Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech?

JONATHAN EHRET and ANDREA BÖNSCH, Visual Computing Institute, RWTH Aachen University, Germany
LUKAS ASPÖCK, Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany
CHRISTINE T. RÖHR, STEFAN BAUMANN, and MARTINE GRICE, I*f*L Phonetik, University of Cologne, Germany
JANINA FELS, Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany
TORSTEN W. KUHLEN, Visual Computing Institute, RWTH Aachen University, Germany

For conversational agents' speech, either all possible sentences have to be prerecorded by voice actors or the required utterances can be synthesized. While synthesizing speech is more flexible and economic in production, it also potentially reduces the perceived naturalness of the agents among others due to mistakes at various linguistic levels. In our article, we are interested in the impact of adequate and inadequate prosody, here particularly in terms of accent placement, on the perceived naturalness and aliveness of the agents. We compare (1) inadequate prosody, as generated by off-the-shelf text-to-speech (TTS) engines with synthetic output; (2) the same inadequate prosody imitated by trained human speakers; and (3) adequate prosody produced by those speakers. The speech was presented either as audio-only or by embodied, anthropomorphic agents, to investigate the potential masking effect by a simultaneous visual representation of those virtual agents. To this end, we conducted an online study with 40 participants listening to four different dialogues each presented in the three *Speech* levels and the two *Embodiment* levels. Results confirmed that adequate prosody in human speech is perceived as more natural (and the agents are perceived as more alive) than inadequate prosody in both human (2) and synthetic speech (1). Thus, it is not sufficient to just use a human voice for an agents' speech to be perceived as natural—it is decisive whether the *prosodic realisation* is adequate or not. Furthermore, and surprisingly, we found no masking effect by speaker embodiment, since neither a human voice with inadequate prosody nor a synthetic voice was judged as more natural, when a virtual agent was visible compared to the audio-only condition. On the contrary, the human voice was even judged as less "alive" when accompanied by a virtual agent. In sum, our results emphasize, on the one hand, the importance of adequate prosody for perceived naturalness, especially in terms of accents being placed on important words in the phrase, while showing, on the other hand, that the embodiment of virtual agents plays a minor role in the naturalness ratings of voices.

CCS Concepts: • **Computing methodologies** → *Phonology/morphology*; *Intelligent agents*; • **Human-centered computing** → **User studies**; *Natural language interfaces;*

Additional Key Words and Phrases: Embodied conversational agents (ECAs), virtual acoustics, prosody, accentuation, speech, text-to-speech, audio, embodiment

Authors' addresses: J. Ehret, A. Bönsch, and T. W. Kuhlen, Visual Computing Institute, RWTH Aachen, Kopernikusstr. 6, 52074 Aachen, Germany; emails: {ehret, boensch, kuhlen}@vr.rwth-aachen.de; L. Aspöck and J. Fels, Institute for Hearing Technology and Acoustics, RWTH Aachen University, Kopernikusstr. 5, 52074 Aachen, Germany; emails: {lukas.aspoeck, janina.fels}@akustik.rwth-aachen.de; C. T. Röhr, S. Baumann, and M. Grice, IfL Phonetik, University of Cologne, Herbert-Lewin-Straße 6, 50931 Köln, Germany; emails: {christine.roehr, stefan.baumann, martine.grice}@uni-koeln.de.

1544-3558/2021/10-ART21 $15.00
https://doi.org/10.1145/3486580

ACM Reference format:
Jonathan Ehret, Andrea Bönsch, Lukas Aspöck, Christine T. Röhr, Stefan Baumann, Martine Grice, Janina Fels, and Torsten W. Kuhlen. 2021. Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech?. *ACM Trans. Appl. Percept.* 18, 4, Article 21 (October 2021), 15 pages.
https://doi.org/10.1145/3486580

1 INTRODUCTION

Embodied conversational agents (ECAs), i.e., virtual agents using natural language embedded into a virtual environment [10], are used in many domains. They can be embedded in training simulations, e.g., for negotiations [22], job interviews [2, 7], public speaking [14, 24], or teaching [13, 29]. Despite plausibly animating the ECAs for these applications (cf. [31, 40]), developers strive to make the speech, which is a key modality of ECAs, as natural as possible to further facilitate the interactions with these agents. While the most natural option for the speech content is to record a voice actor, this is very labor- and cost-intensive [20]. Therefore, **text-to-speech (TTS)** synthesis is often used (e.g., [36, 38]), which creates speech audio from text input only and can also be used in flexible real-time scenarios. While there exist approaches going even further by, e.g., incorporating more information into the synthesis process, such as **concept-to-speech (CTS)** [23], the present work will focus on the easier-to-use and more common TTS method.

There exists a large body of research comparing synthetic speech with prerecorded speech by trained speakers. For example, Chérif and Lemoine [12] found that anthropomorphic agents with a human voice elicit stronger social presence than those with a synthetic voice. Chateau et al. [11] evaluated the emotional response in participants comparing voice quality levels. Krenn et al. [26] looked into the social effects of synthetic voices incorporating dialects. Davis et al. [17] compared non-native judgments of prosodically expressive as well as neutral human utterances with TTS. Malisz et al. [30] used deep learning techniques to adapt the prominence of individual syllables of synthetic speech to the prominence in natural speech. By this they tried to improve the naturalness of the synthetic speech but did not find significant improvements in naturalness ratings. An in-depth analysis of synthetic voices in human-agent interaction by Seaborn et al. [37] provides a summary of many studies evaluating different dimensions when comparing synthetic and human speech. Other studies investigated the effect of synthetic speech on either computer-sided (e.g., [19]) or user-sided alignment/entrainment (e.g., lexical and syntactic alignment in [35]). Rosenthal-von der Pütten et al. [35] examined not only the effect of synthetic as compared to prerecorded speech but also how the ECA (in their case as robotic representation) was embodied, differing between an actual robot, a virtual robot, and no embodiment at all. They did not find an effect of synthetic speech on human-likeness, which can, however, at least partly be accounted to the robot-like visual representations used. Similar studies (e.g., [15]) also enhance this comparison by a more articulated talking head present in the physical space of the participants, namely a Furhat [1]. While there is an open discussion as to what kind of voice to use for non-human devices like smart speakers [9], we will focus here on anthropomorphic agents. To allow for intuitive interactions with those, we will put naturalness (cf. [18]) at the center of our analysis.

Despite the rapid improvement of TTS technology in recent years, human listeners tend to rate synthetic speech as less natural [27], while modern synthetic voices are reaching the level of human voices [37]. The preference of human voices may partly be attributed to an inadequate *prosody* of the synthesized speech, surfacing, for example, as the wrong placement of lexical stresses, pitch accents, and pauses, sometimes leading to a "broken" rhythm, or by inappropriate intonation contours (e.g., [16]). When comparing synthetic and human voices for virtual agents, Cabral et al. [8] copied the natural human prosody in their synthesis, and Davis et al. [17] used different levels of expressive human prosody, but they both explicitly did not evaluate inadequate prosody as commonly present in off-the-shelf TTS solutions. To close the research gap on the effect of inadequate linguistic prosody for German native listeners, we investigate how prerecorded human speech featuring the same

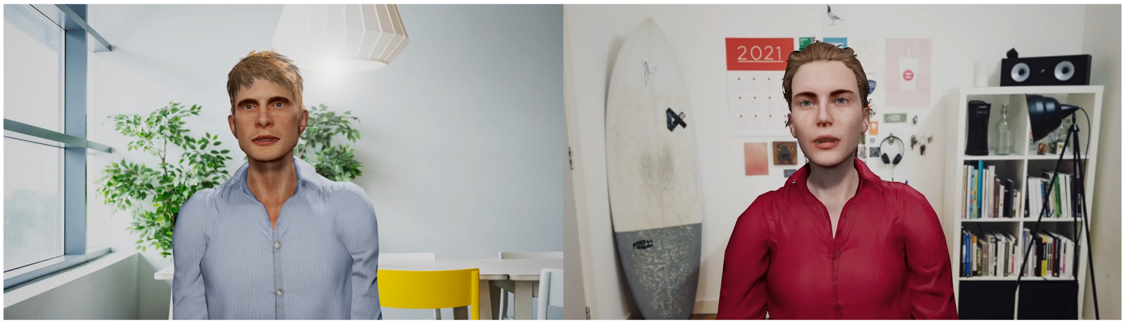


Fig. 1. Side-by-side visualization of two different frames of the used stimuli in the audio-visual condition E_{ECA}. The agents are animated using face recordings of real speakers and engage in a four-sentence conversation of about 30 s length, in this case organizing the next football training ($S4_{\mathrm{training}}$).

inadequate prosody as synthetic speech from off-the-shelf TTS solutions is rated regarding its perceived naturalness compared to TTS, on the one hand, and natural speech with "correct" or adequate prosody on the other. Thereby, we evaluate how strong the influence of such inadequate prosody is with the aim to draw attention to the role of (in-)adequate prosody when using off-the-shelf TTS in ECA research. Our testbed comprises four social contexts representing everyday situations, e.g., making a doctor's appointment, in which two ECAs engage in a four-sentence conversation of about 30 s length. Furthermore, we examine whether seeing virtual representations of the ECAs acting out this speech influences the expectations, and thus the ratings, of naturalness. We expect to find that synthetic speech will be more readily accepted within a virtual environment, since the combination may be felt as matching (cf. Gong and Nass, who found synthetic speech to be best presented with a synthetic face [21]). We call this the *masking effect of synthetic speech* by speaker embodiment. Moreover, we anticipate that the use of embodied ECAs also influences how severely inadequate prosody is assessed. We call this the *masking effect of prosody* by speaker embodiment. Furthermore, we expect the female voice to be judged as more natural in synthetic speech, since most smart speakers nowadays use female synthetic voices [41] and therefore participants are more accustomed to those producing incorrect prosody. To the best of our knowledge no study has been conducted before evaluating this isolated effect of inadequate prosody in TTS in combination with speaker embodiment. Although, as stated by Peeters [33], doing such research directly in virtual reality will increase ecological validity, we had to restrict the presented study to a video-based online survey due to the limitations resulting from the ongoing corona pandemic.

We designed a study varying the *(S)peech* in three levels: synthetic speech as generated by a TTS system (S_{TTS}), speech recorded by a voice actor imitating the less adequate prosody as present in the synthetic stimuli ($S_{\mathrm{human+TTS}}$), and human speech with adequate prosody (S_{human}). We also varied the *(E)mbodiment* of the speakers on two levels between audio-only (E_{audio}) and simultaneously watching ECAs acting out the speech (E_{ECA}) in an audio-visual condition. For our conversations, we used both female and male virtual interlocutors (*(G)ender* with the levels G_{female} and G_{male}).

We test the following hypotheses with respect to perceived naturalness (N):

H1 We expect participants to rate (1) a human voice as more natural than a synthetic voice (even if the prosody is inadequate) and (2) adequate prosody as more natural than inadequate prosody:
$N(S_{\mathrm{human}}) > N(S_{\mathrm{human+TTS}}) > N(S_{\mathrm{TTS}})$.

H2 We expect that watching the ECAs speaking will increase the perceived naturalness of the synthetic speech:
$N(E_{\mathrm{ECA}}) > N(E_{\mathrm{audio}})$ for S_{TTS}.

H3 We expect participants to perceive the female voice as more natural in synthetic speech:
$N(G_{\mathrm{female}}) > N(G_{\mathrm{male}})$ for S_{TTS}.

Table 1. Conversation in the First Scenario (**$S1_{\text{doctor}}$**) Given by a Male ECA (A) and a Female ECA (B)

S1	German (Adequate Prosody)	German (TTS Prosody)	English Translation
A	Guten **Tag**, ich **möch**te gerne einen Ter**min** für eine Kon**TROLL**untersuchung vereinbaren.	**Gu**ten **Tag**, ich möchte gerne einen Ter**min** für eine Kon**troll**untersuchung ver**EIN**baren.	Good morning, I'd like to make an appointment for a check-up.
B	Sehr **ger**ne, aber in **die**sem Monat kann ich Ihnen leider keinen Termin mehr **AN**bieten. Wir sind bereits **VOLL**.	**Sehr** **ger**ne, aber in **die**sem **Mo**nat kann ich Ihnen **lei**der keinen Ter**MIN** mehr anbieten. Wir **sind** bereits **VOLL**.	Very well, but unfortunately I can't offer you any more appointments this month. We are fully booked already.
A	**Scha**de. Wie **sieht** es denn im **FEB**ruar terminlich aus?	**Scha**de. **Wie** sieht es denn im **Feb**ruar ter**MIN**lich aus?	Too bad. What about the schedule for February?
B	**Gut**, hier **sind** noch einige Termine **FREI**. Sie **könn**ten zum Beispiel am **neun**ten Februar um **neun** **UHR** vorbeikommen.	**Gut**, **hier** sind noch **ei**nige Ter**MI**ne frei. Sie **könn**ten zum **Bei**spiel am **neun**ten **Feb**ruar um **neun** Uhr **VOR**beikommen.	It looks good, here we still have some free dates. For example, you could come by on the ninth of February at nine o'clock.

Accented syllables are written in boldface and the nuclear accent in bold capitals. The *adequate* prosody was used for S_{human}, whereas *TTS prosody* was used for $S_{\text{human+TTS}}$ as well as S_{TTS}. For the latter, inadequate nuclear accents are highlighted in red. An English translation of the text is given in the right-hand column. The other scenarios can be found in the appendix in Table 2.

2 ONLINE STUDY

We designed a 3 × 2 within-subject study, comparing the three different levels of *Speech* and the two levels of *Embodiment* of the speakers.

2.1 Materials

We designed four dialogues between a woman and a man consisting of four sentences per dialogue portraying a short telephone call in German of about 30 s each. This allowed us to place the participants as passive observers between the interlocutors. The *Scenarios* were designed to represent everyday situations like making a doctor's appointment ($S1_{\text{doctor}}$), organizing a board game night with friends ($S2_{\text{gaming}}$), booking a flight ($S3_{\text{travel}}$), or organizing the next football training ($S4_{\text{training}}$). The dialogue for scenario $S1_{\text{doctor}}$ is given in Table 1, with the accented syllables in boldface and the *nuclear* accent in bold capitals (see Table 2 in the appendix for the other scenarios). A nuclear accent is the final pitch accent in an utterance that determines the interpretation or pragmatic meaning of the utterance. Table 1 shows a distribution of accents representing a possible adequate prosody. The adequacy of prosody (especially in terms of accent placement, which is of major interest in our study) was checked in a brief informal survey prior to the experiment. Additionally, the prosody as produced by the TTS system is given, with inadequate accents in red. Since the experiment is conducted in German, we also provide an English translation, however, not specifying the accents since they are language-dependent.

We tested different commercial TTS engines and decided in favor of *Google Cloud TTS* using the voices *de-DE-Wavenet-F* as female and *de-DE-Wavenet-B* as male voice since they yielded the audibly most pleasing results while generating on average 2.5 misplaced nuclear accents per dialogue. In the last sentence of the example in Table 1, e.g., the TTS engine placed the nuclear accent on the first syllable of the final verb (*VORbeikommen*, "come by"), representing both a wrong position of lexical stress (which should be on the second syllable, i.e., *vorBEIkommen*) as well as an inappropriate position of the nuclear pitch accent (which should be on the noun *Uhr* "clock," as in the left-hand column of Table 1). These stimuli were used for the S_{TTS} level.

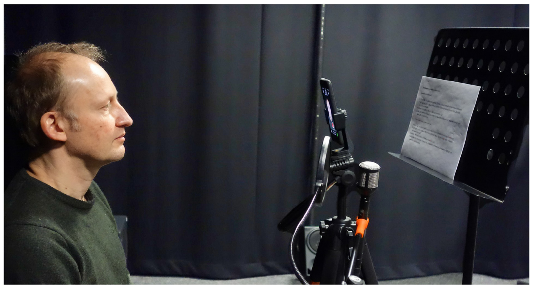


Fig. 2. A voice actor speaking his part in the dialogue being recorded with an AKG C451E microphone with pop filter and an iPhone SE for facial tracking.

Additionally, we recorded a trained 36-year-old female speaker and a trained 51-year-old male speaker with an *AKG C451E* microphone (with CK4 Capsule) at around 50 cm distance to the speaker (see Figure 2) in an acoustically optimized recording room (reverberation time $T_{30} < 200$ ms) reading out the dialogue once with adequate prosody (S_{human}) and once imitating the prosody as produced by the TTS engine ($S_{\text{human+TTS}}$). For the imitation, the actors listened to the sentences produced by the TTS engine a few times and then spoke along with it.

While recording audio, we also captured the facial movements of the speakers to animate the respective ECAs during rendering. In our tests, using a recording based on Apple's *TrueDepth* Sensor turned out to work better than capturing the face using purely RGB-video-based solutions like *OpenFace 2.0* [5] as proposed in [39] or animating the face based on speech only, for example, using *Oculus Lipsync*.[1] Therefore, we used the *Live Link Face* app[2] for iPhone, which records face animations in 100 Hz and writes them into a file, so the activation of the different facial blend shapes can be used later on for rendering (see Figure 2). Since the sentences for $S_{\text{human+TTS}}$ were spoken in sync with the audio of S_{TTS}, we were able to use the face tracking for both conditions. By this process we minimized any qualitative visual differences between the speech conditions.

The audio for both *Embodiment* levels was processed with the *Virtual Acoustics*[3] framework to generate a binaural signal of the virtual sound source approximately 70 cm away from the listener. A static artificial reverberation was added approximating the reverberation in a medium-sized room ($V = 56\,\text{m}^3$, $T_{30} \approx 430$ ms).

For E_{ECA} we used two human models generated with Reallusion's *Character Creator 3* (see Figure 1). The models were rendered in *Unreal Engine 4.22* in front of a static background and lit according to lights estimated from the background. For the conversations we tried to convey the impression of a hands-free phone call, using cuts between the frontal perspectives as depicted side by side in Figure 1. We decided to use this presentation since we assumed that this kind of cut sequence should be known from movies and allowed participants to listen to the agents from a frontal direction.

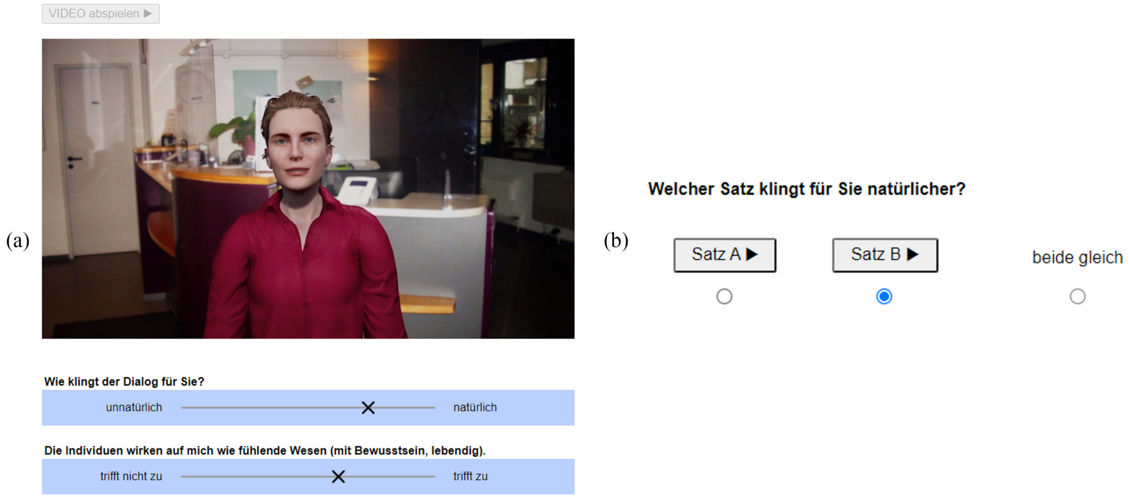
2.2 Procedure

The study was conducted as an online questionnaire realized using the *SoSci Survey* platform [28] and made available to participants at www.soscisurvey.de. The study consisted of two parts with two different tasks. In the first part, participants had to rate the naturalness of 24 stimuli (3 *Speech* conditions × 2 *Embodiment* conditions × 4 *Scenarios*). The evaluation was carried out for each stimulus on a separate page. According to the *Embodiment* condition, 12 stimuli were presented as audio-only and 12 stimuli as video. Participants were able to control when to start a stimulus, but it could only be played once. Each stimulus was rated on two **visual analog scales**

[1] https://developer.oculus.com/downloads/package/oculus-lipsync-unreal/.

[2] https://apps.apple.com/us/app/live-link-face/id1495370836/.

[3] http://www.virtualacoustics.org/.



Wie klingt der Dialog für Sie?

unnatürlich ✕ natürlich

Die Individuen wirken auf mich wie fühlende Wesen (mit Bewusstsein, lebendig).

trifft nicht zu ✕ trifft zu

Welcher Satz klingt für Sie natürlicher?

Fig. 3. Screenshots of the study forms: (a) first part with the video stimulus currently playing and both visual analogue scales filled in; (b) form of the second part, with two buttons playing one stimulus each and radio buttons to select either stimulus or *both equally* (German: *beide gleich*).

(VASs) to evaluate two aspects of naturalness (see Figure 3(a)). The first scale was used to directly collect (*N*)aturalness ratings, answering the question "How does the dialogue sound to you?" (German: *Wie klingt der Dialog für Sie?*). Participants provided the judgments by placing a roll bar on the continuous horizontal scale (VAS) with the left pole labeled *"unnatural"* and the right pole labelled *"natural."* The second scale was used to collect (*A*)liveness ratings. Participants had to judge to what extent the statement "The individuals appear to be sentient (conscious and alive) to me" (German: *Die Individuen wirken auf mich wie fühlende Wesen (mit Bewusstsein, lebendig) "does not apply"* (left pole) or *"does apply"* (right pole) to them. This question is one of five items of the Social Presence Survey [4], connecting the naturalness here to this well-established measure. The responses on both scales were encoded as interval data ranging from 0 (left pole) to 100 (right pole). Hence, the higher the ratings or values, the higher the degree of perceived naturalness/aliveness. The stimuli were presented in randomized order for each participant.

After finishing this part of the questionnaire, an intermediate questionnaire asked the following questions in random order:

(1) "What aspects did you in particular focus on in the videos?":
multiple choice for *speech*, *individuals*, *lipsync*, *gaze*, *environment*, *other*

(2) "Would you want to interact directly with one or both individuals?":
VAS from *"No, not at all"* (0) to *"Yes, absolutely"* (100)

(3) "Which of the two individuals would you prefer to interact with?":
single choice for *male*, *female*, *both equally*

(4) "Would you prefer to see the individuals talking instead of just hearing them?":
VAS from *"No, not at all"* (0) to *"Yes, absolutely"* (100)

(5) "Which version of the dialogue was easier to follow?":
single choice for *video*, *audio-only*, *both equally*

In the second part of the study, participants had to make forced choices; i.e., they had to choose which of two audio stimuli sounded more natural to them (German: *Welcher Satz klingt für Sie natürlicher?*). Therefore, participants were able to listen to each stimulus as often as necessary by clicking on it. After having listened to both stimuli at least once, participants had to pick either one stimulus or choose *"both equally"* (see Figure 3(b)). The stimuli used were individual sentences from the first part of the study. Participants had to rate six pairs, in which the same sentence was spoken with a different *Speech* level (i.e., S_{human} vs. $S_{\text{human+TTS}}$, S_{human} vs. S_{TTS}, and $S_{\text{human+TTS}}$ vs. S_{TTS}) by both speakers (or *Genders*). Additionally, three pairs with identical *Speech* level were used to compare the naturalness of the speakers' *Gender* (G_{female} vs. G_{male}). Since we had not recorded the same sentence spoken by both voice actors, we used sentences with similar length. Finally, we added four filler pairs comparing identical stimuli (two filler stimuli for each speaker) to identify insufficiently attentive participants. Participants failing to rate fillers more than once with *"both equally"* were excluded from the analysis. Hence, in total 13 sentence pairs had to be rated: 6 comparisons with different *Speech* level (but same *Gender*) + 3 comparisons with identical *Speech* level (but mixed *Gender*) + 4 filler sentences. Ratings for the nine (non-filler) sentence pairs were part of the analysis.

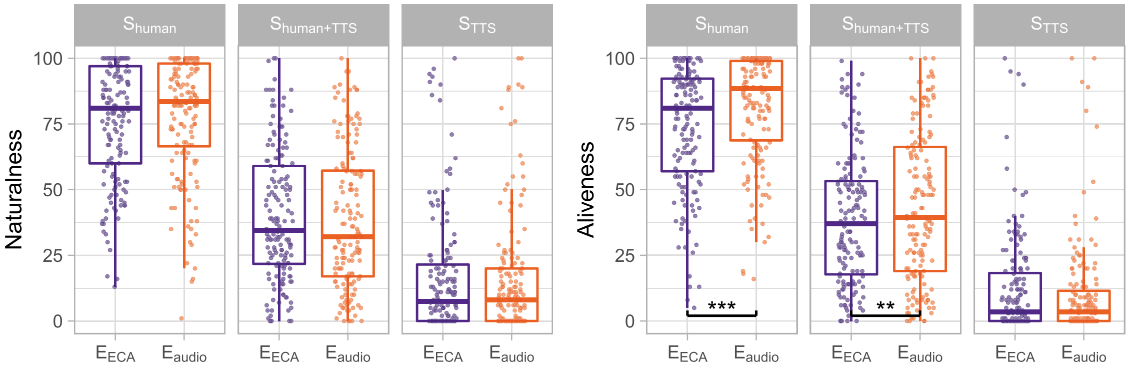
The procedure of the study was as follows: after reading a description about the content and purpose of the study, participants were asked to use regular stereo headphones and conduct an audio calibration using a sequence of TTS samples of numbers and letters. In this sequence, the participants had to adjust the audio volume so that only the numbers were comprehensible, without understanding the less loud letters in between. This created comparable hearing conditions for all participants, independent of hardware and potential background noise. In very quiet environments, the calibration led to a minimum playback volume of around 50 dBA.

Next, example exercises for both parts were shown, so participants were aware of the procedure before being asked to give informed consent and filling in a demographics questionnaire. The remainder of the study was split in two parts. Both parts began with three warm-up conditions (taken from the study conditions), so participants got familiar with the controls of the exercise and were also introduced to the entire range of the stimuli. The study ended with two free-answer fields asking for suggestions to improve the naturalness of the dialogues and asking for general feedback.

2.3 Participants and Analysis

Forty native speakers of German took part in the experiment, which were primarily recruited via university mailing lists. One participant rated more than one of the filler sentences not with *"both equally"* and was therefore excluded from the analyses. Eight of the remaining participants answered one of the fillers incorrectly; however, those participants were kept for the evaluation. The remaining 39 participants (25 female) had a mean age of 30.3 years (standard deviation (SD) = 13.4 years), and all of them reported normal hearing and normal or corrected vision. Twelve of the participants reported to have at least a basic knowledge of linguistics (one of them reported being advanced). Furthermore, six of the participants grew up in a bilingual environment. Participants took between 24 to 38 minutes to complete the entire study.

For the statistical analysis, we performed linear mixed-effects models and generalized linear mixed-effects models by using the *lmer()* and *glmer()* functions from the "lme4" package [6] for R [34]. Linear mixed-effects models were calculated to test for statistical significance of the naturalness (N) and aliveness (A) ratings on the visual analog scales in the first part of the experiment. The models included *Speech* (S_{human}, $S_{\text{human+TTS}}$, S_{TTS}), *Embodiment* (E_{audio}, E_{ECA}), and *Scenario* ($S1_{\text{doctor}}$, $S2_{\text{gaming}}$, $S3_{\text{travel}}$, $S4_{\text{training}}$) as fixed factors and assume random intercepts and slopes for *Speech* by participants. Generalized linear mixed-effects models were performed for the statistical analysis of the distributions of naturalness choices between (1) different *Speech* levels and (2) different speakers (*Gender*: G_{female}, G_{male}) in the second part of the experiment. The models included the type of *Comparison* (i.e., either between (1) different or (2) identical *Speech* levels) as fixed factor and also assume random intercepts and slopes for *Comparison* by participants. We additionally tested all models against a model with the same random effect structure including *Participant Gender* (P_{female}, P_{male}) as another fixed effect. Correlations were computed using Pearson correlation coefficients.



(a)

(b)

Fig. 4. Boxplots of the ratings of (N)aturalness (a) and (A)liveness (b) (on a scale from 0 to 100), split by *Speech* and *Embodiment*. Boxes indicate quartiles with whiskers at full range, excluding outliers. Additionally all individual data points are shown. Differences between all *Speech* levels were significant ($p < .001$); other significances are shown, $***\ p < .001$, $**\ p < .01$.

2.4 Results

Overall results of the first part of the experiment are depicted in Figure 4 in terms of boxplots and individual data points of the naturalness and aliveness ratings split by *Speech* and *Embodiment*. The figure shows that dialogues with adequate prosody spoken by a human voice (S_{human}) are clearly perceived as natural and alive, while the perceived naturalness and aliveness strongly decrease for dialogues with inadequate prosody ($S_{\text{human+TTS}}$, S_{TTS}). However, in the latter conditions human voices are still perceived as more natural and alive (medium scores) than synthetic voices, which received the lowest scores.

In the following, we will first report the effects registered by the statistical analyses of the naturalness (N) and aliveness (A) ratings that are significant by the $|t| > 2$ criterion (corresponding to the established significance level of $p < .05$, cf. [3]). Subsequently, we will report for both rating scales significant contrasts based on pairwise comparisons that exhibit a significance level of at least $p < .001$ unless otherwise specified. Statistical analyses of the naturalness (N) ratings (936 observations) register significant effects of *Speech* [$\chi^2 = 121.15, p < .001$] and *Scenario* [$\chi^2 = 10.58, p > .001$] as well as of the interactions *Speech:Scenario* [$\chi^2 = 4.47, p < .001$] and *Speech:Embodiment:Scenario* [$\chi^2 = 2.61, p < .05$]. Likewise, statistical analyses of the aliveness (A) ratings (936 observations) register significant effects of *Speech* [$\chi^2 = 169.49, p < .001$] and *Scenario* [$\chi^2 = 6.85, p < .001$] as well as of the interaction *Speech:Scenario* [$\chi^2 = 3.2628, p < .01$]. Furthermore, aliveness ratings additionally reveal significant effects of *Embodiment* [$\chi^2 = 13.30, p < .001$] and of the interaction *Speech:Embodiment* [$\chi^2 = 8.54, p < .001$]. Likelihood ratio tests comparing the presented models with a model including *Participant Gender* as another fixed factor revealed no significant effects (N: $\chi^2 = 15.55, p = .9$; A: $\chi^2 = 9.54, p = .99$).

Pairwise comparisons of the effect of the *Speech* levels confirm a significant decrease in the perception of naturalness and aliveness from S_{human} to $S_{\text{human+TTS}}$ to S_{TTS}. Accordingly, we found N and A to be strongly correlated, $r(934) = .85, p < .001$. Further pairwise comparisons reveal that dialogues of scenario $S4_{\text{training}}$ are in general rated significantly more natural and alive than dialogues of scenario $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ (cf. Figure 5). For dialogues with the $S_{\text{human+TTS}}$ *Speech* level the N and A ratings of scenario $S3_{\text{travel}}$ are also significantly higher than the ratings for scenarios $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$. Moreover, for the naturalness ratings only, these differences between scenarios $N(S1_{\text{doctor}})$ and $N(S2_{\text{gaming}})$ vs. $N(S3_{\text{travel}})$ and $N(S4_{\text{training}})$ are enhanced in the $N(E_{\text{audio}})$ condition. Further effects of *Embodiment* are registered for the aliveness ratings: pairwise comparisons reveal that dialogues presented as audio-only ($A(E_{\text{audio}})$) are in general rated as more alive than dialogues presented as

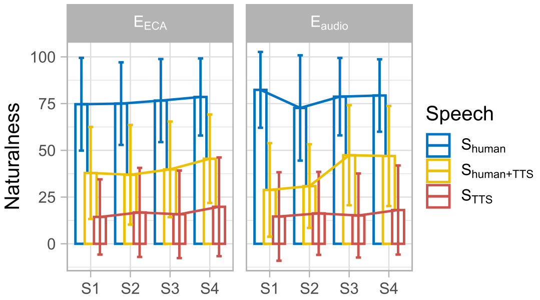


Fig. 5. Mean and standard deviation of the ratings for Naturalness (on a scale from 0 to 100) for each *Scenario* ($S1_{\text{doctor}}$, $S2_{\text{gaming}}$, $S3_{\text{travel}}$, $S4_{\text{training}}$), split by *Embodiment* and shown per *Speech* level.

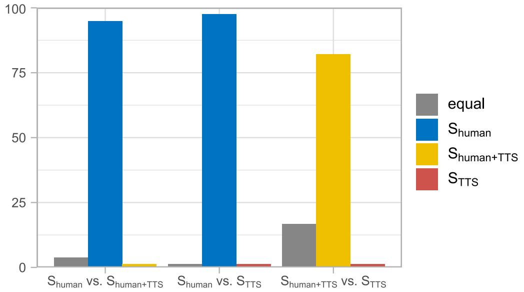


Fig. 6. Percentage of audio sample rated as more or equally natural in the second part of the study when comparing different *Speech* levels.

video ($A(E_{\text{ECA}})$). This effect is enhanced in the conditions with human voices ($A(S_{\text{human}})$: $p < .001$; $A(S_{\text{human+TTS}})$, $p < .01$).

Results of the second part of the experiment are depicted in Figures 6 and 7. Figure 6 shows the percentages of audio samples rated as more natural when comparing different *Speech* levels with each other. Listeners reliably chose utterances spoken by a human voice (S_{human}, $S_{\text{human+TTS}}$) as the more natural variant in all conditions. More precisely, a human utterance with adequate prosody (S_{human}) was preferentially selected (over 96%) whenever available. A human utterance with inadequate prosody ($S_{\text{human+TTS}}$) was only rated as more natural in comparison with a synthetic utterance (S_{TTS}). However, in the latter comparisons 17.1% of the cases were also rated as being equally natural. A likelihood ratio test comparing the generalized linear mixed-effects model including the type of *Comparison* (between different *Speech* levels: 234 observations) as fixed factor with a null model having the same random effect structure (see above) revealed that *Comparison* had a significant effect on whether the more or less natural variant was perceived as more or equally natural ($\chi^2 = 11.98, p < .01$). A further model comparison including *Participant Gender* as another fixed factor revealed no significant effect.

With respect to the comparison of the naturalness of different speakers (*Gender*: G_{female}, G_{male}), listeners' choices were overall less clear and more ambiguous when comparing the identical *Speech* levels with each other. In the S_{human} and S_{TTS} conditions listeners perceived both voices as equally natural in 59% of the cases. If listeners decided between the female and male voice, the female voice is more often rated as more natural in the S_{human} condition ($G_{\text{female}} = 33.3\%$ vs. $G_{\text{male}} = 7.7\%$), while the male voice is more often rated as more natural in the S_{TTS} condition ($G_{\text{female}} = 12.8\%$ vs. $G_{\text{male}} = 28.2\%$). In the $S_{\text{human+TTS}}$ condition, the ratings are quite balanced, although the male voice is most often rated as more natural ($equal = 35.9\%, G_{\text{female}} = 25,6\%, G_{\text{male}} = 38.5\%$). A likelihood ratio test comparing the generalized linear mixed-effects model including the type of *Comparison* (between the identical *Speech* levels: 117 observations) as fixed factor with a null model having the same random effect structure (see above) revealed that *Comparison* or rather the *Speech* level had no significant effect on whether the male or female speaker was perceived as more or equally natural ($\chi^2 = 5.29, p = .71$). However, a further model comparison including *Participant Gender* as another fixed factor revealed a significant effect ($\chi^2 = 5.33, p < .05$). Figure 7 shows the percentages of audio samples (different speakers (*Gender*: G_{female}, G_{male})) rated as more or equally natural when comparing the same *Speech* levels with each other split by the participants' gender (*Participant Gender*: P_{female}, P_{male}). The graph resembles the overall results for the different choices by the two participant groups: female listeners more often rated both voices as equally natural ($P_{\text{female}} = 58.7\%$ vs. $P_{\text{male}} = 38.1\%$). With respect to the speakers' gender, female listeners judge the male voice more often as more natural ($G_{\text{female}} = 16\%$ vs. $G_{\text{male}} = 25.3\%$), while male listeners judge the female voice more often as more natural ($G_{\text{female}} = 38.1\%$ vs. $G_{\text{male}} = 23.8\%$).

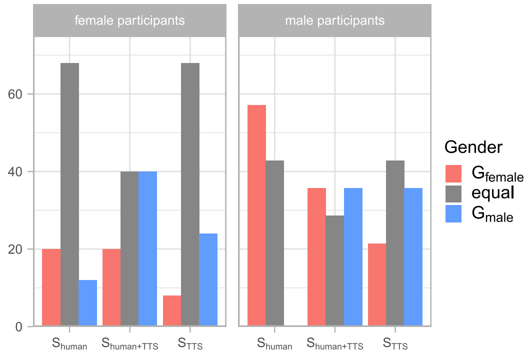


Fig. 7. Percentage of audio sample rated as more or equally natural in the second part of the study. The data is shown per *Speech* level and split between female and male participants.

The intermediate questions asked after the first part of the study revealed how many participants paid attention to the following aspects in the videos: *speech* (89.7%), *lipsync* (43.6%), *gaze* (28.2%), the *individuals* (25.6%), the *environment* (15.4%); 7.7% of the participants specifically mentioned that they focused on intonation and pronunciation, albeit not given as the default answer option. Furthermore, the data shows interesting differences in attention levels between participant gender for *lipsync* ($P_{\text{female}} = 56.0\%$, $P_{\text{male}} = 21.4\%$) and *gaze* ($P_{\text{female}} = 36.0\%$, $P_{\text{male}} = 14.3\%$), while 100% of the male participants stated that they paid attention to *speech*. In general, the participants had diverse opinions on whether they would want to interact directly with one or both individuals (VAS mean = 46.9, SD = 28.4) and showed no clear preference for an interaction with one of the individuals (*both equally* = 48.8%, $G_{\text{female}} = 25.6\%$, $G_{\text{male}} = 25.6\%$). Furthermore, the VAS answers correlated with their ratings of naturalness, $r(934) = .14, p < .001$, and aliveness, $r(934) = .17, p < .001$. Similarly, participants did not show a clear tendency on whether they would prefer to see the individuals talking instead of just hearing them (VAS mean = 45.6, SD = 32.8). Accordingly, most participants answered that they were able to follow both versions of a dialogue equally well (71.8%), showing a slight preference for the audio-only (20.5%) version (vs. 7.7% for video).

3 DISCUSSION

Our first hypothesis **H1** can be confirmed: participants in general rated (1) a human voice as more natural and alive than a synthetic voice and (2) adequate prosody as more natural and alive than inadequate prosody. This also means that inadequate prosody triggered lower naturalness and aliveness ratings for the human voice: $N(S_{\text{human}}) > N(S_{\text{human+TTS}}) > N(S_{\text{TTS}})$. It is interesting to see that $S_{\text{human+TTS}}$ (average $N(S_{\text{human+TTS}}) \approx 39.3$) has been judged to be closer to S_{TTS} (average $N(S_{\text{TTS}}) \approx 16.4$) than to S_{human} (average $N(S_{\text{human}}) \approx 77.2$). While this effect might be considered an artifact of participants not being accustomed to using VASs, it is also reflected in the direct comparisons between the *Speech* levels (cf. Figure 6). Here, the two levels with inadequate prosody ($S_{\text{human+TTS}}$ vs. S_{TTS}) were rated significantly more often as equally natural than in the comparisons with the prosodically adequate condition (S_{human}). This finding indicates a strong (unfavorable) impact of suboptimal prosody on perceived naturalness in general.

Our second hypothesis **H2** cannot be confirmed since there was no significant difference in naturalness and aliveness ratings for synthetic speech due to the level of *Embodiment*. This means that a masking effect neither for synthetic voice nor for inadequate prosody based on lowered expectations toward virtual agents was found in our study. However, and surprisingly, only *listening* to dialogues spoken by human voices (instead of also watching the ECAs speaking) significantly *increased* the perceived level of aliveness. This might be due to the fact that the perception of aliveness of the human voice is potentially masked by seeing an ECA re-enacting this speech.

This aspect was also emphasized by a participant's free comment ("If you *see* an avatar you obviously cannot rate it very human-like anymore"). Since the measure of rating the individuals as more "conscious and alive" was taken from the Social Presence Survey [4], this result is somewhat surprising since normally it is expected that social presence increases with better visualization (i.e., photographic, anthropomorphic, and behavioral realism, cf. [32]). Additionally, it needs to be conceded that the movements (gesture, lipsync, etc.) of the ECAs, although being recorded from real humans, may not be perfectly convincing. This was also the reason we decided to not use photo-realistic rendering for the ECAs to avoid mismatches in rendering and behavioral realism [25].

Furthermore, there was a significant difference of the naturalness ratings between $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ compared to $S3_{\text{travel}}$ and $S4_{\text{training}}$, especially due to the clearly diverging judgments in the $S_{\text{human+TTS}}$ condition. In a post hoc analysis we found that those scenarios that received lower ratings show more severe misplacement of accents: in $S1_{\text{doctor}}$, as mentioned above, the combination of a wrong position of a lexical stress *and* of the nuclear accent (on *VORbeikommen*, "come by") was probably felt as the most serious mistake, closely followed by the wrong nuclear accent placement on the adverb *SIcherheitshalber* ("in order to be on the safe side") instead of *ERwin* in $S2_{\text{gaming}}$ (see Table 1 and Table 2 in the appendix). A nuclear accent on an adverb is generally rare in German, at least if it is not used contrastively. The difference in naturalness ratings between the four scenarios is again more pronounced in the audio-only condition. This finding is in line with our expectation that seeing an artificial agent speaking will affect what participants expect from them with regard to speech quality.

Our third hypothesis **H3** cannot be confirmed either: participants perceived the female voice as least natural in synthetic speech. This could potentially originate from the same rationale mentioned earlier, namely that female synthetic voices are more common. One participant stated that "machine voices that you already know from navigation systems or platform announcements sound very unnatural due to previous experiences with those." Two participants even stated that they already knew this specific synthetic voice from *Siri*. It is, however, interesting to note that male participants were in general less inclined to rate both voices as equally natural, preferring the natural female and the male synthetic voice.

When asked for suggestions for improvement on the presented dialogues, 6 participants proposed to formulate them in a more natural manner, and 10 participants noted that they should be performed with a faster pace and should be pronounced less precisely (5 participants). Still, a quarter of the participants (10) rated the intermediate question whether they "want to interact with the individuals" with a mark of 66% or higher. While the variance of the answers to this question was very high (11 participants rated it lower than 33%), it is interesting to see that the answers to this question correlate with the answers given to the naturalness and aliveness questions by the same participants. This outcome suggests that a high degree of naturalness is important to facilitate interactions with conversational agents.

With this study we aim to raise awareness of the negative effect of inadequate prosody when utilizing synthetic speech for ECA research. Extending the observations by Seaborn et al. [37] that synthetic and human voices are not yet on a par, we added one previously neglected yet important dimension for this discrepancy, namely inadequate prosody, which is quite common in synthesized utterances. While our work did not focus on manually fine-tuning the off-the-shelf synthesis to produce adequate prosody for the same sentences, the difference in naturalness ratings of the human speech (S_{human} vs. $S_{\text{human+TTS}}$) and the differences between the scenarios with different severeness of accent misplacement ($S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ vs. $S3_{\text{travel}}$ and $S4_{\text{training}}$) indicate that naturalness of speech is strongly decreased by inadequate prosody. To this end, we strongly recommend practitioners to pay close attention to inadequate prosody when it comes to ECAs' speech. In case of using natural speech, despite the labor and cost intensity, trained native speakers may produce the best results. In case of using synthetic speech, practitioners should try to reduce or even omit inadequate prosody by either reformulating sentences to achieve a more suited prosody or by manually fine-tuning the synthesis results. This recommendation is especially true since a decrease in naturalness due to inadequate prosody potentially decreases the desired effect when embedding ECAs, for example, in training as mentioned in the introduction.

Nevertheless, these findings would need to be reproduced using embodied conversational agents interacting with the participants in virtual reality since this would also improve the ecological validity of the linguistic results as stated by Peeters [33]. Furthermore, while we conducted the study in German due to the higher availability of native speakers, it would be interesting to reproduce it in English since the German synthetic voices are potentially inferior to the English ones, the latter being further developed.

4 CONCLUSION

We conducted a within-subject online study to evaluate the effect of (1) inadequate prosody (as generated by off-the-shelf TTS solutions) produced either by human speakers or off-the-shelf TTS synthesizers (here: Google Cloud TTS) and (2) embodiment of the ECAs acting out the speech on the perceived naturalness of speech in a virtual environment. Our results show that inadequate prosody has a strong effect on naturalness ratings. However, the results relating to speaker embodiment (comparing audio-only with interlocutors presented as ECAs) were inconclusive, suggesting a minor role of the embodiment of ECAs in naturalness ratings of voice. From the four scenarios used, those that displayed serious mistakes in the placement of nuclear accents were judged as considerably worse than the other scenarios. This leads to the conclusion that a high level of naturalness in human as well as synthetic speech for ECAs can only be achieved if the correct placement of pitch accents is ensured.

APPENDIX

A SCENARIOS S2 TO S4

Table 2. Conversation of the Second to Fourth Scenario Given by a Male ECA (A) and a Female ECA (B)

S2	**German (Adequate Prosody)**	**German (TTS Prosody)**	**English Translation**
A	**HAL**lo, habt ihr beide morgen **Zeit** für einen ge**müt**lichen **SPIE**leabend? Ich hätte mal wieder **Lust** auf eine Runde **Sied**ler von Ca**TAN**.	**HAL**lo, **habt** ihr beide morgen **Zeit** für einen ge**müt**lichen **SPIE**leabend? Ich **hät**te mal wieder **Lust** auf eine **Run**de Siedler von Ca**TAN**.	Hello, do you both have time tomorrow for a cozy game evening? I would like to play “Settlers of Catan” again.
B	**Das** ist ja eine **tol**le I**DEE**, aber **mor**gen sind wir leider schon ver**PLANT**.	**Das** ist ja eine **tol**le I**DEE**, aber **mor**gen sind wir **lei**der schon ver**PLANT**.	That’s a great idea but we already have other plans for tomorrow.
A	**SCHA**de. Wie **wä**re es denn alterna**tiv** mit einem Abend gegen **En**de nächster **WO**che?	**SCHA**de. **Wie** wäre es denn alterna**tiv** mit einem **A**bend gegen **En**de **näch**ster **WO**che?	Too bad. What about an evening towards the end of next week as an alternative?
B	**Ja**, das klingt für **mich** erst einmal **GUT**, aber ich müsste **ER**win sicherheitshalber noch fragen. Ich melde mich **mor**gen zu**RÜCK**.	**Ja**, das **klingt** für mich **erst** einmal **GUT**, aber ich **müss**te Er**win** **SI**cherheitshalber noch fragen. Ich **mel**de mich **mor**gen zu**RÜCK**.	Yes, that sounds good to me, but I have to ask Erwin to be on the safe side. I’ll get back to you tomorrow.

S3	**German (Adequate Prosody)**	**German (TTS Prosody)**	**English Translation**
A	**HAL**lo, ich **woll**te mich nach **Flü**gen für **ei**ne Person von **Ham**burg nach **MEL**bourne erkundigen - am **liebs**ten **BUSI**ness Class.	**HAL**lo, ich **woll**te mich nach **Flü**gen für eine Per**son** von **Ham**burg nach **Mel**bourne er**KUN**digen - am **liebs**ten **BUSI**ness Class.	Hello, I want to book a flight for one person from Hamburg to Melbourne - preferably business class.
B	**GER**ne – an welches **Da**tum hätten Sie dabei ge**DACHT**?	**GER**ne – an **wel**ches **Da**tum hätten Sie dabei ge**DACHT**?	You’re welcome - which date do you have in mind?
A	Ich müsste am **neun**ten oder **zehn**ten **Mai** in Australien **AN**kommen. Wenn es **geht** bereits am **VOR**mittag, sodass ich noch im **Hel**len in mein Ho**TEL** komme.	Ich **müss**te am **neun**ten oder **zehn**ten **Mai** in Aus**TRA**lien ankommen. Wenn es **geht** bereits am **VOR**mittag, sodass ich noch im **Hel**len in mein Ho**TEL** komme.	I have to arrive in Australia on the 9th or 10th of May. I would prefer landing in the morning, to be able to arrive in the hotel during daytime.
B	Alles **KLAR**, dann werde ich Ihnen **gleich** mal ein paar **Mög**lichkeiten he**raus**suchen und **ZU**kommen lassen.	Alles **KLAR**, **dann** werde ich Ihnen **gleich** mal ein paar **Mög**lichkeiten heraussuchen und **ZU**kommen lassen.	All right, then I’ll select some options and send them to you right away.

S4	**German (Adequate Prosody)**	**German (TTS Prosody)**	**English Translation**
A	**HAL**lo, **wann** und **wo** soll ich Dich **mor**gen zum **FUSS**balltraining abholen?	**HAL**lo, **wann** und wo soll ich Dich **mor**gen zum **Fuss**balltraining **AB**holen?	Hello, when and where should I pick you up tomorrow for our football training?
B	Wie **wä**re es mit **sie**ben Uhr **drei**ssig an der **Bus**haltestelle vor der **AN**nakirche? Das **liegt** ja bei **dir** auf dem **WEG**.	**Wie** wäre es mit **sie**ben Uhr **drei**ssig an der **Bus**haltestelle vor der **AN**nakirche? **Das** liegt ja bei dir auf dem **WEG**.	How about 7:30 at the bus stop in front of the Annakirche? That’s on your way.
A	**Su**per, das **PASST**. Aber sei **dies**mal bitte **pünkt**licher als die **LETZ**ten beiden Male, sonst sind wir wieder zu **spät** und müssen **fünf** **STRAF**runden laufen.	**Su**per, das **PASST**. **A**ber sei **dies**mal bitte **pünkt**licher als die **letz**ten beiden **MA**le, **sonst** sind wir wieder zu **spät** und müssen **fünf** **STRAF**runden laufen.	Great, that works out. Please be more punctual this time compared to the last two times; otherwise, we’ll be late again and have to run five penalty laps.
B	Alles **klar**, ich werde mir **MÜ**he geben. Ich **stel**le mir gleich einen **We**cker damit ich **pünkt**lich **-LOS**gehe.	Alles **klar**, ich werde mir **MÜ**he geben. Ich **stel**le mir **gleich** einen **We**cker **da**mit ich **pünkt**lich **LOS**gehe.	All right, I’ll do my best. I’ll set an alarm right away to make sure I leave on time.

Accented syllables are written in bold face and the nuclear accent in bold capitals. The *adequate* prosody was used for S_{human} whereas *TTS prosody* was used for $S_{\text{human+TTS}}$ as well as S_{TTS}. For the latter, inadequate nuclear accents are highlighted in red. An English translation of the text is given in the right-hand column.

REFERENCES

[1] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems*. 114–130. https://doi.org/10.1007/978-3-642-34584-5_9

[2] Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. 2013. The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment*. 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

[3] R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 4 (2008), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

[4] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and Virtual Environments* 10, 6 (2001), 583–598. https://doi.org/10.1162/105474601753272844

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*. 59–66. https://doi.org/10.1109/FG.2018.00019

[6] Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[7] Tobias Baur, Ionut Damian, Patrick Gebhard, Kaéka Porayska-Pomsta, and Elisabeth André. 2013. A job interview simulation: Social cue-based interaction with a virtual character. In *International Conference on Social Computing*. 220–227. https://doi.org/10.1109/SocialCom.2013.39

[8] João Paulo Cabral, Benjamin R. Cowan, Katja Zibrek, and Rachel Mcdonnell. 2017. The influence of synthetic voice on the evaluation of a virtual character. In *Proceedings of Interspeech*. 229–233. https://doi.org/10.21437/Interspeech.2017-325

[9] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings on ACM Human-Computer Interaction* 3, 223 (2019), 19. https://doi.org/10.1145/3359325

[10] Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. 2000. *Embodied Conversational Agents*. MIT Press.

[11] Noël Chateau, Valérie Maffiolo, Nathalie Pican, and Marc Mersiol. 2005. The effect of embodied conversational agents' speech quality on users' attention and emotion. In *International Conference on Affective Computing and Intelligent Interaction*. 652–659. https://doi.org/10.1007/11573548_84

[12] Emna Chérif and Jean-François Lemoine. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice:. *Recherche et Applications en Marketing (English Edition)* 34, 1 (2019), 28–47. https://doi.org/10.1177/2051570719829432

[13] Jacquelyn J. Chini, Carrie L. Straub, and Kevin H. Thomas. 2016. Learning from avatars: Learning assistants practice physics pedagogy in a classroom simulator. *Physical Review Physics Education Research* 12, 010117 (2016), 1–15. https://doi.org/10.1103/PhysRevPhysEducRes.12.010117

[14] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1143–1154. https://doi.org/10.1145/2750858.2806060

[15] Michelle Cohn, Patrik Jonell, Taylor Kim, Jonas Beskow, and Georgia Zellou. 2020. Embodiment and gender interact in alignment to TTS voices. In *Proceedings of the Cognitive Science Society*. 220–226.

[16] Anne Cutler. 1980. Errors of stress and intonation. In *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*, V. A. Fromkin (Ed.). New York, Academic Press, 67–80.

[17] Robert O. Davis, Joseph Vincent, and Taejung Park. 2019. Reconsidering the voice principle with non-native language speakers. *Computers and Education* 140 (2019), 103605. https://doi.org/10.1016/j.compedu.2019.103605

[18] Aline W. de Borst and Beatrice de Gelder. 2015. Is it the real deal? Perception of virtual characters versus humans: An affective cognitive neuroscience perspective. *Frontiers in Psychology* 6, 576 (2015), 1–12. https://doi.org/10.3389/fpsyg.2015.00576

[19] Ramiro H. Gálvez, Agustín Gravano, Stefan Beňuš, Rivka Levitan, Marian Trnka, and Julia Hirschberg. 2020. An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars. *Speech Communication* 124 (2020), 46–67. https://doi.org/10.1016/j.specom.2020.07.007

[20] Kallirroi Georgila, Alan W. Black, Kenji Sagae, and David Traum. 2012. Practical evaluation of human and synthesized speech for virtual human dialogue systems. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. 3519–3526.

[21] Li Gong and Clifford Nass. 2007. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research* 33, 2 (2007), 163–193. https://doi.org/10.1111/j.1468-2958.2007.00295.x

[22] Jonathan Gratch, David DeVault, and Gale Lucas. 2016. The benefits of virtual humans for teaching negotiation. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*. 283–294. https://doi.org/10.1007/978-3-319-47665-0_25

[23] Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. 1997. Semantic and discourse information for text-to-speech intonation. In *Concept to Speech Generation Systems*. 47–56.

[24] Ni Kang, Willem Paul Brinkman, M. Birna Van Riemsdijk, and Mark Neerincx. 2016. The design of virtual audiences: Noticeable and recognizable behavioral styles. *Computers in Human Behavior* 55 (2016), 680–694. https://doi.org/10.1016/j.chb.2015.10.008

[25] Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. 2015. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology* 6, 390 (2015), 1–16. https://doi.org/10.3389/fpsyg.2015.00390

[26] Brigitte Krenn, Stephanie Schreitter, and Friedrich Neubarth. 2017. Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI & Society* 32, 1 (2017), 65–77. https://doi.org/10.1007/s00146-014-0569-0

[27] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics* 14 (2020), 593732. https://doi.org/10.3389/fnbot.2020.593732

[28] D. J. Leiner. 2021. SoSci Survey (Version 3.2.28) [Computer software]. https://www.soscisurvey.de.

[29] Jean-Luc Lugrin, Marc Erich Latoschik, Michael Habel, Daniel Roth, Christian Seufert, and Silke Grafe. 2016. Breaking bad behaviors: A new tool for learning classroom management using virtual reality. *Frontiers in ICT* 3, 26 (2016), 1–21. https://doi.org/10.3389/fict.2016.00026

[30] Zofia Malisz, Harald Berthelsen, Jonas Beskow, and Joakim Gustafson. 2019. PROMIS: A statistical-parametric speech synthesis system with prominence control via a prominence network. In *10th ISCA Speech Synthesis Workshop*. 257–262. https://doi.org/10.21437/SSW.2019-46

[31] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25. https://doi.org/10.1145/2485895.2485900

[32] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI* 5, 114 (2018), 1–35. https://doi.org/10.3389/frobt.2018.00114

[33] David Peeters. 2019. Virtual reality: A game-changing method for the language sciences. *Psychonomic Bulletin and Review* 26, 3 (2019), 894–900. https://doi.org/10.3758/s13423-019-01571-3

[34] R Core Team. 2015. R: A Language and Environment for Statistical Computing. http://www.r-project.org/.

[35] Astrid M. Rosenthal-von der Pütten, Carolin Straßmann, and Nicole C. Krämer. 2016. Robots or agents-neither helps you more or less during second language acquisition. In *International Conference on Intelligent Virtual Agents (IVA'16)*. 256–268. https://doi.org/10.1007/978-3-319-47665-0_23

[36] Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner. 2011. Open source voice creation toolkit for the MARY TTS platform. In *12th Annual Conference of the International Speech Communication Association*. 3253–3256. http://mary.dfki.de/.

[37] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human-agent interaction. *Computing Surveys* 54, 4 (2021), 1–43. https://doi.org/10.1145/3386867

[38] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

[39] Stef Van Der Struijk, Hung-Hsuan Huang, Maryam Sadat Mirzaei, and Toyoaki Nishida. 2018. FACSvatar: An open source modular framework for real-time FACS based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 159–164. https://doi.org/10.1145/3267851.3267918

[40] Isaac Wang and Jaime Ruiz. 2021. Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1648–1673. https://doi.org/10.1080/10447318.2021.1898851

[41] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: Closing gender divides in digital skills through education. https://unesdoc.unesco.org/ark:/48223/pf0000367416.

Received August 2021; accepted August 2021