Comparing Speech and Text Input in Interactive Narratives

Diego Gonzalez
Williams College
Williamstown, Massachusetts, USA
drg4@williams.edu

Andrew S. Gordon
University of Southern California
Los Angeles, California, USA
gordon@ict.usc.edu

ABSTRACT

Intelligent user interfaces are finding new applications in interactive narratives, where players take on the role of a character in a fictional storyline. A recent example is the interactive audio narrative "Traveler", in which a combination of technologies for speech recognition and unsupervised text classification allow players to navigate a branching storyline via open-vocabulary spoken input. We hypothesize that the affordances of audio-based interaction in interactive narratives are different than text-based interaction, and that these differences change the player experience and their understanding of their fictional role. To test this hypothesis, we conducted a controlled experiment (n=39) to compare player interaction in "Traveler" with a text-only variant of the same storyline. We found significant differences in the types of input provided by players, suggesting that interaction modality impacts how players conceive of their relation to narrators of fictional storylines.

ACM Classification Keywords

Human-centered computing: Natural language interfaces

Author Keywords

Interactive Narratives; Natural Language Interfaces; Speech Interaction

USER INTERFACES FOR INTERACTIVE NARRATIVES

Artificial Intelligence (AI) research routinely finds application in computer games and other forms of interactive entertainment, e.g. to control autonomous virtual characters, adjust the difficulty of gameplay, and to procedurally generate content in virtual worlds. In game user interfaces, AI has helped enable new types of game controllers, e.g. motion sensing input devices [13, 5], which may qualitatively change the player's sense of agency in a virtual world. This effect on sense of agency can be seen in computer-based interactive narratives, where AI advances are beginning to enable natural language interfaces that eschew multiple-choice lists or simple verb-object commands in favor of more natural input using text [7, 14] and speech [11, 2, 15].

One example is the Data-driven Interactive Narrative Engine (DINE) [1], a web-based authoring and deployment platform for free-text interactive fiction. After reading a passage of scenario content, players of DINE scenario are presented with a free-text input box to articulate, in natural language, what they would do in the fictional situation. The system analyzes this input, and responds with an outcome that moves the storyline forward. Authors of DINE scenarios design these experiences as static branching storylines, in a style reminiscent of early Choose-Your-Own-Adventure books [12]. The task of interpreting player input is left to the underlying DINE engine, which selects the most appropriate outcome at a given branch point using an unsupervised text classification algorithm that compares statistical word embeddings [8] of player input to the first ten words written for each available outcome.

While DINE's natural language interface can greatly increase players' sense of agency compared to choice-based branching storylines, the success of this approach is highly dependent on the author's ability to anticipate the breadth of actions that players will take in a given situation, as well as the algorithm's ability to route a player's natural language input to the most appropriate outcome. Cychosz et al. [4] describes a series of user studies to identify effective designs for DINE scenarios, focusing on the effect of storyline structure and narrative style on the coherence of the interaction. Most notably, the author's choice of style in presenting character dialogue was a significant predictor of players' coherence ratings. Players were likely to match the author's dialogue style (direct or indirect speech) and tense in their own input, and mirroring direct speech had a negative effect on the classification accuracy of the algorithm. In free-text interfaces for interactive narratives, the success of the technology is dependent on the contributions of both interacting parties (the author and the player).

The interactive digital artwork "Traveler" [16] further explores the use of natural language interfaces in interactive narratives by incorporating speech recognition and by presenting story content as produced audio clips rather than as text. In this piece, players take on the role of Dr. Ramon Pineda, an American physician returning from an international conference in Germany. His homecoming takes a turn for the worse when passing through immigration at LAX airport, where he is whisked away by border agents for questioning and held alongside other travelers affected by increased border security measures. "Traveler" uses interactive narrative to expose players to the dystopian consequences of racial profiling and warrantless border searches, and challenges them to consider what their own actions might be in similar situations. "Trav-

IUI 2018, March 7–11, 2018, Tokyo, Japan.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-4945-1/18/03 ...$15.00.
http://dx.doi.org/10.1145/3172944.3172999

eler" was constructed as a four-scene interactive audio drama. In each scene, the player hears Dr. Pineda describe part of his experience, and is prompted (with a bell sound) to say what they would have Dr. Pineda do next. Player speech is processed using the large-vocabulary speech recognition engine built into recent versions of the Google Chrome web browser, results of which are then passed to an underlying algorithm that selects among four to seven outcomes for a given scene. "Traveler" uses the same unsupervised text classification algorithm as DINE [1] for selecting outcomes, comparing player input to the first ten words of textual representations of the outcome audio clips. In each scene, players are repeatedly prompted for input until the algorithm selects the specific one that transitions the story to the next scene, or to the ending audio clip in the case of the fourth scene.

We hypothesized that the use of audio interaction instead of text interaction in "Traveler" effects the content of player input, which in turn effects the ability of underlying algorithm to route player input to appropriate outcomes. Following the results of Cychosz et al. [4], we expect that players are more likely to mirror the characteristics of the presentation modality in their spoken inputs, e.g. by voicing speech that is directed to other storyline characters, rather than voicing storyline narration. We further hypothesized that this effect, combined with speech recognition errors, will degrade the classification accuracy compared to text-based interaction.

EXPERIMENT

To test these hypotheses, we conducted a human-subjects experiment to compare players' interaction with "Traveler" to a text-only version of the same interactive narrative, where players typed their input rather than speaking.

We recruited 39 English-speaking participants consisting of undergraduate and graduate students enrolled in colleges and universities across the United States, each of whom was completing an academic summer internship at *institution withheld during blind review.* Each participant was randomly assigned to either the text or speech condition. Total participation time was no more than 30 minutes, including the completion of a post-experiment questionnaire.

In the speech condition, 19 of 39 participants interacted with "Traveler" as it was originally conceived as an interactive audio narrative. Sitting at a desktop computer, participants listened to the produced audio clips of this audio drama, and spoke their inputs into the computer's microphone when prompted by a bell sound. In addition, a second microphone was used to record each interaction in the speech condition, which was later transcribed to evaluate the effect of errors in automated speech recognition on the unsupervised text classification algorithm used by "Traveler". An example audio-only interaction in this condition is as follows:

SICK WOMAN: *Help me, please! I'm very sick.* (coughing)

BORDER AGENT: *Not my problem right now.*

NARRATOR: *I couldn't believe what was happening. A sick woman was crying for help but this agent kept yelling at her. That's when I decided I had to do something.* [bell prompt]

```
> hey this woman needs help
```

NARRATOR: *I waited for someone to help the woman.*

SICK WOMAN: (coughs uncontrollably)

WOMAN TRAVELER: *Hey, aren't you a doctor?* [...]

In the text condition, 20 of 39 participants interacted with a new variation of "Traveler" that we authored specifically for this experiment. Sitting at a desktop computer, participants read textual narrative passages that ended in text input boxes, where participants typed their free-text inputs. This text-only version mirrored the content of the audio version, but was modified to read more as an interactive novel than as a script for an audio play. Starting from the script of the audio version, we converted character dialogue lines to narrated quoted speech. The lines of the "narrator" in the audio version (Dr. Pineda) were expanded slightly to convey events and information evident from sound effects and actors' performances in the audio version. The following interaction illustrates how the text version differed from the original audio version:

I could see the sick woman approach a nearby armed officer.

"Help me, please!" she pleaded him, "I'm very sick. I have a fever and severe pain, I need to leave!"

"Not my problem right now." He said, brushing her off, "So just shut up and sit down!"

I couldn't believe what was happening. A sick woman was crying for help but this agent kept yelling at her. That's when I decided I had to do something.

```
> I pleaded with the agent to help her
```

I waited for someone to help the sick woman. It wasn't long before I heard her go back into one of her coughing fits.

"Hey, aren't you a doctor?" asked the woman standing next to me. [...]

In both conditions, the underlying branching storyline structure remained the same. Overall, "Traveler" is structured around four "pages" corresponding to scenes of the narrative: meeting border security agents, an initial interrogation, waiting with other detainees, and a final interrogation. Associated with each scene are four to seven "outcomes" that are presented to the participant in response to their input, selected automatically by comparing the input (text or recognized speech) to the first ten words of each outcome, using the "MaxAvgMaxSim_{10}" algorithm described by [1]. Only outcomes not previously presented to participants are ranked, guaranteeing that participants eventually select the single outcome on each page that advances the storyline to the next page, or to the end in the case of the fourth page.

Table 1 lists the number of outcomes possible on each of the four pages ("outcomes" column), along with the number of inputs collected in both the text and speech conditions ("inputs" column). The number of inputs varied by both page and condition, with means just above two inputs per page in both conditions. A total of 179 inputs were collected from the

20 participants in the text condition, and 230 inputs from the 19 participants in the speech condition.

RESULTS

Our hypothesis was that the modality of interaction would effect the content of participants input, and that differences would subsequently effect the accuracy of the algorithm used to select outcomes. To explore these hypotheses, we conducted a series of three analyses using the 409 interactions collected in our experiment.

Classification accuracy across conditions

First, we investigated whether there was evidence that the classification algorithm performed differently across conditions. To make this comparison, we hand-annotated each of the participants' inputs, noting both whether there existed at least one appropriate outcome on the page and, if so, the most appropriate outcome. For the speech condition, these gold-standard annotations were made on transcriptions of the participants' spoken words (rather than the recognized words), as heard by the annotator in the audio recordings of the participants' interactions.

Table 1 lists the number of inputs for each page that had at least one appropriate outcome available ("supported" column). Overall, inputs in the text condition were more supported by the available outcomes on each page, with 81.6% of inputs supported compared to 68.3% for the voice condition.

Considering only the supported inputs, we used the gold-standard annotations to compute both the percent agreement ("acc." column) and chance-corrected agreement using Cohen's Kappa statistic ("κ" column). Here, chance agreement was simply estimated to be 1 over the number of outcomes on the page, affording comparison of classification performance across pages with different numbers of outcomes. Although performance varied widely across pages and conditions, mean accuracy scores favored the text condition by over ten percent.

Effect of ASR errors on accuracy

Second, we investigated the role that errors in automated speech recognition (ASR) had in degrading the accuracy of the classification algorithm in speech condition. "Traveler" uses the large-vocabulary ASR engine built into recent desktop versions of Google's Chrome browser, backed by state-of-the-art cloud-based ASR models. We computed the Word Error Rate (WER) by comparing the annotator's transcriptions of speech input with the output of ASR, finding a WER of 13.6%, such that 37.6% of speech inputs contained ASR errors. The observed error rate is similar to that seen in previous interactive dialogue systems [10].

We subsequently computed what the classification accuracy of the speech condition would have been given perfect ASR performance ("corrected speech" in Table 1). Percent agreement and chance-corrected agreement improved substantially across each of the four pages, but made up for less than half of the difference in mean accuracy scores as compared to the text condition.

Comparison of input types

Third, we investigated whether there were qualitative differences in the types of inputs provided by participants across conditions. When creating gold-standard annotations to compute classification accuracy, we observed a wide variety of relationships between the participants' input and the various voices and perspectives of characters in the story. Sometimes participants would enter words as if they were speaking directly to storyline characters, e.g. the border agent or the sick woman. Other times, participants adopted the voice of the narrator (Dr. Pineda), describing their intended actions as a continuation of his own past-tense narration. Some participants mirrored the input conventions of traditional interactive fiction [9] by issuing commands to Dr. Pineda, while others spoke directly to Dr. Pineda as if listening to his story.

To investigate whether there were systematic differences across conditions, we categorized all 409 participant inputs across both conditions into five distinguishing categories:

Narrate story: The participant continues the narration of the the narrator, in the past tense. Example: *"I opened the door"*

Speak to narrator: The participant converses directly with the narrator (or the computer), in the present tense, about the situation. Examples: *"You should to open the door," "I'd like to open the door," "I open the door?"*

Speak to character: The participant converses directly with storyline characters, e.g. in response to questions posed to the protagonist. Example: (Character says, "Do you want to open the door?") *"Yes I do."*

Command: The participant issues a command to the protagonist, as if using their input to control an avatar. Example: *"Open the door"*

Meta comment: The participant provides an input that is to be interpreted above the diegetic level of the story, as if speaking to the computer (or experimenter) about the interaction. Examples: *"How do I tell it that I want to open the door?" "I'm not sure why he would open the door."*

Figure 1 graphs the percentage of per-user interactions assigned to each of these five categories across text and speech conditions. Two significant differences were observed ($p < 0.05$, two-tailed t-test). Participants were significantly more likely to narrate the story (in the narrative voice of Dr. Peneda) in the text condition than in the speech condition. In contrast, participants were significantly more likely to speak directly to the narrator (Dr. Peneda) in the speech condition. Smaller differences seen in other input categories were not significant.

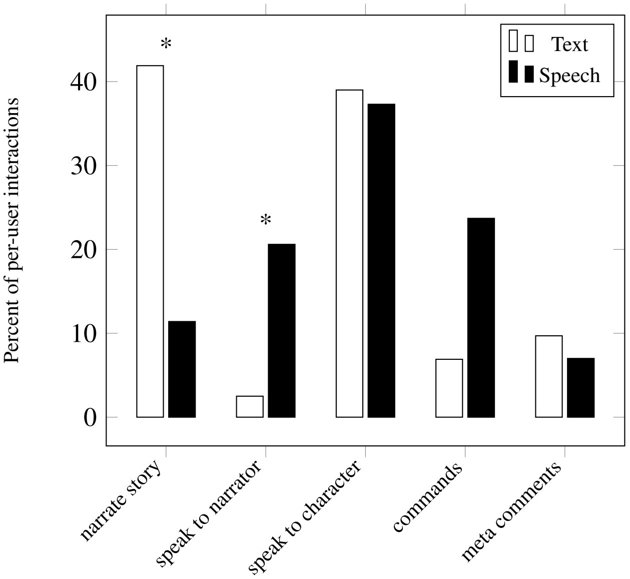
In post-experiment questionnaires, we specifically asked participants to indicate (on Likert scales) how much they felt they were interacting with a computer, the narrator, the protagonist, other characters, or the author of the story. Despite observed differences in the distributions of input types, no statistically significant differences were found in the participants questionnaire responses. In both conditions, participants most felt that they were interacting with the author of the story. Despite observed differences in classification performance, there was not a significant difference in participants ratings of the coherence

		text condition (n=20)				**speech condition** (n=19)				**corrected speech**	
page	*outcomes*	*inputs*	*supported*	*acc.*	*κ*	*inputs*	*supported*	*acc.*	*κ*	*acc.*	*κ*
1	4	48	38	.526	.368	28	21	.810	.746	.857	.810
2	7	56	45	.711	.663	69	57	.474	.386	.509	.427
3	3	40	32	.469	.203	41	37	.108	-.338	.135	-.297
4	7	35	31	.742	.699	46	42	.595	.528	.643	.583
mean	5.25	44.75	36.5	.612	.483	46	39.25	.497	.330	.536	.381

Table 1. Comparison of input classification accuracy

of the experience, or the believability of the fictional storyline events. Only one significant difference in subjective experience was evident across conditions: participants in the text condition more strongly agreed with the statement, "I found it easy to come up with responses" (p < 0.05).

Figure 1. Comparison of input type, p < 0.05 (*)



DISCUSSION

The affordances of text versus speech input have been compared in prior studies [6] and in various interactive applications, such as the tagging of smartphone photographs [3]. The contribution our study is to explore how these affordances are changed by the unique features of interactive narrative. One of the interesting features of interactive narrative, in general, is that there is a wide range of roles that the participant can play in the interaction. As seen in "Traveler," the participants may alternatively see themselves as the protagonist in the storyline, the narrator of this storyline, or as an audience member in the telling of the story. Our third analysis provides evidence that each interaction modality promotes some roles over others. In the speech condition, participants are less likely to adopt the role of the narrator, and more likely to address the narrator directly as a storyteller. In the speech condition, the "voice" of the narrator is, literally, that of another person – a professional voice actor, in this case. In contrast, the "voice" of the narrator in written fiction is that of the reader, either imagined when reading silently or vocalized when reading out loud. When a participant narrates the story using text, they are continuing a monologue already begun through the process of reading. When a participant narrates the story using speech, they are interrupting someone else's telling of a story.

The affordances of each interaction modality shape the distribution of input types provided by participants, with consequences for classification accuracy. In the text condition, roughly 80% of the inputs are either narration or direct speech to storyline characters, and roughly 80% of inputs were supported by one of the available outcomes on each page. Given that nearly all of the authored outcomes in "Traveler" consist of narration or the direct speech of storyline characters, these categories of inputs are more likely to invoke coherent outcomes at correspondent levels of discourse. Furthermore, this parity in interaction supports the underlying classification algorithm, which relies on the similarity of word-level embeddings between inputs and outcomes. The relative diversity of input types in the voice condition make it more difficult for authors to anticipate the breadth of inputs, resulting in fewer inputs that can be supported by authored outcomes. As participants move further afield from narration and direct speech, the similarity-based classification algorithm suffers in performance.

While these results favor text interaction given the technology, we believe that these results also chart a path forward for interactive audio narratives. To facilitate both the ability of authors to cover the input space and the algorithm's ability to select appropriate outcomes, future interactive audio narratives should strive to reduce the diversity of participant input categories, either through instruction or design. While tutorial "how-to-play" instruction might suffice, we hypothesize that removing the voice of the narrator from the audio production would inhibit participants from speaking above the diegetic level of discourse. By moving the genre closer to that of an interactive radio drama than an audio book, some scenario design innovations will be necessary to allow for player agency beyond conversational speech acts, e.g. by designing scenarios where the topic of conversation is which actions to take in the given situations.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1560426. The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

1. Jenna Bellassai, Andrew S. Gordon, Melissa Roemmele, Margaret Cychosz, Obiageli Odimegwu, and Olivia Connolly. 2017. Unsupervised Text Classification for Natural Language Interactive Narratives. In *Proceedings of the 10th International Workshop on Intelligent Narrative Technologies, Snowbird, Utah, October 5-6, 2017*.

2. Marc Cavazza, Jean-Luc Lugrin, David Pizzi, and Fred Charles. 2007. Madame Bovary on the Holodeck: Immersive Interactive Storytelling. In *Proceedings of the 15th ACM International Conference on Multimedia*. 651–660.

3. Mauro Cherubini, Xavier Anguera, Nuria Oliver, and Rodrigo de Oliveira. 2009. Text versus Speech: A Comparison of Tagging Input Modalities for Camera Phones. In *Proceedings of the 11th Conference on Human-Computer Interaction with Mobile Devices and Services, Mobile HCI 2009, Bonn, Germany, September 15-18, 2009*.

4. Margaret Cychosz, Andrew S. Gordon, Obiageli Odimegwu, Olivia Connolly, Jenna Bellassai, and Melissa Roemmele. 2017. Effective Scenario Designs for Free-text Interactive Fiction. In *Proceedings of the 10th International Conference on Interactive Digital Storytelling, Madeira, Portugal, November 14-17, 2017*.

5. Felix Kistler, Elisabeth André, Samuel Mascarenhas, André Silva, Ana Paiva, Nick Degens, Gert Jan Hofstede, Eva Krumhuber, Arvid Kappas, and Ruth Aylett. 2013. *Traveller: An Interactive Cultural Training System Controlled by User-Defined Body Gestures*. 697–704.

6. Gale L. Martin. 1989. The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies* 30, 4 (1989), 355 – 375.

7. Michael Mateas and Andrew Stern. 2003. Integrating Plot, Character and Natural Language Processing in the Interactive Drama Facade. In *Proceedings of Technologies for Interactive Digital Storytelling and Entertainment (TIDSE), Darmstadt, Germany.*

8. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*. Curran Associates Inc., USA, 3111–3119.

9. Nick Montfort. 2003. *Twisty Little Passages: An Approach to Interactive Fiction*. Cambridge, MA: MIT Press.

10. Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Dogan Can, Panayiotis G Georgiou, Shrikanth S Narayanan, Anton Leuski, and David R Traum. 2013. Which ASR should I choose for my dialogue system?. In *Proceedings of the 14th annual SIGdial Meeting on Discourse and Dialogue, Aug 22-24, 2013, Metz, France*. 394–403.

11. Jeff Orkin and Deb K. Roy. 2014. Understanding speech in interactive narratives with crowd sourced data. In *Proceedings of the 8th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2012*.

12. Edward Packard. 1979. *The Cave of Time*. New York: Bantum Books.

13. Daniel Shapiro, Josh McCoy, April Grow, Ben Samuel, Andrew Stern, Reid Swanson, Mike Treanor, and Michael Mateas. 2013. Creating Playable Social Experiences through Whole-Body Interaction with Virtual Characters. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

14. Reid Swanson and Andrew S. Gordon. 2012. Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 16 (Sept. 2012), 35 pages.

15. David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. 2015. New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling. In *Proceedings of the 8th International Conference on Interactive Digital Storytelling (ICIDS), Copenhagen, Denmark.*

16. Mike Treanor, Nicholas Warren, Mason Reed, Adam Smith, Pablo Ortiz, Laurel Carney, Loren Sherman, Elizabeth Carré, Nadya Vivatvisha, D. Harrell, Paola Mardo, Andrew Gordon, Joris Dormans, Barrie Robison, Spencer Gomez, Samantha Heck, Landon Wright, and Terence Soule. 2017. Playable Experiences at AIIDE 2017. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'17), Snowbird, Utah, October 5-9, 2017.*