The Effect of Human Prosody on Comprehension of TTS Robot Speech

Adam K. Coyne[1,2] and Conor McGinn[1,3]

Abstract— The ability to interact verbally with humans is a key requirement of many social robots. It is common however for robot speech to lack contextual human-like prosody, making it intelligible but seeming inexpressive and cold. We investigated the effect that applying human-like prosody to synthetic speech had on aural comprehension during human-robot interaction. A text-to-speech system was used to generate synthetic sentences in two conditions: 'default' and 'human'(informed by voice actor). A speech-in-noise experiment was then performed that required participants to transcribe perceived sentences spoken by a robot in both test conditions. Overall, we found no significant difference in comprehension between sentences spoken using the synthetic voice with prosody and the unaltered synthetic voice, however significant differences in comprehension were detected for shorter sentences (n=50), and among participants that learned English in a different country to the native dialect of the voice actor (n=26). In both of these cases, participants found the voice with human-like prosody harder to comprehend. These findings suggest that introducing human-like prosody to synthetic speech in human-robot interaction, under certain circumstances, may lead to the voice becoming less intelligible. This motivates further research and adds to the growing body of literature on the multifaceted role that voice plays in human-robot interaction.

I. Introduction

Language can be considered be the primary channel of communication between humans, and as such can be of critical value in human-robot interaction. The open-ended nature of language makes it a universal, versatile tool for issuing commands and receiving feedback from robots, but it can also pave the way for less utilitarian, more social paradigms of interaction.

As well as explicit semantic content that enables people to exchange information rapidly and effectively, speech also contains implicit information, providing insight into the background of the speaker (i.e. age, nationality, gender), their affective state, and identity [1]. Much of this implicit information is captured in the prosody of the voice. Prosody is understood to be a vocal nonverbal signal, which "in perceptual terms, accounts for how something is said"[4] [3]. It refers to the rhythm, stress, intonation, etc. of speech, and "plays a role in the comprehension of spoken language by human listeners" [4]. It is language- and culture-dependent [5], [6], [7], and is a contributor to accent, both foreign [8], [9] and native [10], [11], [12].

[1]Robotics and Innovation Lab, Dept. of Mechanical and Manufacturing Engineering, Trinity College Dublin, Ireland
[2]kavana21@tcd.ie
[3]c.mcginn@tcd.ie
[4]quoted from [2, p. 1063]

While some HRI studies have utilized human voices, namely through tele-operation or using pre-recorded soundbites, most tend to use text-to-speech (TTS) engines that generate synthetic speech in accordance with the semantic text content provided and the markup language supported by the software. These TTS programs provide robots with different voices, accents and prosody, offering the potential for robots to appear across a spectrum of gender, age, and expressiveness.

While it may seem intuitively advantageous for a robots' voice to exhibit high levels of prosody, there are several factors that could negatively affect the experience. On one hand, if the TTS system approaches human speech in terms of naturalness, it is possible an uncanny effect is observed, which may in turn cause a mismatch between how the robot sounds, and how it appears and behaves. Secondly, while changing prosody may enhance the robots ability to convey affective expression, artifacts introduced through modification of the intonation and pitch of the voice may affect intelligibility of the speech signal. It has been shown that prosody can increase intelligibility, but only if the listener is culturally familiar with the prosody source, [7] potentially making the robot's speech harder to comprehend if not.

Recent HRI research has found that changes to the prosody of a robot's voice can enhance its ability to express affective states, however the effect of prosody on intelligibility, while understood for human speech, remains almost entirely unexplored for robot speech, leaving robot designers with little insight into how best to implement such a feature.

In this paper, we aim to explore how people comprehend robot speech that is tuned to imitate human prosody. Using a mainstream TTS engine and experimental methods drawn from the field of linguistics research, the intelligibility of unaltered, 'default' sentences was compared to 'prosodic' sentences manually tuned to match the prosody of human actors. We examined factors that we postulated might affect intelligibility of prosodic speech; the length of the sentence uttered (long/short), and whether or not the listener was a native English speaker that learned English in the same region as the human actor used to model the prosody of the synthetic speech. We anticipate that the findings of this research will be of broad interest to the HRI community, especially to designers and researchers engaged in experiments involving speech interaction.

II. Prior Work

The ability to produce prosodic speech to differing extents is a feature in many speech synthesis engines [13], often through the use of the World Wide Web Consortium standard Speech Synthesis Markup Language (SSML) [14], [15]. Despite speech synthesis being essential for speech-capable robots, recent research by McGinn and Torres indicated that HRI researchers often do not give much consideration to the voice the robot is given during HRI experiments, and which has been shown to affect the mental models that they subsequently form of the robot [16], [17].

Existing research into robot speech prosody has primarily been towards the goal of conveying emotion. As reviewed in 2016 by Crumpton and Bethel [18], conveying emotion through robot speech has been done both by fine-tuning prosody in speech sections (often using SSML) or by applying more global prosody settings to entire statements (with other markup tools such as EmotionML [19]). Other studies explore the impact of prosody on naturalness and robot acceptance, notably through the prosodic feature of intonation. In a study by van Straten et al. [20] robot intonation was shown to impact the affective state of children with autism spectrum conditions, and more recently, Velner, Boersma, and de Graaf [21] studied how humanlike intonation for robot speech could affect how it was perceived. The less-explored visual component of prosody has also seen initial research in the context of human-robot interaction by Hill and Vaikiotis-Bateson, through a virtual talking head "adaptable to robots" [22, p. 63].

The use of prosody to increase robot speech intelligibility was explored in a preliminary study by Lee [23]. In a within-subjects study, comprehension was measured for 21 participants through recall of weather forecast information expressed by the robot, and differences were compared between three levels of overall pitch in the robot's voice (presented in fixed order to participants).

The study of robot speech being inherently multidisciplinary (drawing from robot engineering as well as linguistics and audiology), robust methods for assessing intelligibility are seldom encountered, despite being common in the study of speech synthesis [24], [25], including when pertaining to prosody [26], [27], [28]. For assessing intelligibility of speech in the fields of audiology and linguistics, speech-in-noise (SIN) methods are often employed [29], [30], [31], in which speech is overlaid with noise to increase the difficulty of the listening task to distinguish between control conditions, avoiding "ceiling performance" [31, p. 2]. For this purpose, standard sentence lists such as the Bamford-Kowal-Bench (BKB) list [32] are used. The BKB sentence lists were used in the development of several speech-in-noise tests [33] such as HINT [34] and BKB-SIN [35]. BKB sentence lists and similar lists have also used in context of speech prosody [36], [37].

III. Methodology

A. Preparation of Experiment

Stimuli for this experiment took the form of videos. In each video, a social robot expressed a sentence in one of the four synthetic voice conditions (either male or female, and either 'default' or 'human'). To increase the difficulty of perception of the sentence, the audio in each video was overlaid with speech-shaped noise. This approach, known as speech-in-noise testing is widely used in audiology testing as playing clear audio of each sentence to participants may lead to near-100% comprehension rates, obscuring any differences in test conditions–adding noise avoids "ceiling performance" [31, p. 2]. For this task, speech-shaped noise provided by Harvard Speech Corpus [38] was added to the videos at a signal-to-noise ratio of 4dB. To avoid a harsh start and stop, the noise was set to fade in before the sentence was heard and fade out after.

Synthetic speech generation: Synthetic speech was generated using the Google Speech Synthesis API (June 17 2021 release). This was selected because it was among the most natural-sounding commercial TTS system available, it appeared to be widely used in the HRI community [39], [40], [41], and it supported speech synthesis markup language (SSML), the markup language that allows relatively fine grain control of speech prosody.

First, synthetic speech was created by passing the sentences in the sentence corpus to the TTS system without modification. This produced utterances with relatively generic prosody, that generally lacked contextual expressiveness.

To generate speech with human-like prosody, recordings were first made of 5 voice actors who were instructed to speak each of the sentences in the sentence corpus in an expressive way. All actors spoke with a dialect that was native to the region where the experiment was performed. An informal survey was then conducted in which 5 college students were asked to rank each voice recording by "most-to-least natural"; the top-ranking recording of each sentences was retained and subsequently used as the model to construct the synthetic humanly-prosodic voice. Using Audacity, the speech signal of each of these model sentences was analyzed for pitch, timing (i.e. duration of pauses between words) and rate (i.e. speed that words were pronounced). This information was encoded in SSML and was subsequently used by the TTS engine to generate the synthetic prosodic speech. Pitch was controlled for each individual word using the pitch attribute of the prosody element, timing was controlled using the break element, and the rate attribute of the prosody element was used to slow down or speed up words. This is illustrated in the following example SSML snippet:

```
<s xml:lang="en-UK">
  they <prosody pitch="-2st">waited</prosody> for
  <prosody pitch="+2st" rate="slow">one</prosody>
  <prosody pitch="+1st">hour</prosody>
```

```
  <break time="0.1s"/>
  <prosody pitch="+1st">while the</prosody>
  <prosody pitch="-3st">train</prosody>
  <prosody pitch="+3st">stopped</prosody>
  <prosody pitch="+1st">at the</prosody>
  <prosody pitch="-1st">station.</prosody>
</s>
```

Recordings using both 'default' and 'human' configurations were made in two voice conditions: one male ("en-US-Wavenet-D") and one female ("en-US-Wavenet-G").

Visual presentation: Videos were created showing a social robot speaking each sentence. The Stevie robot [42] was selected for the experiment due to its availability to the researchers as well as its ability to move its mouth while speaking, which was deemed important for the purpose of a video. For this experiment, videos of Stevie talking were recorded on a chroma key backdrop which was replaced with a white background, and overlayed with the generated audio sentences to produce videos for each sentence and test condition.

Sentence corpus and measures: Sentences used in the experiment were taken from the Bamford-Kowal-Bench (BKB) sentence list [32]. Sixteen BKB sentences in total were used. Sentences were only chosen if they were not deemed to be age or culturally dependent. Sentences 1-8 were extracted verbatim from the BKB sentence list, while sentences 9-16 were constructed by combining two standard BKB sentences using connective words. This allowed us to explore how prosody affected the comprehension of relatively larger and shorter sentences. The 16 sentences used in the study are given in Table I.

The BKB lexicon was designed to measure speech comprehension. Each sentence contains 3 or 4 "keywords", and each correctly-transcribed keyword counts as a single point, providing a score out of 3 or 4 for each sentence. For the dual-clause sentences, being composed of two basic sentences, the score was out of 6, 7 or 8. For each sentence, scores were normalized and represented as a comprehension rate between 0 and 1, representing the proportion of correctly transcribed keywords.

TABLE I: List of the shorter (1-8) and longer (9-16) sentences used in the experiment. In accordance with the BKB lexicon, one point was awarded for each word in capital letters that was correctly identified.

	Single-clause (extracted from BKB sentence corpus)
1	YELLOW PEARS were LOVELY
2	Some MEN SHAVE in the MORNING
3	The BIG FISH GOT AWAY
4	SHE'S CALLING her DAUGHTER
5	THEY KNOCKED on the WINDOW
6	THEY WALKED ACROSS the GRASS
7	HE DROPPED his MONEY
8	The DOGS GO for a WALK
	Dual-clause (composite of 2 sentences from BKB sentence corpus)
9	THEY WAITED for ONE HOUR while the TRAIN STOPPED at the STATION
10	The FIRE is VERY HOT but the ROOM'S GETTING COLD
11	The LADY WASHED the SHIRT since the WASHING MACHINE BROKE
12	The OLD MAN WORRIES that a TREE FELL ON the HOUSE
13	The FOOTBALL GAME'S OVER and PEOPLE are GOING HOME
14	The TWO FARMERS are TALKING as THEY SIT on a WOODEN BENCH
15	The DAUGHTER LAID the TABLE, and the CHILDREN are ALL EATING
16	While the TRUCK CLIMBS the HILL, the DRIVER WAITS by the CORNER

B. Procedure

Participants were first asked to read an information sheet and provide written informed consent in accordance with ethics requirements. Afterwards, participants were seated at a PC and asked to put on high quality over-ear headphones.[5] The experiment was implemented in Python using PsychoPy libraries [43]. The experiment was programmed to play sixteen videos of the robot, one for each expressed sentence. Following each video, the participant was required to type the perceived sentence, transcribing as completely and accurately as possible. Once completed, the participant clicked a button to advance to the next video. Screenshots showing different stages of the testing procedure are given in Figure 1.

Videos were presented in random order to each participant, with no repeated sentences and an even distribution of factor permutations (prosody type,). Subsequent participant stimuli were counterbalanced, ensuring each permutation was evenly represented across the sample.

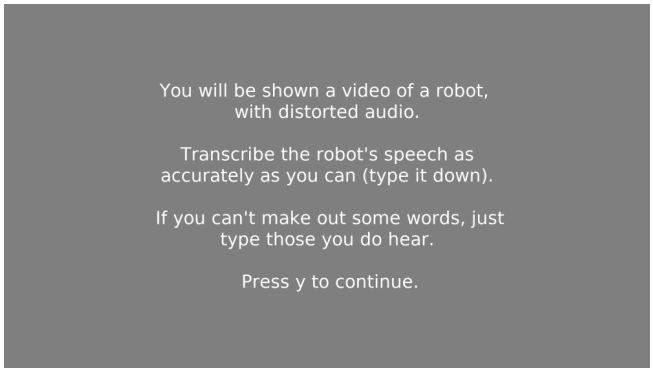
Videos were presented in random order, with randomly selected prosody type for each ('default' vs 'human'), selected in advance to ensure even distribution of test factors (prosody type and sentence length). Test factors were also counterbalanced for subsequent participants—as each participant would only hear each sentence once , this ensured each sentence was expressed in all conditions the same number of times over the course of the experiment.

Upon completion of the experiment, the participant was asked to complete a demographics form. The participant was supervised by a researcher during the experiment and was thanked for their participation upon completion. When the experiment was finished, data was logged to CSV files on the host PC.

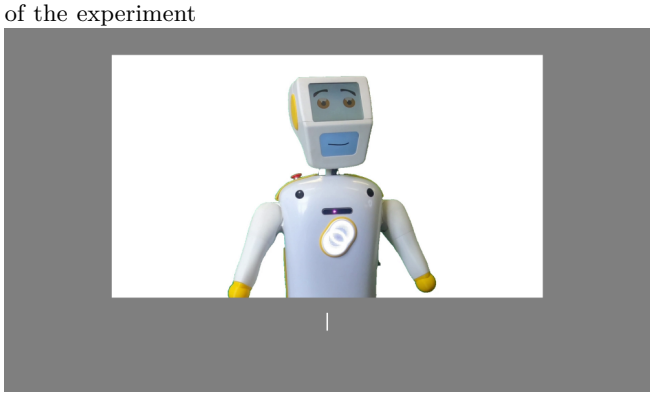
C. Participants

Participant recruitment was conducted in two locations; a university campus and a gift shop in a department store. For each participant, country of origin, country where English was learned, and English language fluency were recorded, as well as whether the participant had any conditions that might impair hearing or speech comprehension.

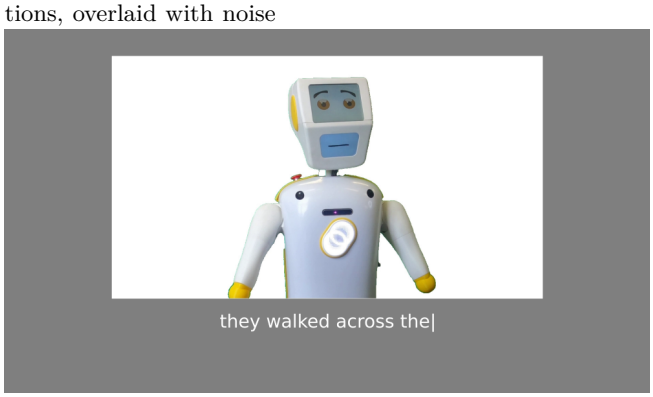
[5]Precautions were taken in regards to the ongoing COVID-19 pandemic; all equipment was thoroughly wiped using antibacterial wipes before participant contact, and personal protective equipment (PPE) was provided.



(a) Test guidelines shared with participants at the beginning of the experiment



(b) A video of the robot taking was shown to the participant with randomly selected and counterbalanced speech conditions, overlaid with noise



(c) Participants could begin to transcribe what they heard before the end of the video

(d) When the video ended, participants could finish transcribing what they heard and move on to the next video

Fig. 1: Screenshots of the testing software.

D. Statistical analysis

In order to analyze differences in means for comprehension rates for both prosody test conditions over the two studied factors of sentence length and familiarity with local English, a mixed-measures 3-way analysis of variance (ANOVA) was conducted. Sentence length and prosody type were within-subjects factors, and familiarity with local English was a between-subjects factor. This form of analysis was selected so that both first-order and interaction effects between factors could be studied, and was calculated in R [44].

IV. Results

In total, 53 participants volunteered for the experiment, of which one that did not complete the experiment, resulting in 52 total participant data points. 11 participants were recruited in the gift shop (mean age of 26, standard deviation of 7.1, 4 female, 7 male), and 41 from a college campus (mean age of 22, standard deviation of 3.5, 10 female, 14 male, 3 other). 31 participants reported being native English speakers, and 26 participants were fluent English speakers that learned English in the local region.

Analysis of first-order effects, with both sentence lengths and both voice types aggregated for each participant, shows a significant difference in the comprehension of 'default' speech ($\mu = 0.742$) and 'human' speech ($\mu = 0.679$); $F(50) = 10.329$, $p = 0.002$. Two second-order interactions with prosody levels were shown; with the factor of sentence length; $F(50) = 5.2935$, $p = 0.026$, and listener locality; $F(50) = 5.1114$, $p = 0.028$. No significant third order interaction was identified between all three factors of prosody level, listener locality, and sentence length; $F(50) = 0.51082$, $p = 0.48$.

Visualisations of results are shown in Figures 2 and 3.

V. Discussion

In this study, we did not find that robot speech exhibiting prosody was easier to comprehend. Rather, the results indicated that the presence of human-like prosody made the speech less intelligible.

One plausible explanation for this effect is that, while prosody might enhance second order effects in human-robot interaction (boosting expressiveness, conveying affect, etc.), possibly increasing robot acceptance, changes it introduces to the pitch, rate and timing of the spoken sentence may come with the trade-off of making comprehension harder, not easier. This is correlated with the prior finding that casual human speech was shown to yield less intelligible prosody than other, more deliberate ways of speaking [45]. Furthermore, as prosody can provide contextual cues, sentences uttered with little context may have been harder to understand. This is supported by the finding that shorter sentences were significantly harder to comprehend when spoken in the 'human' voice condition. As the two clauses of the longer sentences were semantically linked, it may have been

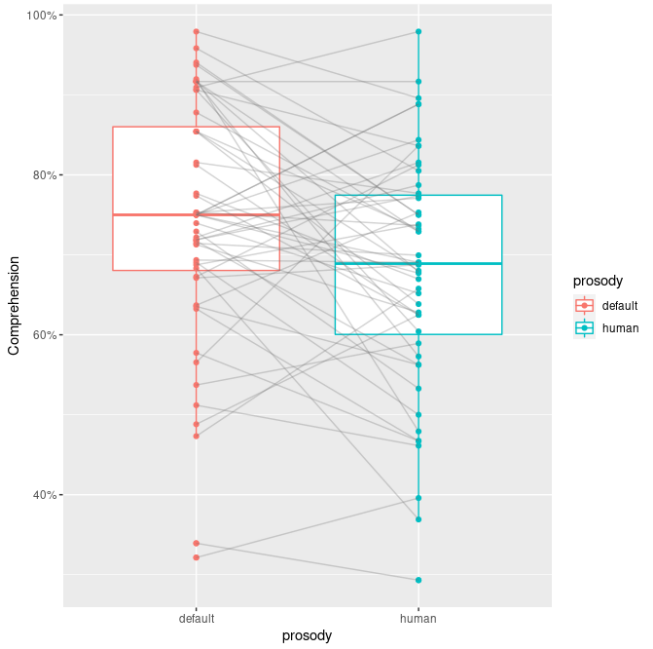


Fig. 2: Box plot of overall paired comprehension ratios for 'default' and 'human' prosody conditions

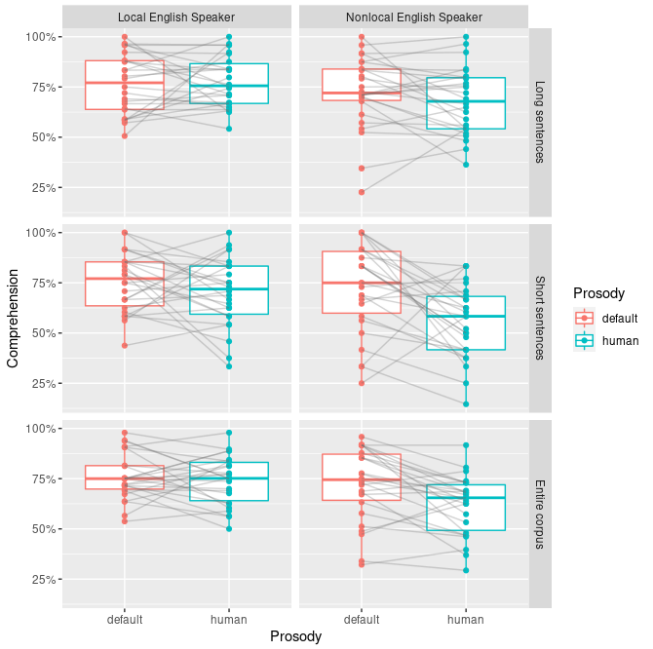


Fig. 3: Box plots of paired comprehension ratios for for 'default' and 'human' prosody conditions, along both levels of participant English dialect locality (local and nonlocal) and sentence length (short, long and overall)

easier to reconstruct misheard words, whereas the above issue is compounded for shorter sentences.

The interaction effect identified between prosody levels and locality on comprehension can be visually observed in Figure 3; the degree to which 'human' prosody negatively affected comprehension was substantially more severe for non-locals (graphs on the right), with substantially smaller differences in means for locals (graphs on the left). This also supports the above explanation; participants who did not learn English in the local region found it more challenging to comprehend prosodic speech modeled on an actor with a dialect from that region. An intuitive reason for this is the region-dependent nature of prosody; prosody that matches an accent the listener is closely familiar with is less likely to cause comprehension issues, but contextual cues provided by prosody from a foreign dialect may be lost, and the resulting speech less intelligible to the listener. This explanation is consistent with findings in the literature regarding foreign prosody comprehension [7].[6]

While not accounting for the observed interaction effects, an alternative explanation for our findings may be connected with limitations in the TTS engine's ability to produce prosodic speech. While efforts were made to model the synthetic speech as closely as possible on natural speech, commercial TTS remain limited in their ability to reproduce the richness of human speech. Although the pitch of whole words could be modified, it was not possible to control the pitch contour, leading to the loss of, for example, any rising inflection on individual words. As a result, it is possible that comprehension was decreased was because the prosody was "not sufficiently human-like".

We know from previous research that prosody has an impact on intelligibility as well as naturalness and acceptance by humans; these results provide analogous evidence that prosody also affects robot speech comprehension, and suggests that design efforts to increase the latter may negatively affect the former. These findings provide new insights into the use of prosody in robot speech, and make clearer its potential benefits and drawbacks when used in robot design and human-robot interaction.

As prosody is context-dependent, robot developers should consider how much context will apply to a robot's utterances, and envisage discarding human-like prosody if there is little context to utterances in its use case. Perhaps more importantly, developers should consider their target audience and culture; depending on the manner of speaking that an audience is accustomed to, a mismatch may result in a decrease in overall comprehension.[7] What our results suggest in this regard

[6] In a 2000 experiment by Pennington and Ellis, it was shown that for Cantonese speakers listening to English, some of the disambiguating benefits of prosody were diminished as "they did not seem to have the relevant knowledge of how [prosody] resolves ambiguity" [7, p. 187].

[7] A deeper factor yet to consider is that, depending on the context, being perceived as a certain ethnicity may trigger discriminatory, even xenophobic biases in the human observer [46], [47]. This has been a topic of conversation in robo-ethics in recent years [48], [49].

is that tailoring robot prosody to match the cultural expectations of the audience may increase acceptance and anthropomorphism without sacrificing intelligibility, but developers should remain aware of the broader social and ethical implications in assigning ethnic traits to a robot, notably if this would contribute to existing racial biases.

Limitations and future work

Several limitations were identified in this experiment, the first being those of the text-to-speech engine. Although mainstream in use and availability, the TTS engine we used did not allow absolute pitch specification, and instead pitch must be determined iteratively relative to the default output for each sentence. Also, as mentioned earlier in the section, the software did not allow us to implement pitch contour control. Although unlikely to be the only reason for the observed difference, especially considering the difference in results between the studied test conditions, the decrease in comprehension when prosody was modified may have been influenced by this shortcoming–more control over prosodic alterations are encouraged in future studies. In the same regard, future experiments building on this might explore the mapping of other human prosody elements aside from pitch and duration, such as loudness and timbre, experimenting with different TTS systems that might offer more varied (or improved) parameter control, or in the case where source prosody is immediately available, exploring non-TTS methods of speech generation such as voice filtering [50].

A single type of noise (speech-shaped noise from the Harvard speech corpus [38]) was used at a fixed signal-to-noise ratio in this experiment. Future work may explore a variety of noise types, either contextual (such as environmental background noises) or generated to match the audio characteristics of each individual synthetic voice. Also, although not necessary for speech-in-noise testing, a range of different signal-to-noise ratios would provide more thorough results.

We explored the impact of cultural background in regards to robot speech comprehension, based on the prosody of local English, showing preliminary indications that a foreign prosody can have an adverse effect on comprehension. The relatively small sample size in this study made it difficult to study distinctions between groups any further; future research may investigate the effects of local prosody for both local and foreign participants, restricted to only native English speakers. Pushing this approach further, as accents have been shown to have an effect on robot perception [51], of interest would be a larger study involving multiple prosodies from different world origins, and their intelligibility by a more diverse pool of participants.

Finally, the robot used in this context was virtual, presented to participants in the form of a video on a PC screen. Employing a physical embodied robot would more accurately simulate the experience of interacting with a robot in a real-world use-case. However, we also anticipate drawbacks of this embodied approach, namely in terms of overall repeatably since ensuring the same level of control over acoustics parameters during the experiment would be notably weaker.

VI. Conclusion

In our study, we explored the effect of applying human-like prosody to robotic speech had on aural comprehension. Our results suggest that under certain circumstances, such as when expressing short utterances with little context, or when the prosody is foreign to the listener, human-like prosody may adversely affect comprehension by comparison to default prosody generated automatically by the TTS engine. Another interpretation of these findings may be that the ability of state of the art TTS engines remains limited, and even under ideal circumstances, synthetic speech generated to have contextual humanlike prosody remains difficult to comprehend.

For robotics developers who are considering enhancing the prosody of their robot platform, the takeaways we propose from this study are a greater awareness of confounding role that prosody plays in the comprehension of robot speech, and its close relationship with context and its cultural provenance.

We hope to also highlight the relatively small overlap between the field of human-robot interaction and those of linguistics, audiology and acoustics, and encourage HRI researchers to embrace experimental methods from these fields when pertinent to ensure scientific rigor.

References

[1] P. Belin, "Voice processing in human and non-human primates," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 361, no. 1476, pp. 2091–2107, 2006.

[2] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in Proceedings of the 16th ACM International Conference on Multimedia, MM '08, (New York, NY, USA), p. 1061–1070, Association for Computing Machinery, 2008.

[3] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: Guide to Algorithms and System Development. Prentice Hall, 2001.

[4] L. Mary, "Significance of prosody for speaker, language and speech recognition," in Extraction and Representation of Prosody for Speaker, Speech and Language Recognition, pp. 1–18, Springer, 2012.

[5] P. Warren, "Prosody and parsing: An introduction," Language and Cognitive Processes, vol. 11, pp. 1–16, 4 1996.

[6] D. Hirst and A. D. Cristo, Intonation systems: a survey of twenty languages. Cambridge University Press, 1998.

[7] M. C. Pennington and N. C. Ellis, "Cantonese speakers' memory for English sentences with prosodic cues," Modern Language Journal, vol. 84, pp. 372–389, 2000.

[8] M. Jilka, "Testing the contribution of prosody to the perception of foreign accent," in New Sounds 2000: Proceedings of the Fourth International Symposium on the Acquisition of Second-Language Speech, pp. 199–207, 2000.

[9] P. B. D. Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," Phonetica, vol. 63, pp. 247–267, 12 2006.

[10] S. Peppé, J. Maxim, and B. Wells, "Prosodic variation in southern British English," Language and Speech, vol. 43, pp. 309–334, 2000.

[11] U. Gut and J.-T. Milde, "The prosody of Nigerian English," in Speech Prosody 2002, International Conference, John Benjamins Publishing Company, 2002.

[12] E. Couper-Kuhlen, "The prosody of other-repetition in British and North American English," Language in Society, vol. 49, pp. 521–552, 9 2020.

[13] Z. Yin, "An overview of speech synthesis technology," in 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), pp. 522–526, 2018.

[14] P. Baggia, P. Bagshaw, M. Bodell, D. Z. Huang, L. Xiaoyan, S. McGlashan, J. Tao, Y. Jun, H. Fang, Y. Kang, et al., "Speech synthesis markup language (SSML) version 1.1," 2010.

[15] P. Baggia, "Speech standards: Lessons learnt," in Human 4.0-From Biology to Cybernetic, IntechOpen, 2020.

[16] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 211–221, IEEE, 2019.

[17] I. Torre, A. B. Latupeirissa, and C. McGinn, "How context shapes the appropriateness of a robot's voice," in 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 215–222, IEEE, 2020.

[18] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," International Journal of Social Robotics, vol. 8, pp. 271–285, 4 2016.

[19] M. Schröeder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato, "Emotion markup language (EmotionML) 1.0," World Wide Web Consortium, Recommendation REC-emotionml-20140522, 2014.

[20] C. L. van Straten, I. Smeekens, E. Barakova, J. Glennon, J. Buitelaar, and A. Chen, "Effects of robots' intonation and bodily appearance on robot-mediated communicative treatment outcomes for children with autism spectrum disorder," Personal and Ubiquitous Computing, vol. 22, pp. 379–390, 4 2018.

[21] E. Velner, P. P. Boersma, and M. M. de Graaf, "Intonation in robot speech: Does it work the same as with people?," in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 569–578, ACM, 3 2020.

[22] H. Hill and E. Vaikiotis-Bateson, "Using graphics to study the perception of speech-in-noise, and vice versa," in Auditory-Visual Speech Processing (AVSP '05), pp. 63–64, 2005.

[23] J. Lee, "Generating robotic speech prosody for human robot interaction: A preliminary study," Applied Sciences, vol. 11, p. 3468, 4 2021.

[24] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," Speech Communication, vol. 55, pp. 572–585, 5 2013.

[25] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?," in Twelfth Annual Conference of the International Speech Communication Association, pp. 1837–1840, 2011.

[26] K. E. Silverman, "On customizing prosody in speech synthesis: Names and addresses as a case in point," in Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993, 1993.

[27] N. Campbell, "Evaluation of speech synthesis," in Evaluation of text and speech systems, pp. 29–64, Springer, 2007.

[28] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," Language and Speech, vol. 40, no. 4, pp. 391–409, 1997.

[29] D. Von Hapsburg, C. A. Champlin, and S. R. Shetty, "Reception thresholds for sentences in bilingual (spanish/english) and monolingual (english) listeners," Journal of the American Academy of Audiology, vol. 15, no. 1, pp. 88–98, 2004.

[30] L. H. Mayo, M. Florentine, and S. Buus, "Age of second-language acquisition and perception of speech in noise," Journal of speech, language, and hearing research, vol. 40, no. 3, pp. 686–693, 1997.

[31] K. J. van Engen, B. Chandrasekaran, and R. Smiljanic, "Effects of speech clarity on recognition memory for spoken sentences," PLoS ONE, vol. 7, 2012.

[32] J. Bench, Åse Kowal, and J. Bamford, "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," British Journal of Audiology, vol. 13, pp. 108–112, 1 1979.

[33] S. Sharma, R. Tripathy, and U. Saxena, "Critical appraisal of speech in noise tests: a systematic review and survey," International Journal of Research in Medical Sciences, vol. 5, p. 13, 12 2016.

[34] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," The Journal of the Acoustical Society of America, vol. 95, pp. 1085–1099, 2 1994.

[35] E. Research, "The BKB-SIN test booklet," 2005.

[36] D.-J. Shin and P. Iverson, "Training Korean second language speakers on English vowels and prosody," The Journal of the Acoustical Society of America, vol. 19, pp. 060048–060048, 2013.

[37] I. Hove and V. Dellwo, "The effects of voice disguise on f0 and on the formants," Proceedings of IAFPA 2014, 2014.

[38] P. Demonte, "Speech shaped noise master audio - HARVARD speech corpus," Oct 2019.

[39] M. J. Hossain, S. M. A. Amin, M. S. Islam, and Marium-E-Jannat, "Development of robotic voice conversion for RIBO using text-to-speech synthesis," in 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), pp. 422–425, IEEE, 9 2018.

[40] J. James, B. T. Balamurali, C. I. Watson, and B. MacDonald, "Empathetic speech synthesis and testing for healthcare robots," International Journal of Social Robotics, 9 2020.

[41] V. Seib, R. Memmesheimer, and D. Paulus, "A ROS-based system for an autonomous service robot," in Robot Operating System (ROS), pp. 215–252, Springer, 2016.

[42] L. Taylor, A. Downing, G. A. Noury, G. Masala, M. Palomino, C. McGinn, and R. Jones, "Exploring the applicability of the socially assistive robot Stevie in a day center for people with dementia," in 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), pp. 957–962, IEEE, 2021.

[43] J. W. Peirce, "PsychoPy–psychophysics software in python," Journal of Neuroscience Methods, vol. 162, pp. 8–13, 5 2007.

[44] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2021.

[45] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and N. Veilleux, "A methodology for analyzing prosody," The Journal of the Acoustical Society of America, vol. 84, pp. S99–S99, 11 1988.

[46] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," British Journal of Social Psychology, vol. 51, pp. 724–731, 12 2012.

[47] A. Esposito, T. Amorese, M. Cuciniello, M. T. Riviello, and G. Cordasco, "How human likeness, gender and ethnicity affect elders' acceptance of assistive robots," in 2020 IEEE International Conference on Human-Machine Systems (ICHMS), pp. 1–6, IEEE, 9 2020.

[48] C. Klein and D. Allan, "Robot racism? Yes, says a study showing humans' biases extend to robots," 2019.

[49] R. Sparrow, "Robotics has a race problem," Science, Technology, & Human Values, vol. 45, pp. 538–560, 5 2020.

[50] A. Gabryś, G. Huybrechts, M. S. Ribeiro, C.-M. Chien, J. Roth, G. Comini, R. Barra-Chicote, B. Perz, and J. Lorenzo-Trueba, "Voice Filter: Few-Shot Text-to-Speech Speaker Adaptation Using Voice Conversion as a Post-Processing Module," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7902–7906, May 2022. ISSN: 2379-190X.

[51] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. MacDonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," International Journal of Social Robotics, vol. 3, pp. 253–262, 8 2011.