Real-Time Visual Prosody for Interactive Virtual Agents

Herwin van Welbergen[1,2](✉), Yu Ding[2], Kai Sattler[1,3], Catherine Pelachaud[2], and Stefan Kopp[1]

[1] Social Cognitive Systems Group, CITEC, Faculty of Technology, Bielefeld University, Bielefeld, Germany
[2] CNRS-LTCI, Télécom-ParisTech, Paris, France
[3] Department of Psychology, University of Bamberg, Bamberg, Germany
hvanwelbergen@techfak.uni-bielefeld.de

Abstract. Speakers accompany their speech with incessant, subtle head movements. It is important to implement such "visual prosody" in virtual agents, not only to make their behavior more natural, but also because it has been shown to help listeners understand speech. We contribute a visual prosody model for *interactive* virtual agents that shall be capable of having live, non-scripted interactions with humans and thus have to use Text-To-Speech rather than recorded speech. We present our method for creating visual prosody online from continuous TTS output, and we report results from three crowdsourcing experiments carried out to see if and to what extent it can help in enhancing the interaction experience with an agent.

Keywords: Visual prosody · Nonverbal behavior · Realtime animation · Interactive agents

1 Introduction

Our heads move almost incessantly during speaking. These movements are related to e.g. the neurological and biomechanical coupling between head and jaw, the prosodic structure of speech, the content of speech and pragmatics such as turn-taking [10]. In this paper we focus on synthesizing and evaluating *visual prosody*: the movements of the head related to the prosodic structure of the accompanying speech [8]. It is important to endow Intelligent Virtual Agents (IVAs) with visual prosody, not only to make their movement and behavior more human-like, but also because visual prosody has been shown to help listeners in understanding speech [16].

Recently, many approaches to motion synthesis based on speech prosody have been proposed (see Sect. 2 for an overview). However, to the best of our knowledge, none of these approaches deal with *interactive* IVAs in live interactions with humans. Going beyond simple playback of prerecorded audio scripts, *interactive* IVAs have to rely on flexible Text-To-Speech (TTS) output. In such

W.-P. Brinkman et al. (Eds.): IVA 2015, LNAI 9238, pp. 139–151, 2015.
DOI: 10.1007/978-3-319-21996-7_16

agents, visual prosody must be generated in *real-time* from running synthetic speech. That is, the preparation of a motion segment (including the generation of TTS for it) should take less time than playing back the motion segment. Beyond that, we aim to use visual prosody in an incremental behavior realizer [17]. In such a realizer, utterances are generated in chunks much smaller than a full sentence, e.g. in a phrase. A visual prosody module that is of any use in such realization scenarios should work in an *online* fashion. That is, it should be able to deal with motion synthesis using only the current and previous phrases and make use of only little (if any) look-ahead.

In this paper, we present the first online TTS-based system for visual prosody. After discussing related work, we present our approach to creating online visual prosody for a behavior realizer. Finally, we report results from three crowdsourcing experiments carried out to measure if and to what extent this model can help in enhancing the interaction experience with an IVA. In particular, we compare our approach (1) to not using speech related head motion at all (as is common practise in most behavior realizers), (2) using motion captured head motion that is unrelated to the spoken content, and 3) feeding TTS to a state-of-the-art offline visual prosody model [5].

2 Related Work

The generation of speech-accompanying head movements has been tackled in several projects before. Lee and colleagues (see [12] for a recent overview of their work) have provided a nonverbal behavior generator (NVBG) that generates head and other motion on the basis of speech. Their work is complementary to ours: the NVBG generates motion based on speech content and pragmatics rather than on speech prosody. Other computer animation systems provide head motion on the basis of speech prosody. Typically these systems work offline (e.g. in [3–5,15]): they take a spoken sentence and generate head motion that is fluent and at the same time fits to the prosodic structure (e.g. speech pitch (f0), energy, or syllable boundaries) of the sentence. To capture the temporal evolution of sequential data (here, head rotations) these systems typically make use of a Hidden Markov Models (HMMs) as visual prosody models [3,15]. However, such visual prosody models suffer from the limitations of HMM independence assumptions.[1] This limitation can be attenuated by a variant of HMM, called parameterized (contextual) HMM, where state emission probabilities and state transition probabilities are defined by contextual parameters (e.g. prosody features) at each time step. Ding et al. [5] proposed a fully parameterized (contextual) HMM as visual prosody model, which not only embeds the advantage of HMMs but also overcomes the limitations of classical HMMs.

Others have worked on 'live' synthesis of visual prosody, where head motion is synthesized directly on the basis of microphone input. Levine et al. [14] generate

[1] The state at time t is independent of all the previous states given the state at time $t-1$; the observation at time t is assumed independent of all other observations and all states given the state at time t.

a live stream of gesture (including head movement) from speech f0 and energy, summarized per syllable. Gesture synthesis is achieved using a HMM that selects the right gesture phase to perform at each syllable. In later work, rather than modeling the mapping between motion and speech prosody directly, Levine and colleagues [13] provide a two-layered model, which outperformes [14] in perceived realism. This model uses an inference layer (using a Conditional Random Field) that models the relationship between prosodic features and more abstract motion features (such as temporal and spatial extend, velocity, curvature). At synthesis time, a control layer selects the most appropiate gesture segment using a pre-computed optimal control policy that aims to minimize the difference between desired and selected gesture features while maximizing the animation quality. Le et al. [11] have implemented a live visual prosody system that synthesizes head, eye and eyelid motion on the basis of speech loudness and f0. The head motion synthesis of this model is implemented as a frame-by-frame selection of head posture on the basis of the prosodic features and the head posture on the previous two frames. It uses a Gaussian Mixture Model (GMM) to maximize the combined probability density of the head posture, velocity and acceleration with the prosodic features. This system outperforms some of the other online and offline systems ([3,4,14]) discussed in this section in terms of preference ratings by subjects. To the best of our knowledge, none of these live visual prosody systems have been tested with TTS rather than real human speech.

3 Online TTS-Based Visual Prosody

As live synthesis fits with our goal of implementing online visual prosody in incremental behavior realization scenarios, we decided to implement our online TTS-based visual prosody model on the basis of a live visual prosody model. Since we are currently exploring the feasibility of TTS-based visual prosody, we opted to implement a modified version of [11] and test it with TTS input, as their system is easy to implement, yet provides synthesis results that are beyond the quality of several existing visual prosody models.

Le et al.'s visual prosody model works as follows. Its speech features are f (f0) and l (loudness). Its motion features $\kappa \in \{r, p, y, v, a\}$ include the euler angles r_t (roll), p_t (pitch) and y_t (yaw) of the head at frame t, and the head velocity v_t and acceleration a_t, defined as:

$$v_t = \left\| \begin{bmatrix} r_t \\ p_t \\ y_t \end{bmatrix} - \begin{bmatrix} r_{t-1} \\ p_{t-1} \\ y_{t-1} \end{bmatrix} \right\|, a_t = \left\| \begin{bmatrix} r_t \\ p_t \\ y_t \end{bmatrix} - 2 \begin{bmatrix} r_{t-1} \\ p_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} r_{t-2} \\ p_{t-2} \\ y_{t-2} \end{bmatrix} \right\| \quad (1)$$

It makes use of five GMMs, one for each motion feature, modeling the joint probability density of that motion feature with the two speech features:

$$P(\boldsymbol{X}) = \sum_{i=1}^{m} c_i \frac{1}{\sqrt{(2\pi)^3 |\boldsymbol{\Sigma_i}|}} e^{-\frac{1}{2}(\boldsymbol{X}-\mu_i)^T \Sigma_i^{-1} (\boldsymbol{X}-\mu_i)} \quad (2)$$

Here $\boldsymbol{X} = (\kappa, f, l)^T$, m is the number of mixtures, c_i, μ_i and Σ_i are the weight, mean and covariance matrix of the i-th mixture respectively. The GMMs are trained using the Expectation-Maximation algorithm. At synthesis, the next head pose is then found using:

$$(r_t^*, p_t^*, y_t^*) = \arg\max_{r_t,p_t,y_t} \prod_{\kappa_t \in \{r_t,p_t,y_t,v_t,a_t\}} P(\kappa_t, f_t, l_t) \tag{3}$$

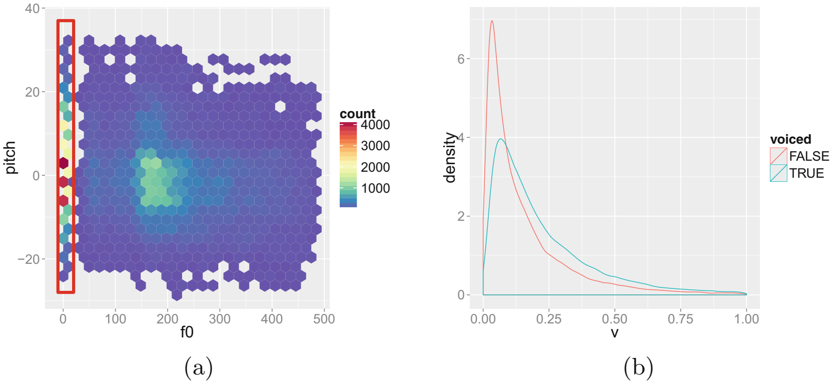
Given the head poses from the previous two frames $(r_{t-1}^*, p_{t-1}^*, y_{t-1}^*)^T$ and $(r_{t-2}^*, p_{t-2}^*, y_{t-2}^*)^T$, this simplifies to:

$$\begin{aligned}(r_t^*, p_t^*, y_t^*) = \arg\max_{r_t,p_t,y_t} \prod_{\kappa_t \in \{r_t,p_t,y_t\}} P(\kappa_t, f_t, l_t)\times \\ P\left(\left\|\begin{bmatrix} r_t \\ p_t \\ y_t \end{bmatrix} - \begin{bmatrix} r_{t-1}^* \\ p_{t-1}^* \\ y_{t-1}^* \end{bmatrix}\right\|, f_t, l_t\right)\times \\ P\left(\left\|\begin{bmatrix} r_t \\ p_t \\ y_t \end{bmatrix} - 2\begin{bmatrix} r_{t-1}^* \\ p_{t-1}^* \\ y_{t-1}^* \end{bmatrix} + \begin{bmatrix} r_{t-2}^* \\ p_{t-2}^* \\ y_{t-2}^* \end{bmatrix}\right\|, f_t, l_t\right)\end{aligned} \tag{4}$$

Thus, the method favours poses and velocities that are likely in combination with the prosodic features, and additionally, by taking into account the previous two frames of animation and the likelihood density of velocities and accelerations, smooth motion trajectories.

We provide several modifications to this model. Le et al. use a customly recorded motion capture corpus in which an actor is asked to read from a phone balanced corpus for 47 min. As neither this corpus nor their trained models are publicly available, we opted to train the model from the IEMocap corpus [2] instead. The corpus contains dialogs between two actors in 8 improvised and 7 scripted scenarios. In each dialog, the movement of one of the actors is recorded using motion capture. Each scenario is recorded for five male and five female actors. We performed Canonical Correlation Analysis (CCA) on the head motion and speech, which revealed significant differences for the synchrony of head motion and speech, both between different actors and between their scripted and improvized sessions. Based on this observation we decided to train the model based on the speech and motion of one female actor (actor F1) in her scripted scenarios. We choose to train on the scripted scenarios, as they have a higher CCA coeficient and we hypothised that they therefore may have less head motions that are unrelated to speech prosody. In total, the training set contained 7.4 min of speech and head motion.

Similarly to [11], we use openSMILE [6] to extract audio features. However, rather than using loudness and f0, we opted to use RMS energy and f0, as we found that RMS energy correlates more with the head postures and is easier to obtain directly from TTS software. As we have less training data available, we decided to use a higher sample rate (120 Hz instead of 24 Hz as used in [11]) to obtain more training samples. Figure 1a shows a histogram of the head pitch and f0 in the training data. In 31 % of the samples in the training data (the red box in Fig. 1a), the f0 is 0, that is, there is either a silence or the speech is not voiced. As it turned out to be difficult to fit a GMM to this speech pattern we decided to learn the GMM only on the voiced parts of the speech, which left us with 5.1 min of training data. As we observed that the head velocity is low during unvoiced speech (Fig. 1b), we decided to keep the head still whenever $f0$ is 0 during synthesis. Based on cross validation results, we set the number of GMMs to 11 in all mixture models ([11] uses 10). Rather than using gradient descent search to find the optimum of Eq. 4, we opted to use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for optimization (we make use of the algorithm as provided by WEKA [9]) as it converges faster than gradient descent and works with different (TTS) voices without requiring manual tweaking of the optimization parameters for each voice.



(a)

(b)

Fig. 1. Left: Histogram of head pitch and f0; right: probability density of head velocity during voiced and unvoiced speech.

4 Evaluation

To measure if and to what extent implementing online visual prosody in a behavior realizer can help in enhancing the interaction experience, we conducted three experiments. The first two experiments explore whether an IVA that exhibits visual prosody is perceived as warmer, more competent and/or more human-like. In the third experiment, we check whether subjects perceive the online visual prosody as fitting to the accompanying speech.

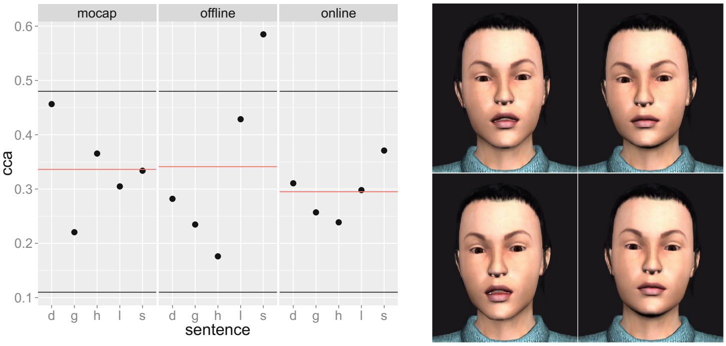
4.1 Experimental Conditions

We use four experimental conditions to measure the performance of online visual prosody. In the *none* condition no visual prosody is used, which is the common practise in current behavior realizers. In the *mocap* condition the head movement of the IVA is steered by motion capture motion from a speaker in the corpus. That is, the IVA replays a real human speaking motion, but it is not produced in concordance with the accompanied TTS output. Motion capture segments that may represent visual prosody are selected from the IEMOCAP corpus [2] using the following criteria:

1. The head should, on average, face in the direction of the interlocutor. This corresponds with motion segments that have their mean pitch, yaw and roll in the ranges $< -4, 4 >$, $< -5, 5 >$ and $< -6, 6 >$ degrees respectively.
2. Extreme head poses are to be avoided, the head pitch, yaw and roll should be in the ranges $< -8, 8 >$, $< -15, 15 >$ and $< -12, 12 >$ degrees respectively.
3. Extreme rotational velocities (greater than $240°/s$) and accelerations (greater than $60°/s^2$) are to be avoided.

These criteria are meant to prevent excessive movements that too obviously do not correspond to TTS. The numerical values were obtained by manual inspection of the histograms of head rotations in the corpus.

In the *offline* condition the head movement of the IVA is steered by a state-of-the-art offline visual prosody model [5]. This model synthesizes head and eyebrow motions on the basis of the f0 and RMS energy of speech. In the experiments we make use of only the head motion generated by this model. This is a challenging baseline as it can take into account more information when synthesizing head motion than the online model. In the *online* condition the head movement of the IVA is steered by the online visual prosody model (see Sect. 3).



(a) (b)

Fig. 2. Left: CCA-coefficients of the stimuli; right: prototypical head movements.

4.2 Stimuli

For all stimuli, behavior is steered using AsapRealizer [17], with the ‘Armandia’ virtual character (see Fig. 2b) and the ‘Hannah’ US-American voice frome

CereProc.[2] We introduce Armandia to the subjects as a virtual assistant that helps manage their appointments. We created four appointments for Armandia to introduce to the subjects (see Fig. 3 for an example). For each appointment, we created videos for all four conditions. The eye gaze of the IVA is directed at the camera in all interaction segments. Videos for the stimuli in all conditions are provided at http://www.herwinvanwelbergen.nl/visual_prosody.

To evaluate the speech-gesture synchrony of our stimuli we calculate the CCA-coefficient between the euler angles of the head and the f0 of speech (see Fig. 2a). This, or similar linear correlations have been used to demonstrate the synchrony between head posture and speech prosody in related work (e.g. [3,16]) and the CCA-coefficient is proposed as an objective measurement for visual prosody by Mariooryad and Busso [15]. The mean CCA-coefficients (indicated with the red lines) are below the CCA-coefficients found in real speech (top black line), but well above chance level (lower black line). Interestingly, the CCA-coefficients are also high in the mocap condition. We hypothesize that this is because both speech and gesture are rhythmic signals and some linear correlation can always be found between two such signals.

All studies are conducted online. Subjects are recruited using the CrowdFlower crowdsourcing platform[3] and got paid for participating. Participant recruition was limited to English speaking countries. At the start of each of the experiments, subjects are shown a video of Armandia reading a login code to them. Subjects have to enter this code to proceed to the rest of the study. This video serves both to let the subject get used to Text-To-Speech and our IVA and to make sure that they understand what is being said (e.g. their audio is enabled and at a high enough level, they understand some English). We introduced several mechanisms (discussed in detail in each experiment) to filter out subjects that provided nonsense answers to minimize their time spend on the experiment and maximize their profit. Subjects were given the option to provide free-text feedback after each experiment. At the end of each study, subjects were debriefed on its purpose.

4.3 Evaluating Warmth, Competence and Humanlikeness

Our experimental design (including the questionnaires) to measure warmth, competence and human-likeness is based upon the design used in succesful laboratory studies by Bergmann and colleagues (e.g. in [1]), in which these factors are compared for several gesturing strategies (including not gesturing at all).

Subjects are instructed that they are to evaluate a virtual assistant that helps manage their appointments. We use a between-subject design: each subject is shown videos of Armandia in one condition. In this experiment, the login code is read using that condition. After logging in, subjects are shown four videos of Armandia discussing an appointment in one or two sentences. Each video is followed by two-choice comprehension questions (see Fig. 3 for an example).

[2] https://www.cereproc.com/.

[3] http://www.crowdflower.com/.

Appointment:	Your plane to Hawaii leaves on Saturday at 10 am, so you should take the train at 7:10, what do you think about that?
Question:	What leaves at 10 am?
Possible answers:	Train, Plane

Fig. 3. Example appointment and comprehension question.

The comprehension questions are used to make sure subjects pay attention to the videos and filter out those that did not understand what was said in them. After watching the four videos subjects are asked to rate how well 18 adjectives (see Table 1) fit Armandia's behaviour on a 7-point Likert scale ranging from not appropiate to very appropiate. The 18 adjectives are intertwined with three test adjectives that have a more or less clear answer (we used 'blond', 'dark-haired', 'english-speaking'). We used a pilot study with 10 subjects from our laboratories to select a set of test adjectives that is best understood by the subjects and to establish baselines for correctly answered comprehension questions.

Results. In total 260 subjects participated in the study, 232 of these finished the questionnaire. We filtered out subjects that did not watch all videos (6), did not rate the test adjectives correctly (48) or could not answer more than 6 out of 8 of the comprehension questions correctly (7). This left us with 171 participants (101 female, 70 male, aged between 18 and 71; $M = 39.7, SD = 12.2$).

To measure the reliability of our warmth, competence and human-likeness factors, we calculated Cronbach's α. All α valuses were above 0.7, which justifies combining these items into one mean value as a single index for this scale (see Table 1).

Table 1. Reliability analysis for the three factors.

Factor	Items	Cronbach's α
Warmth	pleasant, sensitive, friendly, likeable, affable, approachable, sociable	.927
Competence	dedicated, trustworthy, thorough, helpful, intelligent, organized, expert	.925
Human-likeness	active, humanlike, fun-loving, lively	.846

We conducted a one-factorial ANOVA and found no significant difference between the conditions in warmth ($F(3, 167) = .284$, $p = .837$), competence ($F(3, 167) = 1.095$, $p = .889$) nor human-likeness ($F(3, 167) = .722$, $p = .828$). Figure 4 (right) shows the distribution of the factors in each condition.

To check the consistency of the ratings on the questionnaire we conducted principal component analysis (PCA) with orthogonal rotation (varimax) on the 14 questionnaire elements relating to warmth and competence. We selected only these factors, as they are consistently found as universal dimensions of social judgement [7], while human-likeness is added as a more experimental factor, which is not necesaraly orthogonal to warmth and competence in [1]. Two components had eigenvalues of over Kaiser's criterion of 1 and in combination explained 70.8 % of the variance. Figure 4 (left side) shows the factor loading after rotation and the mean and standard deviation of each factor. The items that cluster on the same components suggest that one corresponds to competence and the other to warmth and that each item clusters to its expected component.

item	**warmth**	**competence**
pleasant	**.705**	**.406**
sensitive	**.763**	.082
friendly	**.802**	.322
likeable	**.765**	**.429**
affable	**.732**	.322
approachable	**.754**	**.433**
sociable	**.824**	.300
dedicated	**.493**	**.625**
trustworthy	**.407**	**.760**
thorough	.222	**.858**
helpful	.196	**.796**
intelligent	**.521**	**.680**
organized	.238	**.845**
expert	.394	**.699**

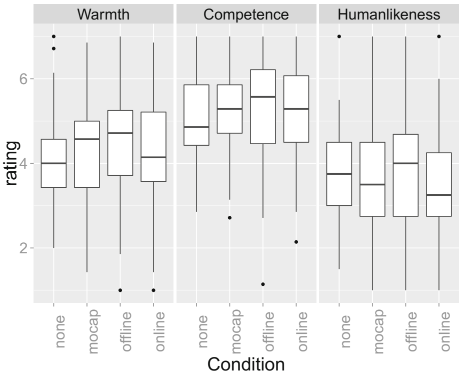


Fig. 4. Results of the evaluation of warmth, competence and human-likeness.

Over 42 % (72) of the subjects used the possibility to provide free-text feedback. Most of the comments (42, in all conditions) were on the quality of the TTS. Several subjects (11, in all conditions) commented that aspects of movement were missing (e.g. emotion, smiles, blinking, gaze). Only one of the subjects commented on the head motion.

4.4 Warmth, Competence and Humanlikeness Revisited

We hypothized that because of the comprehension task, many subjects may have been too focused on understanding the speech to notice the head movement. Therefore we ran a second experiment, in which we removed all comprehension questions. Furthermore, we added a question at the end on which motions were perceived by the subjects (head, lips, blinks, breathing).

Results. After filtering out subjects in the same way as for experiment 1, 176 subject remained. Of these we focus our analysis on the 142 (77 female, 65 male, aged between 20 and 68; $M = 37.8$, $SD = 10.7$) that either correctly perceived head movement when the head moved in their condition, or reported no head movement when they were assigned the none condition (analysis on the aforementioned 176 shows similar results). Of these 142, 53 subjects have participated in the first experiment. This might bias the results in favour of our models, as it creates a partial within-subject condition for some of the participants.

As in the previous experiment, we calculate Cronbach's α on the factors, and again combining them into a one mean value as a single index for each factor was justified. We conducted a one-factorial ANOVA and found no significant difference between the conditions in warmth ($F(3, 138) = .733$, $p = .534$), competence ($F(3, 138) = .923$, $p = .432$) nor human-likeness ($F(3, 138) = .919$, $p = .433$). The means and standard deviations of the factors are almost identical to those of the first experiment.

4.5 Evaluating the Match Between Speech and Head-Motion

To evaluate how well the head motion generated by our model is perceived to fit to the speech, we asked subjects to order the videos of the different conditions for all four appointment sentences: participants were instructed to give each of the four videos a unique ranking number (1st, 2nd, 3rd, 4th), but we did not enforce this in the user interface of the study. This allowed us to filter out participants that did not bother to read the instructions and (arguably) might not give very serious rankings. In this experiment, the login code was read in the no-motion condition.

Results. In total 125 subjects participated in the study. Of these, 95 completed the study. We filtered out subjects that did not watch the videos completely (8), did not provide a unique ranking for each video (27), participated in the previous experiments (7) or reported mistakes in filling out the ranking (1). This left us with 52 participants (30 female, 22 male, aged from 17 to 64; $M = 38.06$, $SD = 11.95$) for the analysis of our results.

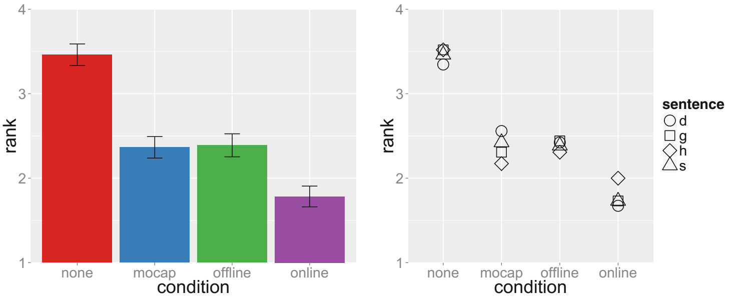


Fig. 5. Left: overview of the rankings, whiskers indicate the 95 % confidence intervals; right: spread of the rankings per appointment sentence.

We performed a repeated measures ANOVA to compare the ranking of the conditions and their interaction with the appointments. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of condition, $\chi^2(5) = 12.257$, $p = .031$ and the interaction between appointment and condition $\chi^2(44) = 62.156$, $p = .038$, therefore the degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\epsilon = .929$, $\epsilon = .934$ respectively). The results show a significant main effect between the conditions $F(2.788, 339.144) = 45.919$, $p < 0.001$, $\eta_p^2 = .474$. Post hoc tests (using Bonferroni adjustment for multiple comparisons) show that the no-motion condition is always ranked significantly lower as all other conditions ($p < 0.001$), online visual prosody is always rated significantly higher as all other conditions ($p < 0.002$) and offline and mocap are inbetween and their ratings do not significantly differ ($p = 1.00$). No interaction between appointment and condition was found ($F(8.407, 428.741) = 1.207$, $p = .291$). An overview of the rankings for each condition and their spread over different appointment sentences is given in Fig. 5.

5 Discussion

The evaluation showed that the online visual prosody model can provide head motion that is perceived to fit better to TTS than using a state-of-the art offline method for visual prosody with TTS, using motion capture from a different speech segment, or using no motion at all. Surprisingly, motion synthesized with the offline visual prosody model was not perceived as fitting better to speech than motion captured motion that is unrelated to the speech. It could well be that generating motion that is perceived to fit to TTS requires different motion qualities (e.g. being more robotic) than generating motion that fits to real human speech. Recall that the offline model is a more intricate model than the relatively simple online model and might capture aspects of human speech (for example prominence) that are not available in TTS. The online model might thus outperform the offline model with TTS-speech because it is more robust in generating head motion that is coherent to speech when some human-like qualities of speech are missing.

We did not find any effect of visual prosody on perceived warmth, competence or human-likeness. There could be several reasons for this: (1) visual prosody might not affect perceived competence, warmth, nor human-likeness, (2) the effects are relatively small and cannot be found in a between-subject crowdsourcing study where we do not control screen-size, attention, outside distractions, sound quality, etc., (3) other factors are far more important for perceived competence, warmth or human-likeness than prosodic head motion (e.g. speech quality, lipsync quality). We aim to tease apart which of these reasons explain our results in further studies. Our experimental design to assess warmth, competence and humanlike-ness was based on a successful laboratory study on the perceived effects of gesture [1]. To assess (2), we plan to both run our study in the laboratory and the laboratory study of Bergmann et al. [1] in a crowdsourceing experiment. Point (3) is supported by the comments of subjects on the quality of the TTS and the lack of other facial motion. Using visual prosody on more than one modality has been shown to enhance the perceived human-likeness of an IVA for real speech [5,15]. In future work we thus aim to enhance the online visual prosody model to include more modalities such as eye and eyelid movement (e.g. using the online model of [11]) and eyebrow movement (as in [5,15]) and assess if those help us in enhancing the human-likeness, warmth and/or competence of TTS-driven real-time visual prosody.

Acknowledgements. We would like to thank Kirsten Bergmann and Philipp Kulms for their feedback on the design of the study and their help with the evaluation of the results. This work was partially performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02. It was also partially funded by the EU H2020 project ARIA-VALUSPA; and by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition, managed by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the contents of this publication.

References

1. Bergmann, Kirsten, Kopp, Stefan, Eyssel, Friederike: Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In: Safonova, Alla (ed.) IVA 2010. LNCS, vol. 6356, pp. 104–117. Springer, Heidelberg (2010)
2. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
3. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Comput. Animation Virtual Worlds **16**(3–4), 283–290 (2005)
4. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. Trans. Graph. **24**(2), 331–347 (2005)
5. Ding, Y., Pelachaud, C., Artières, T.: Modeling multimodal behaviors from speech prosody. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (eds.) IVA 2013. LNCS, vol. 8108, pp. 217–228. Springer, Heidelberg (2013)
6. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Conference on Multimedia, pp. 835–838. ACM (2013)
7. Fiske, S.T., Cuddy, A.J.C., Glick, P.: Universal dimensions of social cognition: warmth and competence. Trends Cogn. Sci. **11**(2), 77–83 (2007)
8. Graf, H.P., Cosatto, E., Strom, V., Hang, F.J.: Visual prosody: facial movements accompanying speech. In: Automatic Face and Gesture Recognition, pp. 381–386. IEEE Computer Society (2002)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
10. Heylen, D.K.J.: Head gestures, gaze and the principles of conversational structure. Int. J. Humanoid Rob. **3**(3), 241–267 (2006)
11. Le, B.H., Ma, X., Deng, Z.: Live speech driven head-and-eye motion generators. Trans. Visual Comput. Graphics **18**(11), 1902–1914 (2012)

12. Lee, J., Marsella, S.: Modeling speaker behavior: a comparison of two approaches. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 161–174. Springer, Heidelberg (2012)
13. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. Trans. Graph. **29**(4), 124:1–124:11 (2010)
14. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. In: SIGGRAPH Asia, pp. 1–10. ACM, New York (2009)
15. Mariooryad, S., Busso, C.: Generating human-like behaviors using joint, speech-driven models for conversational agents. Audio Speech Lang. Process. **20**(8), 2329–2340 (2012)
16. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. **15**(2), 133–137 (2004)
17. van Welbergen, H., Yaghoubzadeh, R., Kopp, S.: AsapRealizer 2.0: the next steps in fluent behavior realization for ECAs. In: Bickmore, T., Marsella, S., Sidner, C. (eds.) IVA 2014. LNCS, vol. 8637, pp. 449–462. Springer, Heidelberg (2014)