Eliza meets the Wizard-of-Oz: Blending Machine and Human Control of Embodied Characters

Steven Dow
HCI Group
Stanford University
353 Serra, Stanford CA
spdow@stanford.edu

Manish Mehta, Blair MacIntyre
School of Interactive Computing
Georgia Institute of Technology
85 5th Street, Atlanta, GA
{mehtama1, blair}@cc.gatech.edu

Michael Mateas
Computer Science Department
UC Santa Cruz
1156 High Street, Santa Cruz, CA
michaelm@cs.ucsc.edu

ABSTRACT
What authoring possibilities arise by blending machine and human control of live embodied character experiences? This paper explores two different "behind-the-scenes" roles for human operators during a three-month gallery installation of an embodied character experience. In the Transcription role, human operators type players' spoken utterances; then, algorithms interpret the player's intention, choose from pre-authored dialogue based on local and global narrative contexts, and procedurally animate two embodied characters. In the Discourse role, human operators select from semantic categories to interpret player intention; algorithms use this "discourse act" to automate character dialogue and animation. We compare these two methods of blending control using game logs and interviews, and document how the amateur operators initially resisted having to learn the Discourse version, but eventually preferred having the authorial control it afforded. This paper also outlines a design space for blending machine and human control in live character experiences.

Author Keywords
Embodied characters, artificial intelligence, Wizard-of-Oz methods, interactive drama

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

General Terms
Design, Experimentation

INTRODUCTION
Interactive conversations with embodied characters show tremendous potential for education and entertainment [1,2,8,9,17]. To date, purely algorithmic means of emulating face-to-face conversation show promise, but must overcome numerous technical challenges from recognizing audience speech and gesture input, to determining user intentionality, to performing character actions based on the conversational and emotional contexts [7,27,29]. Creating working prototypes capable of yielding player feedback can take years [23,24]. Alternatively, controlling embodied characters with human actors can be emotionally engaging, but faces the practical constraints of theatrical performance [20,30]. Maintaining character consistency and showmanship requires interactive entertainment venues, such as Disneyworld, to employ dedicated professional casts and crews and limits their ability to satisfy large daily audiences [1,2,3].

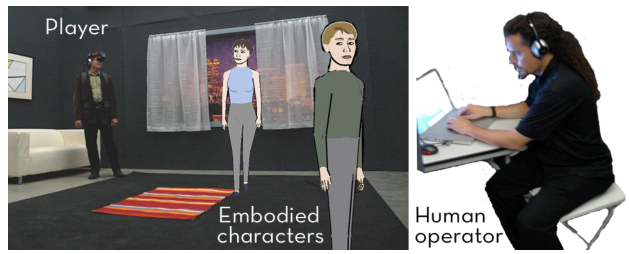

Figure 1: In the AR Façade experience, the player (left) has an interactive conversation with two embodied characters (center); nearby, a human operator (right) observes player interaction and provides input for the machine algorithms.

This paper proposes a design space for controlling embodied characters through a blend of human operators and machine algorithms. One dimension explores how to blend control: on one extreme, the operator does everything, and on the other, the system is entirely automated. Another dimension of the design space is the acting ability of the operator. Disney's character experiences currently use "high-expertise" actors; this research explores using amateurs or non-performers. This paper expands on these dimensions of the design space and examines how amateur operators behave using two different blends of control.

We report on the experiences of nine amateur operators in the embodied character experience *AR Façade* [11,12]. In this experience, a player engages in live conversation inside a full-sized augmented reality (AR) apartment with a married couple, life-size embodied characters named Trip and Grace, who respond interactively to the player's actions and speech (see Figure 1). A nearby human operator controls the experience by observing player actions and providing input to machine algorithms that decide what the characters do and say.

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

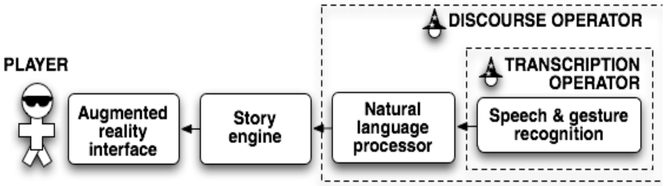

Figure 2: Architecture for AR Façade, highlighting the extent of responsibility for discourse and transcription operators.

Two different methods of blending machine and human control were deployed and observed during a three-month real-world gallery deployment. In the *Transcription* version, human operators transcribe player statements and press buttons corresponding to pre-designed game gestures. The natural language parser (NLP) then processes the text into semantic categories, which pass through the rest of an adaptive story system. In the *Discourse* version, the NLP is removed and human operators manually select semantic categories embedded in the story infrastructure (See Figure 2). In this role, human operators interpret player intentions.

A comparison of these two roles is based on interviews with the nine undergraduate students who performed as human operators and game logs from 140 unique player episodes. The analysis provides evidence for the following:

- Both human-machine control methods successfully leveraged amateur operators to engage audiences in live conversations with embodied characters;
- The Transcription method was easy for amateur operators to learn, but often suffered delays and misinterpretations in producing character responses;
- The Discourse method required more effort for operators to learn the categories of player intention, but became the preferred interface as amateur operators adopted a greater degree of control of conversational flow between players and characters.
- The amateur operators provided subtle and actionable design insights for subsequent iterations of the system.

This paper discusses the goals, challenges, and technical approaches for developing embodied characters; examines human-controlled entertainment experiences and "Wizard-of-Oz" methods in HCI; proposes a design space for blending human and machine control in embodied character experiences; explains *AR Façade's* technical architecture, including the two versions of blending machine and human control; and describes an empirical investigation during a three-month gallery installation. The analysis summarizes game logs and interviews with amateur operators, and discusses the implications of blending control of machine and human operators for real-time performances.

CONTROLLING EMBODIED CHARACTERS

Embodied characters are physical or animated representations of agents designed to be conversational in behavior [8,29]. The intricacies of human speech, gesture, facial expressions, emotions, and behavior provide research focus, both for sensing player input [15,16,19] and controlling characters [14,28]. The goal is to establish a high-bandwidth, high-speed, highly emotional feedback loop between the audience and embodied characters. The design and development of embodied character experiences typically follows one of two broad strategies. One path strives for the creation of artificially intelligent computer software agents. Another approach can be traced to live performances—such as puppetry [14]—where human actors directly command the interactive features of characters. This section discusses each strategy and outlines a design space for blending the two approaches.

Machine Control of Agents and Interactive Story

The research on fully-automated embodied characters seeks to recognize speech and gesture input, determine user intentionality, and choose character actions based on the conversational context. Research projects often address a specific machine-learning sub-system, including: natural language processing [19,24], verbal and non-verbal behavior generation [6,35], player modeling [31], narrative beat sequencing [27], and character animation [9,28].

Combining the necessary AI sub-systems into a completely automated interactive conversation with characters has met mixed results. Such experiences often suffer from slow reaction times, misinterpretations of human language and emotion, and uncoordinated facial, gestural, and verbal cues [9,26,29]. While some systems circumvent natural language misinterpretations by requiring a command language [29] or restricting the conversational possibilities to a narrow context (e.g., ELIZA [34]), these approaches can undermine the illusion of a real conversation.

Other research places embodied characters into a narrative arc with plot, motivation, and desires. Rather than strive for open-ended conversation, Mateas argues that providing context for player action—in the form of material and formal constraints—gives the player a greater sense of agency [22]. Mateas and Stern employed this principle in the game *Façade*, which integrates generative character animation and behavior, drama management, story memory, and natural language processing [23]. The *AR Façade* system goes a step further, embedding the player in an augmented reality version of the game and enabling speech and physical gesture interaction. Empirical studies of the AR version demonstrate increased player *presence*—a sense of being there—but not necessarily *engagement*—a sense of deep involvement. Some players preferred being "outside" of the drama [11].

Technological improvements do not insure embodied character experiences will be engaging for players. Wardrip-Fruin et al. argue that audiences often suffer a mismatch of expectations when interacting with narrative systems and that designers need to "transition" audiences to understand the underlying computational model [33]. Since embodied characters appear and behave like humans, players often expect real conversation. Players can be disappointed when these expectations are not met. This paper proposes leveraging human operators during live experiences to help narrow this expectation gap.

Emulation through Human Actors and Wizards

Human-controlled embodied characters avoid the technical challenges of automatically recognizing and responding to audiences, but currently require full-time actors to portray characters' voices and mannerisms. Disney's "Living Character Initiative" seeks to provide audiences with live improvisational interaction with robotic and animated characters, using trained actors to control movement, expression, and voice. Examples include Turtle Talk with Crush [2], Ratatouille's Remy [1], and the Muppet Mobile Lab with Bunsen & Beaker [3]. The challenges of human-controlled embodied characters are similar to those in theatre productions—from hiring trained actors and maintaining day-to-day energy, to scheduling issues.

Another strategy enables amateurs (i.e. non-professional actors) to deliver high-quality experiences. In role-playing games, human operators intervene as "game masters", facilitating the story world and deciding the outcome of game events not controlled by players or determined by chance [32]. While these games are traditionally played in face-to-face settings, recent role-playing games leverage game masters to guide the player experience in pervasive games [18,25]. Game-mastering practices can offer insights into the roles operators can play in embodied character experiences.

In HCI research, people commonly emulate part of an interactive system. This "Wizard of Oz" (WOz) method shortcuts the prototyping process for novel user interaction techniques, including speech, gesture, multimodal, context-aware, and location-based applications [10,13,16,21]. The WOz method is well suited for speech interfaces, but as Dybkjær et al. point out, a human operator can provide smoother speech interaction than can be realistically achieved with technology alone [13]. Klemmer et al.'s research on speech interfaces simulates technology constraints by artificially inserting random errors on top of human input [21]. In *AR Façade*, rather than simulating realistic system error, human operators typed player statements as quickly and accurately as possible, placing priority on player engagement.

Blending Machine and Human Control

To overcome challenges presented by purely algorithmic-based and human-based approaches, we propose seven design dimensions for blending machine and human control of embodied character experiences (Figure 3). Different combinations of these dimensions reveal design opportunities. Human operators can fulfill different roles (e.g., separate AI subsystems) within a complex interactive conversational system. One or more operators with varying acting abilities can perform onsite or remotely. The system can offer various levels of semantic inclusiveness, from a limited number of conceptual categories to the entirety of human language and gesture. These preliminary dimensions are not orthogonal; a choice on one dimension can limit choices on others.

A creative example of blending machine and human control is the animatronic character, Quasi the Robot [4]. A professional actor produces Quasi's voice by speaking through a voice modulator. Notably, the operator also controls the robot's gestures by selecting emotional characteristics from a palette (e.g., sad, happy, excited, frustrated). Rather than having direct control of Quasi's ears, eyes and arms, the operator selects an emotional state, and (relatively simple) algorithms do the rest.

AR Façade's Discourse method of control also represents high-level semantic categories. However, rather than presenting the human operator with characters' future emotional states, the discourse interface presents player intentions; story algorithms infer the characters' emotions.

SYSTEM ARCHITECTURE FOR AR FAÇADE

The immersive and interactive drama *AR Façade* is an augmented reality version of Mateas' and Stern's 2005 interactive drama, *Façade* [23]. In this experience, players enter a virtual apartment with two embodied characters—Trip and Grace—and have a conversation as if they were

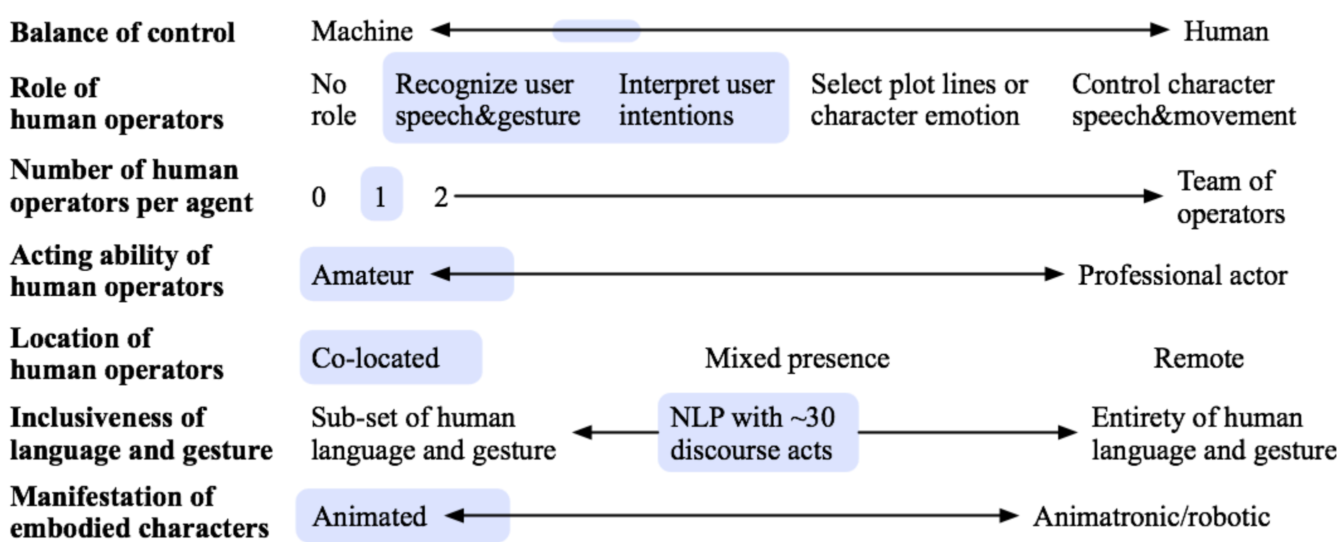

Figure 3: A design space for blending machine and human control of embodied character experiences (highlights denote exploration in this research).

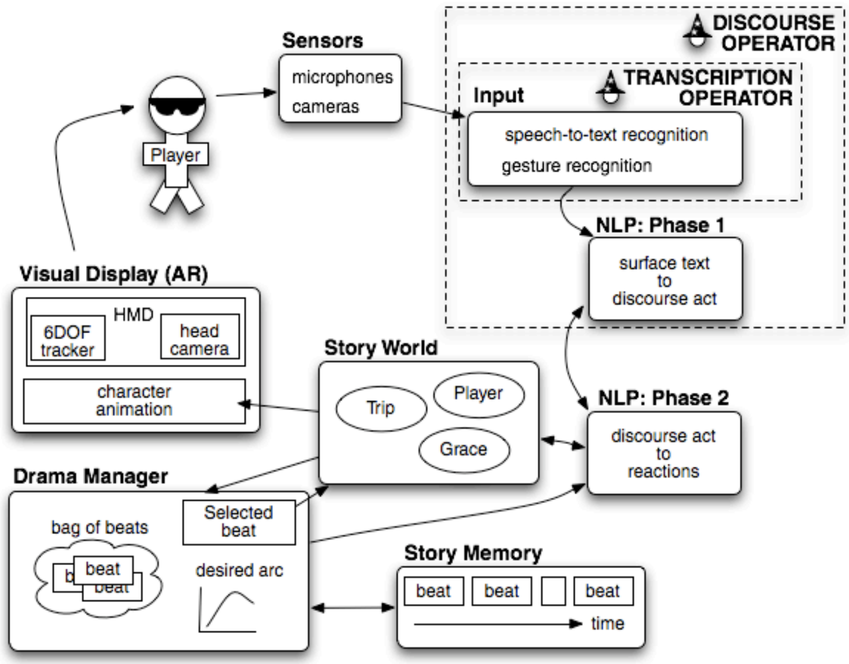

Figure 4: System architecture for *AR Façade* showing the extent of control for Transcription and Discourse operators.

old friends. One unique contribution of *Façade* is the architectural support for authoring dramatic beats, or short story segments that can be dynamically sequenced based on player input and the desired dramatic arc, thus attempting to combine player interaction and structured narrative. Rather than directly mapping player input to character output, *Façade's* AI story engine models the characters' emotional states and attempts to choose lines of dialogue based on local and global contexts. The AR version of *Façade* goes further to "embody" the player by replacing desktop (mouse and keyboard) interaction with unconstrained speech and gesture interaction. *AR Façade* comprises the following modules (see Figure 4):

- A speech and gesture recognizer (provides textual representation of player utterances and discrete physical actions by the player);
- A natural language parser (classifies text input to one of 30 or so parameterized discourse acts);
- A drama manager (sequences dramatic beats based on the current state of the world and memory of previous beats);
- Two procedurally-animated character agents (executes beat-specific behaviors through language, movement, facial expression, eye gaze, etc.); and,
- A non-photorealistic rendered 3D virtual story world graphically overlaid in real-time on live video feed from a tracked camera in a physical space.

The natural language parser (NLP)—central to this paper's discussion—processes player surface text into one or more "discourse acts" with additional parameters (see Figure 5, right). For example, if the player says, "you look beautiful Grace," the NLP would identify the flirt discourse act and select Grace as a parameter.

Façade's discourse acts (agree, disagree, pacify, criticize, refer to items in the room, etc.) are specific to the story, setting, and the emotional tenor of the experience. Despite limitations inherent in representing a small subset of human language, the discourse categories are still general enough to encapsulate much of the dialogue that arises during game play. The story infrastructure uses the discourse act selected by the NLP, queries the current beat from the drama manager, and sends possible reactions to the character agents where procedural animations are handled.

This paper explores two methods for human operators to control the system: Transcription and Discourse selection. In the Transcription version the AI engine's natural language processor (NLP) and drama manager primarily handle the player's experience. In the Discourse version, the NLP is deactivated and the wizard directly selects "discourse acts," or player intentions. While both methods still require a significant story infrastructure, the Discourse method obviates the need for *Façade's* NLP.

Transcription Interface

In the Transcription interface, human operators handle the player's speech and gesture input (see Figure 5). The Transcription interface has a series of buttons for handling object references and specific player gestures. The interface includes a text field at the bottom for typing player statements. After the operator enters text—essentially serving as a speech-to-text converter—*Façade's* NLP takes over and calculates the most appropriate discourse act. *Façade's* NLP imposes a 35-character limitation to simplify the text analysis problem. The operators must also type everything the player says within this buffer limit.

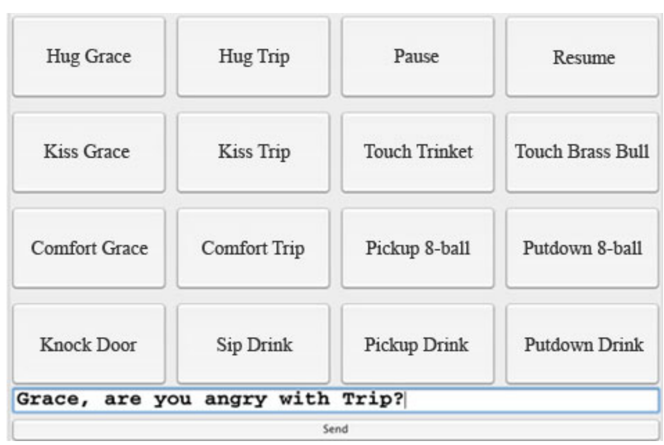

Figure 5: In the transcription interface, operators type player statements in the text box and press buttons for player actions.

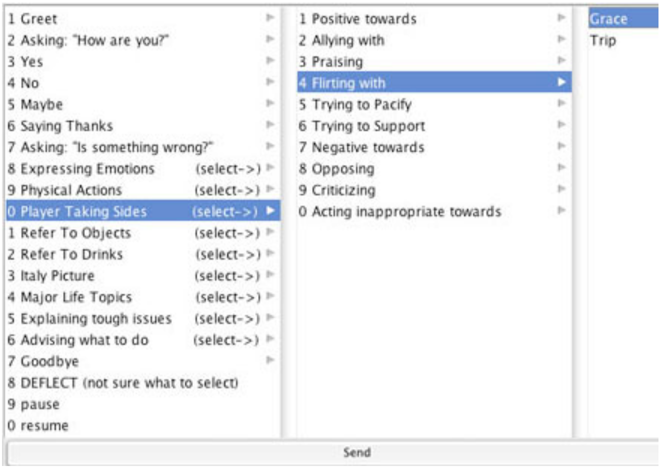

Figure 6: In the discourse interface (right)—operators select pre-authored categories corresponding to player intention.

Discourse Interface

In the Discourse method, the NLP has been removed and human operators directly trigger *Façade's* higher-order discourse acts. Rather than typing out what the player says, the operator selects an item that matches what they think the player *intends*. The Discourses interface (see Figure 6) organizes all 30 discourse acts and their parameters in a hierarchy across three columns. Items can be selected through number keys, arrow keys, and the mouse. The most common discourses are placed directly in the first column (e.g., Greet) with key parameters in the second column. Less common discourses are organized under a categorical heading, so that discourse acts fall in the second column and parameters in the third column, as shown in Figure 5.

As a scenario: if the player says, "you look beautiful Grace," the human operator would go to the category called *Player Taking Sides*, then select *Flirting With*, and then select *Grace*. Under the hood, the system directly triggers the same discourse act representation as would be selected by the NLP. Most player statements are open to interpretation and it's the job of the human operator to decide on the appropriate discourse category.

METHOD

This research explores two different operator roles for amateur operators with the goal to uncover design tradeoffs. A three-month installation of *AR Façade* at a public art and technology gallery garnered game play data for 140 episodes of the experience. This paper focuses on the nine part-time employees at the gallery who served as amateur operators and carried out the experience for players. All nine operators were recruited and hired by the gallery and shared similar characteristics: 19-23 years old, female, and majoring primarily in art-related topics. None of the amateur operators had prior experience performing as a real-time wizard and none had extensive computer experience.

Procedure

The amateur operators were trained to give the player a short description of the experience, to demonstrate the gesture-based interactions, and to help the player into the equipment (i.e., head-mounted display (HMD), backpack computer, and headphones). For each *AR Façade* episode, the operator walked the player to the front door of Trip and Grace's apartment and then disappeared behind the wall to perform either the Transcription or Discourse task. The two methods were visible as separate tabs of a single program, which communicated wirelessly with the player's wearable machine. The operators were allowed to use either interface (or a combination of the two). The operator could view the experience through two monitors. One monitor displayed the player's view from the HMD; the other provided overhead video of the apartment.

Data gathering

Experimenters recorded all operator activity (button presses, typed text, etc.) and conducted three open-ended interviews: the first occurred before the three-month installation (ten minutes), the second happened two weeks after the opening to adjust the operator interfaces for minor usability problems (ten minutes), and the third interview came at the end of the three months (about one hour). Experimenters also interviewed thirty-three players during the final two weeks of the installation.

FINDINGS

The amateur operators elected to use the Transcription version 84 % of the total usage time (see Figure 7). They employed a combination of the two methods (by switching mid-episode) in 44 of the 140 episodes. Initially, the Transcription interface was viewed as "easier"; the Discourse interface had a steeper learning curve. Operator 1 formed a strategy to learn the Discourse interface, "when there are pauses, I start going through the other part (Discourses) to get to know it." Over time, the operators learned the location of discourse categories in the interface and, as we illustrate below, eventually came to prefer the Discourse method and the level of authorial control it provided.

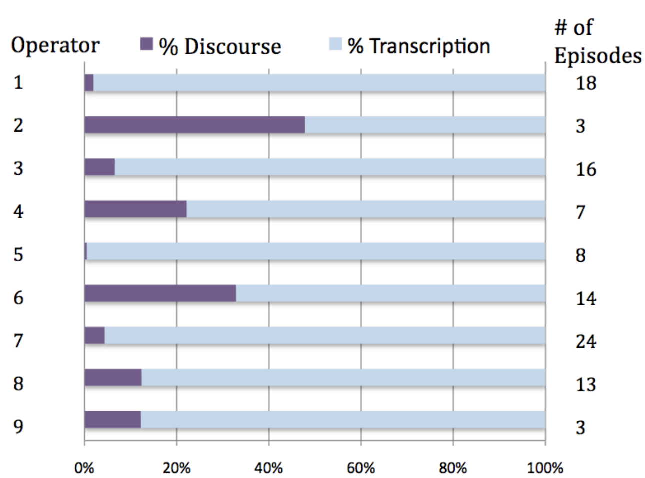

Figure 7: Operators used the Discourse method 16% of the time

The amateur operators successfully performed both tasks during the gallery installation (mostly without direct supervision). One of the biggest challenges across both control methods was simultaneously performing gesture recognition along with speech. Operator 3, for example, claimed it was "hard to concentrate on the TV (monitor) and on the menu at the same time." As a result, gesture recognition suffered with both versions of the operator task. However, given *Façade's* emphasis on conversational game play, gesture recognition was less essential to the player experience. Beyond the context of this research study, simple solutions to this simultaneous input problem exist, such as adding a secondary operator to perform the gesture recognition task.

Issues with the Transcription Method

In the Transcription method, the wizard types the player utterance. While the task is straightforward, several issues arose. For example, spelling errors were common, but easy to eradicate if the operator noticed she typed in the statement incorrectly. Occasionally the operator would type in and enter something that could be misinterpreted by the NLP. For example, one operator typed "hell" instead of "hello," which led the characters to act like the player was aggravated.

Operators had difficulty when players spoke particularly fast or long-winded statements. Players were not constrained by what they could say, nor were they told about the NLP-imposed buffer limit. As a result, players tended to speak freely, using utterances much longer than 35 characters, which the operator must then paraphrase. According to O2: "It's hard to remember the certain amount of letters that you can type…I can't type everything." To deal with the buffer limit, operators developed various coping strategies. Operators would paraphrase the player's statement on the fly, trying not to distort its meaning. For example, when one player said "do you have issues with your parents, Trip?" the operator anticipated that this would be too long for the buffer limit and typed, "you don't like your parents?" This example would have no significant effect on the player experience, as the NLP interprets both statements similarly.

Sometimes operators would split a player utterance as two entries. In one example, the operator typed "but you've been together ten" and then ran into the end of the buffer. She then entered that statement and added the final word "years" as a separate entry. Since the NLP interprets these splits independently, the characters' responses may be strange. By paraphrasing, operators did influence the NLP, but it is unclear to what degree this impacts the player experience. Most players did not suspect the presence of a human operator, and seemed to attribute most errors to the primitive nature of speech recognition or other technical limitations.

Issues with the Discourse Method

In the Discourse entry method, the operator selects categories corresponding to probable player intentions. In general, operators understood what they were supposed to do, as Operator 6 explains:

> It's kind of like guiding…the story. I feel like I have to interpret more what the players are saying. So it's more involved…you have to pay more attention. (O6)

Cognitive Load

Most operators felt the Discourse method required more thinking and attention, in part because it was difficult to immediately know how to categorize statements. Operator 6 said it was her role "to translate what the people are saying so that Trip and Grace can understand it." According to Operator 8, "sometimes I would have to think, okay, should I select the one that says 'the player's angry' or 'the player's trying to criticize'?" While some statements could logically fit into multiple categories, other player statements fell completely outside of *Façade's* discourse lexicon. As Operator 4 recalled:

> One of the girls tried to slap Grace [laughing]…I wasn't sure what to do. I entered 'negative towards Grace' but that's not really the same as physical violence. (O4)

Operator 3 pointed out her approach to this situation: "when (the player) said something that wasn't on the menu, I tried (selecting) something that was close to what they were saying." Other operators dealt with uncertain utterances more passively. Operator 2 would "let it pass by like nothing was said" knowing that the story engine was robust enough to keep the conversation going without explicit player input.

Experimenting with the Interaction and Story

Several operators talked about experimenting with the interaction and introducing discourse acts even if the player said nothing. Operator 8 commented that during lulls in the conversation, she would still enter discourse acts:

> I will select something like 'Therapy', just to offer a little variation... Because some people would be a little passive in their interactions. (O8)

Operator 9 selected unprompted discourse acts, because she wanted to liven up the conversation:

> I clicked 'Have sex' or something because I was hoping that some big explosive thing would happen... I thought that would be fun to see, because Grace seems like kind of an intense chick" (O9)

These operators controlled the story like actors in the Disney living character experiences. The operators wanted to make "interesting" things happen for the player, and so the operators purposefully deviated from just processing input. There were consequences to this experimentation, as Operator 4 found when using the 'Oppose Trip' discourse: "I guess that was too strong of an emotion, because [Trip] kicked him out [laughing]." While problems did occur and the dialogue was not technically "authentic" in terms of the designed *Façade* experience, the operators' agency by and large provided new opportunities for player engagement.

Anticipating Player Statements

Operators formed several useful strategies for dealing with language uncertainty in the Discourse method, especially as the operators became familiar with *AR Façade* story lines. Operator 2 said:

> When (the player) got to a heavier issue, I was sort of expecting ... either the person was gonna comfort or intervene … I was hovering on those two.” (O2)

Likewise, Operator 6 expressed this notion of anticipating the player, saying “I think she's going comfort him, so I'm gonna go hang out in the ‘comfort’ area of the program.” She mentioned her strategy was to wait for the right moment to enter a discourse: “I wait for Trip and Grace to stop saying things before I click (the button), because (entering the discourse) usually cuts them off in the middle of their sentence.” Operator 6 recognized a subtle design feature in *Façade*—character interruptability—and adjusted her performance as she saw fit. Over time, operators could anticipate story lines and player reactions; this provides an advantage over the Transcription method where operators cannot predict verbatim how a player will phrase their next utterance.

Tradeoffs of Transcription vs. Discourse

Comparing the two different methods of blending machine and human control revealed a few important tradeoffs.

Learnability versus Authorial Control

From the operators' perspective, the Transcription interface required less thinking, less pressure to perform, and less investment in the players' enjoyment level. According to Operator 6:

> When you type (the player's statement) you don't have to think about it…the computer will handle it. If it doesn't understand what (the player) is saying then it doesn't understand. (O6)

When operators used the Transcription method, they relied on the NLP to correctly interpret player statements. It was not until the operators gained more experience with *AR Façade* that they realized the system was “not going to recognize everything…” (O8). The Discourse method, on the other hand, required more attention and deliberation, as Operator 3 pointed out, “picking stuff out requires more thinking.” Operator 4 contemplated:

> (Discourse selection) forces your mind to kind of think in a different way, of not just directly translating specifically what they're saying but kind of attributing it to a larger category of emotion or actions. It depends a little more on your interpretation. (O4)

While the Discourse selection method created a greater cognitive demand, it also provided more opportunity for crafting the audience experience. Operators expressed their ability to pick up on the nuances of player emotions, as one operator said

> I can tell when someone feels awkward or when people are getting really annoyed by just like the tone of their voice. (O3)

With this level of insight, some operators claimed, “you could definitely shape the player's experience” (O9). Operator 1 even described the characters as puppets, saying, “I'll make Trip talk about the picture again and hopefully it will guide [the player] over.” Operator 3 went as far to say she could inflict drama “like a voodoo doll…” and that performing as an operator was “almost like playing God.”

Number of Discourse Categories

The proper number of categories for the Discourse method was a point of disagreement among the operators. Some operators felt there was not enough nuance, as Operator 4 pondered, “How should I generalize this emotion?” Operator 3 said “there's not enough vocabulary for everything the players wanted to say …I just wish there were more things.” Meanwhile, Operator 5 said, “I don't think there should be more categories right now. It will be a really long list and you'll be like, ‘Ugh, which one?'” The dispute raises an important tradeoff regarding the number of high-level decision points, summarized by Operator 6:

> I think it would help having more categories but then at the same time that's detrimental to quickly figuring out where things are (in the interface) because that just means more things to look through. (O6)

Conversational Flow

In *AR Façade*, the Discourse method seemed to provide better affordances for keeping the player conversation flowing. For one, the method is less susceptible to technical issues, such as a poor audio connection. As Operator 1 stated, “Even when I didn't understand what they said, I can at least, you know, pick out a keyword and click that” (O1). The flexibility of the Discourse method allowed operators to assert obvious player meanings, but also to identify subtle player intentions that would be missed in surface-level natural language text processing.

Operators had mixed views on whether the Discourse or Dialogue led to faster character response times. According to Operator 1, “typing it out takes a little bit longer than searching and clicking,” while Operator 4 says “it takes a little longer to sort through them and find the right one then it does to just immediately translate what they're saying into text.” We estimate that Discourse selection time is relatively constant, whereas the Transcription time is a function of the utterance length. Long statements not only require more typing, they force the operator to paraphrase.

A detailed conversational analysis could not be conducted, given the open gallery setting and poor audio recordings. Moreover, a time-delay analysis is subject to interpretation, as many player statements have no corresponding character response and, vice versa, many character statements do not follow from a player statement. The player and the character often speak over the top of one another. Nevertheless, a detailed conversational analysis of different control methods could be an interesting area for future lab-based research.

How Operators Affected the Player Experience

Measuring the subjective player experience presents a significant challenge, especially for interactive drama where each episode is very different. The distribution of "discourse acts" provides one quantitative measure of how operators affected the experience. On average, the Discourse method triggered 30.5 discourse acts compared to only 20.8 in Transcription episodes. The fact that human operators were more "active" than the NLP in choosing directions for the conversation says less about player experience and more about the elevated level of operator engagement using the discourse method.

Overall, the distribution of selected discourse acts using the Discourse method is similar to the NLP choices in the Transcription method (see Figure 8). Only 5 of the 26 clusters differed by more than five-percent between the Transcription and Discourse method. The discourse act "Agree" showed the biggest discrepancy with a 30.1% selection rate under Transcription and 15.8% under the Discourse method. This is explained by the fact that the NLP tends to eagerly interpret utterances as "Agree" in addition to more specific meanings. Discourse operators presumably select only the more specific meaning. On average, the percent difference between control methods for each discourse act was 2.8%. Several discourses were used infrequently; for example, "Get Attention" and "Intimate" were used less than 0.2% by both control methods.

Player interviews revealed no major differences between the two control methods. Players seemed unaware of a human element in the system, as revealed in statements like, "I said to Grace that she sounded stressed and I guess the computer took it as 'depressed'... is that a problem with voice recognition?" (Player A). Even when players were impressed with how the system performed, they did not suspect a hidden operator behind the scenes:

> The technology works pretty well for me... I didn't know if there was a mic or anything around the backpack, but I was surprised that it could hear and decipher what I was saying. (Player B)

During interviews, regardless of the particular control alternative used for their episode, players tended to focus on the story, the characters, and about how they felt in the particular social scenario.

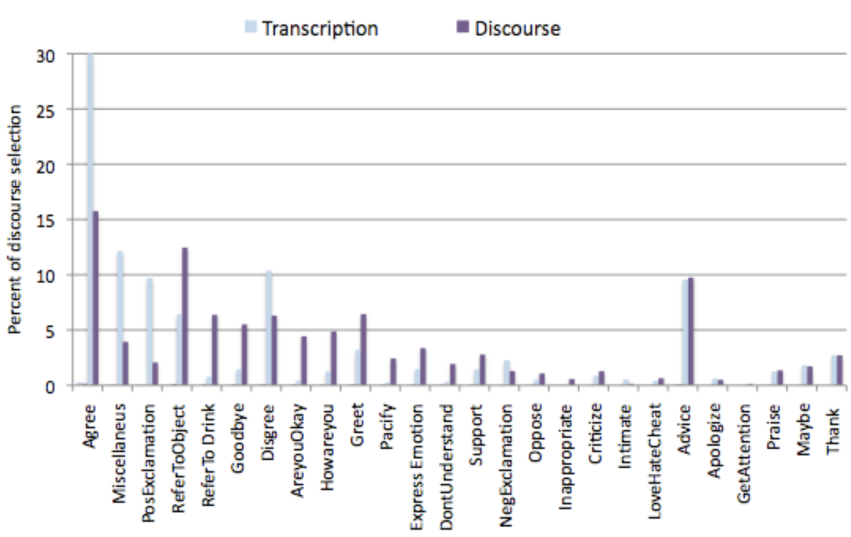

Figure 8: Percent of "discourse act" selection across 140 player episodes for both control methods: Transcription (selected via NLP) and Discourse (selected by amateur operators).

Amateur Operators as Design Partners

While the operators seemed to have fun carrying out the experience for players, they also offered insights about the players and the system. As Operator 7 pointed out, "I can tell a lot about a person in there…if they're outgoing; if they're shy; if they're creative or not; if they're smart." Operator 4 provided a rough description of player styles:

> Most people maintained a very polite kind of distanced role for the most part, but a few people acted surprised. Others were very direct and kind of blunt. And those are usually the people who would end up getting kicked out early. (O4)

Operators also had ideas for improving the underlying system. Operator 8 stated the general observation that the list of discourses included only "extreme choices that didn't really account for all the nuances." Operator 9 pointed out a potential specific issue: "sometimes when you use the criticize button, it criticizes something completely different" and "I pressed marriage and that's when he said something about 'love is blind', which seems irrelevant."

Several of the operators suggested new discourse categories based on what they had observed from players. Operator 3 wanted "more generic phrases, like 'I'm doing fine'", while Operator 8 wanted to be able to express "Can we just drop the subject?" Similarly, Operator 9 offered ideas based on an episode she carried out:

> There's no option for 'Do you want me to leave?' so …they just kept arguing and (the player) was just kind of stuck in this limbo of should I stay or should I go. Maybe if (Trip) had said like, "No, sit down," or something like that then she would have stayed longer and heard more (O9)

Operator 9 also requested that the system include an option for "what?" That way, she thought, "the characters could just repeat what they said." As the amateur operator formed an intimate understanding of the script and the common player behaviors, they were able to provide suggestions that would be useful to system developers during an early-stage prototyping process.

DISCUSSION

This paper explored the affordances offered by two different methods of blending control between machine algorithms and human operators. Both methods of operating *AR Façade* produced similar distributions of discourse acts, and resulted in effective player experiences. Most players enjoyed the installation and most did not suspect the presence of human operators behind the curtain. The amateur operators effectively adapted to both types of control. While the Discourse mode was challenging to learn, it enabled the operators to anticipate player actions and more proactively decide on conversational paths. This provided amateur operators a greater feeling of control over the story and the characters.

Future embodied character systems should explore new forms of operator control. For example, the *Façade* system could expose the emotional state of Trip and Grace and allow the operators to adjust that directly. This study suggests not only would amateur operators be capable of managing it, but they also would enjoy it and be more engaged in giving the players a good experience. The tradeoff here is consistency. The specific personalities of Trip and Grace would be placed in operators' hands, moving them more towards Disney-style expert operators, rather than amateur operators.

Referring back to Figure 3, future work should explore more dimensions of the design space for blending machine-human controlled embodied character experiences. Examining the *role of human operators*, higher-level emotional discourse acts could give human operators more nuanced improvisational ability. For example, authors may select categories like "scared," "startled," "shy," or "anxious," which are related but subtly different. Such emotions could be tied to player or character actions, or some combination of both. For embodied character experiences based on live-action role-playing, operators may simply be considered another type of player. Exploring the *location of human operators*, users may enjoy the experience of "putting on a show" for someone online. The idea of performing embodied characters online, such as with *Facebook's Pet Society* [5], opens up new dramatic possibilities for social networks.

Another dimension of the design space is the *inclusiveness of human language and gesture.* How much expressiveness can the system support? The Discourse method worked effectively with ~30 discourse categories. However, extending to a longer list of discourse categories could present significant learning challenges for amateur operators. Massive hierarchies of discourse categories could overwhelm operators. Perhaps the list of discourse acts could be dynamically updated to highlight only categories relevant to the current conversational context and dramatic story beat. Another possibility would be to develop a hybrid transcription method where operators always paraphrase or use a special command language. Although the learning curve would be steeper than the hierarchical list explored in this paper, operators could potentially become very efficient with practice. Incorporating photographs or symbols to represent actions in the operator interface (such as, a symbol for "hug" or photo of an angry Trip) may improve learnability, and make the operator task accessible to more diverse users.

Two developers spent three years each to create the *Façade* engine, one year alone on the natural language parser [24]. Mixing in human operators can help developers obtain early and authentic feedback on the features of the character, the effectiveness of the story arc, and unexpected player statements. In early prototypes of *Façade*, for example, human wizards could control "paper cutouts" of Trip and Grace to answer early questions about plot sequences and character design. Operators could simulate various software modules, such as the drama manager or beat-goal sequencing, allowing authors to evaluate the choice of discourse acts and detailed story goals, and to potentially understand how information should flow between software components. Player data could be collected throughout the process to understand how players converse in-context and to guide the NLP's development.

CONCLUSION

A gallery deployment of the immersive and interactive story *AR Façade* contrasted two methods of blending control between machine algorithms and an amateur human operator. The Transcription method was easier for amateur operators to learn, but suffered from surface-level misinterpretations from the natural language processer. The semantics-based Discourse method required more time for operators to master the categories of player intention, but eventually enabled a more proactive and prompt delivery of the user experience. This study begins to explore the design space of blending machine and human control of embodied characters for live amateur performance.

ACKNOWLEDGEMENTS

We thank the Beall Center for Art and Technology for hosting our *AR Façade* installation, Georgia Tech's GVU Center for generous financial support, and Björn Hartmann and Scott Klemmer for thoughtful comments on the paper.

REFERENCES

1. Remy and Disney's Living Character Initiative, Jan, 2010. http://www.stitchkingdom.com/2009/04/07/all-about-remy-and-disneys-living-character-initiative/.
2. Turtle Talk with Crush, Jan, 2010. http://disneyworld.disney.go.com/parks/epcot/entertainment/turtle-talk-with-crush/.
3. Mobile Muppet Lab in Epcot, Jan, 2010. http://allears.net/tp/ep/mobilemuppets.htm.
4. Quasi the Robot, Jan, 2010. http://www.interbots.com/characters.html.

5. Pet Society, Jan, 2010. http://www.facebook.com/applications/Pet_Society/11609831134.
6. Badler, N., Allbeck, J., Zhao, L., and Byun, M. Representing and Parameterizing Agent Behaviors. *Proceedings of the Computer Animation*, IEEE Computer Society (2002), 133.
7. Cassell, J., Bickmore, T., Billinghurst, M., et al. An Architecture for Embodied Conversational Characters. *Proceedings of Computer Animation And Simulation*, (1998), 109-120.
8. Cassell, J., Sullivan, J., Prevost, S., and Churchill, E.F. *Embodied Conversational Agents*. MIT Press, 2000.
9. Corradini, A., Mehta, M., Bernsen, N., and Charfuelan, M. Animating an interactive conversational character for an educational game system. *Proceedings of the 10th international conference on Intelligent user interfaces*, ACM (2005), 183-190.
10. Dahlbäck, N., Jönsson, A., and Ahrenberg, L. Wizard of Oz studies: why and how. In Proc. of the Intl. Conf on Intelligent User Interfaces (IUI), 1993.
11. Dow, S., Mehta, M., Harmon, E., MacIntyre, B., and Mateas, M. Presence and engagement in an interactive drama. *Proc. of the SIGCHI conf. on Human factors in computing systems*, ACM (2007), 1475-1484.
12. Dow, S., Mehta, M., Lausier, A., MacIntyre, B., and Mateas, M. Initial lessons from AR Façade, an interactive augmented reality drama. *Proc of ACM SIGCHI international conference on Advances in computer entertainment technology*, 2006.
13. Dybkjær, H., Bernsen, N.O., and Dybkjær, L. Wizard-of-Oz and the Trade-Off between Naturalness and Recognizer Constraints. In EUROSPEECH, 1993.
14. Engler, L. and Fijan, C. *Making Puppets Come Alive: How to Learn and Teach Hand Puppetry*. Dover Publications, 1997.
15. Essa, I. and pentland, A.P. Facial Expression Recognition using a Dynamic Model and Motion Energy. *IN ICCV*, (1995), 360-367.
16. Höysniemi, J., Hämäläinen, P., and Turkki, L. Wizard of Oz prototyping of computer vision based action games for children. *Proc of the Interaction design and children: building a community*, (2004), 27-34.
17. Johnson, W.L., Rickel, J.W., and Lester, J.C. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal on Artificial Intelligence in Education 11*, (2000), 47-78.
18. Jönsson, S. and Waern, A. The art of game-mastering pervasive games. In *Proc. of Advances in Computer Entertainment Technology,* 2008, 224-231.
19. Jurafsky, D. and Martin, J.H. *Speech and Language Processing*. Prentice Hall, 2008.
20. Kelso, M., Weyhrauch, K.P., and Bates, J. Dramatic Presence. In PRESENCE: Teleoperators & Virtual Environments, 2(1), MIT Press (1992).
21. Klemmer, S.R., Sinha, A.K., Chen, J., Landay, J.A., Aboobaker, N., and Wang, A. Suede: a Wizard of Oz prototyping tool for speech user interfaces. *Proc of User interface software and technology*, (2000), 1-10.
22. Mateas, M. A Preliminary Poetics for Interactive Drama and Games. *Digital Creativity 12*, 3 (2001).
23. Mateas, M. and Stern, A. Façade: An Experiment in Building a Fully-Realized Interactive Drama. In Game Developer's Conference: Game Design Track, (2003).
24. Mateas, M. and Stern, A. Natural Language Understanding in Façade: Surface-Text Processing. In *Technologies for Interactive Digital Storytelling and Entertainment*. 2004, 3-13.
25. McGonigal, J. The Puppet Master Problem: Design for Real-World, Mission-Based Gaming. In *Second Person* (eds. Harrigan and Fruin). MIT Press, 2007.
26. Mehta, M., Dow, S., Mateas, M., and MacIntyre, B. Evaluating a conversation-centered interactive drama. *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, ACM (2007), 1-8.
27. Nelson, M.J., Mateas, M., Roberts, D.L., and Jr, C.L.I. Declarative Optimization-Based Drama Management in Interactive Fiction. *IEEE Comput. Graph. Appl. 26*, 3 (2006), 32-41.
28. Perlin, K. Creating Emotive Responsive Characters Within Virtual Worlds. *Proceedings of the Second International Conference on Virtual Worlds*, Springer-Verlag (2000), 99-106.
29. Prendinger, H. and Ishizuka, M. Let's Talk! Socially Intelligent Agents for Language Conversation Training. In *IEEE Trans. on Systems, Man and Cybernetics,* Sept. 2001.
30. Schechner, R. *Performance Studies: An Introduction*. Routledge, 2002.
31. Thue, D., Bulitko, V., and Spectch, M. A Demonstration of Player Modeling in Interactive Storytelling. (2008).
32. Tychsen, A., Hitchens, M., Aylett, R. and Louchart, S., Modeling game master-based story facilitation in multi-player Role-Playing Games. In *Proceedings of the 2009 AAAI Symposium on Intelligent Narrative Technologies II*, pp. 24-32.
33. Wardrip-Fruin, N., Mateas, M., Dow, S., and Sali, S. Agency Reconsidered. *Breaking New Ground: Innovation in Games, Play, Practice and Theory. Proceedings of DiGRA 2009*, (2009).
34. Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM 9*, 1 (1966), 36-45.
35. Young, R.M. and Riedl, M.O. Integrating plan-based behavior generation with game environments. *Proc of ACM SIGCHI Intl Conference on Advances in computer entertainment technology*, (2005), 370-370.