Evaluating a Conversation-Centered Interactive Drama

Manish Mehta[1], Steven Dow[1], Michael Mateas[2], Blair MacIntyre[1]

1. College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0760
{steven, mehtama1, blair}@cc.gatech.edu

2. Computer Science Department
University of California, Santa Cruz
Santa Cruz CA 95064
michaelm@cs.ucsc.edu

ABSTRACT

There is a growing interest in developing technologies for creating interactive dramas [13, 22]. Evaluating them, however, remains an open research problem. In this paper, we present a method for evaluating the technical and design approaches employed in a conversation-centered interactive drama. This method correlates players' subjective experience during conversational breakdowns, captured using retrospective protocols, with the corresponding AI processing in the input language understanding and dialog management subsystems. The methodology is employed to analyze conversation breakdowns in the interactive drama Façade. We find that the narrative cues offered by an interactive drama, coupled with believable character performance, can allow players to interpretively bridge system limitations and avoid experiencing a conversation breakdown. Further, we find that, contrary to standard practice for task-oriented conversation systems, using shallowly understood information as part of the system output hampers the player experience in an interactive drama.

Categories and Subject Descriptors

H.5 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces – *Evaluation/Methodology, Natural Language.*

General Terms

Design, Human Factors, Experimentation.

Keywords

Interactive Drama, Believable Agents, Embodied Conversational Agents, Evaluation.

1. Introduction

In an interactive drama the player enters a virtual world, interacts with autonomous, believable characters and, through her interaction, influences both the characters and the overall development of the story [1]. An interactive drama is in some sense a pure hedonic experience, immersing the player in a dramatic social interaction without providing, as most games do, a clear player goal; the player invents goals for herself as the interaction with the characters unfolds. Interactive drama presents one of the most challenging applications of autonomous characters, requiring characters to simultaneously engage in moment-by-moment personality-rich physical behavior, exhibit conversational competencies, and participate in a dynamically developing story arc. Successful future research in believable agents requires deploying such agents in completed dramas, evaluating the effectiveness of the agents in creating a compelling player experience, and using the results of the evaluation to guide future research.

Conversation-centered interactive dramas, which place the player in rich social situations where the primary interaction is through conversation, offer interesting evaluation challenges. First, methodologies used to evaluate task-based conversation systems are inappropriate, as they employ metrics based on efficiency and task accomplishment; players in interactive dramas don't accomplish tasks, but rather are engaged in a dramatic experience. Second, as most interactive dramas to date have been small prototypes rather than fully-realized experiences, it has been difficult to develop evaluation methodologies. Finally, in an interactive drama, the success of a conversational turn hinges on whether and how the player is able to incorporate the conversational turn into her growing understanding of both the characters and the narrative situation. This inherently qualitative process resists simple approaches to quantifying conversational turn success. Further, this dependence on player interpretation implies that system level technical failures (misunderstood player input and/or the selection of incorrect responses), though useful to know, do not necessarily cause a player-perceived conversational breakdown. The design of the story itself, including the authoring of the conversational content, is instrumental in determining whether the player has an enjoyable experience, and how and whether technical breakdowns impact this experience. Hence, the effectiveness of technical and design techniques used in the interactive drama needs to be related back to the player's perceptions during the interaction. Ideally, we want a player-centric evaluation methodology that starts with the player's experience and analyzes how the technical and design approaches used in the system impact the experience, thus providing insights for creating more engaging player experiences in future systems.

In this paper, we present a qualitative study of Façade, a real-time interactive drama, specifically focusing on the relationship between AI decision making and player perception. Façade is the first fully produced, real-time, interactive drama, combining autonomous characters, artificial intelligence (AI)-based story

Conference'07, May 14--18 2007, Honolulu, Hawai'i, USA.

management, and natural language processing to place the player in a dramatic world. As the first fully-realized interactive drama, Façade provides a nice opportunity to develop evaluation strategies for conversation-centered interactive dramas.

In this study we measure system level technical failures on the understanding side, elicit rich qualitative data (interview data) to understand player interpretations at the points of conversational breakdowns, and correlate player interpretations at these breakdown moments with the corresponding conversational AI decision making. Our approach quite similar to conversational analysis is used to evaluate the effectiveness of the technical and design approaches employed in Façade, and provide guidance for the design of future conversation-based interactive dramas. The main results in this paper are:

a) An intervention designed to elicit subjective opinions about conversational breakdowns in order to correlate player interpretation with the functioning of the conversational AI system in an interactive drama.

b) A counterintuitive finding that, contrary to a task-oriented conversational system, in a social conversational system like Façade, explicitly using the shallowly understood information as part of system output hampers the user experience.

c) Even during system failures, where a player utterance is misunderstood, Façade succeeds in providing enough narrative cues to make the player fit these breakdowns into the experienced narrative flow.

d) Believable character performance is essential for maintaining a positive player experience by keeping player interest alive even during complete conversational breakdowns (breakdowns the player is not able to interpretively bridge).

The rest of the paper is organized as follows. In Section 2, we present existing evaluation techniques for traditional conversational systems and experiential (non task-based) systems. In Section 3, we briefly describe Façade, including both the player experience and the underlying AI architecture. Section 4 describes our evaluation method. We present the details of our findings in Section 5 and discuss the results in Section 6. In Section 7, we conclude with suggestions for future research.

2. Existing Approaches

The conversational systems research community has traditionally employed objective evaluation measures such as task success rate, turn correction ratio, inappropriate utterance ratio, number of turns, concept accuracy, and elapsed time [e.g. 7,16,18]. For question answering systems, language input/answer pairs have been used as an evaluation criterion [6], where the correct understanding is defined in terms of the number of correct replies to the input sentences. These measures are appropriate for task based conversational systems where the purpose of the system is to help the user efficiently accomplish a task. The underlying philosophy behind these evaluation metrics is that conversational interaction can be framed as a simple exchange of clear, well-defined meanings; the success of a conversational turn can be defined by whether the system understood the user's meaning and conveyed a clear meaning back to the user.

These measures, and the assumption underlying these measures, are inappropriate for evaluating an interactive drama. The goal of an interactive drama is not to help the user accomplish a task, but rather to create an engaging, high-player-agency experience. In such an experience, the success of conversational turn hinges on how well the player can incorporate the turn into their growing understanding of the characters and story. An evaluation strategy to assess the design and technical approaches used in an interactive drama must take these factors into account.

In subjective assessments of conversational systems, existing approaches typically extract a user satisfaction measure from a Likert scale questionnaire. The user is presented with a number of statements related to her perception of interacting with the system, and asked to mark her degree of agreement (on a numeric scale) with each statement [e.g. 2,11,17]. However, these approaches fail to capture the rich interpretive processes people employ in understanding conversation, especially important in an experiential interaction with a conversation-centered interactive drama. Moreover, these studies typically separate subjective assessment from objective evaluation, thus preventing subjective assessment from feeding back to the technical approach. In our evaluation approach for Façade, we link the quantifiable technical failures (both true technology failures as well as design bugs) with qualitative player assessment, allowing us to evaluate the effect of the design and implementation of the underlying conversation system on the rich processes of player interpretation.

In recent years, an interest in evaluating the experiential aspects of interactive systems has emerged in the HCI community. For example, Höök, Sengers and Andersson develop an evaluation strategy for interactive art, specifically by using interviews and observation techniques [9]. Stasko et. al. use longitudinal studies to evaluate user experience with ambient displays [20]. Our Façade evaluation draws on the new qualitative evaluation methods being developed in the HCI community, but is unique in adopting these techniques to evaluate conversational believable agents in an interactive drama.

3. Overview of *Façade*

3.1 Story and Player Interaction

Figure 1: Grace and Trip, the central characters of interactive drama Facade

In Façade, the player visits the married couple Grace and Trip at their apartment where she quickly becoming entangled in the high conflict dissolution of their marriage. The game begins with the player in the hallway outside their apartment, where the couple can be overheard arguing. The player's interaction influences both the moment-by-moment development of the

drama as well as the ending. The player may react to the experience with hilarity or anger, or play a number of roles from councilor to devil's advocate. Unlike most games, the player is not given a clear goal; the player invents goals for herself as the interaction with the characters unfolds. For complete details of Façade, see [14].

The player interacts from a first person perspective, moving about the world, manipulating objects, and, most significantly, talking to the characters through unrestricted, typed natural language (the characters respond with spoken dialog). Since the player's interaction effects the long term development of the story, the experience has replay value, in that different interaction approaches will result in different story trajectories. Given the technical and design difficulties of creating real-time, animated, AI-controlled characters that respond broadly and robustly to natural language input, there will inevitably be AI breakdowns in which the characters respond inappropriately to player interaction. Façade was designed to help the player maintain immersion in the experience even in the face of these AI breakdowns.

This study, focusing on the language understanding and dialog management components of the system, examines: (i) Whether and how Façade's design approaches work to maintain player engagement during AI breakdowns, (ii) Which breakdowns tended to hamper player experience and (iii) Which technical and design decisions are responsible for perceived conversation breakdowns.

In order to facilitate our future discussion of conversational AI breakdowns, we first present an overview of Façade's natural language processing (NLP) architecture.

3.2 Facade's NLP Architecture

The Façade system consists of three major components: autonomous characters implemented in the reactive planning language ABL, a probabilistic, agenda-based drama manager, and a natural language processing system which is used by the characters to understand player utterances and decide how to respond to these utterances [15].

The Façade natural language processing (NLP) system makes use of broad, shallow, author-intensive techniques to understand natural language typed by the player. It accepts surface text utterances from the player and decides what reaction(s) the characters should have to the utterance. For example, if the player types "Grace isn't telling the truth", the NLP system is responsible for determining that this is a form of criticism, and deciding what reaction Grace and Trip should have to Grace being criticized in the current context.

The NLP system is divided into two phases: phase I maps surface text into discourse acts, while phase II maps discourse acts into one or more character responses. At phase I, Façade employs shallow semantic parsing to map surface text to discourse acts. The semantic parser is implemented using a forward chaining rule system. Phase I processing is a strong many-to-few mapping – the huge set of all possible strings a player could type is mapped onto a small (~30) set of discourse acts. Discourse act representations are relatively flat. Rather than consisting of complex, symbolic constructions supporting compositional semantics, Façade discourse acts are simple frames whose slots tend to be filled with atomic tokens. For example, the discourse act ReferTo, produced when the player makes a reference to something, has only two slots: character, which takes a token representing the character the ReferTo was directed towards (if any), and object, which takes a token representing either a physical object (e.g. WeddingPicture) or an abstract object such as a topic (e.g. RockyMarriage).

Once phase I has recognized one or more discourse acts in the surface text, phase II determines what the reaction will be to the discourse acts. Phase II is the dialog manager, responsible for maintaining multiple conversational contexts and selecting a reaction from among those proposed by the different contexts. In those cases where phase I generates multiple discourse acts, phase II is responsible for deciding which discourse act to respond to, or, if responding to multiple discourse acts, choosing responses that work together [15]. Phase II is responsible for selecting reactions to story topics and object references, as well as advancing the specific conversation happening within the current story beat. As in-depth conversation on a topic requires understanding all the nuances of a topic as well as significant content creation, Façade, as part of its design strategy at phase II, proposes reactions that try to deflect back to main story topics as well as prevent the player from digging too deeply on any one topic. Push-too-far reactions for example, happen if the player "harps" on a specific topic too much by bringing up the topic several times in a short period of time. In a push-too-far reaction, the "bugged" character (the character for whom the topic is a hot button) responds negatively. Specific-deflect reactions respond to multiple (indirect) references to the same topic in those cases where the most specific reaction is not available (perhaps has already been used up) and push-too-far is not yet available. Generic-deflect reactions are chosen if no better reaction is available. Generic-deflect reactions acknowledge that the player said something, but try to believably ignore the utterance and move on with the story. Two example generic-deflects are "Uh huh… Anyway…" and "Yeah… As I was saying…". One of the goals of the study was to evaluate whether the design approach of using deflect reactions were perceived by the players as conversation breakdowns so as to inform the design of future conversation-centered interactive drama systems.

In Façade, "beats" are the representational unit within the architecture that explicitly coordinates detailed character activity in order to achieve dramatic action. The drama manager, with its collection of beats, forms a macro-story machine that abstracts above the moment-by-moment activity to guide the story at the level of beat sequencing; by selecting beats, the drama manager "turns on" specific micro-story machines to handle the detailed character activity and player interaction within the beat. Once a beat has been selected, the characters try to accomplish the beat by making use of beat specific character behaviors. During a beat, the full performance a player experiences consists of both material from within the beat and reactions to player interaction.

A conversational AI technical failure in Façade could be in the form of an input language understanding error whereby the NLU phase I doesn't detect the right discourse act which in turn could result in characters addressing a wrong topic in response. It could also be in the form of an improper character performance where they are not able to communicate their intentionality through the

Table 1: The table shows different conversation snippet corresponding to player comments during the interview

T: The way you keep talking about italy, damn. G: Thaw me out... G: Trip, you think you're so romantic... but no you're trying -- (interrupted) P: I said one thing about italy. T: Huh? What, what -- what was that? **(a)**	T: Ah! G: Adam, you're saying I'm... not communicative? No, that's just wrong, I'm the one here who is able to actually say things... P: what are you actually trying to say G: Look, why don't we talk about us, our relationship **(b)**	G: Chris, what I really just need right now... is for you to just be my friend... P: yes P: of course T: Uhh, God! What the hell has been going on in here? P: i'll always be your friend T: Heh... should I just go back to the kitchen? **(c)**
T: Ah! G: It's like I don't know who you are anymore. P: Maybe you should talk in private. G: Anne, you're saying I'm... not communicative? No, that's just wrong, I'm the one here who is able to actually say things... **(d)**	G: I'm sure I can return most of this, and try to start over again on this room... P: Why would you do that? G: Ah, yes, I've been waiting for someone to say that! T: What are you talking about? G: Trip, she is just being honest about my decorating, which I appreciate. **(e)**	P: oh grace! G: Well, come in, make yourself at -- (interrupted) P: why are you hiding? P: thank you G: Uh... (clears throat) um... **(f)**

designed verbal and non-verbal behaviors. However an AI level technical failure might not cause a breakdown in player's experience. Our study seeks to understand the relationship between AI technical performance and player experience.

4. Study Procedure

We have been engaged in a large study comparing how different forms of mediation effect player interaction within an AI-based interactive drama. Towards this end, we have built an augmented reality version of Façade, employing a Wizard of Oz for the speech and gesture recognition [24] as well as a desktop version of Façade that uses (Wizard of Oz) speech recognition. In our study, each participant played all three versions including the original desktop keyboard based version of Façade. Elsewhere we have reported on the presence experienced by the player under the different mediation conditions, and the effect of presence on player engagement [23]. In this paper, drawing on the qualitative and quantitative data collected in the larger study, we focus on examining the effect of conversational AI failures on the player's perception of the success or failure of the conversational turn. Here we briefly describe the study procedure.

We recruited twelve participants through Craigslist.org and local game forums in the Atlanta area. As we were not looking for statistically significant results and wanted to use qualitative analysis to reflect on the system errors encountered during the game session, 12 participants provided us tractability in conducting the labor intensive qualitative analysis. We enrolled a range of genders (balanced 50/50), races, education levels, and ages (from 18 to 33 with an average age of 25.8). For each participant, the study lasts about three hours. We paid participants $10 per hour rounded up to the nearest hour for their time. Players signed a consent form and listened to an explanation of Façade. Each participant played Façade three times, once for each interface variation (making six possible orders, balanced out to account for learning effects). Our analysis did not reveal any significant differences across the three interfaces with regards to conversational breakdowns. Therefore, we present combined results across all mediation conditions.

We recorded video of the game episodes and interviews, from both what appeared on screen and from multiple third-person perspectives, in order to capture player emotions and physical actions. Throughout each episode a researcher logged unusual player reactions and apparent conversation breakdowns (or unusual interactions the player might conceivably interpret as a conversation breakdown). After each play, participants were interviewed about their experience. For part of each interview the interviewer reviewed the logged moments on a video monitor, collecting retrospective protocols of the player's experience. We chose the retrospective over a standard talk-aloud protocol in order to keep from interrupting the rapid flow of the game. Our earlier lab study [8] verified the effectiveness of retrospective analysis for gathering subjective player data following each game episode. During the interview, the video was played back and the player described their reactions, interpretations, expectations and player goals at the logged unusual conversation moments. We also collected player and character dialog logs, position and rotation logs, and AI processing logs used for the quantitative analysis of technical failures.

5. Findings

5.1 Phase I Analysis

In order to determine the different types of technical failures present in phase I (semantic parsing of surface text to discourse acts), we looked at the AI logs to determine the discourse acts recognized for every player utterance. The goal is to come up with taxonomy of phase I failures, in order to relate these technical failures to player perception. We identified 5 categories of phase I processing.

i) Correct: The discourse acts produced by phase I adequately captures the pragmatics of the player utterance. For example:

Player: Grace, how are you

Phase 1 output: DAGreet Grace (a greeting discourse act directed at Grace)

ii) Wrong discourse act: The discourse acts produced by phase I completely miss the pragmatic intent of the player utterance. For example:

Player: What's with the sticky notes?

Phase I output: DAGreet Trip (A greeting discourse act directed at Trip).

The player refers to sticky notes that are on the wall of the apartment over by the work table (Grace and Trip both work in

the same advertising agency). Though talking about sticky notes is not part of the story content of the system, talking about work (and their conflicts over work) is. Phase I should have produced a reference to the "work" story topic (the physical sticky notes are related to work); instead, it erroneously produced a greeting discourse act.

iii) Doesn't understand: Phase I doesn't understand the player utterance, and thus doesn't produce a discourse act. For example:

Player: Grace, needs to feel loved

Phase I output: (no output)

Here the system should have produced an "explain" discourse act indicating that the player is explaining that Grace doesn't feel loved (explanations about the marriage dynamic, e.g. that someone is loving, controlling, lying, feels loved, afraid, etc. are part of the semantic domain of the drama), but instead failed to understand the utterance at all.

iv) Conflicting discourse acts: Phase I produced multiple discourse acts that are in direct opposition to each other. For example:

Player: No, it is fine

Phase I output: DAAgree Trip, DADisagree Trip (both that the player agrees and disagrees with Trip)

As we will see below, this category of phase I technical failures can lead to large perceived conversational breaks, as phase II may choose to respond to the discourse act that is the exact opposite of what the player intended.

v) Typing Issues: Phase I errors caused by misspellings (e.g, "you are rigth" instead of "you are right") or by the player believing (incorrectly) that they can split a long utterance over multiple conversational turns.

The results of the complete phase I analysis are shown in Table 1. On average, phase I produced correct discourse acts 73% of the time, a wrong discourse act 9% of the time, failed to understand anything 7% of the time, produced conflicting discourse acts 3% of the time, and suffered from typing issues 5% of the time.

5.2 Player Reaction Analysis

In order to correlate player conversational interpretations with the functioning of the underlying AI system, we first transcribed the retrospective protocols (13 hours of protocols). As explained earlier, these protocols focused on unusual player reactions and apparent conversation breakdowns noted by a researcher. We then employed a common qualitative analysis method, Grounded Theory [21], to analyze the retrospective protocols. Using grounded theory principles, we started the analysis process by making notes for each player comment. These notes include what aspect of the system players were commenting on (e.g., NLU, character reaction, design approach etc) and what the player was thinking about that system aspect (e.g, player said that it was difficult to communicate because of a particular design approach). After finishing the note taking exercise for the interviews of all 12 players, these notes were then used to iteratively come up with a base set of categories. We used these categories to tag each player statement. These categories (subset shown in Table 3) were then organized into higher level major concepts by grouping related categories into a single concept. We then looked at phase I codes (categories (i)-(v) above) and the selected reaction at phase II corresponding to these notable moments to correlate these major concepts with the technical processing. Next, we report on the major concepts and reactions identified in the retrospective protocols, and describe their relationship with the underlying processing occurring in phases I and II. Table 2 contains representative conversation snippets that illustrate the major concepts and reactions described below.

Table 2: The table provides a sampling of ~25 different categories used for coding player interviews

Categories	Player Sentences
Forms back story	"Trip seems terrified with his wife. I felt like he was not faithful"
Relates to interpers. conversation norms	"It's logical for him not to come to grace with this problem b/c he's the one feeling pressure"
Negative feelings	Seems like no where , hard to follow sometimes
Timing Issue s	"I was a beat behind or they were a beat ahead"
Likes character reaction	"His reaction was funny.... Made me laugh"
Characters don't want to discuss topic	"I think I brought up a topic which she didnt want to talk about."
Characters mind is occupied	"He doesnt respond..... may be his mind is on something else"
Strategies for repeated NLU misunderstandings	"Tthey aren't listening to me… I thought I would try to play along"
Likes character reaction in case of NLU error	"I was sticking up for her and all of a sudden I am going to hell. Wait a second this isnt cool. I thought this was funny too. "

5.2.1 Narrative Interpretation of Character Reactions

Players are adept at creating elaborate back-stories to make sense of character reactions. Façade successfully provides the player with ample material (hints at conflicts and topics relevant to the story) for creating these back-stories. For example, in response to the conversational snippet in table 2-a, player 2 described:

"I guess saying I want to go to Italy pissed him off, brings up bad memories, it didn't turn out good for them…. So me saying Italy reminds him of going there and something atrocious must have happened there".

The same formulation of an elaborate back-story was described by Player 12: "It seemed like we had a relationship in the past." (P12, table 2-b).

When the characters would avoid certain topics, players are able to ascribe topic avoidance to the character's mental state, saying that characters "didn't want to talk about [it]" (P10) or that they "didn't want to admit certain things about themselves" (P11). Players feel that at times the characters can't address them because they are "listening to each other" (P8) and are thus pre-occupied. Players use interpersonal conversational norms to understand character reactions, for example: "I think she saw trip walking back in and was like let's stop talking about it" (P3, table 2-c).

In correlating these responses with phase I and II processing, successful narrative interpretations often occur in one of two cases: when the system selects deflect responses (see deflect reaction, section 3) or PushTooFar reactions (see PushTooFar reaction, section 3). Deflect reactions are often selected when the player utterance is not understood or because the characters are performing high-intensity, uninterruptible dialog. PushTooFar reactions are selected when the player refers to the same topic several times. One of the design challenges with creating an interactive drama is giving the autonomous characters the ability to discuss a topic in depth. Limits in natural language processing make it difficult to distinguish the many nuanced meanings that might surround a topic, while limits in authoring effort make it difficult to provide non-repeating content for a topic (when characters in an interactive drama repeat content, this immediately kills the believability of the characters – in Façade, the characters never repeat a line of dialog). PushTooFars and deflects are mechanisms for, hopefully believably, limiting the depth in which a player can drill down on a topic. The successful player interpretations found in our study indicate that this is a successful design strategy.

Players also use narrative interpretation to bridge some of the phase I processing failures. We find examples of successful narrative interpretation being used to "cover" conversational breakdowns for all phase I error categories. While, as we will see below, narrative interpretation isn't always successful, it does illustrate the power of providing the player with sufficient narrative material to allow them to use their human intelligence to fill in the gaps of AI failures.

5.2.2 Reactions to Shallow Semantic Understanding

Players sometimes notice that the characters seem to understand a concept related to the one they expressed, rather than the specific, player expressed concept. For example, in the dialog snippet in Table 2-d, Player 2 said: "I didn't see how that related at all to what I said… talking in private has nothing to me being not communicative." In such situations, player feels that the characters are addressing a related concept, rather than addressing their specific comment or question: "I talked about the city view but he sort of talked about decorating....." (P6). Though the characters understand a related concept, and almost get it right, players "wanted a deeper conversation" (P9).

In analyzing the associated phase I processing, we find that these perceived breaks in conversation occur when the system maps many specific meanings onto one system-understood meaning (e.g. understanding "not happy" as "depressed" or "talk in private" as "not communicative"; both "depression" and "communication" are within the domain of the dramatic world, while the related concepts are not). However, the shallow representation of semantics is not enough on its own to account for perceived conversational breaks; in the data there are many instances of the nuances of player input being mapped to a simpler meaning, but only a subset result in perceived breaks. The selected phase II reactions provide the key. In all cases where shallow semantics result in perceived breaks, the selected reactions included the characters specifically describing their understanding of the player utterance (e.g. Table 2-d). Rather than subtly training the player in the system's level of understanding, as was intended, such character reactions appear to destroy the illusion of a real conversation.

5.2.3 Reactions to Reverse Meanings

Players sometimes feel that characters are reacting to a meaning that is the exact opposite of that intended by the player:

"She is thinking I am saying her decorating is bad when I am saying the exact opposite I don't know how they could have interpreted that as negative." (P2, table2e)

Players generally have strong negative reactions to such reversals of meaning, experiencing not only a break in the conversation, but frustration that the characters are hearing the opposite of what they intend to say. In the associated phase I and phase II processing, such reversals of meaning occur when phase I produces conflicting discourse acts (error category iv from section 5.1), and phase II selects the wrong meaning. In general, phase II prefers responding to negative discourse acts (disagreements, criticism, opposition, etc.) rather than positive ones, since negative discourse acts provoke more dramatically interesting responses. Given this phase II heuristic, if phase I incorrectly recognizes a negative discourse act in addition to a positive one for a positive utterance, phase II will tend to respond to the wrong discourse act (the one that is opposite of the player's intended meaning). Interestingly, analysis of player responses to phase I errors reveals that failure to understand any discourse act (a phase I category iii error) results in much smaller perceived conversational breaks due to narrative interpretations made by the player. That suggests that in cases of a conflicting phase I discourse acts, a better technical strategy is to ignore player input (treating it as not understood) rather than risk phase II responding to the wrong act. This is easily accomplished by a) having a pre-defined list of conflicting discourse acts and b) producing no discourse act when conflicting discourse acts are recognized. The success of this approach, however would require another study correlating player experience using both the old and the new phase I approaches.

5.2.4 Reactions to Conversation Pacing

Players commented on the need for appropriate 'timing' in order to interact well with Façade. Some players felt that the conversation moved at a fast pace like a "run-away train" (P3) causing characters to "move on" (P5) before they had sufficient opportunities to address a particular topic.

Players felt that losing the timing during the interaction prevents them from having a better experience: "At this point I was feeling a bit removed b/c I lost my timing… Once I lost it, it kind of felt like it was gone for the whole rest of the time….." (P12). Some players felt that characters "paused" in response to player's input and it was not possible to say anything at any time as it "disrupts the flow" (P2) .

The need for players to establish appropriate timing with system, which has been anecdotally observed in various public demonstrations of Façade, relates to the concept of entrainment [4] used in dialog systems. Players who are able to have richer interactions with Façade are able to adapt themselves to the interaction flow. An interesting direction for future work is developing adaptation strategies for dynamically adjusting

conversational pacing so as to achieve entrainment in interactive dramas.

5.2.5 Reactions to Reference Problems

In order to resolve which character a player utterance is directed at, Façade's NLP system adopts the following strategy. If the player directly indicates in the utterance who the utterance is directed at (e.g. "Trip, I don't like that"), use that reference. If the player doesn't explicitly direct the utterance at a character and the player has only one character in her view cone (is looking at only one character), direct the utterance at that character. Finally, if neither condition holds, the utterance is directed at the character who most recently spoke. This heuristic results in player confusion when the character reference shifts in the middle of a player utterance. For example:

Trip: No we need -- we need to talk about us both, not just one.

(Player starts typing "ok" in response to Trip, however Grace says the following while the player is typing)

Grace: Adam, you -- you blame me for all this, don't you?

Player: ok

Phase 1 output: DAAgree Grace

In this case, the player is attempting to agree with Trip. Since the implicit character reference switches during the player utterance, the player ends up agreeing with Grace, which in this case results in a strong negative interaction (since the player is agreeing that Grace is to blame for all the marriage problems). Players notice these difficulties: "That was weird. If I am talking after him then it means I am talking to him" (P5). In our data, this issue occurs in only 1% of the conversational turns, but causes large conversational breaks (like reverse meanings, players end up saying something, often provocative, completely different than they intend).

5.2.6 Believable Character Performance Maintains Engagement

Even in situations where the player is aware that the system has incorrectly understood their input, character reactions can still maintain engagement. Not that this is a different case than 5.2.1, in which players employ narrative interpretations to mask a conversational break. In this case, the player experiences a conversational break, yet the details of the character performance maintain engagement. Player comment on humor "His reaction was funny, made me laugh" (P5), and on the "mood and tone" (P10) of the character responses: "She is pissed at me when I asked her, Look at her, she is staring at me, straight faced" (P7, Table 2-f). These various positive comments indicate that, even during conversational breakdowns, believable verbal and non-verbal character performance can help maintain engagement.

6. Discussion

Researchers have argued that embodied conversational agents are more comprehensible if they provide visible cues to support users in constructing narrative explanations for the agent's actions [19]. During their interaction with Façade, players generate interpretations to bridge AI technical failures and to integrate the characters reactions into the ongoing narrative. Through deflection strategies and believable character performances, Façade's design appears to provide the necessary support for making sense of character limitations within the story context. However, inaccurate interpretations of player statements based on a shallow understanding can result in negative reactions by players.

Developing a natural language system that can understand all the nuances related to a topic is notoriously hard. Moreover, in-depth conversation on a topic requires significant content creation. As a design choice, Façade uses different strategies to both deflect back to the main conversation as well as limit the depth in which players can drill down on any one topic. Our study indicates that these deflections, when carried out using real-time, believable character performance, are successfully integrated into the player experience. Players feel as though the deflections are an inherent part of the story and thereby help sustain player interest in the storyline. Further, believable character performance can maintain player interest and engagement even when the player is fully aware that a conversational breakdown is occurring.

Traditionally, to counter speech recognition and natural language processing errors, task-based conversational systems have employed the strategy of seeking confirmation of the understood information in order to move the system forward. As a design choice, traditional systems have always informed the user of what the system has understood [5,10]. In contrast, our results indicate that, in an interactive drama, directly employing shallowly understood player meanings as part of verbal output can produce strong negative reaction from players. Using the understood information gives the player a clear perception of the understanding capabilities of the characters. This tends to break the illusion necessary for the player to feel they are having a real conversation with the characters.

Human machine conversational research [3, 12] points out that people readily adapt to and emulate the conversational styles of their partners. This research suggests that dialog efficiency and user satisfaction can be increased if spoken dialog systems adapt to the user's choice of terms rather than staying within their own fixed vocabulary. Entrainment failures experienced by Façade players indicate another adaptation opportunity. Real-time, non-turn-based dialog systems should dynamically adjust conversational flow so as to maximize player entrainment.

7. Conclusion

In this paper, we evaluate a conversation centered interactive drama by uncovering the rationale and richness behind participants' subjective experience at moments of perceived conversation breakdown. Furthermore, using player perception as a guide, we examine the relationship between the processing occurring in the NLP system and the player's perception of the conversation at points of (potential) conversation breakdown. We thus present a new methodology for evaluating non-task-based, real-time conversation systems.

We employ our approach to evaluate the technical and design strategies in the interactive drama Façade. Our results indicate that Façade succeeds as an experience as it provides players, even at moments of technical failures, a) sufficient narrative cues to integrate the characters' reaction in the ongoing narrative and b) enough opportunities to maintain their interest through believable character performance. Our findings emphasize that these design strategies would be useful for future conversation centered drama systems as they help maintain player interest despite AI technical failures.

We show that using shallowly understood information as part of the characters' verbal output hampers player experience. Our study also raises the possibility of future research directed at dynamically adapting real-time conversational pacing so as to maximize player entrainment.

REFERENCES

[1] Bates, J. Virtual Reality, Art, and Entertainment. Presence: Teleoperators and Virtual Environments, 1(1), 133-138, 1992.

[2] Bernsen, N. O. and Dybkjær, L.: Structured interview-based evaluation of spoken multimodal conversation with H.C. Andersen. In *Proceedings of* ICSLP 2004, 2004.

[3] Brennan, S. E. The grounding problem in conversation with and through computers. In S. R. Fussell & R. J. Kreuz (Eds.), Social and cognitive psychological approaches to interpersonal communication pp. 201-225, 1998.

[4] Brennan, S. E. Lexical entrainment in spontaneous dialog. Proceedings, International Symposium on Spoken Dialogue, pp. 41-44. Philadelphia, PA: ISSD-96, 1996.

[5] Gorin A., Riccardi G., and Wright J., How may i help you?, *Speech Communication*, 23, 113–127, (1997).

[6] Hirschman L. Human language evaluation. In *Procedings of ARPA Human Language Technology Workshop*, pp. 99–101, 1994.

[7] Hirschman L. and Pao C. The cost of errors in a spoken language system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pp. 1419–1422, (1993).

[8] Knickmeyer, R. L. and Mateas, M. Preliminary Evaluation of the Interactive Drama Façade. In *Ext. Abstracts CHI2005,*.

[9] Höök, K., Sengers P. and Andersson G. Sense and Sensibility: Evaluation and Interactive Art. In *Procedings of CHI 2003*, ACM Press (2003), 241 – 248.

[10] Lamel L., Rosset S., Gauvin J. L., Bennacef S., Garnier-Rizet M., and Prouts B. The Limsi Arise System, *In Proc. of IEEE 4th workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 209–214, 1998.

[11] Larsen L. *On the Usability of Spoken Dialogue Systems*, Ph.D. Thesis, Aalborg University, July, 2003

[12] Laurel, Brenda K. (1991): Computers as Theatre. Reading, MA, Addison-Wesley Publishing

[13] Magerko, B. and Laird, J.E. Building an Interactive Drama Architecture. In *Proceedings of TIDSE* 226-237, 2003.

[14] Mateas, M. and Stern, A. Façade: An Experiment in Building a Fully-Realized Interactive Drama. *In Game Developer's Conference: Game Design Track*, 2003.

[15] Mateas, M. *Interactive Drama, Art, and Artificial Intelligence* Ph.D. Thesis. Technical Report CMU-CS-02-206, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. December 2002.

[16] Polifroni J., Hirschman L., Seneff S., and Zue V. Experiments in evaluating interactive spoken language systems. In *Proceedings of the DARPA Speech and NL Workshop*, pp. 28–33, 1992.

[17] Rosis F. de, Cavalluzzi A., Mazzotta I. and Novielli N. Can embodied conversational agents induce empathy in users? In *Proceedings of AISB'05 Virtual Social Characters Symposium.* Hatfield, April 2005.

[18] Sanders, G., A., Scholtz, J. *Measurement and Evaluation of Embodied Conversational Agents,* In Embodied Conversational Agents, MIT Press, Cambridge, 2000.

[19] Sengers P. *Narrative Intelligence.* In Human Cognition and Social Agent Technology, 2000.

[20] Stasko, J, Miller T., Pousman Z., Plaue C., and Ullah O.. Personalized Peripheral Information Awareness through Information Art. In *Proceedings of UbiComp 2004,* ACM Press (2004), 18 – 35.

[21] Strauss A, Corbin J. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Sage, 1990.

[22] Young R. M., Riedl O. M., Branly M., Martin R. J., Saretto C.J. An architecture for integrating plan-based behavior generation with interactive game environments, *Journal of Game Development*, 1, 2004.

[23] Dow, S., Mehta, M., MacIntyre, B., and Mateas, M: Presence and Engagement in an Interactive Drama Accepted to Conference on Computer Human Interaction, 2007

[24] Dow, S., Mehta, M., Lausier, A., MacIntyre, B., and Mateas, M.: Initial Lessons from ARFaçade, An Interactive Augmented Reality Drama, In *ACM SIGCHI Conf. on Advances in Computer Entertainment,* 2006.