Speech, Ink, and Slides: The Interaction of Content Channels

Richard Anderson, Crystal Hoyer, Craig Prince, Jonathan Su, Fred Videon, Steve Wolfman
Department of Computer Science and Engineering
University of Washington
{anderson, clhoyer, cmprince, jonsu, fred, wolf}@cs.washington.edu

ABSTRACT
In this paper, we report on an empirical exploration of digital ink and speech usage in lecture presentation. We studied the video archives of five Master's level Computer Science courses to understand how instructors use ink and speech together while lecturing, and to evaluate techniques for analyzing digital ink. Our interest in understanding how ink and speech are used together is to inform the development of future tools for supporting classroom presentation, distance education, and viewing of archived lectures. We want to make it easier to interact with electronic materials and to extract information from them. We want to provide an empirical basis for addressing challenging problems such as automatically generating full text transcripts of lectures, matching speaker audio with slide content, and recognizing the meaning of the instructor's ink. Our results include an evaluation of handwritten word recognition in the lecture domain, an approach for associating attentional marks with content, an analysis of linkage between speech and ink, and an application of recognition techniques to infer speaker actions.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**]: Evaluation and methodology.

General Terms
Human Factors

Keywords
Digital Ink, Ink recognition, Speech Recognition, Presentation

1. INTRODUCTION
A growing number of systems support the use of digital ink on electronic slides for the delivery of presentations [1][5][8][10]. The motivation for adding ink on slides as a new information channel is to increase speakers' flexibility in presenting material and to support links between spoken utterances and slide content.

We report on an empirical study of ink and speech usage in lecture presentation. The study examined digital archives of five distance learning courses taught at our institution. The archives included complete audio, video, ink, and slide content. We examined usage patterns, paying particular attention to the relationship between speech, ink, and slide content. One of our main goals was to assess how well we could link together the different information channels and determine if — based on real usage data — there was value in combining audio, ink, and slide text in lecture analysis.

Our interest in understanding how ink and speech are used together is to inform the development of future tools for supporting classroom presentation, distance education, and viewing of archived lectures. We want to make it easier to interact with electronic materials and to extract information from them. Our approach is to analyze real lectures and to determine how to use the different information channels in analysis of the lectures. We want to provide an empirical basis for addressing challenging problems such as automatically generating full text transcripts of lectures, matching speaker audio with slide content, and recognizing the meaning of the instructor's ink. Our long run goal is to develop better tools for working with captured lectures, including improved search and navigation tools for lecture viewing, and tools to make digital ink more accessible for blind students.

We begin with background on the study and related work. Our first analysis results are presented in Section 5 where we report on handwriting recognition rates. We evaluated how well a commercial handwriting recognition system performs in the lecture domain since words written while lecturing are potentially harder to recognize than, *e.g.*, private notes. We also analyzed how often words were spoken as they were written to assess the potential for combining speech and handwriting recognition. In Section 6 we present and evaluate initial analysis tools for *attentional* ink, the ink strokes such as circles, underlines, and check marks that link speech to slide content. We wanted to determine how well we could recognize this ink and match it to slide content, since this would be a key step in a higher level analysis of ink and speech. Section 7 describes a detailed qualitative investigation of instructors' use of ink and speech together in five short transcripts of speech matched with ink strokes. The examples all exhibit a pattern of forging a direct tie between spoken phrases and terms on the slides with ink. Finally, Section 8 addresses the problem of recognizing episodes of inking that correspond to instructors correcting mistakes on slides. It is certainly useful to be able to recognize this type of action, but the real motivation was to study the more general problem of activity

inference using ink and potentially speech. Our results are very encouraging for using ink and speech in the analysis of lectures. The recognition rates for handwriting and attentional ink are fairly good, so ink analysis on its own is going to be powerful. We also establish that there is a very tight coupling between the speaker's speech and use of ink which suggest that there will be added benefits in coordinated approach.

2. Related Work

There has been a substantial body of work on capturing and analyzing lectures. Classroom 2000 pioneered work in this area, including integrating digital ink in the presentation, capture, and replay. Other deployments of educational capture systems include Authoring on the Fly [10] and the Cornell Lecture Browser [17]. There has been substantial interest in extracting information from different channels of multimedia materials. Of particular relevance is the work of Chu and Chen which considers implicit and explicit correlation of channels [9].

We are interested in understanding free form use of digital ink. Lopresti [16] discussed ink as a multimedia data type, and Gross [12] and Landay [15] have studied ink with primary emphasis on human to machine communication. Shilman's work [19] on analyzing free form notes is particularly relevant, since it addresses the same type of classification problems as we face. Bargeron [7] has done work on ink based annotation which has parallels to our work with attentional marks. Adler and Davis's [2] work on multimodal sketching is relevant since they study the simultaneous use of speech and ink in design.

3. Ink and Speech Study

We focused on five courses offered in a distance learning Master's program. The courses were video conferenced between two sites, with the instructor in one room and students in both. The instructor lectured from a Tablet PC using Classroom Presenter [5]. Presenter, a presentation system that integrates ink and slides The slides with ink were shown both in the local lecture room and the remote room. Complete audio, video, and ink archives were created of the classes. This has provided us a rich source of data to work with.

The five courses in the study were: Compilers, AI, Databases, Programming Languages and HCI, taught by instructors A, B, C, D, and E respectively. We have full archives of the last four courses. For the first course, we have only one lecture archived in suitable form for analysis in this study. All instructors were experienced in teaching their subjects, and taught from PowerPoint slides. All instructors made significant use of digital ink to write on their slides.

This study was conducted using the recorded versions of the lectures. We constructed a custom replay tool shown in Figure 1. We also analyzed the recorded ink strokes directly with other custom tools. All ink examples in this paper are from actual classes. The speech results in this paper were generated manually by listening to the audio and are not from automatic speech recognition. We chose not to attempt automatic speech recognition at this stage since we felt we could get valuable information about the *potential* of joint speech-ink analysis at far less cost by establishing that there is a tight link between how ink and speech are used. In particular, we were able to study questions such as "are words spoken at the same time they are written?" and "do attentional markings link a phrase of speech to slide content?" without relying on automatic speech recognition.

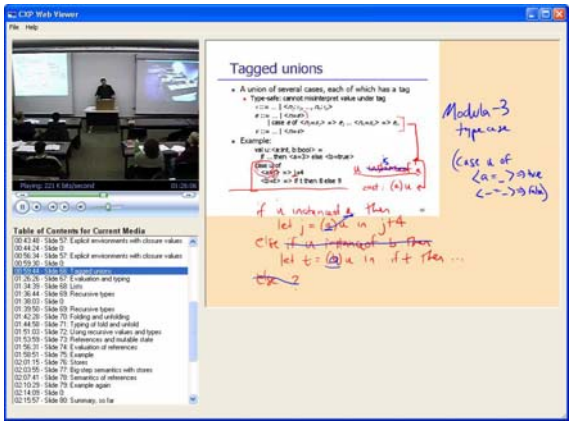


Figure 1 Replay tool used in this study, showing navigation controls, video, and a slide with ink.

4. Use of digital ink

Instructors make rich use of digital ink when they write on slides. In previous work [4] we categorized various usage patterns, and argued that the meaning of ink is often dependent on its spoken context. We have found it useful to break ink into three types: textual, diagrammatic, and attentional. Attentional ink is used to link speech with slide content. Examples of attentional ink shown in Figure 2 include underlines, circles, and arrows. These marks draw attention to the current topic, indicate emphasis, or show connections. Attentional ink is very common, often accounting for more than 50% of inking episodes. The other ink shown on the slide in Figure 2 is textual: short formulas which drew their meaning from speech. Much of the text written on slides is similar: short phrases which depend on context.

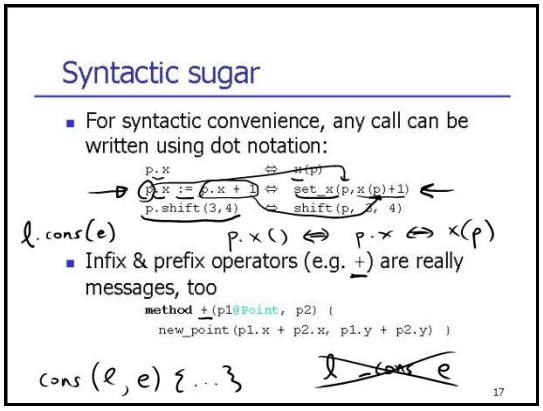


Figure 2 Slide showing attentional and textual ink

5. Handwriting recognition

The first analysis tool we investigated was handwriting recognition to understand how well individual word recognition would work in the lecture domain. Digital ink is an attractive medium for handwriting recognition because ink capture is clean and precise and includes substantial metadata on stroke timing, order, and direction. There is a tremendous body of work on handwriting recognition (see [18] for a bibliography of 255 entries), and recognition is excellent under the right

circumstances. However, the lecture domain presents a potential challenge, writing practices inscrutable to recognizers, and a new opportunity, coordinated recognition of speech and ink, that merit exploration.

5.1 Base recognition

We investigated the question of how well an off-the-shelf handwriting recognizer would work on ink in the lecture domain. The lecture domain presents particular challenges for recognition. The professor may write hastily and sloppily, the writing surface might be at a difficult angle, and the excitement of lecturing might interfere with writing. Our prior experience also shows that written words in lecture are often technical words or abbreviations that do not appear in the dictionary. In our empirical evaluation we took writing samples from the five study courses, and ran them through the Tablet PC recognizer. Our goal was to see what could and could not be recognized, to get a feel for where the boundary is, how much potential there is for using an off-the-shelf recognizer in this domain, and understand how to select or customize recognizers for future analysis work.

Segmenting and recognizing text To analyze the Tablet PC's recognizer's effectiveness on lecture ink, we first isolated the text from attentional and stray ink. Using a custom-built ink serialized format (ISF) viewer we manually segmented the ink, building a corpus containing all the isolated text in the study, and ran only this corpus through the recognizer. It was important to isolate the textual ink because current recognizers do not filter out non-word ink and therefore try to process it as text. (Note that in Section 6.1 below, we discuss automating this segmentation.) Version 1.5 of the Tablet PC SDK was used for all experiments.

We also took advantage of the Tablet PC's ability to segment collections of ink into groups words. Slides with several distinct episodes of writing were first divided into individual episodes and then run through the recognizer. This prevented the recognizer from trying to recognize an entire slide as one sentence. Most of the ink on slides is not in sentence format but rather sentence fragments or just words.

Recognition Analysis After feeding the text segments through the recognizer; four team members coded the accuracy of the results. Every slide was evaluated by one coder who recorded what they thought the instructor was trying to write. If the top recognition result matched the coder's result, the episode was labeled "exact." If the coder's result was in the recognizer's alternates list, the episode was labeled "alternate." If the list of results did not contain the coded result but did contain a close match (i.e. only punctuation marks prevented a match), the episode was labeled "close." Otherwise, the episode was a mismatch and was labeled "none." Slides with only handwritten formulas or code with words embedded in those formulas (such as "list" and "int") were ignored in our results. Table 1 summarizes the results.

Table 1 Text recognition results per instructor.

	Exact	Alternate	Close	None
A	16 (88%)	1 (6%)	0 (0%)	1 (6%)
B	146 (59%)	26 (10%)	6 (2%)	71 (29%)
C	18 (42%)	5 (11%)	1 (3%)	19 (44%)
D	262 (61%)	45 (11%)	9 (2%)	111 (26%)
E	408 (79%)	46 (9%)	2 <(1%)	58 (11%)
Total	512 (56%)	126 (14%)	18 (2%)	260 (28%)

These results show that the off-the-shelf Tablet PC recognizer has a high success rate. Figure 3 shows some specific cases prevalent in the lecture domain that were successfully recognized: (a) the word "Queries" written indistinctly using both cursive and print letters, (b) the word "marketing," misspelled, and (c) the word "speed" written awkwardly at a steep angle and misspelled.

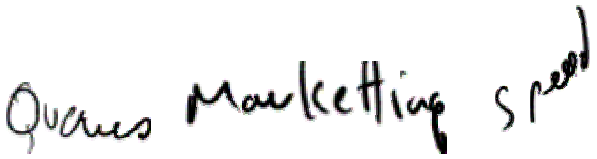


Figure 3 Successful recognition results: (a) "Queries," (b) "Marketing," and (c) "Speed."

Failure factors The above discussion illustrates that the recognizers are able to recognize text in a variety of situations. There were also important factors that caused the recognizer to fail. Some of these factors (such as writing at an angle) did not prevent the above examples from being recognized, which indicates that context (word angle), word choice (some words are more distinct) and instructor handwriting greatly influence the recognizer's ability.

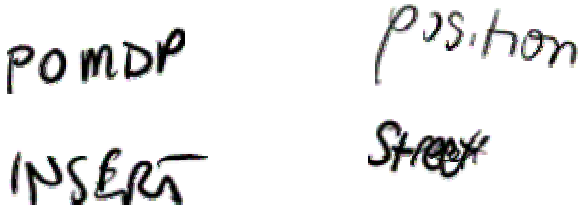


Figure 4 Unsuccessful recognition results: (a) "POMDP," (b) "position," (c) "INSERT," and (d) "Street."

Factors which decreased the likelihood of recognition were: non-dictionary words, steep angles, bad or non-traditional handwriting, and double ink. Figure 4 illustrates some of these failures. In Figure 4 (a), the instructor wrote the acronym "POMDP" which was not recognized (although it is clear to viewers), because it is not in the recognizer's dictionary and because it is in all caps. The failure on "position" is somewhat surprising, it is at a slight angle, but we conjecture that the problem is that the letter "p" is nonstandard. Other words containing "p" by this instructor also failed. Instructors each have different handwriting, but we found that this affected recognition only when non-traditional or caps letters were used for the entire word, such as "INSERT" in Figure 4 (c). Finally, instructors sometimes wrote over previously written ink which confuses the recognizer, Figure 4 (d).

5.2 Joint Writing and Speech Recognition

Analyses of ink and audio streams used in concert have the potential to inform each other. To explore this potential, we propose three methods for combining handwriting and speech recognition techniques and provide evidence suggestive of these methods' potential. The three methods are: directly combining speech and handwriting recognition on phrases that are both spoken and written, recognizing and matching terms on one channel (ink or audio) that would be impossible to accurately recognize in the other, and triangulating across the channels to disambiguate recognized words.

In the lecture domain, it is plausible that many words are spoken as they are written; if this is the case, aligning and combining likely word recognition results from the two streams promises improved recognition. As shown above, baseline handwriting recognition in the lecture domain is good. Furthermore, because of the persistence of ink and the extra effort required to write versus speak a word, the words an instructor writes are often critical terms. For example, Figure 5 shows a written answer to an exercise in one class which the instructor spoke quickly (in six seconds) and then carefully rewrote (in fourteen seconds) for emphasis. Improving recognition of these critical words is particularly important for generating video indexes.

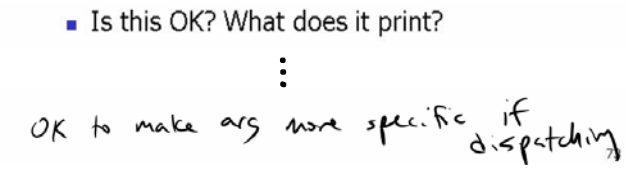


Figure 5 An inked answer to a question posed on the slide. (The example referred to in the question has been elided for space.)

To understand whether coordinated analysis could improve recognition, we investigated whether a sample of written words were also spoken. We generated our sample to be representative of isolated English text on the slides by selecting written words from each of the five courses but filtering out episodes consisting purely of formalisms (e.g., code fragments and grammars), diagram labels, or artifacts of study (e.g., sample sentences for natural language processing). We built the sample by selecting slides with writing at random from the corpus of isolated text described above, continuing in round-robin fashion across the courses until we had at least ten slides from each instructor except A (whose data we exhausted). We replayed each episode and recorded whether the written words were spoken exactly and the time lag between writing and speaking or vice versa, if any.

Table 2 shows the results of our study of spoken and written words. Most episodes — written words or phrases — were spoken exactly, and *all* episodes were spoken at least approximately, with most approximations simply adding or dropping stop words (*e.g.*, "the" or "and"). Most written episodes were spoken during the writing process although speech often ended before writing was finished or started after writing began. The vast majority of episodes occurred within two seconds of corresponding speech.

Table 2 Coexpression of speech and writing. The left columns show the count of writing episodes that were also spoken for each instructor, "approximately" or "exactly." The right columns show the time gap in seconds between speech and writing.[1]

	Exact	Approx	None	Simul	0-2s	> 2s
A	1 (100%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
B	9 (75%)	3 (25%)	0 (0%)	12 (100%)	0 (0%)	0 (0%)
C	9 (82%)	2 (18%)	0 (0%)	10 (91%)	1 (9%)	0 (0%)
D	12 (86%)	2 (14%)	0 (0%)	10 (71%)	4 (29%)	0 (0%)
E	9 (56%)	7 (44%)	0 (0%)	7 (44%)	4 (25%)	5 (31%)
Total	40 (74%)	14 (26%)	0 (0%)	40 (74%)	9 (17%)	5 (9%)

[1] Students spoke three episodes from E (all "> 2s" entries); otherwise, all episodes were spoken by the instructor.

This strong correlation between speech and writing indicates significant potential for handwriting recognition to support and inform speech recognition. These results apply only to the unfiltered episodes, about half of the total episodes in the sample[2]; however, we felt the unfiltered episodes were both the most important and the most easily isolated episodes, making them the most amenable for automation.

Coordinated use of ink and audio streams also promises potential to recognize words that could not be accurately recognized otherwise. In several cases, instructors wrote technical terms or names of people, products, and companies that could not be transcribed from speech but might be recognized by combining writing and speech. Figure 6 shows words with challenging spelling that the instructor both wrote and spoke: a list of paper authors, a company, and a research product. Another potential synergy between ink and audio is connecting acronyms and abbreviations to their expanded forms. Figure 7 shows a written acronym and an abbreviation which were also spoken in full.

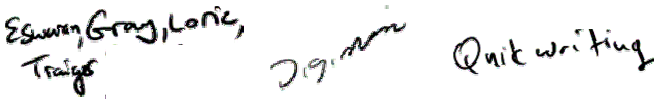


Figure 6 Three episodes unrecognizable from speech alone: "Eswaran, Gray, Lorie, Traiger," "DigiMine," and "Quik Writing."

Figure 7 A written acronym, "Java 2 Enterprise Edition", and abbreviation, "straight", both expanded in speech.

Finally, the techniques used for parallel, multilingual document corpora may lead to more ambitious coordination of speech and ink. For example, bilingual corpora have been used successfully to assist in word sense disambiguation, selecting a word's intended meaning from among a set of definitions[9]. Ambiguities arising in speech (*i.e.*, homonyms) may be unambiguous in writing or vice versa. Figure 8 shows one example of potential disambiguation in each direction (speech to writing and writing to speech). Although examples such as these were rare in our data, courses with less prepared slide content and correspondingly more writing might present more disambiguation opportunities.

Figure 8 An ambiguous written word ("1962"/"1912") that was spoken clearly ("Nineteen sixty-two") and an ambiguous spoken word ("serial"/"cereal") that was written clearly ("serial").

6. Attentional Marks

Attentional marks are ink annotations, such as circles and underlines, that provide linkage between spoken context and slide content [4]. These markings serve a variety of purposes including resolving deictic references (as with physical pointing gestures), grouping related slide elements, and emphasizing important points. In each case, an attentional mark draws attention to a piece of slide content. Recognizing which ink constitutes attentional

[2] We filtered out approximately 50 episodes. The number is imprecise because the filtered episodes were never segmented.

marks and determining the content referenced by each attentional mark would facilitate automated analysis of lecture content by identifying points given special emphasis in lecture, anchoring speech acts to slide content, and providing timestamps for discussion of individual bullets and words.

6.1 Processing attentional marks

We begin by describing our efforts at identifying and recognizing attentional ink. The basic problem is to identify ink strokes that are attentional and then match them with the underlying slide content.

Identifying attentional marks We want to automatically segment ink into textual ink, diagrammatic ink, and attentional ink. The problem, is in general, very hard. Our experience is that even though humans can generally agree on which strokes are attentional marks, there are some tricky cases that require semantic knowledge of the ink to classify correctly. However, at this stage, we would like to be able to recognize the common cases that constitute the bulk of attentional marks: check marks, underlines, and circles.

Our observations indicate that the majority of attentional markings consist of very few strokes: circles, underlines, etc. Furthermore, when an attentional mark comprises more than one stroke, these strokes are usually drawn within close spatial and temporal proximity.

We therefore segregate the ink on a slide into distinct groups of strokes based on temporal proximity. Each stroke group is a candidate attentional mark. We then filter out non-attentional ink, such as text or diagrams, by removing groups formed from a large number strokes.

Recognition of attentional marks Matching attentional ink to slide content requires classifying the type of the ink because each type picks out slide content in a different way. Variations in representation of attentional marks make classification difficult. Different instructors draw the marks in different ways; *e.g.*, some instructors draw their circles rounder than others.

In this study, we focused on a subset of types of attentional marks that represent the majority of attentional ink: boxes/circles, underlines, and bullets (ticks, dots, check marks, arrows). Recognizing these common cases allows us accurately map a large portion of attentional ink to slide content. With this data, we can then investigate the value of these mappings.

Matching with source content The final step of processing that occurs on attentional ink is to determine what content the ink picks out from a document. Source documents in our study were PowerPoint. To facilitate mapping, we capture the source text and geometry, which we represent as a collection of nested rectangles, labeled with the source text. Similar data would be derivable for most source formats.

To extract the slide geometry, we developed a tool that scans through each PowerPoint slide and pick out all the words. For each word, we construct a rectangle representing the bounding box of the word, and associate it with the word content. As a result, we have a representation of each slide as a collection of words and their locations on the slide. Then, based on the type of attentional mark we have recognized and the ink's location on the slide, we can match it appropriately with the content around it.

6.2 Recognition Results

To analyze the effectiveness and accuracy of our attentional mark processing, we studied attentional mark usage in four courses offered in our department. From these lectures, we chose 71 slides each with a large and diverse amount of attentional ink.

We first processed all the ink from these selected slides using the methods described above, generating a list of the attentional marks that occurred on each slide, and what content each mark was referring to. Then, we gave half of these slides to two coders and had them independently code each slide for the types of attentional marks they contained and what content each mark referenced. The other half of the slides was analyzed similarly by two other coders.

We compared our automatically generated results to those given by our four coders to determine the accuracy of our automatic processing. We classified the results into exact matches, exact matches up to leading or trailing punctuation (only periods and commas), close matches (off by fewer than 2 words) and non-matches. The results are summarized in Table 3.

Table 3 Accuracy of attentional mark classification and identification of referenced slide content.

	Exact	Exact to Punctuation	Close	Non-Match	
Circles	70	13	6	17	106
Underlines	207	22	44	66	339
Bullets	52	0	0	35	87
	329	35	50	118	532

Our results indicate a good recognition rate for a baseline. Overall, the exact match rate was 62%, and the match rate (including "exact to punctuation" and close matches) was 78%. Close and especially "exact to punctuation" matches were often of high quality. Figure 9 shows two examples where there were discrepancies at the punctuation level.

Figure 9 (a) Automatic recognition was "λ(x1, x2)." while manual recognition was "λ(x1, x2)". (b) Automatic recognition was "("gossips")" while manual coders differed, recognizing "gossips", ""gossips"", and "gossips")".

Among non-matches, there were a few common problems on the level of mapping recognized attentional marks to content. These failures occurred when there were discrepancies between the slide geometry extracted from PowerPoint and the geometry humans constructed when viewing a slide.

One such difficulty arose when the geometry was too coarse. Currently, geometry recognition stops at the word level. Usually, this choice corresponds with instructor intent to highlight a word or a group of words. However, instructors occasionally pick out particular characters or groups of characters within a word, which requires finer geometry segmentation as in Figure 10.

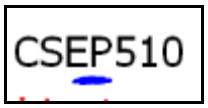


Figure 10 An underline emphasizing the letter "P" within a word.

Another challenge in the matching process is when the geometry generation fails to separate words where humans would expect them to be split as in Figure 11. In this example, the recognizer did not segment the phrase into five separate words because there was no inter-word space. However, as shown by the underlines, the instructor clearly wanted to emphasize each word in the phrase separately. In this case, none of the underline marks were recognized since none of them underlined a significant portion of the entire phrase "this.inkCollector.Ink…".

Figure 11 Underlines intended for four of the five words in the phrase. Our recognizer interpreted the entire phrase as one word.

Our system already does well in recognizing certain attentional marks and matching them to slide content. However, we could further improve the recognition by solving the problems detailed above. Furthermore, recognition of the less common types of attentional marks (brackets, overbars, etc.) will be necessary to create a complete summary of attentional mark usage in lectures.

7. Ink and speech in Formula Tracing

One of the goals of this work is to understand how the speech and ink channels are tied together. We would like to construct static summaries of both the speech (a text transcript) and the slide content (annotated slides). Two questions are: "How can an analysis of inking aid in constructing a text transcript?", and "How can an analysis of speech aid in annotating slides?" To study these questions, we took one short inking example from each of the instructors, and manually created speech transcripts. These are shown in Figures 12 through 16 with the instructors' speech and the length of the episode, in seconds, in the captions. The examples were chosen to have roughly the same length and amount of ink. The transcripts were created by hand and do not include disfluencies. The approximate end of ink strokes are indicated with capital letters in the transcripts. Synchronization in the replay tools is too rough and speech too fast to precisely anchor the starts and ends of strokes to corresponding speech.

7.1 Formula Tracing Examples

Figure 12 Instructor A: "Take a look, if the starting time (A) of that instruction plus the delay (B) it takes is less than or equal to the current cycle (C) it's no longer active so pull it out of the active queue." [0:11]

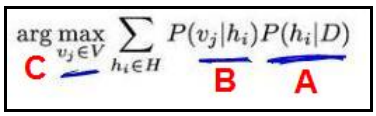


Figure 13 Instructor B: "You multiply the probability of that hypothesis (A) given the data, times the probability of getting that particular classification (B) where vee-jay is the classification given that hypothesis, take the sum and then find the classification that maximizes that quantity (C)" [0:22]

Figure 14 Instructor C: "Transaction one releases its read lock on x (A) before it sets its write lock on y (B), so at this moment in time (C), after transaction one has released its read lock on x, there are no locks in the system, so transaction one can set a read lock on y and read y (D) set a write lock on x and write x (E) and it finishes up and commits (F)." [0:30]

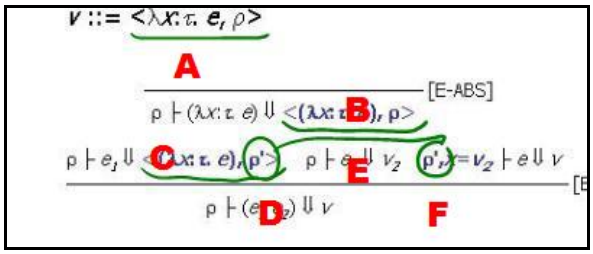


Figure 15 Instructor D: "So I will change my values to be this pair (A) written in angle brackets of the function and the environment. Now my lambda rule (B) will remember the environment when I evaluate my function (C) I will get back the closure, not just the function code and then I can use rho prime (D), this thing I remembered, here (E) (F)." [0:27]

Figure 16 Instructor E: "(A) Inside the form initialization, all I did is create the ink collector (B) and tie it to the form (C). This time I did, in fact, remember to turn the ink collector on (D) and initialize the recognizer (E)." [0:23]

7.2 Analysis of examples

The examples above all exhibit a pattern of underlining terms on formal text. This pattern occurred often in all five study courses, and we have observed it in many other courses. The pattern is to make frequent use of underlining while narrating a formula, program, or other formalism. In the examples above, strokes were written about once every five seconds.

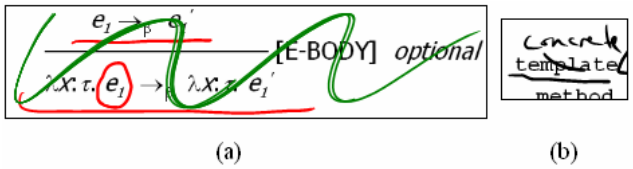
In all of the examples, the ink provides a strong link between speech and content. For every stroke but one, there is a spoken phrase and a term on the slide that can be matched with the ink stroke. The exceptional ink stroke (A on Figure 16) is unusual because it has no corresponding speech. The linkages could be used to label slide content with its corresponding speech, for example, in Figure 12 , S(op) could be labeled "start time".

In Figures 12, 13, and 16 the discussion is parallel to the slide content, with no verbal references to ink. Figure 14 has a direct reference to ink when Instructor C says "at this moment in time" and draws the down arrow labeled C. This example is slightly different from the others in that the instructor is simulating a process, and not just describing something. Figure 15 has two cases of deictic reference which are resolved with ink. At the start of the example, Instructor D says "I will change my values to be *this* pair", and underlines the term. The slide had two other terms which match the reference; so, the ink resolution was necessary. The second case occurs at the end of the same example where "rho prime" is identified with strokes D, E, and F.

8. Recognizing Corrections

Our last example is to look at recognizing particular activities. In this case we identify correcting mistakes on slide. The broad question is can we recognize patterns of activity by analyzing the ink and speech channels. We might want to do this in order to do a break down of the lectures into activity regions (e.g., lecturing, examples, questions, code walk thoughts, collective brain storming) for navigation or summarization. Here, we look at a specific activity, an instructor marking a correction on a slide. This occurs frequently, and recognizing these could be useful to inform the instructor of corrections to make on the slides prior to using the slides again.

Not only is this identification task interesting, it is also by no means trivial. First, compared to other inking, slide corrections occur only a few times per slide deck. As a result, there is not a large body of data upon which to base a classification system. Second, the intent of the ink is often ambiguous or uncertain. Ambiguity is the biggest challenge for the recognition of slide corrections. At times it is difficult to determine if an ink stroke is part of a correction or if it was used during a presentation as part of an explanation. For example, Figure 17 (a) shows both a formula crossed out and Figure 17 (b) shows crossed out words with replacements above them. In both these cases, these are not corrections, but explanations of the material.



(a) (b)

Figure 17 Ambiguous marks that appear to be corrections but were actually part of the explanation of the slide.

In addition, writing is often hasty and imprecise, causing ambiguities of meaning outside the context of the lecture. For example, Figure 18 shows a sentence that was apparently crossed with a replacement written to the right; however, this is actually a sloppy underline for emphasis with explanation to the right.

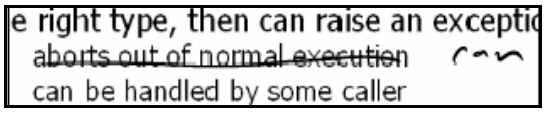
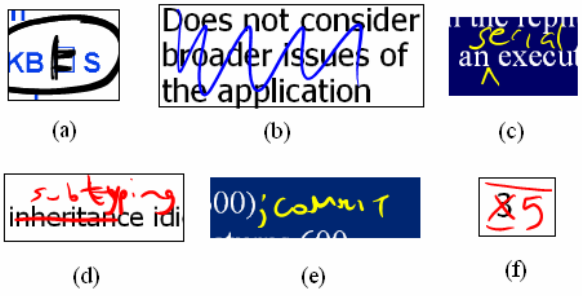


Figure 18 Messy handwriting can make a mark appear to be a correction when viewed later.

Without actually listening to the presentation it is impossible to know if these ambiguous ink strokes are corrections or not. However, this implies that the speech channel can be used in the future for this disambiguation.

8.1 Identifying corrections

We developed a proof-of-concept application for the identification of slide corrections. The first step in this process was to determine the types of slide corrections made by individuals. We analyzed slides from four instructors, and Figure 19 shows the classes of correction marks we observed.



(a) (b) (c)

(d) (e) (f)

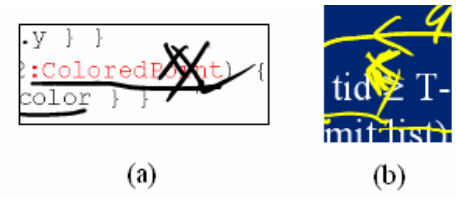
Figure 19 Six classes of corrections: (a) Overwrites, (b) Squiggles, (c) Text-inserts, (d) Strike-throughs, (e) Added text, and (f) Cross-outs.

After identifying the types of corrections, we defined features to help classify these marks. The first feature is the percentage of a stroke that overlays slide content. The intention was to identify ink written over text, which our analysis suggests correlates with slide corrections. The second feature is the number of times the stroke crosses its mean height, which was designed to identify "squiggles" (see Figure 19 (b)). We reduced noise substantially by filtering out attentional marks (except strike-throughs) before classification. Attentional marks rarely form part of corrections, yet they account for the majority of ink strokes. We used the classifier from Section 6 above to perform the filtering.

Finally, we built a simple decision tree using the features above with hand-tuned parameters to determine if a stroke was a slide correction.

8.2 Examples

In this section we will outline a sampling of the more interesting corrections we were able to recognize with our system. Figure 20 shows two of our system's surprising recognition successes In Figure 20 (a), the system recognized the two X's over the word "ColoredPoint" and flagged them as a correction. This is unusual because only a small portion of the word is covered by the X's, unlike most corrections where the entire word is crossed out. In Figure 20 (b), our system detected a symbol crossed out by the instructor although the correction (a "≥"changed to a "≤") is extremely messy. This particular correction is so cluttered that it could easily be overlooked by a human reviewing the slides.



(a) (b)

Figure 20 Two surprising successes of correction identification.

Figure 21 shows our squiggle detection at work. Although most of the stroke is not over slide content we were still able to identify this correction mark because our classifier saw that it had many alternations (see Section 8.1) and covered at least some slide content and thus classified it as a possible correction.

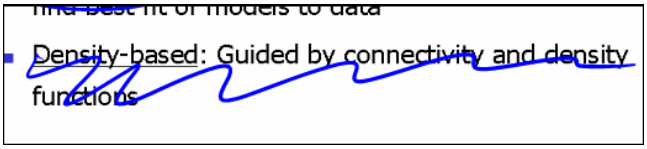


Figure 21 A squiggle correctly detected as a slide correction.

9. Conclusions

In this paper, we have described our study of speech and ink use in lecture presentation. To recap our results:

- Handwriting recognition: In spite of being a difficult domain, commercial handwriting recognition works fairly well on real lecture data. Across five instructors our recognition rate was 68% (77% with alternates).
- Speech-writing co-expression: We measured the frequency that words were spoken, when they were written. Remarkably, in the 54 examples of writing we examined, all of them were spoken.
- Attentional inking: We developed methods for identifying the slide content referred to with attentional marks. Our measured results had a high correspondence with human coders: 61% exact and 78% close or better.
- Formula tracing: Underlining when describing formulas and code is a pattern we observed in all instructors. In all cases this reflected direct ties between phrases in speech and terms in the slide content, indicating potential for augmenting transcripts and screenshots.
- Correction recognition: we were able to identify episodes that were likely to correspond to corrections on slides. This shows that some activities can be recognized by analyzing ink.

Taken together these results give a strong basis for using ink and speech together in the analysis of recorded lectures. We have shown that the basic ink analysis is tractable and gives good results, and that there are strong ties between ink and speech where recognition across channels could be mutually reinforcing.

10. REFERENCES

[1] Abowd, G., Classroom 2000: An experiment with the instrumentation of a living environment. IBM Systems Journal, Voume 38, Number 4, 1999.

[2] Adler, A, and Davis, R., Speech and Sketching for Multimodal Design, Intelligent User Interfaces'04, pp. 214-216, 2004.

[3] Altman, E., Chen, Y., and Low, W., Semantic Exploration of Lecture Videos, ACM Multimedia'02 pp.416-417, 2002.

[4] Anderson, R. J., Anderson, R. E., Hoyer, C. L., and Wolfman, S., A Study of Digital Ink in Lecture Presentation. CHI'04, pp. 567-574, April, 2004.

[5] Anderson, R. J., Anderson, R. E, Simon, B., Wolfman, S., A., VanDeGrift, T., and Yasuhara, K., "Experiences with a Tablet PC Based Lecture Presentation System in Computer Science Courses," SIGCSE 2004, pp. 56-60, 2004.

[6] Bacher, C., and Muller, R., Generalized Replay of Multi-Streamed Authored Documents, Proceedings of ED-Media, Freiburg, 1998.

[7] Bargeron, D., and Moscovich, T., Reflowing digital ink annotation, CHI'03, pp.385-392, 2003.

[8] Berque, D., Bonewrite, T., and Whitesell, M., Using Pen-Based Computers Across the Computer Science Curriculum, 35th ACM SIGCSE, pp. 61-65, 2004.

[9] Chu, W-T., and Chen, H-Y., Cross-Media Correlation: A Case Study of Navigated Hypermedia Documents, Multimedia'02, pp. 57-66, 2002.

[10] Fridland, G., Knipping, L., Rojas, R., E-Chalk: Technical Description, Technical Report B-02-11, FU Berlin, Institut fur Informatik, May 2002.

[11] Gale, W. A., Church, K. W., and Yarowsky, D. "Using bilingual materials to develop word sense disambiguation methods." Int'l. Conf. on Theoretical and Methodological Issues in Machine Translation, pp.101–112, 1992.

[12] Gross, M. D., and Do, E. Y., "Drawing on the Back of an Envelope: a framework for interacting with application programs by freehand drawing," Computers & Graphics, 24 pp. 835–849, 2000.

[13] Jarrett, R., and Su, P., Building Tablet PC Applications, Microsoft Press, 2002.

[14] Liao, C., Liu, Q., Kimber, D., Chiu, P. Foot, J., and Wilcox, L., Shared Interactive Video Teleconferencing, ACM Multimedia'03, pp. 546-554, 2003.

[15] Landay, J. A., and Myers, B. A., Sketching Interfaces: Toward More Human Interface Design, IEEE Computer, Vol 34, No. 3, pp 56–64, March 2001

[16] Lopresti, D., Ink as Multimedia Data, Proceedings of the Fourth Intl. Conference on Information, Systems, Analysis and Synthesis, July 1998, Orlando, FL, pp. 122-128.

[17] Mukhopadhyay, S., and Smith, B., Passive Capture and Structuring of Lectures, ACM Multimedia '99, Orlando, Fl, pp. 477–487, 1999.

[18] Plamondon, R., and Srihari, S., N., On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey, IEEE PAMI, 22(1), pp. 63-84, January 2000.

[19] Shilman, M., Wei, Z., Raghupathy, S., Simard, P., and Jones, D., Discerning Structure from Freeform Handwritten Notes, ICDAR 2003.