Beyond "Next slide, please": The use of content and speech in multi-modal control

David Franklin, Shannon Bradshaw and **Kristian Hammond**

Intelligent Information Laboratory
Northwestern University
{franklin, bradshaw, hammond}@ils.nwu.edu

Abstract

The Intelligent Classroom is an automated lecture facility where one of the primary goals is that speakers be able to control it by interacting with it as they would with a human A/V technician. In this paper we describe our research in imbedding Microsoft Powerpoint into the Intelligent Classroom. In particular we discuss how we use two modes of sensing (Computer Vision and Speech Recognition) to provide hands-free control of slide presentations. We look at utterances and gestures that serve as commands. And, more interestingly, we discuss some probabilistic techniques that can be used to match slide content to the speaker's words, providing new and useful ways of performing a presentation.

1 Introduction

In the Intelligent Classroom, we are attempting to enable new modes of user interaction through the use of multiple sensing modes and plan recognition – the Classroom uses the inputs from its cameras and microphones to determine what the speaker is trying to do, and then takes whatever actions it deems appropriate. (It controls automated VCRs and slide projectors and also produces a video of the presentation.) One of our goals is for the speaker to be able to interact with the Classroom as he would with an A/V technician: sometimes in commands (through speech and/or gesture,) and often by just going about his presentation and trusting the Classroom to do the right thing.

This paper discusses how we have incorporated Microsoft's Powerpoint presentation software into the Intelligent Classroom. We have implemented a few simple command modes and have also utilized some probabilistic techniques from Information Retrieval (IR) to match the speaker's words to the content of slides as an indicator of when it is appropriate to change to another slide. The IR techniques allow the speaker to move through a presentation without having to pause to advance slides; the system listens to his words and determines when it is time to change slides. More interestingly, these techniques allow the speaker to skip around in his presentation by describing the slide that he wishes to skip to – this would prove invaluable for answering audience questions. In the extreme, this could even support a speaker who wished to do an unstructured lecture from a large body of slides.

In this paper, we first look at several different command-based control techniques and how they are implemented for the Classroom. Secondly, we look at a simple probabilistic approach to determining what slide the speaker is lecturing from based on what words he says. Then we look at a number of simple extensions to this that render this useful in the context of giving a presentation. Then we present an example of the system at work. Finally, we summarize this work and look at some future directions for the research.

2 Command-based approaches

One mode of controlling Powerpoint in the Intelligent Classroom is to treat particular speaker actions as commands to advance to the next slide (or to go to the previous slide.) These can be purely speech-based: the speaker says "Next slide" or some synonym. They can also be purely vision-based: the speaker touches a next slide "button" on the projection screen or makes a particular gesture that is understood to mean "next slide." Later in this section we will discuss some weaker clues that rely on other contextual information to be useful.

2.1 How does the speaker say "Next slide"?

One potential difficulty with spoken commands is that the particular sequences of words are not always intended as a command. For instance, if the speaker says "... and then the Classroom will go to the next slide when I ask it to," does he intend for the Classroom to actually go to the next slide? A human A/V technician would not find this to be ambiguous at all

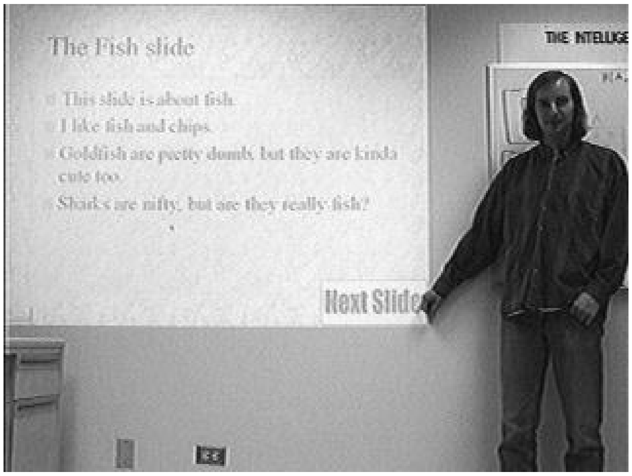

Figure 1: The speaker touches the next slide "button".

– if the speaker intended his utterance as a command, he would make it clear by emphasizing the command words in some way. (He would probably either voice the words in a commanding tone or pause before saying them.) Unfortunately, current speech recognition software is not capable of detecting such word voicing subtleties. But, as our software is able to detect pauses in the speech, so in this way we can distinguish "next slide" as a command from "next slide" as just speech.

In the Intelligent Classroom, we have already done work on visually recognizing a number of speaker actions (Flachsbart 1997). These include the actions of touching a particular location on the projection screen, and pointing in a particular direction. We have designed Powerpoint slide templates with next slide "buttons." So, when the Classroom observes the speaker touching this portion of the slide, it knows to advance to the next slide in the presentation. (Figure 1 shows a speaker touching the next slide "button.") Similarly, the gesture of pointing down and to the right is understood as a command to advance to the next slide. (If you imagine the slides as being a sequence running left to right, the next slide is to the right. We chose downward pointing because it seems like a safe gesture in that it is unlikely to be confused with any other speaker actions.)

There are some other actions that can serve as weaker indicators of the speaker's intentions. If observed in isolation, they do not provide sufficient evidence to act on, but, if observed in combination with other contextual clues, they can provide sufficient evidence to prompt action. Some examples of these kinds of actions are: a "magic wand" hand (various dramatic gestures toward the screen,) a prolonged stare at the display, a lengthy pause, a question from an audience member, and the speaker moving to a particular location on the floor.

2.2 Implementation in the Classroom

Without looking too deeply at the Intelligent Classroom itself (this is discussed in greater detail in (Franklin 1998) and (Franklin & Flachsbart 1998)) we will briefly discuss how these behaviors are implemented. These details are important in considering how Powerpoint becomes part of the much larger Intelligent Classroom, which must deal with a much greater number of speaker actions and numerous A/V components. A system just designed to control Powerpoint could make due with much less.

The Classroom has a library of plans for things the speaker might do during his presentation. These plans include both the speaker's actions (the speaker's *process*) and how the Classroom should act in response (the Classroom's processes). For example, there is "going and writing on the chalkboard plan" which contains the speaker's actions of moving to the chalkboard, stopping at it, beginning to write, and finishing. It also contains Classroom processes for adjusting the lights and controlling the camera that films the presentation. If all the processes for a plan stay properly synchronized and run to completion, the plan is said to have succeeded.

These plans are intended to represent a common understanding of how a speaker and A/V technician would interact. From the perspective of the speaker, a plan tells him (in the speaker's process) what he needs to do to accomplish the goal, with the understanding that the Classroom will take care of its own responsibilities in the plan (the Classroom's processes). From the perspective of the Classroom, a plan tells it what it needs to do if it observes the speaker doing his part. Essentially, the Classroom can view all the slide-advancing plans as saying: "if the speaker wanted me to go on to the next slide, he would ...."

When the Classroom observes that the speaker has taken some action, it looks to see if that action is a part of a plan that the speaker was already known to taking part in. If not, the Classroom will look in its plan library for plans that could explain why the speaker did what he did. Through the observation of a number of actions, and the use of some simple plan recognition techniques, the Classroom is able to determine what the speaker is doing, and what it needs to do about it.

3 A probabilistic approach

In this section we look at the foundations for a probabilistic approach to determining what slide a speaker is lecturing from, based on the words he speaks. In the next section, we will look at what extensions are necessary to make it useful and show how it can be integrated with the command-based approaches. Bayes' Law provides the probabilistic foundation for the approach and the different behaviors of the system depend on how we define the various components of the equation.

$$P(s|w) = \frac{P(s)P(w|s)}{P(w)} \tag{1}$$

The s's refer to the indices of slides and the w's refer to words. In addition:

- $P(s|w)$ is the probability that we should now be viewing slide s, given that we just heard word w.
- $P(s)$ is the probability (prior to hearing w) that we should be viewing slide s now.
- $P(w|s)$ is the probability that we would hear word w, if we were in slide s now.
- $P(w)$ is the probability that we would hear word w now.

If we assume that the words in the slides are representative of the words that will be used to describe them, we can compute $P(w|s)$ by dividing the number of times word w appears in slide s by the total number of words in slide s. $P(w)$ is the weighted sum of all the $P(w|s)$ (weighted by $P(s)$):

$$P(w) = \sum_{s \in S} P(s)P(w|s) \tag{2}$$

Unless the speaker is just reading the slides, our assumption does not hold. But, through the use of the techniques discussed later, we can compute values for $P(w|s)$ and $P(w)$ that are sufficiently accurate for our needs.

Equation 1 computes a probability distribution across all the slides. The system can be confident when the distribution contains a single spike that approaches probability 1.0. The system will be less confident when there is no clearly dominant slide in the distribution. Essentially, the system changes which slide is being shown when it is confident that another particular slide should be shown.

3.1 Controlling system behavior through $P(s)$

How we define $P(s)$ determines to a large extent the slide-switching behavior of the system. If $P(s)$ is simply a function on s then the system will treat certain words as indicators that it should switch to particular slides; whenever that word is spoken, the system will skip to that slide. While this will sometimes yield the correct response, it does not take the speakers current place in the slide presentation into account. This can result in rapid vacilations between slides and other annoying behaviors when an unfortunate speaker happens to speak some of the "indicator" words without intending to go to the words' slides.

As a result, we define $P(s)$ using various information about the state of the presentation, such as: the current slide probability distribution, the current slide, time spent on the current slide, and time spent in the presentation so far. For example, to make the system step forward through the slides sequentially, expecting to advance every k spoken words, we could use the following definition (where $P'(s)$ is the previous probability distribution.):

$$P(s) = \frac{1}{k}P'(s-1) + (1 - \frac{1}{k})P'(s) \tag{3}$$

4 Extending the probabilistic approach

In this section we look at what extensions we have made to the basic probabilistic approach in order to make it work with real data. First we look at some techniques from Information Retrieval that we have adopted. Then we discuss an adjustment to the definition of $P(w|s)$. Finally, we look at some processing we can do on the slides to extract key phrases that the speaker might use.

4.1 Lessons from Information Retrieval

Early in the development of the system, we ran into a couple of the problems that face the designers of all Information Retrieval systems. First, the system has no *a priori* way of knowing that "transmission" is a better indicator word than "the." As a result, a slide that makes heavy use of the word "the" is likely to be switched to whenever the speaker makes heavy use of "the." Second, the system has no way of knowing that "weight of terms" and "term weights" have a great deal in common. As a result of this, a lot of good indicators can be missed.

For these problems, we have used approaches that have served the IR researchers reasonably well. We use a stop-list (consisting of prepositions, articles and conjunctions,) allowing the system to ignore words that do not provide any useful information about what the speaker is talking about. Also, we perform some morphology to find the roots of words. These two techniques reduce the number of words that the system needs to consider in deciding whether to switch slides. In the short term, this improves the system's performance. Hopefully, in the long term the smaller word set will allow us to actually use some of the more complicated techniques suggested in our Future Work.

4.2 Redefining $P(w|s)$

The assumption underlying our original definition of $P(w|s)$ was that the words in the slide are representative of the words that will be used to describe them. If taken literally (which our definition did,) this means that, in discussing a slide, the speaker must only use the words on the slide! (If he uses other words, $P(s|w)$ will become 0.0 for the current slide.) We would rather assert just that he is more likely to use the words on the slide.

So we changed our definition of $P(w|s)$ to reflect a weaker assumption that a certain proportion of the words used to discuss a slide will be found in the slide. The other words will either be unknown (not appearing anywhere in the slide collection) or be from other slides. Since the unknown words do not give any clues as to which slide we should be viewing, they, like the stop-listed words, are ignored. For the other words, we just want to say that they are less likely to occur than they are in the body as a whole. The new definition (now called $P^*(w|s)$) is a weighted average of $P(w|s)$ and the result of dividing number of times w appears in all the slides by the total number of words in the slides. (We use weights of 0.7 and 0.3 respectively, but the precise value does not affect the system's behavior profoundly.)

4.3 Reasoning about slide content

While the previously mentioned IR techniques are used to reduce the words that the system must worry about, we have used a couple "smarter" techniques to find key words and phrases that will serve as excellent indicators. The simplest of these is to look at the different objects that make up the slide (things like graphs, tables and pictures) and use possible descriptions of the objects as indicators. For example, if the speaker says "...and then, in this table ...," he is certainly referring to a slide with a table in it. For each of the different objects that can be embedded in a Powerpoint slide, we have enumerated a number of key phrases that could refer to it.

We have also looked at some grammatical approaches for finding phrases that might be used when discussing a particular slide. In these approaches, we locate complex verb and noun phrases in the slide content. Then we use a number of simple operations to build a list of semantically similar phrases. For example, in discussing a slide containing "document comparison," the speaker might instead say "comparison of documents" or "comparing documents."

When designing slides, a speaker tends to be as concise as possible – putting as much content as possible into just a few words. As a result, slide titles and many slide points use long noun phrases like "current information management technology." But, when speaking, a speaker tends to use more natural language, avoiding the stilted language of the slide. We have written a number of rules that perform the kinds of transformations that a speaker is likely to make.

To find these new phrases, the system first locates the verb and noun phrases used to derive them. The system tags the words from the text of a slide title or slide point with their parts of speech. Then, it looks for verb phrases (defined as one or more verbs preceded or followed by zero or more adverbs) and noun phrases (defined as zero or more adjectives followed by one or more nouns).

We treat each verb phrase as a key phrase, and, when possible, make another key phrase with the adverbs moved to the opposite end. Since we do morphology when matching words and phrases, we do not have to worry about verb conjugations (i.e. past versus present tense, 1st versus 3rd person.)

We break each noun phrase into smaller units: the adjectives, the modifier nouns and the subject noun. From the noun phrase "current information management technology," we get "current," "information management" and "technology" for the three units. Given these, we construct key phrases using the following rules:

- the modifier nouns serve as a key phrase (as in "information management.")
- the modifier nouns can be moved outside the adjective and subject noun (as in "current technology for information management.")
- different subsets of the modifier nouns can be used (as in "information technology.")

Each complex noun phrase generates a number of key phrases which are then treated as "words" in the slide. Whenever one of the key phrases for a particular slide is spoken, it is very likely that the speaker wishes to be lecturing from that slide. These phrases provide much stronger evidence than the individual words do. In fact, as we build up more rules, we hope to be able to rely solely on the key phrases – perhaps using the other approaches only as a last resort. Also, we are looking at how to build phrases using the verb forms of nouns within a noun phrase (as in "current technology used to *manage* information.")

5 Building a useful system

In the previous sections we have discussed a number of methods for determining when to change slides (and what slide to change to.) But, a useful system really needs to utilize all of them simultaneously. So, in this section we discuss how we have combined these methods, using the probabilistic approach we discussed earlier and discuss an interesting new lecture technique that the system facilitates. The methods we discussed earlier fall into two classes: methods that determine what words (or phrases) the system pays attention to, and methods that determine how likely it is that the speaker wishes to change slides.

5.1 Reasoning about words

Before starting a slide presentation, the system analyzes the contents of the slides to determine which words and phrases to use in reasoning about when to change slides. The result of this preprocessing is a list of words and phrases (and all the various frequency information that is needed to compute the various word probabilities.)

First we enumerate all the words used in the slides, using stop-listing to filter out useless words and using the morphological techniques to convert the words to their root forms. Then we look for the key phrases (those found using the language clues and those inferred based on embedded objects) in the slides and add them to our list of words. Finally, given this word list, we go through the slides one more time and compute the frequencies of each word and phrase (how many times it appears in each slide and how many times it appears in the body of slides.) The frequencies include the embedded objects' phrases as if the phrases were a part of the text of the slide. These frequencies are then used to compute the values for $P(w|s)$ and $P^*(w|s)$.

Every time the speaker says a word, the system performs the morphology to find its root, sees if it is one of the enumerated words (or completes one of the key phrases,) and, if so, recomputes the probability distribution using Equation 1. Also, to reflect that it is most likely closer to the time to switch to the next slide, it adjusts the probability distribution using Equation 3.

5.2 Dealing with "next slide" commands

During a slide presentation, the various commands (or command-like actions) are used to alter $P(s)$ (the probability that we should be viewing slide s at the moment.) For commands that are absolutely certain, this will result in the slide immediately being changed. For more ambiguous command actions, the change in the probability distribution might not be enough to cause a slide change, though it will cause it to be more likely to change in the immediate future.

When the speaker performs a "next slide" command, the system sets $P(s'+1)$ to be 1.0 (where s' is the index of the current slide) and sets all the other $P(s)$ to be 0.0. (This is analogous to saying: "I am absolutely certain that we should be viewing slide $s'+1$.") The system alters $P(s)$ in the same way for the command actions of touching the next slide "button" and performing the next slide gesture.

Dealing with the less certain command actions is more complicated. We want such actions to result in a change in the probability distribution that indicate that it is more likely (than it was earlier) that the speaker wishes to lecture from the next slide. A simple way of achieving this is to simply subtract some constant from $P(s')$ and add that to $P(s'+1)$. It is unclear if this technique will prove sufficient in the long run – we need to do more experimentation.

5.3 Skipping around in the slides

Another type of command that has not been discussed yet is where the speaker asks the system to skip to a different slide, providing some sort of description of the slide. For example, the speaker might say: "skip to the conclusion slide" or "skip ahead to the slide that discusses ...." In these commands, the speaker indicates that he wants to go to a different slide and then proceeds to describe it in some way. In the first example, he apparently believes that "Conclusion" is sufficient to uniquely identify the desired slide, while in the second example, he indicates that it is a later slide than the current and then describes its contents in detail.

We have identified a number of forms that these commands can take, and have further identified what particular information we can infer about the desired slide (i.e. going forward or backward.) In dealing with this kind of command, the system adjusts the probability distribution based on the particular information it has inferred. So, for skipping ahead, $P(s)$ will be set to 0.0 for all $s \leq s'$ and, for the other values, 1.0 will be distributed among them. Then these slides will be strongly favored when the speaker describes the slide he wants to switch to. As a result, it should only take a few words before the system is able to switch to the desired slide.

A logical extension of the skipping technique is to support free-form lectures where, rather than preparing a formal lecture, the speaker has a large body of slides which the system chooses from as the speaker talks. In such a lecture, the slides can be accessed in any order, allowing the speaker to be very spontaneous in his lecture. To implement this, $P(s)$ will no longer favor the "next" slide, but will instead reject the slides that have already been accessed. This free-form technique would be useful in circumstances where there are libraries of examples (such as: favorite stories or illustrations, or dictionaries of words, animals or commands.)

6 Using the system

Having described the different components that make up the Powerpoint control system, we now take a look at the system in action. For this example, we took the first six slides of a recent talk given by one of the authors (shown across the top of Figure 2) and ran the system as he lectured from those slides. We used IBM's ViaVoice speech recognition software – the transcript of the presentation (as understood by the computer) is shown on the right side of the figure. Underlined words mark the moments where the system decided to change slides. The bar charts on the left represent the probability distribution across the six slides during the course of the presentation. (Time increases down the page in synch with the text on the right.)

In the results of the speech recognition, we get a pretty good idea of what was said (ViaVoice promises accuracy of about 95%.) For the first slide switch, the combination of hearing "explosion" and "on line" was sufficient to switch the slides. But the system ran into a problem in the transition to the fourth slide: "One approach to term weighting is *tfidf*" became "one approach to turn waiting is t f idea of." None of the key words or phrases were correctly recognized. As a result, the speaker had to command the system to go on to the next slide (output as ":next-slide.") At that point, the presentation went smoothly on to completion.

In analyzing these results, we see that, for the most part, the system performed admirably – if it heard one or two of the key words or phrases, it was able to switch slides in a timely fashion. The results also showed how a summary discussion (as in the discussion of the first slide) has the potential to mislead the system. Finally, although the automatic slide switching should be considered the primary technique, being able to fall back on the command-based ones was shown to be very important.

7 Related work

The most interesting research that combines computer vision and speech recognition to facilitate interaction with a human user is being done in Intelligent Environments research. These projects free users from being tied to the keyboard and the mouse. Instead, they allow users to communicate their intentions through speech and gesture – mediums that people usually find more natural and efficient.

At MIT, the Intelligent Room serves as an intelligent command post for disaster relief planning and as a laboratory tour guide (Coen 1998). This project allows the user to command it and ask questions in natural language. Also, through the use of multiple cameras, it is able to track users as they move about and to determine what they are pointing at on one the room's projection screens. HAL, the Intelligent Room's successor, promises to allow many new modes of interaction.

At IBM research, the "VizSpace" project serves as a test-bed for deviceless multi-modal user interfaces (Lucente, Zwart, & George 1998). Through the combination of speech and gesture, its users can manipulate the objects on its display: including adding and removing objects, moving them and resizing them. They hope that this technology will prove useful in designing user interfaces for new kinds of applications.

The Intelligent Classroom and these other projects make use of existing person-tracking computer vision and speech recognition software – the research is in combining them to create an intelligent user interface that allow users to interact with computers in exciting new ways.

8 Future work

In the near future, there are some ways we plan to extend the system to make it more accurate in selecting which slide to switch to and more quick in detecting that it is time to switch slides. One extension will be to expand our techniques for analyzing slide content to find key phrases (as discussed earlier.) Another, is to implement a similar probabilistic approach to follow the individual points on a slide.

Within an individual slide, we can use the structure of the slide to determine when a speaker is more likely to want to switch slides. For reasonably disciplined speakers, we assume that each point will be discussed in order. Therefore, while it is unlikely that they will want to change slides in the midst of one, when they are at the end of a slide, it is very likely. To follow along with the points in a slide, we will use the probabilistic approach discussed earlier, but instead of using entire slides within the whole body of slides, we will use the slide's points within the slide's body.

So, when the system switches to a new slide, it calculates an initial probability distribution for the points in the slide, strongly favoring the initial point. Then, when the speaker talks, his words are used to adjust

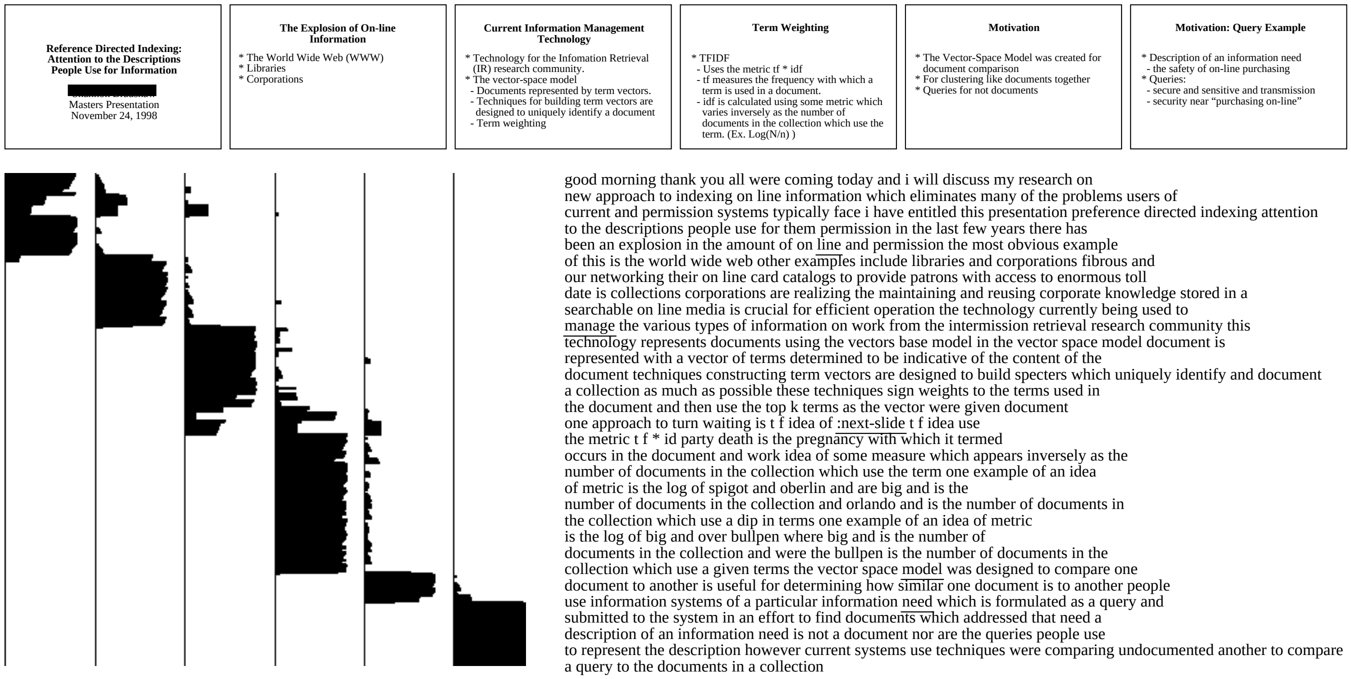

Figure 2: Components of the experiment: The Powerpoint slides (across the top), the words spoken during the presentation (on the right) and the probability distribution

the probability distributions for the current slide and for the slide presentation as a whole. Most of the techniques that we have already discussed are directly applicable to looking at the points within the slide. The speaker's place within the current slide will be considered in determining whether to change slides.

9 Conclusion

This paper describes our efforts in providing multimodal control of Microsoft's Powerpoint presentation software as we embedded it into the Intelligent Classroom. We looked at a command-based approach and a probabilistic approach based on word frequencies, and showed how they could be integrated together. The resulting system provided new and useful ways of performing a slide-based presentation.

This research is part of an overall philosophy and agenda to build systems that exist in the background and are rarely (if ever) explicitly queried or commanded. Instead, such systems wait patiently, acting only when they recognize that they are in a situation where they may be of service. Trusting a program to know when it is time for the next slide is a first step in the path towards a world in which machines are able to truly do what we want, rather than just what we ask.

References

Coen, M. H. 1998. Design principles for intelligent environments. In *Fifteenth National Conference on Artificial Intelligence.*

Flachsbart, J. 1997. Gargoyle: Vision in the intelligent classroom. Master's thesis, University of Chicago.

Franklin, D., and Flachsbart, J. 1998. All gadget and no representation makes jack a dull environment. In *Proceedings of AAAI 1998 Spring Symposium on Intelligent Environments. AAAI TR SS-98-02.*

Franklin, D. 1998. Cooperating with people: The intelligent classroom. In *Fifteenth National Conference on Artificial Intelligence.*

Lucente, M.; Zwart, G.-J.; and George, A. D. 1998. Visualization space: A testbed for deviceless multimodal user interface. In *Proceedings of AAAI 1998 Spring Symposium on Intelligent Environments. AAAI TR SS-98-02.*