Distributed Collaborative Writing: A Comparison of Spoken and Written Modalities for Reviewing and Revising Documents

Christine M. Neuwirth, Ravinder Chandhok, Davida Charney†, Patricia Wojahn and Loel Kim

Carnegie Mellon University
Pittsburgh, PA 15213, USA
Tel: +1-412-268-2468
E-mail: prep-project+@andrew.cmu.edu

†Penn State University
University Park, PA 16802, USA
Tel: +1-814-865-9703
E-mail: irj@psuvm.psu.edu

ABSTRACT

Previous research indicates that voice annotation helps reviewers to express the more complex and social aspects of a collaborative writing task. Little direct evidence exists, however, about the effect of voice annotations on the writers who must use such annotations. To test the effect, we designed an interface intended to alleviate some of the problems associated with the voice modality and undertook a study with two goals: to compare the nature and quantity of voice and written comments, and to evaluate how writers responded to comments produced in each mode. Writers were paired with reviewers who made either written or spoken annotations from which the writers revised. The study provides direct evidence that the greater expressivity of the voice modality, which previous research suggested benefits reviewers, produces annotations that writers also find usable. Interactions of modality with the type of annotation suggest specific advantages of each mode for enhancing the processes of review and revision.

KEYWORDS: computer-supported cooperative work, collaborative writing, annotations, voice.

INTRODUCTION

A recent study of speech versus text as media for reviewing documents found benefits for producing voice annotations for reviewing documents [2, 9]: whereas subjects using speech were more likely to comment more on higher-level concerns than those using text, subjects using text were more likely to comment more on lower-level concerns than those using speech. When subjects using text did comment on higher-level concerns, their comments were judged to be less useful. The study did not, however, examine the effects of the spoken annotations on the *recipient*. Recipients of voice annotations may be at a disadvantage when compared to recipients of written annotations. For example, recipients of a voice message are more limited in their ability to review, skim and otherwise abstract the content. While speech is faster than writing for the producer, it can be slow and tedious for the receiver to process [4]. It is important to investigate the costs and benefits of voice annotations for the writer as well as for the annotator because, as Grudin [5] notes, people will not use systems that have high costs unless those costs are balanced by high benefits.

CHI94-4/94 Boston, Massachusetts USA
© 1994 ACM 0-89791-650-6/94/0051...$3.50

To explore ways to balance the relative costs and benefits of spoken comments, we built an interface to provide some structure for the voice annotations and to allow recipients of annotations some degree of control over this very high bandwidth information, without sacrificing ease of use for the annotator. In addition, we undertook a study with two goals: to compare the nature and quantity of comments produced in voice and in writing, and to evaluate how writers responded to comments produced in each mode. To examine these issues, we conducted an experiment in which writers were paired with annotators who commented on their texts. The annotators made either written or spoken annotations and the authors revised on that basis.

THE INTERFACE DESIGN

In the present study, participants used a voice interface in a prototype collaborative writing environment, called the PREP Editor. The PREP Editor provides a document with multiple columns [11; 10]. Unlike columns for printing, these columns are like margins--they provide a space for communicating about a document.

When designing support for annotations, we considered two aspects of the user interface: production and reception. To produce comments in the PREP Editor, the user creates an annotation column and specifies a default mode of annotation from among three available modes: text, drawing, and voice. Then to create annotations, the user simply clicks in the annotation column near the relevant document content, an operation that resembles putting pen to paper in the margin of a hard copy document. For reception of voice annotations, we worked on ameliorating problems in (1) accessing voice annotations, (2) listening to them, and (3) revising in response to them.

Accessing. In a study of tape recorder users, Degan et al. [3] reported that users' biggest frustration was difficulty in accessing individual ideas. We hypothesized that allowing reviewers to associate individual comments with particular parts of a document, as described above, would alleviate this somewhat. In addition, the PREP Editor provides two ways

to access individual annotations. First, users can review annotations "at-a-glance" as they scroll through the document. A constraint-based layout system maintains the side-by-side alignment of each annotation with its target as shown in Figure 1 below.

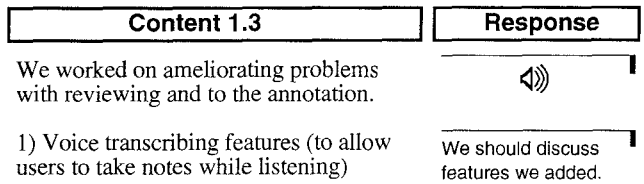

Figure 1: Annotation column (with voice and text comments) next to a content column in the PREP Editor.

Users play voice annotations by clicking on the iconic representation. This method was hypothesized to be most useful for authors who want to work with comments while revising a text. The second access method is via a "down to next" command that jumps to the next comment, where, if it is a voice comment, it automatically plays. This method was hypothesized to be most useful for obtaining a quick overview of all the comments.

Listening. To alleviate some of the problems in listening to comments, we designed a direct manipulation "sound palette" interface, depicted in Figure 2. The buttons across the top of the sound palette correspond to buttons on a standard tape player (from left to right): record, stop, play, rewind, pause, and fast forward. An important feature of the rewind button is that the sound resumes playing automatically after the user lets up on the button. This feature is common in transcription machine interfaces and allows users to relisten to the last bit of sound easily. Across the bottom of the sound palette, a progress bar shows how much of the sound has played. When a user clicks on a location in the progress bar, the sound begins playing from that point in the recording. Thus the progress bar can fast forward, rewind, or loop the sound by direct manipulation. While the mouse button is down in the progress bar, a short sample (<2 seconds) of the sound continuously cycles. By moving the mouse the user can "skim" through the sound. All these features are designed to help users replay sounds rapidly.

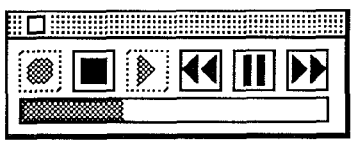

Figure 2: Sound palette.

Revising. When working with a written comment, an author can read part of the comment, revise the text, read the next part of the comment, revise, and so on. To facilitate this process for voice comments, the interface pauses the playback whenever a user takes certain actions such as selecting, revising, or creating text. This allows authors to respond to voice comments and then resume the playback easily.

In sum, the design of the interface was intended to preserve the ease with which reviewers are able to produce voice annotations while ameliorating problems authors might experience in using such annotations. The experimental study assessed how well the interface achieved its goals and explored the relative effectiveness of the two modes for authors and reviewers.

EXPERIMENTAL METHOD

Participants. The participants in the study consisted of twenty authors paired with twenty reviewers. The authors were members of the School of Computer Science, recruited from a pool of people who by self-report had manuscripts that were close to being ready for review. They participated by submitting an on-line version of their manuscripts, by suggesting appropriate reviewers, and by revising the manuscripts based on a reviewer's comments. The reviewers participated by annotating a manuscript. The authors and reviewers were not compensated in any way.

Materials. Authors' manuscripts included grant proposals, conference papers, book chapters, technical reports, and journal articles. All participants used the PREP Editor.

Design. The study employed a 2x2 mixed factorial design. The first factor, production mode, was varied between subjects: half the reviewers produced annotations in voice mode and half in written mode. The second factor, reception mode, was varied within subjects. As discussed below, annotations were manipulated so that each writer received half the comments in voice and half in text. This design allowed us to see, for example, whether comments produced in voice but received in writing might be superior to those received in voice.

Procedure. Each pair of participants took part in three sessions. During the annotation session, the reviewer commented on the draft. In the revision session, the writer read and listened to the reviewer's comments and revised. In the evaluation session, the reviewer evaluated the responsiveness of the writer's revised draft to the comments he or she had made.

Prior to the annotation session, reviewers were asked to read a hard-copy of the manuscript (without making comments) within twenty-four hours of their session. During that session, reviewers had one hour to annotate the manuscript. They were asked to follow their usual manner of reviewing, except that they were asked to use the PREP Editor. Additionally, their mode of commenting was restricted to either voice or text. Reviewers were randomly assigned to either voice or written production mode. Following these sessions, every other annotation was converted to the alternative mode. In the voice condition, half of the comments were transcribed into written text and substituted for the original comments In the text condition, reviewers made only written annotations during their initial session and were then asked to record half of their annotations.

Prior to the revision sessions, the documents were prepared so that the writer's manuscript would appear in one column and the annotations (half in text, half in voice) would appear in a second column. During the session, writers had one hour in which to listen to or read their reviewer's annotations and to revise their manuscripts. After completing their revisions, writers filled out two surveys. First, writers reported their impressions of both the reviewer and the review in terms of competence, personal integrity and likability. In the second survey, writers rated their preferred mode for receiving comments on audience, purpose, substance, organization, style, and grammar.

During the evaluation session, reviewers listened to or read the comments they had made and rated the revised draft on its responsiveness to each comment. In this session, reviewers saw a PREP Editor screen containing 5 columns: 1) the initial document; 2) comments made during the reviewing session; 3) the revised document; 4) changes between the initial and revised text (with added text in italics, deleted text underlined); and 5) a 7-point Likert scale for responsiveness ratings. All participants were asked to "think-aloud" throughout their sessions; all sessions were audio- and videotaped.

ANALYSIS

In both modes of production, reviewers sometimes recorded more than one problem within a single chunk of the manuscript. Independent raters parsed the reviewers' chunks into units containing a single annotation. The inter-rater reliability of the two coders as indicated by signal detection analysis was .83. On average, 28% of the chunks were identified as containing more than one annotation.

The overwhelming majority (92%) of the resulting annotations concerned the communication of a problem (e.g., "Typo. Should be built."); 5% of the annotations were compliments (e.g., "This is great!"); 3% were extraneous remarks (e.g., "Let me put this comment somewhere else.") and irrelevant statements (e.g., "I'm hungry."). Agreement concerning these classifications between two trained raters on 15% of the data was .88 by Cohen's Kappa. Extraneous and irrelevant remarks were excluded from further analysis.

RESULTS AND DISCUSSION

We looked first at the effects of production modality on the cognitive and social aspects of reviewers' annotations. Then we examined the effects on writers' perceptions of reviewers as well as on writers' performance. Finally, we looked at writers' preferences for mode of annotation.

Cognitive Aspects of the Annotations

While it seems unlikely that the mode of producing annotations would affect reviewers' abilities to identify and judge the problems in a text, it might affect their ability to communicate those problems. We examined three factors that might be affected: the number of problems communicated, the type of problems communicated, and the characterization of the problems communicated.

Number of problems communicated. Producing annotations in voice instead of in writing might increase the number of problems communicated because speaking is easier and faster than writing. Given constraints of time or motivation, reviewers making voice comments may make more comments than they would if they were writing. On the other hand, previous research by Kiesler et al. [8] indicates that people tend to compensate somewhat for the slowness of typing by making their main point in fewer words. The present study supports the latter position, as indicated by comparing the number of annotations produced to the relative number of words per comment.

Regardless of whether they were producing voice or written annotations, reviewers averaged about one annotation every two minutes: the mean number of annotations reviewers produced in an hour in voice was 32.0 (S.D. = 29.9) and in writing, 28.9 (S.D. = 12.7). The variance was high and no significant effect for mode of production was found (but see the next section). Reviewers making voice annotations, however, produced roughly two-and-a-half times more words per annotation than reviewers making written annotations: 62.9 words per annotation (S.D. = 37.1) vs. 23.7 words per annotation (S.D. = 7.6) [$F(1,18) = 10.76$; $p<.01$]. Given that most of the annotations identified problems (28.3, S.D.=21.1 for voice; 27.5, S.D.=11.3 for writing), these data indicate that reviewers in both conditions communicated about the same number of problems to writers, with large between-subject variability.

Type of problem communicated. The mode of producing annotations might also affect reviewers' choices of the type of problem to communicate. Research on cognitive processes in revision indicates that representing a text problem is a demanding process [7]. Part of the difficulty lies in the nature of the problems themselves, which can range from well-defined, relatively easy-to-communicate problems (such as violations of a spelling rule) to more difficult to define, difficult-to-communicate problems (such as inappropriate tone). The more difficult problems are those which are harder to designate with a specific term, those whose effect is more diffuse, and those for which there are fewer readily specifiable solutions.

To examine whether the mode of production might affect the types of problems reviewers communicated, annotations were categorized as focusing on problems of (1) mechanics, (2) style, (3) organization, (4) substance (e.g., accuracy of content, adequacy of the reasoning, etc.), (5) purpose/audience, and (6) other (e.g., references to figures, external procedures). Inter-rater reliability of two coders on a subset of the data was .80 by Cohen's Kappa. Table 1 shows the distribution of annotations by problem category and mode of production. We analyzed the data with a 2x6 repeated measures ANOVA, with production mode as the first factor and problem category as the second factor.

Overall, reviewers in both production modes focused more heavily on particular kinds of problems: substance comments were by far the most frequent, followed by

comments on style, and then the rest [F(5,90) = 23.02; p<.001]. Reviewers identified about the same total number of problems in both production modes (an average of 30 comments, or about 5 of each type), but mode of production did influence the type of comments produced, as indicated by a significant interaction [F(5,90) = 3.22; p<.05]: In most categories, reviewers in voice mode produced more comments than reviewers working in text mode. This pattern is reversed, however, for substance comments--to a great enough degree that the total number of comments in the two modes balances out.

	Mode of Production		
Problem	Voice	Keyboard	*Mean*
Mechanics	3.0 (4.5)	1.8 (1.5)	*2.4*
Style	8.3 (9.1)	4.6 (4.8)	*6.5*
Organization	3.9 (6.1)	2.9 (3.8)	*3.4*
Substance**	9.0 (6.3)	15.8 (7.2)	*12.4*
Purpose/Aud.*	3.5 (2.5)	1.6 (1.8)	*2.6*
Other	3.9 (10.6)	2.2 (2.3)	*3.1*
Mean	*5.3*	*4.8*	

Table 1. Mean (S.D.) number of reviewer comments by problem identified (*p < .05; **p < .01)

These results seem in marked contrast to a previous study of voice vs. written annotations [2; 9]. Although the coding categories are not directly comparable, in that study, across modalities, copy-editing annotations (corresponding to "mechanics") were by far the most frequent. Moreover, reviewers working in the written modality were more likely to make more low-level annotations than reviewers in voice modality while reviewers in voice modality were more likely to make more global annotations than reviewers in written modality. The reviewers in the study reported here, however, included faculty as well as graduate students, whereas the reviewers in the previous study were all MBA students. Research in writing processes indicates that less experienced writers tend to focus on low-level errors when revising the texts of others [7]. Also, reviewers in the previous study handwrote rather than typed their written comments, possibly resulting in an even greater degree of production difficulty in the written mode [6]. There is some evidence that, at least for very young writers, focus on low-level aspects in the composition of original texts can be shifted to higher-level concerns when they are freed from the mechanical production difficulties of written language by being allowed to dictate rather than write [1]. Further research is required to determine whether greater expertise or technological differences influenced the reviewers in the current study to produce more comments on substance in the written mode.

Characterization of Problems. The mode of production might also affect how reviewers characterize the problems that they communicate. Previous research indicates that writers' characterizations of problems can range from "spare representations that contain little information about the problem to richly elaborated diagnoses that offer both conceptual and procedural information about the problem" [7]. Writing, even for mature writers, remains a complex activity and people often complain they cannot put ideas to paper fast enough to keep up with their thoughts. Thus, the mode of production may influence the way reviewers characterize problems, with reviewers using voice communicating more elaborated representations and those using keyboards producing sparer representations.

Although there are many ways comments could be analyzed for differences in how problems are characterized, research on persuasion suggests that one important way is whether or not the comment includes a reason for a recommended change [cf. 12]. To examine this possibility, annotations from five subjects in each condition were randomly selected and coded as to whether or not they contained reasons for a recommended change. Inter-rater reliability of two coders on a subset of the data was .71 by Cohen's Kappa.

While reviewers using voice did not produce a greater absolute number of annotations with reasons than reviewers using keyboards (means of 12.8 vs. 9.8 annotations respectively), annotations with reasons represented a greater proportion of the total annotations made in voice (59%) than of the total annotations made with keyboards (28%) [F(1,8) = 10.57; p<.01]. Taken with the results reported above, this finding supports the hypothesis that the effort required to produce a comment influences its length and its content. Later, we consider whether these differences in the comments also influence their reception by writers.

Social Nature of the Annotations

Reviewers often use various language mechanisms to maintain social ties (e.g., techniques to avoid offending writers). In this section, we review results reflecting the social quality of the annotations.

Politeness. The production mode of annotations might affect how politely reviewers communicated problems to the writers. Politeness mechanisms such as equivocation and mitigation generally take more words to express than their unequivocal or unmitigated counterparts. Therefore the slowness of typing compared to speech might result in reviewers in the written production mode expressing themselves in more succinct and unmitigated ways, with the result that their comments might be seen as less polite.

Annotations were categorized into four levels of politeness: compliments; mitigated suggestions; direct, unmitigated suggestions; and impolite comments. Inter-rater reliability of two coders on a subset of the data was .88 by Cohen's Kappa. Table 2 shows the distribution of comments in these categories as a function of mode of production. We analyzed the data with a 2x4 repeated measures ANOVA, with factors production mode and politeness category. The main effect for politeness was significant [F(3,54) = 40.62; p<.001]. Overall, subjects produced many more mitigated and unmitigated problem identifications than rude remarks or compliments.

The mode of production did have the expected effect on politeness, as indicated by a significant interaction [F(3,54) = 12.47; p<.001]. As illustrated in Table 2, subjects who produced annotations by voice were more likely to use mitigated language in identifying problems ($p < .05$); subjects who keyboarded were more likely to use unmitigated language ($p < .05$).

	Mode of Production		
Politeness	Voice	Keyboard	*Mean*
Compliment	2.8 (8.2)	0.5 (0.9)	*1.6*
Mitigated*	21.0 (11.6)	10.6 (4.8)	*15.8*
Unmitigated*	6.6 (9.5)	16.0 (8.7)	*11.3*
Rude	0.7 (2.2)	0.9 (2.2)	*0.8*
Mean	*7.8*	*7.0*	

Table 2. Mean (S.D.) number of annotations as a function of level of politeness (*p < .05)

Writers' Assessments of Reviewers. With less mitigating language in their feedback, reviewers working in text mode might be evaluated less favorably by writers than reviewers working in voice mode. We examined writers' subjective assessments vis-a-vis their reviewers along three dimensions: perceived competence (e.g., expert-inexpert), personal integrity (e.g., fair-unfair), and likability (e.g., friendly-unfriendly). We focused on these dimensions because previous research suggested that these perspectives on the source of an annotation can influence how persuasive the writer finds the annotation [12].

The mode of production could plausibly affect each of these factors. Research on the human speech production system indicates that it is frequently overtaxed, with the result that speech is filled with errors and unplanned pauses. And research in persuasion indicates that dysfluencies in delivery (vocalized pauses such as "uh," articulation difficulties, and so forth) may result in lower judgments of the speaker's competence, with judgments of personal integrity and likability unaffected [12]. Research on computer-mediated communication suggests that reviewers producing annotations in voice might be perceived as having more personal integrity and as being more likable than those producing annotations in writing [cf. 13].

While no significant effect was found for production mode on writers' assessment of reviewer competency, writers' assessment of reviewers' personal integrity was more likely to be lower when reviewers produced their comments by writing rather than speaking [F(1,18) = 7.46; p<.05] (see Table 3). Producing comments in writing also marginally ($p = .06$) lowered assessments of reviewers' likability. In all, these findings support the hypothesis that writers' evaluations of reviewers will be less positive when reviewers produce written annotations than when they produce spoken comments. The fact that we found this effect even when writers selected their own reviewers is striking. This effect might be even larger when writers do not select their reviewer (e.g., a thesis advisor or supervisor), or when the reviewer is unknown (e.g., blind journal reviews).

	Mode of Production	
Assessment	Voice	Keyboard
Competence	5.4 (0.5)	5.2 (1.4)
Personal Integrity*	5.5 (0.2)	4.4 (1.1)
Likability†	5.8 (0.1)	5.2 (0.8)

Table 3. Mean (S.D.) ratings of reviewers' competency, personal integrity and likability (*p < .05; †p=.06)

Writers' Responsiveness to Annotations. The next question we explored is how writers responded to comments in each mode. Because the reviewers themselves were the most competent judges of how well their comments were addressed, we asked them to rate the revision for degree of responsiveness to their comments. Recall that regardless of how the annotations were produced, writers received half of them as voice annotations and half as written annotations. We therefore analyzed reviewers' assessments both by production mode and by reception mode, but neither factor influenced the responsiveness of the revisions. In all cases, the ratings averaged about 4.6 on a 7-point scale, indicating a reasonable degree of responsiveness across the board.

We take this result as indicating that voice annotations--at least given the technology employed here--are no harder to respond to effectively than written comments. This suggests that the other social and cognitive aspects of production and reception may be decisive in selecting appropriate writing aids.

Writers' Preference for Modality. We used 7-point scales to assess writer preferences for modality for receiving various types of annotations: Mechanics, Style, Organization, Substance, and Purpose/Audience. The ratings for each type are displayed in Table 4 as a function of mode of production (with higher scores indicating greater preference for voice). We analyzed the data with a 2x5 repeated measures ANOVA, with the first factor, mode of production and the second, type. As illustrated, ratings for the two modes of production were the same overall--and were fairly neutral. However writers did significantly prefer to receive certain types of comments in particular modes [F(4,72) = 10.83; p<.001]. Regardless of the mode in which comments were produced, writers preferred to receive comments about Purpose/Audience and Style in voice, and preferred to receive comments about Mechanics in writing.

The analysis also indicated that the mode of production and preference for reception interacted [F(4,72) = 4.27 p<.05]. Authors were more likely to prefer receiving comments on Organization in voice if they had been produced in voice and in writing if they had been produced in writing. Conversely, authors were more likely to prefer receiving comments on Purpose/Audience in writing if they were produced in voice and in voice if they were produced in writing.

Problem	Mode of Production: Voice	Keyboard	*Mean*
Mechanics	2.6 (1.0)	2.7 (1.3)	*2.7*
Style	4.7 (1.8)	4.4 (1.0)	*4.6*
Organization*	4.4 (1.4)	3.3 (0.9)	*3.9*
Substance	3.6 (1.4)	3.9 (0.8)	*3.8*
Purpose/Aud.*	3.6 (1.4)	4.9 (0.9)	*4.3*
Mean	*3.8*	*3.8*	

Table 4. Mean (S.D.) rated preferences by writers for receiving comments in voice (*p < .05)

Reviewers were required to produce all their comments in one modality, so we did not assess their modality preferences. A previous study found that reviewers preferred voice for producing high-level comments and writing for producing low-level comments [2; 9]. Of course, preferences can be conditioned by circumstances and these results may not generalize to other situations such as different acoustic environments, social situations, or user activities. For example, in the present study, voice comments were received in a private office. Writers might view voice annotations far less favorably if they had to listen to them in less private places in which it would be inappropriate or awkward to speak aloud or to hear someone's criticisms broadcast (e.g., at desks with only dividers between them). It is interesting that we found no systematic preference for having either the same production and reception modality or disjunct modalities. This result suggests that our methodology for transcribing comments did not produce noticeably artificial comments.

Implications For Interface Design

Prior to the study, each aspect of the interface underwent many changes through an iterative design process. The study itself also resulted in numerous qualitative observations about ways in which the interface design worked well or could be improved.

Producing annotations. At the time of the study, the grain-size for the unit of text which could be annotated was the paragraph (i.e., annotations could be attached to the beginning of a paragraph). Both reviewers and writers found this grain-size to be too large. In response, we have implemented a smaller-grain size: annotations can be attached to any region of text.

Receiving annotations. Authors used the "rewind" feature so they could relisten to parts of the comments they didn't catch or understand. It is questionable whether authors' favorable attitude toward voice annotations would transfer to interfaces lacking this feature. Since this capability seems crucial, it seems worthwhile to enhance this function. For example, our informal observations indicate that users may find a graphical representation of sound waves even more useful than a uniform progress bar, so that they can use speech pauses to detect points of interest.

Responding to annotations. The majority of authors used a "read-revise" strategy to work their way through the comments, occasionally making a "mental note" to return to an annotation later. On the other hand, a small number of authors read through all annotations before making a single revision. This latter group, however, did make evaluations of the annotation during the first pass (e.g., "off base," "good comment," etc.). Thus, both groups would probably benefit from having a way to mark annotations to which they want to return.

Remembering the content of annotations. While authors can rapidly skim written comments to remind themselves of the contents, authors had to replay voice comments to do so. One solution to this difficulty might be to provide authors with a convenient way of jotting a few notes while listening to a voice comment. We have implemented keyboard commands so that an author can access and control the voice annotations without having to move from the keyboard to the mouse to control the sound palette. These commands may make it easier for authors to make quick notes for themselves about the annotations as they listen.

CONCLUSIONS

This study complicates the previous picture of the utility of the voice modality for supporting collaborative writing activities. The results can be summarized as follows:

1. The mode of production affected the type of problem that reviewers communicated: While all the reviewers in the study produced more comments on problems of substance than any other type of problem, reviewers in voice mode were likely to produce more comments about purpose and audience than reviewers in keyboard mode, while reviewers in keyboard mode were likely to produce more comments about substance.

It may be that the written text, which more readily permits review of what has been written, reflection upon it, and revision, may facilitate comments that involve complex substantive issues. If production modality does influence the types of problems communicated, then writing tools offering both modes may need to provide guidelines for choosing the most appropriate mode to work in for encouraging evaluation at the appropriate level.

Interestingly, the type of comments produced by subjects in this study seems to differ markedly from those produced in a previous study [2, 9]. It is reasonable to hypothesize that some of the contrasts between the two studies stem from differences in subject sample, written mode (i.e., word processor vs. pencil and paper), and task motivation.

2. The mode of production affected how reviewers characterized problems. While reviewers in both modalities produced about the same number of annotations overall, the number of words per annotation was far greater in speech. This difference can be accounted for, in part, by the greater frequency of reasons and by the greater number of words used to produce mitigated statements. A higher proportion of the annotations produced in voice contained reasons why the reviewers thought something was a problem and polite language that mitigated the problem.

3. The mode of production affected how writers perceived their reviewers. Writers' evaluations of their reviewers were likely to be less positive when reviewers produced written annotations than when they produced spoken.

4. The study failed to find an overall difference in reviewers' assessments of how responsive writers were to annotations produced or received in the two modalities. Future analyses are planned to examine whether the nature of the annotations and writers' perceptions of reviewers interacted with responsiveness.

5. Despite the previous research findings that spoken annotations would likely be tedious to listen to and more difficult to process [4; 5], writers using the PREP Editor interface for voice annotations were generally favorably disposed or neutral to voice annotations for most types of comments, except low-level mechanical ones.

In this study, authors chose their reviewers and reviewers were constrained to produce comments in only one modality. More research is needed that varies both conditions of producing annotations and the social relations between the writer and reviewer and looks at annotation interfaces for other sorts of documents (e.g., CAD drawings, blueprints, videos, etc.).

The results presented here suggest that it will be useful to explore the effects of tools offering both voice and text modalities further, especially tools incorporating the ability to switch between modes easily when producing annotations. There are outstanding technological challenges associated with providing users the same functionality with voice annotations as they have with text annotations. In the PREP Editor, for example, it is possible for authors to make annotations on their reviewers' written annotations. This feature is used frequently by distributed collaborative writing groups to discuss particular annotations. Providing a similar functionality for voice annotations requires addressing issues of how an author selects a region of a voice comment on which to comment and how a voice annotation that itself has a voice annotation should be played. This study gives some foundation for future interface design and uses of voice in collaborative systems.

ACKNOWLEDGMENTS

This work was supported by a grant from the Information Networking Institute, sponsored by Bellcore. Development of the PREP Editor is supported by a grant from the National Science Foundation (grant number IRI-8902891) and by a grant from Apple Computer, Inc. Other members of the PREP Editor project group (Jim Morris, David Kaufer, Paul Erion, and Dale Wiggins) contributed user interface ideas. We thank the anonymous CHI reviewers, Jörg Haake, Jörg Geisler, and Jolene Galegher for insightful comments on an earlier draft. The PREP Editor prototype is available via anonymous ftp.

REFERENCES

1. Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.
2. Chalfonte, B., Fish, R. S., & Kraut, R. E. (1992). Expressive richness: A comparison of speech and text as media for revision. *In Proceedings of the CHI'92 Conference on Computer-Human Interaction,* (pp. 21-26). ACM Press.
3. Degan, L., Mander, R., & Salomon, G. (1992). Working with audio: Integrating personal tape recorders and desktop computers. *In Proceedings of CHI'92 Conference on Human-Computer Interaction,* (pp. 413-418). ACM Press.
4. Gould, J. D. (1978). An experimental study of writing, dictating, and speaking. In J. Requin (Ed.), *Attention and Performance* VII (pp. 299- 319). Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Grudin, J. (1988). Why computer-supported cooperative work applications fail: Problems in the design and evaluation of organizational interfaces. *In Proceedings CSCW '88 Conference on Computer-Supported Cooperative Work* (pp. 85-93). ACM Press.
6. Hawisher, G. E. (1989). Research and recommendations for computers and composition. In G. E. Hawisher & C. L. Selfe (Eds.), *Critical perspectives on computers and composition instruction* (pp. 44-69). New York: Teachers College.
7. Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Eds.), *Advances in applied psycholinguistics, Vol. II* (pp. 177-249). Cambridge, England: Cambridge University Press.
8. Kiesler, S., Zubrow, D., Moses, A. M., & Geller, V. (1985). Affect in computer-mediated communication: An experiment in synchronous terminal-to-terminal discussion. *Human-Computer Interaction,* 1, 77- 104.
9. Kraut, R.E, Galegher, J., Fish, R.S., & Chalfonte, B. (1992). Task requirements and media choice in collaborative writing. *Human-Computer Interaction,* 7, 375-407.
10. Neuwirth, C. M., Chandhok, R., Kaufer, D. S., Erion, P., Morris, J., & Miller, D. (1992). Flexible diff-ing in a collaborative writing system. *In Proceedings of the Fourth Conference on Computer-Supported Cooperative Work* (CSCW '92) (pp. 147-154). ACM Press.
11. Neuwirth, C. M., Kaufer, D. S., Chandhok, R., & Morris, J. H. (1990). Issues in the design of computer-support for co-authoring and commenting. *Proceedings of the Third Conference on Computer-Supported Cooperative Work* (CSCW '90) (pp. 183-195) ACM Press.
12. O'Keefe, D. J. (1990). *Persuasion: Theory and Research.* Newbury Park, CA: SAGE Publications.
13. Sproull, L., & Kiesler, S. (1991). *Connections.* Boston, MA: MIT Press.