Using Avatars for Improving Speaker Identification in Captioning

Quoc V. Vy and Deborah I. Fels

Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada
{qvy, dfels}@ryerson.ca

Abstract. Captioning is the main method for accessing television and film content by people who are deaf or hard-of-hearing. One major difficulty consistently identified by the community is that of knowing who is speaking particularly for an off screen narrator. A captioning system was created using a participatory design method to improve speaker identification. The final prototype contained avatars and a coloured border for identifying specific speakers. Evaluation results were very positive; however participants also wanted to customize various components such as caption and avatar location.

Keywords: inclusive design, speaker identification, avatars, captioning

1 Introduction and Background

Film and television represent a significant method of cultural distribution. The ability to access such content is important for all people. Captioning is the main method used by people who are deaf or hard-of-hearing to access this content. It is the verbatim transcription of audio content using text descriptions and symbols. The most popular form of captioning is on television called Closed Captioning (CC) in North America or Subtitling in Europe. In North America, Line 21 CC appears as white mono-spaced text on a black background and is usually located near the bottom of screen. Text may be either as all uppercase or, more recently, as mixed-case lettering. The most common method for speaker identification in CC is using text descriptions consisting of two chevrons, speaker's name or function, followed by a colon (e.g., >>ANNE:).

Text descriptions are ineffective as they require prior knowledge (e.g., names of characters, especially when off-screen) and additional cognitive effort to associate a name and other visual indicators (e.g., lips moving) to the speaker. Among the few studies that have been carried out to address this issue and other non-speech information (NSI) are [1] and [2]. Harkins et al. [1] recommended using explicit descriptions for NSI and King [2] found that colour use for speaker identification did not improve much when compared to placement of captioning near the speaker.

In this paper, a graphical solution to speaker identification for captioning is described along with user comments and reactions from a *formative pilot study*. Four participants (two deaf, two hard of hearing) were used for prototype development and evaluations. Twenty casual hearing observers were added during final prototype evaluation for additional comments and feedback.

2 Needs and Requirements Analysis

The proposed system was developed using a participatory design (PD) method and involved individuals who were deaf, hard-of-hearing or hearing. We used an activity and mapping technique [3] to gather and understand user's setting and needs. The activity consisted of asking two deaf participants (one male, one female) to watch a favourite television show with CC at home.

A Diagnostic Mapping was produced from the activity and a follow-up interview revealed some common themes. For example, both users complained that CC was missing non-speech information, such as speaker information. Another common theme was a preference for italicized text and brackets to indicate narration and sound effects. A Virtual Mapping was then created to find some possible solutions to the issues identified. For example, using images and symbols to indicate different NSI and using the black bars, found on standard 4:3 screens with widescreen 16:9 content, to accommodate the additional information. Results from mappings indicated that deaf users rely heavily on captioning to be accurate and of sufficient quality to represent or indicate audio information that they would otherwise unable to obtain.

The proposed system is not designed to be used with existing captioning technology found on television as there are some technical limitations. It would be better applied in a digital implementation such as on a computer or the Internet.

3 Prototype Development and Iterations

A paper-based prototype was drawn using a "pencil before pixel" [4] design from a crude mock-up of the system using sticky notes on a screen. Screenshots or "avatars" of characters were placed adjacent to the captioning to visually identify the speaker, together called a "captioning panel". According to Law of Proximity [5], placement of the captioning panel was associated with the location of characters on screen to further aid identification of speakers. Some reactions from deaf participants were that they were excited and thought that this design was "different", "great", and "helpful". Furthermore, they liked the use of avatars and found that avatars helped indicate "who was talking" and improve "their understanding" of content.

The next step was to create image-based prototypes using actual content from a particular movie, in this case Transformers [6]. In this iteration, graphical and coloured elements were introduced to provide redundancy to further distinguish between speakers. For example, a coloured border matching the character’s primary wardrobe colour surrounded its corresponding avatar. Both deaf participants liked the coloured border and thought it assisted in their ability to identify the correct speaker.

Although both participants were positive about the avatar and colour border, there was also considerable divergence in deaf participants’ expressed needs and preferences. For example, participants wanted to place the avatars and there respective captions in different top/bottom locations and left/right order. As a result, user preferences were implemented which allowed viewers to change the location of the captioning panels, the order of avatars and captions, the size of avatars and text used for captioning, and the transparency of caption background.

4 Initial Evaluation and Discussion of Final Prototype

For the final prototype evaluation, we wanted to gain some additional perspectives from the hard-of-hearing and hearing communities. Two hard-of-hearing (HOH) participants were added to the participant pool and 20 hearing individuals (twelve females and eight males) acted as casual observers. All participants (deaf and HOH) and hearing observers were shown a 4:45 minute video clip of the Transformers movie with the final prototype (see Figure 1).

In Figure 1, four characters are shown on the screen, but only two of them are speaking. The captioning panel is located at the bottom of the screen and depicts the avatars of characters who are speaking, along with their names, a coloured border matching their respective wardrobes, and their corresponding dialogue located to the right of the avatar. The position of the dialogue and avatar is relative to the position of that character on the screen (e.g., Samuel is located on the left side of the screen while Optimus is on the right and mostly off-screen).

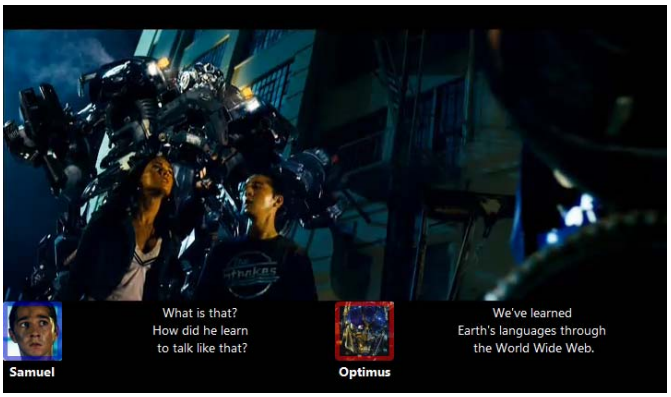

Fig. 1 - Screenshot depicting avatars, coloured borders and positioning for speaker identification. Text captions represent the dialogue occurring in this scene.

All participants thought that using avatars to represent who is speaking and having a multi-level display for captioning was innovative and an improvement over existing text-based methods. They also liked having the name of speaker below the avatar, as well as the use of italicized text and brackets for indicating sounds effects and narration. Participants liked having avatars regardless of ordering, but they preferred that the avatars be located on the left-side of the corresponding captions.

There were a variety of comments from deaf and HOH groups and some unexpected differences. In particular, some graphical elements were helpful for some, but not others. For example, the coloured border was helpful for deaf participants, but not for some of the HOH participants or casual observers. While studies by [2], [7] and [5] found that using colour for speaker identification was not helpful, this study shows that it may be beneficial and desirable by deaf users. Reasons for not liking the coloured border may be due to the use of other methods for identifying speakers such as the image or "avatar" and location of captions.

In this prototype, captions were maintained on screen as long as possible to maximize reading time. However, this caused some difficulty for some deaf participants as the captioning did not always match the onscreen visuals such as lips movements. All participants did not know where to look and were unable to follow along easily. A common suggestion was to highlight the current dialogue similar to karaoke. Highlighting word per word as in karaoke could interfere with reading efficiency as it forces people to read at that particular rate. A more effective implementation might be highlighting the entire captioning instead of individual words. Nonetheless, further research regarding optimizing reading time with speaking time is required.

Some participants from deaf and HOH groups found that they were overwhelmed with the amount of information available on screen. They initially found the screen “too busy” and the captions to appeared “too fast” making it difficult to read. This was caused by the overlapping of multiple dialogues being displayed simultaneously and for the extended duration. As a result, this increased the information that viewers had to absorb in a short time. However, after subsequent views participants were no longer "overwhelmed" as much. It seems that participants were too accustom to conventional CC that they required time to learn this new system and overcome automatic behaviours and expectations for reading the existing style of CC.

Further research and formal user studies are required to determine the effects on perceptual load, the readability of captions, the ability to see and understand video content, and the enjoyment levels. The ability to change various sizes, order and locations of avatars and captioning is a good start in finding an optimal and improved method of access to cultural content for people who are deaf or hard of hearing.

Acknowledgements. Funding for this research was provided by a SSHRC Community-University Research grant. We gratefully acknowledge the participants in the study who provided their time and effort throughout this research. Finally, we thank Dr. Andrew Clement and Dr. Eric Harley for their advice and support.

References

1. Harkins, J. E., Korres, E., Singer, B. R., & Virvan, B. M. (1995). Non-speech information in captioned video: A consumer opinion study with guidelines for the captioning industry. Washington: Gallaudet Research Institute.
2. King, C. M. (1996). CAP-Media website. Retrieved Jan. 5, 2009 from http://www.cap-media.com.
3. Keld Bødker, F. K. (2004). Participatory IT design: Designing for business and workplace realities. Cambridge: MIT Press.
4. Baskinger, M. 2008. Pencils before pixels: a primer in hand-generated sketching. *Interactions.* 15(2). 28-36.
5. Quinlan, P. T. & Wilton, R. N. (1998). Grouping by proximity or similarity? Competition between the Gestalt principles in vision. Perception 27(4). 417–30.
6. Spielberg, S. (Producer), & Bay, M. (Director). (2007). *Transformers* [Motion Picture]. United States of America: DreamWorks.
7. Rashid, R., Vy, Q., Hunt, R., Fels, D.I. (2008). Dancing with words. International Journal of Human-Computer Interaction. 24(5). 505 - 519.