Multimodal Joke Generation and Paralinguistic Personalization for a Socially-Aware Robot

Hannes Ritschel, Thomas Kiderle, Klaus Weber, Florian Lingenfelser, Tobias Baur and Elisabeth André

Human-Centered Multimedia, Augsburg University, Universitätsstr. 6a, 86159 Augsburg, Germany
{ritschel,kiderle,weber,lingenfelser,baur,andre}@hcm-lab.de

Abstract. Robot humor is typically scripted by the human. This work presents a socially-aware robot which generates multimodal jokes for use in real-time human-robot dialogs, including appropriate prosody and non-verbal behaviors. It personalizes the paralinguistic presentation strategy based on socially-aware reinforcement learning, which interprets human social signals and aims to maximize user amusement.

Keywords: robot humor · non-verbal behavior · personalization.

1 Introduction

Humor increases interpersonal attraction and trust in interpersonal communication. It regulates conversations, eases communication problems and helps to cope with critique or stress [19]. Giving robots this ability is an opportunity to create socially intelligent embodied agents, but is also a serious challenge. Humor is complex and generative approaches are faced with many research questions: recognizing the context, estimating the appropriateness in the corresponding situation, generating the humorous content and communicating it successfully.

In the last years, some of these challenges have been investigated. Robots have been sent to theaters [13], presenting Japanese Manzai [11] and stand-up comedy for entertaining the human audience [12]. First steps in personalizing the show to the audience have been made [37], selecting presented contents intelligently according to the audience's visual or auditory reactions. Typically, humorous contents are scripted in advance, outsourcing this complex task to the human. But keeping the diversity of humor, interaction scenarios and dialog topics in mind, an automatic generation of robot humor is desirable. A combination of dynamically generated humor and personalization is desirable [24].

In the text-only domain, several humor generators have been implemented, such as STANDUP [15] for punning riddles. However, humor is multilayered [17]: apart from the text itself, appropriate non-verbal behaviors, such as facial expression and gestures, significantly contribute to successful joke presentation. Text-To-Speech (TTS) systems do not yet generate humor-specific prosody, nor does the robot add non-verbal behaviors automatically. When presenting humorous contents, it is all about timing, pronunciation and appearance, e.g. to

create tension just before telling the punchline of a joke. While recent research investigated the dynamic generation of multimodal ironic robot behaviors, one can find more humor markers in the literature which have not been applied systematically to robotic joke telling yet.

Machine learning is used successfully to adapt the robot's show to the spectators' preferences by selecting scripted contents intelligently. Reinforcement Learning (RL) has become very popular for optimizing social robots' linguistic contents [31,28,30] in recent years, also with focus on humor [37,38].

Building on the STANDUP punning riddle generator, we present an approach for dynamic multimodal joke generation. Our contribution is (1) the identification and (2) systematic implementation of human humor markers for a social robot. Furthermore, we (3) personalize its prosody based on the spectator's audiovisual reactions in real-time. This socially-aware learning process has the ultimate goal of increasing the individual spectator's amusement by tweaking multimodal joke presentation beyond its linguistic content.

2 Related Work

The related work is split up into two sections. First, we take a look at robots entertaining an audience, with focus on scenarios, contents and automated adaptation mechanisms, which help maximizing the spectators' amusement. Afterwards, we outline different humor *markers* in order to equip robots with a natural joke telling strategy. These multimodal social cues have been identified in human interaction to support joke telling and presenting humorous contents.

2.1 Robots Presenting Humor

A very popular comedy show is the traditional Manzai from Japan, which is performed by two entertainers. Hayashi et al. [11] implement a Manzai dialog with two robots. A noise level meter measures human applause and laughter. Both signals are transformed into an estimate of the audience's current amusement (*burst out*, *laugh*, *cool down*), which is used to synchronize the communication and movement of each robot with its comedy partner and the audience. Similar is done by Utemani et al. [34], where the content of the show is determined dynamically by keywords from the audience. After searching for newspaper articles on the internet, they are transformed into a Manzai dialog. The generated dialog coordinates both robots' movements, facial expressions and speech output.

Knight et al. [13] use a robot to present a sequence of scripted jokes. During its stand-up comedy, convex programming is used to adapt the presented content to the audience by measuring the current entertainment level. Laughter and applause are recorded with a stage microphone. The audience can evaluate jokes by holding up green or red cards, which are recorded with a camera. Based on this feedback, the robot dynamically selects the next jokes, which are associated with several attributes. Katevas et al. [12] focus on a robot's non-verbal behaviors during a stand-up comedy show. It presents scripted jokes and utilizes gaze

Table 1. Multimodal markers of verbal humor

Modality	**Markers**
Prosody	Pitch ↑[4][6][3][20][39], Volume ↑[4][6][39][1], Speech rate ↑[20][39][1], Break at punchline [4][6][3][1], Combination of limited pitch range, minor pitch change (syllables, utterance), Syntax and content in the setup of punning riddles [5]
Speech	Laughter [7][21][2]
Facial expr.	Smile [7][21][2] and gaze at the face areas involved in the smile (eyes, mouth) [9][8], Change gaze target to another person [12]

and gestures (e.g. pointing or looking at somebody) to react to acoustic and visual feedback from the audience. Based on the audience's reactions, which are captured with a camera (people's faces) and microphone (laughter/applause), the robot's timing, non-verbal behaviors and answers are adapted.

While aforementioned experiments address larger audiences, research in the context of single spectators optimizes the show for individual preferences. Weber et al. [37] adapt a robot's multimodal joke presentation to the human's sense of humor. The robot presents scripted jokes from different categories and combines them with sounds and grimaces. A reinforcement learning approach selects the best combination while the spectator's smile and laughter are used to shape the reward signal. Ritschel et al. [26] focus on verbal irony in human-robot smalltalk. Based on the user's input the robot dynamically transforms its response into an ironic version with the help of Natural Language Generation (NLG). The authors identify and add typical verbal (prosody) and non-verbal (gaze) robot behavior to successfully support the human in recognizing the presence of irony.

2.2 Multimodal Expression of Humor

The presentation of humor often involves appropriate prosody and non-verbal behaviors, which do not exist out-of-the box with current TTS systems and robots. For example, rolling eyes, winking, extra-long pauses and exaggerated intonational patterns, are crucial in order to support the human in identifying the presence of irony (see [26] for an overview of human irony cues and how to apply them to a social robot). In general, a robots' verbal and non-verbal behaviors have to be synchronized with the text. This also applies to humor: Mirnig et al. [17] point out that humor is multilayered and that the additional robot's modalities contribute to the presentation. According to the authors, adding unimodal verbal or non-verbal, humorous elements to non-humorous robot behavior does not automatically result in increased perceived funniness.

There is no comprehensive overview and transfer of human social cues to the context robot joke telling humor yet. Thus, a collection of *markers* (i.e., characersitics) for verbal humor is compiled in Table 1. It includes conversational humor and canned jokes, but excludes verbal irony. Common markers include increasing pitch, volume and speech rate, often combined with a break at the joke's punchline. Also, an "atypical" prosody with very limited pitch range, as

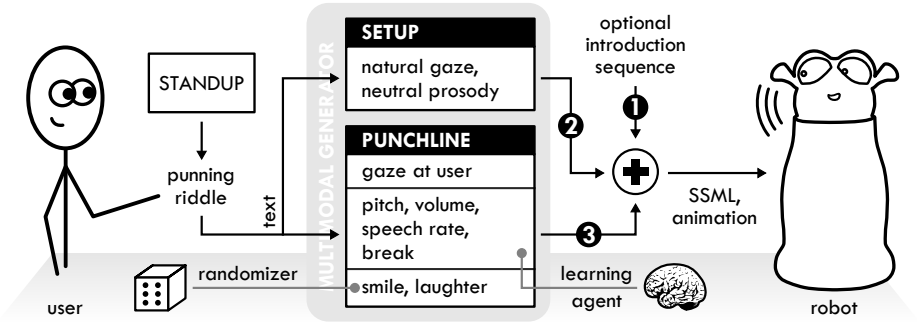

Fig. 1. Overview of the multimodal joke generation approach

well as special linguistic syntax for the setup of riddles is reported in the literature. Sometimes, short laughter or giggling is performed by the speaker. Facial expressions include smiling or targeted gaze behavior.

3 Multimodal Joke Generation

Figure 1 illustrates the general approach of the socially-aware robotic joke teller. A punning riddle is generated, which consists of two parts: the *setup* (question) and *punchline* (answer). Most of STANDUPs joke types are used (for details and examples see [29]). These two text snippets are transformed into a multimodal robot performance based on selected markers from Table 1. Speech Synthesis Markup Language (SSML)[1] is used to add prosody and to embed laughter sounds. The robot's face is animated to include gaze behavior and smile. In order to create variety in the created multimodal output, the parameters are randomized to a certain degree. The implementation uses a Reeti robot[2], which offers basic actuators in its face to control the eye ball rotation, eye lids, mouth, ears and head rotation. Since Reeti's internal TTS system does not support SSML the robot's speech is realized with Cerevoice[3], using the male *William* voice.

3.1 Text Generation

As an optional first step, the robot can add an introduction sequence, such as "Did you know this one?" or "The following punning riddle is a real pearl of comedy!" When embedded in a human-robot dialog scenario this aims to set the stage for the robot's performance by announcing its intention to tell a joke.

Afterwards, the setup and punchline are generated by STANDUP [15], which is a rule-based generator for punning riddles. It uses different schemas and templates in combination with information about pronunciation and semantic relationships of words. Puns can be generated for given topics or keywords. The

[1] https://www.w3.org/TR/speech-synthesis/
[2] http://reeti.fr/index.php/en/
[3] https://www.cereproc.com/en/products/academic

Listing 1.1. Generated SSML

```
<speak>
  <s>What do you get when you cross a choice with a meal?</s>
  <break time="1500ms"/>
  <s><prosody pitch="high" rate="fast" volume="loud">A pick-nic.</prosody></s>
  <spurt audio="g0001_019"></spurt>
</speak>
```

Fig. 2. The robot's gaze and facial expressions: a saccade (left), its neutral facial expression when centering on the spectator (middle) and smile (right).

following text templates are used in the scenario at hand: *cross* (e.g. "What do you get when you cross X with Y?"), *call* (e.g. "What do you call X that has Y?"), *difference* (e.g. "How is X different from Y?"), *similarity* (e.g. "Why is X like Y?") and *type* (e.g. "What kind of X is Y?"). Here is a generated example: "What do you call a washing machine with a September? An autumn-atic washer."

3.2 Setup

Prosody The setup's linguistic markers with regards to the syntax and content (see Table 1) are already applied during STANDUP's text generation process. In human joke telling the spoken language is typically accentuated with a combination of limited pitch range, minor pitch change within syllables or the whole utterance (see Table 1). However, since the realization of the presented markers heavily depends on the robot's hard- and software, the implemented prosodic markers are limited by the produced TTS output. Unfortunately, neither the SSML *range* nor the *emphasis* tag show an audible effect with Cerevoice and the *William* voice. Thus, the question's text is directly converted into SSML without additional tags (see the first sentence in Listing 1.1).

Gaze The robot's gaze behavior during the setup aims to mimic natural human gaze behavior in order to contrast the following punchline. Saccades are frequently implemented in embodied agents since they represent the most noticeable eye movements. They centre the gaze to an object of interest, which causes a rapid shift in eye rotation [32]. In order to mimic this behavior, the robot focuses random points near the spectator's position (see Figure 2) so that it does not stare at the user the whole time.

3.3 Punchline

Prosody Speakers typically take a significant break just before telling the punchline (see Table 1). In SSML pauses are specified with the *break* element and a value in milliseconds for best control over their duration. Since there is no clear information about its duration in the literature the robot uses a random value in the range between 1.5 and 2.0 seconds.

The punchline is often presented with a different pitch, volume and speech rate than the setup (see Table 1). The predefined SSML attribute values *low*, *medium* and *high* are used for pitch manipulation. For volume, the values *soft*, *medium*, *loud* work well. The speech rate is modified with *slow*, *medium* and *fast*. More extreme variants, such as *x-low* or *x-high* result in more synthetic and less natural sounding output. They impair the robot's comprehensibility, in particular compared to the neutral prosody during the setup.

Laughter The speaker occasionally marks the presented humor with laughing (see Table 1) or giggling after the joke. In Cerevoice *vocal gestures* can be embedded with the non-standard SSML *spurt* tag. These audio samples include different types of laughter, ranging from short giggling to long laughter sounds. When used excessively after every punchline, this appears probably unnatural for the audience, especially if the same sample is used over and over again. Based on the insights by Attardo et al. [2] the probability for laughing is set to 30 %. See Listing 1.1 for an SSML sequence with all markers included.

Gaze Joke tellers often gaze at the face areas involved in the spectator's smile (i.e., eyes and mouth) when presenting the punchline (see Table 1). While the robot mimics natural gaze behavior during the setup, the punchline is accompanied by its head and eyes focusing on the spectator. To this end, the robot's head and gaze center on the spectator in front (see Figure 2). We did not implement the speaker's change of gaze between different spectators, as observed and used in [12], since the scenario at hand addresses a single person audience.

Smile Smiling is a frequent human marker when presenting the punchline of a joke (see Table 1). Based on the insights by Attardo et al. [2] the robot uses this marker with a probability of 80 % by raising its lip corners. In order to emphasize the smile even more, the robot's large ears are raised (see Figure 2). After the joke is finished, the robot's face returns to its neutral facial expression.

4 Personalization

Previous experiments adapted the scripted content of robot comedy shows (see Section 2) and indicated that personalization can lead to significantly higher amusement. Building on this, we implement a socially-aware reinforcement learning approach [22] for optimizing the robot's paralinguistic joke presentation strategy beyond the linguistic content. While STANDUP provides the opportunity to generate different categories of jokes and since this type of adaptation

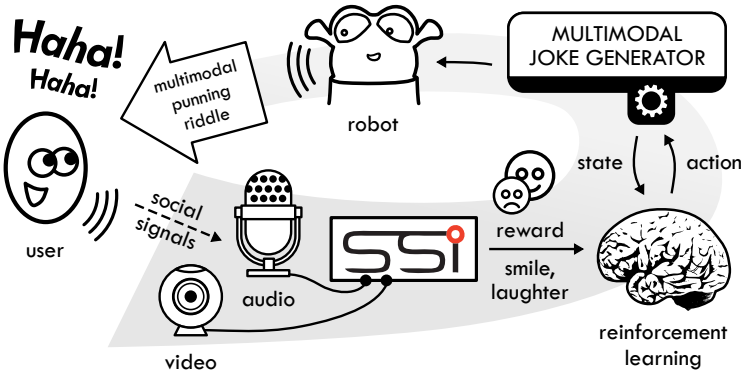

Fig. 3. Overview of the personalization process

has already been investigated we focus on the paralinguistic aspects exclusively in order to reduce the complexity for an evaluation and to prevent side effects.

4.1 Overview

Figure 3 illustrates the approach. A multimodal punning riddle is generated and presented. Meanwhile, the user's audiovisual reactions are recorded and interpreted with the Social Signals Interpretation (SSI) framework [36], which detects human laughter and smile. This data serves to compute the reward signal for a RL agent and to optimize the usage of markers for the next joke.

The presented process also aims to algorithmically improve the RL approach by [37] in terms of scalability. A RL agent needs to sequentially explore which action a_t is the best one to execute in a given state s_t according to the environment's reaction $\mathcal{R}_t$ at timestep t. The discrete set of actions $\mathcal{A}$ and discrete set of states $\mathcal{S}$ needs to be as small as possible in order to reduce the required learning time. In contrast to [37] and inspired by [27,23] we encode the robot's prosodic markers directly in the state space instead of modeling them as actions. This allows to model the learning task more compactly.

4.2 Problem Modeling

The robot should learn quickly since the first impression is crucial for the assessment of the audience [37] and preferences may change over time [27]. Therefore, our real-time personalization uses RL with linear function approximation, using a learning rate $\alpha = 0.25$ and discount factor $\gamma = 0.2$. Initially, the exploration rate is set to $\epsilon = 0.5$ and decreased by 0.05 after each time step t.

State Space Each state represents the robot's current prosodic presentation strategy. It encodes how the prosodic markers are applied during generation of the robot's multimodal performance, i.e. *pitch*, *speech rate*, *volume* and *break*.

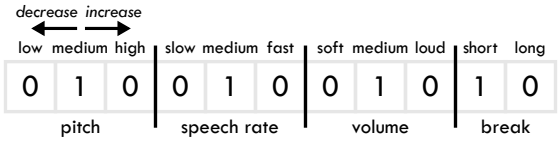

Fig. 4. Initial state (1 = active, 0 = inactive)

Smile and laughter are always used according to their probabilities (see Section 3.3). A state s_t is defined as a four-tuple $(pitch, speech\ rate, volume, break) \in \mathcal{S}$.

The state is converted into the vector $\boldsymbol{\phi}(s_t)$, which is divided into different sections (see Figure 4). All components in the vector are associated with a specific manifestation of the respective marker: digits are associated with the pitch attributes (*low*, *medium*, *high*), with the speech rate (*slow*, *medium*, *fast*), with the volume (*soft*, *medium*, *loud*) and with the length of the pause (*short*, *long*). Every marker is one-hot-encoded, i.e. only one manifestation per marker can be active. Figure 4 illustrates the initial state as an example.

Action Space Two actions exist for every marker: *increase* (↑) and *decrease* (↓), which allow increasing or decreasing pitch, speech rate, volume or break time. Moreover, the action *nop* does not change anything: when the optimal presentation strategy has been found no changes should be made anymore. Switching directly between minima and maxima is excluded by intention. Otherwise, this could result in the robot's behavior appearing strangely. Overall, the available action space $\mathcal{A}$ is defined as the set $\mathcal{A} = \{pitch\,\updownarrow, rate\,\updownarrow, volume\,\updownarrow, pause\,\updownarrow, nop\}$. Actions, which do not have an effect on the state, are excluded (apart from *nop*): if a marker is already set to its minimum or maximum value, it cannot be decreased or increased any further.

Reward As for the choice of rewarding feedback, laughter has for years been identified as a crucial part of social interaction by traditional conversation analysis [10]. Additionally it is the most evident reaction towards a successful punchline within a joke and is therefore a key element to estimate the user's amusement. Naturally, audible laughter is accompanied by a visual component, i.e., a smiling expression in the facial modality. These human social signals are processed (see Section 4.3) to compute the average probability of smiles $\mathbb{E}_{\text{smile}}$ and laughter $\mathbb{E}_{\text{laughter}}$ from the punchline until 2500 ms after the end of the joke. The additional time is essential to give the user time to understand the joke and to include delayed human reactions into the learning process [37]. Since smiles occur more frequently in humorous situations than laughter [2] the reward function $\mathcal{R}_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ at time step t is based on their weighted probabilities:

$$\mathcal{R}_t = \frac{3}{4} \cdot \mathbb{E}_{\text{smile}} + \frac{1}{4} \cdot \mathbb{E}_{\text{laughter}}.$$

Algorithm At every RL time step t, the robot selects one of the available actions $a_t \in \mathcal{A}$ according to state $s_t \in \mathcal{S}$, executes it, senses the user's reactions

and uses those obtained social signals to compute the reward signal $\mathcal{R}_t$. By employing linear function approximation, the robot has to learn a weight vector $\boldsymbol{\omega}$. The weight vector is used to compute a value $Q(s_t, a_t, \boldsymbol{\omega})$ for every action $a \in \mathcal{A}$ by calculating the dot product of the vector $\boldsymbol{\omega}$ and vectorial representation of the current state $\boldsymbol{\phi}(s_t)$:

$$Q(s_t, a_t, \boldsymbol{\omega}) := \boldsymbol{\phi}(s_t) \circ \boldsymbol{\omega}, \forall s_t \in \mathcal{S}, \forall a_t \in \mathcal{A}$$

In order to allow for learning non-linear dependencies between state values, we make use of the Fourier basis as described in Konidaris et al. [14]. Moreover, to find the optimal weight vector $\boldsymbol{\omega}$, the agent uses the reward $\mathcal{R}_t$ to update the weight vector $\boldsymbol{\omega}_t$ until the strategy converges to the optimal one [33]:

$$\Delta\boldsymbol{\omega}_t = \alpha\big(\mathcal{R}_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \boldsymbol{\omega}_t) - Q(s_t, a_t, \boldsymbol{\omega}_t)\big)\boldsymbol{\phi}(s_t)$$

4.3 Sensing social signals

First, we train a custom model offline for recognizing laughter from the audio modality and smiles from video images. To describe the paralinguistic content of voice, Mel Frequency Cepstral Coefficients (MFCC) spectral, pitch, energy, duration, voicing and voice quality features (extracted using the EmoVoice toolbox [35]) are employed. These features were used within an Support Vector Machine (SVM) model trained on excerpts of the Belfast Storytelling Database [16], which contains spontaneous social interactions and dialogs with a laughter focused annotation. Person independent evaluation of the model on the training database showed an unweighted accuracy of 84% for the recognition of laughter frames. For detecting smiles in the video, we apply transfer learning to fine-tune a deep convolutional neural network (VGGFace) by retraining it on the AffectNet facial expression corpus [18].

Based on the trained models audiovisual laughter recognition is carried in real-time during the interaction. The robot continuously captures the spectator's social signals with a headset microphone and webcam and analyzes them with the SSI framework [36]. Bursts of laughter are detected on a frame by frame basis: the audio signal is analyzed within a one-second sliding window that is shifted every 400 milliseconds, resulting in a decision rate of 2.5 Hz. The overall activity is monitored by applying a voice activity transformation to the signals via hamming windowing and intensity calculation. Coherent signal parts (i.e. frames) in which the mean of squared input values – multiplied by a Hamming window – that exceed predefined thresholds for intensity are identified as carriers of vocal activity and therefore serve as input for feature calculation and subsequent classification. Video is captured at a rate of 15 frames per second. Each frame is classified with the neural network model described above and the probabilistic results are averaged with the same sliding window as the audio modality to gain equally clocked classification results from both input signals.

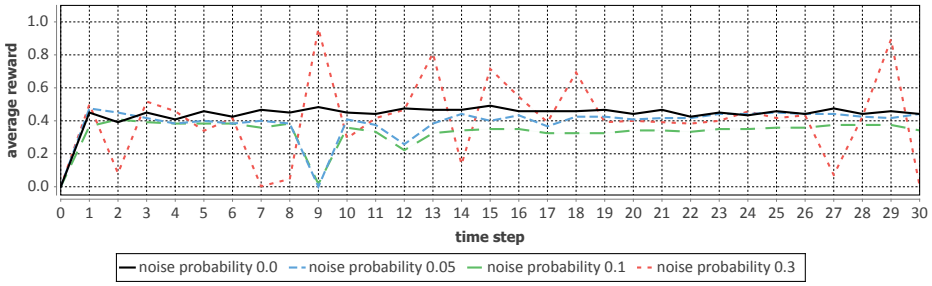

Fig. 5. Simulation results

4.4 Simulation

Several experiments were conducted as a first evaluation of the personalization approach. This is important to check whether the learning approach is implemented correctly and able to adapt to human preferences algorithmically.

Simulated User Each artificial spectator is initialized with a random preference with regard to the robot's pitch, speech rate, volume and break. These values are unknown by the learning agent, which uses the approach from Section 4.2 to find the best presentation strategy. While the simulation does not use the real social signal processing component from Section 4.3 the reward is calculated based on simulated amusement. If a feature from the state space matches the simulated user's actual preference, the neutral reward of 0 is increased by 0.25.

Noise In general, a simulation cannot emulate realistic human behavior. Inspired by [27] we address this issue by adding two kinds of noise: (1) non-deterministic user reactions and (2) sensor hardware and processing noise of the social signal component. The first aims to randomize the spectator's amusement, which will be dependent on more than the robot's joke presentation strategy in real interaction. This is realized by adding a random value the interval $[-1.0, 1.0]$ to the reward. The second addresses noise from the camera and microphone, which result in a wrong interpretation of the sensed human social signals.

Results The plots in Figure 5 averages over 30 trials, each consisting of 30 time/learning steps. This is analogous to a study with 30 participants with the robot telling 30 jokes to each of them. The learning task is non-episodic for each artificial user: there are no terminal states, the agent is provided no initial knowledge and the learned policy is reset between each trial. Performance is evaluated for 0 % (baseline), 5 %, 10 % and 30 % of noise, which randomizes the reward as described above. Learning without noise results in a pretty stable reward by about 0.5. With increasing noise the overall performance decreases as expected. In average, the reward is still very similar to the baseline most of the time, which indicates that the learning approach is able to cope with noise and outperforms table-based reinforcement learning approaches, such as in [27].

5 Conclusion

We have presented a multimodal joke generation approach for social robots. After identifying appropriate human paralinguistic and non-verbal cues from the literature we provided details on how to implement them for the embodied agent, including gaze, prosody, smile and laughter. Furthermore, we have introduced and simulated a reinforcement learning approach to personalize the robot's paralinguistic presentation strategy for the individual spectator, who is recorded with a microphone and a webcam during the show. This input is analyzed by a deep convolutional neural network and a Support Vector Machine to detect human visual smiles and audible laughter, which serve as a reward for the reinforcement learning process, aiming to maximize human amusement.

Our ultimate goal is to evaluate and to embed this socially-aware generation and personalization process in human-robot dialog, where the variety of conversation topics require to dynamically generate humorous contents on-the-fly. We believe that in real-world interaction scenarios, augmenting the robot with an adaptive artificial sense of humor will increase perceived social intelligence and thus overall result in an improved interaction experience. Future work will also investigate whether the generation and personalization of appropriate non-verbal sounds [25] is able to support a robot's humor presentation, too.

Acknowledgments

This research was funded by the European Union PRESENT project, grant agreement No 856879.

References

1. Archakis, A., Giakoumelou, M., Papazachariou, D., Tsakona, V.: The prosodic framing of humour in conversational narratives: Evidence from greek data. Journal of Greek Linguistics **10**(2), 187–212 (2010)
2. Attardo, S., Pickering, L., Baker, A.: Prosodic and multimodal markers of humor in conversation. Pragmatics & Cognition **19**(2), 224–247 (2011)
3. Audrieth, A.L.: The art of using humor in public speaking. Retrieved March **20**, 2005 (1998)
4. Bauman, R.: Story, performance, and event: Contextual studies of oral narrative, vol. 10. Cambridge University Press (1986)
5. Bird, C.: Formulaic jokes in interaction: The prosody of riddle openings. Pragmatics & Cognition **19**(2), 268–290 (2011)
6. Chafe, W.: Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. University of Chicago Press (1994)
7. Gironzetti, E.: Prosodic and multimodal markers of humor. The Routledge handbook of language and humor pp. 400–413 (2017)
8. Gironzetti, E., Attardo, S., Pickering, L.: Smiling, gaze, and humor in conversation: A pilot study. Metapragmatics of Humor: Current research trends **14**, 235 (2016)
9. Gironzetti, E., Huang, M., Pickering, L., Attardo, S.: The role of eye gaze and smiling in humorous dyadic conversations (03 2015)

10. Glenn, P.J.: Initiating shared laughter in multi-party conversations. Western Journal of Communication (includes Communication Reports) **53**(2), 127–149 (1989)
11. Hayashi, K., Kanda, T., Miyashita, T., Ishiguro, H., Hagita, N.: Robot manzai: Robot conversation as a passive–social medium. International Journal of Humanoid Robotics **5**(01), 67–86 (2008)
12. Katevas, K., Healey, P.G., Harris, M.T.: Robot comedy lab: experimenting with the social dynamics of live performance. Frontiers in psychology **6**, 1253 (2015)
13. Knight, H.: Eight lessons learned about non-verbal interactions through robot theater. In: International Conference on Social Robotics. pp. 42–51. Springer (2011)
14. Konidaris, G.D., Osentoski, S., Thomas, P.S.: Value function approximation in reinforcement learning using the fourier basis. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011. AAAI Press (2011)
15. Manurung, R., Ritchie, G., Pain, H., Waller, A., O'Mara, D., Black, R.: The construction of a pun generator for language skills development. Applied Artificial Intelligence **22**(9), 841–869 (2008)
16. McKeown, G., Curran, W., Wagner, J., Lingenfelser, F., André, E.: The belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In: Affect. Comp. and Int. Interaction. pp. 166–172. IEEE (2015)
17. Mirnig, N., Stollnberger, G., Giuliani, M., Tscheligi, M.: Elements of humor: how humans perceive verbal and non-verbal aspects of humorous robot behavior. In: International Conference on Human-Robot Interaction. pp. 211–212. ACM (2017)
18. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
19. Nijholt, A.: Conversational Agents and the Construction of Humorous Acts, chap. 2, pp. 19–47. Wiley-Blackwell (2007)
20. Norrick, N.R.: On the conversational performance of narrative jokes: Toward an account of timing. Humor **14**(3), 255–274 (2001)
21. Pickering, L., Corduas, M., Eisterhold, J., Seifried, B., Eggleston, A., Attardo, S.: Prosodic markers of saliency in humorous narratives. Discourse processes **46**(6), 517–540 (2009)
22. Ritschel, H.: Socially-aware reinforcement learning for personalized human-robot interaction. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018. pp. 1775–1777. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM (2018)
23. Ritschel, H., André, E.: Real-time robot personality adaptation based on reinforcement learning and social signals. In: Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017. pp. 265–266. ACM (2017)
24. Ritschel, H., André, E.: Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning. In: Proceedings of the Workshop on NLG for Human–Robot Interaction. pp. 12–16 (2018)
25. Ritschel, H., Aslan, I., Mertes, S., Seiderer, A., André, E.: Personalized synthesis of intentional and emotional non-verbal sounds for social robots. In: 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019. pp. 1–7. IEEE (2019)
26. Ritschel, H., Aslan, I., Sedlbauer, D., André, E.: Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans.

In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 86–94. AAMAS '19, IFAAMAS (2019)
27. Ritschel, H., Baur, T., André, E.: Adapting a robot's linguistic style based on socially-aware reinforcement learning. In: 26th IEEE International Symposium on Robot and Human Interactive Communication. pp. 378–384. IEEE (2017)
28. Ritschel, H., Janowski, K., Seiderer, A., André, E.: Towards a robotic dietitian with adaptive linguistic style. In: Joint Proceeding of the Poster and Workshop Sessions of AmI-2019, the 2019 European Conference on Ambient Intelligence, Rome, Italy, November 13-15, 2019. CEUR Workshop Proceedings, vol. 2492, pp. 134–138. CEUR-WS.org (2019)
29. Ritschel, H., Kiderle, T., Weber, K., André, E.: Multimodal joke presentation for social robots based on natural-language generation and nonverbal behaviors. In: Proceedings of the 2nd Workshop on NLG for Human–Robot Interaction (2020)
30. Ritschel, H., Seiderer, A., Janowski, K., Aslan, I., André, E.: Drink-o-mender: An adaptive robotic drink adviser. In: Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction. pp. 3:1–3:8. MHFI'18, ACM (2018)
31. Ritschel, H., Seiderer, A., Janowski, K., Wagner, S., André, E.: Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2019, Island of Rhodes, Greece, June 5-7, 2019. pp. 247–255 (2019)
32. Ruhland, K., Andrist, S., Badler, J.B., Peters, C.E., Badler, N.I., Gleicher, M., Mutlu, B., McDonnell, R.: Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In: Eurographics 2014 - State of the Art Reports. pp. 69–91 (2014)
33. Sutton, R.S., Maei, H.R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., Wiewiora, E.: Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 993–1000. ACM (2009)
34. Umetani, T., Nadamoto, A., Kitamura, T.: Manzai robots: entertainment robots as passive media based on autocreated manzai scripts from web news articles. Handbook of Digital Games and Entertainment Technologies pp. 1041–1068 (2017)
35. Vogt, T., André, E., Bee, N.: Emovoice - A framework for online recognition of emotions from voice. In: Perception in Multimodal Dialogue Systems. pp. 188–199 (2008)
36. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time. In: 21st International Conf. on Multimedia. pp. 831–834. ACM (2013)
37. Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., André, E.: How to shape the humor of a robot - social behavior adaptation based on reinforcement learning. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 154–162. ICMI '18, ACM (2018)
38. Weber, K., Ritschel, H., Lingenfelser, F., André, E.: Real-time adaptation of a robotic joke teller based on human social signals. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018. pp. 2259–2261. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM (2018)
39. Wennerstrom, A.: The music of everyday speech: Prosody and discourse analysis. Oxford University Press (2001)