Contextual Non-verbal Behaviour Generation for Humanoid Robot Using Text Sentiment

Amol Deshmukh[1], Alexandre Mazel[2] and Mary Ellen Foster[1]

Abstract— **This paper describes an approach to synthesise non-verbal behaviours for a humanoid robot Pepper using spoken text. Our approach takes into account the sentiment of the spoken text and maps the appropriate gesture and sound relevant to that text in a parameterised manner. This work forms a basis for our planned user study where we will evaluate this approach.**

I. Introduction

Expressiveness is the key attribute which helps humans to understand and accept robots better during interaction. Especially the interactions which are short-lived the urgency of producing the appropriate representation of the message to be conveyed is important. Such short-lived human-robot interactions are frequent in public spaces where the level of acoustic noise tends to be high enough to make it difficult to hear and understand speech. In these situations the social robot needs to provide service with efficacy. This approach could enable the failure of one modality for example speech that cannot be heard due to high noise to be compensated by the other modalities (e.g., gestures can be seen irrespective of acoustic noise). The approach presented in this work aims to synthesise the presentation of non-verbal behaviours from a social robot in a parameterised manner taking into account the sentiment of spoken text.

This work is being carried out in context of the MultiModal Mall Entertainment Robot (MuMMER) project, a four-year, EU-funded project with the overall goal of developing a humanoid robot, using SoftBank's Pepper as a platform, that can interact autonomously and naturally in the dynamic environments of a public shopping mall [1]. The overall idea underlying MuMMER is that for a robot to be successful in such a situation, it must be *entertaining* and *engaging*: that is, it must possess the social intelligence to both understand the needs and interactive behaviour of the users, as well as to produce appropriate behaviour in response. When the robot is able to support such smooth interactions, this should provide a sufficiently engaging experience that will stand up to repeated visits in a long-term deployment context. The approach proposed in this paper could endow humanoid robots with the ability to generate natural non-verbal behaviours enriched with social signals in context to the sentiment of the spoken text without the need of adding manual scripts into the dialogue.

II. Background

Many of the most popular social robots, such as SoftBank's Nao and Pepper have few or no moving parts in their faces, and are therefore not equipped to display facial expressions. Also, as mentioned earlier, often the acoustic context of an interaction can make spoken interaction problematic, particularly in noisy public spaces. Thus, the use of gestures, and other bodily displayed cues, plays a critical role in managing social human-robot interaction [2]. A number of previous studies have examined how various parameters can influence the users' reactions to the non-verbal behaviour of a virtually or physically embodied conversational agent, we discuss some in this section.

Purely emotional body expressions of a social robot such as raising the hands to show emotions such as joy, anger, or fear have been successfully used in a range of robot interaction contexts [3]. Salem et al. found that a robot is evaluated more positively when non-verbal behaviours, such as hand and arm gestures, are displayed along with speech, even if they do not semantically match the spoken utterance [4]. The model proposed by Amaya et al. [5], for example, transforms neutral animations into emotional animations by using "emotional transforms" which affect the speed and spatial amplitude of the animation. Yamaguchi et al. [6] defined a set of rules for modifying basic motions of a virtual character to express basic emotions, such as joy and sadness, and found that amplitude, position, and speed were the main parameters. The approach described by Kim et al. [7] explored how controlling the size, velocity, and frequency of robot gestures could affect user perception of the robot's personality. It was found that all of these factors had an effect on the perceived robot personality, and that this factor in turn affected users' subjective impressions of the robot.

The model developed by Pelachaud [8] for gesture expressiveness adopts six parameters, including spatial extent, temporal extent, fluidity, power, overall activation, and repetition. In perceptual tests, the six parameters were found to be recognisable and also combine to produce movements with different qualities. The work by Xu et al. [9] proposes a parametrized behaviour model with specific behaviour parameters for bodily mood expression, and applied the model to two concrete behaviours such as waving and pointing of the Nao robot. The most important parameters for creating readable mood expressions were found to be hand height and amplitude, head position, and motion speed [10].

In a work that is particularly relevant to our approach, the authors proposed a model for upper-body gestures of a

[1]University of Glasgow, School of Computing Science, Glasgow, UK amol.deshmukh@glasgow.ac.uk

[2]Innovation Department, Softbank Robotics Europe, Paris, France amazel@softbankrobotics.com

social robot [11] to add the ability to modulate functional gestures, such as pointing, to incorporate affective content. In their system, the speed and amplitude of a functional gesture are modified with the goal of projecting a particular affective impression, as expressed by valence and arousal. The work by Rodriguez et al. presented an approach for a gesticulation movements for a humanoid robot depending on sentiment processing taking into account simple head postures, voice parameters, and eye colours as expressiveness enhancing elements [12]. However, the authors did not get significant differences during their user study on adaptive gestures displayed by the robot, perhaps due to lack of interaction context.

While the previous studies listed above considered a range of gesture parameters, all included speed and amplitude parameters in some form. This is not surprising, as these are two dimensions that have been shown to be crucial for controlling gestures for artificial agents [13] and they are, indeed, the two dimensions that are considered in this work.

III. Approach

In our previous work on gestures the results indicated the role of personality as a mediation variable between gestures of different speed and amplitude of a robot [14], the occurrence of a similarity attraction effect for the majority of the observers involved in the experience [15], and the understandability of the gestures displayed [16]. Also changing the speed and amplitude of gestures was associated, to a statistically significant extend, with changes in the users' perception of those gestures [17]. Our current work aims at investigating how our previous findings may change when the gestures are accompanied by spoken speech and other non-verbal cues such as sounds in context of the sentiment attached to the spoken text.

A. Sentiment extraction

To extract the sentiment from the text we use the IBM Watson Tone Analyzer[1] service which uses linguistic analysis to detect emotional and language tones in text and produces a ToneCategory for example: anger, fear, joy, and sadness (emotional tones); analytical, confident, and tentative (language tones). In addition the service also returns a ToneScore in the range of 0.5 to 1. A score greater than 0.75 indicates a high likelihood that the tone is perceived in the content.

B. Sentiment mapping with gestures

We use this emotional tone to map the appropriate gesture in context of spoken text and the ToneScore as a parameter to generate the right intensity (amplitude and speed) for that particular gesture as follows. The variants for these gestures can be generated by mapping the values of the speed λ to the ToneScore. For each of the resulting gestures, another variation can be obtained by modifying the differences $\Delta_i(t) = \theta_i(t) - \theta_i(t-1)$, where $\theta_i(t)$ is the angle between the two mechanical elements connected by joint i at frame t. In particular, the values of the $\Delta_i(t)$ are multiplied, for all values of i and t, by a factor α the *amplitude* hereafter. The different values of α are adopted from the ToneScore of extracted text.

The approach is to use a set of around 100 human-designed animations, representing animation with neutral emotion. Some have specific meaning, like the 'me' animation when Pepper is pointing to his chest, or the 'you' animation when he's pointing in front of him, as usually human is facing the robot. Others animations which do not have a direct meaning, for instance a movement of arms with no context, we call these "empty animations". The animations are selected based on keywords found in the text, e.g., the word 'I' will be mapped to the 'me' animation. Though the length of each animation can be extracted directly from the movement keyframes (for instance one of the 'me' animation takes 1.35 second to be played), currently the length of each word as spoken by the TextToSpeech is unknown, so we use an heuristic to have a rough idea of the length of the sentences when it will be said by the robot, and we concatenate animation to fit this estimated length. We use in priority keywords recognised animation - like the 'me' from above - and fill unknown words with some random animation from a set of "empty animations". Thus we obtain a long animation corresponding to the full sentence, this animation is ready to be tuned to map sentence's emotion and well synchronised.

We then transform the animation on the fly by applying 3 modifiers: change of amplitude, speed ratio, and add position offset. The way we decided to apply modifiers have been humanly designed based on previous experiment (see [14], [15], [16], [17]) and classical animation rules from computer graphics animation, as they are summarised for instance on the Douglas Dooley website[2]. When the ToneCategory is "Joy", we will add some offset to Head and Hip pitch to show an open pose with head up and chest out. The amount of this change of joint angle will be related to the ToneScore. This simple approach have shown to add good significant results in the past [18]. We also speedup the animation based on the ToneScore, and amplify the movement.

On the opposite, in case a ToneCategory is interpreted as "sadness", the pose will be closed with a head pointing to the ground, a slowed down animation and limited amplitude movement, with the 3 modifiers manipulated in proportion to ToneScore. The same parameters are used for "anger" and "fear", but this time an extra modifier is added to produce some non-regularity in the timing of each movement key as suggested by the 'R' parameter in the DESIRE transfer model [19].

C. Sentiment mapping with sounds

We use the B.E.S.T non-verbal sounds tool kit, which has a library has 20 sounds for each category for emotion for example, *'anger', 'disgust', 'enjoyment', 'fear', 'interest', 'sadness', 'shame' or 'surprise'* [20]. Each sound for all

[1] https://www.ibm.com/watson/services/tone-analyzer/

[2] Douglas Dooley, Robot Appeal, *"Animation concepts I learned at Pixar that might be applied to robots."*, http://www.ezmicro.com/robot/index.html.

the categories are evaluated and rated on valence, arousal scores and organised accordingly to intensity level between 1-5, where 1 is low intensity and 5 is high intensity. Our approach is to map the ToneCategory to the emotion category of the sound and the intensity level with ToneScore from text sentiment analyser. The sounds library also has backchannel sounds for example, *'acknowledgement', 'agreement', 'disagreement', 'encourage', 'notsure' or 'askunderstood'*. The backchannel categories can be used in the following situations for example:

- 'acknowledgement': Robot acknowledges what user said/done (e.g., "Hmm")
- 'agreement': Robot agrees with what user said/done (e.g., "Yes!")
- 'disagreement': Robot disagrees with what user said/done (e.g., "No!")
- 'encourage': Robot encourages user to say/do more (e.g., "Go on!")
- 'notsure': Robot is not sure if it understood the user (e.g., "ahem?")
- 'askunderstood': Robot wants to ask if user understood it (e.g., "OK?")

We envisage to use these sounds along with spoken speech and gestures, for example, a sound that follows the sentence should modulate and explain the meaning of the sentence - whereas a sound that occurs before the sentence is then instead "explained" by the sentence. Overall, the sounds will be most useful if the aim is so provide a relatively subtle or "malleable" response to the human. E.g., if we combine an already highly positive sentence such as "Excellent, you are doing a great job!" with a positive sound, it is unlikely that the sound will be able to push the overall meaning of the sentence+sound combination much further into the positive. In that case, it also might not matter much if the sound comes before the sentence or after the sentence.

For back-channeling types of situations, sounds might be more helpful with respect to giving a neutral "Hmm" or "OK" a positive or negative touch. In addition, the sounds could help to make the robot more interesting, and draw more attention to the robot. From that angle, the approach is to place the sounds before the sentence. The rationale would be that the sound helps to return the subject's attention back to the robot as a social entity, as well as what it has to say. We plan to investigate this approach through a user study.

IV. Conclusion and future work

In this paper we described our approach to generate multimodal non-verbal behaviour for a humanoid robot Pepper. Our planned approach provides a systematic way to combine speech, gestures and sounds in context of sentiment attached to spoken text by the robot. Although this is still work in progress, in the future we want to implement our approach and carry out a user study with users to evaluate how users perceive the robot's behaviour during an interaction. We want to investigate which modality is the most effective in communicating the desired message and how they correlate with user's perception. The main hypothesis we would like to investigate in this user study is: Multi-modal behaviour from the robot i.e. combining speech, sound and gestures is most preferred by users during an interaction. The results from the user study will inform the design of the final social strategy for Pepper robot to be developed and deployed by MuMMer project in the shopping mall in Finland.

Acknowledgement

This research has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 688147 (MuMMER, `http://mummer-project.eu/`).

References

[1] M.E. Foster, R. Alami, O. Gestranius, O. Lemon, M. Niemelä, J.-M. Odobez, and A.K. Pandey. The MuMMER project: Engaging human-robot interaction in real-world public spaces. In *Proceedings of the Eighth International Conference on Social Robotics (ICSR 2016)*, November 2016.

[2] C. Breazeal, C.D. Kidd, A.L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 708–713. IEEE, 2005.

[3] M. Zecca, Y. Mizoguchi, K. Endo, F. Iida, Y. Kawabata, N. Endo, K. Itoh, and A. Takanishi. Whole body emotion expressions for kobian humanoid robot-preliminary experiments with different emotional patterns. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 381–386. IEEE, 2009.

[4] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[5] K. Amaya, A. Bruderlin, and T. Calvert. Emotion from motion. In *Graphics interface*, volume 96, pages 222–229, 1996.

[6] A. Yamaguchi, Y. Yano, S. Doki, and S. Okuma. A study of emotional motion description by motion modification and adjectival expressions. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, pages 1–6. IEEE, 2006.

[7] H. Kim, S.S. Kwak, and M. Kim. Personality design of sociable robots by control of gesture design factors. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 494–499. IEEE, 2008.

[8] C. Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009.

[9] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A. Neerincx. Bodily mood expression: Recognize moods from functional behaviors of humanoid robots. In *Proceedings of the 5th International Conference on Social Robotics - Volume 8239*, ICSR 2013, pages 511–520, Berlin, Heidelberg, 2013. Springer-Verlag.

[10] J. Xu, J. Broekens, K. Hindriks, and M.A. Neerincx. The relative importance and interrelations between behavior parameters for robots' mood expression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 558–563. IEEE, 2013.

[11] Greet Van de Perre, Hoang-Long Cao, Albert De Beir, Pablo Gómez Esteban, Dirk Lefeber, and Bram Vanderborght. Generic method for generating blended gestures and affective functional behaviors for social robots. *Autonomous Robots*, 42(3):569–580, Mar 2018.

[12] Igor Rodriguez, Adriano Manfré, Filippo Vella, Ignazio Infantino, and Elena Lazkano. Talking with sentiment: Adaptive expression generation behavior for social robots. In *Workshop of Physical Agents*, pages 209–223. Springer, 2018.

[13] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Proceedings of International Gesture Workshop*, pages 188–199, 2005.

[14] B.G.W. Craenen, A. Deshmukh, M.E. Foster, and A. Vinciarelli. Do we really like robots that match our personality? The case of Big-Five traits, Godspeed scores and robotic gestures. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[15] B.G.W. Craenen, A. Deshmukh, M.E. Foster, and A. Vinciarelli. Shaping gestures to shape personalities: The relationship between gesture parameters, attributed personality traits and Godspeed scores. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[16] A. Deshmukh, B.G.W. Craenen, M.E. Foster, and A. Vinciarelli. The more I understand it, the less I like it: The relationship between understandability and Godspeed scores for robotic gestures. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[17] Amol Deshmukh, Bart Craenen, Alessandro Vinciarelli, and Mary Ellen Foster. Shaping robot gestures to shape users' perception: The effect of amplitude and speed on godspeed ratings. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 293–300. ACM, 2018.

[18] A. Beck, A. Hiolle, A. Mazel, and L. Caamero. Interpretation of emotional body language displayed by robots. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, 2010.

[19] A. Lim, T. Ogata, and G. Okuno. Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music. In *EURASIP Journal on Audio, Speech, and Music Processing*, 2011.

[20] Arvid Kappas, Dennis Küster, Pasquale Dente, and Christina Basedow. Simply the best! creation and validation of the bremen emotional sounds toolkit. In *International Convention of Psychological Science*, 2015.