively simple - we identify the focus phrase in the user's utterance, then hypothesize about a possible question word that is used with that phrase [Inoue *et al.*, 2016; Inoue *et al.*, 2020b]. For example, if the user says "'I ate *pasta* yesterday"', ERICA's response would be "'What type of *pasta*?"'. ERICA uses sentiment responses to indicate positive or negative reactions to the user's utterance. Although the responses are rather limited, the power of the dialogue system is that it can handle a wide range of topics. We conducted experiments with this system and elderly people and found that the system could achieve basic listening skills [Inoue *et al.*, 2020b].

For the job interview system, ERICA acts as the interviewer providing questions to interviewees. Our motivation is to allow young job seekers to experience a job interview in a "safe environment" before undertaking the real thing. The interview is of a fixed structure. The dialogue system uses keywords uttered by the user and bases its response on how well the user answers the question. The system also asks the user to elaborate on certain key phrases. For example, if the user mentions that they have an interest in machine learning during the interview, ERICA asks them to provide more information about it, through a template-based response [Inoue *et al.*, 2019a].

The experiment we conducted involved comparing this dialogue system to a job interview with a fixed structure format, using ERICA as the robot for both conditions as well as comparing with a virtual agent interviewer to see the effect of embodiment [Inoue *et al.*, 2020a]. We found that users had a better impression of the interview and quality of questions with the proposed dialogue system. Furthermore, ERICA had more presence than the virtual agent interviewer. These results helped us confirm that ERICA could play a more formal

social role, as a contrast to attentive listening.

We also emphasize the relative importance of ERICA's behaviors in these two contrasting roles. Turn taking in attentive listening can be made faster - subjects quickly resume the conversation even if they are falsely interrupted since ERICA's responses are short, which cannot be done as a job interviewer. We also used nodding backchannels in the job interview rather than verbal backchannels, reflecting the more formal scenario.

The speed dating scenario is arguably the most difficult of all the scenarios, since it is a mixed-initiative dialogue and ERICA has to provide more depth to the conversation. Furthermore, ERICA should not only be able to handle many topics but should be capable of speaking about several of them in some depth. This is a scenario which perhaps requires a sophisticated language generation model and we are currently investigating techniques which can achieve this.

The other scenarios are not strictly conversations but show how ERICA can be used for different types of dialogue. One early scenario we created with ERICA was as a laboratory guide [Inoue *et al.*, 2016]. The user would ask questions about our laboratory and ERICA would provide the answer. The dialogue manager here was not so sophisticated, as it simply searched for keywords. However we used this scenario as a testbed for measuring human behavior during the dialogue to predict user engagement and then modifying ERICA's dialogue accordingly [Inoue *et al.*, 2019b]. We also combined ERICA with WikiTalk [Lala *et al.*, 2019b] where users ask ERICA to read out desired topic pages from Wikipedia. This shows how ERICA can also be used for a question-answer type role.

6 Key findings for android interaction

From our experiences with ERICA during this project, we have developed some findings which were found to be important in respect to android interaction. These rules can also apply to dialogue with non-android robots and virtual agents but we feel they are more pronounced in android interactions, due to the need for realism in situated interactions.

1. **The rhythm of conversation is as crucial as correct dialogue.** With situated android dialogue, our aim is to make the conversation as close as possible to human-human conversation. This also applies to the rhythm and flow of the interaction. Turn-taking which considers the timing of the responses plays a large role in maintaining conversation flow. Similarly, well-timed backchannels provide a rhythm in ERICA's listening behavior which gives the user feedback of the conversation. One outcome of our system design is that both these behaviors can be tuned to support different social roles. We have found that a more formal role such as a job interviewer requires more conservative turn-taking and less verbal backchannels. Both these models can be parameterized so we can easily tune these behaviors for ERICA. We have observed that this parameterization makes a large difference to the suitability of ERICA for her social role.
2. **Non-linguistic phenomena can support even simplistic dialogue models.** We receive many informal comments about the realism of ERICA's backchannels, often before comments about the dialogue itself. During this project we have found them to be the aspect that draw the most attention. Surprisingly, even though our backchannel model has not changed for a long time, it still draws interest even when our dialogue models have somewhat improved. This shows that good non-linguistic functionality can make up for an unsophisticated dialogue model.
3. **The novelty aspect makes analysis of subjective dialogue experiments difficult.** We found that our intuition about users' experiences with ERICA often do not match with their self reports. Particularly in our attentive listening studies, there are several sessions where ERICA hardly provides any meaningful responses to the user, but is rated highly for her dialogue skills. Our assumption is that for a lot of people, talking with an android is a new and enjoyable experience in itself, independent of what the android actually has to say. User evaluation of dialogue is difficult even in non-android systems but the additional factor of a humanoid robot is difficult to isolate in experiments.
4. **Short utterances are meaningful.** ERICA's text-to-speech system contains a large number of pre-recorded utterances such as fillers, backchannels and laughs as well as generic utterances such as "'I see"' (*sou desu ka* in Japanese) which are imbued with some emotion. In our attentive listening system the variety of short utterances appear to be an important feature for users. Even a simple "yes (*hai*)" carries additional meaning depending on its paralinguistic properties and we have the option to control the tone of the conversation by selecting the appropriate utterance.

The first two lessons are based on the idea that android dialogue management is not just about perfect natural language processing and conversation with an android is more than simply combining a chatbot system with a physical body. It also consists of parallel behaviors, verbal and non-verbal, that convey to the user that the android is engaged with the conversation and the dialogue matches the social affordances given by its physical appearance. This difference is also what sets aside androids from merely becoming fancy smart speakers and which we find ourselves focusing on more as ERICA matures.

7 Conclusion

In this paper we provided a brief summary of the research we have conducted in human-android dialogue research with ERICA. We showed the ability for ERICA to perform several social roles and the considerations we have to make when designing for different types of conversations. We provide some findings which are particularly pertinent for dialogue with androids and can hopefully generalize to other types of robots and agents.

Acknowledgments

This work was supported by JST ERATO Grant number JPMJER1401 and JSPS KAKENHI Grant number JP19H05691.

References

[Andrist *et al.*, 2013] Sean Andrist, Bilge Mutlu, and Michael Gleicher. Conversational gaze aversion for virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 249–262. Springer, 2013.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fong *et al.*, 2003] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[Glas *et al.*, 2016] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. Erica: The ERATO intelligent conversational android. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 22–29. IEEE, 2016.

[Haring *et al.*, 2013] Kerstin Sophie Haring, Katsumi Watanabe, and Celine Mougenot. The influence of robot appearance on assessment. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 131–132. IEEE, 2013.

[Inoue *et al.*, 2016] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. Talking with ERICA, an autonomous android. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 212–215, 2016.

[Inoue *et al.*, 2018] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Transactions on Signal and Information Processing*, 7, 2018.

[Inoue *et al.*, 2019a] Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. A job interview dialogue system with autonomous android erica, 2019.

[Inoue *et al.*, 2019b] Koji Inoue, Divesh Lala, Kenta Yamamoto, Katsuya Takanashi, and Tatsuya Kawahara. Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue, 2019.

[Inoue *et al.*, 2020a] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. Job interviewer android with elaborate follow-up question generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 324–332, New York, NY, USA, 2020. Association for Computing Machinery.

[Inoue *et al.*, 2020b] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 118–127, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[Ishiguro, 2016] Hiroshi Ishiguro. Android science. In *Cognitive Neuroscience Robotics A*, pages 193–234. Springer, 2016.

[Jokinen *et al.*, 2013] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30, 2013.

[Kontogiorgos *et al.*, 2019] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 133–140, 2019.

[Lala *et al.*, 2016] Divesh Lala, Pierrick Milhorat, Koji Inoue, Tianyu Zhao, and Tatsuya Kawahara. Multimodal interaction with the autonomous android erica. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 417–418, New York, NY, USA, 2016. Association for Computing Machinery.

[Lala *et al.*, 2017] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127–136, 2017.

[Lala *et al.*, 2018] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 78–86, 2018.

[Lala *et al.*, 2019a] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*, pages 226–234, 2019.

[Lala *et al.*, 2019b] Divesh Lala, Graham Wilcock, Kristiina Jokinen, and Tatsuya Kawahara. Erica and wikitalk. In *IJCAI*, pages 6533–6535, 2019.

[Lee and Kawahara, 2009] Akinobu Lee and Tatsuya Kawahara. Recent development of open-source speech recognition engine Julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pages 131–137. Asia-Pacific Signal and Information Processing Association, 2009 Annual . . . , 2009.

[Nishio *et al.*, 2007] Shuichi Nishio, Hiroshi Ishiguro, and Norihiro Hagita. Geminoid: Teleoperated android of an existing person. *Humanoid robots: New developments*, 14:343–352, 2007.

[Oh *et al.*, 2006] Jun-Ho Oh, David Hanson, Won-Sup Kim, Young Han, Jung-Yup Kim, and Ill-Woo Park. Design

of android type humanoid robot albert HUBO. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1428–1433. IEEE, 2006.

[Pioggia *et al.*, 2007] Giovanni Pioggia, ML Sica, Marcello Ferro, Roberta Igliozzi, Filippo Muratori, Arti Ahluwalia, and Danilo De Rossi. Human-robot interaction in autism: FACE, an android-based social therapy. In *RO-MAN 2007-the 16th IEEE international symposium on robot and human interactive communication*, pages 605–612. IEEE, 2007.

[Ramanathan *et al.*, 2019] Manoj Ramanathan, Nidhi Mishra, and Nadia Magnenat Thalmann. Nadine humanoid social robotics platform. In *Computer Graphics International Conference*, pages 490–496. Springer, 2019.

[Roddy and Harte, 2020] Matthew Roddy and Naomi Harte. Neural generation of dialogue response timings. *arXiv preprint arXiv:2005.09128*, 2020.

[Stivers *et al.*, 2009] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[Tanaka and Campbell, 2014] Hiroki Tanaka and Nick Campbell. Classification of social laughter in natural conversational speech. *Computer Speech & Language*, 28(1):314–325, 2014.