Mixing Speech and Semantic Free Utterances: A challenge for Natural Language Generation

Matthew P. Aylett
CereProc Ltd.
Edinburgh, UK
matthewa@cereproc.com

ABSTRACT

Speech synthesis can be regarded as a rendering process. Just as a graphic is rendered, so speech is rendered. As such, much of the underlying control, e.g. expression, emotion, emphasis, is delegated to the system driving the speech synthesis. In this paper we explore the challenge of merging semantic free utterances (SFUs), such as groans, yells, sighs, and for robots beeps clicks and non vocal noises, with speech synthesis. We highlight the problems in designing and synchronising speech and SFUs and present a set of design questions and challenges for the higher level system that is required to generate the combined content. It is unclear where this process needs to be carried out in a conversational system, and we argue the natural language generation (NLG) system should be responsible for controlling SFUs and speech output.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**;

KEYWORDS

speech synthesis, social robots, personification

1 INTRODUCTION

Semantic free utterances (SFUs) have been used extensively in films portraying robots. Yilmazyildiz et al [7] categorise SFUs into four categories: gibberish speech, such as the speech generated by the Sims; non-linguistic utterances, such as squeaks and whirring noises; paralinguistic utterances, such as moans and laughter; and musical utterances, where musical is the basis of designing musical tones or musical instrument sounds for communicating. However, the use of SFUs in human-robot interactions is *"very much in the stages of infancy compared with other areas of HRI"* p32 [7].

The use of SFUs in films is primarily to convey character, often characterising a fictional robot as *cute* or *childlike*, with meaning communicated by context (i.e. WALL.E) or by co-present human actors (i.e R2D2). Commercial robots such as Anki's Cosmo and Vector [6] used SFUs in this filmic manner to characterise their robots as *cute*, while academic work on HRI SFUs has focused more on how the sounds communicate affect and meaning e.g. [4, 7, 8].

For social robots, where there is an assumption the robot needs to effectively communicate to carry out tasks and supply services, rather than use SFUs, speech synthesis has been widely used for its ability to convey complex information clearly and unambiguously. McGinn and Torre point out that the voice chosen also has a key impact on perceived robot character to the extent that *"giving a mismatched voice to a robot might introduce a confounding effect in HRI studies."* p211 [5].

In a pilot study[2], we explored how SFUs may be added to synthetic speech to enrich or create a non-natural but appropriate robot characterisation for a table-top social robot, Haru [3]. The SFUs and the speech synthesis voice were designed separately. Subjects did not appear to merge the SFUs with the robot's voice, instead regarding them as background noises. Furthermore, the addition of SFUs did not alter the perceived personality, age or naturalness of the rendered audio. However, the study was not carried out in an interactive setting and care must be taken in generalising these results. Furthermore embodiment was limited to a picture of a human or a robot and, while this did affect the perception of age and personality, real time movements could not support the perception that the SFUs were generated by the same agent generating the voice. The full set of materials can be heard here *https://tinyurl.com/w9chwa6*.

The prompts were generated by hand. For example:

```
Right answer, [Agreement SFU], you have been practising.
[Agreement SFU], okay, let's play another game.
```

With stimuli generated with and without the SFUs. It was decided not to replace vocal content with the SFU but this would have been option. The locations of the SFUs were chosen by hand as well as a set of sentences balanced over the different SFUs available. This raises the question of how might a higher level NLG system deal with the generation of SFUs. We do not feel qualified to answer this question but we are able to pose some of the design questions and challenges we would envisage in their generation.

2 QUESTIONS AND CHALLENGES IN AUTOMATICALLY GENERATING SFUS

A naive view of NLG is that it takes a semantically described frame such as Loves(John, Mary) and generate the appropriate text *"John loves Mary."* this would then be passed to a speech synthesiser to render the text. The NLG does not have to concern itself with how the speech is rendered, rather it simply creates grammatically correct meaningful text. The speech synthesiser would then decide prosody and emphasis and tone of voice. To a certain extent this does work, default prosody is often acceptable. However, non default

prosody is directly related to underlying meaning (which is not part of traditional NLG output), neither is it independent of text realisation. So in this case the speech synthesis system needs to be told what to do by something, as much as the NLG system may need some idea of vocal style i.e. Sarcastic(Loves(John, Mary). SFUs are a particular case of this lack of independence, similar in many ways to the use of filled pauses in conversational speech.

- Should NLG be rendering SFUs as part of textual output? Based on the semantic and conversational context given to NLG by a dialogue manager? Or should we pass the buck further up stream and expect the dialogue manager to decided this sort of rendered conversational detail?
- In some cases SFUs can directly replace text (Such as musical tone signalling agreement replacing 'Okay'. But when should it replace and when should it augment such text?
- Extensive use of text output might not be required, and may disrupt the use of SFUs. If R2D2 spoke, the purpose of his SFUs would become undermined. To what extent can SFUs and speech be mixed at all? Should NLG be able to generate non-natural language?
- Creating sounds as SFUs is a intensive design process which must match the speech synthesis. However it will also be affected by how intelligible the SFUs are and how frequently they are used. This would depend on higher level systems such as NLG which suggests you would need to design your NLG at the same time as designing the SFUs and the speech synthesis. This is a significant challenge.

3 DISCUSSION

There is still little consensus as to whether a robot should have a natural or unnatural sounding voice [5]. Aylett et al [1] argue instead that the robot should be seen as a performer and that a voice for performance is not constrained by standard assumptions of naturalness. This argument suggests we do not have to regard mimicry of a natural voice as the overall objective as it removes the advantage of *not being real*. SFUs could potentially be used to increase user engagement and powerfully characterise an artificial system. This raises a key question of how a system should control, insert, and mix SFUs with speech. Furthermore, how NLG, as the traditional controlling system for speech synthesis, should address this requirement.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No 780890 (Grassroot Wavelengths).

REFERENCES

[1] Matthew P Aylett, Benjamin R Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, alt08.

[2] Matthew P Aylett, Yolanda Vaquez-Alvarez, and Skaiste Butkute. 2020. Creating Robot Personality: Effects of Mixing Speech and Semantic Free Utterances. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[3] Randy Gomez, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura. 2018. Haru: Hardware design of an experimental tabletop robot assistant. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. ACM, 233–240.

[4] Eun-Sook Jee, Yong-Jeon Jeong, Chong Hui Kim, and Hisato Kobayashi. 2010. Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics* 3, 3 (2010), 199–206.

[5] Conor McGinn and Ilaria Torre. 2019. 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)*. IEEE, 211–221.

[6] Chris Wiltz. 2018. From Cozmo to Vector: How Anki Designs Robots With Emotional Intelligence. *Plastics Today* (2018).

[7] Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst. 2016. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction* 32, 1 (2016), 63–85.

[8] Cristina Zaga. 2017. Something in The Way It Moves and Beeps: Exploring Minimal Nonverbal Robot Behavior for Child-Robot Interaction. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 387–388.