Synthesising and evaluating cross-modal emotional ambiguity in virtual agents

Matthew P. Aylett[12] and Blaise Potard[2]

[1] University of Edinburgh,
matthewa@inf.ed.ac.uk,
WWW home page: http://homepages.inf.ed.ac.uk/matthewa/
[2] CereProc Ltd. Edinburgh

Abstract. Emotional ambiguity, when more than one emotion appears present at a given time, or several emotions are superimposed, is common in human interaction and effects such as irony can be intentionally created through a mismatch of such emotional signals. High quality emotional speech synthesis offers a means for testing the effect of combining differences in vocal emotion, facial expression and text content in a virtual agent. In this paper we combine high quality emotional speech synthesis with a video rendered non-naturalistic virtual agent. Vocal emotion and text content combined to increase or decrease the emotional valence (positivity) of an utterance, while emotional facial expressions did not affect valence, but interacted with vocal emotion altering emotional activation in the lax and stressed vocal condition.

Keywords: speech synthesis, unit selection, expressive speech synthesis, emotion, prosody, facial animation

1 Introduction

In this work we address the challenges of synthesising and evaluating cross-modal emotional ambiguity in virtual agents by:

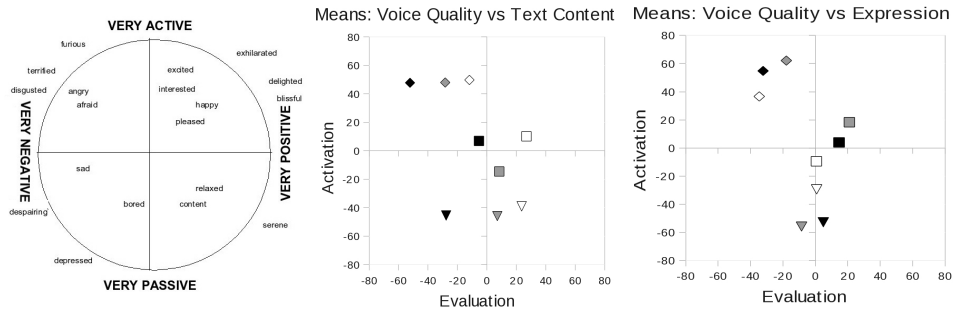
1. Evaluating utterances using a parametric *activation/evaluation* space[1–3] (Figure 1a). This allows the evaluation of magnitude across two dimensions, activation - how active or passive a subject rates an utterance, evaluation - how positive or negative a subject rates an utterance. The experiment was carried out online using 12 native English speakers.
2. Combining three modalities: Textual content, emotional speech synthesis based on stressed/lax voice quality changes[4–7] and synthesised angry/neutral/happy facial expressions using a non-naturalistic animated head[8](Figure 2), and evaluating the interactions between them.

Our research questions are as follows:

RQ1: Are negative and positive features of the three modalities additive in the evaluation domain?

RQ2: Does a mismatch of features across modalities produce an ambiguous emotion? If so can it be distinguished from a neutral rendition?

2 Results



a) b) c)

Fig. 1. a) Activation/Evaluation Space (radius 170). b) Mean activation/evaluation of materials by voice quality and text type. Voice quality significantly affected valence($F(2, 26)=17.93$, $p<0.001$) and activation($F(2, 26)=98.53$, $p<0.001$), Text type significantly affected valence only ($F(2, 26)=25.47$, $p<0.001$) c) Mean activation/evaluation of materials by voice quality and facial expression. Significant interaction between expression and voice quality ($F(4, 26)=4.02$, $p<0.005$). Diamond - Stressed VQ, Square - Neutral VQ, Triangle - Lax VQ. White - Positive Text/Happy Expression, Grey - Neutral Text/Neutral Expression, Black - Negative Text/Angry Expression.

Happy and Angry Facial Poses

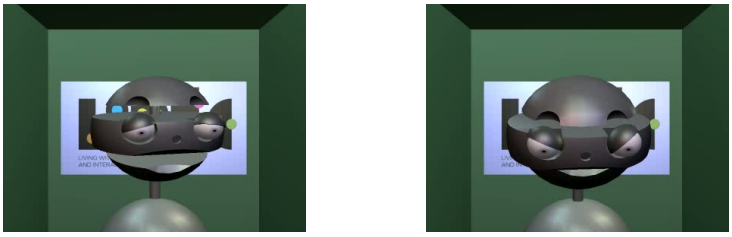


Fig. 2. *Virtual EMYS*

3 Discussion

Results show that we can merge speech style and text content to create a different perception of the emotion in a message wich is additive in the evaluation dimension. In contrast, facial expression had no significant effect on valence. Instead it interacted in a complex way with voice quality, significantly affecting activation when it mismatched the underlying voice quality.

A mismatch between text content and voice quality does make the emotion more ambiguous (closer to the centre point of the activation/evaluation space). In addition, the happy facial expression causes an increase in activation for the lax condition but a decrease for the stressed condition[3][4].

[3] Four videos are available at: *http://homepages.inf.ed.ac.uk/matthewa/iva2012emys/*

[4] This research was funded by the Royal Society through a Royal Society Industrial Fellowship.

References

1. Schlosberg, H.: A scale for judgement of facial expressions. Journal of Experimental Psychology **29** (1954) 497–510
2. Plutchik, R.: The Psychology and Biology of Emotion. Harper Collinns, New York (1994)
3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., M. Sawey, M., Schröder, M.: FEELTRACE: An instrument for recording perceived emotion in real time. In: ISCA Workshop on Speech and Emotion. (2000) 19–24
4. Gobl, C., Chasaide, A.N.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication **40** (2003) 189–212
5. Schröder, M., Grice, M.: Expressing vocal effort in concatenative synthesis. In: ICPhS. (2003) 2589–92
6. Aylett, M.P., Pidcock, C.J.: The cerevoice characterful speech synthesiser sdk. In: AISB. (2007) 174–8
7. Aylett, M., Pidcock, C.: UK patent GB2447263A: Adding and controlling emotion in synthesised speech. (2012)
8. Ribeiro, T., Paiva, A.: The illusion of robotic life: principles and practices of animation for robots. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. HRI '12, New York, NY, USA, ACM (2012) 383–390