RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review

Dongwook Yoon[1, 2] dy252@cornell.edu
Nicholas Chen[1] nchen@microsoft.com
François Guimbretière[2] francois@cs.cornell.edu
Abigail Sellen[1] asellen@microsoft.com
[1]Microsoft Research
21 Station Road, Cambridge CB1 2FB, UK
[2]Cornell University, Information Science
Ithaca, NY 14850

ABSTRACT
This paper introduces a novel document annotation system that aims to enable the kinds of rich communication that usually only occur in face-to-face meetings. Our system, RichReview, lets users create annotations on top of digital documents using three main modalities: freeform inking, voice for narration, and deictic gestures in support of voice. RichReview uses novel visual representations and time-synchronization between modalities to simplify annotation access and navigation. Moreover, RichReview's versatile support for multi-modal annotations enables users to mix and interweave different modalities in threaded conversations. A formative evaluation demonstrates early promise for the system finding support for voice, pointing, and the combination of both to be especially valuable. In addition, initial findings point to the ways in which both content and social context affect modality choice.

Author Keywords
Annotation; multi-modal input; voice; speech; pointing gesture; pen interaction; collaborative authoring; asynchronous communication.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION
The production of documents where collaborators iteratively review and exchange feedback is fundamental to many workplace and academic practices. Effective communication of edits, questions and comments is central to the evolution of a document [27] as well as playing a role in maintaining group dynamics [2].

Working face-to-face (F2F) has many advantages for these kinds of collaborative processes. F2F collaborators enjoy a shared context in which they can verbally explain details and gesture over documents with each other, often taking notes as they do so. These different ways of communicating are interleaved seamlessly and often support each other. It is therefore no surprise that many of the most important discussions about documents occur in face-to-face meetings [8,20].

UIST '14, October 05 - 08 2014, Honolulu, HI, USA

http://dx.doi.org/10.1145/2642918.2647390

Figure 1. RichReview running on tablet. Hovering the pen over screen leaves traces of gesture (blue blob on top). Inking can be done on expansion space (middle). Voice recording is shown as waveform (bottom).

One challenge of F2F meetings (and related techniques like video conferencing) is that they constrain collaborators to be co-present (temporally), which may not always be possible or desirable. In response, people often collaborate asynchronously over communication channels such as ink markup, textual annotations, and email. These techniques, compared to F2F interactions, undermine the production of an implicit shared context and largely restrict communications to a single modality.

In this paper, we describe RichReview, a document annotation system that brings some of the richness and expressivity of F2F discussion to asynchronous collaboration scenarios. RichReview allows collaborators to quickly produce and consume annotations consisting of voice, ink, and pointing gestures on top of ordinary PDF documents.

On the creation side, RichReview introduces a unified, fast, and minimally intrusive set of interactions for creating multi-modal annotations on text. The three different modalities that are supported can be freely interleaved and

combined. RichReview additionally leverages automatic speech recognition (ASR) to segment audio recordings at the word level to simplify the process of trimming or cleaning up audio.

On the consumption side, RichReview's interface streamlines navigation and access to the contents of these rich annotations. Any visual representations of RichReview annotations are recorded with time-stamps to support quick navigation. For instance, voice annotations are rendered as waveforms annotated with an ASR-generated transcript. This provides an interface to easily browse and access any point in the audio stream. RichReview also leverages the time-synchronization between voice, ink, and gesture streams by providing users the ability to use one modality to index into another.

We conducted a formative study investigating how people use (and do not use) RichReview features when discussing documents. RichReview's support for ink, voice and pointing gestures was widely used. Users also took advantage of the freedom and flexibility that RichReview afforded by structuring and responding to annotations in a variety of ways. Based on our results, we discuss design implications for future implementations, including enhancing time indexing and making the system more practical at scale.

RELATED WORK

RichReview has its roots in the Wang Freestyle system [10] which pioneered the use of a combination of speech and ink to annotate a document. It also builds on research that shows that combining modalities allows people to communicate more efficiently and with more depth [19]. However, RichReview's support for annotation production goes beyond existing work by capturing pointing gestures in addition to speech and ink. RichReview also provides improved support for consuming annotations in the form of new visualizations and interactions. The contribution of this work, therefore, is to build on the work of others as we shall outline below. At the same time we wish to broaden the flexibility and expressiveness with which annotations can be made without added complexity for the user.

Ink

Freeform ink annotations are pervasive and used extensively for document work because they are fast to create, can be interleaved with the reading process [20], and are highly flexible in the information they represent [11]. As a result, several annotation systems in the literature have employed ink as a primary modality. The collaborative editor MATE [7] supported the use of ink both for low-level editing commands as well as serving as a general medium for communication. Similarly, the XLibris reading device, which supported pen input, was used to explore various use scenarios of collaborative ink annotations [12]. Unlike these previous systems, ink is not the exclusive annotating modality of RichReview. Rather, it is used in conjunction with gestural and speech-based annotations. Moreover, RichReview employs contemporary techniques such as TextTearing [26] to alleviate issues with limited writing space on digital documents.

Gestures

People often use gestures in a deictic role (i.e. pointing to areas of interest) to streamline discussion around a document [1]. Gestures help people establish a common understanding and offer a shortcut to verbose verbal descriptions [4]. As such, they are an integral complement to spoken language [13]. BoomChameleon [22] is one of the first systems to explore pointing gestures in the form of the Flashlight tool which allowed users to refer to regions of interest in 3D environments. RichReview employs a Spotlight tool to achieve similar functionality in textual documents. A key difference between Spotlight and Flashlight is that Spotlight traces can be used as an index to rapidly jump into the middle of an annotation.

Speech

Speech has been shown to be a uniquely strong medium for identifying high-level problems with a document. Because speaking is faster than writing or typing, it is an efficient way to display complex concepts. Chalfonte and Kraut have also shown that spoken annotation's expressiveness and richness are more suitable for describing structural or semantic issues in comparison with written annotation [3,8]. Furthermore, Neuwirth et al. found that speaking, when compared with writing, generates more detailed explanations and nuance that can lead to better perceptions of comments at the receiving end [16].

Despite these advantages, speech-based annotation is rarely employed. Ethnographic studies of writing [18] make no mention of the use of audio commenting features available in word processing packages. One reasons may be that, as Grudin has noted, speech is slower and more difficult to access than text, which can undermine its use in collaborative applications [6].

One way voice annotation systems have dealt with this accessibility problem is by using ink strokes as navigational indices into an audio stream [21,24,25]. This approach is also used in commercial applications such as the LiveScribe Pulse SmartPen and Microsoft OneNote. Another strategy is to use automatic speech recognition to produce a textual transcript to enable faster browsing and access to the underlying audio [23]. Given the diversity of annotation strategies, RichReview includes both techniques to maximize navigational flexibility: either ink or text can be used to browse and navigate through speech annotations. This combined approach for accessing speech content is similar to what is used in the NoteVideo [14] online lecture browser. It bears noting that all annotation elements, regardless of whether they are voice, ink strokes or Spotlight, can participate in cross-indexing operations.

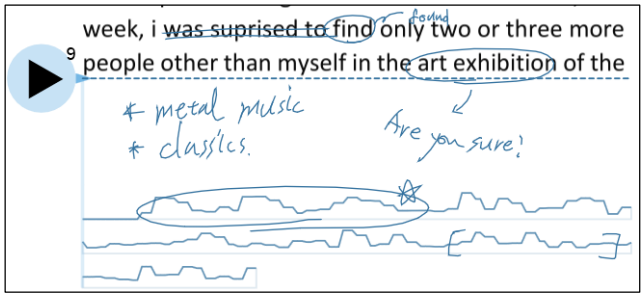

Figure 2. A mixture of static ink annotations along with the playback control for a multi-modal annotation.

RICHREVIEW DESIGN PHILSOPHY

Based on our analysis of previous work, it was apparent that a successful, fully functional, multi-modal annotation tool has yet to be developed. However, equally apparent was that such a system needed to be sensitive to the way that people create and use annotations. With that in mind, we proposed the following goals:

- ***Limiting System Complexity*** Introducing multiple annotation modalities runs the risk of bringing in additional complexity and overhead. The added overhead could then affect all annotation activities. Therefore a significant part of the design of our system was focused on ensuring that annotations are created and consumed in ways that are lightweight and fluid. Satisfying this design goal argued against locking the user into interaction modes or adding additional interaction steps. RichReview employs a simple and consistent set of interactions for creating any kind of annotation.
- ***Versatility and Choice*** The literature provides many examples showing that the optimal modality for communicating varies by its content and purpose. For example, inking is popular for lightweight copyediting, and a combination of voice and pointing can be useful to describe structural issues of a writing. For this reason, a second design goal is to allow users to employ a flexible mix of annotation modalities.
- ***Balancing Emphasis on Production and Consumption*** The success of groupware is contingent on the balance of benefits to different stakeholders [6]. Thus an annotation system that supports collaborative tasks must focus as much on improving the ability of recipients to skim, access, and revisit annotations as it does on supporting the creation of them in the first place. Given that non-textual content can be difficult to access and skim, we place an emphasis on techniques that assist users in consuming rich annotations.

RichReview was expressly designed for use with tablet devices, since tablets are a preferred form factor for active reading activities [15]. Moreover, current tablet devices contain the necessary hardware to capture the three input modalities we are interested in.

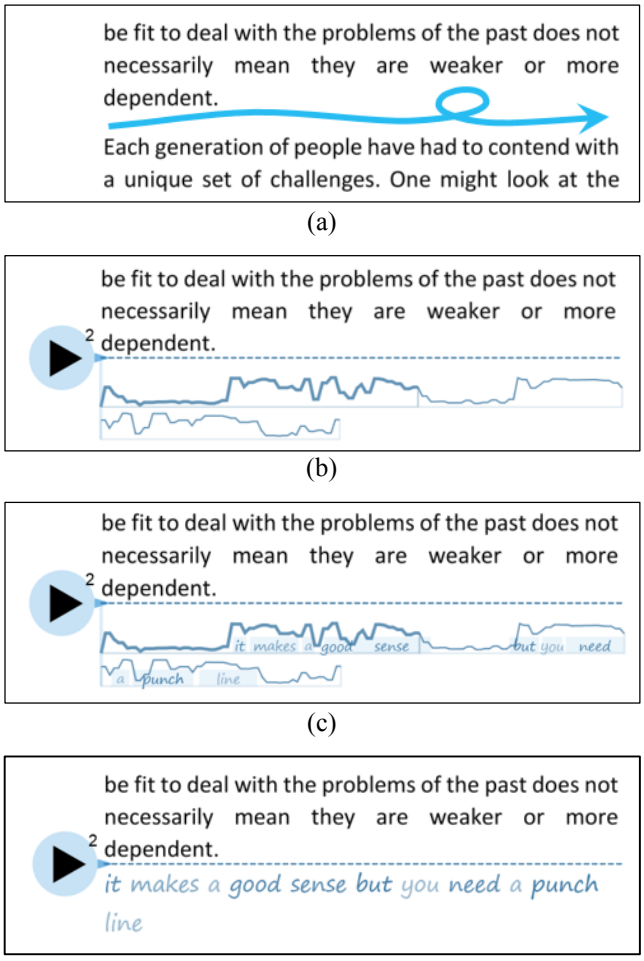

Figure 3. Recording and visualizing speech: (a) Pigtail gesture to begin and anchor recording (b) Waveform during replay (c) Waveform with word overlays (d) Transcription with varying opacities based on recognition confidence.

CREATING ANNOTATIONS IN RICHREVIEW

Ink Annotations

RichReview retains the paper metaphor in which inking can be performed anytime without entering a special mode. When there is insufficient space for writing, TextTearing interactions [26] can be used to create additional writing space in between lines of text. Users can execute this by drawing a horizontal line at the approximate place, followed by a pigtail in the *vertical* direction. All annotations (including the multi-modal ones described later) are anchored to the nearest line of text, graphic, or expansion region on the page (Figure 2). When the position of these elements shift in response to re-layout, the anchored annotation also moves. Annotations created in this way can also be collapsed so that the original layout of the document is preserved.

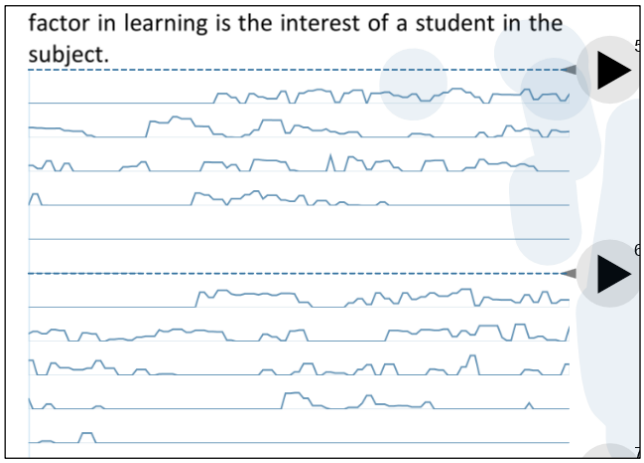

Figure 4. List of recording structures voice annotation by topic (footage from the real-user data, P7).

Multi-Modal Annotations

RichReview's multi-modal annotation recordings capture voice in conjunction with ink and pointing gestures. RichReview requires users to explicitly start and stop the recordings to dispel privacy concerns associated with a system that is always-on. A recording session is started using an underline followed by a pigtail that extends in the *horizontal* direction (Figure 3 (a)). This gesture was selected due to its similarity to the gesture for TextTearing in order to reinforce the idea that annotation activities commence with an underline followed by a pigtail. The location of the underline specifies the anchor point of the annotation and creates a small playback control icon in the margins at the same vertical position on the page. The icon doubles as a marking menu containing commands for working with the annotation.

Capturing Voice

When the recording session begins, a small amount of extra space is inserted between the lines of text where the initial underline gesture is drawn. Inside this space, a waveform representation of the captured audio grows from left to right. Upon reaching the end of the line, the space expands slightly downward and the waveform continues into the new space.

RichReview also includes features to help users add structure to their speech annotations. Similar to Audio Notebook [21], users can structure their voice annotations by creating time-indexed ink notes that the recipient can later use to jump into import parts of the annotation. Also, performing an annotation creation gesture while a recording session is active ends the active recording and immediately starts a new one. This is useful when the annotation moves to a different topic; the interaction saves the user from the interruption incurred from stopping the recording and creating a new one. Performing the annotation creation gesture over an existing annotation appends a new recording to the existing one (Figure 4).

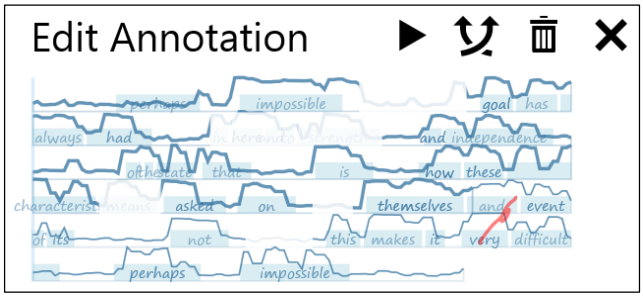

Figure 5. Voice editing user interface presented with waveform. Editing symbol in red deletes a word from the voice recording.

Capturing Pointing Gesture

Pointing at a location in a document is a fast and lightweight way of supporting discussion with reference to specific parts of a document [1]. RichReview provides the *Spotlight* interaction to reproduce this capability. With the Spotlight interaction, hovering the pen over the page while recording creates a circular translucent region at the pen's position (Figure 1). When recording ends, translucent trails of where the Spotlight has been are shown on the document.

Another way that a user can communicate the location of the region of interest is through the creator's viewpoint (i.e. what the creator was looking at during the recording). Similar cues are used in F2F collaboration by observing a collaborator's gaze. To convey this information to recipients, RichReview records viewpoint adjustment operations such as panning and pinch-to-zoom gestures for later playback.

Audio Post-Processing

When a recording is complete, the captured audio is passed to an automatic speech recognizer running in the background. When the transcription is complete, the words in the transcript are shown over the portion of the waveforms corresponding to when they were spoken (Figure 3 (c)). Displaying words over the existing waveform maintains visual continuity with unadorned waveform representation. However, when improved readability of the transcript is desirable, users can display the transcription on its own (Figure 3 (d)) by selecting an option from the marking menu.

The transcribed audio can be used to trim or tidy up the audio in the audio editing tool (Figure 5). In the editing tool, crossing through words or portions of the waveform or transcript grays out those sections of the recording and removes them from the recording. Crossing through a deleted section reverses the deletion. Edits made with the tool are automatically snapped to word boundaries so that the result does not slice the audio in the middle of a word.

CONSUMING ANNOTATIONS IN RICHREVIEW

The ink, gesture, and audio (and transcript) within a recording share the same timebase. RichReview leverages

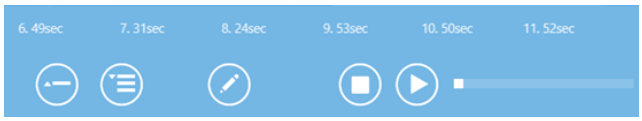

Figure 6. Recording list and media control. A list of annotations, sorted in order of creation, runs across the top. Buttons are used to collapse, expand, edit, stop and play annotations, respectively.

time synchronization to provide a rich rendering of the way the original annotation was created and lets users quickly jump to a specific parts of the annotation stream.

Basic Playback

Basic access to annotation recordings is through the play icon at the annotation anchor or media control at the bottom of the screen (Figure 6). During playback, ink is rendered in a grayed out form if playback has not reached the point where it was created (Figure 7) and then drawn with a colored stroke afterwards. Spotlight traces are rendered as an animated, translucent circle.

Enhanced Navigation of Rich Annotations

Although the basic playback controls are sufficient for the linear consumption of annotation content, they can be inadequate for random access. For example, users may wish to skim through annotations or visit a specific part of an annotation. RichReview offers several features that support these more complex navigation tasks.

Cross-Modal Indexing

For example, users can tap on a point in the waveform or transcript to skip to the corresponding point in the annotation recording. The waveform can be useful for finding gaps in the audio, which often delimit sections within an annotation stream. For finer-grained navigation, the transcript (Figure 3 (d)) can be used to visit parts of an annotation based on words of interest. The need for random-access to audio is critical in light of the fact that speech-to-text technology can still be quite error-prone; generally it is not possible to use the transcript on its own to consume speech content. Therefore, it is imperative that users have a way to quickly jump to and listen to the actual audio.

Ink strokes and Spotlight trails can similarly be used to index into an annotation. One important design decision we made was to show the entirety of the ink and Spotlight traces at all times, so that they can be promptly accessed when needed. On the one hand, this choice does not preserve the exact appearance of the page during annotation creation. However, we believed that giving the ability to skip forwards into an annotation outweighed this concern. We distinguish between strokes that have been made and those that have yet to appear by rendering strokes in different colors.

We also explored other ways of leveraging the links between different modalities that were not as useful. For

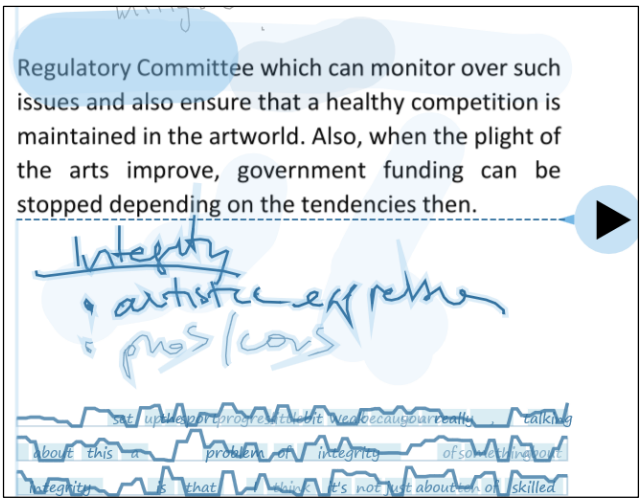

Figure 7. Spotlight trails along with dynamic ink. Recorded strokes are dynamically replayed as playback advances. Grayed out strokes will come in the future. The speech annotation here is structured by writing keywords while speaking.

example, in early prototypes, we highlighted portions of the waveform if they corresponded to times when inking or Spotlight was active. We found that these highlights were not very useful because the highlights provided few cues about the specific objects on the page to which they referred.

Viewpoint Control Options

Users can control whether their viewpoint is locked to match the creator's viewpoint during playback. When locked, the recipient sees what the creator sees. Otherwise, the extent of the creator's viewport is shown as a box.

Thumbnail View

Finally, RichReview provides a high-level overview of all annotations using a space-filling thumbnail (SFT) [5] layout. Pinching more than four fingers zooms the document view out to SFT view. The overview gives a sense of the busiest pages and paragraphs and annotations can be quickly accessed by tapping on one of the page thumbnails.

RESPONDING TO ANNOTATIONS

Given the iterative nature of collaborative writing tasks, RichReview provides collaborative annotation features that allow users to respond to existing annotations made by peers. These features help "close the loop" when people collaborate on a document. In RichReview, annotation entities, such as waveforms, ink, playback controls, and Spotlight traces are color-coded by user identity. However, tracking and showing user identity is only a small part of providing multi-user support.

RichReview differs from other collaborative annotation systems in that it is possible to respond to an annotation using a different modality. The way RichReview enables

Figure 8. Red user inserted a voice annotation in the middle of existing Blue user's voice transcript (footage from the real-user data, P7). Red user's Spotlight is anchored on the transcript.

this is to treat ink and audio annotations in the same way as the underlying body text of the document. There are two benefits of this design decision.

First, mark-up operations that could be applied to the original document can also be applied to annotation entities. For example, ink can be used to circle portions of a waveform (Figure 9) and the Spotlight can be used to bring attention to a part of the transcript (Figure 8).

Second, treating annotations like the underlying text provides multi-modal support for discussions between collaborators conducted through threaded comments. For instance, users can create an expansion space in the middle of an audio waveform and insert a comment using ink (Figure 9). Inserting a voice comment under an existing inking space is also possible (Figure 10, top).

In some cases, interleaving annotations with the body text can break the flow of reading due to annotations having a large visual footprint. In these situations, RichReview allows users to collapse or expand annotations. An option in the marking menu accessed through the play icon allows this to be done on a per-annotation basis. Collapse-all and expand-all buttons in the bottom toolbar can also be used to switch back and forth between the original document layout and the fluid layout.

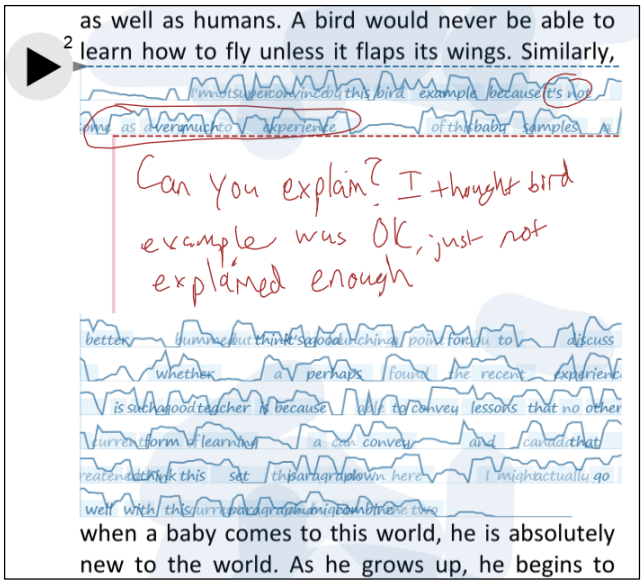

Figure 9. Circling on waveform to designate a part of the voice (footage from the real-user data, P10)

These multi-user features that allow users to converse and engage in discussion through annotations are illustrative of how our initial design goals of interactional consistency and flexibility pervade the entirety of our system.

IMPLEMENTATION

RichReview is a C++ Windows Store app. We use the MuPDF library to extract the location of graphical elements for anchoring annotations as well as to rasterize the PDF document itself. The graphics are performed using a mix of DirectX (Direct3D and Direct2D) and XAML UI elements. Speech recognition is performed using the built in Speech Services in Microsoft Windows. We primarily tested RichReview on a Lenovo Thinkpad Tablet 2 tablet, which has a 10.1'' display screen, an inductive pen digitizer and 5-points multi-touch. However, we have also used RichReview successfully on a range of other pen-enabled tablet devices. We generally prefer to use a Bluetooth headset supporting 16 KHz wideband audio for the audio capture since writing can introduce undesirable rattling noises when recording using the built-in microphone on the tablet.

EVALUATION

We wished to determine whether users could successfully employ the features of RichReview to make comments on a document. Moreover, we wanted to investigate how these features were actually used. Therefore, we conducted a qualitative, formative user study using our prototype system.

Study Design

To prompt realistic feedback from participants, we designed a task based on a representative classroom situation. We asked the participants to assume that they were working as

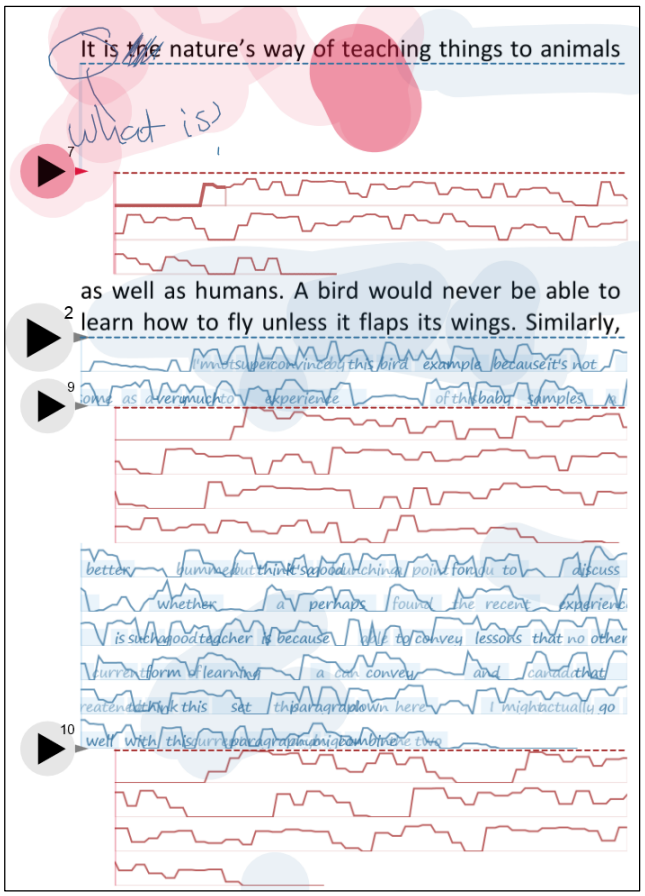

Figure 10. New annotations can be inserted under existing expansion space or in the middle of existing waveform (footage from the real-user data, P2); here the red user has replied to the blue user's existing voice comment.

a teaching assistant (TA) for an introductory, undergraduate writing class, commenting on a student's essay assignment.

Operating under the assumption that the social relationship between collaborators may influence annotation behavior, we told half of our participants that their annotations would be shared with the instructor of the course. We told the other half that their comments would be shared with a peer grader. Note, the point of splitting our participants into two groups was to expand our coverage of possible usage scenarios rather than to carry out a controlled comparison or test a hypothesis.

Procedure

Our study consisted of a practice session that helped participants familiarize themselves with RichReview interactions followed by an open-ended session with full-fledged tasks. During the practice session, we introduced each feature of the system to each participant by demonstrating a specific use case, and then letting them try out the features first hand. For example, the Spotlight feature was introduced in the context of referring to a location of document while recording. Next, we gave each participant a set of practice tasks to carry out. We ended the practice session by having participants use the audio editing functionality: participants were asked to pick one of the most problematic recordings amongst the ones they had created, and edit it so that could be better understood. The practice session took less than 40 minutes.

The open-ended annotation session consisted of two parts. Participants started with the production part where they gave constructive feedback on an essay for 15 minutes. They were asked to create at least 6 comments on an essay concerning any kind of writing issue. Then, participants performed the consumption part, which looked at how rich annotations are consumed and discussed. We asked participants to listen to a set of pre-made annotations, and then respond with constructive feedback for 20 minutes. In both tasks, participants were not required to use any specific interaction techniques. We concluded the evaluation with a 10 minute of semi-structured interview session. In total, the study tasks required approximately 90 minutes to complete.

Materials

The materials used in the study were sample essays to the "Analyze an Issue" portion of the Graduate Record Exam used in the United States for entrance to graduate school. These essays tend to be around 500 to 650 words long (approximately 1 page) and of moderate writing quality. We picked two different essays, counter balanced across participants to rule out text dependencies. Each participant used the same essay across the production and consumption tasks, in order to save reading time.

The pre-made annotations in the consumption task were composed of various discussion topics and consisted of diverse modality combinations. These were based on real annotation data captured in earlier pilot tests of our system.

Participants

We recruited 12 participants from student mailing lists at a large university. The average age of our participants was 21.3 years old. All but one participant was a native English speaker and all had experience with collaborative writing tools. The most frequent discussion channel for their writing tasks was e-mail (5.33 hours/week), followed by F2F meeting (3.67 hours/week). The participants received $15 for taking part in the study.

RESULTS

Broadly speaking, the results of our study demonstrated that participants could successfully employ RichReview to communicate complex ideas about a document. Voice and Spotlight introduced additional expressiveness and efficiency on top of the communication capability of the legacy collaboration tools that were based on textual means. Moreover, cross-modal commentaries and indexing features helped users achieve fluid modality combination and lightweight annotation access.

Experience of Using RichReview

Users compared the overall experience of using RichReview to having a collaborator virtually present. P3, when talking about the Spotlight feature remarked, "It (Spotlight) was like I was talking to someone in person when I point to an area." Further evidence of this was the fact that in annotations, participants often used the pronoun "you" reinforcing the sense that they were talking *through* the computer rather than talking *to* the computer.

Annotation Production

Ink Annotations

Direct inking without additional forms of recording was the most widely used form of annotation. All of the participants used ink for simple mark-up such as circling, underlining, question marks, brackets, proofreading symbols, connecting lines, and personal notes that they wanted to revisit. This highlights the importance of static inking for lightweight interaction.

Voice Annotations

Voice recording was used by all participants when they wanted to make a comment that was longer or more detailed. Participants praised voice's speed and expressiveness. As P4 said, "Now I can hear someone's voice and understand completely what they're trying to say versus just seeing their note and trying to interpret." All participants except P1 and P8 used voice in conjunction with writing and the Spotlight. These two users used voice on its own.

Participants structured their voice annotation in many different ways. One way was to write ink as a visual guidance for the verbal description. For example, participants first made underlines on body-texts or wrote down key points in white space while reading, and then used voice or Spotlight to refer to these points while recording. Another way was to write keywords *during* the recording session to allow their hypothetical recipient navigate to a certain topic in the recording by tapping on a corresponding keyword. Additionally, P3, P5, P8, P9, P11 used the feature where additional annotations could be appended onto an existing one; they used this feature when they had multiple points to talk about in a single paragraph.

Another interesting observation was that participants tended to use voice when they disagreed with an idea and ink when they agreed with it. The reason for this was because when they disagreed, they would use voice to provide a detailed explanation for their disagreement (P6, P7, P10, and P11). This suggests that support for voice annotations could be the best method for achieving group maintenance goals [2].

Spotlight Annotations

Participants frequently used Spotlight when speaking. The Spotlight feature was used to refer not just to the underlying text, but also to other people's ink marks and a part of waveform or transcripts. Annotations about paragraph structure or logical inconsistencies were often accompanied by the spatial cues that Spotlight conveyed. Overall, the feature was seen to be a powerful deictic tool: As P3 said, "I liked that feature (Spotlight) a lot, because it could direct somebody while recording to the specific spot that they are talking about."

Participants did raise some implementation issues, however. P1, P9, and P10 reported that Spotlight was sometimes recorded inadvertently when the pen hovered over the screen for other reasons. P1 complained that the Spotlight trail was too thick to point to a specific line of text or word. In future iterations of the system, these issues could be addressed by filtering out spurious hovering gestures and by changing the blob size.

Socially-Driven Modality Choices

Besides the annotation content and purpose, we found that the social factors affected which communication modalities participants felt comfortable employing. For example, P2 and P8 regarded simple scribbles, such as circling or checkmarks, as an impolite or casual form of annotation, choosing instead to leave voice comments. Similarly, P1 and P8 thought that writing a complete message was more polite than voice. In this case, they claimed written comments were easy to understand and that voice is a "lazy form (P8)". While we cannot draw any conclusions on this basis, this does raise some interesting research questions for future research on annotation and target users.

Editing Audio

Most participants found the voice editing interface easy to use and efficient especially for removing long pauses or utterances such as "Um". This suggests that automatic detection and trimming of the pauses might be useful. However, considering that some users depend on long pauses as a navigation cue for time indexing operations, removing these also might be problematic. Ultimately the long term usefulness of these features would need to be assessed in real practice.

Annotations Use and Organization: Annotation Positioning

Most annotations were placed immediately under the relevant text. If it was about a sentence or keyword, participants would position the annotation under a line in the middle of a paragraph. However, P1 and P8 were reluctant to break the paragraph structure, instead placing the recording below the paragraph and making a reference to the targets using inking or the Spotlight. Some annotations do not have an obvious anchor point such as when they are about multiple paragraphs or global writing issues. In these cases, participants usually placed the recordings below the end of the right column, making them hard to distinguish from those relating to the last paragraph. This observation suggests that a distinct space to anchor meta-commentary [27], possibly close to the bottom of the page, might be useful.

Consuming Annotations

Navigation via the visual representation of the audio was actively used to jump into or revisit a voice annotation. Participants were able to use the waveform as a navigation cue effectively when there were salient features they could focus on, such stretches of silence. Other participants sometimes used the words in the transcript as a way to navigate (P7, P9, and P12).

However, most of participants preferred using the waveform over transcription because of the detrimental effect of transcription errors. For instance, P11 recounted one instance where the phrase "kind of this" was recognized as "Kennedy." Although P11 was aware it was a transcription error, the participant found it very hard to ignore.

Participants found Spotlight trails useful for getting a sense of what an annotation was about. However, participants did not use ink or Spotlight trails to index into annotations. We believe there were two reasons for this: First, Spotlight traces became too cluttered; second, users were not familiar enough with the style of annotation to know how the ink or Spotlight element was structured in relation to the audio.

Creating Responses and Discussion Threads

The ability to create rich annotations about existing ones was well adopted by all participants. In general, most responses tended to be placed immediately below the annotation to which it responded. For instance, P2 inserted a voice annotation in the middle of an existing audio stream (Figure 10). P10 made a written reply below a part of existing spoken annotation making use of cross-modal commentary features (Figure 9). They used cross-modal mark-up features for referring to parts of spoken or written annotations. For instance, P7 used the Spotlight to point to the visual representation of the audio (Figure 8), and P10 marked it up with ink (Figure 9).

DISCUSSION AND FUTURE WORK

Taken together, the results of our study show that supporting multiple modalities within an annotation tool can indeed enable creative and expressive ways for users to comment on and discuss documents with others. Our participants sometimes drew on one modality at a time, and other times intertwined their use of ink, speech and gesture. Indeed, the majority of our participants ended up using a mixture of modalities for a variety of purposes. Further, the reasons participants gave for employing specific modalities, such as speech being a better mode for conveying explanations when there was disagreement, were in line with what previous work has identified as specific strengths of non-textual comments and annotations.

Efficacy of RichReview

Of course, determining whether RichReview is in fact *useful* rather than simply *usable* will depend on its deployment in a naturalistic setting, assessed alongside other collaborative tools. In order to prepare the RichReview system for deployment, a number of issues must be addressed. First, ASR accuracy needs to be improved. One possible solution of this issue is to offer a tiered approach using a more sophisticated system in the cloud or even transcribed via crowd sourcing if extra accuracy is desired. Second, the visual clutter from Spotlight traces can be resolved by employing a filtering feature that renders a part of relevant data only. Finally, tighter integration of RichReview and authoring environments (e.g. LaTeX, Microsoft Word) to support the full write-revise-review workflow is critical.

Additional Application Domains

Looking further, we believe that the ideas behind RichReview can be extended to other application domains. For example, making the system practical for larger groups can be a useful extension. One possible application of the RichReview system at scale is to support discussion in large distributed courses such as MOOCs. There are challenges to tackle in doing this, however. For instance, the current presentation of Spotlight turned out to be cluttered when there are many annotations on the page. Another issue was that participants were concerned about fluid document layout and that introducing too many gaps for annotations would adversely affect the readability of the main body text.

RichReview can also be used to discuss other types of documents. For instance, engineering documents like electronic schematics or architectural drawings, or source code are often created through collaborative processes. These types of documents can similarly benefit from the expressiveness that the RichReview system brings. The general obstacles that needs to be resolved when extending RichReview to these other domains is to make it compatible with the unique discussion requirements and structure of the documents used in these other domains.

CONCLUSION

We have shown a system for rich annotation that offers users flexibility, versatility and expressivity, not only in terms of the creation of annotations, but also in consuming them. The core innovation is our design decisions to support commentaries and indexing across different annotation modalities. The fact that early users felt that they could communicate through and around the document using this tool, even achieving a sense of remote presence, shows promise. At the same time, the design, implementation and evaluation of the system so far has raised a number of interesting research questions for further work including improving navigation of dynamic annotations, making the system scalable, and expanding application areas. The possibilities opened up through multi-modal annotation systems are undoubtedly both diverse and compelling.

ACKNOWLEDGEMENTS

Yoon gratefully acknowledges support from the Kwanjeong Educational Foundation. This work was supported in part by a gift from FXPAL. We would also like to thank Robert Corish and Bill Buxton for their design input.

REFERENCES

1. Bickmore, T., Pfeifer, L., and Yin, L. The role of gesture in document explanation by embodied conversational agents. *International Journal of Semantic Computing*, (2008).
2. Birnholtz, J., Steinhardt, S., and Pavese, A. Write here, write now! *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press (2013), 961.
3. Chalfonte, B.L., Fish, R.S., and Kraut, R.E. Expressive richness. *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*, ACM Press (1991), 21–26.
4. Clark, H. and Brennan, S. Grounding in communication. *Perspectives on socially shared cognition*, (1991).
5. Cockburn, A., Gutwin, C., and Alexander, J. Faster document navigation with space-filling thumbnails. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, (2006).
6. Grudin, J. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. *Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88*, ACM Press (1988), 85–93.
7. Hardock, G., Kurtenbach, G., and Buxton, W. A marking based interface for collaborative writing. *Proceedings of the 6th annual ACM symposium on User interface software and technology - UIST '93*, ACM Press (1993), 259–266.
8. Kraut, R., Galegher, J., Fish, R., and Chalfonte, B. Task Requirements and Media Choice in Collaborative Writing. *Human-Computer Interaction 7*, 4 (1992), 375–407.
9. Kraut, R.E., Gergle, D., and Fussell, S.R. The use of visual information in shared visual spaces. *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02*, ACM Press (2002), 31.
10. Levine, S. and Ehrlich, S. The Freestyle System. *Human-Machine Interactive Systems*, (1991).
11. Marshall, C.C. and Brush, A.J.B. Exploring the relationship between personal and public annotations. 2004, 349–357.
12. Marshall, C.C., Price, M.N., Golovchinsky, G., and Schilit, B.N. Collaborating over portable reading appliances. *Personal Technologies 3*, 1-2 (1999), 43.
13. McNeill, D. *Language and gesture*. 2000.
14. Monserrat, T.-J.K.P., Zhao, S., McGee, K., and Pandey, A.V. NoteVideo. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press (2013), 1139.
15. Morris;, M.R., Brush;, A.J.B., and Meyers, B.R. Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. *Horizontal Interactive Human-Computer System*, (2007), 79 – 86.
16. Neuwirth, C.M., Chandhok, R., Charney, D., Wojahn, P., and Kim, L. Distributed collaborative writing. *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*, ACM Press (1994), 51–57.
17. Nicholson, R.T. Usage patterns in an integrated voice and data communications system. *ACM Transactions on Information Systems 3*, 3 (1985), 307–314.
18. Noël, S. and Robert, J.-M. Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like? *Computer Supported Cooperative Work (CSCW) 13*, 1 (2004), 63–89.
19. Oviatt, S. Ten myths of multimodal interaction. *Communications of the ACM 42*, 11 (1999), 74–81.
20. Sellen, A. and Harper, R. The Myth of the Paperless Office. *MIT Press*, .
21. Stifelman, L., Arons, B., and Schmandt, C. The audio notebook: paper and pen interaction with structured speech. *of the SIGCHI conference on*, (2001), 182–189.
22. Tsang, M. and Fitzmaurice, G. Boom chameleon: simultaneous capture of 3D viewpoint, voice and gesture annotations on a spatially-aware display. *Proceedings of the 15th annual ACM symposium on User interface software and technology*, (2002), 111–120.
23. Whittaker, S., Hirschberg, J., Amento, B., et al. SCANMail. *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, ACM Press (2002), 275.
24. Whittaker, S., Hyland, P., and Wiley, M. FILOCHAT. *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*, ACM Press (1994), 271.
25. Wilcox, L.D., Schilit, B.N., and Sawhney, N. Dynomite. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97*, ACM Press (1997), 186–193.
26. Yoon, D., Chen, N., and Guimbretière, F. TextTearing. *Proceedings of the 26th annual ACM symposium on User interface software and technology - UIST '13*, ACM Press (2013), 107–112.
27. Zheng, Q., Booth, K., and McGrenere, J. Co-authoring with structured annotations. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, ACM Press (2006), 131.