Animating synthetic dyadic conversations with variations based on context and agent attributes

Libo Sun, Alexander Shoulson, Pengfei Huang, Nicole Nelson, Wenhu Qin, Ani Nenkova, Norman I. Badler*

Abstract

Conversations between two people are ubiquitous in many inhabited contexts. The kinds of conversations that occur depend on several factors, including the times and locations of the participating agents, the spatial relationship between the agents, and the type of conversation in which they are engaged. The statistical distribution of dyadic conversations among a population of agents will therefore depend on these factors. In addition, the conversation types, flow, and duration will depend on agent attributes such as interpersonal relationships, emotional state, personal priorities, and socio-cultural proxemics. We present a framework for distributing conversations among virtual embodied agents in a real-time simulation. In order to avoid generating actual language dialogues, we express variations in the conversational flow using behavior trees implementing a set of conversation archetypes. The flow of these behavior trees depends in part on the agents' attributes and progresses based on parametrically estimated transitional probabilities. Based on the participating agents' state, a "smart event" model steers the interchange to different possible outcomes as it executes. Example behavior trees are developed for two conversation archetypes: buyer-seller negotiations and simple question-answering; the model can be readily extended to others. Since the conversation archetype is known to participating agents, they can animate their gestures appropriate to their conversational state. The resulting animated conversations demonstrate reasonable variety and variability within the environmental context.

KEYWORDS
conversation model, agent attributes, smart event, behavior trees, crowd simulation

***Correspondence**
Department of Computer and Information Science, University of Pennsylvania,
Philadelphia, PA 19104-6389, USA

1. INTRODUCTION

Research on realistic behaviors for virtual agents includes basic individual human acts such as walking, running and looking. Conversations are a very important component of social interactions among agents. The animation of conversations between agents should not only increase the realism of a virtual environment, but should also improve a viewer's sense of presence by having the inhabitants appear to be socially engaged in each other and their own surroundings.

Some research on human conversation is concerned with what the agents are saying -- the words, meanings and goals of the intended conversation. Agents decide what should be done next according to the meanings of the words, such as continuing the conversation, changing the topic, or ending the conversation. Other research focuses on facial expressions, head motions and eye gaze, since faces are an important channel of communication with several crucial functions such as controlling the flow of conversation, producing speech, emphasizing what is being said, providing backchannel feedback, controlling turn-taking, and so on. Likewise, considerable attention has been paid to arm or body gestures during conversation, both to augment emotional states and to support or accent linguistic utterances.

While these efforts have been instrumental in producing multimodal animated conversations, they are heavily weighted toward producing animated agents who are a human subject's direct and interactive interlocutor: someone highly visible or even dominant in the scene. For computational expediency in situations where the conversational detail is less critical, where the character's main purpose is just looking like it belongs in the environment, or where the agent is just "part of a crowd", conversational motions are often just random gestures, pre-scripted actions, or motion clip playbacks. These can be monotonous if examined more closely (e.g., by allowing the subject to wander at will through the virtual crowd). We would not want all of the conversations to look the same (or random) for every pair of agents, rather, conversations should occur in appropriate places and with types and frequencies appropriate to the surrounding context. For example, conversations:

- In a restaurant may mostly occur among seated customers and occasionally with the waiters.
- On an urban street corner may be sparse, occurring among people standing or pairs of people walking.
- In a crowded marketplace may often involve negotiation or friendly chats.

Our goal, therefore, is to simulate various interactive (unscripted) conversation scenarios at low computational overhead while allowing environment context and agent attributes to guide and affect the evolution of their conversations. This presents several fundamental problems we need to address:

- What *conversational situations* are likely to occur? This problem yields to a relatively simple case analysis based on how two people can interact verbally.
- What dyadic *conversational archetypes* exist? This is less easy to quantify. According to the intended purpose, conversation archetypes may be: debates, instructions, negotiations, task-orientated interactions, media interviews, casual chats, formal meetings, buyer-seller negotiations and so on [28]. Rather than try to understand the intended purpose or goals a pair of agents may have

(especially if they are “extras”: background agents with no specific simulation existence other than to populate the scene with situationally appropriate characters), we adopt a different approach that looks at the likelihood and frequency of certain conversational archetypes occurring in specific contexts.

- What *initiates* a conversational event between agents? We approach this through a “Smart Event” for dyadic conversation and conversation distribution statistics dependent on time of day and locale. Information on such distributions must be invented or else gathered from empirical observations.
- How does the conversation animation *evolve* to illustrate specific yet varying instances of a general conversational archetype? Our approach to this uses behavior trees that access agent attributes, relationships, and emotional states and trigger gesture motions on each character.

The key contributions in our approach can be summarized as follows:

- ➢ A simple classification of dyadic conversations into fourteen situational types dependent on the spatial relationship between two agents;
- ➢ Implementation of selected dyadic conversation archetypes as behavior trees;
- ➢ Temporal and iterative variations among simulated instances of a conversational archetype;
- ➢ Utilization of agent attributes, relationships and emotions that may be used to influence animations.

The discussion is organized as follows: in the next section, we review related work in the role of conversations in crowd simulation, computational models for dyadic conversations and conversational archetypes. Section 3 describes the framework architecture of our dyadic conversation model. In Section 4, we present the triggering of dyadic conversation between pairs of embodied agents. Section 5 describes how to initialize a dyadic conversation. In Section 6, we introduce the animation of the dyadic conversation model. Section 7 illustrates examples of conversational scenarios. Finally, Section 8 draws conclusions and discusses future work.

2. RELATED WORK

Crowd simulation research covers many tangible aspects of human locomotive behavior such as the realism of the walking motion itself, collision avoidance, navigation and local interactions between agents. Agents have been constructed with varying degrees of perception, memory, planning, attention, psychology and emotion. Agents can react to other agents and their environment to avoid collisions and reach assigned goals [1]. Some simulations allow contextual behaviors appropriate to visiting a train station [2], a museum [3], or an ancient city [4]. Decision networks [5] and constraints [6] are used to focus on more locally contextual meaningful actions. Stocker *et al*. [7] use an extension of Kallmann’s smart objects [8] called “smart events" to efficiently control agent behavior reactions to situations that are meaningful to them. Social aspects may be included for more realistic human interaction. Agents can join in or separate from a group according to their beliefs; and they can walk together towards the same goals [9]. Composite agents [10] are integrated to model emergent crowd behaviors that arise when humans respond to various social and psychological factors, such as aggression, social priority, authority, protection, guidance and so on. "Situation" agents [11] can mediate specific interaction circumstances to avoid deadlocks or awkward avoidance paths.

Virtually all these systems, however, lack specific "conversation" behaviors that might aid in forming a realistic social context for the crowd. Although some "greeting" behavior could be generated when two agents are close enough and know each other; it is very simplistic and would look the same between any two agents in any situation. More realistic conversations with variations are needed to increase the realism of video game crowds, particularly urban environment games such as *Grand Theft Auto* or *The Sims* [12]. More specific efforts to build conversation behaviors lie outside the crowd simulation work, and are often called "Embodied Conversational Agents" (ECA) [13]. These efforts address all visible aspects of conversation such as gesture, facial expression, eye gaze, turn-taking, backchannel signals, and of course, language and expressive (emotional) content. The focus of ECA research has generally been directed toward developing computer animated agents that interact (face-to-face) with a human participant. Whether designed for internet services, tutoring systems, virtual reality experiences or games, an ECA rarely engages in conversations with other ECAs.

In general, there are two main approaches to create computational models for dyadic conversations: through linguistics or through animation. An example of the linguistics approach is given by Moulin and Rousseau [14], who discuss a conversation model that acts like a finite-state machine bound to two conversational agents. The model focuses on three levels: the lowest is "communication" such as maintaining turn-taking, the middle is "conceptual" comprising topic sequences and concept transfer, and the highest is "social" involving the management and respect of social relationships between agents. Cassell *et al*. [15] present a Behavior Expression Animation Toolkit (BEAT) which allows animators to input typed text that they wish to be spoken by an animated human figure. BEAT outputs appropriate and synchronized nonverbal behaviors and synthesized speech in a form that can be sent to a number of different animation systems. A Language Tagging Module is responsible for annotating input text with the linguistic and contextual information that allows successful nonverbal behavior assignment and scheduling so that the gestures are appropriate and consistent with what has been said. By integrating BEAT, O'Sullivan *et al*. [16] describe ongoing development of a framework for adaptive level of detail for human animation, which incorporates levels of detail for not only geometry and motion, but also includes a complexity gradient for natural behavior, both conversational and social. Level of detail Artificial Intelligence (LODAI) is facilitated by a process of role-passing, where agents are given the ability to take on different roles depending on the situation they are in.

The second approach is from the animation perspective. Since there are some situations where the language content is unknown and unperceivable (e.g., it may not be audible over background noise, it may be in a foreign language, or the agents themselves are just "background" characters in a given setting), at least a visual simulation should create the appearance of a relevant conversation event. Jan and Traum [17] give a typical example to simulate conversations ignoring linguistic and speech components. They describe an algorithm that generates believable behaviors for background characters involved in conversation and that supports dynamic changes to conversation group structure. Furthermore, a variety of markup languages have been proposed for behavior planning of animated agents, including conversation behaviors. The most sophisticated are BML and MPML3D. The Behavior Markup Language (BML) refers to a broad effort in controlling communicative channels of virtual agents [18] [19]. The BML project aims to develop a representation framework for describing both nonverbal and verbal real-time behavior that is independent of the particular graphical realization. BML is a standard XML-based interface between behavior planners and behavior realizers. MPML3D (Multimodal Presentation Markup Language 3D) [20] is an XML-based scripting language for controlling the verbal and non-verbal behavior of 3D agents. MPML3D can support interaction-rich scenarios with reactive agents in Second Life and OpenSim. In both languages, the nonverbal behaviors select predefined gestures and facial expressions that are specified, triggered and synchronized with speech. Taking BML as the input, SmartBody [21], an open source modular framework, can realize behavior scheduling, synchronization and animation. Jan *et al*. [22] have presented a model for simulating cultural differences in the conversational behavior of virtual agents. The model provides parameters for differences in proxemics, gaze and overlap in turn taking. Levine *et al*. [23] present a system that generates gestural body animations automatically using speech, rather than text input. A gesture generation system presented by Neff *et al*. [24] can recreate a specific speaker's gesturing style. Pedica and Vilhjálmsson [25] have pointed out that the addition of territorial behaviors can increase believability of a virtual conversant. Jan and Traum [26] present an algorithm to control the positioning and movement behavior of autonomous agents in dynamic conversations based on a social force model. Hostetler [27] also addresses the problem of positioning and orienting agents in a conversational group.

Conversations have many types depending on application requirements. According to the intended purpose, an exchange may be classified into archetypical categories such as debates, instructions, negotiations, task-orientated interactions, media interviews, casual chats, task-oriented communication in noisy environments, formal meetings, buyer-seller negotiation and so on [28]. They can be further distinguished by duration, the participants' attributes, and performatives based on the relationship between agents (age, familiarity, authority level, knowledge, culture background, and so on) [29]. Although there are many kinds of conversation archetypes, we just select specific representative archetypes to illustrate and animate our framework: simple asking-answering, friendly chatting, bargaining and arguing. Our framework can be extended to accommodate other archetypes as needed. Furthermore, we would like a lightweight simulation model so that agents may initiate and end conversations in ways that can be biased in real-time by their social roles and attributes, culture, personality and possible realms of disagreement.

Here we focus on creating a framework for modeling dyadic conversation simulation between two embodied agents situated in a larger setting of other agents and a spatial-temporal context. In this model, we ignore any linguistic and speech components and leave aside facial animation details. The latter may be added through a number of established facial animation models. What remains are head, arm and body motions. These are mostly sufficient for the background characters in a simulation, especially in a crowd [30, 31].

3. FRAMEWORK OF A DYADIC CONVERSATION MODEL

As Figure 1 shows, dyadic conversation model comprises three parts: conversation triggering which is responsible for starting a conversation; conversation initialization which is responsible for computing the relevant conversation parameters for the involved agents; and conversation animation which is responsible for portraying some realistic and diverse agent behaviors.

First, a conversation smart event triggers a conversation for two agents according to the time, the environment context and the number of conversations desired in the scenario. A triggered conversation will be realized when the two agents can approach to each other; conversely the conversation cannot be realized if two agents cannot get close (e.g., something blocks them) or their distance separation is outside the threshold for a conversation. When the conversation is successfully triggered, the conversation archetype and the situation type are determined according to the environment context, agent attributes, estimated probability and the relationship between conversation archetype and situation type. The other conversation parameters are computed based on the chosen types, including the conversation outcome if it exists, the number of iterations for the whole conversation, the duration of each agent’s turn and the proxemics between the two agents engaging in the conversation. As a result, all the conversation parameters are well-defined and initialized for execution, so a behavior tree is constructed to evolve the specific conversation. This behavior tree manages the entire conversation event including bringing the two agents into the correct proxemics positions, alternating the turns, generating appropriate gestures, terminating the conversation and finally releasing the agents from this event allowing them to execute their default behavior or to participate in other activities.

Each conversation archetype is built as a major branch of the behavior tree. The conversation archetype, the situation type and related conversation parameters impact and constrain the conversation flow. Diverse conversations are generated and even the conversations with the same situation type and conversation archetype can show variations due to different possible conversation outcomes, different agent emotions and gestures. Moreover, the actions are stored in the nodes of the behavior tree in a manner consistent with that of a smart event. The agents select the most appropriate actions to execute so that they can update their states and animate their head, arm and body motions. In addition, in order to show more interesting conversation scenarios, one conversation archetype can change into another one based on the current situation such as an agent's emotional state, inter-agent relationships and environment context. If the conversation archetype actually changes, the conversation parameters are re-computed before the transition to guarantee that the conversation is executed successfully and reasonably.

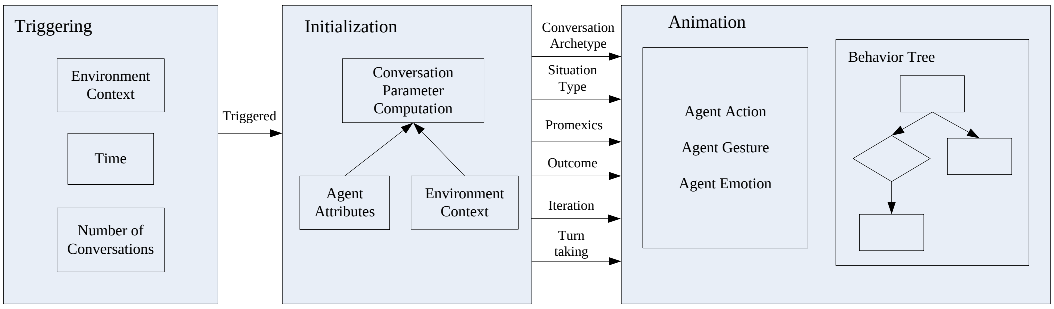


Figure 1. The framework of the dyadic conversation model

4. TRIGGERING OF SYNTHETIC DYADIC CONVERSATIONS

4.1 Conversation Smart Event

Based on the smart event concept [7], the conversation smart event triggers the conversation, computes the relevant conversation parameters, informs two agents involved in a conversation about the beginning of the conversation, possible action sets for agents to select and execute according to the current situation, and the ending of the conversation. What the agents need to do is to select the most appropriate actions according to its attributes and current emotion. Furthermore, the conversation smart event is responsible for checking whether or not the selected agent is in another conversation already. If the agent is in a conversation, then the conversation smart event bypasses this agent. If the agent is not in a conversation, the conversation smart event will consider it as a potential target for a new conversation and informs it to begin a conversation if there are other agents available.

The conversation can be invoked in the following ways:

- Initiated by a Director: the Director specifies two agents to start a conversation whenever desired, where the Director is a process responsible for selecting which events to execute and what agents to involve in those events. The Director can be a human operator (such as a player in a game) or an automated procedure [32].
- Requested by Agent: two agents can request a conversation event when they have a desire to begin a conversation, which will be explained in Section 5.

4.2 The Prerequisite for Triggering a Conversation

Edward Hall (1969) identified four distances or zones that humans set in their daily interactions. These zones include the intimate zone, the personal zone, the social zone, and the public zone (shown in Figure 2). The intimate zone begins with skin surface and goes out about 18 inches, so that people who are emotionally very close will converse at this distance. The personal zone ranges from about 18 inches to approximately 4 feet. Interactions at this distance may still be reasonably close. The social zone ranges from about 4 feet to about 12 feet. Business communications are frequently exchanged in this zone. The public zone runs outward from 12 feet and public speakers often use this distance when they give a speech. For the conversation types we address here conversations start when two agents are at least within their respective social zones. If two agents are any farther away, e.g. they are in the public zone, and then the conversation smart event does not consider them as potential targets. Furthermore, the inter-agent relationship, the emotional state the agents are in and how many conversations are currently taking place affect the probability of triggering the conversation too. The conversation smart event does not consider agents already engaged in other (non-default) events as potential participants. The probability of triggering the conversation depends on the following aspects:

- Two agents are close enough: the conversation is more likely to happen when there are only a few conversations in the scenario;
- Two agents are not close: the conversation can mostly happen only when there are no obstacles or others which prevent them to get close to each other to start a conversation. If the conversation is requested by agent, the inter-agent relationship between them affects the conversation triggering probability, while if a “Director” initiates the conversation the inter-agent relationship has no effect on the conversation probability.. Thus when two agents know each other or they are good friends or even more intimate, the conversation is more likely to happen. Furthermore, if two agents are seated, the probability is lower than that of two agents who are are standing or walking. If one is seated and the other is standing or walking, the probability of starting a conversation is even lower.

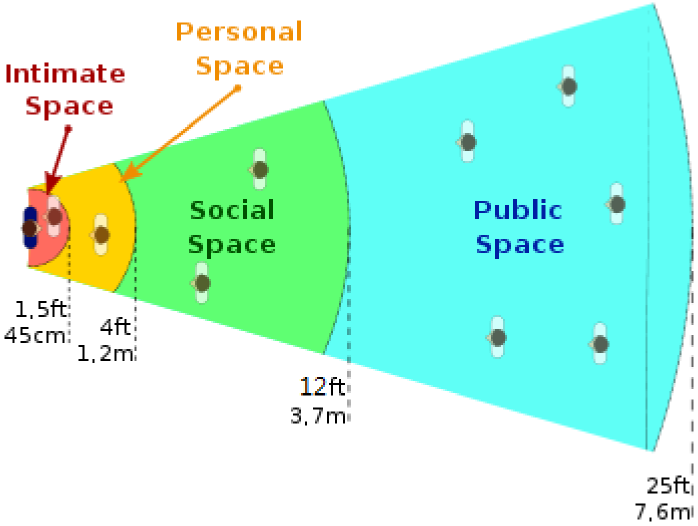


Figure 2. Zone distances

The formula which computes the probability to trigger a new conversation can be given as follows:

$$C_T = \begin{cases} (pM + \dfrac{q}{N} + rR)/3 & \text{required by "Agent"} \\ (pM + \dfrac{q}{N})/2 & \text{triggered by "Director"} \end{cases}$$

where C_T is the probability of the conversation triggering;

M represents whether two agents can approach to each. Its value is 1 if two agents can approach to each other making their distance less than some threshold; otherwise its value is 0;

N represents the number of conversation already in the scenario. The greater the current number of conversations, the lower the probability that a new conversation will be triggered;

R represents the relationship between two agents involved in the conversation. Its value is in the range of [0, 1], where 0 means that two agents are total strangers and 1 means that two agents are very intimate. Note that the value of R is set to 1 if an agent needs to ask a question of another even though the two agents do not know each other;

p, q, r are corresponding weights for each item and are in the range of [0,1] respectively. Their differences are not significant unless the influence of some factor needs to be specifically emphasized.

When the computed value of C_T is larger than 0.4, a conversation is triggered, otherwise, no conversation is triggered.

5. CONTEXTUAL SELECTION AND INITIALIZATION OF SYNTHETIC DYADIC CONVERSATIONS

5.1 Conversation Archetypes and the Transitions Among Them

Although generic conversations must be triggered, it is essential to know what sorts of conversations are possible, which ones are desirable or relevant in the context of participating agent attributes, and how spatio-temporal factors such as location and the time of day influence conversation choices and probabilities. We must elaborate these conversational features next.

We mainly consider four conversation archetypes: *simple asking-answering*, *friendly chatting*, *bargaining* and *arguing*. They are mutually exclusive and one conversation archetype can transition to another archetype based on the current situation and an estimated probability. The transitions between these four conversation archetypes are shown in Figure 3: two agents can greet each other (simple asking-answering) and if they are happy and do not have other events to attend to in a short time, they can begin friendly chatting until the conversation is over.

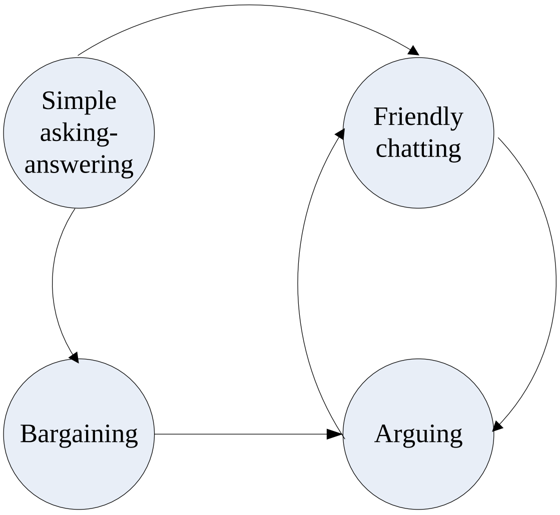


Figure 3. Transitions between four conversation archetypes

5.2 Situation Type

Conversations can occur in many ways between two agents who are standing, sitting, walking or talking on the phone. We define the postural state of two agents when starting a conversation as the *situation* type. A simple case analysis can be based on {body posture agent A} × {body posture agent B} × {facing direction}. More specifically, the combinations can be quantized as {standing, sitting, walking} ×

{standing, sitting, walking} × {facing, offset, parallel}. In addition, we consider the case of one agent using a cell phone to converse with an "invisible" second agent. This yields 14 situation types:

1. Both standing, facing each other;
2. Both standing, facing about 45° from forward toward the other agent;
3. Both standing, facing the same direction;
4. Both walking (or jogging) together (side-by-side);
5. Both walking (or jogging) toward each other and talking very briefly (as in a greeting) "en passant";
6. Both seated, facing each other;
7. Both seated, sitting next to each other facing the same direction (e.g., on a bench);
8. Both seated, facing about 45° from forward toward the other agent;
9. One seated and the other standing;
10. One seated and the other walking;
11. One standing and the other walking;
12. One agent is walking using a cell phone;
13. One agent is seated using a cell phone;
14. One agent is standing using a cell phone.

These situation types are useful to distinguish the sorts of body, head and arm motions that must be animated on the agent models. For facing directions, the angle between the agent's bodies will be dictated by the proxemics of their culture and that will in turn affect the head orientation. For cell phone use, the occupied hand will not be engaged in gestures at all as it will be used to hold the phone to the ear or in front of the mouth. Finally, the length of a conversation will be very dependent on the time during which the participating agents are close enough, so that mixed locomotion situations are apt to produce very abbreviated verbal interchanges. (If both moving agents stop it then becomes a different situation type, e.g., "en passant" may transition to a standing conversation).

5.3 The Relationship between Conversation Archetype and Situation Type

The conversation archetype and the situation type are correlated with each other. For each conversation archetype, not all situation types are suitable, and conversely, for each situation type, not all conversation archetypes are appropriate. As a result, when one of these two is determined, the other one should be statistically selected from the relevant possibilities. For example, when the situation type is "Both standing, facing the same direction", the conversation archetype can only be selected from the set "Simple asking-answering" and "Friendly chatting" – "Bargaining" and "Arguing" are not possible. The related situation types and conversation archetypes are listed in Table 1 where the situation type number corresponds with the list in Section 5.2.

Table 1. The relationship between conversation archetype and situation type

	Number to Represent Situation Type

Conversation Archetype	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Simple asking-answering	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y
Friendly chatting	Y	Y	Y	Y	N	Y	Y	Y	N	N	N	Y	Y	Y
Bargaining	Y	N	N	N	N	N	N	N	N	N	N	N	N	N
Arguing	Y	N	N	N	N	Y	N	N	N	N	N	Y	Y	Y

5.4 Determination of the Conversation Archetype and Situation Type

We first determine the conversation archetype and then the situation type. The environment context and agent attributes act as cascaded filters to the conversation archetypes; that is, the distribution of the four conversation archetypes depends on the environmental and temporal context and agent attributes such as their inter-agent relationship and emotions. We compute the probability of each conversation archetype with the following formulas; the conversation archetype with highest probability is selected for the current conversation. For notational simplicity we will refer to the four conversational archetypes by numbers: Simple asking-answering=1; Friendly chatting=2; Bargaining=3; Arguing=4. We will describe each influential term of the formulas in detail in the following sub-sections. We separate conversations into two cases: one with both agents visible and the other with a solitary agent on a cell phone.

➢ Two agents involved in a conversation are both visible in the scenario:

$$P_{CT2_i} = aD_i + bR + cE + dS + eEP_i$$

where P_{CT2_i} is the probability of the i^{th} conversation archetype $i = 1,2,3,4$;

D_i represents the distribution of the i^{th} conversation archetype in some environment context. Its value is in the range $[0,1]$;

R represents the relationship between two agents involved in the conversation. Its value is in the range $[0,1]$, where 0 means that two agents are total strangers and 1 means that two agents are very intimate;

E is determined by the emotion of two agents together. E_i represents a simple one-dimensional “happiness” model of the emotion of $agent_i$ for $i = 1,2$. The value of E_i is in the range $[0,1]$, where the higher the value, the happier the agent. The value of E is given by the following formula where *min* and *max* compute the smallest and largest value respectively:

$$E=\begin{cases}\min(E_1,E_2) & \text{if } \min(E_1,E_2)<0.3\\ \dfrac{E_1+E_2}{2} & \text{if } 0.3\le\min(E_1,E_2)<\max(E_1,E_2)\le 0.7\\ \max(E_1,E_2) & \text{if } 0.3\le\min(E_1,E_2)<0.7<\max(E_1,E_2)\end{cases}$$

S represents whether one or both agents have a required event scheduled (anticipated) in a short time. Its value is 0 or 1, where 0 means that neither of two agents have a required event soon and 1 means that at least one agent has a required event soon;

EP_i represents the estimated probability for the i^{th} conversation archetype. Its value is in the range $[0,1]$ and is randomly determined and regenerated for every conversation archetype computation. That is, EP_i guarantees that different conversation archetypes can be obtained and therefore our approach can show variations among the conversations.

For certain conversations occurring in some environment context between two agents, D and R are static while E and S can change as time passes.

a,b,c,d and e are weights in the range $[0,1]$ respectively. Since some variables, including R , E and S , influence each conversation archetype in different ways, the weights of these terms vary, such as shown in the following tables. Note that the precise numbers are less important than displaying the different influences of these terms on each conversation archetype. While we simply estimated these values, they could be set by observing large sets of actual human behaviors in an analogous environment.

The value of b is shown in Table 2. When the inter-agent relationships are different, the value of b for each conversation archetype is correspondingly different.

Table 2. The value of b for different inter-agent relationships

Relationship	**Conversation Archetype**			
	Simple asking-answering	Friendly chatting	Bargaining	Arguing
family members	0.6	1	0	0
friends	0.75	1	0	0.1
coworkers or classmates	1	1	0	0.1
buyer-seller	1	0	0.8	0.2
strangers	1	0.4	0	0.1

The value of c is shown in Table 3: its value changes according to an agent's changing emotions.

Table 3. The value of c with different E

	Conversation Archetype			
E	Simple asking-answering	Friendly chatting	Bargaining	Arguing
$E < 0.3$	1	0.1	0.1	0.4
$0.3 \le E \le 0.7$	1	1	1	1
$E > 0.7$	0.6	1	0.6	0.6

The value of d is shown in Table 4. It guarantees that the conversation archetype is simple asking-answering when at least one agent has a required event soon.

Table 4. The value of d with different S

	Conversation Archetype			
Scheduled	Simple asking-answering	Friendly chatting	Bargaining	Arguing
$S = 1$	1	0	0	0
$S = 0$	1	1	1	1

- One agent is using a cell phone and the other agent is invisible in the simulated scenario:

$$P_{CT1_i} = a_1 D_i + b_1 R + e_1 EP_i$$

where P_{CT1_i} is the probability of i^{th} conversation archetype; other terms are the same as the first case and a_1, b_1 and e_1 are also in the range $[0,1]$. The value of b_1 is shown in Table 5 which displays the different influences for each conversation archetype.

Table 5. The value of b_1

	Conversation Archetype			
Relationship	Simple asking-answering	Friendly chatting	Bargaining	Arguing
family members	0.6	1	0	0
friends	0.75	1	0	0.1
coworkers or classmates	1	1	0	0.1
strangers	1	0	0	0

After the conversation archetype has been determined, the situation type is decided based on the rules described in the next Section.

5.4.1 Environment Context

The environment context where the conversation occurs influences and constrains the situation types. Take the street, for example: the probability of both agents standing facing each other or both walking (or jogging) together (side-by-side) is much higher than that of both seated facing each other or both seated next to one another facing the same direction. Conversely, in a restaurant the probability of both agents being seated facing each other or both seated next to each other facing the same direction is higher than that of both standing facing each other, walking or standing using a cell phone or both walking (or jogging) together (side-by-side). The distribution of the situation types in different environment contexts can be obtained by observing analogous situations in real life. Table 6 shows a possible situation probability distribution where the number in each corresponding item represents the distribution percentage (among all conversations) of one situation type in each environment context.

Table 6. The distribution of situation types with different environment contexts

	Number to Represent Situation Type													
Environment	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Street	20	20	5	15	15	--	--	--	--	--	10	10	--	5
Office	5	5	--	--	--	25	10	10	15	5	5	5	10	5
Restaurant	5	5	2	5	3	35	20	5	5	3	2	3	5	2
Crossroad	5	5	10	30	20	--	--	--	--	--	10	10	--	10
Marketplace	30	5	5	10	10	10	5	5	2	3	2	5	3	5
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Moreover, the environment the agents inhabit is an important factor that influences the distribution of each conversation archetype. For example, if the agents are walking along the street and if they start a conversation, the possibility of friendly chatting and simple asking-answering may be higher than bargaining or arguing. However, if the agents are shopping in the store, the distribution of simple asking-answering and bargaining may be a little higher than just friendly chatting. The empirical distributions of four conversation archetypes are shown in Table 7; of course the precise numbers are less important than establishing some differences in distribution according to the environment.

Table 7. The distribution of four conversation archetypes with different environment contexts

	Conversation Archetype			
Environment	Simple asking-answering	Friendly chatting	Bargaining	Arguing
Street	30	60	5	5
Shopping Mall	30	20	45	5
Marketplace	25	15	50	10
Restaurant	25	65	--	10
Crossroad	70	25	--	5

…	…	…	…	…

5.4.2 Agent Attributes

Agents are modeled with intrinsic personality types using the five factor model and, in addition, have a set of personal and socio-cultural attributes. We classify agent attributes into five classes according to their influences on conversation, shown in Table 8. The first class includes the attributes which are static for the sake of the simulation, such as age, gender, personality, culture and so on. The second class considers temporal factors, such as calendar (time, day, date), which influence the duration of the conversation. The third class includes transient relationships to other specific people, such as friends, family members, co-workers, buyer-seller and so on. The fourth class includes the attributes which are dynamic for each agent such as emotion and mood. Changes to the fourth attribute class have the most influence on the actions and gestures of the agent during the conversation. For example, an unhappy agent may animate with a more drooping, resigned posture than a happy agent. The fifth class considers other current behaviors and constraints, such as a hand occupied by a cell phone or coffee cup so that it will not be engaged in gestures at all.

Table 8. Five classes of agent attributes

Attributes	Examples
Static	Age, gender, personality, culture
Temporal	Calendar(time, day, date)
Relational	Friends, family members, co-workers, seller, customer, supervisor, teacher-student
Dynamic	Emotion, mood
Behavior and constraint	Hands are occupied with a cell phone or coffee

5.5 Further Conversation Parameter Computation

After the conversation archetype and situation type are determined, other related conversation parameters are computed. The proximity between two agents is determined by the agent attributes, especially by the inter-agent relationship and culture. We use Hall's theory and inter-agent relationship to set the proxemics:

- If two agents are good friends, the personal zone is used;
- If two agents are just acquaintances, the social zone is adopted;
- If two agents are lovers or family members, the intimate zone is used for embracing, touching or whispering;

Furthermore, the distance between two agents is also dependent on the two agents' cultures and social status. In general, the person who is high in social status prefers and needs more space than the person who is low in social status. In general, e.g., Arabs, Italians, Latin-Americans and Africans speak at a closer range than Americans, British and Germans. The angle between their respective forward orientations lies within a range of [180°± 90°]. The value of the angle is dependent on the situation type. That is, when the situation types are different, the angles between their respective forward orientations are different. For example, when the situation type is both walking (or jogging) together (side by side), their respective forward orientations are parallel which means they are facing the same direction; while when the situation type is both seated facing each other, the angle between their respective forward orientations is 180°.When the situation type is walking, standing or seated using a cell phone, only one agent is visible so that we do not need to consider proxemics, though cell phone use in close proximity to others may be undesirable in the first place..

The other conversation parameters, including outcome, iterations and turn taking, are mainly influenced by the conversation archetype. For example, when the conversation archetype is simple asking-answering, iterations are few and the duration one agent is taking for a turn is short. The outcome is only effective when the conversation occurs between a buyer and a seller and it represents whether the transaction is successful or not. Parameters are further determined by the agent attributes, including any relationship between them, their schedules and their emotional states. For example, if two agents are good friends, are very happy when greeting one another, and do not have scheduled work, then the iterations may be many and the duration for each turn may be relatively long. Conversely, if the two agents are strangers, one of them is very unhappy, or one has scheduled work, then the conversation may be very short with quick iterations. Table 9 summarizes the relationship between conversation archetype and outcome, iteration and turn taking.

Table 9. The relationship between conversation parameters and conversation type

	Conversation Archetype			
Computation Parameters	Simple asking-answering	Friendly chatting	Bargaining	Arguing

outcome			success;
			failure;	

iteration	few;	few;
		many;
			few;	few;

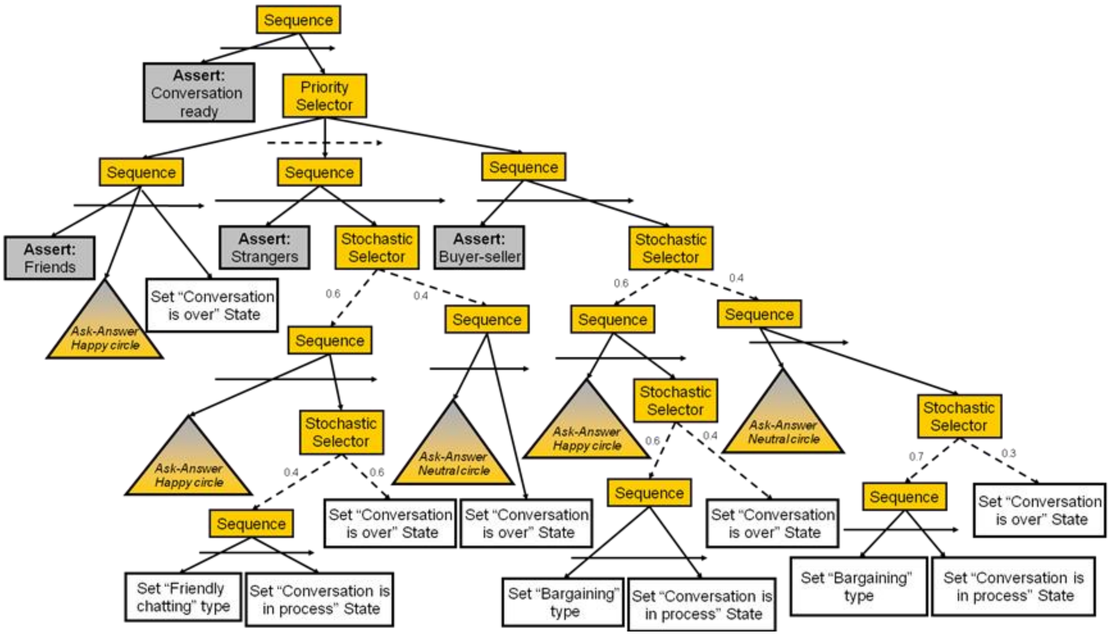
turn taking	short;	short;
		long;
			short;	short;

6. ANIMATION OF SYNTHETIC DYADIC CONVERSATIONS

6.1 Behavior Tree Design

We construct a behavior tree to simulate dyadic conversations. The behavior tree is designed to be very general since it can deal with every conversation archetype. It has many branches from its root, where each branch is for one conversation archetype. So far we have implemented four archetypes though more branches can be added easily when other conversation archetypes are needed. Since it is very appealing for its reusability, we can build the behavior trees very quickly due to the similarities of sub-trees among different conversation archetypes.

Since one conversation archetype can transition into another archetype during the same conversation, it is necessary for the behavior tree to guarantee that the transition between two different archetypes should be successful. Therefore, some variables are needed to record the current phase of the conversation and the current conversation archetype. The variable recording the conversation phase includes three possible values: “Ready”, “Process” and “Over”. “Ready” means the initialization of conversation has been finished and the conversation is going to begin; “Process” means one conversation archetype is finished and it transitions to another conversation archetype which is going to start; “Over” means the conversation ends. Figure 4 shows the design of the sub-tree for “simple asking-answering” and the transitions from the “simple asking-answering” to the “friendly chatting” and the “bargaining” conversation archetypes. Each sub-tree is expanded until each node in every sub-tree is either an assertion or an action. If the behavior tree assertion finds that the conversation archetype has changed, it will execute the branch of the new conversation archetype until it transitions to another archetype or the conversation ends.



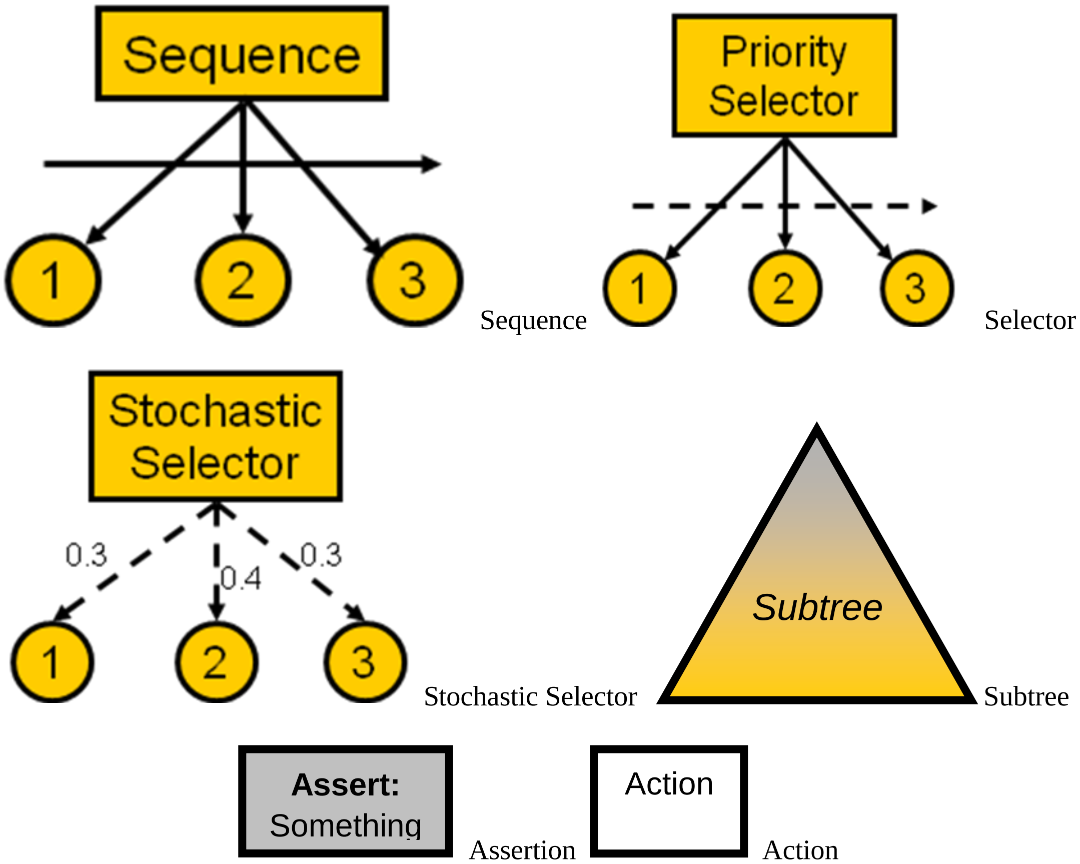


Figure 4. The “Simple asking-answering” sub-tree design

6.2 Gesture, Body Postures and Emotions

A gesture is a form of non-verbal communication in which visible bodily actions communicate particular messages, either in place of speech or together and in parallel with spoken words. The gestures are so much a part of speaking that one is often unaware of them, but if we look around and watch someone talking in informal terms, we are likely to see the hands and arms in motion. Here we are ignoring any linguistic and speech components, so we adopt just gestures and body posture to reflect conversation archetype, nominal gestural variations and the change of the agent’s emotions. For simplicity, we classify the emotion into four kinds: happy, neutral, sad and angry. The behavior tree can output suitable BML tags during execution, which guarantees appropriate gesture choices and their synchronization for both participating agents.

7. EXAMPLES

To illustrate the framework, we construct conversation smart events occurring in a variety of environments. Figure 5 shows conversation distributions are different in different environments. The quadruple (ID, CT, ST, R) describes a conversation with identifier ID, conversation archetype CT, situation type ST and the relationship R between two agents involved in the conversation. The value of ID can be {1, 2 …}; the value of CT is {1, 2, 3, 4} corresponding to {simple asking-answering, friendly chatting, bargaining, arguing}; the value of ST is {1,2,…14} corresponding to the list described in Section 5.2; the value of R is {1, 2, …7} corresponding to {strangers, co-workers, classmates, friends, family members, buyer-seller, waiter-customer}. Conversations may start and end at different times. We use ID to differentiate each conversation. A conversation with the same ID at different times can show its evolution. For example, in the marketplace, the conversation with ID 2 at 1 minute is bargaining between buyer-seller while at 4 minutes, the conversation changes into arguing.

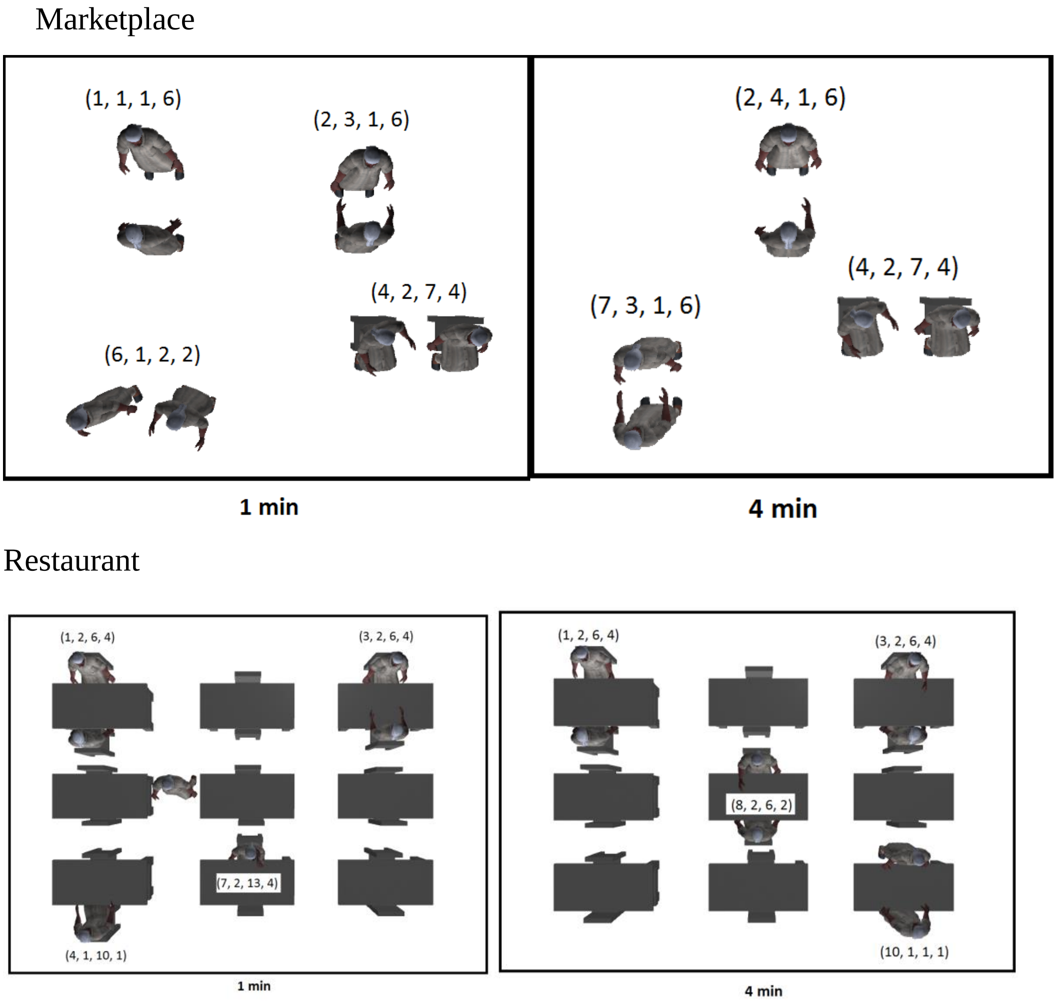
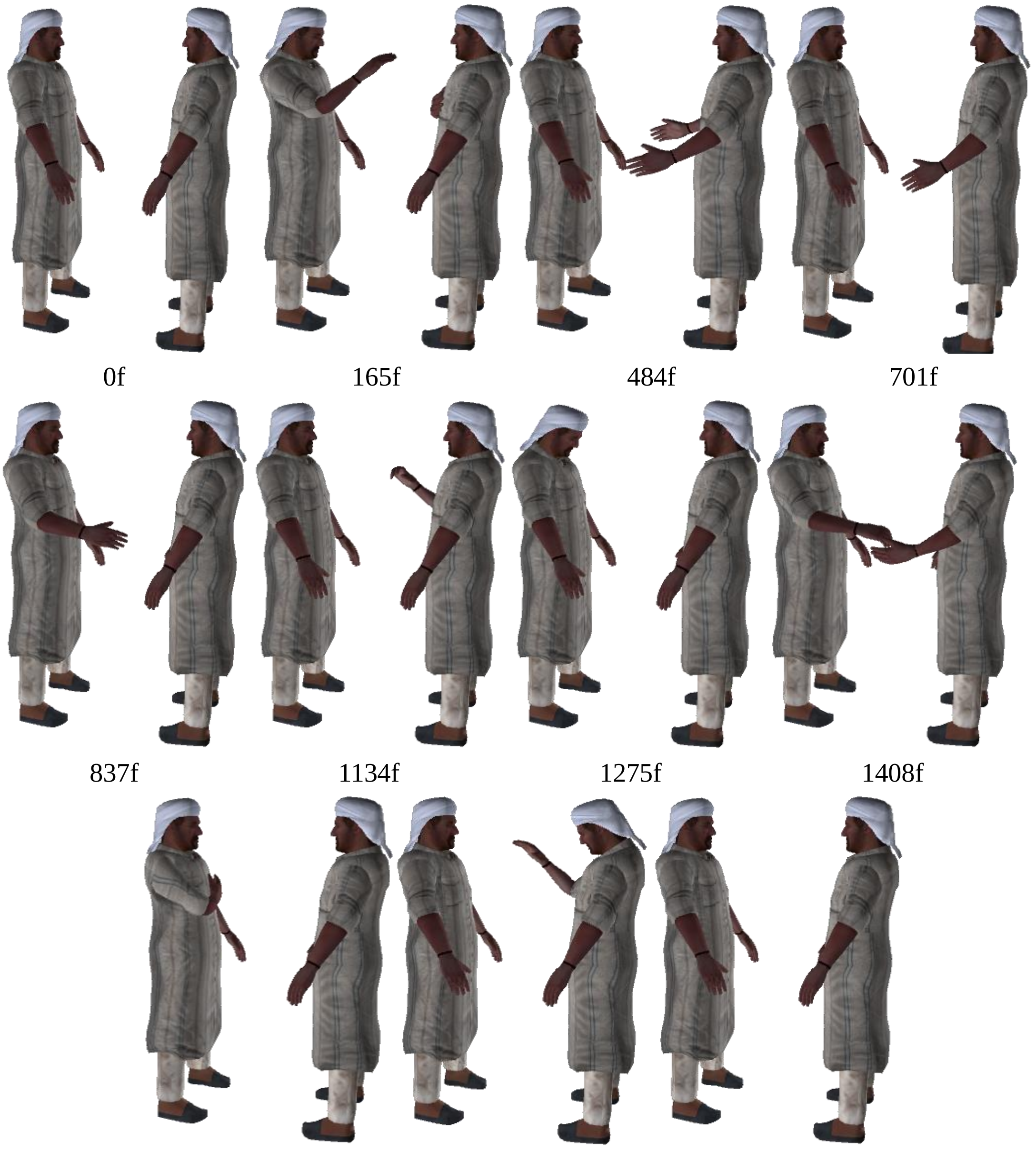


Figure 5. Conversation distributions in different environments

Figure 6 shows the evolution of a conversation between buyer and seller in the marketplace. The outcome of the conversation is indirect transaction success. The conversation archetype at the beginning is simple asking-answering, then it changes into bargaining from 701f; next both buyer and seller compromise and agree with the price so that the transaction is successful. As a result, the buyer gives the money and meanwhile the seller gives the product, which we can see at 1408f.



0f 165f 484f 701f

837f 1134f 1275f 1408f

1593f 1824f 1911f

Figure 6. The conversation between buyer and seller with indirect successful transaction

Figure 7 shows different situation types for simple asking-answering conversations. It displays diversities in conversation instantiation which increases the realism of the simulated scenario.

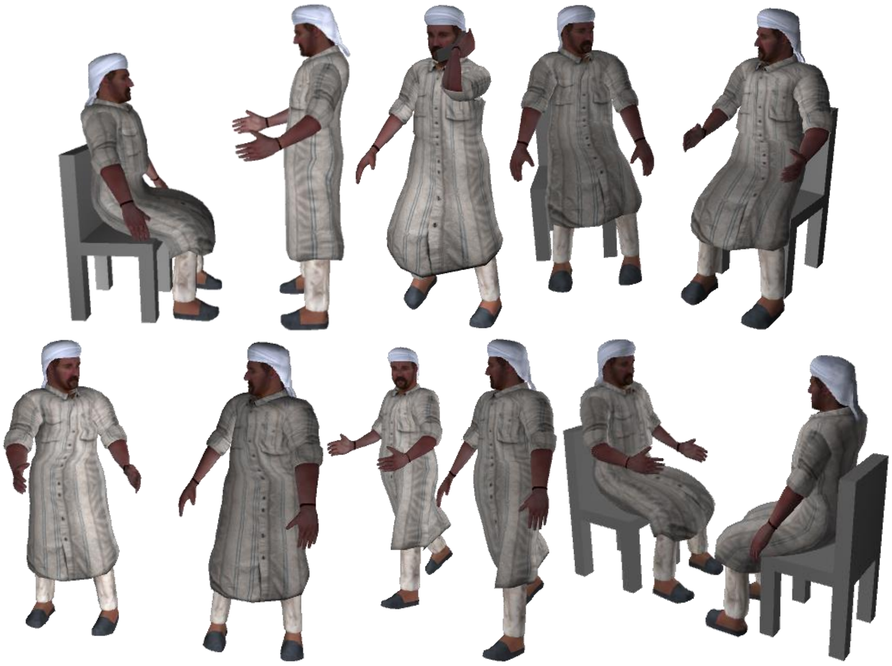


Figure 7. Different situation types for simple asking-answering conversation

Figure 8 shows a marketplace scenario with agents walking, standing, greeting, bargaining and conversing using a cell phone. The behaviors of agents go well beyond just walking and avoiding collisions with other agents and obstacles. They can communicate with other agents at their will, which increases the realism of the scenario and is closer to the fabric of real life interactions.

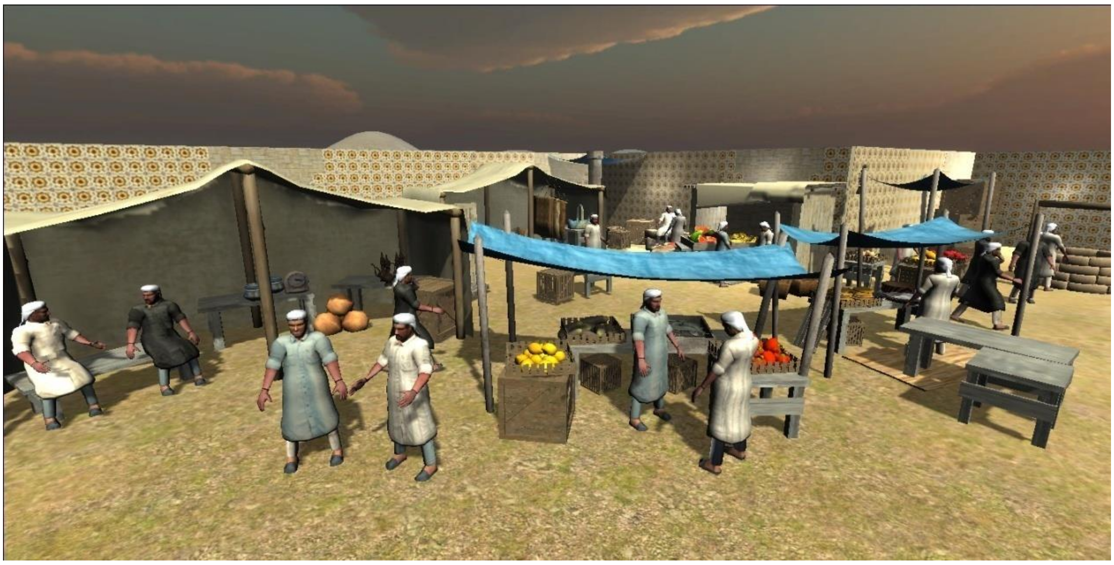


Figure 8. The marketplace scenario

8. CONCLUSION

We have presented a computational framework for synthetic dyadic conversations based on agent attributes and spatio-temporal context. A representative set of conversation archetypes and situations are implemented from the framework using a behavior tree outputting animation commands to the Unity game engine. Our dyadic conversation model shows how environment context, agent attributes and smart conversation events may influence conversation patterns. Agents exhibit different gestures and actions depending on the conversation archetype, situation type and their emotional state. We adopted the ideas of smart events to select visually expressible features of conversations and show some diverse conversation scenarios, which reduced the computation complexity as well as increased the realism of the scenarios.

The marketplace example demonstrates that our framework has the potential to show plausible communication acts between pairs of agents to increase the realism of background characters in a visual crowd simulation. Furthermore, the integration of the situation types illustrates increased diversity of the conversations. One anticipated application of the framework is to produce culturally-variable and agent-sensitive visual simulations for police and military training systems. Because the agent attributes of an actual human subject in a virtual reality experience may be given the same structure as that used for the virtual agents, interactions between the real and virtual agents may be mediated in real-time by the comparative priorities and biases of both.

The main objective of future work is to empirically determine how changing the attribute types and probability distributions influence conversations. Statistics for real world environments should also be empirically determined and then used for simulations to allow future validation studies. We will also engage human subjects in navigating the virtual space and interacting with the virtual agents to see how both mutually influence real-time conversation simulations.

ACKNOWLEDGEMENTS

The authors wish to acknowledge contributions to the marketplace scene by Alice Yang, Kaitlin Reese, Daniel Garcia, and Robert Mead. This work was supported by Jiangsu Province Development Foundation under Grant No. BE2009662. The research reported in this document was performed in connection with Contract Number W911NF10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturers or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron.

References

[1] Pelechano N, Allbeck J &Badler N. *Virtual Crowds, Methods, Simulation and Control*. Morgan & Claypool, 2008.

[2] Shao W & Terzopoulos D. Autonomous pedestrians. *Graphical Models* 2007; 69(5–6): 246–274.

[3] Devillers F, Donikian S, Lamarche F & Taille J-F. A programming environment for behavioural animation. *Visualization & Computer Animation* 2002; 13(5): 263–320.

[4] Maïm J, Haegler S, Yersin B, Mueller P, Thalmann D & Van Gool L. Populating ancient Pompeii with crowds of virtual Romans. *Proceedings of the 8th International Symposium on Virtual Reality, Archaeology & Cultural Heritage (VAST)*, 2007.

[5] Yu Q & Terzopoulos D. A decision network framework for the behavioral animation of virtual humans. *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symp. on Computer Animation* 2007; 119–128.

[6] Allbeck J & Kress-Gazit H. Constraints-based complex behavior in rich environments. *Intelligent Virtual Agents, Springer LNCS*, Vol. 6356, 2010; 1–14.

[7] Stocker C, Sun L, Huang P, Qin W, Allbeck J & Badler N. Smart events and primed agents. *Intelligent Virtual Agents, Springer LNCS*, Vol. 6356, 2010; 15–27.

[8] Kallmann M. Interaction with 3-D objects. In N. Magnenat-Thalmann & D. Thalmann, eds., *Handbook of Virtual Humans*. John Wiley & Sons, 2004; 303–322.

[9] Musse S & Thalmann D. Hierarchical model for real time simulation of virtual human crowds. *IEEE Trans. on Visualization & Computer Graphics* 2001; 7(2): 152–164.

[10] Yeh H, Curtis S, Patil S, van den Berg J, Manocha D & Lin M. Composite agents. *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, 2008.

[11] Schuerman M, Singh S, Kapadia M & Faloutsos P. Situation agents: Agent-based externalized steering logic. *Computer Animation and Virtual Worlds* 2010; 21(3-4): 267–276.

[12] Ennis C. Plausible crowd and group formations. PhD Thesis, Trinity College Dublin, 2010.

[13] Cassell J, Sullivan J, Prevost S & Churchill E, eds., *Embodied Conversational Agents*, MIT Press, 2000.

[14] Moulin B & Rousseau D. An approach for modelling and simulating conversations. In D. Vanderveken, S. Kubo, eds., *Essays in Speech Act Theory*. John Benjamins, 2002.

[15] Cassell J, Vilhjálmsson H & Bickmore T. BEAT: The Behavior Expression Animation Toolkit. *ACM SIGGRAPH 2001*; 477–486.

[16] O' Sullivan C, Cassell J, Vilhjalmsson H, Dingliana J, Dobbyn S, McNamee B, Peters C & Giang T. Levels of Detail for Crowds and Groups. *Computer Graphics Forum* 2002; 21(4): 733–741.

[17] Jan D & Traum D. Dialog Simulation for Background Characters. *Intelligent Virtual Agents, LNAI*, Vol 3661, 2005; 65–74.

[18] Kopp S, Krenn B, Marsella S, Marshall A, Pelachaud C, Pirker H, Thórisson K & Vilhjálmsson H. Towards a Common Framework for Multimodal Generation: The

Behavior Markup Language. *Intelligent Virtual Agents, Springer LNCS*, Vol. 4133, 2006; 205–217.

[19] Vilhjálmsson H, Cantelmo N, Cassell J, Chafai N, Kipp M, Kopp S, Mancini M, Marsella S, Marshall A, Pelachaud C, Ruttkay Z, Thórisson K, van Welbergen H & van derWerf R. The Behavior Markup Language: Recent Developments and Challenges. *Intelligent Virtual Agents, LNCS*, Vol. 4722, 2007; 99–111.

[20] Prendinger H, Ullrich S, Nakasone A & Ishizuka M. MPML3D: Scripting Agents for the 3D Internet. *IEEE Transactions on Visualization and Computer Graphics* 2011; 17(5): 655–668.

[21] Thiebaux M, Marsella S, Marshall A & Kallmann M. SmartBody: Behavior realization for embodied conversational agents. *Autonomous Agents and Multiagent Systems* 2008; 151–158.

[22] Jan D, Herrera D, Martinovski B, Novick D & Traum D. A Computational Model of Culture-Specific Conversational Behavior. *Intelligent Virtual Agents*, *LNAI*, Vol. 4722, 2007; 45–56.

[23] Levine S, Theobalt C, & Koltun V. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics* 2009; 28(5):1–10.

[24] Neff M, Kipp M, Albrecht I, & Seidel H. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics* 2008; 27(1):1–24.

[25] Pedica C & Vilhjálmsson H. Social perception and steering for online avatars. *Intelligent Virtual Agents, Springer LNCS*, Vol. 5208, 2008; 104–116.

[26] Jan D & Traum D. Dynamic movement and positioning of embodied agents in multiparty conversations. *Autonomous Agents and Multiagent Systems* 2007; 47–49.

[27] Hostetler T. Controlling steering behavior for small groups of pedestrians in virtual urban environments. PhD thesis, Univ. of Iowa, 2002.

[28] Thórisson K & Jonsdottir G. A granular architecture for dynamic realtime dialogue. *Intelligent Virtual Agents, Springer LNCS*, Vol. 5208, 2008; 131–138.

[29] Poggi I. & Pelachaud C. Performative faces. *Speech Communication*, 1998; 5–21.

[30] McDonnell R, Larkin M, Dobbyn S, Collins S & O'Sullivan C. Clone attack! Perception of crowd variety. *ACM Transactions on Graphics* 2008; 27(3): 1–8.

[31] Ennis C, McDonnell R & O'Sullivan C. Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics* 2010; 29(4): 1–9.

[32] Shoulson A. & Badler N. A framework for utilizing background agents for ambience and adaptive narrative. Fourth International Conference on Interactive Digital Storytelling (ICIDS), Springer LNCS, Vol. 7069, 2011.

AUTHORS' BIOGRAPHIES

Libo Sun is a doctor of engineering and going to work at Southeast University in China. Her research interests include computer animation, virtual reality and crowd simulation.

Alexander Shoulson is a second-year Ph.D. student in the Department of Computer and Information Science at the University of Pennsylvania, advised by Dr. Norman Badler. His research interests include behavioral control and decision structures for virtual humans, crowd simulation and modeling, interactive and adaptive narrative systems, machine learning, and natural language processing. Shoulson received a BA in Computer Science and Mathematics at Hamilton College in 2010.

Pengfei Huang is currently a PhD student of Graphics Lab at the University of Pennsylvania. His research interests include crowd simulation, multi-modal communication model, virtual reality, motion analysis and synthesis, and virtual surgery.

Nicole Nelson is a graduate student of Graphics Lab at the University of Pennsylvania. Her research interests include crowd simulation and computer animation.

Wenhu Qin is a Professor at the vehicle safety and virtual reality lab at Southeast University in China. He has received his PhD in 2005 from Southeast University. He has more than 25 journal papers, 10 conference papers and a book. He has 3 patents. His research interests include vehicle safety, virtual reality, crowd simulation and road traffic accident reconstruction.

Ani Nenkova is an Assistant Professor of Computer and Information Science at the University of Pennsylvania. Her main areas of research are automatic summarization, discourse, and text quality. She obtained her PhD degree in Computer Science from Columbia University in 2006. She also spent a year and a half as a postdoctoral fellow at Stanford University before joining Penn in Fall 2007.

Norman I. Badler is a Professor of Computer and Information Science at the University of Pennsylvania and has been on that faculty since 1974. Active in computer graphics since 1968 with more than 200 technical papers, his research interests center on embodied agent animation and simulation, human-computer interfaces, crowd modeling and control, and computational connections between language and human action. Badler received the BA degree in Creative Studies Mathematics from the University of California at Santa Barbara in 1970, the MSc in Mathematics in 1971, and the Ph.D. in Computer Science in 1975, both from the University of Toronto. He directs the SIG Center for Computer Graphics and the Center for Human Modeling and Simulation at Penn.