Conversation Starters: Using Spatial Context to Initiate Dialogue in First Person Perspective Games

David B. Christian, Mark O. Riedl and R. Michael Young

Liquid Narrative Group
Computer Science Department
North Carolina State University
Raleigh, NC 27695 USA
dbchrist@unity.ncsu.edu, moriedl@unity.ncsu.edu, young@csc.ncsu.edu

Abstract

A first-person perspective role-playing game by its very nature attempts to create a sense of presence by fostering a deep connection between a user and her avatar. However, the current communication interfaces found within such environments endanger this connection by both forcing the player to consider the interface instead of the game and removing control over the character's behavior from the player. In this paper, we first describe an algorithm which allows users to initiate dialogue with artificial agents in a manner that more closely mirrors normal human interaction. We then describe techniques to increase the realism of the behavior of agents while in the presence of a conversation initiation.

Introduction

Many computer role-playing games (e.g. Planescape: Torment, Baldur's Gate, Ultima) are well known for allowing players to become deeply connected with characters through dialogue and shared adventure. First person perspective games (e.g. Unreal Tournament, Half Life) try to develop a sense of immersion by presenting the user with an environment in which their view and the protagonist's view are tied. A few role-playing games take place in a first-person perspective environment (e.g. Deus Ex, as well as older games such as Pool of Radiance), and attempt to combine these two features to create an immersive, dynamic, first-person perspective game.

One problem that such first-person perspective role-playing games run into is creating a complete but intuitive interface for such games. If the interface is too complex or unintuitive, players may lose the feeling of immersion offered by the first-person perspective. Instead, players may fumble for esoteric key sequences that map to very natural behaviors. If the interface is too simple, however, there is danger of losing some of the game detail and user control that connects the user to her character.

This issue comes directly into play when designing interfaces for communication between users and system-controlled characters in a first-person perspective game. In a third-person perspective game, the player often initiates conversation with another character by clicking on her with a mouse. The player then chooses which of several phrases to say to this character, and the character responds in turn.

However, this sort of "point-and-click" interface defeats some of the purpose of a first person perspective game: to keep the user "trapped" within a virtual persona's body, in order to increase the player's sense of presence. An intermediate solution to this problem is to require users to manipulate their own character's location and viewpoint so that characters that they wish to speak to are near the center of the screen and within a certain distance, and then have the user press some sort of selection key. This is a reasonably realistic interface, since real people prefer to hold conversations at short distance from each other (Argyle and Dean 1965). This is, in fact, what most first-person perspective games involving communication do.

Another reasonable model for a first-person interface design can be found in old text-parsing games, such as Sierra's early adventure games. In these games the user typed sentences like "tell Guard about princess," supposedly to pass on everything the player knows about the princess. This more open-ended approach to dialogue meshes better with the first-person environment than either interface described earlier because it puts more control in the hands of the user, while still not requiring the use of the mouse as a deictic reference tool. However, initiating dialogue with a character in such a game required following a very strict grammar and in no way followed realistic human behavior.

In a first-person perspective environment, it is possible to meld simple textual cues with spatial context about the speaker's location and orientation to reason about the intended recipient of an utterance. In this paper, we propose an algorithm that uses this information to allow users to initiate dialogue with agents without directly selecting them first. By focusing almost entirely on spatial context, along with some name recognition, we are able to create a generalized dialogue initiation algorithm without worrying about knowledge representation or social norms. To reinforce this natural interface for communicating, we

$$\text{P(utterance is for me)} = (w_1 * \text{P(label refers to me)} + w_2 * \text{P(gaze is directed at me)} + w_3 * \text{P(within hearing distance)}) /(w_1+ w_2+ w_3)$$

Figure 1. Probabilistic formula for determining the likelihood that an utterance is intended for a specific agent.

also describe some natural reactions for agents who are in the presence of dialogue initiation.

Dialogue initiation

In order to bring a naturalistic human language interface to the first-person perspective environment, we need artificial agents who inhabit human-form avatar bodies to be able to locate the direction from which an utterance originates and to be able to determine whether that utterance is intended for them. Moreover, the behavior of these artificial agents should be similar to that of humans when confronted with dialog-initiating utterances. Thus the problem is twofold: agents must possess a model of conversation initiation that allows them to determine whether they are being addressed and agents must possess the ability to act in a believable fashion so that the interface for starting conversations is natural and invisible. Accordingly, we treat the model and the behavior as separate modules. First we present an algorithm that determines which set of artificial agents should respond (not just which is the best candidate to respond) and then describe how the algorithm is used to generate believable behavior from the artificial agent's avatar bodies. Since artificial agents are part of the 3D world and not merely immersed in it like humans, they possess the ability to extract additional information from the 3D environment. Our model, however, does not account for any information that a human would not be able to determine if he were in the same situation.

The Respondent Search Algorithm

The objective of the algorithm is to identify one or more agents that will respond to a dialog initiation. We assume that an utterance is made by a human player and that the player's intent is to invoke a response from an embodied agent character, who, in this paper, we call the *intended respondent*. We do not concern ourselves with identifying the actual intended respondent, but rather provide an algorithm for determining the set of agents that might consider themselves the intended respondent. We call this set of agents the *respondent cluster*.

Identifying multiple respondents allows for a wider and more realistic range of behavioral reactions to an utterance because spatial context alone may not be sufficient for an agent (or for our algorithm) to completely reduce the respondent set to a singleton. We allow for the possibility that other modules (e.g., those that take into account discourse or social context) provide additional filtering capabilities. We describe behavioral reactions in detail in the following sections.

Our approach to determining the agents contained in the respondent cluster is performed in two parts. First we determine for each agent within hearing distance of the speaker the probability that the utterance was directed towards them. Once this is determined for each agent, we perform a cluster analysis based on the probabilities to determine which agents should respond to the utterance. We have identified three factors for determining whether a dialog-initiating utterance is directed towards an individual: the use of labels (parts of utterances meant to identify the intended referent), the direction of the gaze of the speaker, and the distance between that candidate recipient and the speaker. For every agent within a designated hearing radius, we compute the probability that an utterance is intended for that agent. This computation is based on probabilistic estimates of the three factors. The probabilistic algorithm is presented in Figure 1. We will address each of these factors in turn.

The first factor, the label, is used by the speaker to indicate the intended target of the utterance. The label can be specific, such as the use of a proper name or title – "Hey Fred!" or "Excuse me, bellhop…" – or it can be only specific enough to exclude some candidates, such as the use of gender labeling – "Excuse me, ma'am…" – or it can be left out completely. The agent maintains a function that maps labels used in an utterance to a value between zero and one, indicating how closely a label applies to that specific agent in the context of the utterance. We assume that the agent has the ability to parse utterances and identify any labels that may occur in them. When no labels are used in an utterance, the label contribution to the overall probability is set to one for every potential referent agent, since the lack of label does not eliminate the possibility that the agent could be the intended recipient.

The second factor, the direction of gaze of the speaker, is also a crucial element in determining the intended respondent. In most circumstances, a speaker will look at (or at least look in the general direction of) the person with whom he wishes to converse (Cassell et al 2001). In real life, a human's area of focused vision is quite small, and thus to see something clearly a human has to be looking straight at it. Detecting the focus of another person's gaze is an instinctual skill (Donath 1995). However, in 3D gaming environments, player control of his avatar body cannot be assumed to be precise, given that head rotation is not typically independent from body rotation and that the human avatar and the intended respondent may both be moving through the environment.

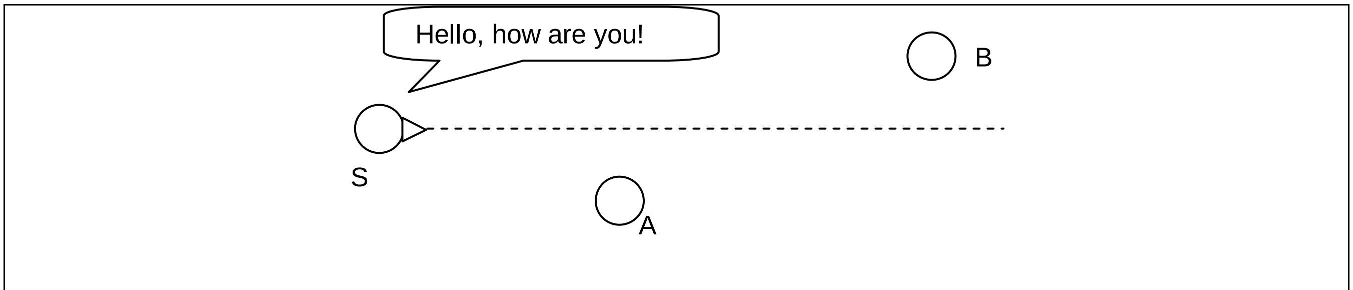

Figure 2. An ambiguous situation invoking response from more than one agent.

Therefore, the agent must compute the line-of-sight vector emanating from the speaker and determine the probability that the speaker is looking at the agent based on the agent's proximity to that vector. We can perform this computation by determining the angular disparity between the line-of-sight and the line between speaker and the agent. The standard distribution function can be applied to map angular disparity to a probability value. Zero disparity will map to the peak of the bell-curve, while a disparity of 180 degrees will map to the fringes of the bell-curve. While there is no concept of foveal vision in first-person perspective games – all objects on the screen are equally in focus – the distribution function can be scaled such that there is a sharp drop-off in probability value for agents that are not within view.

The third factor is distance between the agent and the speaker. There are two reasons that distance is an important factor in determining the intended recipient of a message. First, Americans prefer to be within 3-5 feet of a human they are speaking to, although this is not a hard and fast rule and varies greatly across cultural and social contexts (Argyle and Dean 1965). Consequently, we assume that the speaker will be more likely to initiate dialogue with an agent that is nearby. Second, we assume that the farther away the agent is, the less likely he will hear the utterance since sound drops off exponentially across space. Distance is closely linked with sound-levels. However, since most 3D game environments use text chat as a form of communication, there is currently no direct notion of sound drop-off for player-to-character communication in these kinds of virtual worlds. For now, the distance between speaker and hearer will suffice to approximate this relationship. The algorithm again uses the standard distribution function to map the agent's distance from the speaker to a value between zero and one. The standard distribution should be scaled so that probability drops to zero at the distance at which the ability to discriminate between speech and background noise becomes impossible.

Using the algorithm described above, each agent within the hearing radius of a speaker can determine the probability that any dialog-initiating utterance is intended for him. The algorithm is not complete, however, because we must still determine who will respond. From observations of real-world situations, we know that if the utterance's spatial context is sufficiently ambiguous, more than one recipient may respond. Alternatively, an utterance may actually be directed to all the agents in the room. We wish to preserve this level of realism even though we could ensure that, in all cases, only one agent responds. To determine which agents will respond, we perform cluster analysis on the probability values the agents have computed. The cluster analysis separates the agents into classes based on the likelihood of each agent being the intended recipient. Clustering is performed so that agents with similar computed probabilities are grouped into sets. The cluster containing the agents with the highest probabilities, referred to as the *respondent cluster*, is selected as the set of agents that will respond to the utterance. Under circumstances where the utterance is highly selective – with the use of unambiguous labels, direct gaze, and close proximity – there will normally be only one agent in the respondent cluster. However, under more ambiguous circumstances, our approach does not rule out the possibility of responses from more than one agent.

Our approach is based on observations that when an utterance's intended recipient is ambiguous, human candidates in the real world will attempt to discern who the intended recipient is by looking around, possibly tracking the speaker's gaze, and so forth. If an utterance contains no label or contains a label that could refer to several agents, and the user's gaze is not directed near any agent, it is possible that two characters who are not even in close proximity but who are near the speaker's line-of-sight might both respond as if they each were the intended recipient. In Figure 2, agents *A* and *B* may both be assigned to the respondent cluster, due to the fact that the speaker, *S*, is gazing near but not directly at either of the agents and there are no other distinguishing labels in the speaker's utterance. The next section addresses how we can use the algorithm presented above to determine how agents should behave so as to give the appearance of believability, whether or not the agent is in the respondent cluster.

Believable behavior

The algorithm described in the previous section can be used to produce a set of agents we consider are the most likely intended recipients of an utterance produced by a user. We can now pass on these agent identities to a conversation handler, which handles the actual conversation once initiated. Although we have discovered who these agents are, we have not yet ensured that the agents use this information to react to dialogue initiation in a realistic manner. This section describes the two types of agents who must simulate reaction to the utterance: respondents (i.e. those in the respondent cluster), and non-respondents. Respondents must prepare for conversation, while non-respondent agents must behave as humans would, that is, they must invoke natural behaviors to appear to "discover" what the algorithm has already told them: that they are not the intended audience for the utterance.

Respondents. Respondents are those agents who are in the respondent cluster described in 2.1. As long as there is some agent within the hearing radius from the speaker, there will be at least one agent in the respondent cluster. There are two considerations we make when considering the appropriate reaction of respondents: if there are multiple respondents, and if the respondent's probability of correctness is above a certain "initiation etiquette" threshold.

If there are multiple agents in the respondent cluster, then some algorithm must decide which one is going to speak. More than that, some behavior pattern must make them act believable. Unfortunately, choosing which agent out of multiple parties should respond to a speaker is highly dependent on social context. The representation of this extended context is beyond the scope of this paper. This could be implemented in a separate module with knowledge base access. A sample algorithm might simply have all the multiple agents look at each other and then randomly choose one agent to respond.

There are some times when even humans get confused about the intended recipient of an utterance. For instance, when the intended receiver is neither looked at nor called by name, some confusion typically arises among people that consider themselves possible recipients. This same confusion can also occur when the speaker looks right at the receiver but calls her by the wrong name.

To model this behavior, we set a threshold of "initiation etiquette" for the agents. If an agent is in the respondent cluster, but its probability of being the intended recipient is below the etiquette threshold, the agent acts confused. Before responding, she follows the speaker's gaze, even through herself if necessary, to determine that there is no better candidate behind her. As mentioned previously, following a subject's gaze is a natural and instinctual human behavior. After determining that she is the best candidate, she may respond with a clarification, such as "are you talking to me?"

Non-respondents. Human non-respondents also react to dialogue initiation, and so our agents must as well. When a human initiates a dialogue with another person in a crowded room, people surrounding the intended receiver often look up to determine if the person is looking at them. One can imagine, for example, the following situation:

> Lisa is walking down the hallway when someone says "excuse me, Miss, I really need to talk to you." She turns to look at the origin of the comment, but she sees that the speaker is looking at someone else standing near her, and continues on her way.

If Lisa is an agent, we can see the three factors described in the respondent search algorithm from her (simulated) perspective. The labeling (i.e. "Miss") is ambiguous, the distance between Lisa and the speaker is reasonable, and the speaker's gaze is directed elsewhere. Together, these three factors limit the chances that Lisa would end up in the respondent cluster, and therefore Lisa would not respond. However, crucially, a *human* Lisa would not know that she was not the respondent until she actually looked at the speaker to determine his gaze. If the speaker had been looking at the human Lisa, she would have been compelled by social norms to respond.

Because ambiguity often leads humans to guess they are the recipient even when they are not, we must also make our non-respondent agents simulate this behavior. People gain information about distance and labels faster than they learn about gaze direction, because one must turn to look at the speaker to determine their gaze, while distance and label information are transmitted aurally. We therefore model our non-respondent agents' reaction through a hierarchical system, which examines labeling and then distance to determine whether turning to the speaker is necessary even though an agent will not ultimately respond vocally to the utterance. This model is shown in Figure 3.

```
if(label-correct)
    turn to speaker, follow gaze, ignore;
else if(label ambiguous or missing)
    if(close-enough)
        turn to speaker, ignore
    else
        ignore
else ignore
```

Figure 3. An algorithm for creating believable behavior in non-respondents.

The factors seen in Figure 3, such as label correctness and distance thresholds, are the same ones needed for clustering to discover the intended respondents. Therefore, determining the behavior of non-respondent agents in a room should not be computationally difficult.

Conclusion

In order to preserve the sense of immersion that a user experiences while using a first-person perspective game, it is essential to provide a natural interface for conversing with embodied agents within the 3D game world.

Existing interfaces for initiating dialogue, such as switching modes to “click” on the agent one wishes to speak to or using constrained identifiers and grammar – “tell Guard about princess” – will violate the user’s sense of being a part of the 3D game world. We have developed an algorithm upon which a more natural mode of dialogue initiation can be built. The algorithm uses the spatial context between the speaker and the agents within hearing range to select one or more agents to respond to a user’s utterances. This enables the user to initiate dialog by merely entering text in a more natural way (e.g., without the need for menu systems or mouse “picking”). The labels contained in the user’s utterance, the direction of gaze of the user’s avatar, and the proximity of the agents to the user’s avatar are used to select respondents.

It is not enough to select appropriate respondents, however. We use the algorithm’s outputs to select believable behaviors to perform. Both agents that are in the respondent cluster and those outside of the respondent cluster show behavior that is meaningful to the situation. Agents in the respondent cluster will turn to face the speaker and move to engage him in conversation. In the case that there is more than one respondent, additional modules that use discourse context or social context can be used to refine the respondent selection. Agents outside the respondent cluster will show behaviors consistent with someone who was momentarily confused about his role in the dialogue.

References

Argyle, M. and Dean, J. 1965. Eye Contact, Distance and Affiliation. *Sociometry* 28: 289-304.

Cassell, J., Bickmore, T., Vilhjalmsson, H., and Yan, H. 2001.More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge-Based Systems* 14: 55-64.

Donath, J. 1995. The Illustrated Conversation. *Multimedia Tools and Applications* 1.