

The Responsive Room

Ian Eslick

M.I.T. Media Lab
E15-321
eslick@media.mit.edu

Hope Ginsburg

M.I.T. Visual Arts Program
N51-101
ginsburg@mit.edu

Rachel Kern

M.I.T. Media Lab
E15-384C
rachelk@media.mit.edu

Kevin Wang

Harvard Graduate
School of Education
kevinkw@gmail.com

ABSTRACT

We introduce the Responsive Room, an interactive, ambient environment designed to augment spoken language with light and auditory cues by analyzing patterns in spoken speech. This paper focuses primarily on the use of the Responsive Room for storytelling. To this end we constructed a scripted, theatrical sequence that explores the relationship between story telling, color and sound. We also analyzed the possible automated generation of these sequences from analysis of the spoken speech. An offline analysis extracted important categories from the text such as affect, locale and topic. The system used to analyze speech for major story features is described, as is the theory for mapping these features to colors. The preliminary results of an exploratory experiment with children are discussed.

Author Keywords

Interactive media, affective computing, ambient spaces.

ACM Classification Keywords

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

INTRODUCTION

The Responsive Room was originally conceived to explore the response of children to ambient cueing intended to highlight aspects of a spoken story. Such a room would operate by sensing vocal affect and employing speech recognition to extract keywords, phrases and plot units from the spoken story, and mapping them to various categories such as affect, location, topic, etc. These categories would be passed to a mapping module that evokes audio and visual cues to underscore or enhance the child's experience of the story.

One possible motivation for such an environment is to enhance the experience of reading for children who have

grown up in a world of highly interactive video games and other engaging multimedia.

In part, this specific use of a speech-driven interactive environment was selected in order to account for limitations in current speech recognition technology. In the reading of a story, there is a single speaker (barring interruptions); the words being spoken will have local coherence in topic and affect. Moreover, there is the tendency to speak with exaggerated affect. The use of a Commonsense-based category detector [2] makes it possible to compensate for the noise of speech recognition by averaging the contribution of a number of keywords and phrases over a recent history of utterances.

BACKGROUND

The local inspirations and guides for this work include Hiroshi Ishii's ambientROOM [4] and Bobick et al's "KidsRoom" [1]. The ambientROOM used media such as light, sound, shadow, airflow and water flow as subtle, peripheral communication mechanisms within a space. These ambient media were intended to display information that was not related to the user's primary task, and was only processed in the background of a user's attention. The Responsive Room sits closer to the foreground, accentuating or highlighting aspects of the foreground experience without arresting attention completely.

The KidsRoom was an augmented environment that was far more interactive than the ambientROOM; the room itself was a recreation of a child's bedroom, complete with furniture and appropriate décor. It also included video projection screens, colored lights, speakers, video cameras, and a microphone, all of which was computer-controlled. The KidsRoom was designed to guide children on an interactive, imaginative journey, during which the bedroom morphed into several other worlds, including a forest world, a river world, and a monster world. The real world objects in the room, such as the furniture, changed their functions as one world evolved into the next; for example, the bed became a boat in the river world. The system directed the children through the adventure, and used the video cameras, along with computer vision, to track the kids as they played and interacted in the room. The major differences between our system and the system in the KidsRoom is primarily that our system is designed to accentuate imaginary scenarios instead of provide a basis for them, and secondly that it uses recognized speech as its major information

channel, whereas the KidsRoom did not use speech recognition.

THEORY

Here we describe the three primary areas of interest in our system: the mappings between story elements and color, the mappings between story elements and sound, and the process of establishing story elements from both written and speech-detected text. Our model of affect was based on the commonly referenced set of primary emotions (anger, fear, happiness, sadness, disgust and surprise) developed by Ekman [3].

Color

Association of color with affect is always a controversial topic. The perceptual processing of color in a given context is highly dependent on both visual aspects of the surrounding environment (color, illumination, etc.), as well as the cognitive context in which the color is perceived.

The color in the Responsive Room was determined by three basic principles of color theory: value, temperature and hue. Value, the relative lightness or darkness of the color, and temperature, the relative warmth or coolness of the color, are mapped to the time of day and the type of light in each setting. For example, a scene that takes place in the bright moonlight maps to a light, cool color. The temperature of the color is also influenced by whether the scene takes place inside (warm) or outside (cool). The hue of the color is mapped to the emotional tenor of the scene. Each of Ekman's six primary emotions is mapped to a different hue based on historical and contemporary color theory. For example, sadness is mapped to the color blue.

We also explored two different types of color mappings: iconic and affective. The affective mapping was purely based on emotion with no regard to setting. The iconic mapping took affect into account, as well as considering time of day, location, etc., in determining the lightness or darkness of the color.

Sound

Films, and more recently video games, have provided a rich legacy of knowledge about the relationship between the semantics of scenes (plot elements, actors, relationship, themes) and styles of music. Rather than appealing to a core theory linking sound and emotion, we drew primarily from pre-recorded music samples that were composed to establish specific affects.

In the initial room, we took a set of sound files from the soundtrack to the video game "Ultima 9". This soundtrack was composed to establish specific moods within the game for various unfolding events such as conflict, peaceful repose, scary underworld environments, cheerful spring days, etc., providing a decent lexicon of tracks from which to choose. Moreover, many of the tracks were designed to fade smoothly from one into another as events in the game world changed.

The sound was mapped using two features: location, specifically inside versus outside, and primary emotion. The inside/outside determination was largely arbitrary, but it allowed us to key to specific elements of the Hansel and Gretel story. The sound-to-emotion map was only partial, as the story primarily featured the emotions of sadness, fear, anger and occasional periods of happiness. Thus, we did not need to construct a mapping for the emotions of disgust and surprise.

Category Spotting

To mediate between the noisy results of a speech recognition engine and the precise categories required as an input to the models for color and sound, we used the ConceptNet [8] database and toolkit developed by the Commonsense Computing group at the MIT Media Laboratory. The toolkit uses a large database of commonsense relationships between English phrases, such as "throw person" *causes* "fearfulness". These commonsense relationships can be used to assess what concepts are closely related to a given query word or words. Thus, if speech detection extracts only one key concept such as "witch" or "throw person" from the sentence "The evil witch threw the frightened boy into the toolshed," the affect of all the words connected to the keywords will allow the system to clearly identify "fear" as the dominant affect. This same principle can apply to any category that can be captured by a vector of English words or phrases. A significant fraction of the subject material's concepts must, of course, be represented within the ConceptNet database.

SYSTEM CONSTRUCTION

The prototype system was assembled primarily from off-the-shelf packages for each of the aforementioned stages of processing.

1. **Speech** The speech processing was handled by the Java-based Sphinx-4 -- a product of Carnegie Mellon University [6]. We generated two custom vocabulary files for the system; one for the fairy tale story we told to kids, so we could compare the scripted version to the speech-based version, and the other to perform some experiments on broadening how people might interact with the Responsive Room.

2. **Visual Effects** Room color was produced using a simple window on a computer screen to fill the output of an LCD projector and thus illuminate a wall of the room. Because the room's walls and ceiling were constructed from translucent tracing paper, the color was also projected through the primary wall, spilling over onto adjacent walls, as well as onto the ceiling.

3. **Category Detection** Category detection was based on a generalized reimplement of the ConceptNet toolkit. This rewrite made it easier to control and intercede in various phases of the category detection.

4. **Audio Environment** The audio environment was constructed using multiple instances of the xmms linux-based audio player and the xmms-shell program which provides a text-based command line API to the main player. These were also controlled using a socket protocol to control the xmms and shell processes. Smooth cross fading of two audio tracks was implemented by launching a separate process to manually alter volume in two instances over a fixed time interval.

These facilities were split between a Java system (speech & visual effects) and Lisp (affect detection & audio control). The number of different subsystems led to a somewhat unstable system that was partially compensated for by the use of Common Lisp as the controlling language. Several problems encountered during the experiments were quickly fixed at the Lisp command line.

EXPLORATORY EXPERIMENT

To evaluate the experience of children in this room, we set up a simple exploratory experiment.¹ The experiment consisted of two children listening to a reading of Grimm's "Hansel and Gretel" fairy tale while sitting in the Responsive Room. The children were four and seven years of age. A set of light and sound cues was designed and manually assigned to specific points in the story in order to simulate the response of a live speech system. Separately, we ensured that the story features for these cues could be generated by the category spotting system. The quality of processing the text was at about a 75% accuracy level to the hand-designed features. During the reading, we switched the underlying color theory used to create the light cues, using iconic cues in the first half of the story, and affective cues in the second half of the story. We stopped and asked the same set of questions of the children after each half of the story to determine what, if any, difference it made to them.

We used the same audio map throughout the experiment, taking a combination of location (inside versus outside) and affect, and mapping each combination to a specific track on the Ultima 9 video game soundtrack and a Katrina Finch CD. The location-emotion pairs and their associated audio track names are as follows:

- [*inside fear*] --> "paws"
- [*outside fear*] --> "undead"
- [*outside sad*] --> "moonglow"
- [*outside happy*] --> "mountain-dance"
- [*inside happy*] --> "Britain"

¹ By MIT policy, the results of these experiments cannot be published or used for anything other than classroom use.

- [*outside anger*] --> "good-v-evil"

We used time-of-day, type of lighting, and emotion as bases for color mapping in the Hansel and Gretel story. The following scene in Hansel and Gretel would produce a light warm yellow: Hansel and Gretel return to their kind father's house after besting the wicked witch and escaping from the forest. The rationale is as follows: the time of day is afternoon, which produces a light, warm color. The warmth of the color is further enhanced by the fact that the scene takes place indoors. Finally, in this scene, Hansel and Gretel are happy, which yields a yellow hue.

Experimental Results

The kids who took part in the experiment expressed that they enjoyed the experience of listening to the story in the Responsive Room. Our interview with them afterwards elicited such comments as, "It made me feel like it's really happening," in reference to the colors, and "It's pretty cool", in reference to the music.

The kids took note of the changing colored light first – it was the first thing mentioned when they were asked what they noticed about the room. The purpose for the lights seemed apparent to the seven-year old; he noted, "Every time a part of the story changed, the lights turned a different color". He also commented that "because of the colors, I'll know if bad things or good things are happening". The four-year old was less attuned to the motivation behind the changing colors of the lights, and thought that the storyteller turning each page of the story controlled the color changes. The four-year old did take note of the music, however, saying that the music was happier in the part of the story when Hansel and Gretel killed the witch. The children also commented that the effects were creepy and a little bit scary.

Interestingly, the original hypothesis with which we approached the iconic vs. affective color maps appeared to be ill-founded. Neither child noticed any difference between the different maps. This implies to us that children are paying attention more to changes in the environment than to the current state of the environment. This could be explained in many ways. A single color in the room provides little contrast and so the listener rapidly habituates to the presence of a single color. The sound cues may capture more attention than the color cues. The timing of stimulus changes may have been a more salient feature in that these changes matched up closely with plot changes in the story.

FUTURE DIRECTIONS

This experiment raised a host of questions about augmenting imagination-based experience. We need to better understand the relationship between light, sound, timing, aspects of the story and the user experience. A much larger sample population and sequence of

experiments would be needed to establish the role of different modalities.

For example, we chose a specific model of affect-to-color mapping and a more theatrical notion of cueing color for our experiments. An important variation to try in future experiments is the impact of well-executed theatrical style cueing which requires foreknowledge of interesting events before they occur as compared to a moving average view of one or more categories of the story (presence of a character, current location, day and night, etc.) that slowly evoke color and sound changes. The later style is more practical for modern automated sensing and analysis methods.

The theory of color can be employed in many different kinds of mappings. For example, we might create an N-dimensional space where each axis of the space is a topical category mapped to a specific set of color values. Based on the location established by the current category estimates, a color can be estimated as a linear interpolation of the axis values. Color contrast can be employed by introducing multiple colors and multiple N-dimensional spaces, but would need to be cross-constrained to account for the interaction of multiple colors. In fact, the perceptual nature of color could provide an interesting landscape for exploring the grammar created from a set of semantic features and the constraints imposed by perceptual constraints on combinations of colors.

Other future work could involve adding more features to the Responsive Room. During the planning stages, we discussed several other features, but did not have the time or the resources to implement them. One possibility was adding scents to the room as an olfactory reflection of location. Another suggestion was to include a ceiling that could be raised and lowered to enlarge or constrict the physical space of the room. One challenge of adding these additional features is that it would greatly complicate the ability to control an experiment for only one of the attributes.

In analyzing speech, category spotting using ConceptNet works surprisingly well. Our accuracy of almost 75% is reasonable for establishing non-theatrical cues because the impact of a mistake is less salient to the listener. Analysis of the failures demonstrates room for improvement along several dimensions:

- There are concept gaps in the database that can be filled in.
- A-priori identification of major protagonist and antagonist figures in a story can significantly improve interpretation of the affect of specific events.
- Recognition of negatives is not currently supported by the system. The phrase “did not die”, unless represented as a unitary concept, will provide the category implications of the concept “die” instead of “survive”.

A final observation from the design of the Responsive Room is that there is a truly significant gap between modern speech recognition systems and the requirement for accuracy in following spoken text content. One near term approach to bridge this gap might be to feed the text of a story into the machine so that it can try to match the speech data to an a-priori model of the story. Given a guess as to where in the story the speaker is, the system can analyze the high quality text to plan theatrical or ambient cues based on the natural language features of the story.

CONCLUSION

We conclude that there are some effects experienced in the room that are of definite value towards our original goal of accenting experience and which deserve further exploration. What is less clear from these early explorations is how the technology issues can be sufficiently dispensed with in such a way that they fit into a natural pattern of interaction.

Some design constraints that we did derive from this process include:

- Color is less important than transitions between colors.
- Transitions move the ambient color to the foreground, and are thus more salient to participants. Slower transitions of color may evoke a very different, possibly background, feeling.
- Music is a more powerful determiner of mood than color in the theater setting, but more disruptive in a discussion setting. We might consider experimenting with other forms of music that are more amenable to spoken conversation or perhaps generated music or mixed samples that can shift moods more smoothly
- We only used projections of solid colors to express the dominant affect in each scene. Because several different emotions were present in each scene, it may be interesting to explore using a mixture of colors in our projections to represent the mixture of emotions. We could do this by projecting an image of several different colors mixed together, or some other pattern that employs several colors at once.

ACKNOWLEDGMENTS

We thank the professor, staff and students of MAS.834, the members of the Commonsense Computing Group at M.I.T., and the Carnegie Mellon Sphinx Group for making their tools available online and easy to use.

REFERENCES

1. Bobick, A. F., Intille, S. S., Davis, J. W., Baird, F., Pinhanez, C. S., Campbell, L. W., Ivanov, Y. A., Schütte, A. and Wilson, A. The KidsRoom: A Perceptually Interactive and Immersive Story Environment. *Presence* 8(4): 369–393, August 1999.

2. Eslick, I. and Vercoe, S. Commonsense-based Category Detector, unpublished work.
3. Ekman, P. Facial Expression of Emotion. *American Psychologist*, 48, 384-392. 1993.
4. Ishii, H., and Ullmer, B. Tangible Bits: Towards Seamless Interfaces Between People, Bits and Atoms. *ACM CHI '97*. 234-241, ACM Press, 1997.
5. Itten, Johannes. *The Art of Color: The Subjective Experience and Objective Rationale of Color*. Rheinhold Publishing Corporation. New York: 1961.
6. Lee, K. F., Hon, H. W. and Reddy, R. An Overview of the SPHINX Speech Recognition System. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-38(1), 35-44, 1990.
7. Liu, H., Lieberman, H., and Selker, T. A Model of Textual Affect Sensing Using Real-World Knowledge. *Proceedings of the 2003 International Conference on Intelligent User Interfaces, IUI 2003*. 125-132, ACM, 2003.
8. Liu H. and Singh P. ConceptNet – a practical common sense reasoning tool-kit, *BT Technology Journal* 10 (2004)
9. BasicTips.com: Articles and Tips for Webmasters. Understanding Color Emotion Triggers, Part 1A. http://www.basictips.com/tips/article_78.shtml