CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces

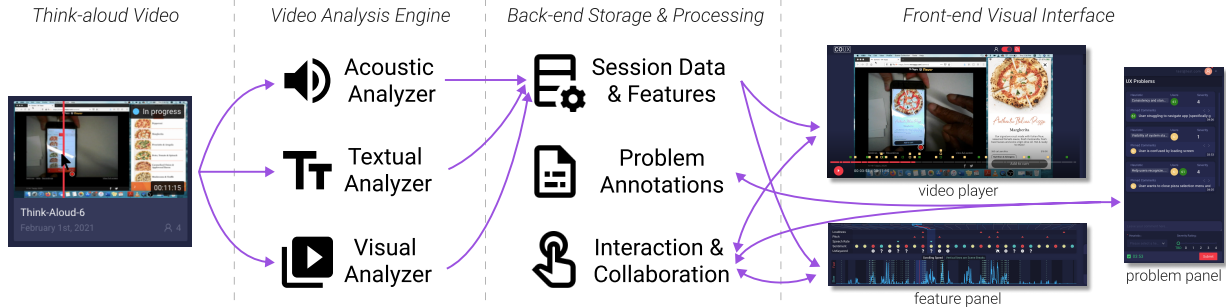Ehsan Jahangirzadeh Soure[†], Emily Kuang[†], Mingming Fan[*], and Jian Zhao[*]

Fig. 1. CoUX is a collaborative visual analytics tool to support multiple UX evaluators with analyzing think-aloud usability test recordings. From an input video, a video analysis engine extracts various types of features, which are stored on a back-end and presented on a front-end visual interface to facilitate the identification of usability problems among UX evaluators. Moreover, the front-end, consisting of three interactively coordinated panels, communicates with the back-end to support individual problem logging and annotation as well as collaboration amongst a team of UX evaluators.

Abstract—Reviewing a think-aloud video is both time-consuming and demanding as it requires UX (user experience) professionals to attend to many behavioral signals of the user in the video. Moreover, challenges arise when multiple UX professionals need to collaborate to reduce bias and errors. We propose a collaborative visual analytics tool, CoUX, to facilitate UX evaluators collectively reviewing think-aloud usability test videos of digital interfaces. CoUX seamlessly supports usability problem identification, annotation, and discussion in an integrated environment. To ease the discovery of usability problems, CoUX visualizes a set of problem-indicators based on acoustic, textual, and visual features extracted from the video and audio of a think-aloud session with machine learning. CoUX further enables collaboration amongst UX evaluators for logging, commenting, and consolidating the discovered problems with a chatbox-like user interface. We designed CoUX based on a formative study with two UX experts and insights derived from the literature. We conducted a user study with six pairs of UX practitioners on collaborative think-aloud video analysis tasks. The results indicate that CoUX is useful and effective in facilitating both problem identification and collaborative teamwork. We provide insights into how different features of CoUX were used to support both independent analysis and collaboration. Furthermore, our work highlights opportunities to improve collaborative usability test video analysis.

Index Terms—User experience, usability problems, think-aloud, video analysis, machine learning, visual analytics, collaboration.

✦

1 INTRODUCTION

Digital products have become increasingly feature-rich and often require users to navigate through an ever-growing number of onscreen elements, such as pressing a sequence of buttons to place an order on a smartphone. The increasing complexity of digital interfaces makes it challenging to achieve compelling user experience (UX). UX professionals often need to work collaboratively to identify and resolve UX problems via in-depth user evaluations. Of many evaluation approaches, usability testing with *think-aloud protocol* is widely used [12, 36] and considered as the single most useful method [41]. When using think-aloud protocols, participants verbalize their thoughts while performing actions. This allows UX evaluators to gain insights into their thought processes that is inaccessible to mere observations [33].

- *Ehsan Jahangirzadeh Soure and Jian Zhao are with the University of Waterloo. E-mails: {ejahangi,jianzhao}@uwaterloo.ca.*
- *Emily Kuang is with the Rochester Institute of Technology. E-mail: emily.kuang@mail.rit.edu.*
- *Mingming Fan is with the Hong Kong University of Science and Technology and the Rochester Institute of Technology. E-mail: mingmingfan@ust.hk.*
- *† These authors contributed equally. ∗ Corresponding authors.*

Despite being useful, analyzing recorded think-aloud videos is tedious, challenging, and time-consuming [6, 12, 14, 43]. First, UX evaluators need to make decisions by attending to multiple behavioral signals in both the visual and audio channels and conducting multiple tasks simultaneously in a fast pace [6], such as observing participants' actions, listening to their verbalized thoughts, inferring usability problems, and taking notes. Moreover, to increase the reliability and completeness of the analysis, UX evaluators are recommended to work collaboratively [14, 15] to overcome the *evaluator effect* [23]—the fact that different UX evaluators may uncover or interpret usability problems differently. Unfortunately, fewer than 30% of UX evaluators have a chance to collaboratively analyze the same usability test session due to practical constraints (e.g., limited company resources [6, 14]).

To mitigate these challenges, we propose a collaborative visual analytics tool, *CoUX*, to assist a team of UX evaluators with identifying, determining, and discussing usability problems in think-aloud usability test videos for digital products. Our approach is partially inspired by recent studies extracting acoustic and textual features (e.g., loudness, pitches, and sentiment) from a video to help identify usability problems [10, 11, 13]. We further leverage various machine learning techniques to detect acoustic and textual features directly from the audio (without manual transcripts), as well as user interactions (e.g., scrolling speed and scene breaks) from the video frames. To better support UX evaluators' decision making, CoUX segments a video into meaningful chunks based on the semantics exhibited in the think-aloud audio, extracts various visual, acoustic, and textual features, and visual-

izes the information collectively on multiple synchronized timelines. This design allows UX evaluators to easily attend to multiple streams of information likely indicating problems, to discover problems that might be otherwise overlooked, and make informed decisions about the occurrence and severity of the problems.

More importantly, CoUX is empowered with a collaborative decision support for the discussion and consolidation of analysis results amongst multiple UX evaluators. We draw on insights from studies of collaboration amongst UX evaluators and collaborative visualization (e.g., [14, 22, 63–65, 68]). CoUX allows UX evaluators to analyze a video independently, and then enter a collaborative mode to discuss and summarize their analyses, minimizing the evaluator effect [23]. In independent analysis, detected usability problems and their severity levels, as well as UX evaluators' reasoning, are automatically organized in a chat box like interface. During collaboration, UX evaluators are enabled with interactive and visual support from CoUX to make decisions collaboratively by discussing their findings in structured conversational threads and consolidating the results, synchronously or asynchronously.

Our design of CoUX is grounded in design considerations derived from the literature and our interviews with two UX professionals. For evaluation, we conducted a user study with six pairs of UX practitioners on collaborative think-aloud video analysis tasks. The results indicate that CoUX helped improve the completeness and reliability of their analyses with an effective support for discovering, determining, and discussing UX problems. CoUX allowed them to spot problems that they might otherwise have neglected, and encouraged focused conversations to seek clarification from and respond to their partners.

In summary, we make the following contributions: (1) a video analysis pipeline that extracts multiple acoustic, textual, and visual features from a think-aloud recording to facilitate UX problem identification; (2) a visual analytics tool, CoUX, that supports problem identification, annotation, and collaboration for UX evaluators in an integrated environment; (3) Insights into the results of a user study with six pairs of UX practitioners on collaborative think-aloud video analysis tasks.

2 RELATED WORK

Our work is inspired and informed by related work in three areas: usability testing analysis, machine learning for user experience research, and collaborative visual analysis.

2.1 Usability Test Analysis Tools

Numerous commercial tools have been developed to support UX evaluators with conducting usability test and reviewing test session data. The first category is *offline tools* that need to be installed on local machines, such as Morae [61], Noldus Viso [42], and Silverback [57]. These tools allow UX evaluators to review sessions with functionalities like note-taking and marking events on the video progress bar, on top of basic usability test support such as screen recording, survey administration, and results exporting. However, many offline applications have been retired due to the emerging trend of remote and online user testing platforms [37]. These *online platforms* allow for more flexible collaboration, such as UserTesting.com [62] and FullStory [16]. While these tools support a range of user testing and analysis functions, their data analysis capabilities are limited to session playback, note-taking, tagging, and mouse point clouds. In contrast, we design CoUX to meet the increasing demand for online, remote, and collaborative tools that support usability test session review with advanced analysis support.

In the research community, several prototypes have been developed to facilitate UX problem identification. Usability Problem Inspector [1] was designed for UX evaluators to inspect a test session on the fly and was shown to be effective at helping evaluators find important usability problems in an interface design. However, to better understand the user's behavior and interactions, UX evaluators often have to repeatedly review the usability test recording to pinpoint the problems. Skov and Stage conducted an empirical study of a conceptual tool to demonstrate its usefulness for problem identification with a group of usability evaluators [58]. VA2 [4] supports evaluation session analysis by combining multiple sources of information including interaction logs, think-aloud speech, and eye-tracking data. However, unlike CoUX, collaborative

features and online remote access are not explored. Several other visual analytics tools support better understanding of users' behaviors based on large interaction logs [8, 48]. However, none of them focus on reviewing think-aloud recordings.

In sum, the above tools primarily provide basic functions for analyzing the content of a test session, such as playback, note-taking, tagging, and some user interaction visualization (e.g., click heatmap), and offer limited collaborative features, such as sharing notes or clips. In contrast, CoUX adopts computational methods to extract rich features from the audio, transcript, and video content of the test session and visualizes these features as auxiliary information to better inform the analysis process. Additionally, CoUX considers the specific collaboration needs among UX evaluators such as discussing and resolving conflicts in detecting UX problems and rating the problem severity.

2.2 AI-Assisted UX Data Analysis

Recently, researchers began to leverage artificial intelligence (AI) to assess the usability of digital interfaces [46] and detect UX problems [20, 21, 28, 47]. For example, user interaction events were utilized to create machine learning (ML) classifiers to detect usability issues of websites [20, 47] and virtual reality applications [21]. In addition, user interaction paths were compared to construct graph-based AI models to detect potential UX problems [28]. Although these automatic methods were promising, they were primarily based on users' interaction logs, which only indirectly reflect some aspects of UX problems and lack a true understanding of the UX problems. In contrast, UX evaluators tend to use multi-modal information from both the acoustic and visual channels of a test session to pinpoint and interpret problems [6].

To address the limitations of AI, VisTA is equipped with AI as an assistant by detecting and highlighting video segments containing potential UX problems [13]. It extracts features such as negative segments and abnormal pitches, which are indicators of UX problems [11]. We employ a similar philosophy to overcome the constraints of AI. We further extract the speech and textual features from think-aloud usability test recordings and present them to UX evaluators to assist with their analysis in CoUX. Moreover, we take a step further to extract additional features from the video such as scrolling speed.

Unlike VisTA that is designed to support a single UX evaluator, CoUX is able to support both individual analysis and collaboration among UX evaluators.

2.3 Collaborative Visual Analysis Tools

One critical challenge for UX problem detection is the vague evaluation procedures, which can lead to bias or unclear problem criteria [23]. Thus, different UX evaluators could detect different sets of problems when assessing the same interface, known as the *evaluator effect* [23]. Most evaluators perceive this effect when merging their individual findings with teams [24]. Thus, collaboration and involvement amongst UX evaluators are integral to both increasing the reliability [23] and improving completeness of the problems identified [55]. However, few systems have been developed to adequately support collaborative analysis of usability test sessions. When designing CoUX, we strive to support UX evaluators' collaboration for detecting problems, annotating or assessing problem severity using usability heuristics [40], and initiating discussion in one integrated environment.

Moreover, the design of CoUX draws on insights from both co-located (e.g., [26, 35]) and distributed (e.g., [22, 51, 63–65, 68]) collaborative visualization tools, while these tools do not focus on analyzing think-aloud sessions. In particular, we are inspired by prior work on the support of coordination and synthesis in collaborative analysis activities. Robinson explored the co-located synthesis of findings from paired participants after each had completed an asynchronous individual analysis phase [50]. They found that establishing common ground and role assignment are critical aspects of collaborative synthesis. Mahyar and Tory extended this concept to link common work within a visualization tool to support collaborative sensemaking of documents [35]. CoUX follows these principles by employing both an individual and a collaborative analysis modes, further with the ability to merge problem annotations and severity ratings, helping establish common ground.

Visualizing the analysis history is another strategy for coordination and synthesis, especially in asynchronous collaboration. Sarvghad et al. found that collaborative data analysis can benefit from displaying data dimension coverage of history [51, 52]. Similarly, KTGraph highlights of previously investigated data in a graph visualization to support collaboration [68]. CoUX supports coordination by showing previously annotated UX problems on a video timeline. Also, visual cues of segments of the video timeline are changed based on the state of the problems identified, such as in the uninitiated or in-progress phases.

Furthermore, allowing analysts to use tags and links to organize their comments and identify others' contributions improves final analytic results [65, 67]. Accordingly, CoUX enables user-generated comments and tags for identified problems to explicitly communicate the intent, uncertainty, and progress of their discussion via conversational threads.

3 DESIGN OF COUX

Our main goal is to support UX evaluators in making decisions of usability problems and generating reliable annotations via collaboration. Towards this, we conducted 30-minute semi-structured interviews with two experts (E1 and E2) who are experienced in UX research. E1 is an assistant professor in information science at a university, whose research applies mainly qualitative methods. They complete the majority of their data analysis through Google Sheets [19]. E2 is a UX researcher at a start-up company with over four years of experience practicing UX. As part of his daily job, he uses Zoom [69] and Gong.io [18] to conduct and analyze user evaluation sessions. The goal of these interviews was to understand the current practices and challenges of UX evaluators in analyzing video-recorded usability test sessions and assess their needs for a new collaborative decision making and video analysis tool.

3.1 Design Considerations

Based on our interview findings and prior work, we derived the design considerations for CoUX.

D1: Leverage various information about the video to enhance the robustness of problem identification. Research has indicated that users tend to verbalize their thoughts with abnormal speech features (e.g., abnormal loudness, pitch, and speech rate) when they encounter problems [11]; Their verbalizations also contain more negative sentiments, questions, and verbal fillers [11]. Moreover, UX evaluators can identify more usability problems when these features are presented during analysis [13]. When discussing her video analysis strategies, E1 said: *"I do analyze the speech features but I don't have a good automatic tool to do so."* E2 also mentioned that he observes *"hesitation and pauses in users' speech"* to decide whether they encounter a usability problem. Furthermore, UX evaluators also correlate these verbalizations with the visual content of the recordings. In an international survey of UX professionals, 95% of them believed that the user's actions (e.g. scrolling on the interface, pressing the wrong button) were helpful in identifying usability problems [12]. CoUX supports these needs for determining UX problems by employing machine learning to automatically extract acoustic, textual, and visual features from the recording, which are then presented collectively on its interface.

D2: Provide an integrated environment for both video review and problem logging to ease the problem annotation. In addition to displaying useful information, it is critical to provide a seamless user interface for both video review and problem annotation. Previous studies have shown that UX evaluators often have to review recordings and take notes in separate applications, such as spreadsheets, text editors, and presentation tools [14]. This finding was echoed by E1 who usually stores all the videos in a separate folder while all the analysis and coding is done on a spreadsheet. As a result, she finds that *"organizing and sorting through the files has been tricky."* E2 experiences a similar problem as he reviews the videos on Zoom cloud recordings but keeps his annotations in a separate document. Using separate applications leads to difficulty when trying to pinpoint specific problems during discussions. E1 said that *"we don't have a way to solve timestamps so we just have to manually track it down and put it on a cell and then when we want to review it, we have to retreat to that specific segment in the video."* E2 mentioned *"sometimes the design*

lead wants to see exactly how the user reacted so I need an easier way to show her the snippet of the recording." To address these challenges, CoUX provides an integrated environment with both video reviewing and problem logging functions, allowing UX evaluators to become more organized and efficient during usability test video analysis.

D3: Support collaboration between UX evaluators with both individual and collaboration modes. UX evaluators may have their own biases and limitations when analyzing usability problems, which is known as the "evaluator effect" [23]. Thus, it is important to have multiple evaluators collaborate with each other. Indeed, collaboration amongst evaluators has been found to enhance both the reliability [23] and thoroughness [55] of the problems identified. To serve this purpose, collaboration typically happens among two or more evaluators who first perform independent analysis of the same data [14, 23]. E1 stated that she and at least one other coder would annotate the same video individually by hiding the columns on a spreadsheet. E2 also described reviewing the video individually at first before sharing results with colleagues, which is in line with this best practices process. We aim to design CoUX by following this workflow with two modes: an individual mode for independent problem identification and a collaboration mode for problem merging, decision making, and discussion. This mitigates the confirmation bias since evaluators rely on their own judgment for initial assessments and decisions before seeing others' results.

D4: Allow for both synchronous and asynchronous communication between UX evaluators. Maintaining effective communication between UX evaluators is critical to achieve successful collaboration during the analysis of usability problems. Research has shown that the most frequent form of collaboration is short discussions at the outset of analysis [14]. This was reiterated by E1: *"after we finished coding, we'll highlight the disagreements and then during our meeting time we'll discuss and resolve those highlights."* E2 also mentioned that he discusses the results with the team in short meetings after the session. This type of synchronous communication should be supported by CoUX, e.g., with an instant messaging feature. Further, in the event that a synchronous meeting is not possible, which is not uncommon in practice, E1 and her collaborators would leave comments on the spreadsheet and tag the other person. Thus, asynchronous communication should also be supported to allow the messages to be viewed and discussed at a later time. Thus, we aim to adopt a similar workflow where UX evaluators can discuss and decide both synchronously and asynchronously using comments in a thread and consolidate their opinions using interactive visual support from CoUX.

4 COUX SYSTEM

4.1 System Overview

We developed the CoUX system based on the aforementioned design considerations. As shown in Fig. 1, CoUX consists of a back-end storage & processing and a front-end visual interface, both of which require data extracted from a video analysis engine.

The video analysis engine contains three modules for extracting different types of features from the session recording, including the *Acoustic*, *Textual*, and *Visual* Analyzers (**D1**). The outputs of the video analysis engine are uploaded into the *Session Data & Features* storage hosted in the back-end. The back-end also contains the *Problem Annotations* and *Interaction & Collaboration* storage. The Problem Annotations storage saves all the inputs from UX evaluators regarding the usability problems, while the Interaction & Collaboration storage supports all the actions that the UX evaluators perform in the front-end.

The front-end is composed of three interactively coordinated views: the *Video Player*, *Feature Panel*, and *Problem Panel*. The Video Player allows UX evaluators to play, pause, and rewind the session recording, as well as view a timeline of their annotations above the video progress bar. The Feature Panel presents all the extracted features and highlights the ones that correspond to the current timestamp of the video. Lastly, the Problem Panel allows UX evaluators to enter descriptions of problems that they identified, the design heuristics or principles violated (e.g., Nielsen's heuristics [38], Norman's principles [44]), custom tags, and their severity ratings [40]. The interface includes a toggle for UX evaluators to switch between *individual* and *collaboration* modes (**D3**).

In the individual mode, the Problem Panel displays the comments entered by a single UX evaluator. In the collaboration mode, the Problem Panel also displays the comments of other UX evaluators and allows for both synchronous and asynchronous communication through the chat functionality (**D4**). The three above views together are shown on the same CoUX interface, which provides UX evaluators an integrated environment for both video review and problem annotations (**D2**).

4.2 Video Analysis and User Feature Extraction

To assist UX evaluators with thorough identification of usability problems, CoUX analyzes think-aloud videos by segmenting them into small meaningful chunks and extracting various features related to the user in the video (**D1**). The video segments are automatically detected using the Auditok library [56] at periods of silence characterized by the lack of acoustic activity. By doing so, the entire long video is cut into small "bite-size" portions to facilitate UX evaluators' analysis, each of which may correspond to one or few usability problems. Each segment is then transcribed using the Google Speech Recognition API [66]. The audio, transcript, and video of the segments are used to extract three main categories of user features: acoustic, textual, and visual.

- **Pitch:** Users tend to change their pitch when they encounter a problem while thinking aloud [11]. For the corresponding audio of each segment, we computed the frequency of the speech using the "sound to speech function" in the Praat-Parselmouth library [27]. Based on the mean and the standard deviation of the pitch over the entire session, a segment is categorized as containing abnormal pitch if at least 10% of the values are over two standard deviations away from the mean. Thus, it is given one of the three values: 1 for abnormally high, 0 for normal, and -1 for abnormally low.
- **Loudness:** Loudness has been shown as another useful speech feature for analyzing usability test sessions [7]. We utilized the "sound to intensity" function in Praat-Parselmouth [27] to extract the intensity of the sound (in dB). The detection of abnormalities and assigned values are the same as the pitch feature.
- **Speech Rate:** We computed the speech rate by dividing the number of words spoken in a segment by its duration, where the number of words was counted based on the transcript. Only abnormally slow speech is detected based on prior research showing that users slow down when encountering an issue [11]. Thus, each segment is labelled 1 for abnormally slow or 0 for normal.
- **Negations:** Negations in users' think-aloud verbalizations may indicate that they encounter a usability problem [11]. To determine if users said a negation, we applied a keyword-matching to the transcripts to detect the following words: *no*, *not*, *don't*, *doesn't*, *didn't*, *can't* and *never* [10, 11].
- **Questions:** Questions are another type of indicator for usability problems, indicating a user may be in doubt. Similar to negations, we utilized a keyword-matching algorithm containing the following words: *what*, *which*, *why*, *how*, and *where* [10, 11].
- **Verbal Fillers:** Verbal fillers indicate hesitations in the user's speech, which may demonstrate a problem. We also utilized a keyword-matching algorithm containing the words: *um*, *uh*, and *like* [10, 11].
- **Sentiment:** The sentiment of a user's speech (e.g., positive, neutral, or negative) is another source of useful information for problem identification [11, 17, 60]. We used the Valence Aware Dictionary and Sentiment Reasoner library [25] to detect the sentiment based on the transcripts for each video segment. Based on the compound score (between -1 and 1), a segment is labelled as positive $((0.2, 1])$, negative $([-1, -0.2))$, or neutral $([-0.2, 0.2])$.
- **Scrolling Speed:** When using a digital product, the amount of scrolling may reflect a user's confusion. For example, frequently scrolling back and forth on a webpage could indicate that a user has difficulty in understanding the interface [3]. Thus, we extracted the scrolling speed (in the amount of pixel movement per second) for each segment using the dense optical flow algorithm from OpenCV [45], resulting in a continuous time-series.
- **Scene Break:** Frequent switching of views may also indicate that the user has difficulty locating the desired item on a digital interface [3, 20]. We used the OpenCV-based video scene detection library [5],

which performs a comparison of sequential frames in a video and detects substantial changes in content. This results in a series of timestamps of these scene breaks.

These features are meant to provide extra information to help UX evaluators review think-aloud sessions of digital products and make decisions regarding usability problems. The features are selected based on our interviews and the literature as mentioned above. However, whether this feature set is complete or not is still an open question.

4.3 CoUX User Interface

For better work organization, CoUX features a project management page showing all the videos that need to be analyzed upon logging into the system. Clicking on any video opens the main CoUX interface. This interface consists of three key components (Fig. 2): (a) a Video Player for viewing the recorded think-aloud sessions, (b) a Feature Panel for displaying various extracted features based on the analysis in Sec. 4.2, and (c) a Problem Panel for logging discovered usability problems and discussing them with other UX evaluators.

4.3.1 Problem Identification

Effectively identifying potential UX problems is the key objective of reviewing a think-aloud video. On the left, CoUX comprises all necessary elements for problem identification based on various information extracted from the video (**D1**). First of all, an integrated video player (Fig. 2-A) is provided to prevent any switching between different tools, which is the largest element on the screen to facilitate the video browsing. The player supports all regular functionalities like play, pause, forward, and rewind. Further, similar to the YouTube chaptered design, the player progress bar shows the automatically-generated segments (Fig. 2-a1) that split the video into "bite sizes" (see Sec. 4.2).

Below the player, a couple of visualizations are placed on the Feature Panel (Fig. 2-B) to facilitate the use of all the extracted features while reviewing the video. CoUX distinguishes discrete and continuous features, and displays them on two sub-panels. First, discrete features (i.e., all the acoustic and textual features) are visualized in the Feature Matrix (Fig. 2-b1), where rows indicate the features and columns represent the video segments. All values in the matrix are shown as icons and colors instead of text to allow UX evaluators to quickly scan and recognize the feature values that could signify a problem. For example, represent neutral, negative, and positive sentiments; represent filler words (e.g., um, uh), negations, and questions; and represent high and low anomalies. Second, continuous features (i.e., the visual features) are shown in a Feature Chart (Fig. 2-b2), where the scrolling speed is implemented as a line chart and the scene breaks are represented as vertical green lines.

These features serve as auxiliary data for the video to enhance the thoroughness of problem identification by UX evaluators. While the video is playing, CoUX dynamically highlights the corresponding segments in both the Feature Matrix and Feature Chart, with a lighter blue background. In addition, a red vertical line representing the playhead moves on the Feature Chart while the video is playing. In contrast, the column width of the Feature Matrix does not reflect the time length of each segment (instead, a fixed width). Thus, a Sankey visualization [49] (Fig. 2-b3) is placed between the player progress bar and the matrix to indicate the correspondence. Similarly, a red curve is shown on the Sankey to indicate the playhead. This design increases the readability and scalability; if each column width of the Feature Matrix maps exactly to the segment length, some columns could be too narrow to display any readable features whereas others could be very wide, wasting the space. Lastly, all the above visualizations are clickable, which facilitates navigation to different parts of the video.

4.3.2 Problem Annotation

Once an evaluator identifies a UX problem, they can log the problem with the Problem Panel (Fig. 2-C), integrated seamlessly within CoUX (**D2**). When an evaluator starts to type in the chatbox-like interface at the bottom of the panel (Fig. 2-c1), the video automatically pauses so that they do not need to manually click the video controls. Annotations can be bound to video playtime by checking the time check
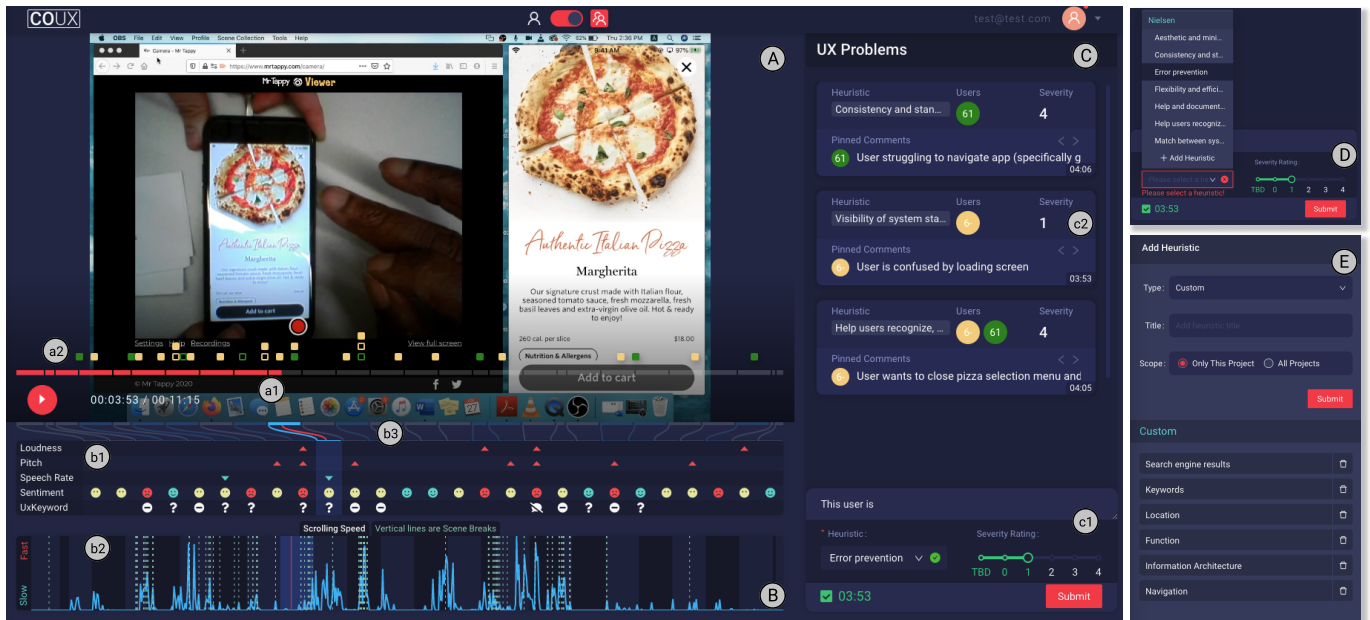
Fig. 2. The CoUX user interface, showing a realistic study session of two UX evaluators (see Sec. 5) analyzing a think-aloud video recording of a food delivery mobile app: (A) a Video Player for viewing the video; (B) a Feature Panel for displaying various extracted features to assist the analysis; and (C) a Problem Panel for logging discovered usability problems and discussion. (D) Problem annotation via a dropdown for common heuristics tags (e.g., Nielsen heuristics [38] and Norman principles [44]) and a slider for problem severity rating [40]. (E) A popup panel for adding custom tags.
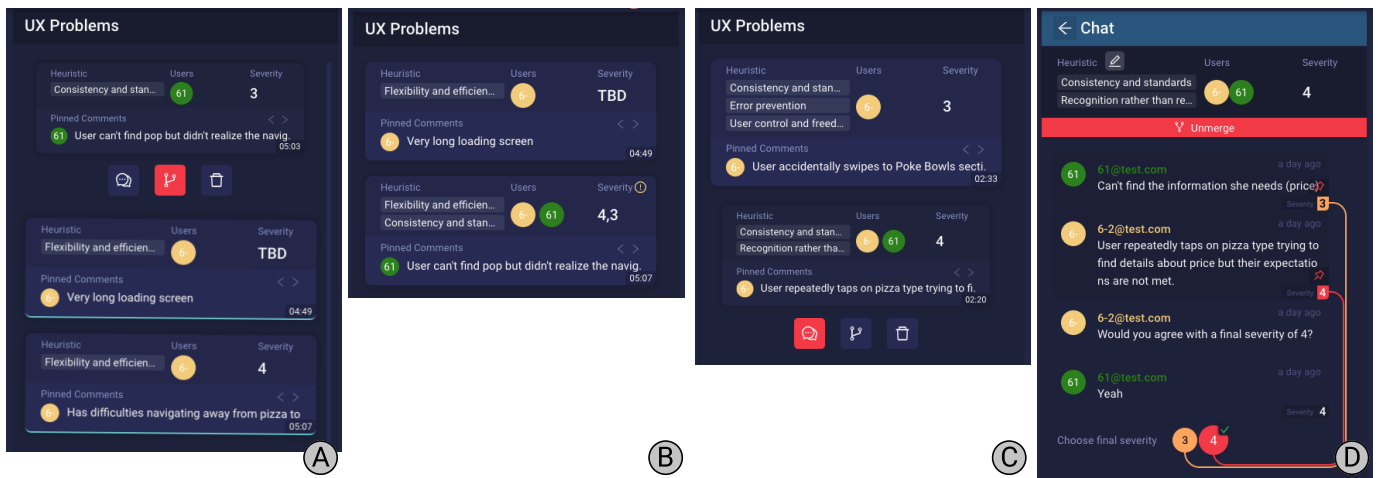

Fig. 3. CoUX supports the collaboration among UX evaluators via a chat thread design: (A) merging two Annotation Cards, (B) merged results, (C) showing the conversation between a pair of UX evaluators, and (B) the Discussion Panel of the selected card.

box ✅ 00:52 . Evaluators can add comments or descriptions for the identified problem, and select predefined heuristic tags from a grouped dropdown list and a severity level (0–4, where 4 indicates the highest severity) [40] with a slider (Fig. 2-D). CoUX supports common tags including Nielsen's heuristics [38] and Norman's principles [44]. Moreover, evaluators can add their custom tags via a popup panel (Fig. 2-E). These tags can be created within custom groups and set to either applicable to a specific video or all videos in a project.

After an annotation is submitted, CoUX adds an Annotation Card (Fig. 2-c2) to the Problem Panel, which displays all the cards bound to the active video segment. Each Annotation Card shows the problem tags, severity, comments/descriptions, and corresponding evaluators. Moreover, the Annotation Timeline (Fig. 2-a2) updates with a new solid Annotation Square 🟩 pinned onto the video progress bar, which shows an overview of all problems with colors indicating their creators.

4.3.3 Problem Discussion and Collaboration

CoUX supports an individual mode and a collaborative mode to mitigate the "evaluator effect" (D3). In the individual mode, an evaluator can oversee their own Annotation Cards and Squares. When switching to the collaborative mode with the mode button 🧑🔴🧑 on the top, evaluators can navigate to each other's identified problems by simply clicking the corresponding elements and start a discussion to consolidate their annotations. Evaluators can still create new problem annotations in this collaborative mode. The discussion/consolidation is moderated via chat threads, similar to Slack [59], to support both synchronous and asynchronous collaboration (D4).

As it is possible that evaluators created different annotations about the same underlying problem during the individual mode, CoUX allows them to merge the Annotation Cards. To do so, an evaluator first clicks a card and then three buttons pop up: discuss 💬, merge 🍴, and delete 🗑 (Fig. 3-A). When the "merge" button is clicked, a shaking animation highlights mergeable cards. In addition to a merged Annotation Card (Fig. 3-B), a new Annotation Square is added on the Annotation Timeline while the previous squares become hollow ⬜. Currently, merging is only allowed for problems in the same video segment.

Moreover, clicking the "discuss" button 💬 enters an in-situ Discussion Panel (Fig. 3-D) of this Annotation Card (Fig. 3-C). They can then add new comments and propose a different severity rating for the problem, or discuss the merging if applicable. New and existing

comments are displayed based on their timestamps. Evaluators can also pin important comments. A thread visualization helps evaluators review all the proposed severity ratings (Fig. 3-D). If an annotation has a conflict in the severity rating (i.e., more than one severity ratings are proposed), evaluators are asked to determine the final severity for the annotation; otherwise, this problem remains unresolved, with a warning icon `Severity ⓘ` associated with the Annotation Card. Evaluators can also add or remove heuristics by clicking on the edit button `Heuristic ✎` on the top of the panel. These Annotation Cards on the Problem Panel provide an informative summary about a problem. Each card shows all the tags, severity ratings, participating evaluators, and pinned comments in a carousel view (Fig. 3-C). For merged cards, evaluators can also unmerge them through a button `⑂ Unmerge`.

5 USER STUDY

To investigate how UX evaluators would use CoUX in their analysis of usability test videos, we conducted a user study to assess the usefulness and effectiveness of the system functions. Specifically, our exploration was guided by: **RQ1** - How does CoUX support evaluators in analyzing think-aloud sessions? **RQ2** - How do teams work together and communicate during their analysis through CoUX? **RQ3** - What are the general challenges in collaborative UX video analysis?

5.1 Participants and Apparatus

We recruited 12 participants (two males, nine females, and one not disclosed, aged 23–32) via social media and mailing lists. They were UX designers ($N = 4$), UX researchers ($N = 4$), and UX/HCI graduate students ($N = 4$). On average, they had three years of experience in UX ($SD = 2.2$). Eleven (91.7%) self-reported being very familiar or extremely familiar with identifying usability problems, with one participant being moderately familiar ($M = 4.17, SD = 0.55$). The participants were recruited in pairs. They had all collaborated with their partners before on at least one project. Seven (58.3%) were very or extremely familiar with their partner, with the rest being moderately familiar ($M = 3.83, SD = 0.80$).

Participants completed the study remotely with their own computers while communicating with the moderator through video-conferencing software. Participants were asked to make the application window full screen throughout the study. Participants used the largest screen available. The average display size was 20 inches ($SD = 7.12$).

5.2 Study Videos

We collected two recorded usability test sessions in which users were instructed to use digital products with the think-aloud protocol. In the practice video (length: 3 minutes 34 seconds), a user was asked to find a photo of an instruction manual for an early telescope on a Science and Technology Museum's website. In the study video (length: 11 minutes 15 seconds), a user was asked to complete three tasks on a Food Delivery Mobile App, including: (1) find the Wegmans store on the Amherst St.; (2) buy 10 bottles of classic Coke and 10 bottles of Sprite, and some full sheet pizzas with any topping while staying under a budget of $100; and (3) change the pick up order to delivery instead. These videos were chosen since they are representative of digital interfaces: one for a desktop website and the other for a smartphone application. There were also numerous usability issues in both videos which promoted discussions between the participants and their partners.

5.3 Task and Design

Each pair of participants conducted the study together and was asked to review the study videos and identify usability problems using CoUX. There were two phases in the study session: (1) an *Individual* phase and (2) a *Collaborative* phase. In the individual phase, participants identified usability problems and submitted the annotations of these problems independently. In the CoUX interface, they could only see the problem cards that they had inputted. In the collaborative phase, the problem annotations of both partners were revealed to each other. Then, they were asked to review each other's annotations, merge cards as desired, and discuss the problems before reaching a final decision. Splitting the session into two phases was based on the recommendation

Table 1. Usage statistics of various functions in CoUX.

Function Usage	Mean (SD)
Clicks on Feature Matrix (Fig. 2-b1)	7.6 (7.1)
Clicks on Feature Chart (Fig. 2-b2)	1.2 (2.2)
Number of problem annotations (Fig. 2-c1)	18.3 (7.0)
Number of problem merges (Fig. 3-A)	1.9 (1.7)
Number of comments per chat thread (Fig. 3-D)	2.8 (0.7)
Number of discussed problem annotations	6 (3.6)
Number of additional problems found after collaboration	4.2 (2.7)

that to serve the purpose of improving reliability, collaboration should happen among two or more usability practitioners who first perform independent analysis of the same dataset [14].

5.4 Procedure

To begin, each pair was given a short video tutorial about CoUX. Participants were able to ask any questions about the study and the system. They were then introduced to the usability test video review task, and instructed to assume that developers of the products will have limited time to address the problems identified in the session. This assumption resembled the fact that UX practitioners often have limited time to analyze test sessions [12, 43] and allowed for a more realistic evaluation of the extracted features and collaboration support in CoUX.

After the tutorial, the participants completed a practice trial by first analyzing the museum video individually for five minutes, then collaborating with their partner for another five minutes. This allowed them to become familiar with the system and the full procedure of the two-phase task. In the study session, participants were first asked to identify usability problems with the food delivery app individually for 25 minutes and then filled out a short survey based on the 5-point Likert Scale, which sought to understand the usefulness of each feature and the ease of use of the annotation functionality in CoUX. After a short break, they had 15 minutes for the *Collaboration* phase where they discussed each other's problems and tried to consolidate them into a final set. At the end, each pair of participants independently completed another short survey about their collaboration experiences. These survey questions were based on previous findings about collaborative analysis [23, 55]. When performing both the individual and collaborative tasks, participants were asked to communicate only within CoUX. This would allow them to fully explore and use CoUX during the study. We then conducted a semi-structured group interview to collect their feedback about the system. All the interview sessions were video-recorded, and participants' interactions with the system (e.g., clicks, video-playing behaviors) were logged. The study lasted about 90–100 minutes and participants received $25.

6 RESULTS

We first present participants' general user experience of CoUX (Sec. 6.1) and then how they used the features during their individual analysis (Sec. 6.2) and collaboration (Sec. 6.3) respectively, based on our RQs. Participant x in the study pair n is labeled as *Pn-x*.

6.1 General User Experience

Overall, participants felt that CoUX was a useful tool to support their analysis of a usability test video recording. Fig. 4 presents participants' ratings on different aspects of CoUX. They agreed that they could find all the functionalities easily in the interface to perform their tasks ($Md = 4, IQR = 1$). Table 1 shows the usage statistics of the main functions of CoUX. This suggests that all functions were used by participants, in particular the extracted features, problem annotations, and chat threads, which will be explained in the following sections.

Moreover, participants appreciated that CoUX integrated analytics, collaboration, and communication features together in one integrated environment. *"Usually we were using Google sheets [19] to coordinate and it was getting quite difficult, because we had to follow up with another person... it was messy and difficult but right now, it seems quite easy [with CoUX]."-P2-1* Eleven (91.7%) participants agreed or strongly agreed to recommend CoUX to others ($Md = 4, IQR = 0$).

Q1. Find system functions quickly   1 7 4   **Q2. Sentiment is useful**   3 2 5 2   **Q3. UX Keyword is useful**   2 2 4 2 2   **Q4. Loudness is useful**   6 5 1   **Q5. Pitch is useful**   2 4 4 2

Q6. Speech rate is useful   2 3 4 3   **Q7. Scrolling speed is useful**   1 4 4 3   **Q8. Scene breaks are useful**   4 2 3 2 1   **Q9. Entering annotations is easy**   1 5 5   **Q10. Adding heuristic is easy**   4 2 3 3

Q11. Adding severity rating is easy   1 1 3 7   **Q12. Confident on identified problems**   4 7 1   **Q13. Confident on severity rating**   3 4 5   **Q14. Confident on NO additional problems**   4 3 3 2   **Q15. Decided more easily & quickly**   1 5 5 1

Q16. Decided more fairly & comfortably   2 2 6 2   **Q17. Liked seeing others' annotations**   4 8   **Q18. Identified more problems**   4 8   **Q19. Improved the reliability**   1 4 7   **Q20. Would recommend CoUX to others**   1 9 2
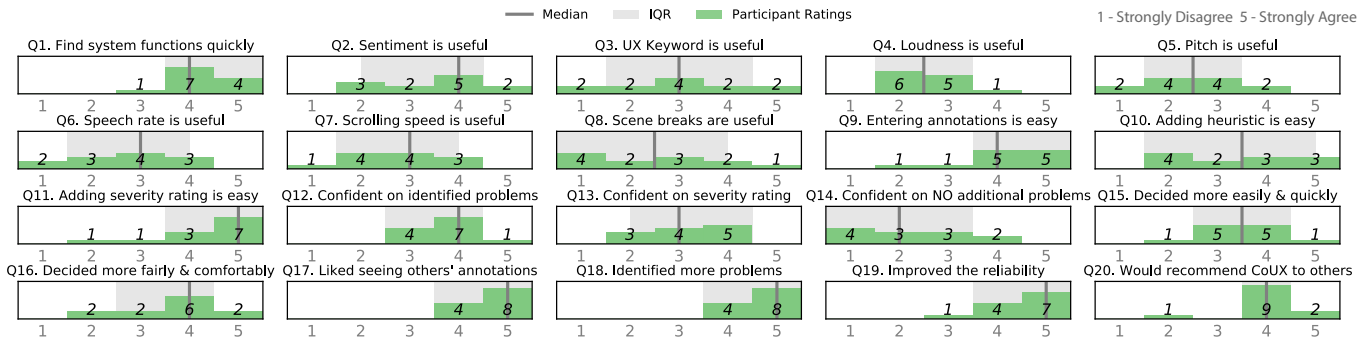
Fig. 4. Participants' questionnaire ratings (Likert 1-5) after the individual phase (Q1-14) and after the collaboration phase (Q15-Q20).

Participants also compared it to previous tools that they had used. *"Actually I use a similar tool just like this [Lookback [34]], but it doesn't offer any analytic information, like loudness or pitch."-P3-1*

6.2 Individual Analysis (RQ1)

6.2.1 Problem Identification: Feature Panel

In general, the Feature Panel supported individual problem identification in three ways. First, participants used it as **hints or warnings** to get alerted about certain segments while reviewing the video. *"The reason I looked at those is more as a hint or warning to see what is coming up. I paid more attention to that segment if there's a red face."-P4-2* Second, they used it as **anticipations of problems** to help better skip ahead without missing important problems when they were under time pressure. *"When I heard... 10 minutes left, for the video that I haven't watched, I picked the more highlighted ones to directly find any problems so that's when I found those markers to be helpful."-P4-1* Lastly, they used it as **checks or anchors** to pinpoint areas to revisit in their second-pass. *"When I finished the whole video I prefer to go back to see the angry face again, so it definitely helped me to pick up on something that I might have missed in the process."-P1-2* This was echoed by *P5-1*: *"I used that panel (feature matrix) to see if there is anything I missed, I think they just confirm that I found the major areas of concerns that the software interface had."*

Sentiment was rated as the most useful feature ($Md = 4, IQR = 1.25$). One reason was that this feature was perceived to be accurate by the participants. *"I feel like this is providing useful insight and... the sentiment easily got me to the areas in the video where the user was confused."-P5-1* Another reason was that *"the icons and sentiments are very easy to understand."-P4-1* One (*P3-2*) mentioned that even though the negative sentiments often point them to areas with a usability problem (*"I think 90% of them are accurate"*), they were still cautious about using the neutral sentiments and relied on their own judgment for those segments. Thus, human judgment was still exercised as *P6-2* pointed out: *"It was showing a neutral sentiment but I thought the user was actually very frustrated... so I still relied on my own intuition."*

UX keywords were generally viewed positively and rated as the second-most useful features ($Md = 3, IQR = 2$), and were perceived to be accurate. For example, *"The UX keywords matched up with what my impression was while watching the video."-P5-1* Also, *P6-1* appreciated the accuracy of the question marks, e.g., clicking on it navigated *P6-1* back to the segment where the user of the app was confused.

Speech features, i.e., loudness ($Md = 2.5, IQR = 1$)**, pitch** ($Md = 2.5, IQR = 1$)**, and speech rate** ($Md = 3, IQR = 1.25$), were perceived as less reliable than sentiment and UX keywords. Participants thought that these features could be augmented with the context and other information. *"The pitch was interesting, but I feel like I still have to listen to a combination of their tone and the context."-P6-2* Moreover, participants felt these features were new and needed more explanations about how they were determined. *"I'm not sure I can trust this stuff because I'm not sure [they] were based on what logic."-P1-2*

Scrolling speed ($Md = 3, IQR = 2.5$) **and scene breaks** ($Md = 2.5, IQR = 2.25$) were appreciated by some participants who used the peaks as possible indicators of users' confusion. *"I actually almost paid all of my attention on the scrolling speed. Compared to the icons,*

I definitely prefer to look at the visualizations."-P1-2 *"I looked at the peak of the waves of the scrolling speed just to double-check what was happening."-P6-1* Also, *P3-1* thought the scrolling speed was useful, but would like to see numbers instead of a relative scale from "slow" to "fast." The scene breaks were used as *"indicators of task changes to skip through or go back to review further."-P6-1*

However, these two features were relatively less used. One reason was that *"they are too far away from the video so while I'm watching the video I couldn't see that information."-P2-2* In addition, *P4-2* mentioned *"I didn't really look at the scrolling speed, because I think this is not super relevant to this task."*

Interestingly, participants also used a **combination of different features** to better locate segments that contained usability problems. *"In the same column, if there were two or more icons that show abnormalities, I paid more attention to it because it's more obvious."-P1-2* The above observations confirmed the success of **D1**, which was to leverage various information about the video to support evaluators' analysis.

6.2.2 Problem Annotation: Problem Panel

Participants used the **problem annotation** function of CoUX (Fig. 3) extensively as they recorded the usability problems in this area. On average, participants entered 18.3 problems in this recording ($SD = 7.0$). Overall, participants felt that it was the most important component of the interface and entering annotations was *"pretty clear and straightforward"* ($Md = 4, IQR = 1$). *"The chat box... is really useful because it's very clear and very easy to use."-P1-1*

In addition, they liked the functionality to attach **heuristics** ($Md = 3.5, IQR = 2.25$) and **severity rating** ($Md = 5, IQR = 1$) to each problem description. *"I like the heuristic function that I can select from the heuristics which are already there and I can add my own options as well, so that was also very helpful."-P2-1* However, we also received some mixed feedback about having to attach a heuristic to every annotation. *"I think the heuristics are great, but I don't think it should be mandatory. I usually make notes of activities and those aren't things that I would tie to a heuristic."-P6-2* In this case, the system could be modified to allow heuristics to be optional.

Five out of the six pairs used the **custom tagging** function extensively as it allowed them to add tags that were *"more relevant to the actual video, like here there is the older adult and accessibility issues that are more specific than Nielsen's."-P2-1* Below are some custom tags participants added: "Information Architecture", "Navigation", "Test Condition", "Older adults preferences." These custom tags were both concrete and diverse, reflecting their unique experiences and expertise.

The above feedback also demonstrates the benefits of CoUX by providing an integrated platform to assist problem annotation, as guided by **D2**. *"We can finish the analysis and make the comments in one screen instead of using lots of applications."-P2-2*

6.3 Collaborative Analysis (RQ2)

6.3.1 Effects of Collaboration

Overall, participants felt that this collaborative session was an important component of their analysis process. They felt that it allowed them to reach decisions more easily and quickly ($Md = 3.5, IQR = 1$), as well as more fairly and comfortably ($Md = 4, IQR = 1$).

Participants pointed out two main benefits from the collaboration support. First, collaboration helped them **improve the *completeness* of their results** ($Md = 5, IQR = 1$). As shown in Table 1, participants identified on average 4.2 more problems after the collaboration phase. This is 39.3% more compared to the number of problems they identified in the individual analysis. *"Some of the problems I didn't recognize but she did, so her annotation reminded me that here is a problem, and I can write the feedback on it."-P1-1* Second, collaboration allowed them to **improve the *reliability* of their results** ($Md = 5, IQR = 1$). On average, participants commented on 6 (57.1%) problems that they had not previously identified (Table 1), demonstrating that they conducted a more robust analysis. *"To obtain the most unbiased feedback and the most unique ideas without biasing each other, this was really helpful."-P5-1* Similarly, having more people analyze the problem allowed for different perspectives of the issues to be explored. *"We actually noticed that the same area in the video has problems, but we focused on different aspects of the problem."-P3-2* One participant was unsure about the severity rating, but looking at her partner's annotation of the same problem gave her *"confidence that three is a good rating for this issue and it's not too high."-P2-1* Such feedback demonstrates the successful implementation of **D3**, which was to support collaboration between UX evaluators with both individual and collaboration modes for purposes of improving completeness and reliability.

6.3.2 Usage of Collaboration Support

Annotation Timeline. In the collaborative phase, participants viewed and discussed each other's annotations of usability problems. They liked **seeing the problem annotations of their partner** ($Md = 5, IQR = 1$) and used the annotation timeline extensively to **navigate to each annotation**. *"For the collaborative session, I pretty much clicked on all the boxes (annotation squares) to get what I needed."-P5-1* Participants also used the blinking annotation squares (Fig.2-a2) as indicators: *"In terms of the flashing squares, I would go through and check them for new changes."-P6-2*

Problem Merging. Participants used the merging function when they identified the same problem in the same segment. This allowed them to **have a focused conversation** regarding a certain problem in one place. On average, each pair merged 1.9 problems ($SD = 1.7$) in the test video (Table 1). *"At the three-minute timestamp, there were two cards where we were exactly talking about the same issue of the [user] tapping on 'My Cart' and it was not responsive, so I merged them together."-P6-1* Some participants mentioned that they might need to merge multiple problem annotations of the same underlying problem in different segments as one overarching usability issue.

Chat Threads. For each usability problem, participants left on average 2.8 comments ($SD = 0.7$) in each thread (Table 1). Participants utilized the chat threads (Fig. 3-D) in four ways. First, they used it as a **record or documentation** of their discussion. Having the discussion on the same interface as the video allowed the participants to review the particular segments for the usability problems that their partners identified. *"We can see the process of what we discussed and what we talked about so we won't forget what we say and review the video at the same time."-P1-1* *"The threads are like a place where we can document the final decision."-P3-1* *"This is really useful for when we want to have a clear trail of the analysis."-P6-1*

Second, participants used it as a means to **seeking clarification from their partners** on their usability problem descriptions. For example, *P2-2* left comments like *"Can you explain more about this problem?"* In another case, *P6-1* mentioned a usability problem about *"the scrolling area is limited without clear color indication."* p6-2 then utilized the thread to reply *"What do you mean by this? Do you mean on screen?"* In this way, they were able to sync up on the specific element of the interface that caused the problem.

Third, it was utilized for **consolidating the heuristics and severity rating**. For example, *P3-1* commented *"The user was looking for pop but missed the 'drink' section, maybe the word 'drink' is not associated with the word 'pop' in her point of view"* and tagged it with the "consistency and standards" heuristic. *P3-2* found the same issue but offered an alternative heuristic, *"This could be because there are no drink pictures on this page"* and tagged it with the "visibility" heuristic. *P3-2* then utilized the chat to communicate with *P3-1*: *"I agree the severity is 2, but it is more of a visibility issue..."* After *P3-1* agreed, they removed "consistency" and selected the final severity rating.

Lastly, these threads were viewed as a **precursor to video-call/in-person meeting**. *"This is good for me to quickly see what my partner and I agreed on and then we can skip those in the meeting."-P4-2*

Collaboration Modes. Although participants collaborated *synchronously* in the study, they pointed out that the tool would also allow them to collaborate **asynchronously** by both leaving notes and following up with their partners' annotations in the corresponding thread whenever they had time. *"A beneficial scenario would be if we're working in different countries so we can't analyze and talk in real-time."-P6-2* Further, participants mentioned additional functions to support asynchronous collaboration, such as revision history (*P1-1, 2-1, 3-1*) and e-mail notifications of new comments (*P1-2, 2-2, 4-2*).

These results confirm the support of **D3** and **D4** by CoUX, via enabling effective and seamless collaboration on UX problem analysis.

6.4 Challenges (RQ3)

From this study, we also learned about the potential challenges in collaborative think-aloud video analysis, which could shed light on future research avenues.

Managing Disagreements. Although participants appreciated the support of CoUX for merging problem annotations, they pointed out that managing disagreements is a difficult task in nature because each evaluator has their own interpretations. Evaluators could disagree on **whether there is a usability problem**. For example, Pair 4 had a disagreement on whether a problem was actually an issue with the app interface. *"[P4-1] just put she thought that this app is too much trouble, but from my understanding, it's because maybe she's not familiar with the iOS keyboard."-P4-2* They could also disagree on **the severity of a usability problem**. *"I think collaboratively agreeing on what the final severity rating is more difficult, after this session, I still think there's problems that are still open-ended and unagreed upon."-P5-1* Similarly, Pair 1 disagreed on the interpretation of the actions of the user in the video while she was adjusting the quantity of the drinks. *"So we disagreed on what the user was trying to do but agreed it is the efficiency of use problem."-P1-1* To manage these disagreements, participants usually *"left comments with explanations of... why we think that that is a problem."-P2-1* Although managing disagreements could be time-consuming and difficult, participants believed that discussions, such as those in the chat threads of CoUX, could lead to more robust analysis. *"Having the debates during collaboration are actually where the meat of the analysis is."-P6-1*

Workspace Awareness. Although participants collaborated synchronously in the study, they sometimes worked on different portions of the video and thus missed the "real-time" element of a synchronous collaboration. *"It feels like we're not on the same page, because when I work on the first card, she is probably working on the second card, so we cannot get the real-time feedback [on the same card]."-P1-1* Thus, future work should explore ways to increase workspace awareness.

Chat vs. Conference Calls. Many participants considered chat threads as a "light-weight" communication and enjoyed its flexibility that allowed them to respond whenever they had time. *"I think communicating via the chat is totally fine, I would rather have a new thread for every issue."-P5-2* *"I would prefer this because... I can come back to it and comment whenever I have time instead of getting on a call and having long discussions."-P2-1* However, some participants felt that using a video or voice call could be more efficient for resolving disagreements. *"You'd save more time by just hopping on a quick call rather than keep typing explanations [in chat threads] over and over again."-P5-2* Additionally, some participants also suggested a combination of both depending on the depth of their conversations. *"I think it would only be useful to have a call if we actually really disagreed about something and we couldn't come to a consensus in the chat."-P5-2* Thus, in collaborative interfaces, the trade-offs between light-weight chat threads and more heavy-weight video-calls need to be further explored in the future.

7 DISCUSSION

In this section, we discuss the key lessons and observations from our study and the limitations of our work. We also point out some design implications obtained from the study and potential future directions.

Problem Identification. CoUX visualized behavioral signals indicating usability problems [11, 20, 28, 47] from the acoustic, textual, and visual information of the videos on its Feature Panel. Participants creatively used these features to *become alerted about potential upcoming problems when playing the video*, to *skip less important portions of the video if pressed for time*, and to *facilitate their revisitation of the video in their second-pass analyses*. These usages of the Feature Panel demonstrated the flexibility of CoUX for analyzing usability test videos with different time budgets, which is a common challenge [12, 36, 43].

In general, participants trusted the sentiment and UX keywords more than other features, because they could intuitively draw connections to usability problems. For relatively new features (e.g., speech, scrolling speed, and scene breaks), participants did not fully trust them as they did not clearly see the underlying logic. Also, their existing experience affected their perception of the features; for example, they felt speech features could be affected by an individual's speaking behavior and thus were unreliable. These usage patterns indicate that participants *actively* scrutinized and interpreted the features instead of *passively* accepting them. This is encouraging as it suggests that CoUX supports, rather than replaces, UX practitioners' independent analysis.

Problem Annotation. Participants rated highly about being able to annotate the problems, attach UX heuristics violated, and provide severity ratings all in one integrated system. While existing commercial tools allow UX practitioners to attach problem descriptions while watching a video (e.g., [18, 34, 62]), no such tools provide the ability to attach UX heuristics and severity ratings at the same time, which are two important pieces of information to have when analyzing usability problems [13, 32, 39, 53]. Furthermore, participants enjoyed being able to create their custom tags for UX problems when they felt the standard heuristics were too generic to describe the problems accurately, which echos previous findings [29].

Collaboration. In the collaborative mode, CoUX visualizes the problems identified by their partners, which helped participants catch the missing ones and become more confident about those identified by both. Moreover, CoUX allowed them to have focused conversations regarding whether to merge different interpretations of the same problem. The chat threads supported both synchronous and asynchronous collaboration, and provided a track record of their analysis history, allowing for revisiting how they arrived at a decision. As a result, participants felt their analysis was more complete and reliable, suggesting that CoUX helps UX practitioners reduce their limitations and biases.

Our study also revealed challenges for further investigation. First, while participants felt chat threads would be sufficient in many cases to resolve their disagreements, a video/audio call could avoid back-and-forth textual chats. As the video/audio call is more disruptive, it remains an open question of whether and how to integrate it into CoUX.

Second, displaying the provenance and history of analysis and data is critical to increase the team awareness and collaboration efficiency [35, 51, 65, 67, 68]. While CoUX supports this via various aspects such as the Annotation Squares, chart threads, etc., future work should investigate ways to enhance the capability of CoUX with advanced visualizations, such as graphs [67, 68] and analysis coverage [35, 51].

Third, managing redundant information is important and can be time-consuming. To address this issue, CoUX supports problem annotation merging. Future work can employ machine learning to recommend duplicates of annotations and suggest auto-merging.

Limitations. Our research took a first step to designing an integrated visual analytics tool to support both individual and collaborative analysis of think-aloud videos. Although our exploratory study with six pairs of UX practitioners revealed its potential as well as the analysis patterns, a more comprehensive controlled experiment would allow us to quantitatively compare against existing tools. Potentially, we could use CoUX to analyze the recorded study sessions of itself.

Second, while CoUX visualizes various features from the video, other behavioral signals, such as facial expressions and body language, could suggest usability problems [9, 12, 30]. Meanwhile, certain features, such as the scrolling speed, might only be relevant to certain products, such as mobile apps and websites. Thus, future work should consider extracting other types of behavioral signals and display only relevant features to UX practitioners based on the context. The features could be further configurable between continuous and discrete form, which would give UX practitioners the flexibility to switch as desired.

Third, we used one think-aloud video for our exploratory study (in addition to one video for training), which was of specified length and only included tasks related to navigating an application. For longer videos, UX evaluators may use the Feature Panel more often to quickly navigate to segments with indicators of usability problems. They may also need to conduct their analysis in multiple rounds, resulting in more asynchronous communication. We also found potential misconceptions between the perceived usefulness of various extracted features and the actual performance suggested by the literature (e.g., [11]). Future work should collect more usability test videos for quantitatively assessing the features and understanding the potential misconceptions. As the accuracy of the features could affect users' impressions and usage of an AI [13, 31], we could encode and visualize the uncertainty of the features if the accuracy is low (e.g., using the color transparency).

Lastly, although our study only evaluated the collaboration between pairs of UX practitioners, CoUX allows three or more UX practitioners to collaborate simultaneously. It is, however, still an open question of what challenges UX practitioners might encounter when collaborating with three or more colleagues.

Design Implications. Our study generates several design implications for developing future collaborative analysis tools for UX evaluators. First, as discussed earlier, our participants had concerns about speech features and other new features to them, while our algorithm already accounted for individual's speaking behaviors, etc. Thus, when employing machine learning to assist the decision making, it is necessary to best convey the meanings of extracted features, in particular when such meanings are counter-intuitive, to UX practitioners.

Second, as UX practitioners continue to use CoUX, it will be able to accumulate their custom problem tags over time and even suggest relevant tags for them to consider. These custom tags could reflect on UX practitioners' experience and expertise, and future systems should promote such customization and even support sharing custom tags with their colleagues to complement each other's analysis.

Third, in some cases, participants felt that it was hard to reach an agreement. As the ultimate goal of the collaboration is to expose users to different perspectives and to increase the completeness and reliability of their analysis, reaching an agreement is not always necessary [2, 54]. The future design of such systems should explicitly signal to UX practitioners that disagreements with discussions are acceptable. Moreover, the tool must distinguish the disagreements *with* discussions from the disagreements *without* discussions, as the latter should be highlighted to encourage evaluators to discuss about.

8 CONCLUSION

Informed by the literature and a formative study with UX experts, we designed a visual analytics tool, CoUX, to support UX practitioners in both independently analyzing a usability test video and collaborating with each other. CoUX extracts acoustic, textual, and visual features from think-aloud videos using machine learning and includes a chatbox-like interface for problem annotation and discussion among others. We conducted an exploratory user study with six pairs of UX practitioners in collaborative video analysis tasks. The results show that CoUX helped them improve the completeness and reliability of their analyses. The results also show different features allowed them to spot problems that they might otherwise have neglected and to have focused conversations to seek clarification from and respond to their partners. In sum, our work has taken a first step to creating an integrated environment to support the analysis and collaboration of usability test videos among UX practitioners and highlighted further research directions.

REFERENCES

[1] T. Andre, H. Hartson, and R. Williges. Determining the Effectiveness of the Usability Problem Inspector: A Theory-Based Model and Tool for Finding Usability Problems. *Human factors*, 45:455–82, Feb. 2003. doi: 10.1518/hfes.45.3.455.27255

[2] B. Bailey. Judging the Severity of Usability Issues on Web Sites: This Doesn't Work. https://www.usability.gov/get-involved/blog/2005/10/judging-the-severity-of-usability-issues.html, Oct. 2005.

[3] R. Benbunan-Fich. Using protocol analysis to evaluate the usability of a commercial web site. *Information & management*, 39(2):151–163, 2001.

[4] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl. VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):61–70, Jan. 2016. doi: 10.1109/TVCG.2015.2467871

[5] B. Castellano. Scenedetect: A cross-platform, OpenCV-based video scene detection program and Python library.

[6] P. K. Chilana, J. O. Wobbrock, and A. J. Ko. Understanding usability practices in complex domains. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, pp. 2337–2346. ACM Press, Atlanta, Georgia, USA, 2010. doi: 10.1145/1753326.1753678

[7] L. Cooke. Assessing Concurrent Think-Aloud Protocol as a Usability Test Method: A Technical Communication Approach. *IEEE Transactions on Professional Communication*, 53(3):202–215, Sept. 2010. doi: 10.1109/TPC.2010.2052859

[8] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. Richter Lipford, and R. Chang. Recovering Reasoning Processes from User Interactions. *IEEE computer graphics and applications*, 29(3):52–61, May 2009. doi: 10.1109/MCG.2009.49

[9] H. B.-L. Duh, G. C. Tan, and V. H.-h. Chen. Usability evaluation for mobile device: a comparison of laboratory and field tests. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pp. 181–186, 2006.

[10] M. Fan, Y. Li, and K. N. Truong. Automatic Detection of Usability Problem Encounters in Think-aloud Sessions. *ACM Transactions on Interactive Intelligent Systems*, 10(2):1–24, June 2020. doi: 10.1145/3385732

[11] M. Fan, J. Lin, C. Chung, and K. N. Truong. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Transactions on Computer-Human Interaction*, 26(5):1–35, Sept. 2019. doi: 10.1145/3325281

[12] M. Fan, S. Shi, and K. N. Truong. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *Journal of Usability Studies*, 15(2):85–102, 2020.

[13] M. Fan, K. Wu, J. Zhao, Y. Li, W. Wei, and K. N. Truong. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):343–352, Jan. 2020. doi: 10.1109/TVCG.2019.2934797

[14] A. Følstad, E. L.-C. Law, and K. Hornb. Analysis in practical usability evaluation: A survey study. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems - CHI '12*, pp. 2127–2136. ACM Press, Austin, Texas, 2012. doi: 10.1145/2207676.2208365

[15] A. Følstad, E. L.-C. Law, and K. Hornbæk. Analysis in usability evaluations: An exploratory study. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, NordiCHI '10, pp. 647–650. Association for Computing Machinery, New York, NY, USA, Oct. 2010. doi: 10.1145/1868914.1868995

[16] FullStory. FullStory — Robust Analytics, Session Replay, Heatmaps, Dev Tools, and more. https://www.fullstory.com/platform.

[17] F. Goldman-Eisler. Psycholinguistics: Experiments in spontaneous speech. 1968.

[18] Gong.Io. Gong.io. https://www.gong.io/, 2021.

[19] Google. Google Sheets: Free Online Spreadsheets for Personal Use. https://www.google.ca/sheets/about/, 2021.

[20] J. Grigera, A. Garrido, J. M. Rivero, and G. Rossi. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, 97:129–148, 2017.

[21] P. Harms. Automated usability evaluation of virtual reality applications. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(3):1–36, 2019.

[22] J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: Sup-

porting asynchronous collaborative visualization. *Communications of the ACM*, 52(1):87–97, Jan. 2009. doi: 10.1145/1435417.1435439

[23] M. Hertzum and N. E. Jacobsen. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 15(1):183–204, 2001. doi: 10.1207/S15327590IJHC1501_14

[24] M. Hertzum, R. Molich, and N. E. Jacobsen. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, 33(2):144–162, Apr. 2013. doi: 10.1080/0144929X.2013.783114

[25] C. J. Hutto and E. Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, pp. 216–225. Ann Arbor, MI, 2014.

[26] P. Isenberg and D. Fisher. Collaborative brushing and linking for co-located visual analytics of document collections. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, Euro-Vis'09, pp. 1031–1038. The Eurographs Association & John Wiley & Sons, Ltd., Chichester, GBR, June 2009. doi: 10.1111/j.1467-8659.2009.01444.x

[27] Y. Jadoul. Praat-parselmouth: Praat in Python, the Pythonic way. https://parselmouth.readthedocs.io/en/stable/, 2019.

[28] J. Jeong, N. Kim, and H. P. In. Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *International Journal of Human-Computer Studies*, 139:102364, 2020.

[29] C. Jiménez, C. Rusu, S. Roncagliolo, R. Inostroza, and V. Rusu. Evaluating a Methodology to Establish Usability Heuristics. In *2012 31st International Conference of the Chilean Computer Science Society*, pp. 51–59, Nov. 2012. doi: 10.1109/SCCC.2012.14

[30] J. O. Johanssen, J. P. Bernius, and B. Bruegge. Toward usability problem identification based on user emotions derived from facial expressions. In *2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, pp. 1–7. IEEE, 2019.

[31] R. Kocielnik, S. Amershi, and P. N. Bennett. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. Association for Computing Machinery, New York, NY, USA, May 2019. doi: 10.1145/3290605.3300641

[32] E. L.-C. Law and E. T. Hvannberg. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction*, NordiCHI '04, pp. 241–250. Association for Computing Machinery, New York, NY, USA, Oct. 2004. doi: 10.1145/1028014.1028051

[33] C. Lewis. *Using the "Thinking Aloud" Method in Cognitive Interface Design*. IBM T.J. Watson Research Center, 1982.

[34] Lookback. Lookback: Simple and powerful user research. https://lookback.io/, 2021.

[35] N. Mahyar and M. Tory. Supporting communication and coordination in collaborative sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1633–1642, 2014. doi: 10.1109/TVCG.2014.2346573

[36] S. McDonald, H. M. Edwards, and T. Zhao. Exploring Think-Alouds in Usability Testing: An International Survey. *IEEE Transactions on Professional Communication*, 55(1):2–19, Mar. 2012. doi: 10.1109/TPC.2011.2182569

[37] K. Moran and K. Pernice. Remote Moderated Usability Tests: Why to Do Them. https://www.nngroup.com/articles/moderated-remote-usability-test-why/, Apr. 2020.

[38] J. Nielsen. 10 Usability Heuristics for User Interface Design. https://www.nngroup.com/articles/ten-usability-heuristics/, 1994.

[39] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pp. 152–158. Association for Computing Machinery, New York, NY, USA, Apr. 1994. doi: 10.1145/191666.191729

[40] J. Nielsen. Severity Ratings for Usability Problems. https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/, 1994.

[41] J. Nielsen. Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/, 2012.

[42] Noldus. Record & annotate - Recording options and easy annotation. https://www.noldus.com/viso/record-annotate, 2020.

[43] M. Nørgaard and K. Hornbæk. What do usability evaluators do in practice? an explorative study of think-aloud testing. In *Proceedings of the 6th Conference on Designing Interactive Systems*, DIS '06, pp. 209–218. Association for Computing Machinery, New York, NY, USA, June 2006. doi: 10.1145/1142405.1142439

[44] D. Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[45] OpenCV. Optical Flow function. `https://docs.opencv.org/3.4.13/d4/dee/tutorial_optical_flow.html`, 2020.

[46] A. Oztekin, D. Delen, A. Turkyilmaz, and S. Zaim. A machine learning-based usability evaluation method for elearning systems. *Decision Support Systems*, 56:63–73, 2013.

[47] F. Paternò, A. G. Schiavone, and A. Conti. Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–11, 2017.

[48] H. Richter Lipford, F. Stukes, W. Dou, M. Hawkins, and R. Chang. Helping Users Recall Their Reasoning Process. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pp. 187–194. Salt Lake City, Utah, USA, Oct. 2010. doi: 10.1109/VAST.2010.5653598

[49] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 233–240. IEEE, 2005.

[50] A. C. Robinson. Collaborative synthesis of visual analytic results. In *VAST'08 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings*, pp. 67–74, Dec. 2008. doi: 10.1109/VAST.2008.4677358

[51] A. Sarvghad and M. Tory. Exploiting analysis history to support collaborative data analysis. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, pp. 123–130. CAN, June 2015.

[52] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing Dimension Coverage to Support Exploratory Analysis. *IEEE transactions on visualization and computer graphics*, 23(1):21–30, Jan. 2017. doi: 10.1109/TVCG.2016. 2598466

[53] J. Sauro. Rating the Severity of Usability Problems. `https://measuringu.com/rating-severity/`, July 2013.

[54] J. Sauro. How to Assign the Severity of Usability Problems. `https://measuringu.com/severity-ratings/`, June 2016.

[55] A. Sears. Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction*, 9(3):213–234, Sept. 1997. doi: 10.1207/s15327590ijhc0903_2

[56] A. Sehili. Auditok: A module for Audio/Acoustic Activity Detection.

[57] Silverback. Silverback 3. `https://silverbackapp.com/`, 2019.

[58] M. B. Skov and J. Stage. Supporting problem identification in usability evaluations. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, OZCHI '05, pp. 1–9. Computer-Human Interaction Special Interest Group (CHISIG) of Australia, Narrabundah, AUS, Nov. 2005.

[59] Slack. Slack: Where work happens. `https://slack.com/`.

[60] S. L. Sporer and B. Schwandt. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(4):421–446, 2006.

[61] TechSmith. Morae 3 Tutorials. `https://www.techsmith.com/tutorial-morae-current.html`.

[62] UserTesting. Usertesting: The human insight platform. `https://www.usertesting.com/`.

[63] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, Nov. 2007. doi: 10.1109/TVCG.2007.70577

[64] M. Wattenberg and J. Kriss. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, July 2006. doi: 10.1109/TVCG.2006.65

[65] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala. CommentSpace: Structured support for collaborative visual analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 3131–3140. ACM, 2011. doi: 10.1145/1978942.1979407

[66] A. Zhang (Uberi). SpeechRecognition: Library for performing speech recognition, with support for several engines and APIs, online and offline.

[67] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan. Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):261–270, Jan 2017. doi: 10.1109/TVCG.2016. 2598543

[68] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):340–350, 2017.

[69] Zoom. Zoom - Video Conferencing, Web Conferencing, Webinars, Screen Sharing. `https://zoom.us/`, 2021.