1

Effects of Number of Voices and Voice Type on Storytelling Experience and Robot Perception

Authors

Sophia C. Steinhaeusser, University of Würzburg, sophia.steinhaeusser@uni-wuerzburg.de
Birgit Lugrin, University of Würzburg, birgit.lugrin@uni-wuerzburg.de

The Abstract

Social robots as storytellers have great potential in regards to replay every possible voice recording but also produce an endless variety of expressive synthetic voices themselves to illustrate different story characters. In an online study, we compared type of voice (human vs. synthetic) and number of voices to investigate the effects on storytelling experience and robot perception. Results show that recipients' transportation into the story was higher using a single voice compared to using different voices independent of type of voice. The same pattern was found for perceived robot intelligence. Anthropomorphism was higher for human than for synthetic voices. Further, animacy and likeability differed between type of voices in dependence of their number. No significant differences were found for warmth, competence, discomfort, or perceived safety. In general, when using human voice in robotic storytelling a single voice for the whole telling seems to be preferable. Further, when focusing on the storytelling experience in terms of transportation a single voice should be preferred, independent of voice type. Illustrating different characters by different voices is only suggested when utilizing synthetic voices.

1.1 Introduction

Our voice is our most important communication medium [48]. This is even true while communication technology is rapidly evolving. For example thinking of the communicative possibilities provided by smartphones, voice calls are still the most common way of mobile communication [16]. One reason for our preference on communication by voice are the inherent cues providing information about our interlocuter. For instance, the human voice provides information on a speaker's sex [4], attractiveness [7], personality [4, 38], credability [55], and emotions [68, 34]. Especially for emotional expression our voice is our primary communication instrument [52]. For example, anger is conveyed by an increase in loudness, fundamental frequency, and intensity [34, 65], whereas sadness is recognized when speaking more slowly with a high pitch and intensity [65].

Thus, it is rather not surprising that humans also want to communicate with machines by naturally and intuitively using their voice [48]. In turn, some technologies have to rely on verbal communication. This is not only true for voice assistants, but also for social robots. For effectively supporting humans in a task-related as well as social manner social robots need to engage on both a cognitive but also on an emotional level [9]. Therefore, a social robot and its behavior should appear comprehensive and human-like [51]. "Robot voice is one of the essential cues for the formulation of robot social attributes" [21, p. 230]. Just like humans, robots can also express emotions through their voices. Doing so fosters a richer social interaction, and supports the user's interest in the interaction [8].

One of the many fields in which robots are deployed is the robotic storytelling (see e.g., [61, 1, 58]). Especially within a story context, emotions are crucial to understand a story [45]. Thus, robot storytellers should be provided expressive voices. For example, Kory Westlund et al. [35] found that when a robotic storyteller's voice included a wide range of intonation and emotion compared to a flat voice without expressiveness, children who listened to the expressive robot showed stronger emotional engagement during the storytelling, greater inclusion of new vocabulary acquired during the storytelling into retelling of the story and greater fidelity to the original story when retelling. In addition to imitating the expressiveness of human voices, social robots as storytellers offer other speech-related capabilities that surpass those of human storytellers. While humans have limited abilities disguising their voice to portray different story characters, robots are able to not only playback every possible voice recording but also generate a sheer endless amount of synthetic voices. Comparing single synthetic voice usage to the adjustment of robotic voice to illustrate different characters, Striepe et al. [60] indicated that adapting a robot's voice to story characters in terms of pitch and speed improves recipients' narrative presence. However, synthetic voices are sometimes negatively remarked by recipients (see e.g. [57]) and human voices are preferred

over synthetic voices [24, 17]. Nevertheless, while researchers develop systems automatically producing character-matched synthetic voices for robotic storytelling [42], there is no research on using different human voices to illustrate individual characters in robotic storytelling to the best of our knowledge yet. Therefore, we aim to shed light on this knowledge gap by investigating the effect of number of voices and type of voice – human respectively synthetic – in robotic storytelling on both storytelling experience and robot perception.

1.2 Related Work

Audio books, "typically defined as a recording of a text read aloud by the author, a professional narrator, or a synthetic voice" [30, p. 124], have become one of the most successful and fastest growing formats of story reception [43]. Especially for this long-form content the speaker quality and skill are decisive for evaluating the listening experience [13, 15]. For example, narrators can manipulate their voice in terms of volume, pitch, intensity, or pauses to illustrate a story's text [15, 65]. Furthermore, giving different characters different voices enhances the listening experience and eases keeping track of a story's characters [40]. In more detail, there are four styles of narration in terms of voicing. Most commonly each character is represented by a different voice acted out by the speaker, called *fully voiced* narration. For instance, reading the books from the *Harry Potter* series by J. K. Rowling out loud, Jim Dale adopts various voices with adapted tones and pitches for all characters in the series [40]. In contrast, in *partially voiced* narration only the primary characters are represented with distinguishable voices, while the other characters share a voice. Even less voices are utilized in *unvoiced* narration in which a story is read straight without acting out distinguishable voices. Last, when multiple narrators are taking part in an audio book production, different speakers can represent different characters, which is called *multivoiced* narration [12]. Since many voice actors are needed to create a multivoiced audio book, the production process can be very expensive and time consuming. Using synthetic voices is a more affordable way and could be a more feasible approach for audio book producers [47, 53]. Yet synthetic voices are already used to produce audio books [30]. While synthesized speech is difficult to understand at first, intelligibility increases within the first five sentences of exposure largely and linearly [64].

1.2.1 Voice Types in Human-Robot Interaction

Synthetic voices as part of the voice processing technology are mainly used to transmit messages comprising information from a machine to a human [48]. Synthetic speech is also a key element of human-robot interaction [31]. First of

all, the "presence of voice is [a] strong trigger for anthropomorphic perception" [p. 204][25], the attribution of human characteristics to robots. For instance, emotive synthesized speech can improve the attribution of empathy to a robot compared to a flat synthesized voice [32]. Moreover, the choice of voice alone can already affect the interaction with and perception of a robot considerably. For instance, Dou et al. [21] reported that a humanoid robot gained highest ratings on competence when using a male voice, whereas highest scores on warmth were obtained using a child voice.

Although synthetic voices are common within HRI, researchers also work with pre-recorded human voices played back by the robot (see e.g. [61, 41]). Since humans prefer real human voices over synthetic ones [24, 59, 44] this choice can also positively affect an interaction. Inter alia, human voices are perceived as being more expressive [11], likeable [5], truthful, involved [44], credible, pleasant [36, 46], and appealing [36]. While some researchers reported no difference between human and synthetic voices regarding persuasion [59, 44], Rodero [46] reported that human voices are more persuasive in conveying advertising messages. These positive effects seem to be at least partially transferred to a robot using human voice. Comparing a synthetic and human voice used by the *Alpha* robot, Xu [67] found that when speaking with a human voice Alpha was rated more trustworthy, but was as attractive as with a synthetic voice. Choice of voice also influences how people approach a robot. Walters et al. [66] reported that participants' desired comfortable distance to a robot was significantly greater when the robot used a synthetic compared to a human or no voice. Even a mismatch in terms of human-likeness between an agent's voice and motion does not reduce the human voice's positive effect on pleasantness [24]. Thus, Ferstl et al. [24] recommend the highest possible realism of voice for virtual and robot-like agents, even when this produces incongruence between the modalities.

Especially regarding storytelling human voices seem to be preferable due to being considered as more suitable for emotional communication [46, 47]. Comparing a storytelling performed by the synthetic voice of Amazon's *Alexa* to a pre-recorded human female voice, Rodero [47] reported stronger emotional responses in the human voice condition presumably due to to a deeper level of processing. In addition, higher physiological levels of attention, arousal, and valence as well as increased self-reported enjoyment, engagement, imagery and recognition of information were observed in the human voice condition. Similar results were found for robotic storytelling. Examining participants' body language Costa et al. found exhibited facial expressions indicating emotions and non-verbal arm and head gestures indicating engagement when using a human compared to a synthetic voice during storytelling with the *Aesop* robot [17]. This might be explained by the positive relationship between perceived robot anthropomorphism and engagement as well as narrative presence found by Striepe et al. [60]. According to this finding, human-likeness, e.g. in voice, seems to be an advantage for robotic storytelling. Another but somewhat similar explanation is provided by Rodero and Lucas, who introduce the *human*

emotional intimacy effect, proclaiming that people experience closeness and connection when listening to a human voice which in turn leads to a solid and positive emotional response [47]. This theory is in line with Mayer's *voice principle*. This principle states that "people learn more deeply when the words in a multimedia message are spoken in a human voice rather than in a machine voice" [39, p. 345], therefore human voices are suggested more effective for teaching than synthesized ones. Following Rodero [47] this effect might be transferred to emotional story content.

However, the predominance of human over synthetic voices seems to vanish with the continuous technical improvement of synthetic voices. Comparing narration in a multimedia learning environment via a modern or older text-to-speech engine or a recorded human voice, Craig and Schroeder reported better perceptions of human voice considering engagement and human-likeness. Nonetheless, there were no significant differences in learning outcomes, credibility, perceptions, or cognitive efficiency to the modern text-to-speech system [19]. Utilizing similar voice conditions with a virtual pedagogical agent the modern text-to-speech engine even had a greater training efficiency and produced more learning on transfer outcomes than the human voice while being rated as equally credible [18]. "This provides consistent evidence against the voice effect." [18, p. 15] Comparing different synthetic voices newer methods achieve results in likeability closer to human voices than older engines [5]. Even regarding long-form content such as storytelling several synthetic voices outperform human voices [13].

This might explain newer findings in robotic storytelling. For instance, Goossens et al. [27] indicated no significant difference in terms of engagement and story difficulty between *NAO*'s original synthetic voice and a pre-recorded human voice. Moreover, children who listened to NAO's original voice performed significantly better concerning vocabulary acquisition. In a similar study using the *Pepper* robot Carolis et al. [20] obtained similar results. Children felt more positive emotions and reported a higher user experience in terms of pleasantness, story understanding, and image clarity when listening to the text-to-speech voice. Thus, new approaches and engines might provide high-quality synthesized speech suitable for robotic storytelling.

1.2.2 Number of Voices in Robotic Storytelling

While voice type used in HRI has been heavily researched, the number of voices used is less explored. Yet, machine learning frameworks are already created to identify synthetic voices matching story's characters. Based on the idea that common children stories mainly include similar principal characters such as a young innocent female such as Snow White or Ariel, a young heroic male such as Prince Charming or Robin Hood, and an older villain such as the Evil Queen or Ursula synthetic voices can be clustered following salient attributes such as age and gender but also evilness, intelligence, pitch, and so on [29]. Using Naive Bayes, Greene et al. [29] modeled a relationship be-

tween these attributes and synthetic voices to predict appropriate voices for children's stories' characters. However, their approach was not yet evaluated in a perceptual study. Also using an algorithmic approach Min et al. [42] were able to map synthetic voices to characters in a robotic storytelling outperforming a baseline of random selection. In addition, the authors reported that gender and character differentiation correlate positively with naturalness in the robotic storytelling.

Using several manually pre-shaped versions of NAO's synthetic voice Ruffin et al. [50] implemented a robotic storytelling of an African tale in which each character was given a distinct voice and LED color. Again, the resulting storytelling sequence was not evaluated. Similarly, Striepe et al. [60] adapted the synthetic voice of the *Reeti* robot in terms of speed and pitch to create a fully voiced robotic storytelling. A user study revealed no significant differences between one voice and character-adjusted voice concerning narrative engagement and anthropomorphism of the robot. However, narrative presence was higher when the robot modified its voice to illustrate the story's characters indicating a positive effect of character illustration via voice.

While the limited body of research on the use of different synthetic voices in robotic storytelling shows positive consequences of character illustration, to the best of our knowledge, no comparable studies have been conducted on the use of different human voices.

1.3 Contribution

In contrast to the attempts of developing systems which automatically produce synthetic voices matching characters in a robotic storytelling [42], there is no research on using different pre-recorded human voices for robotic multivoiced storytelling.

Based on previous work by Striepe et al. [60] and also findings from audio book research [40] the use of multiple character-illustrating voices should improve the storytelling experience. The listener's transportation, "the extent that [they] are absorbed into a story" [28, p. 701], as a key element of narrative engagement [10] is therefore assumed to increase when a robotic storyteller uses distinct voices for different characters compared to an unvoiced narration, regardless of the type of voice used.

H1: Transportation is higher during a multivoiced than during an unvoiced storytelling.

Regarding type of voice used in HRI mixed results are found. Since positive effects of synthetic voice usage were yet only found whithin children (see e.g., [20, 27]), we form our hypotheses based on the *human emotional intimacy effect* [47], suggesting that a human voice enhances the storytelling experience more than a synthetic voice, regardless of number of voices.

H2: Transportation is higher when a robotic storyteller uses human compared to synthetic voices.

Concerning robot perception, voice is an important "anthropomorphic ability" [22, p. 183]. Previous studies reported beneficial effects of using human voices in HRI (see e.g., [67, 24]). Especially, human voices carry more emotional cues [47] that are important for perceiving a human-like social entity [23].

H3a: Perceived anthropomorphism is higher when a robotic storyteller uses human compared to synthetic voices.

H3b: General robot perception is improved when a robotic storyteller uses human compared to synthetic voices.

H3c: Social robot perception is improved when a robotic storyteller uses human compared to synthetic voices.

Regarding the number of voices, no suggestions can be made about robot perception due to the limited scope of related work. Thus, the effects of multi-voiced compared to unvoiced robotic storytelling were examined exploratory. Also, the interaction between type and number of voices was investigated exploratory.

RQ1: Does the number of voices used influence the perception of a robotic storyteller?

RQ2: Is there an interaction between type and number of voices used in a robotic storytelling in terms of storytelling experience and robot perception?

1.4 Method

To investigate the effect of number and type of voices used in robotic storytelling on storytelling experience and robot perception a 2 (one vs. three voices) x 2 (synthetic vs. human voice) between groups design was applied in an online study. The study was approved by the local ethics committee of the Institute for Human–Computer-Media at the University of Würzburg (vote #140222).

1.4.1 Materials

To analyze the influence of robot voice type and number, four settings were implemented using the social robot *NAO V6* [56] and *Choregraphe* version 2.8.6 [2].

We chose the story "Conversation on a Bench in the Park" [1] which is conceptualized as a dialogue between two men moderated by a narrator. The men are talking on a park bench about a murder from the past. At the end

[1] https://www.kurzgeschichten-stories.de/t_2774.aspx

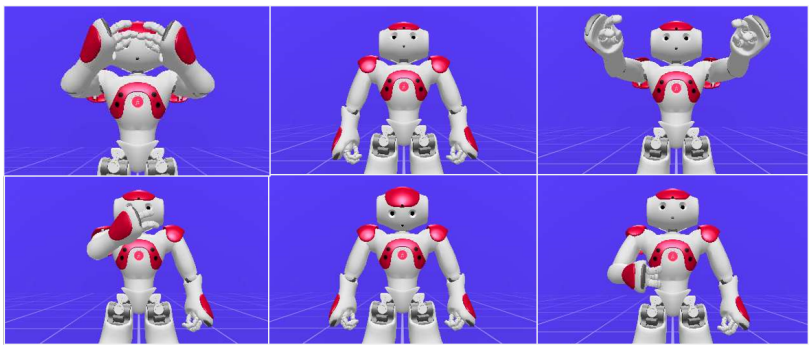

FIGURE 1.1
Emotional gestures expressing fear, disgust, joy, surprise, sadness, and anger (f.l.t.r.) implemented using Choregraphe.

of the story, the narrator reveals that possibly one of the men, called Sammy, might have been the murderer himself while the other man was the responsible police inspector. The story was annotated by four independent raters in terms of basic emotions per sentence. Doing so, a set of six recurring gestures shown in Figure 1.1 resembling the six basic emotions was added to the storytelling. In addition, NAO's line of sight was manipulated to clarify the change of speaker. During the narrator's passages, NAO looked directly into the camera, while its gaze fixed the men to its left respectively right during their passages. NAO was seated during the whole storytelling.

For the human voice conditions, three male voice actors were recruited, with the narrator being spoken by a younger man and the two men being spoken by older men. For the story version using one human voice (*oneHum* condition) the narrator's voice was recorded reading the whole story out loud in an unvoiced style, whereas all three voices were recorded in a multivoiced manner for the version using three human voices (*threeHum* condition). For the synthetic voice conditions, NAO's internal text-to-speech (TTS) module was used which offers a *voice shaping* option that modifies the voice not only in speed but also tone and thus allows to modificate NAO's voice to generate new voices of, e.g., different gender and age [2]. For the single synthetic voice version we used NAO's standard TTS voice (*oneSyn* condition). The standard TTS voice was also used for the narrator in the three synthetic voices version (*threeSyn* condition). For the two old men, the TTS voice's tone was adjusted to generate two synthetic male voices following suggestions of Traunmueller and Eriksson [62].

Combining non-verbal and verbal features described above, four versions of the story with the same non-verbal behavior but differing in voice usage were implemented and video-taped[2]. The resulting video stimuli had a length of

[2]www.soundandrobotics.com/chX

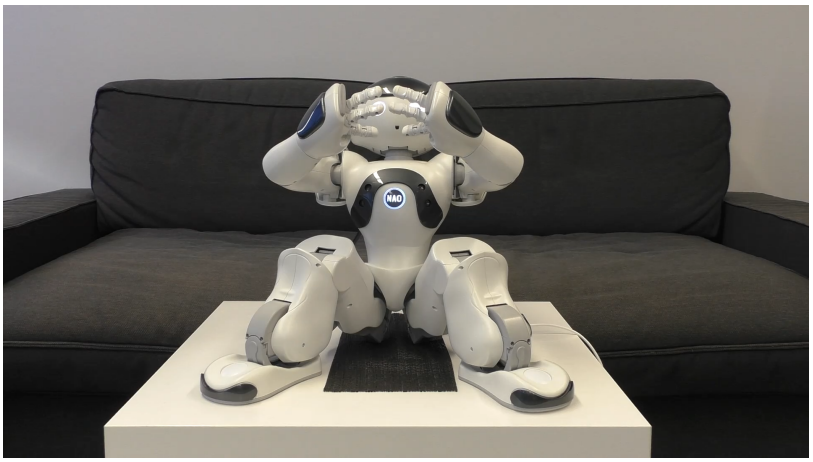

FIGURE 1.2
Freeze frame showing NAO expressing fear when narrator is speaking.

approximately 3:50 minutes. Camera angle and picture section are displayed in Figure 1.2.

1.4.2 Measures

To analyze participants' storytelling experience their *transportation* was measured using the *Transportation Scale Short Form* (TS-SF) [3]. It includes six items, e.g., "I could picture myself in the scene of the events described in the narrative" which are answered on a seven-point Likert-Scale anchored by 1 - "I totally disagree" and 7 - "I totally agree". Two items concern story characters and thus were adopted to our story, e.g. "As I listened to the story, I could vividly picture Sammy". Appel et al. [3] reported reliability of .80 to .87, whereas Cronbach's alpha of .91 was calculated for the current sample.

General robot perception was operationalized by the *Godspeed* questionnaire series [6], comprising five scales measured on five-point semantic differentials: (1) *Anthropomorphism* comprising five items, e.g., "machinelike" versus "humanlike", (2) *Animacy* including six items, e.g. "mechanical" versus "organic", (3) *Likeability* with five items, e.g., "unfriendly" versus "friendly", (4) *Perceived Intelligence* comprising five items, e.g., "foolish" versus "sensible", and (5) *Perceived Safety* including three items, e.g., "anxious" versus "relaxed". Bartneck et al. [6] reported reliability of .88 to .93 for *Anthropomorphism*, .70 for *Animacy*, .87 to .92 for *Likeability*, and .75 to .77 for *Perceived Intelligence*. Reliability was not reported for the *Perceived Safety* scale. For the current sample, Cronbach's alpha of .76 was calculated for *Anthropomorphism*, .80 for *Animacy*, .86 for *Likeability*, .83 for *Perceived Intelligence* and .76 for *Perceived Safety*.

To get deeper insights into NAO's perceived *anthropomorphism*, the *multidimensional questionnaires to assess perceived robot morphology - anthropomorphism scale* (RoMo) by Roesler et al. [49] was applied. The questionnaire includes four scales targeting (1) *Appearance*, (2) *Movement*, (3) *Communication*, and (4) *Context*. Since we did not manipulate the robot's appearance and context and due to our focus on the robot's speech we only used the *Communication* scale which comprises ten items on verbal and non-verbal expression such as "How human-like is the speech rhythm of the robot?". The items were answered using a slider anchored by 0% - "not at all" and 100% - "fully". Cronbach's alpha for the current sample was .89.

Social robot perception was operationalized using the *Robotic Social Attributes Scale* [14]. The questionnaire includes the three factors (1) *Warmth*, (2) *Competence*, and (3) *Discomfort*, each comprising six items in the form of adjectives, for example "emotional", "reliable", and "strange", which are evaluated on a 9-point Likert-scale anchored by 1 - "definitely not associated" and 9 - "definitely associated". The authors reported reliability of .92 for warmth, .95 for competence, and .90 for discomfort, while for the current sample values of .87 for warmth, .91 for competence, and .78 for discomfort were calculated.

Last, participants were asked to provide gender and age. They were also asked on their previous experiences with the NAO robot, namely whether they had seen it in pictures or videos or had already interacted with it.

1.4.3 Study Procedure

The online survey was hosted using *LimeSurvey* version 444 [37]. When accessing the website, individuals first gave informed consent to take part in the study. After being randomly assigned to one of the four conditions, they watched the respective video described in section 1.4.1. Afterwards, they filled in the questionnaires on transportation, general robot perception, perceived anthropomorphism, and social robot perception. Last, participants provided demographic data and were thanked and debriefed.

Participation in the study took about 15 minutes. We recruited our participants from the students enrolled at the University of Würzburg using the internal online-recruitment system. For their participation they received credits mandatory for obtaining their final degree.

1.4.4 Participants

Overall, 145 persons with a mean age of 21.32 (SD = 2.23) years took part in the study. While 28 participants identified as male (age: M = 22.00, SD = 2.09), the majority of 117 participants self-reported being female (age: M = 21.16, SD = 2.24). No one self-indicated as a diverse gender. Only 21 participants stated to have never seen the NAO robot before, whereas 124 already saw it in pictures or videos. In contrast, only 46 participants had already interacted with the NAO robot in person.

TABLE 1.1
Descriptive data per condition.

	oneHum		oneSyn		threeHum		threeSyn	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Transportation[a]	2.73	1.23	2.42	1.06	2.31	0.97	2.11	0.96
Anthro.[b]	1.99	0.58	1.66	0.57	1.43	0.43	1.62	0.57
Animacy[b]	2.72	0.54	2.35	0.57	2.44	0.57	2.50	0.71
Likeability[b]	3.54	0.68	3.21	0.77	3.11	0.59	3.32	0.84
P. Intelligence[b]	3.41	0.60	3.29	0.76	3.04	0.57	3.12	0.84
P. Safety[b]	3.06	0.89	3.34	0.81	3.05	0.89	3.07	0.85
Anthro. Comm.[c]	38.92	19.32	23.18	15.80	30.79	13.80	27.10	17.44
Warmth[d]	3.93	1.48	3.88	1.62	3.64	1.55	3.88	1.69
Competence[d]	5.08	1.81	5.33	1.83	4.69	1.63	4.72	1.69
Discomfort[d]	2.62	1.38	3.03	1.12	3.26	1.68	3.31	1.57

P. = Perceived, Anthro. = Anthropomorphism, Anthro. Comm. = Anthropomorphism in Communication.
a. Calculated values from 1 to 7.
b. Calculated values from 1 to 5.
c. Calculated values from 0% to 100%.
d. Calculated values from 1 to 9.

Being randomly assigned to one of the four conditions, 36 persons watched the story told by NAO using one human voice ($n_{male} = 9$, $n_{female} = 27$; age: $M = 21.39$, $SD = 2.33$), whereas 41 participants watched the story told using one synthetic voice ($n_{male} = 8$, $n_{female} = 33$; age: $M = 21.34$, $SD = 2.25$). In each case, 34 people watched the video, with three human ($n_{male} = 5$, $n_{female} = 29$; age: $M = 21.32$, $SD = 2.48$) respectively three synthetic ($n_{male} = 6$, $n_{female} = 28$; age: $M = 21.34$, $SD = 1.88$) voices.

1.5 Results

All analyses were conducted using *JASP* version 0.16.0.0 [33]. An alpha-level of .05 was applied for all statistical tests. Descriptive data is presented in Table 1.1. Calculated Levene's tests indicated homogeneity of variances for all scales, $ps > .05$.

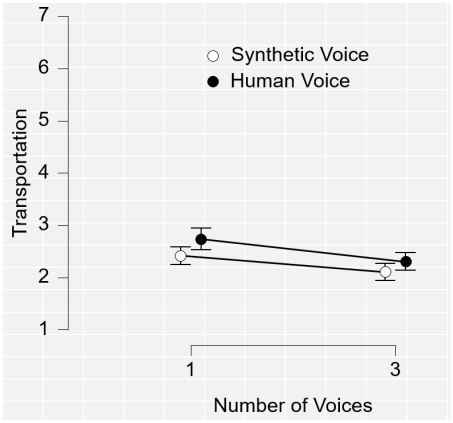

FIGURE 1.3
Descriptive plot for transportation. Error bars display standard error.

1.5.1 Transportation

A two-way ANOVA was used to analyze the effects of number of voices (H1) and type of voice (H2) on recipients' *Transportation* into the story told. Calculated results indicate a significant main effect of number of voices ($F(1, 141) = 4.33$, $p = .039$, $\omega^2 = .02$), whereas no significant main effect of type of voice was shown, $F(1, 141) = 2.39$, $p = .148$, $\omega^2 = .01$. Also, no significant interaction effect was revealed, $F(1, 141) = 0.11$, $p = .739$, $\omega^2 = .00$. Bonferroni-adjusted post-hoc comparisons again revealed no significant interaction ($ps > .05$), but a significant difference comparing number of voices when averaging over the levels of type of voice with higher values for one compared to three voices as displayed in Figure 1.3.

1.5.2 General Robot Perception

First, both anthropomorphism-related scales were analyzed using two-way ANOVAs (H3a & RQ1). Considering *Anthropomorphism* a significant main effect of number of voices ($F(1, 141) = 3.97$, $p = .001$, $\omega^2 = .06$) was indicated, while no significant main effect of type of voice was observed, $F(1, 141) = 0.68$, $p = .450$, $\omega^2 = .00$. Further, a significant interaction was found between number and type of voices, $F(1, 141) = 2.49$, $p = .004$, $\omega^2 = .05$. Bonferroni-adjusted post-hoc comparisons indicate significantly higher values of anthropomorphism for *oneHum* compared to *threeHum* ($p < .001$) as well as for *oneHum* compared to *threeSyn*, $p = .029$. In addition, the difference of higher vaues in *oneHum* compared to the *oneSyn* condition just missed significance, $p = .050$. This disordinal interaction can be seen in Figure 1.4. In contrast, for *Anthropomorphism in Communication* no significant main effect of number of voices ($F(1, 141) = 0.57$, $p = .450$, $\omega^2 = .00$) was shown, while

the main effect of type of voice was significant, $F(1, 141) = 12.18$, $p < .001$, $\omega^2 = .07$. Also, the interaction effect was significant, $F(1, 141) = 4.68$, $p = .032$, $\omega^2 = .02$. Bonferroni-corrected post-hoc comparisons reveal significantly higher values for *oneHum* compared to *oneSyn* ($p < .001$) as well as for *oneHum* compared to *threeSyn*, $p = .022$. This disordinal interaction is displayed in Figure 1.4.

Further, to analyze general robot perception again two-way ANOVAs were calculated (H3b & RQ1). Regarding *Animacy*, neither main effect of number of voices ($F(1, 141) = 0.37$, $p = .544$, $\omega^2 = .00$) nor main effect of type of voice ($F(1, 141) = 2.35$, $p = .128$, $\omega^2 = .01$) were significant. However, ANOVA calculation revealed a significant interaction between number and type of voices, $F(1, 141) = 4.70$, $p = .032$, $\omega^2 = .03$. Bonferroni-adjusted post-hoc tests solely showed significantly higher values for *oneHum* compared to *oneSyn* ($p = .047$) as displayed in Figure 1.4.

Analyzing *Likeability*, no significant main effect of number of voices ($F(1, 141) = 1.74$, $p = .189$, $\omega^2 = .01$) or type of voice ($F(1, 141) = 0.25$, $p = .616$, $\omega^2 = .00$) was indicated. In contrast, the interaction between number and type of voices was significant, $F(1, 141) = 5.29$, $p = .023$, $\omega^2 = .03$. However, Bonferroni-corrected post-hoc comparisons only indicated a trend of descriptively higher values in the *oneHum* condition compared to the *threeHum* condition, $p = .077$.

For *Perceived Intelligence* a significant main effect of number of voices was revealed ($F(1, 141) = 5.30$, $p = .023$, $\omega^2 = .03$), whereas the main effect of type of voice was insignificant, $F(1, 141) = 0.02$, $p = .881$, $\omega^2 = .00$. Also, no significant interaction effect was found, $F(1, 141) = 0.81$, $p = .369$, $\omega^2 = .00$. Bonferroni-adjusted post-hoc comparisons confirmed that when averaging results over the levels of type of voice using one voice achieves higher ratings than using three voices. , as displayed in Figure 1.4, but pairwise comparisons revealed no significant differences between the individual groups.

Last, there was no significant main effect of number ($F(1, 141) = 0.95$, $p = .331$, $\omega^2 = .00$) or type of voice ($F(1, 141) = 1.16$, $p = .284$, $\omega^2 = .00$) for *Perceived Safety*. In addition, no significant interaction effect was observed, $F(1, 141) = 0.86$, $p = .355$, $\omega^2 = .00$.

1.5.3 Social Robot Perception

In order to investigate the effects of number and type of voice on social robot perception (H3c & RQ1), again two-way ANOVAs were carried out. Regarding *Warmth*, neither a significant main effect for number ($F(1, 141) = 31$, $p = .582$, $\omega^2 = .00$) nor for type of voices ($F(1, 141) = 0.32$, $p = .723$, $\omega^2 = .00$) was found. Similarly, the interaction effect was insignificant ($F(1, 141) = 0.77$, $p = .581$, $\omega^2 = .00$) and Bonferroni-corrected post-hoc tests indicated no significant differences in pairwise comparisons. The same pattern was found for *Competence*. No significant main effect was found for number ($F(1, 141) = 2.96$, $p = .088$, $\omega^2 = .01$) or type of voices, $F(1, 141) = 0.22$, $p = .628$, $\omega^2 =$

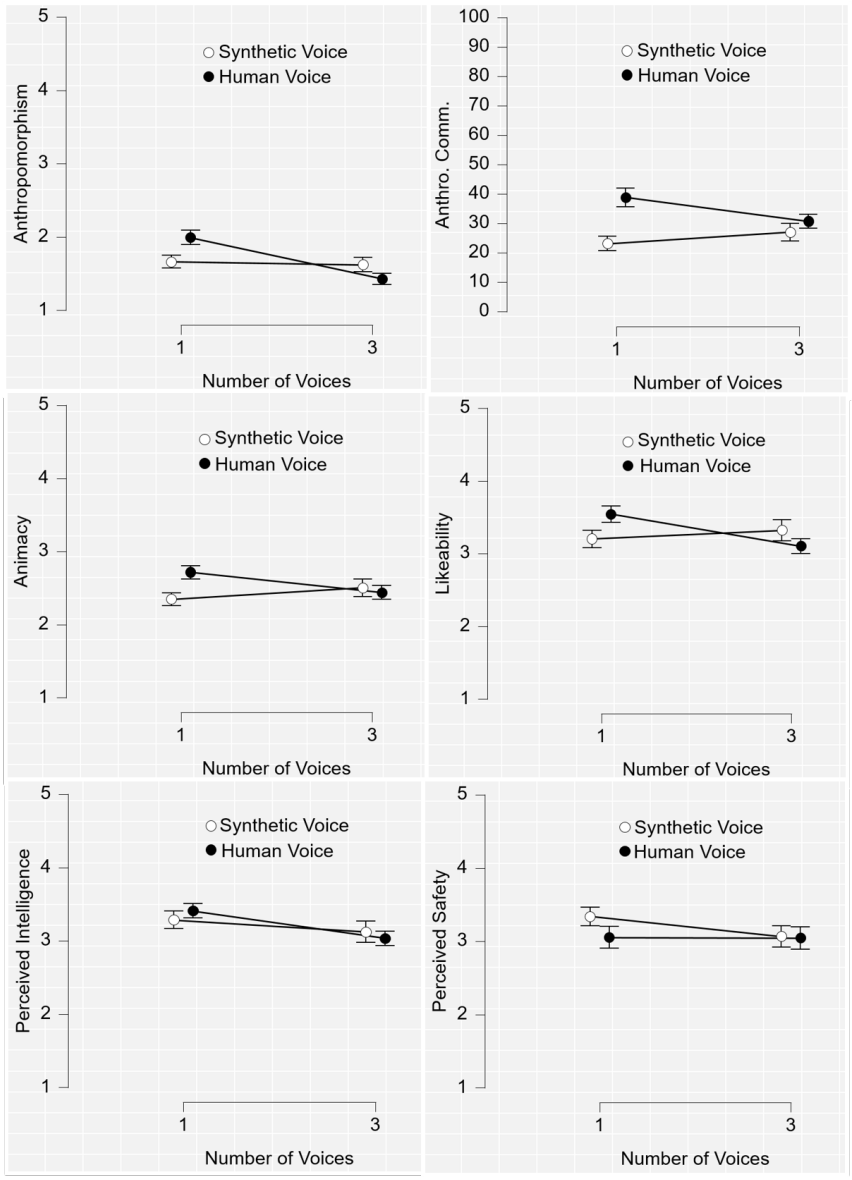

FIGURE 1.4
Descriptive plots for general robot perception. Error bars display standard error. Anthro. Comm. = Anthropomorphism in Communication.

.00. Again, no significant interaction effect was indicated, ($F(1, 141) = 0.14$, $p = .707$, $\omega^2 = .00$) and Bonferroni-adjusted pairwise comparisons revealed no significant differences between the individual conditions. Last, similar results were found for *Discomfort*. Concerning the main effect of number of voices ($F(1, 141) = 3.65$, $p = .058$, $\omega^2 = .02$) only a trend that just missed significance was revealed, whereas no significant main effect of type of voice ($F(1, 141) = 0.89$, $p = .348$, $\omega^2 = .00$) was indicated. Again, no significant interaction was found, ($F(1, 141) = 0.57$, $p = .451$, $\omega^2 = .00$), and Bonferroni-corrected post-hoc comparisons indicated no significant differences between the conditions.

1.6 Discussion

To investigate the effects of type of voice (human vs. synthetic) and number of voices (one vs. three) in a robotic storytelling scenario on both storytelling experience and robot perception an online study was carried out.

Participants' transportation into the story told by the robot was significantly higher for the unvoiced compared to the multivoiced narration for both synthetic and human voices. Thus, **H1** is rejected. This finding is not only in contrast to our presumption but also to findings from audio book research [40] and related research on robotic storytelling using fully voiced narration [60]. One possible explanation might be that humans are looking for consistency. If we listen to a multivoiced audio book we imagine multiple voice actors. Consequently, watching a single robot using different voices is inconsistent in terms of expected number of speakers. This conflict might have impeded participants' transportation into the story. In contrast, this finding could also potentially be due to a poor distinguishability of the voices used. We used only male voices in the human voice conditions and also male-sounding voices in the synthetic voice conditions. Although the narrator was way younger than the other two speaker the voices may have been too similar. Additionally, Min et al. [42] report a positive effect not only of character but also gender distinction. A wider variety of voices should be used in future studies to improve character and gender distinction and elicit naturalness as reported by Min et al. [42].

Further, no significant difference was found between human versus synthetic voice(s). Thus, **H2** is rejected, too. Also, no interaction between number and type of voice was indicated (**RQ2**). Overall, transportation was relatively low in all conditions. Reasons might be the story itself and the NAO robot. The story comprises a dialogue between two men having a conversation about a murder from the past. While the story's genre *crime* is one of the most popular book genres today [54] and should therefore meet most of the participants' interest, the story's structure is relatively uncommon and thus interfering – although the dialogue structure was suiting our research aim. The isochronous

narration includes rather individual information about the murder stringed together than a continuous plot and thus might have been hard to follow. Vaughn et al. [63] suggest a positive relationship between transportation and fluency or easiness of processing a story. In turn, the missing fluency in our story and the mental effort required to follow the storytelling and grasp all relevant information may have hindered participants' transportation into the story. Another impeding factor might have been the mechanical sounds from the robot's motors. Frid et al. [26] reported that "certain mechanical sounds produced by the NAO robot can communicate other affective states than those originally intended to be expressed" [p. 8]. This could have led to confusion among our participants, which in turn worsens fluency of the storytelling and increases mental demand. Moreover, NAO's motor sounds were generally found to be disturbing [26]. Therefore, in future studies these motor sounds should be completely masked as suggested by Frid et al. [26] or the use of other robots should be considered. Additional attention should be paid to the fact that the story used is easy to process and follows familiar structures.

Effects of choice of voice type and number on general robot perception were mixed. Especially for anthropomorphism conflicting results were found. While anthropomorphism in communication was significantly higher when using human compared to synthetic voices in both unvoiced and multivoiced conditions, no such finding was revealed for anthropomorphism measured using the *Godspeed* scale focusing on robot appearance, i.e. "moving rigidly" or "machinelike" [6]. However, for both measures the *oneHum* condition yielded the highest scores, so that **H3a** can be partially accepted. Human voices seem to be preferable for triggering anthropomorphism. Regarding number of voices (**RQ1**), following results from our pairwise comparisons unvoiced narration should be preferred when utilizing human voices, whereas when using synthetic voices both unvoiced and multivoiced narration are acceptable in terms of perceived anthropomorphism.

For animacy, likeability, perceived intelligence, and perceived safety no significant differences were found when comparing human and synthetic voices. Therefore, **H3b** is rejected. This finding is in contrast to the *human emotional intimacy effect*. Even though the synthetic voices used in our study were perceived less human-like as shown above, this lack of human-likeness did not affect general robot perception. Our findings support the claim that modern synthetic voice engines achieve results close to human speech [5, 13] and may have reached a point where they can deliver narration as credible as human voices [19].

Regarding **RQ1**, no difference between unvoiced and multivoiced narration was found for animacy, likeability, and perceived safety. In contrast, unvoiced narration scored higher on perceived intelligence compared to using three voices independently from voice type. While no interaction (**RQ2**) between number and type of voice was found for perceived intelligence and safety, both factors interact in terms of animacy and likeability. While human voice should be preferred for unvoiced narration in terms of animacy, no differences

were obtained between human and synthetic voice for multivoiced narration. For improving likeability, either one human or multiple synthetic voices should be used. However, this is only a trend in the data. Overall, the *oneHum* condition scored highest on all scales except for perceived safety. Thus, unvoiced narration using a human voice is suggested to improve general robot perception but also synthetic voices are acceptable.

Regarding social robot perception, no differences in warmth, competence, and discomfort were obtained between human and synthetic voice usage, thus **H3c** must be rejected. Similarly, no differences were indicated comparing unvoiced and multivoiced narration (**RQ1**). Also, no interaction between both factors was observed (**RQ2**). Neither type nor umber of voice seems to affect social robot perception.

1.7 Practical Implications for Choice of Voice

First of all, the choice of voice for robotic storytelling can be made independently from a robot's perceived sociality in terms of warmth, and more important discomfort and competence. None of the voice conditions tested made our participants feel uncomfortable. In terms of storytelling experience unvoiced narration is suggested, while type of voice can be freely chosen. This is also true if a robot's perceived intelligence shall be improved. The decision between human and synthetic voice becomes relevant only for scenarios in which anthropomorphism is crucial. If high levels of anthropomorphism are desired, human unvoiced narration should be preferred. Otherwise, human and synthetic voices performed almost the same, so that human voices are partially recommended for unvoiced narration, whereas when using synthetic voices suit both unvoiced and multivoiced narration.

1.8 Conclusion

An online study was carried out to shed light on the choice of voice for robotic storytelling not only in terms of type of voice, namely human or synthetic, but also in terms of number of voices used, leading to unvoiced or multivoiced narration. While audio book research reports positive effects of multivoiced narration on the storytelling experience our results suggest a preference for unvoiced narration potentially due to the robot's physical embodiment. Regarding type of voice, our findings support the assumption that modern synthetic voice engines achieve results close to human speech and may have reached a point where they can deliver narration as good as human voices. At this point,

mixed productions of synthetic and natural signals [53] might be a next step to be tested in future work.

Acknowledgments

The authors would like to thank Alisa Dianov, Angela Ast, Annika Büttner, Alisa Ebner, and Jana Luksch for preparing the stimulus material.

Bibliography

[1] Quoc Ahn Le, Christophe d'Alessandro, Olivier Deroo, David Doukhan, Rodolphe Gelin, Jean-Claude Martin, Catherine Pelachaud, Albert Rilliard, and Sophie Rosset. Towards a storytelling humanoid robot. In Association for the Advancement of Artificial, editor, *2010 AAAI Fall Symposium Series*, 2010.

[2] Aldebaran Robotics. Choregraphe, 2016.

[3] Markus Appel, Timo Gnambs, Tobias Richter, and Melanie C. Green. The transportation scale–short form (ts–sf). *Media Psychology*, 18(2):243–266, 2015.

[4] C. D. Aronovitch. The voice of personality: stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of social psychology*, 99(2):207–220, 1976.

[5] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. The perception and analysis of the likeability and human likeness of synthesized speech. In *Interspeech 2018*, pages 2863–2867, ISCA, 2018. ISCA.

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[7] Diane S. Berry. Vocal types and stereotypes: Joint effects of vocal attractiveness and vocal maturity on person perception. *Journal of Nonverbal Behavior*, 16(1):41–54, 1992.

[8] Cecilia Breazeal. Emotive qualities in robot speech. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*, pages 1388–1394. IEEE, 2001.

[9] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. In Bruno Siciliano and Oussama Khatib, editors, *Springer Handbook of Robotics*, pages 1935–1972. Springer International Publishing, Cham, 2016.

[10] Rick Busselle and Helena Bilandzic. Measuring narrative engagement. *Media Psychology*, 12(4):321–347, 2009.

[11] João Paulo Cabral, Benjamin R. Cowan, Katja Zibrek, and Rachel McDonnell. The influence of synthetic voice on the evaluation of a virtual character. In *Interspeech 2017*, pages 229–233, ISCA, 2017. ISCA.

[12] Maria Cahill and Jennifer Richey. What sound does an odyssey make? content analysis of award-winning audiobooks. *The Library Quarterly*, 85(4):371–385, 2015.

[13] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjørn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, 2020. ACM.

[14] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. The robotic social attributes scale (rosas). In Bilge Mutlu, Manfred Tscheligi, Astrid Weiss, and James E. Young, editors, *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 254–262, New York, New York, USA, 2017. ACM Press.

[15] Leilani Clark. Speaking pictures: the role of sound and orality in audio presentations of children's picture books. *New Review of Children's Literature and Librarianship*, 9(1):1–19, 2003.

[16] Coop Italia. Most common ways to communicate with a smartphone in italy in 2019, by type [graph]. https://www.statista.com/statistics/1085033/most-common-ways-to-communicate-with-smartphones-in-italy/, 2019.

[17] Sandra Costa, Alberto Brunete, Byung-Chull Bae, and Nikolaos Mavridis. Emotional storytelling using virtual and robotic agents. *International Journal of Humanoid Robotics*, 15(03):1850006, 2018.

[18] Scotty D. Craig and Noah L. Schroeder. Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, 114:193–205, 2017.

[19] Scotty D. Craig and Noah L. Schroeder. Text-to-speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research*, 57(6):1534–1548, 2019.

[20] Berardina de Carolis, Francesca D'Errico, and Veronica Rossano. Pepper as a storyteller: Exploring the effect of human vs. robot voice on children's emotional experience. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Human-Computer Interaction – INTERACT 2021*, volume 12933 of *Lecture Notes in Computer Science*, pages 471–480. Springer International Publishing, Cham, 2021.

[21] Xiao Dou, Chih-Fu Wu, Jin Niu, and Kuan-Ru Pan. Effect of voice type and head-light color in social robots for different applications. *International Journal of Social Robotics*, 2021.

[22] Brian R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3-4):177–190, 2003.

[23] Friederike Eyssel, Frank Hegel, Gernot Horstmann, and Claudia Wagner. Anthropomorphic inferences from emotional nonverbal cues: A case study. In *19th International Symposium in Robot and Human Interactive Communication*, pages 646–651. IEEE, 092010.

[24] Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. Human or robot? In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*, pages 76–83, New York, NY, USA, 2021. ACM.

[25] Julia Fink. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams, editors, *Social Robotics*, volume 7621 of *Lecture Notes in Computer Science*, pages 199–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[26] Emma Frid, Roberto Bresin, and Simon Alexanderson. Perception of mechanical sounds inherent to expressive gestures of a nao robot-implications for movement sonification of humanoids. *Sound and Music Computing*, 2018.

[27] Nicole Goossens, Rian Aarts, and Paul Vogt. Storytelling with a social robot. *Robots for Learning R4L*, 2019.

[28] Melanie C. Green and Timothy C. Brock. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, 79(5):701–721, 2000.

[29] Erica Greene, Taniya Mishra, Patrick Haffner, and Alistair Conkie. Predicting character-appropriate voices for a tts-based storyteller system. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[30] Iben Have and Birgitte Stougaard Pedersen. Sonic mediazation of the book: Affordances of the audiobook. *MedieKultur — Journal of media and communication research*, 54:123–140, 2013.

[31] S. Hennig and R. Chellali. Expressive synthetic voices: Considerations for human robot interaction. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 589–595. IEEE, 2012.

[32] Jesin James, B. T. Balamurali, Catherine I. Watson, and Bruce MacDonald. Empathetic speech synthesis and testing for healthcare robots. *International Journal of Social Robotics*, 13(8):2119–2137, 2021.

[33] JASP Team. Jasp, 2021.

[34] Tom Johnstone and Klaus R. Scherer. Vocal communication of emotion. *Handbook of emotions*, 2:220–235, 2000.

[35] Jacqueline M. Kory Westlund, Sooyeon Jeong, Hae W. Park, Samuel Ronfard, Aradhana Adhikari, Paul L. Harris, David DeSteno, and Cynthia L. Breazeal. Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, 11:295, 2017.

[36] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 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*, 14:593732, 2020.

[37] LimeSurvey GmbH. Limesurvey, 2022.

[38] Edith B. Mallory and Virginia R. Miller. A possible basis for the association of voice characteristics and personality traits. *Speech Monographs*, 25(4):255–260, 1958.

[39] Richard E. Mayer. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In Richard E. Mayer, editor, *The Cambridge handbook of multimedia learning*, Cambridge handbooks in psychology, pages 345–370. Cambridge University Press, New York, NY, 2014.

[40] Kaite Mediatore. Reading with your ears: Readers' advisory and audio books. *Reference & User Services Quarterly*, 42(4):318–323, 2003.

[41] Ali Meghdari, Azadeh Shariati, Minoo Alemi, Gholamreza R Vossoughi, Abdollah Eydi, Ehsan Ahmadi, Behrad Mozafari, Ali Amoozandeh Nobaveh, and Reza Tahami. Arash: A social robot buddy to support children with cancer in a hospital environment. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 232(6):605–618, 2018.

[42] Hye-Jin Min, Sang-Chae Kim, Joonyeob Kim, Jin-Woo Chung, and Jong C. Park. Speaker-tts voice mapping towards natural and characteristic robot storytelling. In *2013 IEEE RO-MAN*, pages 793–800. IEEE, 26.08.2013 - 29.08.2013.

[43] Jessica E. Moyer. Audiobooks and e-books: A literature review. *Reference & User Services Quarterly*, 51(4):340–354, 2012.

[44] John W. Mullennix, Steven E. Stern, Stephen J. Wilson, and Corrielynn Dyson. Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, 19(4):407–424, 2003.

[45] Seo-Hui Park, Byung-Chull Bae, and Yun-Gyung Cheong. Emotion recognition from text stories using an emotion embedding model. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 579–583. IEEE, 2020.

[46] Emma Rodero. Effectiveness, attention, and recall of human and artificial voices in an advertising story. prosody influence and functions of voices. *Computers in Human Behavior*, 77:336–346, 2017.

[47] Emma Rodero and Ignacio Lucas. Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society*, page 146144482110241, 2021.

[48] David B. Roe, editor. *Voice communication between humans and machines*. National Academy Press, Washington, DC, 1994.

[49] Eileen Roesler, Kenneth zur Kammer, and Linda Onnasch. *Multidimensionale Fragebögen zur Erfassung der wahrgenommenen Robotermorphologie (RoMo) in der Mensch-Roboter-Interaktion [Submitted manuscript]*. 2023.

[50] Margie Ruffin, Jaye Nias, Kayla Taylor, Gabrielle Singleton, and Amber Sylvain. Character development to facilitate retention in a storytelling robot. In Morris Chang, Dan Lo, and Eric Gamess, editors, *Proceedings of the 2020 ACM Southeast Conference*, pages 276–279, New York, NY, USA, 2020. ACM.

[51] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. Effects of gesture on the perception of psychological anthropomorphism: A case study with a humanoid robot. In Bilge Mutlu,

Christoph Bartneck, Jaap Ham, Vanessa Evers, and Takayuki Kanda, editors, *Social Robotics*, volume 7072 of *Lecture Notes in Computer Science*, pages 31–41. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[52] K. R. Scherer. Expression of emotion in voice and music. *Journal of voice : official journal of the Voice Foundation*, 9(3):235–248, 1995.

[53] Meysam Shamsi, Nelly Barbot, Damien Lolive, and Jonathan Chevelu. Mixing synthetic and recorded signals for audio-book generation. In Alexey Karpov and Rodmonga Potapova, editors, *Speech and Computer*, volume 12335 of *Lecture Notes in Computer Science*, pages 479–489. Springer International Publishing, Cham, 2020.

[54] Simon-Kucher & Partners. Welche genres lesen sie unabhängig vom format?, 2020.

[55] Brent K. Simonds, Kevin R. Meyer, Margaret M. Quinlan, and Stephen K. Hunt. Effects of instructor speech rate on student affective learning, recall, and perceptions of nonverbal immediacy, credibility, and clarity. *Communication Research Reports*, 23(3):187–197, 2006.

[56] SoftBank Robotics. Nao: V6, 2018.

[57] Sophia C. Steinhaeusser, Juliane J. Gabel, and Birgit Lugrin. Your new friend nao vs. robot no. 783 - effects of personal or impersonal framing in a robotic storytelling use case. In Cindy Bethel, Ana Paiva, Elizabeth Broadbent, David Feil-Seifer, and Daniel Szafir, editors, *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 334–338, New York, NY, USA, 03082021. ACM.

[58] Sophia C. Steinhaeusser, Philipp Schaper, and Birgit Lugrin. Comparing a robotic storyteller versus audio book with integration of sound effects and background music. In Cindy Bethel, Ana Paiva, Elizabeth Broadbent, David Feil-Seifer, and Daniel Szafir, editors, *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 328–333, New York, NY, USA, 03082021. ACM.

[59] S. E. Stern, J. W. Mullennix, C. Dyson, and S. J. Wilson. The persuasiveness of synthetic speech versus human speech. *Human factors*, 41(4):588–595, 1999.

[60] Hendrik Striepe, Melissa Donnermann, Martina Lein, and Birgit Lugrin. Modeling and evaluating emotion, contextual head movement and voices for a social robot storyteller. *International Journal of Social Robotics*, pages 1–17, 2019.

[61] Hendrik Striepe and Birgit Lugrin. There once was a robot storyteller: Measuring the effects of emotion and non-verbal behaviour. In Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He, editors, *Social*

Robotics, volume 10652 of *Lecture Notes in Computer Science*, pages 126–136. Springer International Publishing, Cham, 2017.

[62] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*, 11, 1995.

[63] Leigh Ann Vaughn, Sarah J. Hesse, Zhivka Petkova, and Lindsay Trudeau. "this story is right on": The impact of regulatory fit on narrative engagement and persuasion. *European Journal of Social Psychology*, 39(3):447–456, 2009.

[64] Horabail Venkatagiri. Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication*, 10(2):96–104, 1994.

[65] Rashmi Verma, Parakrant Sarkar, and K. Sreenivasa Rao. Conversion of neutral speech to storytelling style speech. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6. IEEE, 04.01.2015 - 07.01.2015.

[66] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. te Boekhorst. Human approach distances to a mechanical-looking robot with different robot voice styles. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 707–712. IEEE, 01.08.2008 - 03.08.2008.

[67] Kun Xu. First encounter with robot alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media & Society*, 21(11-12):2522–2547, 2019.

[68] Y. Yogo, M. Ando, A. Hashi, S. Tsutsui, and N. Yamada. Judgments of emotion by nurses and students given double-bind information on a patient's tone of voice and message content. *Perceptual and motor skills*, 90(3 Pt 1):855–863, 2000.