2010

Seeing is Believing: Body Motion Dominates in Multisensory Conversations

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

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

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

Recommended Citation

Ennis, C., McDonnell, R. and O'Sullivan, C. (2010) Seeing is Believing: Body Motion Dominates in Multisensory Conversations. *ACM transactions on graphics*, Volume 29, No. 4, 2010. doi:10.1145/1778765.1778828

Seeing is Believing: Body Motion Dominates in Multisensory Conversations

Cathy Ennis* Rachel McDonnell† Carol O'Sullivan‡

Graphics, Vision and Visualisation Group, Trinity College Dublin

Abstract

In many scenes with human characters, interacting groups are an important factor for maintaining a sense of realism. However, little is known about what makes these characters appear realistic. In this paper, we investigate human sensitivity to audio mismatches (i.e., when individuals' voices are not matched to their gestures) and visual desynchronization (i.e., when the body motions of the individuals in a group are mis-aligned in time) in virtual human conversers. Using motion capture data from a range of both polite conversations and arguments, we conduct a series of perceptual experiments and determine some factors that contribute to the plausibility of virtual conversing groups. We found that participants are more sensitive to visual desynchronization of body motions, than to mismatches between the characters' gestures and their voices. Furthermore, synthetic conversations can appear sufficiently realistic once there is an appropriate balance between talker and listener roles. This is regardless of body motion desynchronization or mismatched audio.

CR Categories: I.3.7 [Computer Graphics]: Three Dimensional Graphics and Realism—Animation;

Keywords: perception, crowds, conversational agents

1 Introduction

Cut-scenes in games have been around for at least two decades [Rouse 1998], and more recently the use of performance capture (where both motion capture and audio are recorded together) has been growing in popularity as it produces the most alive and realistic characters. However, it is not always possible due to the cost of hiring A-list actors or VIP characters (such as Roger Federer or Rafael Nadal in Topspin 3™) to perform motion and voice, or because of the location constraints necessary for high quality audio [Edg 2010]. In these cases, audio and motion are often captured separately. However, little is known about what effects, if any, possible desynchronization has on the perception of the final sequence.

Many developments have been made in real-time crowds over the past years, with video games such as Assassin's Creed involving the investing of significant resources into developing interactive and believable crowds [Bernard et al. 2008]. However, still missing from real-time crowd applications is the concept of realistic conversing groups. In many scenes depicting crowds or groups, it is reasonable to expect a significant proportion of the crowd to be walking or standing in groups, interacting and conversing. While we are acutely sensitive to social cues and rules for conversing with one another, little is known about what we expect from virtual agents in order for them appear realistic. Much research has been carried out into the area of gesture generation for embodied conversational agents [Bickmore and Cassell 2005; Cassell et al. 2001a], but it is still not clear how these gestures are perceived by a user, especially for virtual characters interacting with each other.

Figure 1: *Screenshot of our real-time crowd system with dynamic agents and conversing groups.*

We investigate appropriate desynchronization methods for urban crowds or other scenes where conversing groups would be needed, e.g., cocktail party or concert foyer scenes. In such scenarios, it would be useful to know how to piece together realistic conversations for groups of virtual agents from a finite resource of audio and motion capture data. Therefore, we wish to investigate human sensitivity to visual desynchronization (when the body motions of the individuals in a group are mis-aligned in time) and mis-matched audio (when individuals' voices are not matched to their gestures). Because the application of our results is intended to be for background (i.e., non-hero) characters, we do not consider facial animation for the purpose of this study. However, our results should also be applicable to scenarios with smaller numbers of characters, or high level of detail groups in large crowd scenes. In these cases, it may be necessary to further examine the role of facial animation along with body motion.

How do people react when presented with a virtual conversation with visual and aural information? Do they rely on one sense more than another? If the motions of the characters do not match the audio, how realistic will this appear? If one of the bodies is matched to the audio, will this improve plausibility? We conducted the first series of experiments to determine perceptibility thresholds for desynchronization of body motion in both the absence and presence of audio, both omni-directional and when correctly and incorrectly localized in 3D space. Overall, we found that people attend more to

*e-mail: ennisca@cs.tcd.ie

†e-mail: Rachel.McDonnell@cs.tcd.ie

‡e-mail: Carol.OSullivan@cs.tcd.ie

the body motions of characters than to the audio; once the body motions, talker/listener roles and interactions between the characters appear plausible, it is not always necessary to ensure that the gestures and audio beats match. However, it does help to match the audio to the motion of at least one talker. Our results provide guidelines to enhance the believability of synthetic conversations in real-time applications depicting groups and crowds.

2 Related Work

Much research has been carried out in the area of human motion perception and it is well documented that human motion can be recognized from minimal visual cues [Johansson 1973]. Rose and Clarke [2009] investigated how good people are at detecting talker/listener roles using biological motion alone. They found that for conversations depicting most of the six Ekman basic emotions [Ekman 1992], people were able to identify the speaker correctly. This reiterates previous research that shows that people are acutely sensitive to gesturing and its role in interpersonal communication [McNeill 1996]. However, Rose and Clarke's study was carried out for emotional conversations, and does not address the issue of sensitivity to gestures and roles for non-emotional, natural conversations.

People use many different types of gestures to reinforce or clarify the point they are making while conversing [Ekman and Friesen 1969]. These range from explicit gestures that physically describe an action or thing, to more implicit gestures that are almost involuntary, such as emotional reactions. It has been well documented that speech and these non-verbal communication methods are very heavily linked [Goldin-Meadow 2005; Kendon 1994]. Krahmer and Swerts [2007] found that words are perceived to be more prominent when they are accompanied by a visual beat cue. Giorgolo and Verstraten [2008] investigated the perception of speech and gesture integration for body motion in the absence of facial motion. They found that people are sensitive to temporal delays between speech and gestures of 500ms or more and that this affected how acceptable participants found video clips of a person talking. This implies that both visual and aural signals are closely integrated in the brain when viewing human conversations. Our work also investigates the effect of mis-matching audio and body motion, but for virtual characters and with more extreme scenarios, where the gesture and motions are from different conversations.

Conversational agents have been an important element of human-computer interfaces. However, little is known about how realistic these agents appear to be. Vilhjámsson and Cassell [1998] presented a method to automatically generate body gestures and conversational properties like turn-taking and feedback. They describe a system where the user communicates with an avatar via text and the avatar reacts with appropriate body gestures, facial expressions and simple body functions. Early perceptual evaluation results showed that users preferred a combination of their model and manual controls when controlling an avatar in conversational situations. Cassell et al. [2001b] implemented a similar tool where an appropriate animation gesture is chosen and applied to the character based on the linguistics of the conversation. Also relevant, though not confined to conversational groups, we previously found that adding plausible groups to a pedestrian crowd scene is important for an increased sense of realism [Peters and Ennis 2009].

Levine et al. [2009] used Hidden Markov Models to generate gestural body animations automatically using speech, rather than text input. Results from user studies found that the model used was preferable to random synthesis of gestures for the same videos. While this study gives us some insight into the perception of communicative behavior, it focusses on the gestures of individuals rather than interactions between individuals. Data driven approaches have been used to generate behavior for virtual agents for various purposes. Lerner et al. [2009] used video based human trajectory examples to animate agents who interact with both each other and with the environment. Neff et al. [2008] used video input to create a statistical model of a person's gesture style, which was then used to output a continuous stream of speech coordinated gestures. More related to our study, it has been shown that people can distinguish real conversations, where the body motions for characters are synchronized with each other, from desynchronized conversations, where motions from different conversations were used on each character [McDonnell et al. 2009]. It was also found that people are more sensitive to desynchronization in polite conversations compared to argumentative scenarios.

3 Motivation

In this paper, we investigate the importance of matching audio to body motion for virtual conversing characters. For real-time crowds, problems can arise with data requiring too much memory, so there is a need to ensure that motion captured data is as reusable as possible. There are also issues with replaying real conversations. In a dynamic scene, characters will be leaving and joining groups at different times, and it will not always be possible to synchronize character animations. Also, when the camera is moving through the crowd and a conversing group becomes salient, the characters will not necessarily be synchronized with each other due to the complexity of finding suitable blending animations for looping or transitioning animations. In these cases, it would be easier to simply match an audio clip to the animation of one character in the group.

Adding a sense of variety in a crowd is an important factor to make the scene appear real. Research has been conducted into methods to create the illusion of variety for appearance and motion of characters [McDonnell et al. 2008]. Variety in the behaviors of characters is also important in order to simulate agents who appear to have individual personalities [Durupinar et al. 2008]. Motion capture data is time consuming to obtain and process for conversational groups. Therefore, any real-time crowd system containing such groups will contain a limited data set, leading to cloned conversations.

We wished to find out if conversations appeared plausible if only the talker in the conversation was matched to the audio. Of interest to us also was whether it would be possible to use unmatched audio for the motions being displayed in a conversation, once the motions are synchronized as they were captured. If people are sensitive to desynchronization of audio and visual information, we wanted to further investigate which modality people would rely on when judging realism, and how implausible a conversation would appear if both audio is unmatched and motions are desynchronized. We conducted a set of experiments in order to answer these questions.

4 Experiment Design

For these experiments, we created a number of real and synthetic conversational scenarios depicting a group of three individuals. The real conversations involved replaying the body motions and audio as captured. Our notations for synthetic conversations are in "B_xA_y" format, where B refers to Body motions and A refers to Audio (see Table 1 for an overview of notational conventions). In our first experiment we used visual only stimuli, so the body motion conditions are suffixed by "A_0". Throughout this study, we will use the terms *synchronized / desynchronized* to refer to the body motion of our characters and *matched / unmatched* when referring to audio (as we do not desynchronize the voices from each other, we simply mis-match the entire audio clip from the body motions).

	Bodies (B)	Audio (A)
$Real$	sync (S)	matching (M)
$B_S A_{NM}$	sync (S)	not matching (NM)
$B_D A_{NM}$	desync (D)	not matching (NM)
$B_D A_{M1}$	desync (D)	matching 1 talker (M1)
$B_{0T} A_{NM}$	0 talkers (0T), 3 list.	not matching (NM)
$B_{1T} A_{M1}$	1 talker (1T), 2 list.	matching 1 talker (M1)
$B_{2T} A_{M1}$	2 talkers (2T), 1 list.	matching 1 talker (M1)
$B_{3T} A_{M1}$	3 talkers (3T), 0 list.	matching 1 talker (M1)

Table 1: *Notation used for describing experimental conditions.*

The first synthetic condition with audio we investigated was **synchronized bodies not matched to audio ($B_S A_{NM}$).** All three body animations were chosen from the same point in time of a conversation, but a different audio clip was chosen at random. For this and all following random conditions, we ensured that the audio never matched the body motions by chance. In the case of body motion desynchronization, the motions were never from the same clip. The rationale for this condition was that, if a high sensitivity threshold was found, this would allow us to add variety to a scene by playing a particular conversation animation with random audio.

Our second synthetic condition was **desynchronized bodies not matched to audio ($B_D A_{NM}$),** where each of the three body animations were chosen at random with a randomly chosen audio clip. If this condition was found by participants to be plausible, it would mean that random body motions and audio clips could be played together, giving a large range of variety. This would allow us to resolve blending problems and would easily facilitate dynamic group formation.

Our next three conditions were **1, 2 and 3 talkers, with audio matched to 1 talker ($B_{1T} A_{M1}$, $B_{2T} A_{M1}$, $B_{3T} A_{M1}$).** Here, either 1, 2 or 3 talker body animations were chosen from different conversations at random (the other characters used random listener body animations) with audio randomly matched to one of the talker motions. We used these conditions in order to determine whether any random animation could be chosen for other characters in a group, once one talker was matched to the audio. This would be an easy way to populate a crowd scene with conversing groups.

For completeness, our final synthetic condition was chosen to investigate the effect of conversational audio being played with no character displaying a talker body motion: **0 talkers, audio not matched ($B_{0T} A_{NM}$).** This contained 3 listener body animations, chosen at random, with a random audio clip playing.

5 Stimuli Creation and Experimental Framework

Two sets of actors participated in recording sessions where we recorded both their voices and body movements. The first set contained three males aged between 25 and 31. The second had three females aged between 22 and 28 (see Figure 2). We chose two groups of actors in order to test if our results were generalizable. Briton and Hall [1995] found that female actors were more expressive with gestures and that female participants rated female gestures higher than male gestures, which were found to be less fluid and more interruptive. We chose groups of three to allow us to capture group dynamics that would not be as present if using groups of two, such as turn taking, interruptions and gaze shifting. All actors were non-professionals and were accustomed to the motion capture setup and environment. We also ensured that actors knew each other and were informed about the topics that they would be discussing in advance, in order to ensure natural, realistic conversations.

Figure 2: *Our audio and motion capture setup with three female actors.*

Motion capture was conducted using a 13 camera Vicon optical system, with 52 markers per actor. The markers were placed on the major joints and at regular intervals on the body, in order to capture accurate body motion. We did not capture finger or face motion as it was conversational body motion that was of most interest to us in this study. An AKG C-414 omni-directional microphone was placed on a tripod in the centre of the actor triangle to record their voices from all directions while they were being motion captured. Also, we placed a Behringer C-2 studio condenser microphone in front of each actor to record only their part of the conversations, using a MOTU-896HD external soundcard. We wanted to collect audio for all actors simultaneously and each actor individually so that we could play audio from the centre of the conversation (Section 6.2), and also position individual audio tracks corresponding to 3D positions of characters on screen (Section 6.4). A clapboard with motion capture markers attached was used to indicate the start of a conversation clip. Actors were instructed to place their feet in pre-specified positions on the edges of a triangle at the start of each capture in order to prevent significant changes in their positions. In advance of each recording, we adjusted the preamp gain for each actor to ensure that no audio distortion occurred due to microphone proximity. Thereby, we minimized the audio recording constraints and the actors were informed that they could move around freely within the motion capture zone once capture had begun.

We captured and recorded two different conversation types: *debates* and *dominant speaker*, as we previously found differences in participant sensitivity to the desynchronization of characters' motions in these two conditions [McDonnell et al. 2009]. Debate conversations were free-flowing in nature, where each actor expressed a strong opinion on the topic being discussed, and interruptions were common. Dominant speaker conversations allowed only one speaker to talk at a time, while the others politely listened and were not allowed to interrupt. In total, we recorded 30 dominant speaker conversations (5 per each of the 6 actors) and 10 debates (5 with the female actor group and 5 with the male). Dominant speaker conversations lasted approximately 30 seconds, while debates lasted between 2 and 3 minutes.

We chose six virtual characters to represent each actor in the experiments (Figure 3). The characters were chosen to approximately match the actors in age, weight and height, to minimize re-targeting errors. Throughout the experiments, we matched the motions of each actors to their virtual character.

Figure 3: *Examples of stimuli used in our experiments: (L) a real female debate and (R) a real male dominant speaker conversation.*

For each dominant speaker conversation, we chose two different temporal offsets from the start of the conversation to begin a 10 second conversation clip. For each debate conversation, we chose six different offsets. For the dominant speaker conversations, we annotated clips to tag each character as either a speaker or listener. For the virtual representations of the real conversations, we played the correct conversational audio and motion capture clips simultaneously. The synthetic conversations were made up of the conditions displayed in Tables 2 and 5 and described in Section 3.

The real-time experimental system was developed using a commercially available animation system and an open-source renderer. This allowed us to seed each experiment randomly for each participant. In order that participants did not always associate a voice with a character, we color modulated the characters at every trial to disguise them. Also, we placed the camera so that one of the characters was centrally focussed, but randomly chose which character to focus on at each trial. See Figure 3 for examples of stimuli.

The experiments were run on a workstation with 4GB of RAM, a Creative SB Audigy 2ZS soundcard and an 8-series G-Force graphics card. The stimuli were displayed on a 24-inch widescreen monitor and participants used Sennheiser HD 202 headphones to listen to the audio (Figure 4). Participants viewed each trial for 10 seconds and were asked to indicate using a mouse click whether they thought that the conversation they viewed was *real* or *synthetic*. We found through interviewing participants that 10 seconds was an adequate time for them to make their decisions. We randomly associated the right or left mouse button with the *real* response so as to avoid any bias towards a particular button-press. Similarly, 50% of our stimuli were real, in order to avoid any bias. After a participant gave his/her response, a cross was displayed to focus attention on the centre of the screen.

Figure 4: *Participant in Localized Audio experiment.*

Block	Factor	AV Condition	Total Trials
Debates	Real	Real ($B_S A_0$)	6
	Synth.	$B_D A_0$	6
Dom. Sp.	Real	Real ($B_S A_0$)	24
	Synth.	$B_{0T} A_0$	6
		$B_{1T} A_0$	6
		$B_{2T} A_0$	6
		$B_{3T} A_0$	6

Table 2: *Experimental design for No Audio experiment, showing total number of trials (50% male actors, 50% female).*

6 Experiments

In order to answer our questions posed in Section 3, we conducted a set of three experiments that examined how audio and body motion affect the perception of virtual conversations. We first conducted a No Audio experiment, to identify the sensitivity of participants to body motion desynchronization alone. We then conducted two experiments with audio. The first explored omni-directional audio, where we played the same mono audio track in both headphones. We wanted to determine whether the addition of audio had any effect on participants' sensitivity to body desynchronization. The final experiment considered localized audio, where we played a separate audio track from each character's 3D position on screen through the headphones. We conducted this experiment to determine whether the addition of richer, localized audio would have any effect on participants' ability to distinguish real from synthetic conversations.

Our experimental design, including factors and conditions tested, can be found in Tables 2 and 5. For our analysis of the experiments, we conducted a 3-way repeated measures ANalysis Of VAriance (ANOVA) with within-subjects factors of *AV condition* and *actor group*. The No Audio experiment analysis had an additional within-subjects factor of *conversation type* (discussed in Section 6.1.1). *AV condition* refers to the various audio visual combinations, for both real and synthetic conversations (which were counterbalanced). *Actor group* refers to the male and female motion captured groups, while *Conversation type* refers to dominant speaker conversations and debates (which were separated into two blocks). For both experiments containing audio, there was a between-subjects factor of *sex of participant* (see Sections 6.2.1 and 6.4.1). Cross experimental analysis was conducted using 2-way repeated measures ANOVAs, with within-subject factors of *AV condition* and *audio signal level*, where the three audio signal levels refer to No Audio, Omni-Audio and Localized Audio (discussed in Section 7). Post-hoc analysis was conducted using Newman-Keuls tests for comparison of means and only significant results at the 95% level are reported.

For each of our experiments, participants were over 18 years of age, naïve to the purpose of the experiment and from a range of disciplines. Ethical approval was granted for all experiments, and participants were recruited via email. They were given a book voucher to compensate for their time.

6.1 No Audio Experiment

In the No Audio experiment, we explored participants' sensitivity to body motion desynchronization in the absence of audio. We hypothesized that debate style conversations would be more plausible than dominant speaker conversations.

Experiment	Block	Effect	F-Test	Post-hoc
No Audio	Overall	Conversation Type	$F_{1,11} = 10.066, p < .05$	Debates more realistic than Dominant Speaker
	Dominant Speaker	AV Condition	$F_{4,44} = 27.193, p < .00005$	Real more realistic than synthetic $B_{3T}A_0$ less realistic than $B_{0T}A_0$ and $B_{1T}A_0$
	Debates	AV Condition	$F_{1,11} = 20.731, p < .0005$	Real more realistic than synthetic
Omni-Directional	Dominant Speaker	AV Condition	$F_{6,102} = 40.63, p < .00005$	Real more realistic than synthetic $B_{1T}A_{M1}$ and B_SA_{NM} most realistic synthetic conditions $B_{0T}A_{NM}$ and B_DA_{NM} next realistic synthetic conditions $B_{2T}A_{M1}$ next realistic synthetic condition $B_{3T}A_{M1}$ least realistic condition
	Debates	AV Condition	$F_{3,51} = 33.63, p < .00005$	Real more realistic than synthetic B_DA_{M1} more realistic than B_SA_{NM} and B_DA_{NM}
Localized	Dominant Speaker	AV Condition	$F_{6,66} = 14.193, p < .00005$	Real more realistic than synthetic B_SA_{NM} most realistic synthetic condition $B_{1T}A_{M1}$ and $B_{0T}A_N$ next realistic synthetic conditions $B_{2T}A_{M1}$ and $B_{3T}A_{M1}$ least realistic synthetic conditions
		AV Condition × Actor	$F_{6,66} = 2.2685, p < .05$	Male voices more realistic than females for $B_{1T}A_{M1}$
	Debates	AV Condition	$F_{3,33} = 28.94, p < .00005$	Real more realistic than synthetic
No Audio vs. Omni-Directional vs. Localized	Dominant Speaker	AV Condition	$F_{4,164} = 86.195, p < .00005$	Real more realistic than all synthetic $B_{1T}A_{M1}$ most realistic synthetic $B_{0T}A_{NM}$ next most realistic synthetic $B_{2T}A_{NM}$ next most realistic synthetic $B_{3T}A_{NM}$ least realistic synthetic
		AV Cond. × Audio Sig. Level	$F_{8,164} = 2.05, p < .05$	$B_{1T}A_{M1}$ more realistic for Omni-Directional than others
	Debates	AV Condition	$F_{1,41} = 132.37, p < .00005$	Real more realistic all synthetic
Omni-Directional vs. Localized	Dominant Speaker	AV Condition	$F_{1,30} = 35.62, p < .00005$	B_SA_{NM} more realistic than B_DA_{NM}
	Debates	AV Condition	$F_{1,30} = 6.22, p < .05$	$B_{1T}A_{M1}$ more realistic than B_SA_{NM}

Table 3: *Significant results for each experiment and cross-experimental analysis. Note: for all analyses except the No Audio experiment, ANOVAs contained between-subject factors as well as repeated-measures factors. Omni-Directional and Localized Audio experiments had a between-subject factor of sex of participant. For all cross-experimental analysis, there was a between-subject factor of audio signal level. The No Audio, Omni-Directional and Localized Audio experiments had 12, 19 and 13 participants respectively.*

We conducted a similar experiment to that described in [McDonnell et al. 2009], where we investigated the effect of body motion alone. In this instance, we used a richer motion capture data set, with both male and female actor groups, and participants were given a more intuitive task. In [McDonnell et al. 2009], we presented participants with a 2 Alternative Forced Choice (2AFC) experiment, where they viewed pairs of conversations and had to choose which was the real one. This time, we asked them to make their judgement viewing one stimulus at a time, in order to determine how realistic they found both the real and synthetic conditions to be in isolation, rather than making a comparison between the two.

Twelve volunteers (10M, 2F) took part in this experiment. It was conducted in two blocks (*debates* and *dominant speaker* conversations) shown in random order. There were four synthetic conditions for the dominant speaker block and one for the debates block as outlined in Table 2. As in all succeeding experiments, participants were first shown an example of a real and synthetic conversation (from a conversation clip not used in the experiments). They were told that the real conversations depicted the body motions played back exactly as captured, while the synthetic ones were altered in some way. We did not influence their decisions by explicitly informing them that conversations were desynchronized. We used the same synthetic conversation conditions as in [McDonnell et al. 2009].

6.1.1 Results

The results for the No Audio experiment can be seen in Table 4, which contain the 'real' rating means and standard deviations across all experiments. For the most part, our results replicated the effects of [McDonnell et al. 2009]. However, unlike our previous results, we did find that conversations where three characters were animated with talker body motions were particularly unrealistic for the dominant speaker conversations (Table 4 and Figure 5, green series). These different results could be due to the fact that we conducted this experiment with a more intuitive task that yielded more sensitivity to desynchronization. There was no effect of actor group, implying that participants found it equally difficult to determine real conversations for both male and female motions.

6.2 Omni-Directional Audio Experiment

In the next experiment, we tested participants' sensitivity to audio that is matched or unmatched to body motion. Does a talker's body motion in a conversation need to match the audio? For the dominant speaker conversations, we hypothesized that when the audio did not match the body motion of the speaker (or when more talking bodies were present than the number of voices heard) participants would find these conversations most unrealistic. Will it look plausible to use audio from a different conversation as long as the body motions are synchronized? How unrealistic do the conversations appear when both bodies are desynchronized and audio is unmatched? Since in the debates block, the conversations contained more complex dynamics and [McDonnell et al. 2009] found that desynchronizing debate body motions appeared quite realistic, we postulated that unmatched audio would have a similar effect regardless of the synchronization of body motion.

To investigate these issues, nineteen new volunteers (11M, 8F) took part in this experiment. As before, it was conducted in two blocks shown in random order. The conditions we tested for both blocks were *actor group* and *AV condition*. A breakdown of the AV conditions for this experiment can be found in Table 5.

Experiment		No Audio		Omni Audio		Loc. Audio	
Block	**Cond.**	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**
Debates	Real	0.83	0.14	0.85	0.14	0.83	0.17
	$B_D A_{NM}$	0.44	0.19	0.32	0.27	0.26	0.16
	$B_S A_{NM}$	-	-	0.37	0.23	0.30	0.15
	$B_D A_{M1}$	-	-	0.52	0.20	0.36	0.21
Dom. Sp.	Real	0.84	0.14	0.87	0.096	0.78	0.13
	$B_{0T} A_{NM}$	0.33	0.27	0.33	0.27	0.31	0.29
	$B_{1T} A_{M1}$	0.40	0.25	0.65	0.25	0.38	0.21
	$B_{2T} A_{M1}$	0.22	0.16	0.19	0.21	0.21	0.26
	$B_{3T} A_{M1}$	0.10	0.13	0.05	0.11	0.10	0.16
	$B_S A_{NM}$	-	-	0.68	0.21	0.58	0.25
	$B_D A_{NM}$	-	-	0.38	0.25	0.33	0.23

Table 4: *Mean 'real' ratings and standard deviations for each experiment. Note: for No Audio experiment ignore A_x, all are A_0.*

Block	**Factor**	**AV Condition**	**Total Trials**
Debates	Real	Real ($B_S A_M$)	18
	Synth.	$B_D A_{NM}$	6
		$B_S A_{NM}$	6
		$B_D A_{M1}$	6
Dom. Sp.	Real	Real ($B_S A_M$)	36
	Synth.	$B_S A_{NM}$	6
		$B_D A_{NM}$	6
		$B_{0T} A_{NM}$	6
		$B_{1T} A_{M1}$	6
		$B_{2T} A_{M1}$	6
		$B_{3T} A_{M1}$	6

Table 5: *Experimental design for both Omni-Directional and Localized Audio experiments, showing total number of trials (50% male actors, 50% female).*

6.2.1 Results

Our statistical analysis showed no effect of *actor group*, which demonstrated that sensitivity to synthetic conversations was independent of the actors used. Also, in contrast to results found by Briton and Hall [1995], there was no effect of the sex of the participant or any interactions, implying that both males and females perceived the non-verbal behaviors similarly for both male and female virtual characters.

Our results from the dominant speaker block of this experiment reveal some interesting results for the effect of audio on participants' perception of synthetic conversations. We found that when there was one talker body motion ($B_{1T} A_{M1}$ and $B_S A_{NM}$), participants perceived these synthetic conversations to be equally real, regardless of whether the gestures of the talkers matched the audio or not (Figure 5, blue series). We also found that, as the number of talkers (depicted by body motions) grew, conversations were found to be progressively less realistic, since the audio only contained one talker. Zero talkers and desynchronized body motions were also found to be unrealistic, but less so than conditions containing more than one talker (Table 4).

For the debates, the most interesting result was that desynchronization can be masked to some degree by ensuring that one character's motion matches the audio, even if the other two characters are given random motions. We also found that when the audio was not matched to any character, synthetic conversations were considered to be equally unrealistic, regardless of whether the body motions themselves were desynchronized.

6.3 Localized Audio Baseline

We next decided to investigate the effect of 3D audio, and first tested how accurate people were at localizing conversational audio. We postulated that people would correctly identify the 3D location of directional audio for virtual conversations. Will different pitches affect the ability of participants to localize audio? Previous research has shown that higher pitched tones are easier to localize than lower pitched ones [Musicant and Butler 1984]. Our six different actors provided a reasonable range of pitch, and we hypothesized that the ability to accurately localize male and female voices would be affected by this pitch difference.

To prepare the stimuli for this experiment, we used the audio recordings from the three Behringer C-2 studio condenser microphones, which recorded the speech of each individual seperately. We used only dominant speaker conversations for this experiment, so only one speaker would be heard during any given trial. Each character was assigned a static audio source at their position from where the 3D audio was played. Each static source had position, direction, orientation and fall-off properties, which could be matched to the position and orientation of the characters in each trial.

For this experiment, each trial had three characters on screen. However, this time the characters displayed no animation and appeared in a standing pose. There was one condition of *actor voice*, in order to determine whether there were differences in participants' ability in localizing the voices of our six actors. In each trial, one actor voice was played from a random character's position on screen, and there were nine repetitions of each actor's voice.

Fifteen new volunteers took part in this experiment (9M, 6F). They were asked to listen to each conversation and decide from which character the voice was coming. They responded by pressing either the 1, 2 or 3 key to indicate the position of the chosen character on screen.

6.3.1 Results

We conducted a repeated measures ANOVA, where the within-subjects factor was *actor voice* and the between-subjects factor was *sex of participant*. We found no main effect of actor voice or sex of participant, which means that both male and female participants could localize audio for a range of actor voices. An independent-samples t-test was conducted to compare the accuracy of participants and we found that participants could correctly localize the audio for each actor voice well above chance ($p < 0.0005$ in all cases).

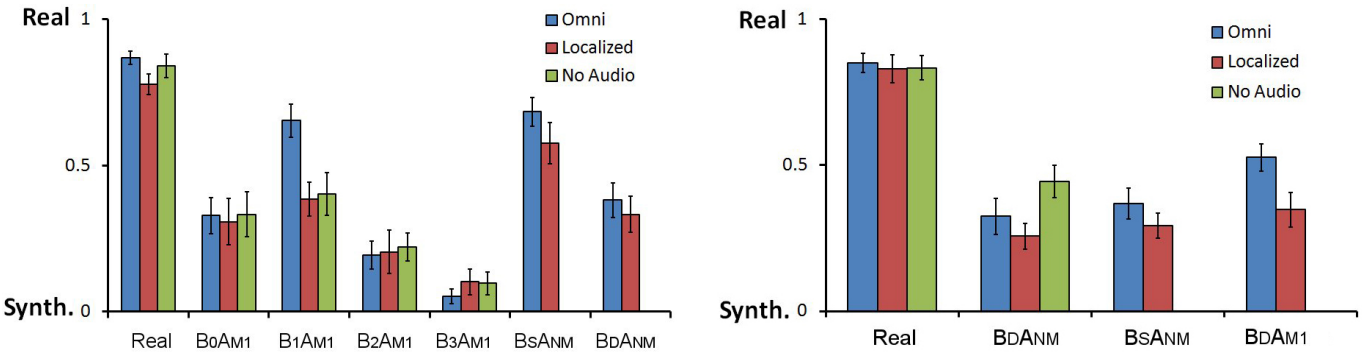

Figure 5: *Results across our 3 experiments for dominant speaker conversations (L) and debates (R), showing mean 'real' ratings and standard error bars for each condition. Note: for No Audio (green) ignore A_x label, all are A_0.*

6.4 Localized Audio Experiment

Our baseline experiment showed that people can accurately localize conversational audio. However, does the addition of a localized audio signal affect participants' ability to recognize synthetic conversations? We hypothesized that the addition of a reliable localized sound source would increase a participant's dependency on audio as a factor when making their decisions, resulting in increased recognition of synthetic conversations in most cases. However, since we found from our Omni-Directional Audio experiment that conversations where there was one speaker matching in body motion and audio ($B_D A_{M1}$ for the debates and $B_{1T} A_{M1}$ for dominant speaker conversations) were more realistic, the addition of a more reliable auditory cue might also make these conditions even more realistic.

Thirteen volunteers took part in this experiment (8M, 5F). We prepared the audio for this experiment in the same manner as in our baseline and the experiment procedure was as outlined in our Omni-Directional Audio experiment. Stimuli were presented to participants in one block each for *debate* and *dominant speaker* conversation types.

The conditions for the dominant speaker block were similar to those for the Omni-Directional Audio experiment (Table 5), except that in this experiment, audio localization was also added. All but the following two conditions contained audio localized correctly for the character speaking. The desynchronized body motion, not matched to audio condition ($B_D A_{NM}$), had audio localized at a random character position. The zero talker body motions, audio not matched condition ($B_{0T} A_{NM}$) now congruently contained no audio.

The conditions for the debates block matched those for the Omni-Directional Audio experiment. The audio localization for the debates was achieved by positioning a sound source at each character's location. For the real conversations, the three sound sources were localized at the correct characters. For the condition when one character's body motions were matched to the audio ($B_D A_{M1}$), the sound source was localized at the matching character, while the other two sound sources were randomly positioned at the two remaining characters. For the remaining conditions ($B_S A_{NM}$ and $B_D A_{NM}$), the sound sources were randomly positioned at each character.

6.4.1 Results

Our results were similar to those with the omni-directional audio, which suggests that the addition of localized audio was not any better than omni-directional audio at helping participants to better distinguish real from synthetic conversations (Figure 5, red series). We found that for the one talker body motion matched to audio condition ($B_{1T} A_{M1}$), that there was a difference in participants' responses, where they found the male characters more realistic than the females. This could be due to the level of expressivity of the talker gestures of the male characters, but warrants further investigation. As expected, we found no effect of sex of participant.

7 Discussion

We compared the results found for the No Audio, Omni-Directional Audio and Localized Audio experiments for matching conditions (see Table 3). We wished to determines the overall effect the addition of an audio signal had on participants' sensitivity to desynchronized talking bodies.

For the dominant speaker blocks, we cross analyzed matching conditions (conditions 1-5 in Figure 5 *(L)*). We found that the addition of audio (omni or localized) did not affect the perception of desynchronization of the talking bodies. The one talker body motion condition was perceived to be more realistic with omni-directional audio than with localized or no audio. We had hypothesized that the localized audio would improve the realism of this condition, but participants actually found it to be synthetic more often. This unexpected result could be due to the localized audio accelerating identification of the speaker in the trial, thereby allowing attention to be shifted to the desynchronized listener motions. Perhaps the addition of head-look modifications (i.e., ensuring that the listeners attend to the speaker) would make this condition more realistic.

Similarly, for the debates we cross analyzed matching conditions (conditions 1 and 2 in Figure 5 *(R)*). Whether the participants viewed the stimuli in the presence or absence of audio (both omni-directional and localized), did not significantly affect their responses. From this, we deduced that audio did not influence the results, even in these more complex dynamic conversations.

Following this analysis, two audio specific conditions remained between the Omni-Directional Audio and Localized Audio experiments, where audio was unmatched to the conversations (last two conditions in Figure 5 *(L)* and *(R)*). For the debates, we found that the *audio signal level* just missed significance ($F_{1,30} = 4.1224, p = 0.051$), which was possibly due to the richer audio signal allowing more accurate identification of desynchronized body motions.

Overall, we found similar trends across all experiments; with or without audio information, participants were able to recognize real conversations with high levels of accuracy. Therefore, when popu-

lating a crowd with conversing agents, for the most salient groups it would be preferable to use synchronized body motions matched to audio where possible. However, when this is not feasible (e.g., when more variety is needed from a limited database), we provide guidelines in Section 8 to help mask desynchronization and maintain realism for such scenes.

It is important to note that the stimuli for this experiment were focussed on by participants for ten seconds each, with no audio or visual distracters. Taking this into account, many of our conditions produced promising results, in particular the one talker matched to audio conditions ($B_D A_{M1}$ for debates and $B_{1T} A_{M1}$ for dominant speaker conversations). When bodies were synchronized with each other, but not matched to the audio ($B_S A_{NM}$), this was also perceived to be quite realistic. Our results could also apply to larger groups, especially for dominant speaker conversations, since attention would remain focussed on the single speaker, regardless of group size. Perhaps adding more characters for debates would only serve as distracters, due to the chaotic nature of the conversations.

8 Guidelines and Future Work

Based on our results, we can propose the following guidelines for developers who wish to add plausible conversational groups to real-time crowd systems:

- Localization of audio does not increase realism of conversing groups, so may not be worth additional implementation effort
- Audio can be plausibly assigned on-the-fly to dominant speaker conversations by ensuring appropriate talker/listener roles, regardless of audio matching or body desynchronization
- Debates will be more difficult to implement, as they will only appear sufficiently plausible if at least one talker in the group is matched to audio
- No special considerations need to be taken into account when using male or female characters

We have implemented three of the experimental conditions in our real-time crowd system (*Real*, $B_D A_{M1}$, and $B_D A_{NM}$ debates). Ten groups of conversers were placed in an open scene, amongst two hundred pedestrian characters (see Figure 1). A sound source was located at each of the characters in each of the groups using the OpenAL audio library (as in the Localized Audio experiment). The addition of conversing groups enhanced the overall realism of the simulation, especially for fixed camera viewpoints. However, with the camera in walk-through or fly-through mode (as shown in the supplemental video), we observed that setting plausible parameters for audio was non-trivial. Finding the correct levels of attenuation, directionality and gain in order to create a plausible simulation was challenging. In particular, there was a mismatch between the 3D audio effects of a large out-door scene and the small screen display. Furthermore, with the large amount of visual and auditory distracters when panning through the scene, the desynchronization in even the worst case ($B_D A_{NM}$) seemed more plausible. The effects of desynchronization and audio parametrization of conversing groups when viewed in different scenarios will be explored in future work. The effects of facial animation will also be investigated.

Using our results, we hope to build a model to generate plausible conversers based on motion captured gestures. Our aim is to create dynamic scenarios where agents can join and leave conversational groups in a plausible manner. Groups will need to be of different sizes, agent positions within a group will vary and plausible way to transition between conversations will be needed. By ensuring that the listeners attend to the talker, we should be able to further increase the realism of synthetic conversations.

9 Acknowledgements

This work was sponsored by Science Foundation Ireland as part of the Metropolis project. We wish to thank the rest of the Metropolis team, in particular Gavin Kearney and Paul McDonald.

References

BERNARD, S., THERIEN, J., MALONE, C., BEESON, S., GUBMAN, A., AND PARDO, R., 2008. Taming the Mob: Creating believable crowds in Assassin's Creed. Presented at Game Developers Conference (San Francisco, CA, Feb 18–22).

BICKMORE, T., AND CASSELL, J. 2005. Social Dialogue with Embodied Conversational Agents. In *Advances in natural multimodal dialogue systems*, 23–54.

BRITON, N. J., AND HALL, J. A. 1995. Beliefs about female and male nonverbal communication. *Sex Roles: Journal of Research 32*, 1-2, 79–90.

CASSELL, J., NAKANO, Y., BICKMORE, T., SIDNER, C., AND RICH, C. 2001. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 114–123.

CASSELL, J., VIHJÁLMSSON, H., AND BICKMORE, T. 2001. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH 2001*, 477–486.

DURUPINAR, F., ALLBECK, J., PELECHANO, N., AND BADLER, N. 2008. Creating crowd variation with the OCEAN personality model. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, 1217–1220.

2010. Edge online: Performance art. http://www.edge-online.com/magazine/performance-art.

EKMAN, P., AND FRIESEN, W. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica 1*, 1, 49–98.

EKMAN, P. 1992. An argument for basic emotions. *Cognition and Emotion 6*, 3, 169–200.

GIORGOLO, G., AND VERSTRATEN, F. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, 31–36.

GOLDIN-MEADOW, S. 2005. *Hearing gesture: How our hands help us think*. Belknap Press.

JOHANSSON, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics 14*, 2, 201–211.

KENDON, A. 1994. Do gestures communicate? A review. *Research on language and social interaction 27*, 3, 175–200.

KRAHMER, E., AND SWERTS, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language 57*, 3, 396–414.

LERNER, A., FITUSI, E., CHRYSANTHOU, Y., AND COHEN-OR, D. 2009. Fitting behaviors to pedestrian simulations. In *SCA '09: Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 199–208.

LEVINE, S., THEOBALT, C., AND KOLTUN, V. 2009. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics 28*, 5, 1–10.

MCDONNELL, R., LARKIN, M., DOBBYN, S., COLLINS, S., AND O'SULLIVAN, C. 2008. Clone attack! perception of crowd variety. *ACM Transactions on Graphics 27*, 3, 26:1–26:8.

MCDONNELL, R., ENNIS, C., DOBBYN, S., AND O'SULLIVAN, C. 2009. Talking bodies: Sensitivity to desynchronization of conversations. *ACM Transactions on Applied Perception 6*, 4, 22:1–22:8.

MCNEILL, D. 1996. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

MUSICANT, A. D., AND BUTLER, R. A. 1984. The influence of pinnae-based spectral cues on sound localization. *The Journal of the Acoustical Society of America 75*, 4, 1195–1200.

NEFF, M., KIPP, M., ALBRECHT, I., AND SEIDEL, H. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics 27*, 1, 1–24.

PETERS, C., AND ENNIS, C. 2009. Modeling groups of plausible virtual pedestrians. *IEEE Computer Graphics and Applications 29*, 4, 54–63.

ROSE, D., AND CLARKE, T. J. 2009. Look who's talking: Visual detection of speech from whole-body biological motion cues during emotive interpersonal conversation. *Perception 38*, 1, 153–156.

ROUSE, R. 1998. Embrace your limitations – cut-scenes in computer games. *ACM SIGGRAPH Computer Graphics 32*, 4, 7–9.

VILHJÁLMSSON, H., AND CASSELL, J. 1998. Bodychat: autonomous communicative behaviors in avatars. In *Proceedings of the second international conference on Autonomous agents*, 269–276.