Cross Modal Evaluation of High Quality Emotional Speech Synthesis with the Virtual Human Toolkit

Blaise Potard[1], Matthew P. Aylett[12], and David A. Baude[1]

[1] CereProc Ltd.,
blaise@cereproc.com,
www.cereproc.com
[2] University of Edinburgh

Abstract. Emotional expression is a key requirement for intelligent virtual agents. In order for an agent to produce dynamic spoken content speech synthesis is required. However, despite substantial work with pre-recorded prompts, very little work has explored the combined effect of high quality emotional speech synthesis and facial expression. In this paper we offer a baseline evaluation of the naturalness and emotional range available by combining the freely available SmartBody component of the Virtual Human Toolkit (VHTK) with CereVoice text to speech (TTS) system. Results echo previous work using pre-recorded prompts, the visual modality is dominant and the modalities do not interact. This allows the speech synthesis to add gradual changes to the perceived emotion both in terms of valence and activation. The naturalness reported is good, 3.54 on a 5 point MOS scale.

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

1 Introduction

Both human faces and human voices convey information about the speaker's emotional state. In order to develop artificial agents making use of both modalities, both emotional speech and emotional facial expression require synthesis. However, whereas a considerable about of previous work has examined these modalities in isolation (see [14] for a review), less work has examined their combined effect. Of this, work where an emotional voice has been used together with emotional facial synthesis has almost exclusively used pre-recorded prompts (e.g. SEMAINE's [1] sensitive artificial listener), and almost exclusively used matching voice and facial expression.

This is partly caused by the lack of available speech synthesis systems that can generate emotional variation. A large proportion of systems evaluated are research systems and not available for general academic use. Two exceptions are OpenMary TTS, a diphone based voice using MBROLA [15], and the commercial unit selection system CereVoice [3], which is freely available for academic research.

Integrating emotional speech synthesis with agent animation systems is also a challenging task requiring synchronisation of lip movements and audio and visual streaming. The Virtual Human Toolkit [9] (VHTK) is a collection of modules, tools, and libraries designed to aid and support researchers and developers with the creation of virtual human conversational characters. More specifically, VHTKs component SmartBody is a character animation library that provides synchronized locomotion, steering, object manipulation, lip syncing, gaze direction, and non-verbal behaviour in real-time.

SmartBody has been used in the current study as it contains all the necessary components for animating a highly realistic talking head. Although the CereVoice SDK is not normally distributed along with SmartBody, support for the CereVoice SDK is built-in, and the mapping from phonemes to viseme used for lip animation in VHTK is already present[3].

The resulting multi-modal system offers both state-of-the-art animation and speech synthesis at a commercial grade quality which can be used by the IVA community to explore multi-modal emotional interaction. In order to facilitate such work this paper presents baseline results which we believe will be invaluable for allowing further comparison and the investigation of the effect of interaction, and alternative graphic and audio renderings on perceived emotion.

We address these challenges of synthesising and evaluating cross-modal emotional ambiguity in virtual agents by: 1) Evaluating utterances using a parametric *activation/evaluation* space, 2) integrating the CereVoice synthesiser with the Virtual Human Toolkit, making the combined system readily available for researchers who wish to explore high quality dynamic emotional expression in their work. Our research questions are as follows:

RQ1: How do negative/positive and active/passive features of the two modalities combine? Are they independent? How much range do they offer?

RQ2: Does combining the emotional change across modalities impact naturalness in comparison to a high quality neutral baseline?

1.1 Positive/Negative Voice Quality Selection in Speech Synthesis

Voice quality is an important factor in the perception of emotion in speech [8]. A stressed (tense) voice quality is rated negatively in the evaluation space, while a lax (calm) voice quality is rated negatively in the activation space [5]. However, unlike speech rate and pitch which can be modified relatively easily using digital signal processing techniques such as PSOLA [18], modifying voice quality is more difficult, especially if it is important to retain naturalness. Rather than modifying speech to create the effect, an alternative approach is to record different voice qualities in sub-corpora and use them directly during concatenative synthesis.

[3] The mapping is currently only available for US accented voices but further accents will become available. Previous studies have shown phoneme-based lip animation is superior to viseme-based approaches [11]. Phone sequences including stress is available from the CereVoice system API and could be incorporated into later releases of VHTK.

This approach has been applied to diphone synthesis [16], however, CereVoice is the first commercial system to use pre-recorded voice quality sub-corpora in unit selection [4]. Previous work has examined the use of sub-corpora of specific emotions e.g. [10] where Happy, Angry and Neutral sub-corpora were used.

As with [16] three styles of voice quality are available: Neutral, the default for the recorded corpora, and two sub-corpora of lax (calm) and stressed (tense) voice quality. Adding XML tags in the speech input of the form:

```
<usel genre='stressed'>Text</usel>
<usel genre='calm'>Text</usel>
```

biases the selection of the units to come from the sub-corpora.

1.2 Evaluating Emotion in Synthesis

In order to evaluate mild changes in emotion and interactions between different modalities, an approach which is parametric rather than categorical is required. We therefore adopt the approach taken by FEELTRACE [7] and evaluate utterances within the *activation/evaluation space*.

FEELTRACE was developed specifically for assessing gradual changes in emotion by allowing subjects to place the emotion in a two dimensional space called the evaluation/activation space. This space is based on previous work in psychology [12, 13] and regards emotions as having two components, a valence which varies from negative to positive, and an activation which varies from passive to active (See Figure 1a). Therefore rather than asking subjects which emotion they perceive in an utterance, the subject chooses a point in this two dimensional space. There is active debate on how well such a space can represent emotional variation (see [2] for a review). Results presented here are not intended to be used to support or validate the model itself, rather the model is used purely pragmatically because of it powerful ability to detect *shifts* in emotion. This allows us to investigate the perceptual effect across modalities.

1.3 The Talking Head

As mentioned above, SmartBody has been used in the current study to generate realistic multi-modal animations. By default, SmartBody relies on the freely available Festival [17] speech synthesiser, but also support other synthesisers such as CereVoice.

Forcing VHTK to use the CereVoice SDK instead of the Festival synthesiser simply requires installing the SDK where SmartBody expects to find it. Some small modifications were performed in the source code of SmartBody to improve the robustness and ease of use of the integration, these modifications were transmitted to the SmartBody team.

We used one of the standard female character of the SmartBody library, Rachel, along with, for evaluation purposes, a custom build of CereProcs US female voice Isabella. 12 neutral sentences were selected for the evaluation. For reference, these sentences were recorded by the *Isabella* voice talent in the 3 voice

modalities (neutral, lax, tense), but in order to ensure realistic Text-to-Speech output, these natural recordings were omitted for the voice build. The recorded voice stimuli were used for a training phase during the evaluation.

The video stimuli were generated to simulate 3 different affects: *neutral*, *happy*, and *angry*. The *neutral* stimuli adopted the neutral stance from SmartBody. The *happy* stimuli were marked by a light smile, tightening of the eyes during speech, and by having the character smile markedly at the very end of the utterance. The angry stimuli were simulated by having the character frown markedly during the utterance.

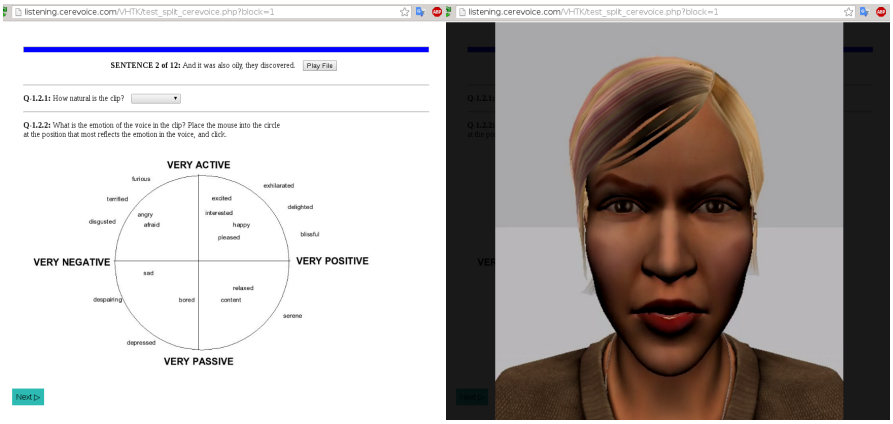
A SmartBody animation was generated that created a long video sequence of the Rachel character uttering the 12 selected sentences in various configurations. In total, the animation contained 108 clips (all 12 sentences in all possible combinations of video / audio modalities, i.e. $3 \times 3 \times 12 = 108$). Note that SmartBody can by default only output sequences of images, and the image output needs to be triggered manually from the user interface. For simplicity, we generated all clips from a single SmartBody script, then split the audio and set of image sequences accordingly. The image sequences were generated at a constant frame rate of 30 image per second, and the audio was generated at a sampling rate of 48kHz. The clips were then compressed into 2 different video format (MP4<h264>/ webm<VP8>) so as to be compatible with most HTML5 browsers (Google Chrome / Internet Explorer 9+ / Firefox / Safari). The audio embedded in the video was compressed in the AAC format at 128kbps.

2 Methodology

We asked subjects to rate the emotion in the synthetic speech by choosing a position in the activation/evaluation space (cf. circle in Figure 1a). We also asked them to rate naturalness on a 5 point scale (Bad/Poor/Fair/Good/Excellent). The experiment was carried out online (see Figure 1) using in total 13 English native speakers recruited through Crowdflower – a UK crowd sourcing evaluation service, similar to Amazon Mechanical Turk. Subjects were asked to use headphones, and to rate the *speech* present in the video clips. There were two factors in the experiment: Facial expression (happy, neutral, angry) **FACE**, and voice quality (Tense/Neutral/Lax) **VQ**.

The listeners were first trained in rating the audio stimuli by practising on short videos clips with the *neutral* facial expression but with audio stimuli of the *neutral*, *tense*, and *calm* voice modalities respectively. These clips were generated similarly to the evaluation sentences described above, except that the evaluation sentences had been recorded and retained during the voice building process, therefore despite being synthesised they were of comparable quality to pre-recorded prompts.

The training of the listeners was performed through a tutorial section in the evaluation process containing an example of each voice modality, with an interface and guidelines identical to the rest of the evaluation; the system would however not proceed until the positions in the activation / evaluation space



(a) Evaluation interface (b) while playing a video

Fig. 1: Online Experimental Setup

chosen by the user fall within an area consistent with the voice quality of each clip.

The evaluation was split into 3 parts: each set contained 3 variants of each of the 12 sentences (36 stimuli). Materials were presented in a randomised sequence to avoid order effects. Each listener was given the option to perform 1 to 3 parts of the evaluation, with the evaluation being terminated for each part as soon as we received responses from 13 listeners. In order to prevent the listeners from being tempted to cheat, the web-based evaluation ensured that each video was played at least once, and participants could only receive their payment using a key that was provided to them at the end of each part, handling some of the issues identified by [6].

Each part of the evaluation took roughly 15 minutes to perform. A majority of the listeners (8) did all 3 parts; some listeners attempted to use the same payment key for several parts of the evaluation and the parts with a wrong key were rejected. In order to maintain a balanced design this resulted in 10 subject responses for each of the 108 stimuli.

3 Results

A by-materials MANOVA analysis with two factors: facial expression **FACE** and voice quality **VQ**, across 3 dependent variables, Naturalness, Activation, Evaluation, were used to analyse the experimental results. Both factors had a significant multivariate effect (FACE: Wilk's Lambda 0.001, $F(6, 40) = 276.774$, $p < 0.001$, partial $\eta^2 = 0.976$), (VQ: Wilk's Lambda 0.235, $F(6, 40) = 7.080$, $p < 0.001$, partial $\eta^2 = 0.515$). Sphericity held for all dependent variables (Mauchly's Test of Sphericity). The interaction was not significant (partial $\eta^2 = 0.125$)

Univariate tests showed a significant effect of FACE and VQ on valence (FACE: $F(2, 22) = 720.390$, $p < 0.001$, partial $\eta^2 = 0.987$, VQ: $F(2, 22) = 25.906$, $p < 0.001$, partial $\eta^2 = 0.702$) and on activation (FACE: $F(2, 22) = 274.699$, $p < 0.001$, partial $\eta^2 = 0.961$, VQ: $F(2, 22) = 5.238$, $p < 0.025$, partial $\eta^2 = 0.323$). For naturalness only **FACE** had a significant effect (FACE: $F(2, 22) = 24.840$, $p < 0.001$, partial $\eta^2 = 0.693$).

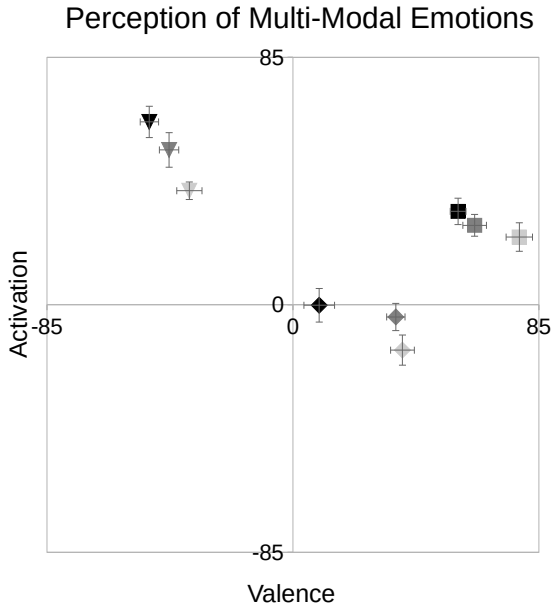


Fig. 2: Mean activation/evaluation of materials by facial expression (FACE) indicated by shape of point, and voice quality (VQ) indicated by shade. Triangle - Angry FACE, Diamond - Neutral FACE, Square - Happy FACE. Black - Tense VQ, Dark Grey - Neutral VQ, Light Grey - Lax VQ. The original activation/evaluation circle shown in Figure 1a is radius 170. All means show moderate variation within this space (< 80). FACE is dominant but VQ also significantly alters the perception of emotion. Error bars show ± 1 standard error.

Figure 2 shows the means by **VQ** and **FACE**. Posthoc pairwise comparisons (Least Significant Difference - LSD), showed all means significantly ($p < 0.005$) different for FACE Valence: Angry $<$ Neutral $<$ Happy, and FACE Activation: Angry $>$ Happy $>$ Neutral. For VQ Valence: Tense $<$ Neutral $<$ Lax, and VQ Activation: Tense $>$ Neutral and Lax.

The mean naturalness overall was 3.54 with a standard deviation of 0.30. This is in line with unit selection results although lower than neutral speech as a unimodal stimuli [5]. The animation of the avatar appears to dominate the

impression of naturalness. In previous audio only experiments, both lax and tense voice qualities led to a small but significant decrease of perceived naturalness [5]. We did not observe this effect, but instead in a post-hoc test a significant difference between the angry face and both the neutral and happy face (LSD $p < 0.001$, Angry mean 3.41 ± 1 SE 0.06, Neutral mean 3.59 ± 1 SE 0.06, Happy mean 3.60 ± 1 SE 0.07).

Nine example videos are available at:
`https://dl.dropboxusercontent.com/u/1618087/rachel_iva_2016.zip`

4 Discussion

The results show that using emotional variation in facial expression and speech can both be synthesised to affect perception of emotion. Compared to previous results with speech only [5] where emotional synthesised speech caused a drop in naturalness, the animated head appears to mitigate this effect with focus moving to the naturalness of animated facial expression.

However, this experiment was not interactive. It is within an application environment, where emotion is key to personifying a character and conveying their underlying motivations, that these baseline results need to be compared. As well as supporting each other cross modally, the ability to mismatch the emotion conveyed by the speech and facial expression could possibly be used to synthesise a sense of irony or underlying tension in the virtual agent which could be useful for games and tutoring applications. It is, however, unclear how such a sophisticated use of emotion might be evaluated in such a context.

In addition to voice quality, the CereVoice system also allows manipulation of pitch, amplitude and speech rate which could be used to support and also alter the perceived effect of the speech. However, altering the synthesised speech over time using all these factors is non trivial if naturalness is to be maintained. As Schröder points out, *"In a dialogue, an emotional state may build up rather gradually, and may change over time as the interaction moves on."* [15, p. 211]. Thus we have a time element, as well as a vocal element, that need to be coordinated to create a successful effect.

5 Conclusions

This is the first study we are aware of that has investigated how high quality commercial expressive speech synthesis interacts with emotional facial expressions. Previous work considered less natural systems (formant, or diphone speech synthesis systems) or used pre-recorded prompts.

We have shown that voice quality, facial expression combine relatively independently to create different perceptions of emotion.

6 Acknowledgements

This research was funded by the Royal Society through a Royal Society Industrial Fellowship and and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 (ARIA-VALUSPA).

References

1. The semaine project. http://www.semaine-project.eu/
2. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review 43(2), 155–177 (2015)
3. Aylett, M.P., Pidcock, C.J.: The cerevoice characterful speech synthesiser sdk. In: AISB. pp. 174–178 (2007)
4. Aylett, M.P., Pidcock, C.J.: UK patent GB2447263A: Adding and controlling emotion in synthesised speech (2012)
5. Aylett, M.P., Potard, B., Pidcock, C.J.: Expressive speech synthesis: Synthesising ambiguity. In: SSW8. pp. 133–138. Barcelona, Spain (August 2013)
6. Buchholz, S., Latorre, J.: Crowdsourcing preference tests, and how to detect cheating. In: Proc. Interspeech. pp. 3053–3056 (2011)
7. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: An instrument for recording perceived emotion in real time. In: ITRW on speech and emotion. pp. 19–24 (2000)
8. Gobl, C., Chasaide, A.N., et al.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40(1), 189–212 (Apr 2003)
9. Hartholt, A., Traum, D., Marsella, S.C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.P., Gratch, J.: All Together Now: Introducing the Virtual Human Toolkit. In: IVA. Edinburgh, UK (Aug 2013)
10. Hofer, G.O., Richmond, K., Clark, R.A.: Informed blending of databases for emotional speech synthesis. In: Proc. Interspeech (2005)
11. Mattheyses, W., Latacz, L., Verhelst, W.: Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. Speech Communication 55(7), 857–876 (2013)
12. Plutchik, R.: The Psychology and Biology of Emotion. Harper Collins College Publishers, New York (1994)
13. Schlosberg, H.: A scale for the judgement of facial expressions. Journal of Experimental Psychology 29(6), 497–510 (1941)
14. Schröder, M.: Emotional speech synthesis: A review. In: Proceedings Eurospeech 01. pp. 561–564 (2001)
15. Schröder, M.: Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: Proceedings Workshop on Affective Dialogue Systems. pp. 209–220. Springer (2004)
16. Schröder, M., Grice, M.: Expressing vocal effort in concatenative synthesis. In: Proc. 15th international conference of phonetic sciences. pp. 2589–2592 (2003)
17. Taylor, P.A., Black, A., Caley, R.: The architecture of the festival speech synthesis system. In: SSW3. pp. 147–151. Jenolan Caves, Australia (1998)
18. Valbret, H., Moulines, E., Tubach, J.P.: Voice transformation using psola technique. In: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. vol. 1, pp. 145–148. IEEE (1992)