Hi robot, it's not what you say, it's how you say it

Jūra Miniota*, Siyang Wang†, Jonas Beskow‡, Joakim Gustafson§, Éva Székely¶ and André Pereira‖
KTH, Royal Institute of Technology
Stockholm, Sweden
Email: { jura*, siyangw†, beskow‡, jkgu§, szekely¶, atap‖}@kth.se

***Abstract*—Many robots use their voice to communicate with people in spoken language but the voices commonly used for robots are often optimized for transactional interactions, rather than social ones. This can limit their ability to create engaging and natural interactions. To address this issue, we designed a spontaneous text-to-speech tool and used it to author natural and spontaneous robot speech. A crowdsourcing evaluation methodology is proposed to compare this type of speech to natural speech and state-of-the-art text-to-speech technology, both in disembodied and embodied form. We created speech samples in a naturalistic setting of people playing tabletop games and conducted a user study evaluating Naturalness, Intelligibility, Social Impression, Prosody, and Perceived Intelligence. The speech samples were chosen to represent three contexts that are common in tabletopgames and the contexts were introduced to the participants that evaluated the speech samples. The study results show that the proposed evaluation methodology allowed for a robust analysis that successfully compared the different conditions. Moreover, the spontaneous voice met our target design goal of being perceived as more natural than a leading commercial text-to-speech.**

***Index Terms*—speech synthesis, human-robot interaction, embodiment, spontaneous speech, intelligibility, naturalness**

I. Introduction

The voice conveys more information than just the meaning of words. Non-verbal vocalizations and prosody are used to reflect the speaker's emotional state and other contextual cues. To enable social robots to effectively communicate, their voices must be contextually appropriate for the given situation. Recent advancements in speech synthesis using neural models have made it possible to create voices that can sometimes sound indistinguishable from human voices [1]. However, these state-of-the-art voices are designed mainly for simple transactions or narrations, such as reading from a document, and they are not yet suitable for the more complex social interactions that social robots are designed for. Some attempts have been made to create naturalistic voices with features like hesitation, fillers, and other elements used by humans to communicate their intentions and emotional states [2], [3]. However, these solutions are still rare in the field of human-robot interaction, and a competent evaluation method for naturalistic voices has not yet been established.

In this paper, we present a text-to-speech model (CustomTTS) capable of generating natural spontaneous speech with prosody control for use in social robot interactions. This model was compared to two other sources of synthetic speech: a leading commercially available neural text-to-speech engine, and natural human speech samples collected from participants while engaged in a collaborative task. Our aim was to gain valuable insights into the performance of these voice sources and their potential applications in social robotics and human-robot interaction.

Our evaluation methodology departs from traditional techniques common in the field of speech synthesis that simply rate and compare voice samples [4], [5]. Instead, we focus on the specific application of creating a synthetic voice for multiparty tabletop interactions by presenting participants with a specific context before evaluating the voices. Our study is designed to evaluate and compare the naturalness, prosody, social impression, and perceived intelligence of these different voice sources, both in embodied and disembodied conditions.

Samples generated by Custom TTS received a more similar score to the human voice compared to the commercial TTS, which has been optimized for clarity rather than human-likeness (See Figure 1). This indicates that the authoring tool produces speech that is more like human speech and more conversational. In addition, our study found that embodiment had a greater effect on the perception of the commercial TTS than on the authored and human voices (See Figure 2).

II. Related work

A. Speech synthesis

In recent years, widely used voice assistants such as Siri, Alexa, and Google Assistant have utilized Text-To-Speech (TTS) voices with a neutral and easily comprehensible speaking style [6]. State-of-the-art TTS is developed on read aloud speech and generally recorded in good studio conditions, which makes it largely optimized for clarity. These voices are suitable for interactions where the agent is meant to respond to commands and not express its own opinions. However, these voices may be limited in social and embodied scenarios, such as tabletop games, where it is important for the agent to convey its attitudes towards what it is saying [7]. In these situations, a more spontaneous speaking style may be necessary, making TTS voices that can display emotional and social cues a crucial aspect of human-robot interactions.

The most common method of building a style-specific TTS voice is to record a training corpus where someone reads typical utterances in the desired manner of speaking. This method has been utilized for creating an emotional TTS by asking voice actors to read drama scripts in a specified emotional state [8]. The same technique has been applied to develop TTS voices with unique personalities for animated characters in speech-enabled computer games [9].

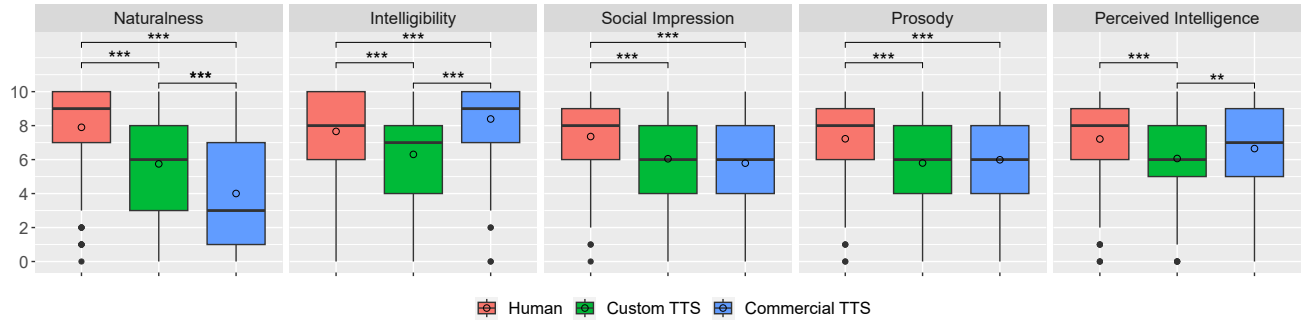

Fig. 1. Mean (represented by circle) and median (represented by the middle bar) ratings for all questions for the three voices. ($*p < .05, **p < .01, ***p < .001$)

RyanSpeech is a corpus that was specifically recorded to train a conversational TTS voice [10]. The professional voice actor was asked to read texts from real-world dialogue settings such as chatbots and task-oriented dialogue systems. However, this method has a limitation, as it can be challenging even for a voice actor to sound spontaneous [11] and convey genuine emotions when reading from a pre-written dialogue script. One way to improve naturalness in TTS voices with a specific style is to train it on unscripted spontaneous data. This approach has been used to build spontaneous TTS voices using podcast data, and it was found to be effective in generating a more natural-sounding voice [12], [13]. The current study also adopts this approach and uses a modified implementation of Tacotron2 to build a spontaneous TTS voice on the Trinity Speech-Gesture Dataset [14], [15]. The TTS training corpus includes annotations for breaths, fillers, and style breaks, as well as normalized measures of mean pitch and speaking rate per style unit. These additional features allow for greater control over the resulting synthesized speech, leading to a more natural-sounding voice [2].

B. Voices for Robots and Virtual Agents

Several studies have investigated the preferred voice for interactive agents and their impact on user perceptions. A study on the preferred voice for a virtual health coach found no significant difference between a general-purpose TTS voice and a limited domain TTS voice in terms of naturalness, conversational aspects, and likeability [16]. Another study found a strong correlation between users' ratings of co-presence and their general liking of the voice, whether recorded or TTS, for a virtual advisor [17]. A comparison of a human voice with a copy-synthesis version of a voice showed that lowering voice quality reduced the perceived expressiveness of the speech but did not affect appeal, credibility, human-like behavior, or fit with the virtual human [18]. A study comparing the use of standard TTS, modern TTS, and human speech as the voice of a virtual agent revealed that human speech was more trusted than either TTS voice [19].

When comparing TTS voices of the robot Sophia and IBM Watson with human speech it was found that human speech was preferred over TTS voices due to the latter not sounding like they meant what they said and lacking cues found in human speech [20]. Another study revealed that speaking style in virtual agents affects naturalness and aliveness ratings, with human-like speaking reducing these ratings [21]. A comparison of unit selection TTS, neural TTS, and human speech as the voice of a virtual agent found that the highly natural-sounding neural TTS did not result in higher social ratings, which the authors attributed to the uncanny valley or the lack of appropriate prosodic realizations and pausing behavior in the TTS [22]. The neural TTS in that study used read speech by a voice actor which is the common practice in speech synthesis. In the current study, we aim to study both the impact of embodiment in the perception of TTS voices, but also to enhance the naturalness of the synthesized speech by using a neural TTS. The neural TTS was trained on unscripted spontaneous speech data, and manipulated by an authoring tool that provides control over breathing, filled pauses, silent pauses, and prosodic realization.

III. Case Study

To obtain naturalistic speech in the desired context, we collected data from people playing a tabletop game. The following sections are a description of the game and how the data was collected.

A. Scenario

The selection of the tabletop game Pandemic [1] as the context for the study was informed by the results of a small preliminary data collection [23] that analyzed the lexical diversity and speech interaction potential of various collaborative tabletop games. Pandemic was identified to provide extremely rich speech interaction opportunities. Pandemic was released in 2008 and has gained recognition as one of the most successful cooperative board games. A short version, "Pandemic Hot Zone - Europe" can be played by 1-4 players in 30 minutes. In the game, players work together to prevent the spread of diseases in Europe by traveling to major cities and curing the diseases represented by colored cubes. The game's physical components, including a board and various

[1] https://zmangames.com/en/games/pandemic/

game pieces, provide opportunities for collecting naturalistic speech data in physically grounded dialog.

B. Data collection

The data collection involved 16 participants who were divided into four groups of four and played the game Pandemic. For the purpose of clarity, in the rest of this paper, we will refer to these participants as *players* to distinguish from the participants in the main perception study. The players had an age range of under 20 to 40, with one player under 20, 12 between 20-30 and three between 30-40. The sample comprised of equal number of men and women, and each group had at least one man and one woman. The players were recruited from a database of people who had previously participated in experiments at the research laboratory. Prior to the experiment, they signed a consent form, agreeing to have their data collected and shared with other researchers in an anonymized form. They were each given a 200 SEK voucher as a token of appreciation for their participation.

The data collection was recorded using five cameras and four Sennheiser ME3 microphones, resulting in 9 audio channels[2]. The Sennheiser ME3 microphones were selected for their ability to provide clear channel separation. The audio was transcribed using Google Cloud's speech-to-text technology.

IV. Voice and embodiment design

This paper evaluates two synthetic voices in addition to the human samples extracted from the previously mentioned dataset. The first synthetic voice was created with Custom TTS and the second is a state-of-the-art commercial neural TTS. The chosen embodiment for evaluating the voices is also described in this section.

A. Custom TTS: Spontaneous TTS with prosody control

We built Custom TTS with the neural TTS engine Tacotron 2 [14] [3], with modifications to utilize the spontaneous corpus used for training and to enable prosody control [2]. A similar TTS system has demonstrated better conversational quality compared to traditional read-speech TTS in both spontaneous speech prompts and chat-bot generated prompts [3].

The training corpus created from the speech recordings of a single-speaker dataset [15] that consists of 25 spontaneous monologues by a male speaker, on average 10.6 minutes long, totaling around 4 hours of speech audio. The speaker spoke in a colloquial style and without interruption, thus the speech contains naturally occurring fillers ('uh', 'um' etc.) which are labeled so that the TTS system would learn to synthesize them. This is a departure from the typical corpus that commercial TTS systems are trained on, which are often scripted and without naturally occurring fillers, thus making the commercial TTS systems not able to sound spontaneous or synthesize fillers. The spontaneous corpus is segmented by automatically labeled breaths [25]. We call each breath-segmented part a breath group and form training utterances by concatenating 2-3 continuous breath groups. This approach is called *breath group bigrams* [26] and has been shown to facilitate more robust training of spontaneous TTS.

To achieve prosody-controllability, we extracted pitch and speaking rate at the breath group level automatically using a prosody analysis tool and input them to the TTS model during training. This is similar to the approach in [2]. We made a graphical user interface where it is possible to insert tokens in the text to synthesize silent pauses, breathing and fillers, and control the speaking rate and pitch of words or phrases to make the controllable aspect of our Custom TTS more accessible [4].

B. Commercial read-speech TTS

For our comparison, we selected Amazon Polly as the Commercial TTS model. Amazon Polly is recognized as one of the leading publicly available TTS systems. To ensure a fair comparison, we chose the 'neural Matthew' voice, which is at the time of this publication listed as Amazon Polly's preferred English male voice. The transcriptions from the data collection described in section III-B were used to generate the commercial voice using the Amazon Polly web interface. This voice will be referred to as the "Commercial Voice" throughout the rest of the paper and was generated without any additional custom tags offered in the TTS interface.

C. Virtual Furhat

For evaluating the different voices, we made use of a virtual social robot from Furhat Robotics. This allowed us to animate the speech of the synthetic and natural voices, and capture the result for evaluation. The speech animation was generated using a custom algorithm that took force-aligned phonetic transcripts and prominence estimates as inputs.

To ensure consistency across all voices, a gender-neutral face was selected for the virtual robot (a modified version of the Furhat "default" character). The animations were played back using a python script via the Furhat remote API and recorded using screen capture.

V. Evaluation

A. Stimuli

The aim of the study is to evaluate the Custom TTS (described in Section IV-A) and compare it to the two baselines: Commercial TTS (described in Section IV-B) and spontaneous human voice. To evaluate the voices, we selected utterances from the dataset described in Section III-B that represent important speech contexts that occur in the game: *Planning an Action, Executing an Action*, and *Reading from a Card*. We focused on these speech contexts as they are the

[2]The dataset features many other modalities and annotations and is divided in 2 groups of players that knew each other, and 2 that did not. Contact us if you would like to request access to the dataset.

[3]We used a PyTorch implementation https://github.com/NVIDIA/tacotron2. For vocoding (waveform generation), we fine-tuned the pre-trained universal model of HiFi-GAN [24] on the corpus.

[4]Technical details and the experiment setup, such as the audio and video files used in the reported experiment can be found here: https://sites.google.com/view/spontaneousrobotspeech

most common types of speech-task contexts during a player's turn.

Read speech is a common source of training data for TTS, making *Reading from a Card* a good baseline for comparison with the more spontaneous speech contexts. *Planning an Action* involves vocalizing intentions and often seeking assistance, which results in the use of fillers and hesitation in the speech. When *Executing an Action*, players use referential speech to accompany their actions, leading to highly spontaneous and expressive speech that is often referential.

From each of the four experiments in the board game dataset, a male and a female player were chosen. From each player three utterances corresponding to the specific speech contexts were selected. A total of 24 samples that were good representatives for the speech-task contexts were selected. The 24 samples were carefully manually transcribed and synthesized using the neural voice from the Commercial TTS and Custom TTS. In the text input to both systems silent pauses and breath pauses were approximated with commas and periods. The Commercial TTS samples were obtained by inputting the transcriptions in the Amazon Web Services online tool 'Amazon Polly'. Custom TTS samples were obtained by using the transcriptions and having an expert work with the synthesis tool described in section IV-A. The expert listened to the recorded human speech in order to achieve a comparable manner of speaking in the Custom TTS. First the overall speaking rate was adjusted to be similar the human version, and style breaks were added in the transcription to make it possible to change the prosody of certain parts. Then the normalized speaking rate and f0 input features were adjusted for each segment to match human realisation in terms of prominence and prosodic turn taking cues. This made it possible to generate speech with prosodic cues like continuation rise and final fall in multi-part utterances, and a slower speaking rate and increased pitch to mark prominence.

For the embodied condition each audio sample was used to generate a video of a virtual robot with lip-sync as described in section IV-C. For each voice condition (Custom TTS, Commercial TTS and human voice), 24 (8 players times 3 contexts) video samples were generated, in total 72 videos.

B. Measures

We chose a short version of the MOS-X questionnaire (MOS-X2 [27]), as the main measure in our study. It is inspired by the original Mean Opinion Scale (MOS) questionnaire and it is a four-item questionnaire that is similarly developed for evaluating shorter stimuli (like samples of voices). It measures four factors: *Naturalness, Intelligibility, Social impression and Prosody*.

Additionally, two questions from the perceived intelligence dimension of the Godspeed questionnaire [28] (*"Incompetent/Competent"* and *"Unintelligent/Intelligent"*) were added to measure the level of perceived intelligence that participants attribute to the voices. In the original Godspeed questionnaire they are semantic differential 7 point scales. They where changed to 11 point scales (0 to 10) to harmonize with the MOS-X questionnaire. The two questions are averaged resulting in a perceived intelligence measure that is used in our analysis (Cronbach-$\alpha = 0.905$).

C. Procedure

The experiment takes approximately 15 minutes to complete. Before the participants completed the survey they were informed that the task would be a study on computer synthesized speech. They were informed about how many screens they would go through and that they should spend about 5 minutes on each screen. We deliberately avoided to use any description that would indicate an embodiment or the presence of a mixture of synthesized and human speech. The survey consisted of three pages where the participants compared speech samples from the two different synthesized voices and the human voice side by side. The voices were speaking the same words and the human voice belonged to the same human player for all three pages. For each of the voices, the participants were first presented with a short description of the context (either 'Someone is playing a board game and they are **reading from a card**', 'Someone is playing a board game and they are **planning an action**' or 'Someone is playing a board game and they are **executing an action**'). Then the stimuli were displayed side by side and they were asked to rate the selected items from MOS-X2 [29] and godspeed [30] for each sample, 6 questions for each of the three voices (see table I). Then they were informed that the voices spoke the same words and they were asked to transcribe what they said as an attention check. This was also served the purpose of promoting additional listens that could strengthen the within-subject nature of our task by for instance making the participants reflect on the intelligibility of each voice and adjusting their answers. The next items in the questionnaire, that also promote multiple listens, were to choose their favourite and least favourite voices and give a open-ended motivation for their choice. Half of the participants were given disembodied (audio only) samples and the other half were given the embodied condition with video and audio. We used a within-subject design to collect open-ended feedback and compare the different voices. However, we used a between-subject design to compare the embodied and disembodied condition to avoid potential order exposure effects.

D. Participants

The participants were recruited from Prolific [5], a crowd-working platform. They were sourced with a filter for gender balance (41 female, 42 male, 9 unknown) and all participants were classified in the system as native English speakers. Additionally, to look for preferences of possible habituation to different accents, the current country of residency of each rater was taken into account when choosing participants. The two synthetic voices had accents from the United States (Amazon Polly) and Hiberno (Irish) English (Custom TTS).

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

Question	Label Left	Label Right
Naturalness: How natural (pleasantly human-like) was the sound of the voice?	Perfectly natural	Extremely unnatural
Intelligibility: Please rate the extent to which it was easy or difficult to understand what the voice was saying	Completely Intelligible	Completely Unintelligible
Prosody: To what extent were the elements of timing, pitch, and emphasis appropriate for the messages?	Always appropriate	Completely inappropriate
Social impression: To what extent was the tone of voice socially and emotionally appropriate for the messages?	Always appropriate	Completely inappropriate
The voice represents someone who is:	Competent	Incompetent
The voice represents someone who is:	Intelligent	Unintelligent

TABLE I
QUESTIONNAIRE PRESENTED TO THE PARTICIPANTS TO RATE EACH VOICE ON AN 11 POINT SCALE FROM LEFT TO RIGHT.

Participants listed the United Kingdom as a country of residence 58 times, the United States 25 times with the remaining 9 unknown. Out of the participants, 48 rated the embodied condition and 44 rated the disembodied condition. In total 92 subjects (each rating 3 different contexts and 3 different voices for each context) participated in the experiment.

VI. RESULTS

A. MOS-X2 and Perceived Intelligence

To examine the quality of the samples, a repeated measures two-way ANOVA was used with two within subject factors, (1) *voice* with the three different voices (*human*, *custom* and *commercial*), (2) *context* with the three different contexts (*reading*, *planning an action* and *executing an action*) and one between subjects factor (*embodiment*). This analysis was applied for each of our five target measures from the items in the questionnaire. The multivariate results of the GLM repeated measures suggests significant differences in the within subjects factor of voice (Wilks' Lambda = .231, F(10,81) = 29.97, $p < .001$), and in the interaction between voice and context (Wilks' Lambda = .645, F(20,71) = 1.95, $p = .021$). The other factors and interactions were not significant.

1) Naturalness: The Human voice got a significantly higher mean score in Naturalness (M = 7.878, sd = .194) compared to both *custom* (M = 5.679, sd = .241) ($p < 0.001$) and commercial (M = 3.976, sd = .296) ($p < 0.001$). The Commercial TTS also got a significantly lower mean score in Naturalness than *custom* ($p < 0.001$) placing Custom TTS above the other synthetic voice and below the human voice in mean score. This was true regardless of context.

2) Intelligibility: Post hoc pairwise comparisons using the Bonferroni correction revealed a significantly higher mean score in Intelligibility for *commercial* (M = 8.459, sd = .149) compared to both *human* (M = 7.612) ($p < 0.001$) and *custom* (M = 6.289), ($p < 0.001$). Custom TTS got a significantly lower mean score than both other voices ($p < 0.001$) placing the human voice in the middle of the two synthesized voices in terms of the intelligibility score. For samples of *context* 'reading', 'commercial' (M = 8.898, sd = .150) had statistically significant higher mean scores in Intelligibility when compared to ''human' (M = 7.203, sd = .262) ($p < .001$) and *custom* (M = 6.467, sd = .255) ($p < .001$). However, there was no significant statistical difference in intelligibility between the human voice and the Custom TTS for 'reading'. For samples of context 'Planning an action' and 'Executing an action', the Custom TTS had a significantly lower mean score ($p < 0.001$) than both other voices. However, there is no statistically significant difference in the mean score for human voice and the Commercial TTS.

3) Social impression: The Human voice got a significantly higher mean score in Social Impression (M = 7.275, sd = .169) compared to both *custom* (M = 6.008, sd = .190) ($p < 0.001$) and commercial (M = 5.813, sd = .249) ($p < 0.001$). There was no statistically significant difference in the mean between the synthetic voices. This is also true for samples of context 'Planning an action' and 'Executing action'. For samples of context 'Reading' the human voice had a significantly higher mean score than the Custom TTS ($p < 0.001$). However, no statistically significant difference could be found between the means of the synthetic voices or between the human voice and the Commercial TTS.

4) Prosody: The Human voice got a significantly higher mean score in Prosody (M = 7.190, sd = .168) compared to both *custom* (M = 5.679, sd = .199) ($p < 0.001$) and commercial (M = 6.017, sd = .214) ($p < 0.001$). There was no significant difference in the mean between the synthetic voices. This is also true for samples of context 'Planning an action' and 'Executing action'. For samples of context 'Reading' the human voice had a significantly higher mean score than the Custom TTS ($p = 0.008$). However, no statistically significant difference could be found between the means of the synthetic voices or between the human voice and the Commercial TTS.

5) Perceived Intelligence: The CustomTTS got a statistically significant lower mean score in Perceived Intelligence (M = 6.018, sd = .179) compared to the human voice (M = 7.137, sd = .168) ($p < 0.001$) and the Commercial TTS (M = 6.736, sd = .186) ($p = 0.006$). There was no statistically significant difference in the overall mean between the human voice and the Commercial TTS . For samples of context 'Reading' there is no statistically significant difference in the mean for the human voice and the Commercial TTS. However, for that context, the human voice and the Commercial TTS have a statistically significant higher mean than the Custom TTS ($p = 0.35, p = 0.03$). For samples of context 'Planning an action' and 'Executing an action' the human voice had a significantly higher mean score than

the Custom TTS ($p < 0.001$). However, no statistically significant difference could be found between the means of the synthetic voices ($p = 0.211, p = 0.276$). There was no significant difference between the means of the human voice and the Commercial TTS for context 'Planning an action' ($p = 0.211$). A significant difference between the means of human voice and the Commercial TTS ($p = 0.018$) for the speech context 'Executing an action' was found.

6) Embodiment: An embodiment between subject effect was found on the Social Impression measure, where the 'disembodied' condition showed an average lower mean score (M = 6.093, sd = .193) when compared to the 'embodied' virtual Furhat condition (M = 6.637, sd = .185) ($p = .045$). Embodiment also appears to boost the naturalness rating of Commercial TTS, see figure 2.

B. Preferred voices

Participants chose Human as their overall preferred voice 48%, Custom TTS 14% and commercial 38%. Interestingly, adding the embodiment closes the gap between human and commercial TTS. Second place is more evenly spread, while the Custom TTS is placed last more commonly (50%) and the natural voice to a lesser degree (18%). Grouping these results regarding context also showed a very similar distribution. We believe that the underwhelming performance of Custom TTS in this aspect regards in part to the order of the questionnaire, given that these questions are asked after the part of the study which involved transcribing what was said in the samples. As we could see from the intelligibility results, this task is probably better fulfilled by recurring to 'commercial' which might explain the extremely significant difference here. Also the many listeners judged the Custom TTS as a noisier version of the human speech. while the Commercial TTS was something completely different, with a clear focus on clarity while speaking.

C. Qualitative results

Participants were also asked to briefly motivate the answer to their favorite and least favorite voice. We analyzed these responses and found the following common themes:

1) Commercial TTS: Almost all participants that chose the Commercial TTS as their favourite stated the clarity and intelligibility of the voice as the main reason for choosing that voice (73 comments), some even stated that the voice sounded 'robotic' but they still preferred its clarity. Some also stated cadence and pacing as reasons for choosing it as the favorite (15 comments). Some participants that commented on why they put a voice as least favourite said they had chosen the Commercial TTS as the least favourite because it sounded boring or unnatural (68 comments).

"Even though the favourite audio sounds like a computer, it was still the easiest to understand. The least favourite was difficult to understand what was said and I had to listen a few times to hear what was being said"

2) Human voice: Almost all of the participants that picked the human voice as their favourite stated in some way that they considered naturalness as the reasons they chose the way they did (95 comments). The participants that picked the human voice as their least favourite and commented on their choice stated that the voice lacked clarity or intelligibility (41 comments).

"I considered the naturalness of the recording to see which sounded like it was a person really playing the game."

3) Custom TTS: Most of the participants that chose the Custom TTS as their favourite stated naturalness as the reason for choosing it (32 comments). Many of them also stated a combination of comprehension and naturalness and/or pitch and tone of the voice. Many participants that chose it as their least favourite stated that the main reason is signal quality (60 comments). Some stated lack of naturalness and timing as well.

"When choosing my favorite and least favorite voices, I considered the speed of the voice and the tone of the voice. "..." [Custom TTS] sounded more charismatic and wouldn't be out of place when playing a board game."

4) Other observations: Some participants stated that they based their choice on the accent of the speaker (16 comments). There was no clear preference for one accent over the other. Another observation is that very few participants stated the context as the reason for choosing a favourite voice.

"...however there seems to be slight twang of an Irish accent in which makes them sound more human."

D. Outlier and Accents

To ensure that the human voices were perceived similarly, a pairwise comparison between the human voices was conducted. One of the human voices stood out because it had a significantly lower mean score in Intelligibility compared to the other speakers. This can be explained by a strong non-native accent that is not present in the other human voices. The samples containing the conditions with that voice were discarded from the tests and the participants reported already reflect this change. Our analysis showed that the accents of the remaining human voices did not affect the results.

VII. Discussion

A. Exploring voice evaluation results

Initially we had the expectations that Naturalness, Prosody, Social Impression, and Perceived Intelligence would be rated higher in the human voice, and the authored voice would outperform the commercial neural TTS in these measures. However, we also anticipated that the commercial TTS, given that it is designed for clarity, would outperform the other two voices in the evaluation of Intelligibility. Our results show that this was partially observed. The human voice either outperformed or tied with one of the synthetic voices in all measurements, except Intelligibility. The Custom TTS was perceived as significantly more natural than the Commercial TTS, but it was rated similarly in prosody and social impression and perceived as less intelligent. The speech that was modeled was spontaneous and often thoughtful and filled with pauses. We hypothesize that the low scores obtained in perceived intelligence may arise from the prevalence of utterances with pauses that demonstrate uncertainty. This can

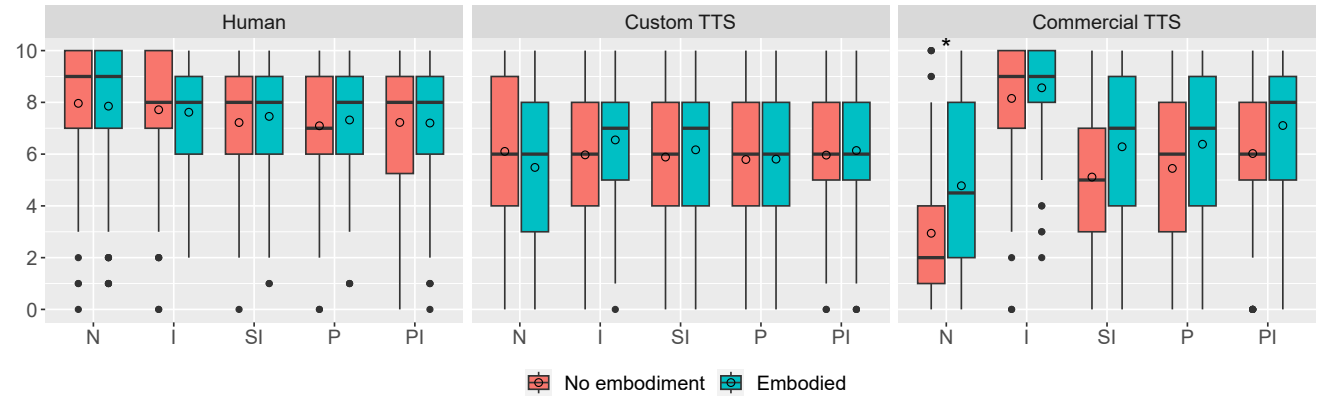

Fig. 2. Mean (represented by circle) and median (represented by the middle bar) scores of all questions for each of the the different voices grouped by embodiment. Error bars show standard deviation. N: Naturalness, I: Intelligibility, SI: Social Impression, P: Prosody, PI: Perceived Intelligence (consisting of the two questions from Godspeed averaged together as described in V-B). $*p < .05$

possibly impact the perceived intelligence of a voice as fast speakers can be perceived as more intelligent. Contrary to our expectations, embodiment did not have an impact on naturalness and perceived intelligence. One possible explanation for that is that there was a lack of emotion expression and more complex non-verbal behaviors in the virtual robot. It has been shown that it is important that the emotion in the speech and facial animation match [31]. However, embodiment had a negligible impact in perceived intelligence and a positive trend in naturalness specifically for the commercial TTS voice where embodiment did appear to affect perception (see Figure 2).

We found a significant interaction between voice type and context which shows the relevance of presenting results separated by context and possibly in introducing context for the evaluation of voice samples. However, the qualitative feedback from users to select their preferred voices did not focus on this aspect. This indicates that a single sentence might not be enough to place participants in the state of mind of a particular context and a better balance between task time and presenting context should be achieved. However, on average participants played each sample several times in each evaluation screen (M = 2.84, sd = 1.99) and took the expected average time in minutes to complete the assignment (M = 14.76, sd = 6.76). This shows that the multi-step evaluation procedure does promote careful listening and provides an appropriate evaluation scheme for iterating on voice evaluation crowd-sourcing experiments. To better assess the use of spontaneous voices in extended interactions, future work should consider incorporating longer dialogue turns. These findings highlight the difficulty in contextualizing evaluations, as has been called for when assessing the quality of modern TTS systems [32].

B. Implications for voice design

Commercial TTS engines are designed to deliver clear and pleasant-sounding speech. However, our study demonstrates that TTS generated from spontaneous speech can also achieve high levels of naturalness, which was our primary design goal. While the Commercial TTS is well-suited for robots that need to provide quick and transactional answers, it is not always necessary for robots to prioritize intelligibility. A natural and playful voice can add value in social domains such as entertainment and healthcare. Furthermore, a voice that expresses uncertainty in the same way as humans can contribute to explainable AI, as it can make the AI appear more transparent in what it knows and what it doesn't know.

VIII. Conclusion

In this paper, we introduce a spontaneous TTS voice (i.e. Custom TTS) that has been trained on spontaneous speech data and can be controlled in terms of fluency and manner of speaking. To evaluate the system we used recordings of humans playing a board game and selected three types of speaking contexts (Reading, Planning an Action, and Executing an Action) to synthesize with our Custom TTS. We performed a user study with embodiment in the form of a virtual robot and without embodiment, comparing the spontaneous speech generated with Custom TTS with two baselines: a commercial TTS and the original human speech.

Our findings demonstrate the potential of training a TTS system on spontaneous speech to create a voice that is more similar to human conversational speech. Moreover, the ability to control fluency and prosody in the TTS system was found to be highly useful. Furthermore, the results highlight how difficult it is to evaluate TTS in terms of prosody and task suitability without a more contextualized and longer evaluation. It is challenging for subjects to envision the usefulness of a robot that doesn't always know everything and consequently needs to sound uncertain at times. In social situations where hesitations and natural spontaneous speech are crucial, such as when playing a game with a robot (or watching a video of such an interaction), where the robot acts as a peer, the advantages of the Custom TTS voice over a Commercial TTS would likely be even clearer. We conclude that for many human-robot application scenarios it is not only important to focus on clarity or on what a robot says, but also how the robot says it.

ACKNOWLEDGMENT

This work was supported by the Swedish Research Council project 2021-05803, and Connected (2019-05003), the Riksbankens Jubileumsfond project CAPTivating (P20-0298), and Digital Futures, project "Advanced Adaptive Intelligent Systems".

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," *Proc. Interspeech 2022*, pp. 4990–4994, 2022.

[3] S. Wang, J. Gustafson, and E. Székely, "Evaluating sampling-based filler insertion with spontaneous tts," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, 2022.

[4] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[5] P. L. Salza, E. Foti, L. Nebbia, and M. Oreglia, "Mos and pair comparison combined methods for quality evaluation of text-to-speech systems," *Acta Acustica united with Acustica*, vol. 82, no. 4, pp. 650–656, 1996.

[6] M. P. Aylett, B. R. Cowan, and L. Clark, "Siri, echo and performance: You have to suffer darling," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–10.

[7] N. G. Ward, *Prosodic patterns in English conversation*. Cambridge University Press, 2019.

[8] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[9] J. Gustafson and K. Sjölander, "Voice creation for conversational fairy-tale characters," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[10] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," *arXiv preprint arXiv:2106.08468*, 2021.

[11] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous tts with neural hmms," in *ICASSP*, 2023.

[12] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data." 2019.

[13] J. O'Mahony, C. Lai, and S. King, "Combining conversational speech with read speech to improve prosody in text-to-speech synthesis," in *Interspeech 2022*, 2022.

[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[15] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 93–98.

[16] K. Georgila, A. W. Black, K. Sagae, and D. Traum, "Practical evaluation of human and synthesized speech for virtual human dialogue systems," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 3519–3526.

[17] A. Abdulrahman and D. Richards, "Is natural necessary? human voice versus synthetic voice for intelligent virtual agents," *Multimodal Technologies and Interaction*, vol. 6, no. 7, p. 51, 2022.

[18] J. P. Cabral, B. R. Cowan, K. Zibrek, and R. McDonnell, "The influence of synthetic voice on the evaluation of a virtual character." in *INTERSPEECH*. Stockholm, 2017, pp. 229–233.

[19] E. K. Chiou, N. L. Schroeder, and S. D. Craig, "How we trust, perceive, and learn from virtual humans: The influence of voice quality," *Computers & Education*, vol. 146, p. 103756, 2020.

[20] K. Kühne, M. H. Fischer, and Y. 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*, p. 105, 2020.

[21] J. Ehret, A. Bönsch, L. Aspöck, C. T. Röhr, S. Baumann, M. Grice, J. Fels, and T. W. Kuhlen, "Do prosody and embodiment influence the perceived naturalness of conversational agents' speech?" *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 4, pp. 1–15, 2021.

[22] T. D. Do, R. P. McMahan, and P. J. Wisniewski, "A new uncanny valley? the effects of speech fidelity and human listener gender on social perceptions of a virtual-human speaker," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–11.

[23] J. Miniotaite and A. Pereira, "Tabletop games as multimodal datasets for social ai," in *Workshop on the representation, sharing and evaluation of MultiModal Agent Interaction (MMAI)*, 2022.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6925–6929.

[26] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7649–7653.

[27] J. R. Lewis, "Investigating mos-x ratings of synthetic and human voices," 2017.

[28] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," 2008.

[29] J. R. Lewis, "Investigating mos-x ratings of synthetic and human voices," 2017.

[30] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, pp. 71–81, 2009.

[31] C.-J. Chang, L. Zhao, S. Zhang, and M. Kapadia, "Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis," *Computer Animation and Virtual Worlds*, vol. 33, no. 3-4, p. e2076, 2022.

[32] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, E. Szekely, C. Tånnander *et al.*, "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.