Multimodal Joke Presentation for Social Robots based on Natural-Language Generation and Nonverbal Behaviors

Hannes Ritschel, Thomas Kiderle, Klaus Weber and Elisabeth André
Human-Centered Multimedia, Augsburg University
Augsburg, Germany
{ritschel,kiderle,weber,andre}@hcm-lab.de

ABSTRACT

Natural Language Generation is a key technology for generating dynamic, humorous contents in human-robot interaction. Due to their embodiment, many social robots offer expressive modalities to support the spoken language. First steps have been taken to synchronize nonverbal and paralinguistic behaviors with conversational humor. This work outlines a multimodal approach for augmenting generated text-based punning riddles with appropriate facial expression, gaze, prosody and laughter for a social robot.

1 BACKGROUND

Natural Language Generation (NLG) is an essential tool for providing a natural human-robot interaction experience [7]. It opens up the ability to react and generate content dynamically during conversations. In addition to the flexbility with respect to the content itself, a generation approach allows to control the robot's linguistic style, its use of figures of speech, stylistic instruments and other language tools – characteristics, which contribute to the perceived personality. One important aspect closely linked to language and personality is humor. It regulates conversations, increases interpersonal attraction and trust in human-human interactions. In the context of conversational agents, it makes interactions more natural, enjoyable and increases credibility and acceptance [15]. For embodied agents, such as social robots, humorous contents are typically scripted, such as in [11–13, 27]. It is desirable to use a generative approach for generating humorous contents during interaction and to address the variety of humor. Many experiments for text-based generation have already been made, including the STANDUP [14] generator for punning riddles. However, text alone is not sufficient for a robot's convincing humor presentation. Appropriate paralinguistic and nonverbal behavior must be added to the text, taking facial expression, gaze, gestures and laughter into account. In this manner, multimodal generation and expression of irony based on NLG and tailored behavior has recently been explored for a *Reeti* robot [22], where linguistic, prosodic and nonverbal markers help the human to identify the robot's use of irony.

The following sections present an approach for augmenting generated textual jokes with nonverbal behaviors from the literature for a social robot. Resulting in a multimodal joke presentation, it aims to be embedded in a human-robot dialog to generate humorous contents on-the-fly.

2 MULTIMODAL JOKE GENERATION

The multimodal joke telling robot is illustrated in Figure 1. It uses the STANDUP [14] joke generator to create a textual punning riddle, which consists of a *setup* (question) and a *punchline* (answer).

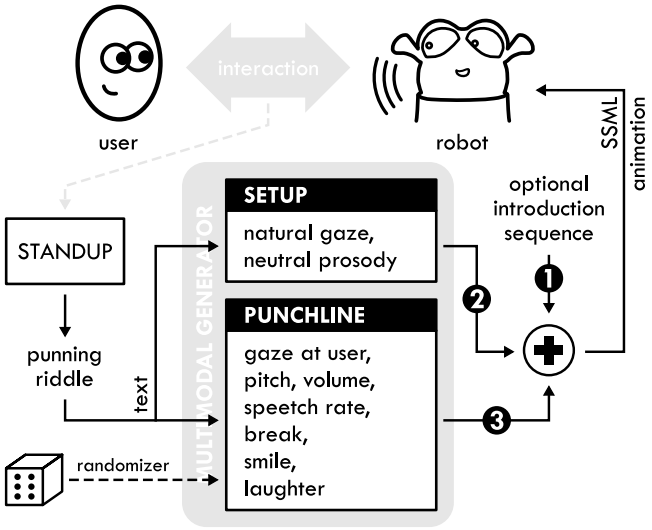

Figure 1: Multimodal joke generation approach

Multimodal markers from the literature are added for the robot's performance, including prosody, laughter, smile and gaze. Their use is randomized to a certain degree in order to keep variety in the generated behavior. As in [22], a *Reeti* robot with an expressive face is used as output medium. Since it does not have arms or legs, gestures cannot be taken into account. Audible tweaks, such as the prosody, are implemented with the Speech Synthesis Markup Language (SSML)[1], which is rendered to audio by the Cerevoice[2] Text-to-Speech (TTS) system using the male *William* voice.

2.1 Text

The NLG part uses STANDUP [14], which generates different types of punning riddles: *cross*, *call*, *difference*, *similarity* or *type*. See Table 1 for a list of examples. It requires three different inputs: *schemas*, *description rules*, and *text templates*, which describe linguistic requirements, provide guidelines for the formulation and determine how the generated contents are aggregated and combined. Since it is common to announce a joke in conversations [2] an introduction sequence can be added, e.g. "Did you know this one?" or "The following punning riddle is a real pearl of comedy!" Afterwards, a randomized set of multimodal cues (see below) is applied to the generated joke with a rule-based approach. The NLG component does not decide or interact with the mutlimodal expression component.

[1] https://www.w3.org/TR/speech-synthesis/
[2] https://www.cereproc.com/en/products/academic

Table 1: Utilized STANDUP joke types

Joke type	Riddle structure
cross	What do you get when you cross X with Y?
call	What do you call a cross between X and Y?
	What do you call X that has Y ?
	What do you call X with Y?
	What do you call X?
difference	What is the difference between X and Y?
	Why is X different from Y that is Z?
	Why is X different from Y?
	How is X different from Y?
similarity	What does X and Y have in common?
	Why is X like Y?
	How is X like Y?
type	What kind of X has Y?
	What kind of X is Y?

Listing 1: Generated SSML

```
<?xml version="1.0" encoding="UTF-8"?>
<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>
```

2.2 Prosody

A combination of limited pitch range, minor pitch change within syllables or the whole utterance [5] is typical when presenting the setup of punning riddles. Unfortunately, the SSML *range* and *emphasis* tag have no audible effect with Cerevoice and the *William* voice. Thus, the joke's question is converted into SSML without additional tags (see first sentence in Listing 1). With regard to timing, a break is typically taken just before telling the punchline [1, 3, 4, 6]. The SSML *break* element is used to insert a short break by specifying the duration in milliseconds. A random value between 1500 ms and 2000 ms is used.

The punchline is often presented with a higher pitch [3, 4, 6, 16, 29], volume [1, 4, 6, 29] and speech rate [1, 16, 29] than the setup. This can be realized by setting the *pitch* attribute to the predefined value *high*. Similarly, the *volume* and *rate* attribute can be set to *loud* or *fast*, respectively. More extreme settings (e.g. *x-loud*, *x-fast*, etc.) sound less natural and impact the robot's comprehensibility.

2.3 Gaze

During the setup, the robot aims to mimic natural gaze behavior by using saccades (see Figure 2), which centre its gaze to an object of interest [26]. This contrasts the following punchline, during which the robot's head and eyes focus on the spectator, as [9, 10] observe that joke tellers gaze at the face areas involved in the spectator's smile (i.e., eyes and mouth) when presenting the punchline.

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

2.4 Laughter

Laughing and giggling are sometimes expressed by the speaker after telling a joke [2, 8, 17]. The *vocal gestures* offered by the Cerevoice TTS system include a list of such samples, ranging from short giggling to long laughter sounds, which can be embedded with the non-standard SSML *spurt* tag. The samples are randomized since it may appear unnatural for the audience if the same sound is used exessively, 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 %.

2.5 Smile

A frequent human marker when presenting the punchline of a joke is smiling [2, 8, 17]. Based on the study by Attardo et al. [2] the robot uses this humor marker with a probability of 80 %: the robot raises its lip corners just before it starts telling the punchline. In order to emphasize the smile even more, the robot's large ears are raised (see Figure 2). If this marker is not used the robot shows a neutral facial expression.

3 CONCLUSION

The presented approach allows transforming text-based, dynamically generated punning riddles into a multimodal robot performance. It is implemented as a rule-based approach which mimics human humor markers for the robot by adding nonverbal and paralinguistic behaviors after generating the text. The presentation of both text and facial expression is synchronized. In future work, the aim is to embed this process in a human-robot interaction scenario, where keywords from the conversation could be fed into the joke generator to create thematically linked jokes on-the-fly.

Due to individual preferences and equivocal insights with regard to humor markers reported in the literature, the personalization of the robot's joke presentation is of interest. While [11–13, 27] manipulate the selection of scripted contents, the focus will be on tweaking the multimodal presentation of the robot's (non-)verbal and paralinguistic behaviors. In this context a socially-aware Reinforcement Learning approach, such as in [18–20, 23–25, 28], will be investigated. Furthermore, the generation of appropriate sounds may support the robot's joke presentation, too [21].

ACKNOWLEDGMENTS

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

REFERENCES

[1] Argiris Archakis, Maria Giakoumelou, Dimitris Papazachariou, and Villy Tsakona. 2010. The prosodic framing of humour in conversational narratives: Evidence from Greek data. *Journal of Greek Linguistics* 10, 2 (2010), 187–212.

[2] Salvatore Attardo, Lucy Pickering, and Amanda Baker. 2011. Prosodic and multimodal markers of humor in conversation. *Pragmatics & Cognition* 19, 2 (2011), 224–247.

[3] Anthony L Audrieth. 1998. The art of using humor in public speaking. *Retrieved March* 20 (1998), 2005.

[4] Richard Bauman. 1986. *Story, performance, and event: Contextual studies of oral narrative*. Vol. 10. Cambridge University Press.

[5] Christy Bird. 2011. Formulaic jokes in interaction: The prosody of riddle openings. *Pragmatics & Cognition* 19, 2 (2011), 268–290.

[6] Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

[7] Mary Ellen Foster, Hendrik Buschmeier, and Dimitra Gkatzia. 2018. Proceedings of the Workshop on NLG for Human–Robot Interaction. In *Proceedings of the Workshop on NLG for Human–Robot Interaction*.

[8] Elisa Gironzetti. 2017. Prosodic and multimodal markers of humor. *Attardo, S.(ed.)* (2017), 400–413.

[9] Elisa Gironzetti, Salvatore Attardo, and Lucy Pickering. 2016. Smiling, gaze, and humor in conversation: A pilot study. *Metapragmatics of Humor: Current research trends* 14 (2016), 235.

[10] Elisa Gironzetti, Meichan Huang, Lucy Pickering, and Salvatore Attardo. 2015. The Role of Eye Gaze and Smiling in Humorous Dyadic Conversations.

[11] Kotaro Hayashi, Takayuki Kanda, Takahiro Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Robot manzai: Robot conversation as a passive–social medium. *International Journal of Humanoid Robotics* 5, 01 (2008), 67–86.

[12] Kleomenis Katevas, Patrick Healey, and Matthew Harris. 2014. Robot Stand-up: Engineering a Comic Performance.

[13] Heather Knight, Scott Satkin, Varun Ramakrishna, and Santosh Divvala. 2011. A savvy robot standup comic: Online learning through audience tracking. In *Workshop paper (TEI'10)*.

[14] Ruli Manurung, Graeme Ritchie, Helen Pain, Annalu Waller, Dave O'Mara, and Rolf Black. 2008. The construction of a pun generator for language skills development. *Applied Artificial Intelligence* 22, 9 (2008), 841–869.

[15] Anton Nijholt. 2007. *Conversational Agents and the Construction of Humorous Acts*. Wiley-Blackwell, Chapter 2, 19–47.

[16] Neal R Norrick. 2001. On the conversational performance of narrative jokes: Toward an account of timing. *Humor* 14, 3 (2001), 255–274.

[17] Lucy Pickering, Marcella Corduas, Jodi Eisterhold, Brenna Seifried, Alyson Eggleston, and Salvatore Attardo. 2009. Prosodic markers of saliency in humorous narratives. *Discourse processes* 46, 6 (2009), 517–540.

[18] Hannes Ritschel. 2018. 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*. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 1775–1777.

[19] Hannes Ritschel and Elisabeth André. 2017. 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*. ACM, 265–266.

[20] Hannes Ritschel and Elisabeth André. 2018. 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*. 12–16.

[21] Hannes Ritschel, Ilhan Aslan, Silvan Mertes, Andreas Seiderer, and Elisabeth André. 2019. 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*. IEEE, 1–7.

[22] Hannes Ritschel, Ilhan Aslan, David Sedlbauer, and Elisabeth André. 2019. 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 (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, 86–94.

[23] Hannes Ritschel, Tobias Baur, and Elisabeth André. 2017. Adapting a Robot's linguistic style based on socially-aware reinforcement learning. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*. IEEE, 378–384.

[24] Hannes Ritschel, Kathrin Janowski, Andreas Seiderer, and Elisabeth André. 2019. 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. CEUR-WS.org, 134–138.

[25] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. 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*. 247–255.

[26] Kerstin Ruhland, Sean Andrist, Jeremy Badler, Christopher Peters, Norman Badler, Michael Gleicher, Bilge Mutlu, and Rachel Mcdonnell. 2014. 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*. 69–91.

[27] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. 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 (ICMI '18)*. ACM, 154–162.

[28] Klaus Weber, Hannes Ritschel, Florian Lingenfelser, and Elisabeth André. 2018. 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*. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2259–2261.

[29] Ann Wennerstrom. 2001. *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.