RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion

Dongwook Yoon[1], Nicholas Chen[2], Bernie Randles[1], Amy Cheatle[1], Corinna E. Löckenhoff[1], Steven J. Jackson[1], Abigail Sellen[2], and François Guimbretière[1]

[1]Cornell University,
Ithaca, NY
{dy252, bmb256, ac2288, cel72, sjj54, fvg3}
@cornell.edu

[2]Microsoft Research,
Cambridge, UK
{nchen, asellen}@microsoft.com

ABSTRACT

New multi-modal annotation tools hold the promise of bringing the benefits of face-to-face contact to remote, asynchronous interactions. One such system, RichReview++, incorporates new techniques to improve access to the embedded multimedia commentary and allows users to annotate with new modalities, like deictic gestures. We conducted a series of field deployments of RichReview++ to characterize how these features benefit students using them for activities in the university classroom. Our first deployment investigated the use of multi-modal annotations as a way for instructors to provide feedback on student term papers. Our second deployment used annotations to support peer discussion about assigned readings in a graduate-level course. We found that presenting voice comments as interactive waveforms seems to facilitate students' consumption of the instructor's voice comments. We also found that gestural annotations clarify voice and give annotators a quick and lightweight way to alter the scope of their voice comments. Based on these results, we suggest ways to best leverage multi-modal annotation tools in education environments.

Author Keywords

Multi-modal annotation; collaborative annotation; anchored comment; instructor feedback; online peer-discussion.

ACM Classification Keywords

H.5.3 Group and Organization Interfaces: Collaborative computing; H.5.2 User Interfaces: Interaction styles; H.5.1 Multimedia Information Systems: Audio input/output.

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA

ACM 978-1-4503-3592-8/16/02…$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819951

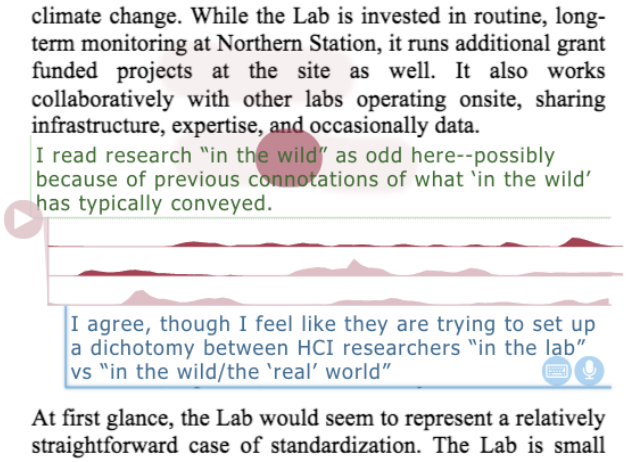

Figure 1. RichReview++ screen shot showing a thread of multi-modal annotations containing text, voice, and gestures.

INTRODUCTION

In modern classrooms, typewritten annotations over digitally shared texts have become widely used for document-centric classroom activities (e.g., instructors giving feedback to students' assignments [35] or student groups discussing about a course material [2, 40]) because they afford a flexible workflow that gives students more time to digest comments at their own pace. However, purely textual annotations cannot deliver nuanced and complicated ideas as effectively as comments delivered in person. In response, researchers have studied different types of multi-modal solutions that leverage expressivity of voice [21, 36], sometimes combined with fluid free-form digital inking [20, 33, 34] or visually rich video recording [15, 28]. In online education, such systems provide students with engaging and effective learning experiences by combining the expressivity and clarity of multi-modal communication [5, 28] with the relaxed workflow of asynchronous interaction [27, 36].

Despite their benefits, multi-modal annotations suffer from some of the disadvantages of multimedia content: First, the encoded multimedia stream, especially voice recorded as audio, is more difficult to access than textual annotation and makes consumption difficult [8]. Second, in transferring the

rich modalities of face-to-face communication into the digital space, not all of the interaction modes have received equal attention. For example, pointing gestures that are used to indicate a region of interest in a face-to-face conversation over documents [1] have not been incorporated into multi-modal annotation tools.

We recently presented RichReview, a new multi-modal annotation system designed to address the problems identified above [37]. RichReview employed a novel approach to capturing gestural annotation by tracking pen-hovering movements in tandem with voice and digitizer writing. Voice comments in the system are visualized as a waveform that serves as an interactive navigational cue to facilitate the review and re-review of audio recording.

Although a small-scale formative laboratory study validated the interface concept [26, 37] behind RichReview, the real-world efficacy and implications of the multi-modal features it introduced (e.g., pointing gesture or audio visualization) have not been fully characterized. This paper addresses this gap and describes the deployment of these features in two different educational contexts: (1) to support instructors' delivery of term paper feedback in an advanced undergraduate seminar; and (2) to support reading assignment discussions in a graduate-level course.

This series of evaluations provides answers to the following questions:

- What practical benefits does the multi-modal integration of voice and ink with gesture offer in realistic settings and for typical classroom activities (e.g., paper revision or peer discussion)?
- To what extent can waveform indexing overcome the problem of diminished accessibility in multi-modal annotations? Are there any caveats?
- What are the broader implications of these features for the use of multi-modal annotation tools for tasks in online classrooms? For example, what are the expected limitations and possible workarounds?

The first deployment examined instructor feedback on a writing assignment in an advanced undergraduate class. To provide access across the range of devices employed by this user population (i.e., laptop, desktop, and tablet), we implemented a cross-platform viewer for RichReview comments. Analysis of the online logs and student feedback suggested that the visual waveform indexing successfully supported repetitive replays of voice comments and that gestures facilitated the mapping of voice annotations to relevant sections of the manuscript. We also found that students rated multi-modal feedback favorably and sometimes even considered it as a proper substitute for attending in-person office hours.

To enrich and extend the findings from the first study, we performed a follow-up deployment where the annotation system was used for peer discussion over assigned readings—another instance of document-centric collaboration in a classroom setting. In contrast to the first evaluation where only the instructor created annotations, peer discussions entail multi-way communication among the students creating annotations. To support these different interaction modes, we built an improved system called RichReview++ that offered a set of new discussion support features. Results showed that students used gestural annotation in a variety of fluid and expressive ways to clarify references in voice comments. We also noted challenges to students employing voice annotation in peer discussion and discuss the resulting implications for multi-modal annotation.

Taken together, these findings motivate further inquiry into the augmentation of other types of annotations with deictic referencing, factors affecting the ease with which people are able to produce voice-based annotations, and the use of multi-modal annotations in other types of course materials and educational formats such as online courses.

RELATED WORK

Multi-Modal Document Annotation

Voice has been employed as the central element of most multi-modal systems for its many beneficial characteristics. Spoken annotation possesses high throughput [8], delivers nuances and emotions [3, 36], clarifies the speaker's intention [12], supports a positive perception of the speaker [24], and addresses higher-level concerns such as semantic and structural aspects more effectively than text-based comments [3, 16, 24]. Other systems incorporated additional modalities to augment or complement voice (e.g., inking as a freeform visual aid [20], or video recording to add rich audio-visual multimedia content [28]). To the best of our knowledge, none of these previous studies have incorporated gesture in the context of document annotation. In the present paper, we evaluate how digitally replicated gesture can bring deictic referencing capability of face-to-face communication (e.g., pointing a finger or waving hands over text [1]) into document-centric tasks (e.g., manuscript feedback and discussion) in real classrooms.

Previous studies in HCI have examined the deictic role of gestures in online communications. Boom Chameleon allowed users to highlight a point of a 3D scene using glowing blobs that are recorded in sync with voice [30]. Fussell et al. transmitted and overlaid pen-drawing traces over remote live-video feeds for a deictic purpose [7]. Harrison et al.'s electronic cocktail napkin system could capture a remote user's hand gesture over an upward facing monitor as a video stream recorded by a downward facing camera [10]. Lee and Tatar comparatively evaluated different deictic markups as a visual aid for collaborative Sudoku puzzles [19]. In contrast, our study focuses on an asynchronous communication scenario where the pointing gestures are animated over written documents [37]. As such, our study provides added insights about the way voice

and gesture annotations are recorded and replayed in the context of the underlying text.

Accessing Annotations

A major challenge in annotating a document using non-textual modalities, particularly those that do not have a clear semantic analog on paper (e.g., voice and gesture) is the relative difficulty in accessing these annotations. For instance, Grudin pointed out that voice is a linear medium that does not readily offer quick skimming and browsing of its contents [8]. Multi-modal annotation systems that use voice as a central mode of interaction inevitably inherit this problem, and researchers have explored a number of potential solutions. Auto captioning has been suggested as a way of computationally transforming audio to a more visually consumable format (i.e., text [23, 32]). However, typical transcriptions are error-prone and can require long processing times. Hindus and Schmandt presented a sound editing system based on binary acoustic structuring (silence or speech) [13], but they did not provide in-depth studies of people interacting with that system. In professional audio/video editing software, audio streams are often represented in the form of sound waves that provides visual cues for fine-grained audio editing. Moreover, Neuwirth et al. hypothesized that visual features of an audio waveform can serve as navigation cues for consuming voice comments [24]. However, the utility of using waveforms to navigate annotations has so far been untested. In this paper, we investigate the benefits of navigating multi-modal annotations using audio waveforms.

When audio is recorded in conjunction with other timestamped data, such as digital ink, it is also possible to navigate the audio using the linked data. For example, ink strokes recorded along with voice can be used to skip to the voice stream corresponding to a particular stroke [29, 33, 34]. We tested a variation of this functionality in which gesture traces [37] are used to navigate the annotation.

Multi-modal Communication in Education

There is a long tradition of leveraging multi-modality in various types of educational settings. When providing feedback, instructors may record video screenshots of themselves editing students' papers [5, 28] or provide anchored audio comments over a PDF document [25]. Online discussion forums now support threaded voice comments to facilitate discussions between students [11, 21]. Our study differs from previous work by investigating how students process feedback from an instructor and participate in discussions with their peers given new gestural annotation and waveform indexing features.

The system we tested in our field deployments is based on RichReview [37], a multi-modal annotation system that combines ink, audio, and gestural annotation to provide a face-to-face communication experience for collaborators reviewing and providing feedback on documents. A prior laboratory study [40] suggested that the new interaction features would provide several useful communication

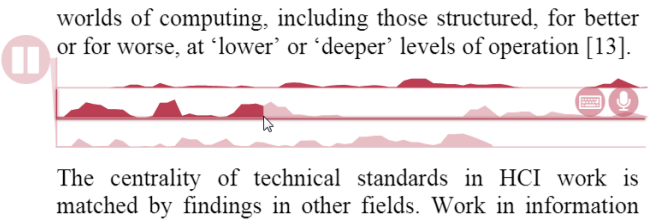

Figure 2. The Waveform Indexing feature. The user can click a waveform to jump through the audio recording.

affordances. The studies in this paper go one step further and test these features against the demands of real educational settings.

THE RICHREVIEW SYSTEM

This paper focuses on a field evaluation of the novel multi-modal annotation features (e.g., gesture and waveform indexing) that were suggested as a part of the RichReview system [37]. We now review two of the key features, gesture recording and time indexing in more detail.

First, RichReview captures gestures via its Spotlight feature. While recording voice, the annotator can hover the pen over the tablet screen to associate page locations with the spoken contents. During annotation playback, the traces of the pen's movements are replayed in the form of a semi-transparent blob moving in sync with voice. In the prior lab study, the Spotlight feature was found to play a deictic role similarly to face-to-face gestures [37].

Second, RichReview offers versatile time indexing features through multiple modalities to enhance accessibility of recorded multimedia comments. It presents voice annotation as lines of sound waveforms, so that users can visually spot regions of interest in the recording and then click on them to navigate through the recording (Figure 2). We made these waveform representations more readily accessible by interleaving them into vertically expanded spaces in between lines of text (Figure 1) [38], rather than abstracting them into audio icons. In addition, the gestural comments are also rendered as semi-transparent traces, so that the user can click a part of a trace to start a replay from the moment when the gesture was created at that position.

PROVIDING FEEDBACK ON STUDENT PAPERS

Our first deployment examined the utility of the new multi-modal annotation features in the context of an instructor providing feedback on a term paper. The main focus in this study was students' perceptions of the effectiveness of the feedback, but we also observed the emotional responses and interpersonal dynamics that emerged from the use of the tool.

Web-based Viewer for RichReview

In the original RichReview system, making full use of the capabilities of system required specialized hardware (Windows tablets). While it would have been possible to provide each student in a small class with such a device, it

was an inherently impractical and non-scalable solution considering deployment costs.

Our solution was to create a new, web-based viewer for documents annotated using RichReview. The instructor used a tablet computer to create her comments (leveraging the full feature set of the system) and uploaded the annotated document to the web. Students could then access instructor comments on their device of choice by visiting a URL the instructor provided. Figure 2 is a screen capture from the web viewer. Our implementation was hosted on a cloud service (Microsoft Azure) and satisfied the security and accessibility requirements of the Family Educational Rights and Privacy Act (FERPA) in the United States for educational electronic data storage and handling.

Deployment Procedure

The system described above was deployed in an undergraduate level Human Development course in the fall 2014 semester at our university. The term-paper assignment involved writing a proposal for a life-span developmental research project. The paper was to include prior literature, research questions, measures, methods, and plans for analysis. Recommended paper length was under 20 pages, and students worked individually. Students in the class submitted paper drafts as PDF files over email. The instructor then commented on each paper with RichReview on a tablet using digital ink, audio, and gestures. Students received feedback 10 days before the final version of the paper was due.

The instructor, who was a member of our research team, spent an average of 20 minutes commenting on each paper, but she noted that weaker papers appeared to take relatively more time than the others. She took advantage of the full range of affordances, often combining multiple modalities (e.g., drawing a flow chart and 'walking' students through the chart by speaking and pointing over the drawing). Also, the instructor attempted to communicate emotion by expressing enthusiasm or encouragement. The absence of a voice editing feature required the instructor to start recordings from scratch if she felt the comment was unclear or got side-tracked. Although she generally regarded this as a shortcoming of the system, she noted that the re-recorded comments were often more cohesive and of better quality.

Participants

Participation was voluntary and students could choose between receiving handwritten notes on a printout or RichReview-based comments. Exit surveys were also optional and the instructor did not have access to student responses. The class was composed of one instructor and 17 students, 16 of which participated in the study and 13 of which answered the exit survey. The participating students were nearly all female undergraduates (mean age: 21.0, SD = 0.44); one student was male, and one was a graduate student. All students used laptops to view the feedback provided.

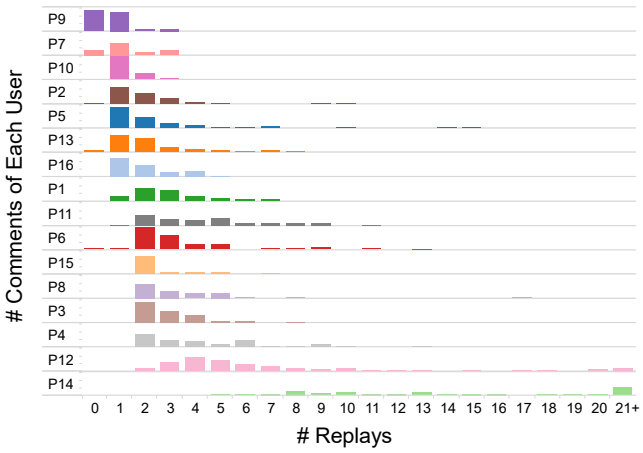

Figure 3. Each row indicates a distribution of the number of comments (y-axis, ranges from 0 to 32, un-normalized) per the number of replays for each user (x-axis, cut off at 21). The rows were sorted in the order of momentum position.

Measures

We monitored and logged students' online activities such as page navigation and voice playback. The exit survey collected subjective ratings on a 5-point Likert Scale. Students were also asked to rate their preference for the different ways of receiving feedback from the instructor. Since students had received paper-based handwritten feedback from the same instructor earlier during the semester, they had a baseline against which to compare RichReview-based feedback. Lastly, the questionnaire also featured free-response questions about the usability of various system features (See the auxiliary material for the copy of the questionnaire).

Results

Essays ranged from 10 to 22 pages in length (M = 15.7 pages, SD = 2.73). The instructor made digital ink markups for typographical edits as well as voice comments for detailed commentaries. On average, the instructor made 51.4 voice comments (SD = 11.7) per student essay with a mean length of 14.9 sec (SD = 15.1).

Students reported that the system was easy to learn and effective. One commented that "This was by far the best experience I've had while revising a paper (P10)". Students reported a willingness toward continued use and recommended its use to peers. To quote P6, "Would definitely use again and would recommend this to others!"

Benefits of Multi-Modal Indexing for Consuming Comments

One recurring theme in the qualitative feedback we gathered was that easy-to-use audio replay was helpful for consuming the recorded comments. We analyzed logs of online activities to take a deeper look into replay patterns. The results showed that audio re-listening was very popular. Most (73.1%) of the voice comments were replayed more than twice, and a few (6.44%) were replayed more than 10 times. As shown in the replay count histogram of Figure 3, more than half of the users (9)

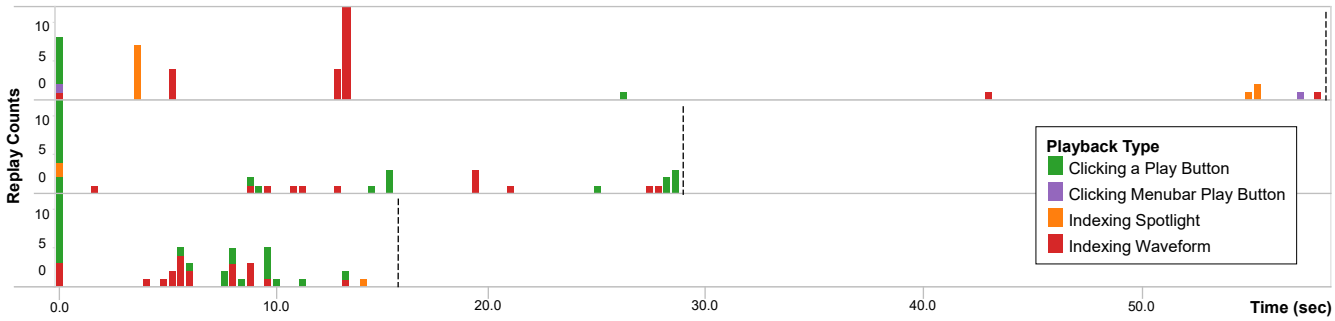

Figure 4. Navigation patterns observed from a user listening three different recorded multi-modal annotations. The x-axis is time separated into 0.4 sec interval (ranges from 0 to 58.2 sec), and the y axis is a number of playback hit in each time interval.

played the majority of the comments they received more than twice. Moreover, the histogram for every student featured a few comments that were replayed many times. P5, for instance, predominantly listened to comments once, but had two comments that were replayed more than 14 times.

Analysis of data revealed that the students exploited various types of indexing features for re-listening, but made particularly heavy use of audio waveforms. Across all playback operations, the play button was used 81.7% of the time, waveforms were used 12.4% of the time, and Spotlight traces were used 4.6% of the time. We also found that the number of replays started by clicking on the waveform for a given voice comment was significantly correlated with the comment's length ($N=823$, $r=0.22$, $p<.001$). A similar pattern did not emerge for the play button. This suggests that for long comments, students perceived waveform-based indexing as useful in jumping straight to relevant passages in the audio.

To further illustrate this point, we show a representative user's (P14) pattern of accessing a set of annotations (Figure 4). With a lengthy comment like comment #1 (58.5 sec), P14 jumped into a specific point multiple times using Spotlight indexing (3.6 sec, 7 times, the orange bar) and Waveform indexing (13.4 sec, 16 times, the red bars). The clusters of bars with both green and red in comments #2 and #3 show that P14 triggered playback using a mix of controls. The tall green bars on the left side of Figure 4 indicate that the comments are almost exclusively played first using the playback button. Finally, the green bars in the middle of recording show that P14 employed a stuttering playback pattern in which P14 repeatedly paused and started the comments in the middle of the audio stream. Similar patterns were observed in the vast majority of the other users (P1, 2, 4, 5, 6, 11, 12, and 13).

These quantitative findings were further corroborated during our survey. Participants reported that they used the waveforms to "skip over the parts you adjusted already (P6)", "listen to specific parts over again (P3)", or "repeat a missed word (P4)". Participants ratings also suggest that the Waveform indexing feature was very helpful for understanding audio comments ($M = 4.63$, $SD = 0.52$, 8 responses). On the other hand, the gesture-based indexing using Spotlight traces was not as popular as the waveform indexing. Only 9 participants noticed the feature and used it. Participants reported that the feature's lack of discoverability was the barrier: "I didn't know you could do that (P8)".

Multi-modal Annotations vs. Longhand Comments

In the exit survey, we asked participants to compare their preferences between RichReview and traditional feedback methods for different types of comments. They preferred the multi-modal comments over written comments for receiving feedback about writing issues related to factual content ($M = 4.27$, $SD = 0.79$) and structure ($M = 4.55$, $SD = 0.52$). There was no significant preference difference between the two methods for comments pertaining to grammatical errors and typos ($M = 3.09$, $SD = 1.45$). This result echoes and confirms previous research on voice-only annotation [3, 16, 24], which found that spoken comments were preferred over text when describing higher-level (structural, or semantic) writing issues in comparison with local problems.

Multi-Modal Annotations vs. Office Hours

Surprisingly, the majority (11) of participants preferred RichReview annotated documents over office-hour meetings ($M = 3.91$, $SD = 1.22$), and believed that they offered an acceptable substitute for in-person meetings ($M = 3.91$, $SD = 1.04$). Qualitative comments offered two explanations for this result. First, students wanted to "incorporate all the comments" and make sure that they were "doing everything that the instructor suggested" (P4). RichReview was useful because recorded comments were hard to miss and could be addressed one at a time. Also, the rapid stream of feedback received when meeting in person made students worry about whether they were missing or misunderstanding the instructor's comments. In contrast, the recorded comments could be replayed multiple times if they were not clear. To quote P4, "I can listen to everything multiple times if I didn't get it which also made it less intimidating". On the other hand, a few students (2 of 13) thought that in-person meetings offered a more immediate interactive dialogue which RichReview's asynchronous interaction did not offer.

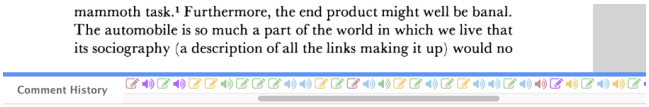

Figure 5. Comment history feature. A user can click one of the chronologically sorted links to existing comments to jump to the relevant page and the selected comment is highlighted.

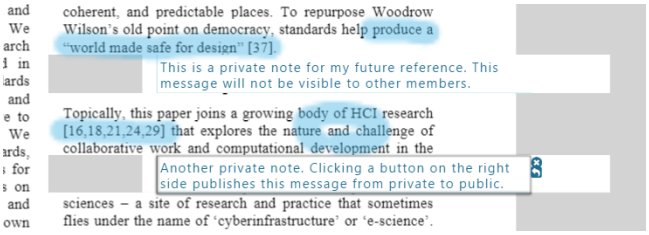

Figure 6. Private Notes and Highlights. Private notes extend beyond the page boundary for a clear visual distinction.

Pointing Gestures

The deployment confirmed previous findings that the Spotlight functionality plays a deictic role in support of voice. Based on qualitative feedback, we learned that the moving blob clarified which part of their paper the instructor wanted them to edit and revise. To quote: "It was helpful to hear the professor's comments as I was reading along with the bouncing ball [the Spotlight]. It helped me focus specifically on the sections that needed improvement. (P1)". All of the participants reported that they could see the instructor's Spotlight annotations during replay. They also rated the feature to be very helpful for understanding the instructor's audio comment (M = 4.62, SD = 0.51).

Conflicting Responses to Visual Richness and Complexity

The qualitative feedback drew out a diverse response to the visual complexity that might be introduced by the interleaved representation of waveforms. P1 and P8 mentioned that having the waveform alongside text lines didn't bother them: "I didn't mind seeing the voice wavelengths on the page (P8)". However, there was a single user who felt that waveforms cluttered the page: "seemed a bit messy… [I] did not like how the sound waves cut into the page (P16)". Other users didn't mention any comforts or discomforts regarding the system's visual complexity.

SUPPORTING PEER DISCUSSION

We followed up the deployment above with a second study, where the multi-modal annotation system was used to support student discussion around course readings. This follow-up study allowed us to examine a different communication pattern encountered in education: students receiving and maintaining awareness about comments from several different peers [9]. Also, in this follow-up study students produced multi-modal annotations in addition to consuming them. Finally, the system we deployed in this study supported text annotations, which enabled us to observe differences in how students perceived the production costs between textual and non-textual annotation modalities [21].

System Changes to Support Peer Discussion

To support peer discussion through multi-modal annotation, students needed the ability to create multi-modal comments in addition to being able to consume them. This necessitated modification of the standard RichReview system, which was stylus-centric, to one that was keyboard/touchpad-centric. In this new system, we supported typed text along with voice and pointing gestures. Students could respond to typed annotations using any of the supported annotation modalities to form multi-modal discussion threads (Figure 1). Like the instructor feedback tool, we did not include the ability to edit voice comments. For clarity, we refer to this new system as RichReview^{++}.

We also added three new features that were designed to facilitate and coordinate the discussion process. First, to enable quick and exhaustive browsing of the comments from multiple peer users, we added a Comment History feature that showed a chronologically sorted list of icons representing all of the comments present in a document (Figure 5). Second, although peer discussions took place around public annotations, the literature suggested that support for private annotations would be important as a staging step before creating public comments and also to support active reading [22]. Consequently, our system added the ability to create Private Notes that only the creator could view and edit. To make the private notes visually distinguishable from the public comments, we extended them over the page bounds, creating a clear visual contrast (see Figure 6). Third, the interview results from the first block indicated that binding a RichReview^{++} annotation to a specific line of text would discourage discussion about the texts as a whole and preclude synthetic thinking that related different articles to each other [39]. Consequently, we added a dedicated page called "General Notes" for these high-level comments. This special page consisted of two columns. The left column was left empty and provided space for high-level comments. The right column was populated with titles of all of the articles in the assigned reading for the week. Using this list of article titles as a high-level representation of the whole topic area, students could insert comments about an article underneath its title and indicate relationships between different articles by connecting them with Spotlights.

Deployment Setting

The RichReview^{++} online discussion system was deployed to a graduate level social science course in the spring 2015 semester. In the class, an instructor taught 18 students, with 4 of them connecting from a satellite campus to the main class via a videoconferencing system. This seminar type class centered on individual readings and class-wide peer discussion activities, which were open-ended and student-directed. Assigned readings for a given week were first

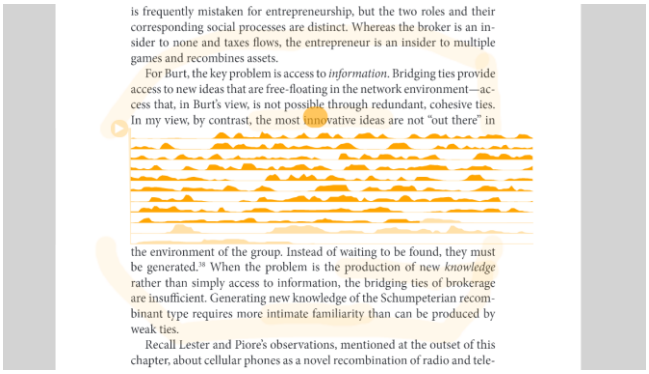

Figure 7. P7 referred to three different phrases and a paragraph using the pointing gesture feature.

discussed online for a week, and then in an offline class-wide discussion session lasting 2.5 hours. Online discussion contributions made up 20% of the grade. The reading materials were composed of either 4-5 different conference papers, 3-4 chapters of a textbook, or a mix of both, totaling 150-250 pages per week.

Students used RichReview^{++} in two two-week long deployment blocks (weeks 6-7, and weeks 11-12 of the semester). At the end of each block of the study, we conducted a 30 minute long semi-structured interview with a focus on the way multi-modal annotation supported discussion activities.

14 of the 18 students (4 females) participated in the study. The participants were mostly graduate students (1 undergraduate) in their mid-20s (M = 26.5, SD = 3.5). The students' familiarity with the course topic occupied the entire spectrum from novice to very familiar. 5 participants were native English speakers, and the others spoke English as a second language. Students' proficiency in English ranged from intermediate to fluent. We placed the 14 participants into 3 groups with 4-5 members each in order to limit the volume of annotations on each group's document. The main reason for doing this was to make sure that RichReview^{++}'s visually rich comments did not take up all of the available screen real-estate. The student groups were balanced regarding background, gender, English proficiency, and campus location.

Results

In the first block of the study, we observed 251 textual comments and 90 voice comments. Of the voice comments, 37% contained pointing gestures. Students produced 322 comment threads and 306 of these consisted of a single comment (Mean word counts = 31.8, Mean recording duration = 19.6). 16 comments had replies. However, the resulting threads were only 2-3 comments long and did not have much back and forth conversation. Three participants created the majority of their comments using the recording feature, three others made mixed use of modalities, and the rest created their comments using typewritten comments. In the second block of the study, participants made a similar number of textual comments (236) but much fewer voice comments (8) than in the first block.

Pointing Gestures

The pointing gesture seemed to be most useful as a visual aid for directing the listener's attention to locations on the page to which the voice recording referred. Out of the 90 voice comments created in the first block of the study, 37 employed the use of pointing gestures, 25 of these referred to a single place of the body texts, and 12 pointed at multiple places (MAX=4, M = 1.41, SD = 0.77). One interesting property of annotations that referred to different locations was the shift in location granularity. For example, P7 moved between pointing to a phrase with a few words (e.g., "I don't want to agree with the statement that 'the most innovative ideas [gesturing over the phrase]' …") and pointing to the entire paragraph ("I agree with this view provided here in this part … [circling the paragraph]") in a single audio recording (Figure 7).

Qualitative data also supported the benefits of pointing in tandem with voice. Students reported that pointing gestures were useful for referring to graphical element of texts (P1) and connecting multiple parts of the document in a single description (P7). To quote P1: "I found it particularly useful for pointing out things like diagrams… there was a few confusing diagrams and, that [the Spotlight feature] allowed me to draw out what was confusing." On the consumption side, students found gestures were helpful for understanding the speaker's intent (P9, 10, 11), which corroborates our findings in the first study.

Barriers to Creating Voice Comments

Students reported that creating voice comments (M = 3.08, SD = 1.38) required more effort than typing (M = 4.33, SD = 0.49). The reasons behind why our students felt it hard to record voice echoed previous findings of Marriott in [21]. These reasons included lack of editing features, self-consciousness (non-native speakers concerned about their accents), and environmental constraints (e.g., working in a library or when a roommate was sleeping). Also, the linear and irreversible nature of voice recording made them feel compelled to keep speaking, which interfered with their thinking (5 of 11 students). The lower use of voice comments in the second block of the study can also be explained by the fact that students were busier at the end of the semester and had less time for creating multi-modal comments.

Discussion Support Features

The most popular feature amongst participants was the Comment History. Students employed the feature to jump through comments to get a general sense of how peers' read the course materials. Private Notes were seldom used because students tended to read and annotate the material on print-outs or using their own archiving tools (such as Google Docs or Evernote). General Notes also did not get used much since students ended up not having difficulty making high-level comments within the article text.

DISCUSSION

Importance of Deictic Annotations

The simplicity of the gestural annotation functionality in the system we tested (essentially a single semi-transparent blob) belies the depth and diversity of how it was used, particularly for deictic purposes. For instance, in the first block of the second study, students could easily modulate the granularity of what was being discussed between a phrase, a paragraph, and multiple parts of a page through small variations in the movement of the Spotlight pointer. In light of the utility of deictic annotation, a weakness of the system we tested was the tight coupling between spotlight and voice annotations. In situations where voice annotation was not used extensively, such as in the second block of the second study, the lightweight mechanism for specifying deictic information was also unavailable. In the future work section, we detail a new strategy that may can combine rich deictic information with modalities other than voice.

Visual Representations for Annotations

Many of the annotation modalities that are employed do not have an obvious visual representation. Such is the case for speech and gestures. Our results showed that giving these annotations a visual component, such as through waveforms or Spotlight traces increased the ease of access. On the other hand, these benefits must be weighed against the dangers of visual clutter from these rich graphical representations that prior studies conducted in the lab suggested may occur [37].

A surprising (and encouraging) finding from our deployment was that only a minority of users reported that the presence of interleaved audio waveforms interfered with their consumption of the main document. It bears mentioning that RichReview included facilities for hiding all annotations and restoring the original layout of the underlying document, which would have been applicable to users who found the waveforms distracting. Unfortunately, the annotation hiding functionality was rarely used during the deployment due to low discoverability and so we currently cannot say for certain whether it represents an effective solution.

With regards to Spotlight traces, none of users mentioned any difficulties arising from their presence. We believe the reason is because we kept the opacity of the traces fairly low (alpha = 0.15). On the other hand, users also reported that the Spotlight traces were too faint to notice and to operate the indexing feature on them. We suspect that adjusting the opacity of the traces has a trade-off between promoting interaction with them and preventing visual clutters from them. One possible solution is to interactively adjust the opacity with which traces are rendered based on where a person hovers the cursor in a document.

Factors Affecting the Production of Voice Annotations

We observed a difference between the two deployments regarding the perceived ease of producing voice annotations. We believe that this result highlights the fact that the role of the annotation creator and task type plays a role in determining how voice annotations are employed.

The instructor in our writing feedback study and the executives in Ehrlich's study [6] successfully leveraged the nuanced and expressive capacity of voice commenting without observing any serious disadvantages. In contrast, students engaged in peer discussion, such as those in Marriott's [21] and Hew and Cheung's [12] studies, and those in our study reported difficulties when producing voice comments.

One explanation for this difference might be that individuals in a supervisory role, such as teachers or managers, are more comfortable generating voice comments. This effect could stem from the fact that their position requires providing feedback and directions for guidance in many situations beyond document annotation. Another factor can be due to the differences in annotation content. Making an insightful contribution in the discussion task required students to engage in a deeper, synthetic type of thinking in contrast to the more analytic thinking required of the instructor when giving feedback. These differences warrant further investigation to ascertain their underlying cause.

Nevertheless, an important lesson for annotation systems is that voice annotations are not all equivalent; general observations like the fact that voice is easier to produce but hard to consume [8] may or may not be accurate depending on the context. Furthermore, knowledge about the relationship between users of the system as well and annotation content allows for the multi-modal annotation system to be tailored to the task at hand.

Better than Being There

Many students using RichReview preferred it to in-person meetings during the instructor's office hours. This result seemed a bit unintuitive given the fact that face-to-face meetings are typically regarded as the gold standard for communication technology. Hollan et al., however, pointed out that the goal of communication technologies is not necessarily to emulate face-to-face meetings [14].

Digital communication tools provide useful functionality not available in face-to-face meetings. For example, RichReview's quick replay, asynchronous workflow, and visually obvious pointing gestures made it easier for students to thoroughly process instructor comments at their own pace. We take this as encouraging evidence that tools like RichReview can leverage digital affordances to provide a superior alternative to face-to-face interaction for specific communication tasks.

Applications in MOOC Environments

One important domain where we believe tools like RichReview would impart benefits is in massively open online courses (MOOCs). In MOOCs, personalized feedback from instructors is less practical due to the high

student to instructor ratio. Therefore, online peer interaction, such as discussion [17] or assessment [18], has been suggested as a scalable solution for providing feedback to a large number of students. Combining the rich set of user data conveyed through multi-modal annotation and the scalability of MOOCs can open up opportunities for applications that leverage the collective intelligence of students [31]. For example, aggregating students' gestural comments on a shared document can reveal collective interests (e.g., heavily gestured paragraphs), and the logical connections between different parts of a page (a phrase and a paragraph are often gestured in the same comment together).

CONCLUSION AND FUTURE WORK

In this paper we conducted two deployment studies to gauge the effectiveness of novel multi-modal annotation features for supporting various activities encountered in university classrooms. Our studies examined the use of the RichReview annotation tool as a mechanism through which instructors provide feedback about student writing and students to discuss about reading assignments.

The analysis of the log data and survey indicated students were able to effectively incorporate waveform indexing as part of a diverse navigation strategy. The waveform feature seemed to help students to fully digest the instructor's comments at their own pace. This resulted in many students' preference for the multi-modal annotation system over traditional commenting methods. Results from the follow-up study provided more evidence that gestural comments were useful for referring to different parts of a document. However the linear and irreversible nature of voice made the commenting task more stressful since students had to think and speak at the same time.

Another insight our data revealed was that rich deictic information have the potential of enhancing textual annotations as well. One possible method to achieve this would be to allow typed comments to refer to different parts of a document instead of the approach in current systems [4, 40] of anchoring the comment to a specific block of text. Deictic phrases such as "this sentence" or "that paragraph" in the comment can be associated with specific locations in the text to which they refer. Later, these connections can then be rendered visually when the comment is being viewed.

Broadly speaking, we found that the multi-modal annotations that RichReview^{++} offers to be useful. However, it is important to note that the technology on its own offers only a partial solution. We believe that understanding the affordances and implications of this type of collaboration tool and using it as part of a broader effort in the classroom is necessary to realize the full potential of the tool.

ACKNOWLEDGEMENTS

Yoon gratefully acknowledges support from the Kwanjeong Educational Foundation. This work was supported in part by gifts from Microsoft and FXPAL.

REFERENCES

1. Timothy Bickmore, Laura Pfeifer, and Langxuan Yin. 2008. the Role of Gesture in Document Explanation By Embodied Conversational Agents. *International Journal of Semantic Computing* 02, 01: 47–70.
2. Aj Brush, David Bargeron, and Jonathan Grudin. 2002. Supporting interaction outside of class: anchored discussions vs. discussion boards. *Computer Cupport for Collaborative Learning: Foundation for a CSCL Communiy*: 425–434.
3. Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. 1991. Expressive richness: A COMPARISON OF SPEECH AND TEXT AS MEDIA FOR REVISION. *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*, ACM Press, 21–26.
4. Elizabeth F. Churchill, Jonathan Trevor, Sara Bly, Les Nelson, and Davor Cubranic. 2000. Anchored conversations: Chatting In The Context Of A Document. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00*, ACM Press, 454–461.
5. Anne Crook, Alice Mauchline, Stephen Maw, et al. 2012. The use of video technology for providing feedback to students: Can it enhance the feedback experience for staff and students? *Computers and Education* 58, 1: 386–396.
6. Susan F Ehrlich. 1987. Strategies for Encouraging Successful Adoption of Office Communication Systems. *ACM Trans. Inf. Syst.* 5, 4: 340–357.
7. Susan R Fussell, Leslie D Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D I Kramer. 2004. Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Hum.-Comput. Interact.* 19, 3: 273–309.
8. Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, ACM Press, 85–93.
9. Carl Gutwin, Gwen Stark, and Saul Greenberg. 1995. Support for Workspace Awareness in Educational Groupware. *Proc Conference on Computer Spported Collaborative Learning*, Association for Computational Linguistics, 147–156.
10. Steve Harrison, Scott Minneman, and Joshua Marinacci. 1999. The DrawStream Station or the AVCs of Video Cocktail Napkins. *Proceedings of the IEEE International Conference on Multimedia Computing*

and Systems - Volume 2, IEEE Computer Society, 543–549.

11. Khe Foon Hew and Wing Sum Cheung. 2012. Examining the use of asynchronous voice discussion in a blended-learning environment. *Proceedings of the European Conference on e-Government, ECEG* 10, 4: 136–140.
12. Khe Foon Hew and Wing Sum Cheung. 2013. Audio-based versus text-based asynchronous online discussion: Two case studies. *Instructional Science* 41, 2: 365–380.
13. Debby Hindus and Chris Schmandt. 1992. Ubiquitous audio: Capturing Spontaneous Collaboration. *Proceedings of the 1992 ACM conference on Computer-supported cooperative work - CSCW '92*, ACM Press, 210–217.
14. Jim Hollan and Scott Stornetta. 1992. Beyond Being There. *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, ACM, 119–125.
15. Juho Kim, Elena L. Glassman, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. RIMES: Embedding Interactive Multimedia Exercises in Lecture Videos. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, ACM, 1535–1544.
16. Robert Kraut, Jolene Galegher, Robert Fish, and Barbara Chalfonte. 1992. Task Requirements and Media Choice in Collaborative Writing. *Human-Computer Interaction* 7, 4: 375–407.
17. Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S. Bernstein, and Scott R. Klemmer. 2015. Talkabout: Making Distance Matter with Small Groups in Massive Classes. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, ACM Press, 1116–1128.
18. Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*, ACM, 75–84.
19. Joon-Suk Lee and Deborah Tatar. 2012. "Good enough" pointing in pervasive computing. *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, 330–337.
20. Stephen R Levine and Susan F Ehrlich. 1991. The Freestyle System. In *Human-Machine Interactive Systems*. Springer, 3–21.
21. Philip Marriott. 2002. Voice vs text-based discussion forums: An implementation of Wimba Voice Boards. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 640–646.
22. C.C. Marshall and a.J.B. Brush. 2004. Exploring the relationship between personal and public annotations. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.*, 349–357.
23. Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. 2013. NoteVideo: Facilitating Navigation of Blackboard-style Lecture Videos. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press, 1139.
24. Christine M. Neuwirth, Ravinder Chandhok, David Charney, Patricia Wojahn, and Loel Kim. 1994. Distributed collaborative writing. *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*, ACM Press, 51–57.
25. Jody Oomen-Early, Mary Bold, Kristin L Wiginton, Tara L Gallien, and Nancy Anderson. 2008. Using asynchronous audio communication (AAC) in the online classroom: A comparative study. *Journal of Online Learning and Teaching* 4, 3: 267–276.
26. Bernie Randles, Dongwook Yoon, Amy Cheatle, Malte Jung, and Francois Guimbretiere. 2015. Supporting Face-to-Face Like Communication Modalities for Asynchronous Assignment Feedback in Math Education. *Learning @ Scale*, ACM, 321–326.
27. Jonathan Grudin Scott LeeTiernan. Fostering Engagement in Asynchronous Learning Through Collaborative Multimedia Annotation.
28. Mary Lourdes Silva. 2012. Camtasia in the Classroom: Student Attitudes and Preferences for Video Commentary or Microsoft Word Comments During the Revision Process. *Computers and Composition* 29, 1: 1–22.
29. Lisa Stifelman, Barry Arons, and Chris Schmandt. 2001. The Audio Notebook: Paper and Pen Interaction with Structured Speech. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, 182–189.
30. Michael Tsang, George W Fitzmaurice, Gordon Kurtenbach, Azam Khan, and Bill Buxton. 2002. 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*, ACM, 111–120.
31. Sarah Weir, Juho Kim, Robert C Miller, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 405–416.
32. Steve Whittaker, Julia Hirschberg, Brian Amento, et al.

2002. SCANMail: a voicemail interface that makes speech browsable, readable and searchable. *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, ACM Press, 275.

33. Steve Whittaker, Patrick Hyland, and Myrtle Wiley. 1994. Filochat: handwritten notes provide access to recorded conversations. *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*, ACM Press, 271–277.
34. Lynn D. Wilcox, Bill N. Schilit, and Nitin Sawhney. 1997. Dynomite: a dynamically organized ink and audio notebook. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97*, ACM Press, 186–193.
35. Thomas Wolsey. 2008. Efficacy of Instructor Feedback on Written Work in an Online Program. *International Journal on E-Learning* 7, 2: 311–329.
36. Elisabeth Yaneske and Briony Oates. 2010. Using voice boards: Pedagogical design, technological implementation, evaluation and reflections. *Australasian Journal of Educational Technology* 26, 8: 233–250.
37. Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, ACM, 481–490.
38. Dongwook Yoon, Nicholas Chen, and François Guimbretière. 2013. TextTearing: Opening White Space for Digital Ink Annotation. *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, ACM, 107–112.
39. Qixing Zheng, Kellogg Booth, and Joanna McGrenere. 2006. Co-authoring with structured annotations. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, ACM Press, 131.
40. Sacha Zyto, D Karger, M Ackerman, and S Mahajan. 2012. Successful classroom deployment of a social document annotation system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, 1883–1892.