Talking Bodies: Sensitivity to Desynchronization of Conversations

Rachel McDonnell
Trinity College Dublin, Ireland, rachel.mcdonnell@cs.tcd.ie

Cathy Ennis
Technological University Dublin, cathy.ennis@tudublin.ie

Simon Dobbyn

Carol O’Sullivan
Trinity College Dublin, Ireland, carol.osullivan@cs.tcd.ie

Follow this and additional works at: https://arrow.tudublin.ie/aircart

Part of the Computer Sciences Commons

Recommended Citation

This Article is brought to you for free and open access by the Applied Intelligence Research Centre at ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
Talking Bodies: Sensitivity to Desynchronization of Conversations

RACHEL McDONNELL, CATHY ENNIS, SIMON DOBBYN, and CAROL O’SULLIVAN
Trinity College Dublin

In this article, we investigate human sensitivity to the coordination and timing of conversational body language for virtual characters. First, we captured the full body motions (excluding faces and hands) of three actors conversing about a range of topics, in either a polite (i.e., one person talking at a time) or debate/argument style. Stimuli were then created by applying the motion-captured conversations from the actors to virtual characters. In a 2AFC experiment, participants viewed paired sequences of synchronized and desynchronized conversations and were asked to guess which was the real one. Detection performance was above chance for both conversation styles but more so for the polite conversations, where desynchronization was more noticeable.

Categories and Subject Descriptors: I.3.7 [Computer Graphics]: Three Dimensional Graphics and Realism—Animation
General Terms: Experimentation, Human Factors
Additional Key Words and Phrases: Perception, graphics, motion capture

ACM Reference Format:

1. INTRODUCTION

When a group of people are talking, verbal and nonverbal cues are used in order to signal turn-taking and appropriate interruptions. Little is known about our sensitivity to these cues for virtual characters. However, groups of idle people conversing are important for the realistic depiction of a crowded scene (Figure 1). Research on virtual crowds in the past has focused heavily on agent locomotion, but rarely have there been efforts to create realistic groups of stationary humans. We propose to use motion-captured data of real conversations in order to create realistic groups of conversing characters. However, if motion data or storage is limited, this may result in many duplicated conversations, which might appear unrealistic. We therefore wish to examine the circumstances under which combining and reusing segments of recorded conversations would appear realistic to the observer.

There are two questions that we wish to address: Firstly, based on body language alone, are humans able to determine if a real conversation looks real? Evidence from neuroscience suggests that from body
motion alone, humans are able to distinguish between speaker and listener roles for certain emotive conversations [Rose and Clarke 2009]. However, since the conversations used in their study involved strong emotions, we are unsure if this will generalize to more natural conversational settings.

Secondly, in order to increase the variety of crowd scenes, will mixing and matching segments of motion captured conversations appear realistic? Previous work in the computer graphics field has focused on creating per-person variety by transplanting limb animations [Ikemoto and Forsyth 2004] or upper and lower body animations [Heck et al. 2006]. We wish to determine if rescheduling animations across characters in a group will result in plausible conversations.

2. BACKGROUND

While much work has been carried out in the area of simulating behavior for virtual crowds [Thalmann 2001; Musse and Thalmann 2001; Paris et al. 2007], little is known about how we perceive this simulated behavior. Of particular interest is how we perceive conversational agents in a virtual scene and what are the important cues that we attend to when viewing realistic conversations.

Our previous work on the perception of group plausibility has looked at pedestrian formations for different urban scenarios. Rules were implemented for positioning and orientating characters in static scenes based on the context of the scene and these were compared to virtual representations of real-life formations [Peters et al. 2008; Ennis et al. 2008]. We found that participants were able to distinguish real from synthetic formations for certain scenes. Following this, the importance of dynamic groups in virtual pedestrian crowd scenes was investigated [Peters and Ennis 2009]. Scenes were populated with different combinations of singles, pairs, and groups of three and participants were asked which scenes were more realistic. We found that groups are an important factor in crowd behavior, but that the addition of groups to virtual scenes needs to be done in an intelligent way to add to the user’s perception of realism.

Research on creating the illusion of variety in crowd scenes has focused on changing the appearance of characters (e.g., de Heras Ciechomski et al. [2005] and Maím et al. [2007]). However, motion variety has received less attention. In previous work, we investigated the sensitivity of users to motion repetition in crowd scenes [McDonnell et al. 2008]. We detailed the durations that clones of walking motions could be displayed without being detected, and found that ten clones could only be displayed for ten seconds.
without being detected. It is possible that gesture repetition in a group could be even more noticeable, due to the fact that gestures can be very distinctive for individuals.

There is much evidence that people can recognize human motion from very small cues [Johansson 1973]. Rose et al. [2009] investigated how speaker/listener roles can be identified from emotional biological motion. They found that speaker identification was above chance for angry and joyful conversations.

The idea of conversational agents has been investigated mostly for human-computer interaction applications. Cassell et al. [2001] presented a method to automatically generate conversational behaviors using appropriate speech, facial expressions, and hand gesturing for two agents face to face in a scripted conversation. The dialog planner runs through the text of the script and identifies cues based on the semantics of content for a range of gestures. The idea of conversational agents has been investigated mostly for human-computer interaction applications. Cassell et al. [2001] presented a method to automatically generate conversational behaviors using appropriate speech, facial expressions, and hand gesturing for two agents face to face in a scripted conversation. The dialog planner runs through the text of the script and identifies cues based on the semantics of content for a range of gestures. In earlier work [O'Sullivan et al. 2002a, 2002b], we employed a behavior generation toolkit based on the method outlined previously to generate levels of detail for conversational behavior for groups of agents. This was achieved using filter rules to remove generated behaviors based on a group's visual saliency within the scene.

While we do not focus on creating a method for automatic conversational behavior generation in this article, we hope to discover more about conversing virtual humans with that aim in mind for the future. In this article, we focus on the perception of conversational body language in the first instance. While the computational cost of rendering face and hand animations is prohibitive for real-time crowd systems, we do intend to investigate the role of these factors in future work.

3. STIMULI CREATION

Three male actors participated in a motion capture session which was conducted on a 13 camera Vicon optical system, with 46 markers per actor. The actors were instructed to stand at the corners of a triangle in the capture space of the motion capture system for the start of each recording. During the recording, they were free to move around close to their starting point. Since we were not recording finger motion, we placed fabric around their fingers to restrict movement in this area. Before the session, a list of topics was given to the actors and they indicated which topics they were most comfortable discussing. All actors were nonprofessionals, familiar with each other, the environment, and the motion capture setup. Conversation was free-flowing and unscripted, which resulted in natural dialog. Recordings ranged in length from one to three minutes.1

Two different styles of conversation were recorded. The first had a dominant speaker, where one person from the group spoke while the others politely listened. Each actor took a turn at being the dominant speaker, resulting in three recordings of this type of conversation. The second style was a debate, where each actor had an affirmative or negative position on the topic being discussed. This style of conversation allowed for interruptions while discussing different opinions. We recorded four debates in total.

Using video footage as reference, each recording of the dominant speaker conversations was annotated for each actor to indicate at which frames of animation they were talking or listening for longer than 10 seconds (Figure 2).

For the virtual scene, we chose three male virtual characters to represent our three actors. We ensured that the characters were the same relative height as the actors, in order to minimize retargeting errors. A white background was used in order to provide good contrast. The camera was placed so that one character was centrally focused and the gestures of the other two were clearly visible (Figures 3 and 4). For each trial, the camera was randomly focused on one of the three characters, in order to avoid any effect of character preference.

1 see video at http://gv2.cs.tcd.ie/mcdonner/TAP09.
4. EXPERIMENT
Twenty-eight naïve volunteers (17 male-11 female) from different educational backgrounds took part in this experiment, and received a book voucher as a token of appreciation. A 2 Alternative Forced Choice (2AFC) paradigm was used in this experiment, where the following were presented in random
Talking Bodies: Sensitivity to Desynchronization of Conversations

Fig. 5. Examples of desynchronized dominant speaker conditions, where 10-second talker or listener (T/L) clips are chosen at random from the annotated recordings.

order: a 10-second clip of a real (synchronized) conversation (Figures 3 and 4), and a 10-second clip of an altered (desynchronized) conversation. Participants were asked to indicate which conversation was the real one, using a left mouse click to indicate the first conversation, or a right click to indicate the second conversation.

The experimental system was developed using an open-source renderer that used DirectX 9.0 and a commercially available animation engine. This allowed us to randomly seed the experiment for each participant to allow for much variation in the stimuli (i.e., repetitions were randomly selected from the recordings). This avoided any effects that may have occurred due to the repetition of particular animation sequences. The experiment was run on a workstation with 2GB of RAM, an 8-series GeForce graphics card on a wide-screen 24-inch LCD monitor.

Participants viewed two counterbalanced blocks, where block 1 showed synchronized and desynchronized dominant speaker conversations and block 2 showed synchronized and desynchronized debates.

In the dominant speaker block, we tested the number of talkers condition, in order to determine if this had an effect on the realism of the desynchronized conversations. We tested: 3 talkers, 2 talkers, and 1 talker, where the others in the group were listeners. We also tested a 0 talker condition where no character was talking, to represent the situation of a break in conversation. Ten-second motion clips were chosen at random from the three dominant speaker recordings to match these conditions (Figure 5). We ensured that the virtual characters were consistently matched to the same actors (e.g., the 3 talker case had a clip from each of the actors rather than 3 from one actor). Therefore, in both the real and altered conversations, the virtual characters had the same personality traits. For the real conversations in this block, 10-second synchronized motion clips were randomly chosen for each character from the 3 dominant speaker recordings. Twelve pairs were shown in total (3 repetitions of each number of talkers).

In the debates block, we tested the topic context condition, in order to determine if desynchronized clips from the same recording would be more acceptable than those from different recordings. As before, we ensured that each virtual character was consistently matched to an actor. For the real debate conversations, 10-second synchronized motion clips were randomly chosen for each character from the 4 debate recordings. Twelve pairs were shown in total (6 repetitions of each topic context).

We hypothesized that desynchronized debates would be less noticeable than desynchronized dominant speaker conversations, due to the fact that breaking turn-taking etiquette may be more noticeable for polite conversations since there should only be one talker at a time. Conversely, debates by nature are more chaotic and tend not to follow turn-taking rules, which may result in them appearing more realistic when desynchronized. We also hypothesized that participants would find 0 or 3 talkers particularly unrealistic, and that desynchronized conversations from out-of-context debates may appear less realistic, due to the level of argumentation not matching.
For each participant, we recorded their ability to identify the synchronized animation sequence for both debates and dominant speaker conversations. We averaged the results over all repetitions for each participant. Since corresponding conditions were not present for dominant speaker and debates, we averaged over the number of talkers for the dominant speaker and the topic context for debates. A repeated measures ANalysis Of VAriance (ANOVA) was then conducted to determine if there was an effect of conversation type. A main effect was found \((F_{1,26} = 5.2, p < 0.04)\), which was due to the desynchronized dominant speaker conversations being easier to detect (70% detection) than the desynchronized debates (61% detection) (Figure 6).

In order to determine if there was an effect of the number of speakers in the desynchronized dominant speaker conversations, a repeated measures ANOVA was conducted. Surprisingly, we found no effect of speaker number, which implies that desynchronized conversations using 0 or 3 talkers did not stand out as more unrealistic than any other type (Figure 7). We feel that this could be due to the fact that humans put their own interpretations on the conversations, making them more tolerant of different contexts.
combinations of talker/listeners. For example, 3 or 0 talker conversations may have been perceived as heated debates or breaks in conversation.

We also found no effect of topic context for the debates, where participants were able to detect anomalies equally for both the same and different debate topics. Again, this could be due to people associating their own story or scenario with the conversations that they viewed, thus making desynchronized conversations appear plausible at times.

Fifty percent of our participants (5 female-9 male) were informed less about the task than the others (6 female-8 male). Half were told only that the desynchronized conversations were “altered in some way,” while the other half were specifically told that they were desynchronized. A between-groups ANOVA on the results showed that there was no difference between the detection abilities of the participants in the first group from those in the second. This implies that even when participants knew what they were looking for in terms of anomalies, this did not change their performance in detection of anomalies.

Differences have been reported between male and female interpretations of conversations (e.g., Hannah and Murachver [1999]). Therefore, we used a between-groups ANOVA to determine if there were differences between detection rates for our male and female participants. We found no effect, which implies that conversation interpretations were consistent regardless of the sex of the participant.

6. CONCLUSION

Our results show that people can detect real synchronized conversations above chance using body motion alone. We also showed that desynchronization of debate or argumentative-style conversations is less noticeable than for dominant speaker (polite) conversations. Therefore, using debate-style conversations played back in different orders could be a good solution to create variety for idle groups in virtual crowd scenes. We feel that head-look alterations could improve the believability of the desynchronized dominant speaker conversations. Future experiments will test this.

Previously, we felt that creating variety for dominant speaker conversational groups would be restricted to assigning a speaker role to only one character in the group at any given time. Interestingly, we found no effect of the number of talkers in the desynchronized dominant speaker conversations. This implies that, based on body motion alone, participants were equally convinced by all combinations. This allows us to increase the variety in our groups by using different talker/listener ratios.

One limitation of our results is that we used a single set of male actors. It would be interesting to test if our results hold true for other actors and combinations of males and females or all female actors. Due to the lack of facial and hand animation in the stimuli presented in this article, it is possible that “listener” gestures may have been mistaken as talker gestures (e.g., head nods, etc.). Therefore, we feel that with the addition of facial and hand animation, participants could be more sensitive to desynchronization. Sound will almost certainly disambiguate speakers, but it is unclear to what extent this it true in a crowd scene. These are other factors that we wish to address in future experiments.

Eye-tracking would be another way in which we could find out more about how we perceive virtual conversing agents. Using an eye-tracking device, we could determine what participants attend to in a virtual conversation. For example, we might find that people pay more attention to head orientation or gestures. This information would allow us to improve the believability of the desynchronized conversations, by altering the salient areas.

In order to create further variety in the groups for crowd simulations, different character positioning within the group and different numbers of characters would be necessary. We would like to investigate if changing the positions of the characters within a group has an effect on the perception of realism of the interactions between characters in a group. Also, if we add or subtract members from a group, to what degree will the animations of the characters need to be altered before the conversation becomes perceptually plausible to a viewer?
REFERENCES


Received July 2009; accepted August 2009