KNOCK-KNOCK! WHO'S THERE? THE LAUGHTER-ENHANCED VIRTUAL REAL-ESTATE AGENT

Bogdan Ludusan, Petra Wagner

Fakultät für Linguistik und Literaturwissenschaft
Center for Cognitive Interaction Technology
Universität Bielefeld
bogdan.ludusan@uni-bielefeld.de

Abstract: We present here an investigation into the use of social laughter in human-machine interaction. We employed an overhearer study, in which the interaction between a virtual real-estate agent and a client visiting an apartment was evaluated. Two conditions were considered, without and with laughter, the only difference between the two cases being the inclusion of laughter, at appropriate locations in the agent's speech, in the latter case. A large pool of participants listened to the recording corresponding to one of the two conditions, while watching a slideshow of the visited apartment. They were asked to judge the interaction of the agent, based on the following dimensions: professionalism, communication, pleasantness, formality and spontaneity, as well as expressing the likelihood of recommending the apartment and the agent to acquaintances. The results showed that the scores of the five dimensions decreased while the apartment and the agent ratings increased, in the laughter case, but only the formality difference reached significance. Nevertheless, a linear regression analysis showed that condition had an effect on the ratings, mainly through interactions with other factors. We discuss these findings and propose further directions of research to be followed.

1 Introduction

Being one of the most often encountered non-verbal vocalisations in spoken interaction [1], laughter is a ubiquitous phenomenon in human conversation, and has been found to benefit conversational success in multiple ways [2]: As a means to express or establish the social setting of an ongoing interaction, it may indicate closeness or affiliation between interaction partners, or convey a feeling of pleasantness with regards to an interlocutor. Further functions of laughter include the softening of previous affirmations, as a remedy of possible interactional misunderstandings, or helping to overcome difficult situations, e.g., involving embarrassment or discomfort. Beyond its function in the expression of social roles, laughter has been found to convey information about discourse structure [3, 4] and to express pragmatic functions [5, 6]. Recently, it has also been suggested that laughter has propositional meaning and can interact with content emanating from spoken language [7].

Despite a general consensus on the widespread use and range of functions that laughter fulfills in in human interaction, their usefulness and optimal integration in human-machine applications is still relatively under-studied (e.g. [8, 9]), and the existing studies tend to focus on the social connection role of laughter rather than communicative aspects of human-machine interaction (e.g., [10]). One underlying reason for this may be that laughter is still difficult to synthesize naturally and convincingly [11], presumably due to its high variability in form and distribution [12]. In fact, for other characteristics typical of spontaneous human interaction such as speech disfluencies, it has been found that their integration into dialogue systems may lead to an increased task performance, but also to lower acceptability ratings if they are not synthesized convincingly. Thus, the communicative benefit of certain dialogue features not necessarily goes hand in hand with subjective user experience in human-machine settings [13].

Based on the above-mentioned findings, we conducted an empirical study that aims to to tackle the potential communicative benefit of laughter present in a virtual agent's speech. The study manipulated the presence of social laughter as a between-subjects variable, and was performed using a crowdsourced overhearer design. Within this approach, we tested the following hypothesis:

> *A virtual agent that uses conversational laughter is perceived as more successful in the interaction with a human than a virtual agent that does not use laughter.*

2 Methods

As the behavioural quality of artificial characters or voices cannot be meaningfully evaluated in a non-contextualized fashion [14], our hypothesis was tested by embedding laughter events in an interactive scenario. Rather than integrating laughter in a dialogue system, we report a crowdsourced overhearer study in which the participants watch a simulated conversation between a virtual real-estate agent and a human client while visiting an apartment together. Our participants either listen to a conversation containing laughter, or not. Our study design makes use of the well-known advantage of the Wizard-of-Oz paradigm, i.e., a straightforward control of the tested independent variable (here: presence or absence of laughter).

Our design decision for a crowdsourced overhearer study does not enable us to measure a number of crucial quality indicators typically assessed in human-machine interaction studies, e.g., task performance or subjective user experience. However, it comes with a number of advantages: It allows us for easily gathering multiple judgements on interaction quality of a single data point, i.e., a single interaction [15], and simplifies an examination of potential user-centered predictors of interaction quality (e.g., gender, prior experience with speech technology).

2.1 Stimuli

A dialogue was written based on first-hand experience, by a native speaker. Then we employed the Google Text-To-Speech platform[1] to synthesize the voice of our virtual agent. As we wanted a state-of-the-art synthesis quality we chose one of the male WaveNet voices (de-DE-Wavenet-B) and each sentence was synthesized with different parameters in order to sound as spontaneous as possible, without introducing any artifacts. The synthesized files were then checked for naturalness by two native speakers and any flagged inconsistencies were corrected. The resulting audio files constituted the materials for the non-laughter condition.

In order to obtain the laughter-enhanced virtual agent, we employed natural laughter productions. For this, we searched the DUEL corpus [16] for instances of social laughter produced by male speakers. They included laughs composed of between one and three laughter syllables, either voiced or unvoiced (snort-like), with low to medium intensities. Laughter having these characteristics was found to give the best perceived naturalness in previous experiments and has been shown to work best when integrated with synthetic speech [17]. We then identified appropriate places where to add laughter in the virtual agent's speech, choosing suitable laughter tokens for each position (e.g. similar in pitch, timbre, expressed function, etc). This resulted in eight sentences having a laugh spliced within them. Thus, the laughter condition contained

[1] https://cloud.google.com/text-to-speech/

the same sentences as the non-laughter case, the only difference being the eight spliced laughter events. Two of these events contained three syllables, four laughs had two syllables and the remaining two were one-syllable long. All, but two of the two-syllable long laughs, were voiced.

After the dialogue lines of the agent have been finalized, we recorded the lines corresponding to the client in the recording studio of Bielefeld University. A female native speaker of German was recorded for the client lines. She familiarized herself with the lines belonging to the client beforehand and she had a transcript of them with her in the recording booth. During the recording, she would reply with her line, after the experimenter would play from outside of the booth the line corresponding to the virtual agent. We used the laughter-enhanced agent lines during the recordings. The speaker was allowed to adapt the script lines to achieve an optimal fit with her spontaneous behaviour, as long as the overall message remained the same. Several takes have been made and the one that sounded the most spontaneous was used in the experiment. Once both recordings (virtual agent/client) were obtained, a video clip was created by combining the recordings with matching images of an apartment, collected from the internet. The images changed, in accordance to where in the apartment the interaction was taking place. Everything, but the existence of laughter in the virtual agent's speech, were identical in the two videos. The video corresponding to the non-laughter condition was 4 minutes 33 seconds long, while the one for the laughter condition was 4 minutes 37 seconds long.

2.2 Setup

The PsyToolkit platform [18, 19] was used for our experiment. The experiment link landed the visitors on an introductory page in which a short description of the task (listening to a dialogue between a virtual real-estate agent and a client and evaluating the interaction of the agent) and its technical requirements (the existence of an audio output equipment on the device used to run the experiment and the supported browsers) were explained and the ethics and consent information listed. After consent was given, the participants were asked to play the video consisting of a slide show of the apartment and the recording between the virtual agent and the client. They were randomly assigned to one of the two conditions.

Following the video, we asked the participants to evaluate the interaction of the virtual agent with its client, from the point of view of five characteristics, as well as to rate the likeliness of recommending the apartment and the agent to their acquaintances. We used a 10-point scale (1-10) and the evaluated five characteristics were the following:

- *professionalism*, ranging between unprofessional (1) and professional (10)
- *communication* skills, ranging between bad (1) and very good (10)
- *pleasantness*, ranging between unpleasant (1) and pleasant (10)
- *formality*, ranging between casual (1) and formal (10)
- *spontaneity*, ranging between scripted (1) and spontaneous (10)

The *ratings of the apartment and of the agent* used the same 10-point scale, ranging from very unlikely (1) to very likely (10). We also asked the participants information about their age, gender (male/female/other), their native language(s), the level of German competence (equivalent level according to the Common European Framework of Reference for Languages or native speaker) and their self-reported level of exposure to voice technology systems – on a 10-point scale, ranging from very low (1) to very high (10). We considered the apartment and agent ratings as proxies of interaction success, while the five characteristics were evaluated in order to better understand which dimensions laughter affects. For a virtual agent to be more successful it should have higher ratings (probably also for the apartment it presents) and, presumably, higher scores in at least some of the measures professionalism, communication skill, pleasantness and spontaneity, as well as a lower formality rating.

Participants were recruited among students of Bielefeld University (receiving credit hours for the time taken to complete the study), colleagues and acquaintances of the authors, as well as by means of the Prolific crowdsourcing website[2], the latter group being paid for their contributions according to regular rates used at the university for taking part in experiments. They had to be residing in Germany and be fluent in German. We excluded four participants, who finished the experiment in less than 5 minutes. Thus, our study comprised 124 participants (mean age: 26.5 years, age range: 18-45 years; gender distribution: 62 males, 61 females, 1 other). An equal number of participants (62) took part in each condition. The full data was included in all the analyses, except in those that included gender as a variable, for which only male and female participant data was considered.

3 Results

The participant responses for the seven questions evaluated in the experiment are summarized in Table 1. It contains the mean and standard deviation for each dimension in the two investigated conditions. The results obtained for the laughter condition show a lower score in all the five characteristics of the interaction, but an increase with respect to recommending the apartment or the agent. Wilcoxon rank sum tests were performed to test the significance of the differences between the two conditions. From the seven dimensions evaluated by the participants, only the difference in formality reached a significant level ($p = 0.002$).

Table 1 – Mean values (μ) and standard deviations (σ) of the scores given by the participants to the seven dimensions evaluated in the experiment, for each of the two conditions (non-laughter/laughter). The statistically significant differences between conditions are marked in bold.

Dimension	**Non-laughter**		**Laughter**	
	μ	σ	μ	σ
Professionalism	8.532	1.544	8.21	1.45
Communication	8.323	1.667	8.21	1.461
Pleasantness	7.5	2.468	7.29	1.859
Formality	**6.242**	2.373	**4.887**	2.45
Spontaneity	5.371	2.376	5.194	2.216
Apartment	8.048	1.693	8.435	1.14
Agent	6.565	2.545	6.952	2.183

In order to better understand how the considered dimensions, as well as the age and the reported speech technology exposure of the participants, interact with each other, we computed pairwise Spearman correlation between each two variables on the whole data. The obtained results are illustrated in Table 2. We observed moderate positive correlations (corresponding to ρ_s values of around 0.5 and higher) between several dimensions: professionalism and communication, communication and pleasantness, as well as between pleasantness and spontaneity. The recommendation of the agent seems to be associated to the same measures that were correlated among themselves (communication, pleasantness and spontaneity), but also with the recommendation of the apartment.

Although these results hint towards a relation between the agent rating and its communi-

[2] https://www.prolific.co/

Table 2 – Spearman ρ_S values for the pairwise correlation between each of the two variables recorded in the study: professionalism (Pro), communication (Com), pleasantness (Ple), formality (For), spontaneity (Spo), apartment (Apt) and agent (Agn) recommendation, age (Age) and speech technology exposure (Spe) of the participants. The values in bold represent significant values.

	Com	**Ple**	**For**	**Spo**	**Apt**	**Agn**	**Spe**	**Age**
Pro	**0.526**	**0.258**	**0.376**	0.139	**0.286**	**0.302**	**0.196**	-0.009
Com		**0.596**	-0.048	**0.399**	**0.334**	**0.553**	0.139	0.004
Ple			-0.089	**0.496**	**0.350**	**0.684**	0.068	0.027
For				-0.036	-0.032	-0.158	0.150	-0.038
Spo					**0.214**	**0.488**	-0.071	**0.188**
Apt						**0.576**	0.168	-0.022
Agn							0.072	-0.042
Spe								0.018

cation skills, its pleasantness and its spontaneity, we see in Table 1 that the scores for the latter group of variables decrease while the agent rating increases, in the laughter condition. Thus, we performed additional analyses, to untangle the effects of the first five dimensions on the rating of the apartment and of the agent, in the two conditions. We fitted a linear regression model with either the agent or the apartment rating as the dependent variable, and professionalism, communication, pleasantness, formality, spontaneity, participant age, gender and speech technology exposure as independent variables. We also included the laughter condition as a predictor, as well as all two-way interactions between laughter condition and the previously listed independent variables, and we employed a sum contrast for the categorical predictors (laughter condition and gender).

The model fitted for the agent rating showed a significant main effect of communication skill ($p = 3.6 \cdot e^{-4}$) and pleasantness ($p = 3.7 \cdot e^{-10}$). Significant effects were seen also for the following interactions of condition: with professionalism ($p = 0.011$), with formality ($p = 0.005$) and with speech technology exposure ($p = 0.022$). Our regression model revealed that an increase of one point in pleasantness resulted in an increase of 0.59 in the agent rating, while an equivalent increase in communication would result in a 0.48 increase in the likelihood of recommending the agent. Although laughter condition had no significant effect on agent rating (as seen also in Table 1), it affected the rating through interactions with other predictors. In the laughter case, a unit increase in the communication skill resulted in a 0.34 increase in agent rating, increasing formality by one point reduced the rating by 0.2, while a unit increase in speech technology exposure increased the rating of the agent by 0.16. Thus, the former two interactions might be responsible for the higher average agent rating observed in the laughter case, despite the fact that the main predictors (communication and pleasantness) were lower than in the non-laughter case. The significant interaction between condition and speech technology suggests that participants that are more familiar with speech technology applications actually appreciate a more human-like virtual agent.

For the apartment rating, the linear model revealed a significant effect of communication skill ($p = 0.002$), age ($p = 0.020$), gender ($p = 0.023$) and condition ($p = 0.049$). An increase of one unit in the communication skills rating resulted in an increase of 0.33 and a decrease of 0.05, for communication skill and age, respectively. Female participants gave, on average, a 0.25 higher rating to the apartment than the male participants. While the results suggest that the use of laughter results in a 0.23 increase in apartment rating, condition also interacts with other predictors. The interaction between condition and speech technology exposure was significant ($p = 0.009$). Similar to the findings obtained for agent rating, a unit increase in speech technology exposure increased the rating of the apartment by 0.15, in the laughter condition.

4 Discussion and Conclusions

We have seen that the use of laughter by a virtual agent in its interaction with a client plays a role, at least when judging from the perspective of a person overhearing the conversation. While we would have expected laughter to have a positive effect on the evaluated dimensions (and a negative one on formality), only the difference in formality was found significant. Moreover, for two of the dimensions involved, professionalism and communication, we can even observe some sort of ceiling effect, possibly indicating that the quality of the interaction and of the synthesized voice are high enough. Despite this quality, our virtual agent was still considered far from natural, based on the lower pleasantness and spontaneity scores.

The fact that the evaluated dimensions did not improve in the laughter case might be due to how laughter was integrated in the agent's speech. Although a significant effort was put into choosing laughter instances that would fit the voice of the agent, a perfect match could not be found. With state-of-the-art laughter synthesis not yet reaching sufficient naturalness [11], current approaches that aim to incorporate laughter in human-machine interaction will have to find ways of better integrating the laughter instances with the agent's voice (e.g. voice conversion). Another reason why laughter did not improve the pleasantness of the agent might be because the laughter was perceived as being disconnected from the neighbouring speech. It has been shown that spectral changes occur to speech preceding laughter [20], and these changes might be used by the listener to infer upcoming laughter. Lastly, we employed here only laughs, but the inventory used by humans in conversation is more diverse and includes other frequent vocalization, such as speech laughs or smiled speech which might have been more appropriate in some circumstances. To achieve this, an integration with systems synthesizing these phenomena (e.g. [21]) would be necessary.

The lack of a stronger effect of laughter might be due also to the actual task. First, while laughter plays numerous roles in conversation, its use in a professional setting may be less frequent. Second, this was an overhearer study and the various dimensions of the interaction were evaluated by a third-party, not by the dialogue interlocutor. We, thus, believe its integration in a more interactive setting would enhance its perceived role, and asking the interlocutor to evaluate the interaction would give us a more accurate assessment of its immediate effect.

Despite the moderate to high correlations found between communication, pleasantness and spontaneity, on the one hand, and agent and apartment ratings, on the other hand, a decrease in the score of the former did not results in a decrease in the latter ratings, in the laughter case. Actually, an increase for the two ratings was observed, probably due to laughter condition having an effect on the apartment rating and to several significant interactions between condition and other predictors, for both types of ratings. Furthermore, the participants in the laughter condition seem to be more confident in their ratings of the apartment and the agent, as seen from the lower standard deviation of their answers, compared to the non-laughter case. An interesting observation to be made here is that, for both ratings, there was an interaction between condition and speech technology exposure, with participants having a higher exposure being more likely to recommend the agent and apartment. This might point to a potential negative bias of human users (or, at least, of those people not very familiar with voice applications, in our case) towards a machine that appears too human-like [22]. Thus, the communicative benefit introduced by laughter in a conversation might have interacted with a possible bias humans may have towards the appropriateness of a machine that is able to laugh. In order to test this hypothesis, we will conduct a follow-up study in which the virtual agent will be replaced by a human interlocutor and the same two conditions will be considered. A comparison of the difference between the two conditions, in this study and in its follow-up will allow us to ascertain the existence of such a bias, in the case of laughter.

Further work will be necessary, not only for a better understanding of how laughter use by machines may be perceived by human interlocutors, but also on the technical implementation of laughter in human-machine interaction systems. While previous research (e.g. [9]) proposed work programmes in this direction, they are mainly focused on the appropriate positioning of laughter in the discourse, as well as the pragmatics of laughter use. Based on the results of our study, we believe an equal amount of attention is necessary at the acoustic level, for an appropriate choice of the laughter event and its seamless integration in the speech stream. Moreover, in order for a dialogue system to properly react to social laughter, it needs to be able to consistently recognize such laughter, which might not be such an easy task either [23].

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 799022. The authors would like to thank Annett Jorschick for helpful discussions regarding the experiment and the statistical analysis, Leonie Schade and Paul Kaminski for helping setting up the dialogue, and all the participants for their aid.

References

[1] TROUVAIN, J. and K. TRUONG: *Comparing non-verbal vocalisations in conversational speech corpora.* In *Proc. of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pp. 36–39. 2012.

[2] GLENN, P.: *Towards a social interactional approach to laughter*, pp. 7–34. Studies in Interactional Sociolinguistics. Cambridge University Press, 2003. doi:10.1017/CBO9780511519888.003.

[3] HOLT, E.: *The last laugh: Shared laughter and topic termination. Journal of Pragmatics*, 42(6), pp. 1513–1525, 2010. doi:10.1016/j.pragma.2010.01.011.

[4] BONIN, F., N. CAMPBELL, and C. VOGEL: *Laughter and topic changes: Temporal distribution and information flow.* In *Proc. of CogInfoCom*, pp. 53–58. 2012. doi:10.1109/CogInfoCom.2012.6422056.

[5] GAVIOLI, L.: *Turn-initial versus turn-final laughter: Two techniques for initiating remedy in English/Italian bookshop service encounters. Discourse Processes*, 19(3), pp. 369–384, 1995. doi:10.1080/01638539509544923.

[6] HEINZ, B.: *Backchannel responses as strategic responses in bilingual speakers' conversations. Journal of Pragmatics*, 35(7), pp. 1113–1142, 2003. doi:10.1016/S0378-2166(02)00190-X.

[7] MAZZOCCONI, C., Y. TIAN, and J. GINZBURG: *What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. IEEE Transactions on Affective Computing*, pp. 1–19, 2020.

[8] EL HADDAD, K., H. ÇAKMAK, E. GILMARTIN, S. DUPONT, and T. DUTOIT: *Towards a listening agent: a system generating audiovisual laughs and smiles to show interest.* In *Proc. of ICMI*, pp. 248–255. 2016. doi:10.1145/2993148.2993182.

[9] MARAEV, V., C. MAZZOCCONI, C. HOWES, and J. GINZBURG: *Integrating laughter into spoken dialogue systems: preliminary analysis and suggested programme.* In *Proc. of the Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pp. 9–14. 2018. doi:10.21437/AI-MHRI.2018-3.

[10] Mancini, M., B. Biancardi, F. Pecune, G. Varni, Y. Ding, C. Pelachaud, G. Volpe, and A. Camurri: *Implementing and evaluating a laughing virtual character. ACM Transactions on Internet Technology*, 17(1), 2017. doi:10.1145/2998571.

[11] Mori, H., T. Nagata, and Y. Arimoto: *Conversational and social laughter synthesis with WaveNet.* In *Proc. of INTERSPEECH*, pp. 520–523. 2019. doi:10.21437/Interspeech.2019-2131.

[12] Vettin, J. and D. Todt: *Laughter in conversation: Features of occurrence and acoustic structure. Journal of Nonverbal Behavior*, 28(2), pp. 93–115, 2004. doi:10.1023/B:JONB.0000023654.73558.72.

[13] Betz, S., B. Carlmeyer, P. Wagner, and B. Wrede: *Interactive hesitation synthesis: modelling and evaluation. Multimodal Technologies and Interaction*, 2(1), p. 9, 2018. doi:10.3390/mti2010009.

[14] Wagner, P., J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Vosse: *Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program.* In *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 105–110. 2019. doi:10.21437/SSW.2019-19.

[15] Edlund, J., J. Gustafson, M. Heldner, and A. Hjalmarsson: *Towards human-like spoken dialogue systems. Speech Communication*, 50, pp. 630–645, 2008. doi:10.1016/j.specom.2008.04.002.

[16] Hough, J., Y. Tian, L. de Ruiter, S. Betz, S. Kousidis, D. Schlangen, and J. Ginzburg: *DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter.* In *Proceedings of the 10th Language Resources and Evaluation Conference*, pp. 1784–1788. 2016.

[17] Trouvain, J. and M. Schröder: *How (not) to add laughter to synthetic speech.* In *Proc. of the Tutorial and Research Workshop on Affective Dialogue Systems*, pp. 229–232. 2004.

[18] Stoet, G.: *PsyToolkit: A software package for programming psychological experiments using Linux. Behavior Research Methods*, 42(4), pp. 1096–1104, 2010. doi:10.3758/BRM.42.4.1096.

[19] Stoet, G.: *PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. Teaching of Psychology*, 44(1), pp. 24–31, 2017. doi:10.1177/0098628316677643.

[20] Ludusan, B. and P. Wagner: *No laughing matter: An investigation into the acoustic cues marking the use of laughter.* In *Proc. of ICPhS*, pp. 2179–2182. 2019.

[21] El Haddad, K., S. Dupont, J. Urbain, and T. Dutoit: *Speech-laughs: an hmm-based approach for amused speech synthesis.* In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4939–4943. IEEE, 2015.

[22] Moore, R. K.: *Appropriate voices for artefacts: Some key insights.* In *Proc. of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, pp. 7–11. 2017.

[23] Ludusan, B. and P. Wagner: *An evaluation of manual and semi-automatic laughter annotation.* In *Proc. of INTERSPEECH*, pp. 621–625. 2020. doi:10.21437/Interspeech.2020-2521.