Prosodic and Visual Naturalness of Dialogs Presented by Conversational Virtual Agents

Lukas Aspöck[1], Jonathan Ehret[2], Stefan Baumann[3], Andrea Bönsch[2], Christine T. Röhr[3], Martine Grice[3], Torsten W. Kuhlen[2] and Janina Fels[1]

Investigating the Naturalness of Virtual Agents

- Conversional virtual agents become more present in our daily life (e.g., on *smart* devices).
- Mostly text-to-speech synthesis (TTS) is used for speech production, which often differs substantially from natural speech [1].
- The effect of TTS in comparison to natural language as well as the role of the embodiment of embodied conversional agents (ECAs) needs to be studied [2].
- In an interdisciplinary team, we are aiming at revealing the impact of adequate and inadequate prosody on the perceived naturalness and aliveness of virtual agents [3].

Pilot Study: Online Experiment

- In a 3 x 2 within-subject study, participants had to watch and listen to four different dialogs (short, made-up telephone calls of around 30 s each).
- Dialogs were presented in three speech conditions:
  1. S_{human}: Human speech with adequate prosody
  2. S_{TTS}: Synthetic speech produced by an off-the-shelf TTS system
  3. $S_{\text{human+TTS}}$: Human speech with same inadequate prosody as S_{TTS}
- Two embodiment conditions:
  1. E_{audio}: audio-only presentation
  2. E_{ECA}: audio-visual presentation of an ECA (video)
- Study implemented on SoSciSurvey platform.

Speech Stimuli: Four Dialogs

- Condition 1: Anechoic audio recordings of these dialogs were made with two trained speakers (AKG C451E/CK4 capsule at ~ 50 cm distance). Facial movement was captured using Apple's *True Depth* sensor of an *iPhone SE* and the *Live Link Face* App (face animation recordings @ 100 Hz).
- Condition 2: TTS-generated dialogs using the Google Cloud TTS engine (female voice: *de-DE-Wavenet-F*; male voice: d*e-DE-Wavenet-B*).
- Condition 3: Anechoic audio recordings of these dialogs including TTS prosody with two trained speakers (same setup as for condition 1, also including face movement tracking).

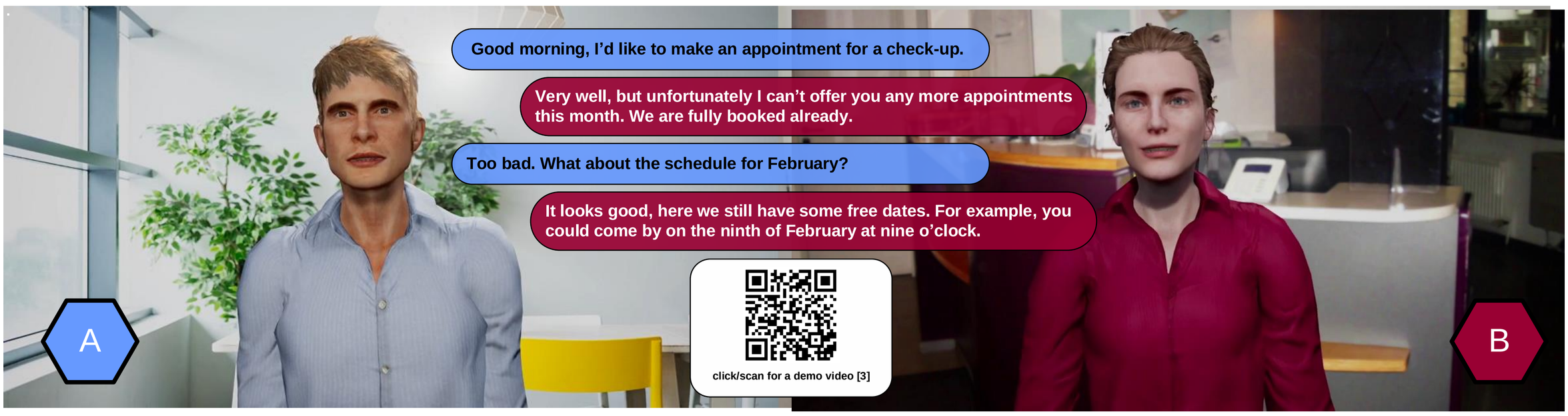


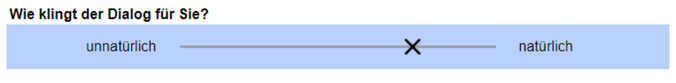
Fig. 1: Side-by-Side view and text example of the dialog situation including Speaker A and Speaker B (translated from German to English).

Virtual Acoustic Environment

- Binaural auralizations for all speech conditions were created using the *Virtual Acoustics* software (VA, http://www.virtualacoustics.org).
- The speaker was simulated as a virtual sound source at a distance of ~70cm.
- A static binaural artificial reverberation of a medium-sized room (V = 56 m³, T_{30} ≈ 430 ms) was added using the *BinauralArtificialReverb* rendering module of VA.
- VA was connected to the Unreal Engine 4.22 in which human models (created with *Character Creator 3)* of the two speakers were rendered in front of a static background (Fig. 1).
- Captured facial expressions were mapped to the faces of the models

Experimental Procedure

- Participants were asked to conduct the study in a quiet environment using headphones. A calibration procedure ensured that output levels of the audio stimuli were between 50 and 60 dB(A).
- The task was to rate the naturalness N of 24 stimuli, 12 (4 scenarios x 3 speech conditions) for E_{audio} and 12 for E_{audio}. Each stimulus could only played once.
- To rate naturalness, the question "How does the dialogue sound to you?" was answered (see below)



- Responses were mapped to a scale from 0 to 100.
- In a second part of the study, participants had to choose which of two audio stimuli sounded more natural to them.

Hypotheses

H1 We expect participants to rate (i) a human voice as more natural than a synthetic voice (even if the prosody is inadequate) and to rate (ii) adequate prosody as more natural than inadequate prosody:

$$N(S_{\text{human}}) > N\ (S_{\text{human+TTS}}) > N\ (S_{\text{TTS}})$$

H2 We expect that watching the ECAs speaking will increase the perceived naturalness of the synt. speech:

$$N(E_{\text{ECA}}) > N\ (E_{\text{audio}}) \text{ for } S_{\text{TTS}}$$

Results and Discussion

In total 39 native speakers of German were evaluated.

- Results for naturalness ratings in Fig. 2.
- **H1** confirmed: S_{Human} is rated more natural than inadequate prosody and synthetic speech.
- Differences betw. speech were significant ($p < .001$).
- **H2** not confirmed: No significant effects between embodiment conditions.
- Second part: Participants reliably chose adequate prosody as more natural and preferred the human voice, albeit not so clearly when comparing both conditions with inadequate prosody (Fig. 3).

Summary and Outlook

- An experimental procedure to investigate speech of ECAs has been implemented.
- The pilot study shows that inadequate prosody has a strong effect on perceived naturalness of speech.
- The results indicate only a minor role of the visual representation of the ECA.
- Future experiments also in VR and using English language, as English TTS is further developed.

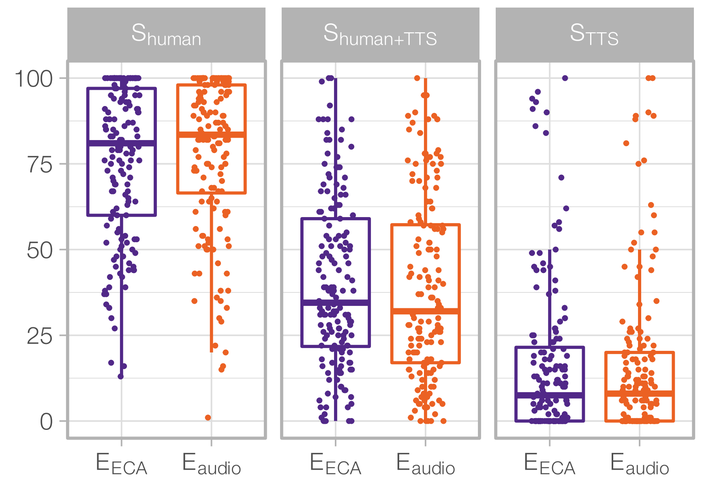


Fig. 2: Boxplots and all individual data points for naturalness ratings of three speech and two embodiment conditions.

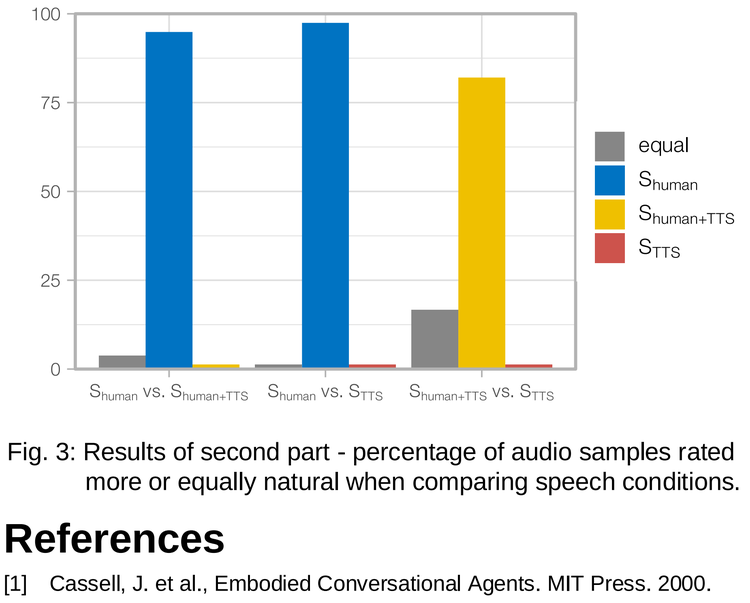


Fig. 3: Results of second part - percentage of audio samples rated more or equally natural when comparing speech conditions.

References

[1] Cassell, J. et al., Embodied Conversational Agents. MIT Press. 2000.

[2] Chérif, E. and Lemoine, J., Anthropomorphic Virtual Assistants and the Reactions of Internet Users: An Experiment on the Assistant's Voice. 2019. https://doi.org/10.1177/2051570719829432

[3] Ehret, J. et al., Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? ACM Transactions of Applied Perception (*conditionally accept*). 2021.

Institute for Hearing Technology and Acoustics
RWTH Aachen University
Kopernikusstr. 5
52074 Aachen
las@akustik.rwth-aachen.de
1

Visual Computing Institute
RWTH Aachen University
Kopernikusstr. 6
52074 Aachen
info@vr.rwth-aachen.de
2

IfL Phonetics
University of Cologne
Herbert-Lewin-Str. 6
50931 Cologne
stefan.baumann@uni-koeln.de
3