Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions

Yun Huang, Yifeng Huang, Na Xue
SALT Lab
School of Information Studies
Syracuse University, U.S.A
yhuang,yhuan114,nxue@syr.edu

Jeffrey P. Bigham
Human-Computer Interaction Institute
Language Technologies Institute
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT
Hearing-impaired people and non-native speakers rely on captions for access to video content, yet most videos remain uncaptioned or have machine-generated captions with high error rates. In this paper, we present the design, implementation and evaluation of BandCaption, a system that combines automatic speech recognition with input from crowd workers to provide a cost-efficient captioning solution for accessible online videos. We consider four stakeholder groups as our source of crowd workers: *(i)* individuals with hearing impairments, *(ii)* second-language speakers with low proficiency, *(iii)* second-language speakers with high proficiency, and *(iv)* native speakers. Each group has different abilities and incentives, which our workflow leverages. Our findings show that BandCaption enables crowd workers who have different needs and strengths to accomplish micro-tasks and make complementary contributions. Based on our results, we outline opportunities for future research and provide design suggestions to deliver cost-efficient captioning solutions.

Author Keywords
Video Caption; Crowdsourcing; Complementary Contributions

ACM Classification Keywords
H.5.2. User Interfaces

INTRODUCTION
Given the large volume of videos on social media and in online learning environments, e.g., Massive Open Online Courses (MOOCs), there is a pressing need to provide accessible online videos for both hearing-impaired users and non-native speaking users. As of 2012, it is estimated that 36 million Americans are hearing impaired [3]. The National Association of the Deaf (NAD) has been taking legal action to require educational organizations to provide captioned online videos to meet the needs of these individuals. In addition, there are approximately 1 million international students enrolled in undergraduate colleges and universities in the United States alone [25]. Prior research finds that positive effects are generated for non-native speaking people when they use video captions as a tool for word, phrase, and context exposure [5].

However, the cost of captioning services for a large volume of videos is prohibitive [30], in part because it takes at least three to four times the length of the video for inexperienced people to create captions from scratch [9], and experienced captionists are much more costly. Machine-generated captions using Automatic Speech Recognition (ASR) technologies usually contain many errors [12]. As a result, captioning all video content being produced currently is intractable, and this ignores the sizable legacy of existing uncaptioned videos.

A promising approach is to use crowdsourcing to help address this problem. Having crowd workers edit machine-generated captions has been shown to be feasible [16], and starting with the ASR output may lower costs [22]. However, requiring that crowd workers be able to hear and be skilled with the language in order to contribute has likely left motivated potential contributors out. Past approaches have not considered that some crowd members with the incentive to help caption videos may be those the captioning is intended to help most, e.g., people with hearing impairments and those who are learning the language (second-language speakers).

In response, we designed and implemented a system called BandCaption that structures the captioning tasks into a workflow that allows crowd workers with different language skills to make useful contributions at the micro-task level. BandCaption starts with speech recognition, and then applies a Mark-Edit-Approve workflow that is similar to Find-Fix-Verify [2], except each stage is allocated to workers with different needs and strengths and different workers can collaboratively contribute to the same task. Our studies show that different user groups made complementary contributions based on their hearing or language strengths and constraints. For example, native speakers, in general, did best at marking all kinds of errors; hearing-impaired users were very sensitive to missing punctuation; and second-language users picked up missing words and irrelevant words well, however they showed limitations at correcting the errors. In our study, native speakers could edit all the errors that second-language

CHI 2017, May 06-11, 2017, Denver, CO, USA

 ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3026032

users failed to correct due to complicated spelling or general unfamiliarity with the spoken language.

The contributions of this work are the following: 1) a system, BandCaption, that enables users who may not be able to hear or who are learning the language to become crowd workers who can make useful contributions to captioning; 2) a Mark-Edit-Approve workflow that enables crowd workers who have different needs and strengths to accomplish micro-tasks; and 3) studies that demonstrate the value of assigning micro-tasks to workers based on their complementary skills.

RELATED WORK

Below we provide an overview of the literature concerning video captioning, and highlight how our work contributes to the literature.

The Utilities of Video Captioning

A significant amount of work has emphasized the importance of video captions. Captioning errors such as missing words and misspellings of technical terms may lead to a lack of understanding or increased misunderstanding, especially for people with disabilities [11]. Studies found that non-native speakers exposed to text (e.g., captions, transcripts, subtitles, etc.) in addition to aural repetition could better understand the information being conveyed as opposed to those who were solely exposed to aural repetition [13], as these users could confirm the information they heard by referring to the captions [7] [10] [29]. Users ultimately could use captions to increase their level of attention, improve information processing, reinforce previous knowledge, and analyze incomplete language. In addition, combining captions with audio-visual resources could also enhance users' listening and reading comprehension of a second language [31] [6] [8] [21], as videos with captions may facilitate vocabulary acquisition and video watching [26].

Recent work shows that captions generated by ASR can aid hearing-impaired individuals in understanding videos, even though these machine-generated captions may contain errors [28]. This implies that hearing-impaired users may not need all of the video captions to be accurate; instead, they may need only specific video segments to have accurate captions in order to better understand the video.

Inspired by the above literature, we designed the BandCaption system to allow users to identify specific caption issues. Our study with the BandCaption system contributes new understandings on the struggles that hearing-impaired users and non-native speakers face when using inaccurate captions.

Creating Video Captions

Online video-captioning processes need to accommodate key factors, such as cost, availability and quality [16]. Currently, there are three main methods of providing video captions, including CART (Communications Access Real-Time Translation) [17], ASR [27], and crowdsourced captioning, e.g., Scribe [18]. CART is currently the most reliable service for caption generation, but it requires trained people who are able to type at the same speed at which they hear. The high cost of CART also substantially affects its viability. ASR offers machine-generated captions but these often contain a significant amount of errors which may impede users from understanding video content [12]. Research explored by Novotney and Callison-Burch verifies the viability of using crowd workers to transcribe videos, in some instances in a different language [24]. Similar solutions are proposed by Munteau et al. with their system, which enables users to edit machine-generated captions "on-the-fly" and in a collaborative manner [22]. Such types of crowdsourcing solutions require a great deal of time, participants, and attention to produce adequate results [18]. In terms of quality, CART is similar to crowdsourced captioning, and both of these are greater than ASR [17]. It is suggested to use a combination of two of these methods to create effective systems [19].

Our BandCaption system combines ASR and crowdsourcing approaches and divides the caption correction tasks into smaller micro-tasks. This targeted problem-solving model could reduce the crowd work effort and resolve caption issues more effectively based on the competence of different crowd workers.

The Find-Fix-Verify Model

The "Find-Fix-Verify" model, as an assembly-line inspired method, has been shown to be effective in other application domains, e.g., crowdsourced writing [2]. Bernstein et al. implemented this strategy by breaking the process of proofreading written content into three stages [2]. By distributing the task to workers, the cost is reduced while quality is controlled through a careful distributing process. The value of breaking a larger task into smaller tasks helps better define the objectives and levels of execution expected. It prevents "errant crowd workers from contributing too much, too little, or introducing errors into the document" [2]. The model has also been used to micro-task assignment systems [4] and has been implemented to leverage the wisdom of crowds [1].

In this work, we implemented a Mark-Edit-Approve model in the design of BandCaption. It is similar to Find-Fix-Verify, where one task is decomposed to micro-tasks. Unlike Find-Fix-Verify, where workers vote on one fix from a set of alternatives as the final fix, our Mark-Edit-Approve model allows subsequent workers to edit earlier workers' edits. Prior research exploring the relationship between "native" and "non-native" speakers in collaborative and educational environments find that 1) native and non-native speakers make and miss different types of errors, and 2) errors made/missed by non-native speakers can be resolved by native speakers commonly through repetitious evaluation [15]. This suggests native speakers and non-native speakers will find different caption errors, and native speakers could fix the errors that non-native speakers could not. Our results show that applying the Mark-Edit-Approve model could provide an opportunity for different crowd worker groups not only to reduce the workload, but also to enable them to make complementary contributions.

SYSTEM DESIGN

The design of BandCaption was informed by the relevant literature we reviewed. The high-level idea is as follows:

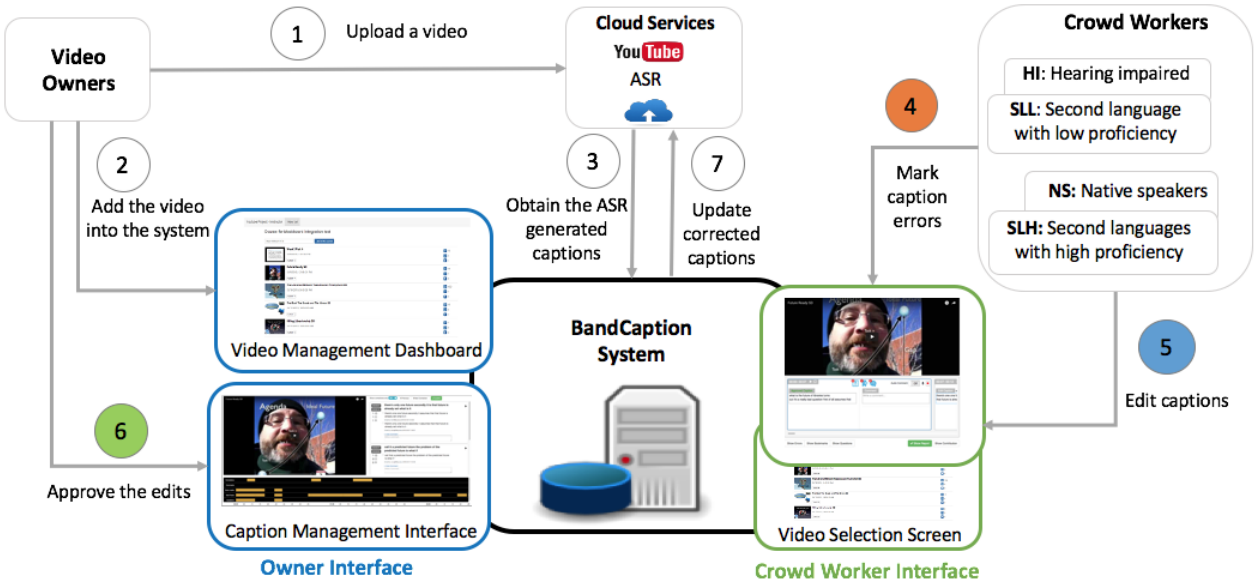

Figure 1. System Model and User Workflow

(1) to break the caption correction task into micro-tasks, so that, instead of simply consuming the video content, hearing-impaired users and non-native speakers can contribute as crowd workers by marking specific captions that need to be edited based on their own needs [28, 31]; (2) to allow users with varying language competency levels to co-edit caption errors, with certain easy edits being contributed by users with lower language proficiency (such as second-language crowd workers) and more difficult errors being edited by native speakers with higher language proficiency [15].

More specifically, we consider four stakeholder groups as our source of crowd workers: *(i)* HI - individuals with hearing impairments; *(ii)* SLL - second-language speakers with low proficiency; *(iii)* SLH - second-language speakers with high proficiency; and *(iv)* NS - native speakers. One scenario could be that the HI and SLL crowd workers mark caption errors, and SLH and NS crowd workers may be more capable to help edit the marked captions. This scenario could leverage different incentives of the crowd workers to crowdsource captions more efficiently.

We expect that, compared to the approach where crowd workers edit captions they think are necessary to correct, our BandCaption system can (1) reduce the total effort of caption corrections by targeting specific errors that are identified by users who rely on captions; and (2) provide a better understanding as to how different user groups use captions, which could help us to devise incentive mechanisms to encourage more contributions to the system.

In the remainder of this section, we will present the workflow and rationale of the interaction design of the BandCaption system. We use the term crowd workers to refer to any users who have the potential to contribute to captioning.

System Model

To bootstrap video captions using ASR technologies, we chose to leverage an existing cloud-based service. We found that YouTube provided ASR-generated captions through open APIs, and YouTube has been widely used to render a variety of videos. Thus, we adopted YouTube's caption service to simplify the creation of the machine-generated video captions.

Figure 1 illustrates the overall workflow. There are two major types of users: video owners and crowd workers. The video owner is in charge of setting up the videos with machine-generated captions in the system, which encompasses the following steps: (1) the video owner uploads a video to YouTube and turns on captioning in the YouTube video's settings; (2) the video owner adds the YouTube video to our system by providing the video URL; (3) the video owner gives permissions to the system to access the ASR-generated video captions, so that they are made available to crowd workers. Then, all the crowd workers can use the crowd worker interface to select a video to watch. If they notice caption errors while watching the videos, they can mark the captions, as illustrated in step (4). The crowd worker who marks the errors may or may not have the skills to edit the captions. Those who have the skills and are willing to edit the captions can edit the captions directly, in step (5). Those who edit the captions may not necessarily mark the errors, though. Whenever a caption correction is submitted by a crowd worker, the owner will be able to approve the correction in step (6). At the same time of approval, the approved captions will be updated to the cloud service in step (7), and crowd workers' interfaces will be updated with the newly corrected captions immediately. It is important to clarify that the mark-edit-approve process for one caption can be iterated multiple times if necessary. Therefore marking and editing are collaborative, and these contributions

can build upon each other to incrementally improve the quality of the captions.

The Crowd Worker Interface

As shown in Figure 2, the interface was created to ensure crowd workers could replay any specific caption segment to mark caption errors or to edit captions. More specifically, machine-generated captions are broken up into segments by the ASR technology; thus, each mark is associated with a video segment. The panel is synchronized with the video display. When the video is paused, the crowd worker can navigate the video segments using the navigation buttons on the sides or the scrolling bar at the bottom. We applied a participatory design approach when designing the user interaction and made three rounds of revisions to the design in the past 9 months. Seven participants iteratively ran tests and evaluated the design by finishing TAM (Technology Acceptance Model) surveys, which were used to measure the perceived usefulness and the perceived ease of use [20]. The current design has incorporated their suggestions for better usability.

First, regarding marking features, in order to allow and encourage crowd workers to mark captioning errors, crowd workers should be able to mark the caption at any time when they are watching a video. This interaction should not interrupt the display of the video, and the marking effort should be kept to a minimum. When crowd workers are limited (e.g., by time or availability), it needs to be effective for workers to evaluate the priority of the caption tasks. Buttons for marking caption errors are positioned so that crowd workers can easily access them while watching for errors. When any of these buttons are selected, a notification counter is displayed, which is also visible to other crowd workers. Overall, the number of users who have clicked these buttons could signify the important captions with a high demand that need to be fixed.

Second, regarding editing features, the machine-generated captions are displayed where crowd workers can directly edit them. If the crowd worker clicks within this area, a "submit" button appears. Before submitting an edit, crowd workers can replay the video at the desired caption/time segment twice to save work from replaying the video for caption editing. When the user is ready to finish editing, clicking the "submit" button will send the edits to the video owner. In the meantime, crowd workers will be able to see which other user last edited the same caption next to the "Edit Caption" title. Once the caption edits are submitted, the video continues the display automatically. After a caption is approved and submitted by the video owner, the original caption will be replaced by the new, corrected caption, and the "Edit Caption" title will be replaced with an "Approved Caption" title.

To promote interaction between crowd workers, we also provide a comment area where they can make comments. Other crowd workers may address these comments, if they think it necessary. An additional feature below the caption-edit area enables crowd workers to filter comments and inputs made personally and/or by other crowd workers. Crowd workers can also view a report of their contributions to the video along with other crowd workers' contributions.

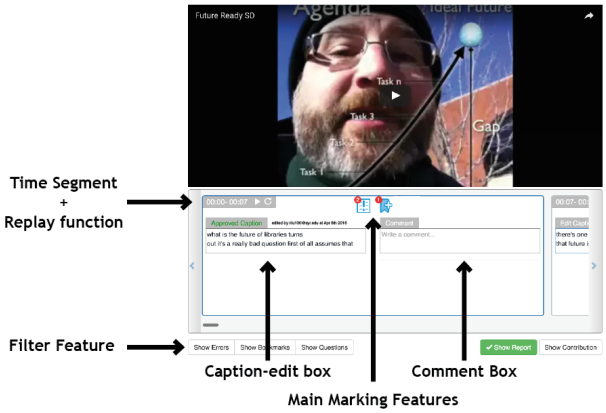

Figure 2. User Interface for Marking and Editing Captions

STUDY DESIGN

To understand if the BandCaption system can serve the design goals in terms of effectively enabling workers to mark caption errors and edit them, we designed a study with different groups of workers who had varying language skills.

For *marking*, we focused on addressing:

Q1: Given the same video, will different crowd worker groups have the same marking performance?

Q2: What are the criteria used by crowd workers to mark caption errors?

For *editing*, we focused on addressing:

Q3: Can different crowd worker groups make complementary contributions?

Q4: Does the BandCaption design save crowd workers' effort in editing the video captions?

We focused on examining the two sessions separately because our design aims to enable independent marking and editing contributions. In real life scenarios, users may mark captions without editing the captions; they may also edit captions that are marked by others, but do not mark captions themselves.

Task and Procedure

For each participant, we began by introducing them to the system, starting from system log-in. We then let the participant navigate the system and explain to us their understanding of each feature. If their understanding was different from the design purpose, we explained the feature to them.

We used two videos that each participant would then review for machine-generated caption errors. Each video was approximately 15 minutes in duration. The videos were about how to create new library services. We intentionally selected these videos because the topic did not involve any specific technologies that could create additional barriers for participants with limited language skills or create biased results in favor of participants that might be more familiar with those technologies.

For each video, each participant was asked to complete a *Mark Session* to identify caption errors. After the *Mark Session* was complete, the researchers asked each participant about what he/she marked and asked him/her to explain why they marked it. Depending on their language skills and competence, the participants might continue with an *Edit Session*. In the *Edit Session*, given a set of identified caption errors, participants who had the competence to correct captions were asked to correct all errors. Upon finishing the corrections, the participants were asked to explain how they made their corrections. During each session, the researchers paid close attention to each participant's behavior, individually and in comparison to other participants.

After they completed these two sessions for the first video, they were then asked to repeat the same two sessions for the second video. This was done to accommodate for the possible effect of a learning curve at the beginning of the first video, as participants might need time to pick up the new system. Fortunately, our final results showed consistent results between the two videos, indicating the system was easy to use.

Recruitment and Participants

We studied four distinct user groups: 1) hearing-impaired users (HI); 2) second-language speakers with lower proficiency (SLL), who were taking entry-level classes at the English Language Institute; 3) second-language speakers with high proficiency (SLH) who had passed the Graduate Record Examination (GRE) with satisfactory scores; and 4) native speakers (NS).

Eventually, we recruited a total of 34 participants in a range of different ages and backgrounds. Each participant was paid 20 dollars for participating in the study.

The HI group consisted of 4 participants who were all legally deaf, about 40 years of age; 2 female and 2 male. They were all native speakers. We were able to recruit these participants by sending emails to a local community group for hearing-impaired people. Recruiting through local organizations allowed us to reach a more ecologically valid population than recruiting from online crowdsourcing platforms like MTurk would have. Because these participants could not hear the audio, they were only asked to mark captions.

The SLL group consisted of 10 participants, all undergraduate non-native-speaking students, about 20 years of age; 7 from China, and 1 each from France, the UAE, and Taiwan; 9 females and 1 male. They were recruited from the level 3 class in the English Language Institute via emailing the mailing list.

The SLH group consisted of 10 participants who were all graduate international students studying in the U.S. (non-native country), about 24 years of age; 9 from China and 1 from India; 5 female and 5 male. To make sure they were at a similar level of English proficiency, we recruited them from the same level of education (i.e., graduate level) and did so via the school's mailing list.

The NS group consisted of 10 native-speaking participants, 6 males and 4 females, with an average age of 23 years old. They were recruited from the school's mailing list.

FINDINGS

In this section, we present participants' marking and editing results, as well as feedback on the study. We also evaluate the effectiveness of the design in collecting marking and editing contributions. As mentioned earlier, there was no significant difference between the results of the two videos, thus we only present the information and results for one, the second video, below.

The Mark Session

We first present the overall marking performance of different groups, as shown in Figure 3. We observed that on average the NS group participants marked more captions (with errors) than the other groups.

Group	Marked Captions		Percentage of Total Captions with Errors	Errors in the Marked Captions		Percentage of Total Errors
	Mean	SD		Mean	SD	
HI	39.6	16.7	23.3%	244	51	35.2%
SLL	28.5	16.2	16.8%	171	98	24.7%
SLH	32.5	17.8	19.1%	189	110	27.3%
NS	78.8	30.4	46.4%	409	14	59..0%

Figure 3. Marking results of each group (out of 177 captions in this test video, 170 captions had a total of 693 errors).

Hearing-Impaired Participants (HI Group)

Because the HI participants had a different group size than the other groups, and showed unique challenges, we present their results separately. The marking results of the HI group are presented in Figure 4, a bitmap which displays the marking locations of the video segments within the video timeframe. Each row is for one participant and each black mark represents the location along the video timeline. The 4 HI participants marked a total of 53 unique captions.

Figure 4. HI group - error marking locations along the video timeline.

Upon completing the study with each participant, we asked the participant for his/her feedback regarding the marking process. When communicating with the HI participants, we wrote on paper (1 HI could read lips). Participant HI-1 highlighted the importance of punctuation, stating that, "*I am sitting here watching the closed captions on my television tonight and realize what a difference proper English makes in comprehending the message. I think because there was no punctuation to separate different thoughts, exclamation points to emphasize feelings, commas, capital letters for names, etc... It was all squished together making it difficult to connect thoughts and understand what was being said*". Participant HI-2 expressed a similar need for punctuation, and

explained that, " *What's hardest for me is there is no punctuation so I do not know what words are connected in a sentence or the starting [of] a new sentence*", while Participant HI-3 supported the importance for punctuation, intently nodding her head when asked if punctuation in captions was important to her.

During the user studies, the researcher observed that participant HI-3 looked frustrated and often sighed while reading the captions. Another researcher also observed physical cues from participants that suggested discomfort and strain. For example, one researcher observed that Participant HI-4 took a long rest after he finished the video. This commonality in relation to HI group participants' emphasis on punctuation illustrates how frustrated HI group participants were when the captions were not provided with the appropriate punctuation. Because of these consistent findings and the observed strain it was placing on HI group participants, we chose to conclude our study with the four HI group participants. In fact, the test video had 177 captions, but 53 (30%) of them had missing punctuation problems. The finding showed that punctuation should become an emphasized standard for more accessible video captions in future system designs.

When reviewing the bitmaps, we found that there were a few common captions identified by all HI group participants in each video. We extracted the common captions identified by all four participants. In particular, there are 9 captions in the video that were marked by all HI group participants. Compared to the average caption length of the whole video ($Mean = 59.16, SD = 12.24$), these captions had more characters ($Mean = 63.56, SD = 2.7$). This finding suggests that if a caption marked by an HI has an unusually long sentence that does not have punctuation, potentially it could be reviewed for missing punctuation.

Non-Hearing Impaired (SLL, SLH, and NS Groups)

In this section, we present the marking results of SLL, SLH, and NS Groups. Figures 5, 6, and 7 show the marking results of the SLL, SLH and NS groups respectively. If we calculate the unique captions marked by different groups, the SLL, SLH, and NS groups marked a total of 97, 109, 158 unique captions respectively.

We ran Fisher's Exact Test to test the independence between different groups and the amount of unique captions marked. The results showed that for the amount of unique captions marked, there was a significant difference between the SLL group and the NS group ($p < 0.0001$). Similarly, there was a significant difference between the SLH group and the NS group ($p < 0.0001$). But there was no significant difference between the SLL and SLH Group ($p = 0.234$). **(Q1)**

For Figures 5 and 6, we found a common caption marked by almost every participant in the SLL and SLH groups. The original caption marked was, "*came out what's Google+ Facebook with stops how many of you don't have to*". Because at this point of the video, there was a noticeably slower change in the speaker's pace, this was probably why all the participants were able to notice the errors in the caption. The caption should be corrected as, "*...came out. What's Google+? Facebook with circles. [laughter]. How many of you don't have to...*".

After the participants finished marking captions for errors, we asked them to share why they marked certain captions, or the general criteria they used for marking. Participant SLH-8 described that, "*I found some words are totally irrelevant with the words said by the speaker, and this kind of errors is the most obvious one for me*", while Participant NS-3 said, "*I marked the tense and grammar errors*". In Contrast, Participant SLL-10 described a focus on missing words, "*I think the most obvious errors for me are the missing words that are not captured by the caption*".

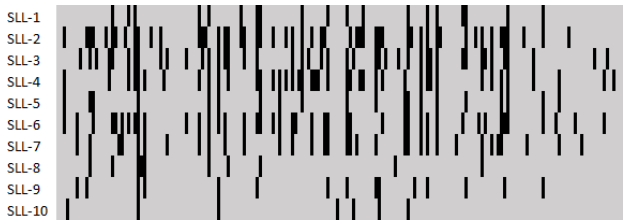

Figure 5. SLL group - error marking locations along the video timeline.

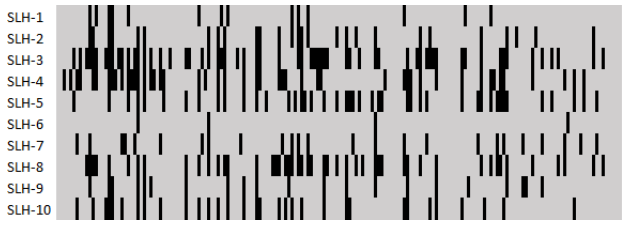

Figure 6. SLH group - error marking locations along the video timeline.

Figure 7. NS group - error marking locations along the video timeline.

According to the collected feedback from the SLL and SLH groups, participants generally found it easy to identify spelling errors first, and then to identify syntactical errors second, which might include words spoken and not displayed in the caption. Some participants in the Group SLL vocalized the difficulty of marking errors: "*Based on my understanding, I am not sure whether the caption is right or not, so I did not mark them*".

In contrast, the majority of the participants in the NS group shared with us that they would sometimes leave an error if they felt it was unimportant, like participant SLH-7 described. Participant NS-6 admitted that, "*I did not mark the punctuations and repeated words*". Considering the importance of punctuation to the HI group, we did not expect such a contrast in their marking criteria.

	Punctuation	Missing Words	Misspellings	Grammatical Errors	Irrelevant Words
HI	4	-	-	2	1
SLL	1	8	5	1	8
SLH	2	8	3	1	8
NS	3	4	6	4	7

Figure 8. The number of participants who applied different marking criteria.

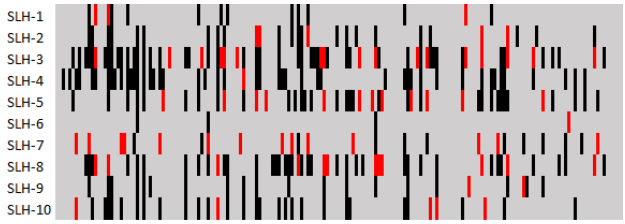

Figure 9. SLH group - comparing the marking and editing results.

To evaluate if the SLL, SLH and NS groups applied different criteria when marking captions, we annotated each participant's quotes to the identified marking criteria. Figure 8 shows the number of participants who applied different marking criteria. We further applied the Chi-squared test to test the independence between groups and error types. As HI Group participants had not identified missing words or misspelling errors in the video, we applied a Chi-squared test with all four groups and only three different error types (Punctuation, Grammar and Wrong/Irrelevant Words) first. The results showed that there was no significant difference between all four groups and the three error types ($\chi^2 = 10.006, df = 6, p = 0.1244$). We then applied a Chi-squared test with only three groups (SLL, SLH and NS), excluding Group HI, and all five error types (Punctuation, Missing Words, Misspelling, Grammar and Wrong/Irrelevant Words). The results showed that there was no significant difference between the three groups and all five error types ($\chi^2 = 6.4955, df = 6, p = 0.59$). **(Q2)**

The Edit Session

In this session, we aim to understand whether the SLH group had any limitations with caption editing, and whether the NS group could make complementary contributions to fulfill the tasks. Thus, we ran the study with the SLH group participants first. In the *Edit Session*, according to the prior literature [15] reviewed previously, we expected that the SLH group might not be able to correct all captions, and the NS group could make complementary contributions by editing the captions that the SLH group failed to correct. Thus, we asked the SLH group to edit as much as possible, then the NS group could run the study by only editing the remaining captions of the SLH group.

Second-Language Speakers with High Proficiency (SLH)

Out of 177 captions in this video, 170 captions had a total of 693 errors, including 387 missing punctuation marks (in 162 captions), 156 missing words (in 58 captions), 131 misspellings (in 75 captions), 7 grammar errors (in 7 captions) and 12 irrelevant words (in 11 captions). As we expected, each of the SLH participants failed to edit some captions, ($Mean = 25.6, SD = 14.24$). Figure 9 highlights the locations in red where the SLH group participants marked in the *Mark Session*, but failed to edit in the *Edit Session*.

When asked what criteria was used to mark errors, participant SLH-4 explained, "*I can correct irrelevant words, and words with similar pronunciation but different meanings*". Participant SLH-5 stated that, "*I add[ed] some punctuations in the captions such as comma[s] and period[s]*", though only 2 SLH participants added punctuation. Overall, the criteria that was used by the SLH participants for marking was extended to correcting errors as well.

We asked SLH group participants why they were unable to correct some of the captions they marked. Participant SLH-9 noted that, "*For some errors, I cannot figure out the correct words, although the caption was obvious[ly] wrong based on my understanding*", while participant SLH-4 felt that the video quality made errors harder to identify, saying, "*The video is more interactive, always has some background noise. For some errors, I can correct despite of the noise. Some [I] cannot* ". When asked about criteria for correcting errors, participant SLH-2 depended on the clarity of the video to correct errors saying, "*I can clearly hear what the speaker said and can spell the correct words*".

Native-Speaking Group (NS)

After the SLH group finished all their sessions, we extracted 10 captions that the SLH group marked but were unable to correct. In the *Mark Session*, the NS group still marked as many captions as they found necessary to correct. But when they entered the *Edit Session*, they were asked to correct the 10 captions we extracted. This would allow us to study whether the NS group could correct the captions left by the SLH group. The results showed that the NS group ultimately corrected most of these captions ($Mean = 7.25, SD = 1.91$).

Similar to the SLH group, the NS group's correcting criteria was almost the same as their marking criteria. Participant NS-4 stated clearly that, "*My correcting criteria is similar to my marking criteria*". In general, NS group participants were sensitive to different types of errors. For example, one type of correction that was unique to the NS group was adding quotation marks. Participant NS-1 noted that, "*I added some quotations when the accent changed*". When asked about the errors that they did not correct, NS group participants expressed that some captions seemed not to have any errors to correct. Participant NS-1 also stated, "*I did not correct some punctuations such as comma[s]*".

Collective Contributions

When comparing the NS participants' total corrected captions to their unmarked captions, as shown in Figure 10, we found that the captions corrected by the NS participants (marked by the SLH participants originally), were mostly different from the captions marked by Group NS participants. The complementary marked captions contributed to a larger amount of total errors marked and thereafter corrected.

More specifically, given the 10 captions with a total of 44 errors, the NS participants corrected missing punctuation (Mean = 3, SD = 5), missing words (Mean = 9, SD = 3), misspellings (Mean = 6, SD = 2), grammar errors (Mean = 2, SD = 1) and irrelevant words(Mean = 2, SD = 1);though only 5 NS participants added punctuation. Our results showed that without the SLH group pertinent errors would had been missed, but without the NS group, those errors might not have been corrected. This finding suggested that for the errors that were marked, different groups of workers would be able to correct them in a collaborative manner. The interaction between multiple distinct groups improves the quality of the captions due to their different language skills. **(Q3)**

	Total Corrected Captions	Captions Not Marked by the NS	Characters in Completed Captions	Saved Typing Characters	Percentage of Saved Characters	Unchanged Captions
NS-1	6	3	492	430	87.40%	4
NS-2	7	4	Captions	496	88.10%	3
NS-3	7	1	560	491	87.68%	3
NS-4	10	6	831	739	88.93%	0
NS-5	7	6	576	513	89.06%	3
NS-6	5	3	396	340	85.86%	5
NS-7	7	3	565	500	88.50%	3
NS-8	4	1	300	282	94.00%	6
NS-9	10	2	771	678	87.94%	0

Figure 10. NS group - editing performance

Saved Effort in Caption Creation

The design of the BandCaption system is intended to allow users who rely on captions to mark specific video segments where they need accurate captions, such that crowd workers do not need to aimlessly edit all captions that may have errors. We now present an analysis to verify the efficiency of our system in terms of saving effort from making captions.

As shown in Figure 11, the system reduced caption creation effort as a result of using ASR for the caption baseline. We compared the total characters of the completed captions that were corrected by the SLH participants to the characters that were changed after these captions were edited. On average, the participants saved typing 93.28% of characters.

	Total Corrected Captions	Characters in Completed Captions	Saved Typing Characters	Percentage of Saved Characters	Unchanged Captions
SLH-1	11	852	803	94.25%	3
SLH-2	22	1715	1623	94.64%	5
SLH-3	45	3567	3354	94.03%	17
SLH-4	47	3685	3493	94.79%	1
SLH-5	35	2876	2722	94.65%	9
SLH-6	3	185	162	87.57%	1
SLH-7	11	875	836	95.54%	14
SLH-8	36	2829	2619	92.58%	12
SLH-9	19	1518	1384	91.17%	2
SLH-10	27	2116	1981	93.62%	5

Figure 11. SLH group - editing performance

We also examined the difference between the amount of changed characters and the total amount of characters in completed captions. We applied a t-test between those two variables for correcting completion of the SLH group and the NS group. The results showed that for both SLH group and NS group participants, there was a significant difference between the amount of changed characters and the amount of characters in completed captions. More specifically, for the NS group, the amount of characters in completed captions ($Mean = 80.228 SD = 4.315$) and changed characters ($Mean = 9.286 SD = 3.186$) had a significant difference ($p < 0.001$); for the SLH group, the amount of characters in completed captions ($Mean = 78.977, SD = 7.958$) and changed characters ($Mean = 4.848, SD = 5.678$) also had a significant difference ($p < 0.001$). **(Q4)**

DISCUSSION

In this section, we reflect on our findings and discuss the design implications. We also acknowledge the limitations of our work and propose future research plans.

Reflections on the Findings

We were driven by the opportunity of dividing crowd workers into different groups to build a more cost-efficient system for caption correction. The Mark-Edit-Approve model offers the opportunity for tasks to be divided amongst different groups of crowd workers, amalgamating their sensitivities to different types of errors (e.g., spelling, punctuation, grammar) to mark and edit errors. Our findings confirmed that it was important to include hearing-impaired individuals in the study to mark captions because their own understanding of what a comprehensible caption should be could be different from other participants. For example, all HIs consistently reported their frustration due to missing punctuation, a type of error that did not seem as important to other participant groups, or in general might seem to be a negligible detail for captions to miss. But HI group participants helped us realize the importance of this type of error; as such, punctuation could become emphasized in different ways through the system, e.g., via user comments or through an algorithmic request.

This insight was reciprocated by the results of our non-native speaking participants, specifically SLL group participants, who had trouble marking and editing captions that involved jokes due to a lack of cultural understanding, or points in a video when the audio was faster than normal, just as Neuman and Koskinen found in their research [23]. The value of our study became clear, as it has helped us identify the importance of such a crowdsourcing system. In order to effectively correct captions using crowdsourcing, the system needs to anticipate and capitalize on the different backgrounds, strengths, and sensitivities of its users. An approach that is adaptive and accessible, it could, or maybe it must, 'feed' users captions based on those different dimensions to create a successful assembly line, or even better, an effective network of crowdsourced caption correcting. So in the case of captions that are being influenced by cultural cues, they can be directed towards native-speaking individuals who can translate and correct the error. This insight is also supported by Kurhila's observations in her research [15].

Design Implications

As discussed in "The Future of Crowd Work", current crowd work usually consists of independent and homogeneous tasks

[14]; the existing platforms assign caption creation to crowd workers as a single task and require crowd workers who can provide quality contributions to have certain language skills. This poses challenges to motivate the crowd workers, as well as realtime challenges such as scaling up to increased demand and making workers efficient enough to generate results ahead of deadlines [14]. Our work provides evidence and offers design implications for future, improved systems which can be used to address these challenges, to leverage the unique skills of crowd workers, and to provide a better workflow.

First, our findings can be used by video websites, e.g., YouTube and MOOCs. These systems can add user background information by either explicitly asking or estimating based on users' watching history. For example, if a user watches Chinese videos more regularly than English ones and frequently searches in Chinese characters, then the user might be a second language speaker, or a user who always turns the captions on might have a hearing impairment. Second, our results demonstrated that different intrinsic motivations could serve as motivators that possibly turn "users" (consumers of the videos) who rely on captions into "contributors" (creators of the video captions). For example, the desire to avoid frustration could form HI participants' intrinsic motivation to mark captions, or curiosity of unknown terms or words could motivate second language users to mark captions for further explanation. Our participants also mentioned that showing the number of marks for a caption (indicating the level of demand) could motivate them to edit captions.

Limitations and Future Work

The majority of our SLH participants were Chinese. Although China currently represents the largest international student population in the U.S. [25], we felt that this might influence our findings, as we did not have as clear an idea of how other second language speakers might contribute to our system as we did for the Chinese participants. Similarly, we realized how limiting the videos used in the study might affect our findings. We plan to run more studies with more participants who have diverse background and test with different types of videos in future work.

CONCLUSION

We present our design and study of a crowdsourcing system that bootstraps video captions using ASR technologies. Our system implements a Mark-Edit-Approve model and allows crowd workers who have different skills and abilities to make marking and correcting contributions collectively and interactively. Our study with 34 participants of 4 different groups contributes new knowledge of how different user groups with varying levels of language skills can make complementary contributions to the system. We also discuss design implications of future crowdsourcing systems and how to better motivate crowd workers.

ACKNOWLEDGEMENT

The contents of this publication were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90DP0061-01-00). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS).

REFERENCES

1. Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. 2013. Crowdsourcing linked data quality assessment. In *International Semantic Web Conference*. Springer, 260–276.

2. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: A Word Processor with a Crowd Inside. *Commun. ACM* 58, 8 (July 2015), 85–94. DOI: **http://dx.doi.org/10.1145/2791285**

3. Debra L Blackwell, Jacqueline W Lucas, and Tainya C Clarke. 2014. Summary health statistics for US adults: national health interview survey, 2012. *Vital and health statistics. Series 10, Data from the National Health Survey* 260 (2014), 1–161.

4. Alessandro Bozzon, Marco Brambilla, and Andrea Mauri. 2012. A Model-Driven Approach for Crowdsourcing Search. In *CrowdSearch*. 31–35.

5. Judy Chai, Rosemary Erlam, and others. 2008. The effect and the influence of the use of video and captions on second language learning. *New Zealand Studies in Applied Linguistics* 14, 2 (2008), 25.

6. Martine Danan. 2004. Captioning and subtitling: Undervalued language learning strategies. *Meta: Translators' Journal* 49, 1 (2004), 67–77.

7. Jrgen Froehlich. 1988. German Videos with German Subtitles: A New Approach to Listening Comprehension Development. *Die Unterrichtspraxis / Teaching German* 21, 2 (1988), 199–203. **http://www.jstor.org/stable/3530283**

8. Thomas J. Garza. 1991. Evaluating the Use of Captioned Video Materials in Advanced Foreign Language Learning. *Foreign Language Annals* 24, 3 (1991), 239–258. DOI:**http://dx.doi.org/10.1111/j.1944-9720.1991.tb00469.x**

9. Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. In *Proceedings of the International Web for All Conference (W4A 2016)*. 10.

10. C Grimmer. 1992. Supertext English language subtitles: A boon for English language learners. *EA Journal* 10, 1 (1992), 66–75.

11. Rebecca Perkins Harrington and Gregg C. Vanderheiden. 2013. Crowd Caption Correction (CCC). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and*

Accessibility (ASSETS '13). ACM, New York, NY, USA, Article 45, 2 pages. DOI: http://dx.doi.org/10.1145/2513383.2513413

12. Timothy J. Hazen. 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *In Proc. Interspeech*.
13. Jing-Fong Jane Hsu. 1994. Computer assisted language learning (CALL): The effect of ESL students' use of interactional modifications on listening comprehension. (1994).
14. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. DOI: http://dx.doi.org/10.1145/2441776.2441923
15. Salla Kurhila. 2001. Correction in talk between native and non-native speaker. *Journal of Pragmatics* 33, 7 (2001), 1083 – 1110. DOI: http://dx.doi.org/10.1016/S0378-2166(00)00048-5
16. Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2014. Accessibility evaluation of classroom captions. *ACM Transactions on Accessible Computing (TACCESS)* 5, 3 (2014), 7.
17. Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time Captioning by Groups of Non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 23–34. DOI: http://dx.doi.org/10.1145/2380116.2380122
18. Walter S. Lasecki, Raja Kushalnagar, and Jeffrey P. Bigham. 2014. Legion Scribe: Real-time Captioning by Non-experts. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*. ACM, New York, NY, USA, 303–304. DOI: http://dx.doi.org/10.1145/2661334.2661352
19. Chia-ying Lee and James R Glass. 2011. A Transcription Task for Crowdsourcing with Automatic Quality Control. In *Interspeech*, Vol. 11. Citeseer, 3041–3044.
20. Paul Legris, John Ingham, and Pierre Collerette. 2003. Why do people use information technology? A critical review of the technology acceptance model. *Information Management* 40, 3 (2003), 191 – 204. DOI: http://dx.doi.org/10.1016/S0378-7206(01)00143-4
21. Paul Markham and Lizette Peter. 2003. The influence of English language and Spanish language captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems* 31, 3 (2003), 331–341.
22. Cosmin Munteanu, Ron Baecker, and Gerald Penn. 2008. Collaborative Editing for Improved Usefulness and Usability of Transcript-enhanced Webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 373–382. DOI: http://dx.doi.org/10.1145/1357054.1357117
23. Susan B Neuman and Patricia Koskinen. 1992. Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly* (1992), 95–106.
24. Scott Novotney and Chris Callison-Burch. 2010. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-expert Transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 207–215. http://dl.acm.org/citation.cfm?id=1857999.1858023
25. Institute of International Education. 2015. Open Doors 2015 Report on International Educational Exchange. (2015). http://www.iie.org/Research-and-Publications/Open-Doors#.V0S1MZMrKAw
26. Jan L. Plass, Dorothy M. Chun, Richard E. Mayer, and Detlev Leutner. 1998. Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of educational psychology* 90, 1 (1998), 25.
27. Lawrence Rabiner and Biing-Hwang Juang. 1993. Fundamentals of speech recognition. (1993).
28. Brent N. Shiver and Rosalee J. Wolfe. 2015. Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 231–238. DOI: http://dx.doi.org/10.1145/2700648.2809857
29. Robert Vanderplank. 1988. The value of teletext sub-titles in language learning. *ELT Journal* 42, 4 (1988), 272–281. DOI: http://dx.doi.org/10.1093/elt/42.4.272
30. Mike Wald. 2013. Concurrent Collaborative Captioning. SERP.
31. Paula Winke, Susan Gass, and Tetyana Sydorenko. 2013. Factors Influencing the Use of Captions by Foreign Language Learners: An Eye-Tracking Study. *The Modern Language Journal* 97, 1 (2013), 254–275. DOI:http://dx.doi.org/10.1111/j.1540-4781.2013.01432.x