Measuring Dramatic Believability

Brian Magerko

Games for Entertainment and Learning Lab
Michigan State University
417 Communication Arts Bldg., East Lansing, MI 48824
magerko@msu.edu

Abstract

The concept of a synthetic character being "believable" in a digital performance has remained a wholly subjective and nebulous descriptor. Little work has been done to provide a useful metric for comparing or contrasting the relative "believability" of different characters. This paper proposes a general method for measuring dramatic believability based on the premise that the believability of a performance to an observer is dependent on the different expectations of that observer. This is meant as an exploratory process of identifying a new, general means of defining believability for use in academic research and character design.

Introduction

Creating *believable* synthetic characters in digital media has been a major focus of research in applied artificial intelligence research (Mateas 1997; Pisan and Nayak 2001; Riedl and Stern 2006a). The believability of these characters is typically supported by the behavior elicited from cognitive, physiological, social, or emotional processes unique to that character's composition. However, as noted by others (Riedl and Stern 2006b), it has proven difficult to clearly define what this term "believability" truly means when discussing (and more importantly comparing and contrasting) character behaviors. "Believability" too often refers to the specific kinds of behaviors a particular agent has rather than a more general metric that is extrinsic of a particular approach and focuses on an observer's perception of that behavior.

Research in virtual reality has defined "believability" in terms of a character's effect on observer behavior in an immersive virtual environment (Bailenson et al. 2005; Garau 2003). The hypothesis is that, while interacting with a synthetic character in a virtual environment, user behavioral data matches that of real-life data as the character becomes more lifelike (e.g. a user's sense of personal space changes when a synthetic character's gaze is more lifelike). While such a definition is certainly useful for discussing behavioral affect, it is not obviously helpful when discussing the overall design of a complete character (as opposed to judging the merits of using particular bottom-up behaviors, like realistic gaze).

A general metric for describing the believability of a character would be useful to both academic researchers as well as designers. At the very least, such a metric would provide an objective common language to discuss the believability of different intelligent agents or character designs. This paper proposes such an experimental metric for the design and evaluation of believable agents. This metric is based on the deconstruction of believability into a) the user's expectation of a performance and b) the fulfillment of that expectation. A definition of these two features is offered below as well as a discussion of the dependence that exists between them in terms of measuring believability.

User Expectation

I contend that the believability of behavior is *a function of what is expected by the observer*. In other words, a character's performance is evaluated through the subjective lens of an individual observer's expectations of that performance. An observer's expectation of a performance can be deconstructed into *external expectation* and *internal expectation*. An observer's external expectation is defined as "the world knowledge that an observer has extrinsic to the performance medium being observed." This includes commonplace knowledge of how the world works, social norms, commonly known facts and skills, etc. Internal expectation is thus defined as "the promises and/or conventions that are provided in the digital world." For example, if a player in a 3D game can move a crate by pushing on it, the game sets up an expectation that "crates can be moved." Another example would be a synthetic character communicating with the player via natural language discourse, setting up the expectation to be able to hold conversations with the character as we normally do with natural language and other humans (or animals).

By considering observer expectation, we can initially understand how this nebulous concept of believability may be grounded. If the player of a computer game expects crates to be moved, either because of conventions used in that game genre or from moving other crates in that particular game, then when the player comes across a similar-looking crate that does not move, that player's expectations are violated. If those violations happen often, then one could determine that the believability of the crate's extremely simple behavior is rather disappointing because of the failed promise that "boxes can be moved." This illustrates my contention that, whatever "believability" may be, a performance is lacking in it if that performance promises something it cannot deliver. The violation or fulfillment of observer expectations is at the heart of the believability of that performance.

Factors of Character Performance

Dramatically believable behavior can be deconstructed into a high-level description of the factors that influence it, as shown in Figure 1. The character definition (e.g. story role, relationships, emotions, etc.) influences which interaction modalities are used (e.g. speech, body language, semaphore, etc.) and performance (e.g. how the actor fulfills story goals, portrays emotions, etc.). The modalities also determine how the character can perform. The environment will influence the character's performance (e.g. Macbeth will play out very differently in a sci-fi spaceship vs. a Scottish castle). The story also has a direct effect (e.g. character's role in the plot, specific performance details). This deconstruction is similar to the observation in the human-computer interaction realm of how user characteristics may impact performance on a task (vs. character and story definition affecting believability) (Coursaris and Kim 2006).

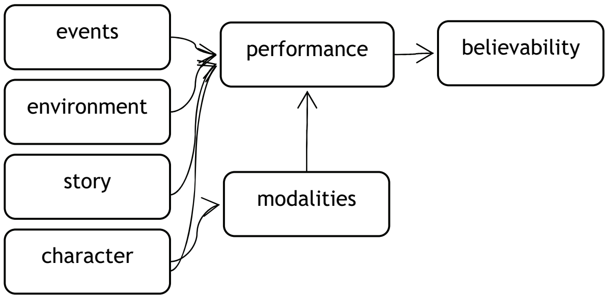

Figure 1. High-level factors of character performance.

These factors can be used to describe a character's involvement in a dramatic situation as well as to identify the ceiling for the performance and how well it is met (our working definition for believability). For example, the character "Bit" in *Tron* (1982) can be described with these factors as a cursory evaluation of its performance in the game. The *environment* is a fantasy analogy of the insides of a computer. The *story* is that a programmer from the real world, the protagonist, finds himself in the computer world and fights to get home against the oppressive rulers. The character *Bit* is a dramatic representation of a Boolean value and acts as a sidekick character through the story. Its color is blue, a benign color in the Tron world. When asked questions, Bit responds with only a "yes" or "no" answer. The different modalities involved in Bit's performance are 3D motion (the character floats around in the air), and answering questions with a "yes" or "no" sound and animation. This character's performance, therefore, is fairly simple (secondary character with highly constrained dialogue and body language). The observer's internal expectation of Bit's performance is likely to be affected by the game's fiction (i.e. being inside of a computer), the character definition (i.e. the "bit" analogy sets very specific expectations), and the performance modalities. The external expectation is likely influenced by the observer's knowledge of computers, implicit knowledge of modern Western narratives in film

Alyx Vance, from *Half Life 2 (2004)*, is a more complex character that can be examined using this same set of factors. Her *environment* is an Orwellian military state. The *story* is about a rogue scientist on the run from the militaristic overlords, aided by Alyx Vance and other rebels. Alyx is a scientist, rebel, and daughter of another key scientist in the story. The interaction *modalities* Alyx employs are 3D motion, natural language interaction (one-sided dialogue with the player, two-sided with other characters), detailed facial animations, and canned speech acts. Alyx's *performance* is very complex, attempting at times to portray as realistic of a human character as possible. The fidelity of her performance depends highly on the situation. She has realistic facial expressions, but does not react appropriately when, for example, shot at by the player or when the player is obviously lost. In terms of the specific interaction modalities, her performance is mixed. The character animation sets a very high ceiling, which is fairly well met. Alyx's responses appear fairly naturalistic, especially compared to other contemporary game characters. Her physical interaction with the world sets a medium expectation (e.g. she can sometimes use a weapon or unlock a door with her tools), which is not well fulfilled (i.e. she has the same limitations of being a scripted character that other game characters routinely exhibit). She also engages with the character at times during discourse, but the player a) can never respond, and b) can never initiate a conversation. Therefore this high expectation has a low fulfillment.

Character Modalities and Performance

The above examples illustrate a simple process of identifying the key high-level factors that influence a performance and an evaluation of the believability of that performance. The performance of a character within each of the utilized modalities is paramount to the believability of that character given the other factors (e.g. Bit's believability is only judged on its 3D motion and limited discourse given the character definition and the fictional world it inhabits). The inclusion of a modality by a character is determined by the character definition within the fiction of the story it inhabits. In turn, these modalities define the dimensions of the character performance which can be used for evaluating the believability of that performance.

Performance modalities include such features as the character's speech acts, physical appearance, blocking, body language, gaze, facial expressions, and motion. Just as character animation or drawings do, each modality has a spectrum of representational fidelity (e.g. discourse can be viewed as the continuum shown in Figure 2).

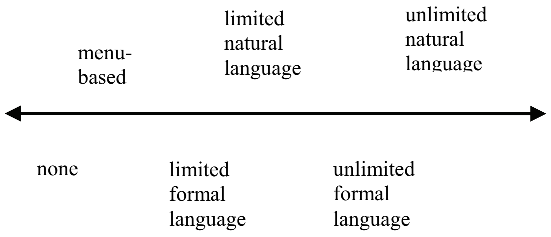

Figure 2. Possible spectrum of fidelity for language interactions with a synthetic character.

A character's performance (i.e. the observable changes within each modality made by the character) is the only observable aspect of the character that can be used by an observer to determine believability (as opposed to character back story or cognition). Therefore, the character performance is to the observer a process of *perceived, constrained decision-making*. As Mateas recognized, "characters are not reality, but rather an artistic abstraction of reality" (Mateas 1997). Therefore, a dramatic performance needs to be believable within the perceived constraints of that character.

Character constraints are both external and internal. *External constraints* deal with both world-level and meta-level issues. World-level issues involve such factors as social status, the character's role in the ongoing narrative, relationships with other characters, etc. Meta-level issues then deal with the interface that the observer has to that character, the application or medium the character is being presented in, etc. Character internal constraints, therefore, involve many AI issues, such as commonsense reasoning; emotional modeling; interpreting visual and audio inputs; mapping beliefs, desires and intentions to actions; directability; story goals for the character; etc.

Defining Believability

We now have a character's performance viewed through two lenses: a) how well the character performs within each modality and b) how well the character adheres to its perceived constraints. This leads us to a more specific definition for how to measure believability. A character's believability can be measured as a *function of the sum across modalities* of a character's adherence to the set of *relevant constraints*, as shown in Equation 1, where the term Diff_A is the difference between observer expectation and what is actually offered for a modality (m_a) summed across all modalities (M_a) considered for some agent A. In other words, we can evaluate a performance one modality at a time, observing how consistent the character's performance within that modality in light of the relevant external and internal constraints, and then sum that result with the results from the other modalities.

$$\mathrm{Diff}_A = \sum_{m \epsilon M_A} 1 / (\mathrm{Max}(m_a) - \mathrm{Value}(m_a))$$

Equation 1.

However, just relying on this Diff_A term alone to describe believability suggests that "always shooting lower for what the player should expect is better" since the difference term is likely to be smaller. A more accurate function would also take the absolute value of observer expectation into account; in other words, a character is "more believable" if it is closer to some standard of realism for that modality. We therefore get the term Max_A:

$$\mathrm{Max}_A = \sum_{m \epsilon M_A} \mathrm{Max}(m_a)$$

Equation 2.

We therefore come to a more complete definition of believability (see Equation 3), which states that the believability of some agent A's performance is a function of a) the difference between what is expected and what is offered and b) proximity of observer expectation to some measure of "reality."

$$B(A) = f(\mathrm{Diff}_A, \mathrm{Max}_A)$$

Equation 3.

Bit's performance as a synthetic character could be measured across two modalities, 3D motion and discourse. The external constraints involved would be the story's fiction, the character's role as a sidekick, and Bit's relationship to the protagonist as a fellow blue entity. The character internal constraints would be Bit's perceived knowledge (in relation to what questions were being asked) and the bit of personality that Bit has. These constraints would then be used for consideration when subjectively evaluating the performance within each of the two modalities used. As shown in Figure 3, Bit does reasonably well at discourse (always answers easily interpretable questions with a limited grammar, but is difficult to do well in a real human-agent interaction) and does very well with 3D motion (the motion is purely flying smoothly in a 3D space with simple rules of physics applied).

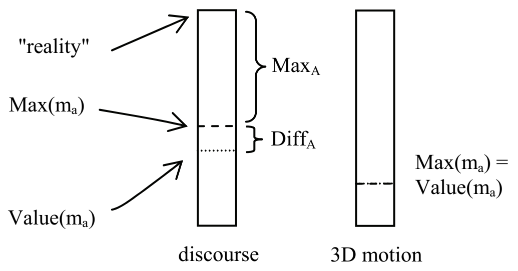

Figure 3. Believability of "Bit" with two modalities.

Discussion

This proposed metric for believability probably raises more questions than answers them. How to evaluate or even apply this method has not yet been answered. Presence research employs two main approaches to experimentally show synthetic character affect on users in a virtual reality: post-experience interviewing and in-experience behavioral data collection. It is hard to argue that behavioral data collection is an appropriate approach for many different possible digital media (e.g. computer games, digital film, digital assistant, etc.), therefore a usability study to gauge the emotional response of an observer through asking questions seems appropriate. It is conceivable that there is some possible mapping between believability and measurable statistics (e.g. recognition or recall from the experience), but that also seems improbable. Designing an appropriate test to use this metric is paramount for both evaluating this method as well as putting it to use. A future goal of this work is to explore both well-defined quantitative and qualitative guidelines to use for comparing performances.

The details for comparison within modalities are also unclear. What determines where the ceiling is for a particular modality? Is there a spectrum of fidelity that can be defined for each? It seems reasonable to assume that there may be absolutes within each, but at what level of granularity? The previous decomposition of natural language (see Figure 2) may be an appropriate tactic for identifying the spectrum across each modality. However, the approach does not clearly generalize to either modalities that have inter-dependencies with other modalities or with modalities that can differ along several dimensions (e.g. natural language approaches may focus mainly on only a specific aspect of the domain versus an approach that does well in all areas).

This issue of how to define modalities is related to the issue of how to determine precisely how much a particular modality is filled. My examples used earlier are purely subjective based on my personal reported perceptions.

Is it possible to formalize the relationship between modalities and other factors? For example, what if the fiction of the world sets up an expectation for a modality that doesn't exist? That problem with believability of a performance is not currently addressed in the methodology described here.

Despite the apparent failings that need to be worked through experimentation, this approach does hold some promise. It provides a standard method for academics and designers to compare and contrast different synthetic characters. Further design and experimentation of this proposed method will hopefully yield a useful metric to be used in this work.

References

(2004). "Half-Life 2." Valve Software Corporation.

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (2005). "The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive and Behavioral Markers of Coprescence in Immersive Virtual Environments." *Presence*, 14(4), 379-393.

Coursaris, C. K. and Kim, D. J. (2006) "A Qualitative Review of Empirical Mobile Usability Studies." *Twelfth Americas Conference on Information Systems*, Acapulco, Mexico.

Garau, M. (2003). "The Impact of Avatar Fidelity on Social Interaction in Virtual Environments," University College London, London.

Lisberger, S. (1982). "Tron." Buena Vista Pictures, USA.

Mateas, M. (1997). "An Oz-Centric Review of Interactive Drama and Believable Agents." School of Computer Science, Carnegie Mellon, Pittsburgh, PA.

Pisan, Y. and Nayak, A. (2001) "Increasing Believability: Agents that Justify Their Actions." *10th IEEE International Conference on Fuzzy Systems*, Melbourne, AU, 1347-1350.

Riedl, M. and Stern, A. (2006a) "Believable Agents and Intelligent Story Adaptation for Interactive Storytelling." *Proceedings of the 3rd International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, DE, 1-12.

Riedl, M. and Stern, A. (2006b) "Failing Believably: Toward Drama Management with Autonomous Actors in Interactive Narratives." *3rd International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, DE.