

1-1-2012

ENHANCED CAPTIONING: SPEAKER IDENTIFICATION USING GRAPHICAL AND TEXT-BASED IDENTIFIERS

Quoc V. Vy
Ryerson University

Follow this and additional works at: <http://digitalcommons.ryerson.ca/dissertations>



Part of the [Computer Sciences Commons](#)

Recommended Citation

Vy, Quoc V., "ENHANCED CAPTIONING: SPEAKER IDENTIFICATION USING GRAPHICAL AND TEXT-BASED IDENTIFIERS" (2012). *Theses and dissertations*. Paper 1702.

This Thesis is brought to you for free and open access by Digital Commons @ Ryerson. It has been accepted for inclusion in Theses and dissertations by an authorized administrator of Digital Commons @ Ryerson. For more information, please contact bcameron@ryerson.ca.

ENHANCED CAPTIONING: SPEAKER IDENTIFICATION
USING GRAPHICAL AND TEXT-BASED IDENTIFIERS

by

Quoc Vu Vy

Bachelor of Science (2008)
Computer Science
Ryerson University, Canada

A thesis presented to

Ryerson University

in partial fulfillment of the requirements for the degree of

Master of Science

in the program of

Computer Science

Toronto, Ontario, Canada

© Quoc Vy 2012

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

// Signed //

Quoc Vy

ENHANCED CAPTIONING: SPEAKER IDENTIFICATION USING GRAPHICAL AND TEXT-BASED IDENTIFIERS

Master of Science (2012)

Quoc Vu Vy

Computer Science
Ryerson University, Canada

Abstract

This thesis proposes a new technique for speaker identification in captioning using three identifiers: image, name and colour. This technique was implemented as a proof-of-concept system called the *Enhanced Captioning: Speaker Identification* (EC: SID). This EC: SID was developed using participatory design and evaluated with people who are deaf or hard-of-hearing. This system evaluation used questionnaires and eye tracking methodologies, and the control was closed captioning, the existing system for North America. The results indicated that there is potential for using graphical and text-based identifiers for speaker identification. The placement of captioning or displaying the name of the speaker may not be effective for indicating who is speaking. The ability to customize these identifiers allows for changes in the content and different needs of users. Further design and evaluation is required to determine the long-term practicality of this graphical speaker identification technique.

Acknowledgements

I would like to acknowledge the following organizations and individuals for their valuable contributions towards this research:

... for *supporting* people who are deaf and hard-of-hearing
(of which some were recruited as study participants),

Canadian Hearing Society

Spadina & Davenport
Toronto, Canada

Canadian Cultural Society of the Deaf

Distillery Historic District
Toronto, Canada

Access Centre

Ryerson University, Canada

The LIFE Institute

Ryerson University, Canada

... for *sharing* your knowledge & insight,

Andrew Clement, PhD

Professor
Faculty of Information
University of Toronto, Canada

Ellen Hibbard

PhD Candidate
Communication & Culture
Ryerson University & York University, Canada

Norman Alm, PhD

Honorary Research Fellow
School of Computing
University of Dundee, United Kingdom

Alan F. Newell, PhD

Professor (Emeritus)
School of Computing
University of Dundee, United Kingdom

Paula Forbes, PhD

Research Fellow
School of Natural and Computing Sciences
University of Aberdeen, United Kingdom

... for *guiding* me through this journey as my research supervisors
(of which I am forever *grateful*),

Eric R. Harley, PhD

Professor
Department of Computer Science
Ryerson University, Canada

Deborah I. Fels, PEng, PhD

Professor
Ted Rogers School of Information Technology Management
Ryerson University, Canada

And for your *love, understanding, and patience* since Day 0,

my family (DV, KN, KV) & my friends (SM, HR, KS)

Table of Contents

Abstract.....	iii
Acknowledgements.....	iv
List of Tables	vii
List of Figures	viii
List of Appendices.....	ix
Chapter 1: Introduction.....	10
1.1 Motivation.....	10
1.2 Problem Statement.....	12
1.3 Thesis Statement	13
1.4 Research Questions	13
1.5 Scope	14
Chapter 2: Literature Review.....	16
2.1 History of Captioning.....	16
2.2 Laws and Regulations	18
2.3 Captioning Technology	19
2.5 Protocols and Guidelines	21
2.6 Other Purposes and Development	23
Chapter 3: System Design	25
3.1 Participatory Design	26
3.2 Investigation and Analysis of Needs	26
3.3 Prototype Development	32
3.4 System Description.....	34
3.4.1 Speaker Identification	36
3.4.2 User Preferences	37
Chapter 4: System Evaluation	38
4.1 Methodology.....	38
4.1.1 Content	38
4.1.2 Data Collection	40

4.1.3 Study Participants.....	40
4.2 Results	43
4.2.1 Eye Tracking	43
4.2.2 Repeated Measures for Understanding, Distraction, and Preferences	49
4.2.3 Preference of EC Components (Crosstabs).....	51
4.2.4 Purpose of EC Components (Crosstabs).....	53
4.3 Discussion	55
4.3.1 Participatory Design for EC: SID.....	55
4.3.2 Eye Tracking Behaviour of Captioning Systems.....	57
4.3.3 Use of Identifiers	57
4.3.4 Using Label for Speaker Identification	59
4.3.5 Limitations.....	60
Chapter 5: Conclusion	64
5.1 Summary of Findings.....	64
5.2 Contributions.....	66
5.2 Future Work	66
Bibliography	90
Glossary	95
Acronyms and Abbreviations.....	95
Terminologies.....	95

List of Tables

Table 1 Properties of Scenes	39
Table 2 Descriptives for Preference of Closed Captioning by Hearing Status	41
Table 3 Descriptives for Preference of Speaker Identification by Hearing Status	41
Table 4 Descriptives for Understanding of Speaker's Name	49
Table 5 Descriptives for Distraction of Speaker's Name	50
Table 6 Descriptives for Preference of Speaker's Name	50
Table 7 Preference of EC: SID components.....	51
Table 8 Preference of Image for SID by Hearing Status	52
Table 9 Preference of Image for SID by Hearing Status	52
Table 10 Preference of Label for SID by Hearing Status.....	52
Table 11 Preference of Staggered Dialogue by Hearing Status.....	53
Table 12 Purpose of Avatar by Hearing Status	53
Table 13 Purpose of Label by Hearing Status	54
Table 14 Purpose of Dialogue by Hearing Status.....	54

List of Figures

Figure 1 Closed Captioning (left) and Subtitles (right) on Television.....	11
Figure 2 Twitter messages about missing captioning for Disney's Up (2009)	11
Figure 3 Intertitles in Silent Films	16
Figure 4 Family watching television circa 1958.....	17
Figure 5 Character Set for Closed Captioning (left) and Subtitles (right)	20
Figure 6 Example of Teletext using Ceefax on BBC1	20
Figure 7 Example of a cinematic cut-scene from Warcraft III (2003)	23
Figure 8 Diagnostic Mapping	28
Figure 9 Captioning Panel	29
Figure 10 Black Bars (left) vs. Single Black Bar (right) for Captioning	30
Figure 11 Captioning Panel	35
Figure 12 Enhanced Captioning: Speaker Identification Styles	37
Figure 13 Scenes from Transformers (2007) for System Evaluation	39
Figure 14 AOI for Closed Captioning: Scene A (left) and Scene D (right)	44
Figure 15 AOI for Enhanced Captioning: Scene A (left) and Scene D (right)	45
Figure 16 Sum of Fixation Duration: Scene * AOI_Category (Style = EC)	47
Figure 17 Sum of Fixation Duration: AOI_Category (Style = EC)	47
Figure 18 Sum of Fixation Duration: Scene * AOI_Category (Video, Captioning)	48
Figure 19 Comments from regarding understanding and distracting SID components	58
Figure 20 Comments from regarding preference of SID components.....	59

List of Appendices

Appendix A: Enhanced Captioning: SID	68
Source Code.....	68
List of Files	68
Appendix B: System Evaluation.....	69
Research Ethics Board.....	69
Study - Information Sheet.....	70
Study - Consent Agreement	72
Questionnaires	73
Pre-Study Questionnaire	73
Trial Questionnaire	76
Comprehension Questionnaire	78
Post-Study Questionnaire.....	79
Captioning of Scenes	80
Scene A - Narration	80
Scene B - Aircraft	81
Scene C - Car Dealership	82
Scene D - Transformers.....	84
Appendix C: Copyright Permission	86

Chapter 1: Introduction

1.1 Motivation

In order for society to be inclusive, people who are *deaf* or *hard-of-hearing* should be given equal opportunity to participate, as with others. For simplicity, the term *deaf* includes people who are *deafened*, or consider themselves as *Deaf* (see [Glossary](#)). According to the *World Federation of the Deaf*, there are approximately 70 million people who are deaf, which is 1% of the world's population at 7 billion (UN 2011). Similarly, the *Canadian Association for the Deaf* considers 1% of Canadians are deaf and 9% are hard-of-hearing (CAD 2007), which if combined is approximately 3.45 million of the 34,482,800 people in Canada (Statistics Canada 2011). This represents a large group of people whose experience is limited when consuming popular culture (e.g. film, television, theatre, and music) in their existing form.

Cultural content is commonly available and consumed through television and more recently the Internet, where accessing these media has quickly become a part of everyday life (CRTC 2009). For example, Canadians watch an average of 30.6 hours per week (CRTC 2011), while Americans spend more than 33 hours per week (Nielsen 2012). In addition, streaming content is becoming more popular on mobile devices (e.g. smartphones and tablets), which is increasing the opportunity of consuming this content everywhere. Although providing access for people who are deaf or hard-of-hearing to this content may be challenging, their access is essential towards having a more inclusive society.

In order to access cultural content, people who are deaf or hard-of-hearing rely on captioning. Captioning is a text-based audio transcription for representing speech and *non-speech information* (e.g., speech prosody, emotion, speaker identification, sound effects, and music) by using text descriptions and symbols. The most popular form of captioning is found on television: *closed captioning* (CC) in North America, and *subtitles (for the hard of hearing)* in Europe (see Figure 1). The practical difference between these two systems is that closed captioning only uses white text on a dark

background, whereas subtitles may use different colours for indicating non-speech information (e.g., speaker identification and sound effects).

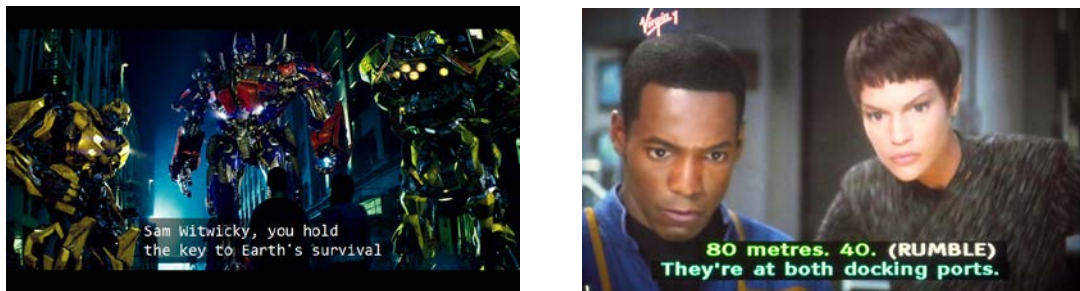


Figure 1 Closed Captioning (left) and Subtitles (right) on Television

Although using text descriptions and symbols may be appropriate for dialogue, it is ineffective for representing non-speech information, which may be crucial for presenting a coherent and entertaining experience (Zdenek 2011). Despite these limitations, the availability of content that is captioned is still a considerable step towards a more inclusive society. Therefore, content available today should be at least captioned, which is the expected standard from people who are deaf or hard-of-hearing. For example, the rental copy of Disney's *Up* (2009) was initially released without any captioning or audio description, as these were considered "bonus features". After receiving several complaints from the deaf and blind community (see Figure 2), this "manufacturing error" was fixed for subsequent copies of this and future films.

@TheFarmerjoe

<http://twitter.com/#!/TheFarmerjoe/status/5701896583>

"She went on to say that it was a marketing decision to remove all special features, apparently they saw SDH as a special feature."

@MarleeMatlin

<http://twitter.com/#!/MarleeMatlin/status/5807301183>

"Thank you Rich Ross and Disney for some awesome #captionaction. The "UP" captions on DVD/rental versions will be restored on future copies."

Figure 2 Twitter messages about missing captioning for Disney's *Up* (2009)

1.2 Problem Statement

As there are still issues with the accessibility of content, much effort from the deaf community is towards the quantity of captioning, and not the quality of captioning. Although practical and beneficial, a common misconception is that captioning is complete, and no further development is required. However, the use of text descriptions and symbols for conveying non-speech information does not provide an equivalent experience.

The most prominent example of this limitation is with music, where the title of the song and/or music notes are displayed, such as (♪ Beethoven's Symphony No. 5 ♪) or MUSIC: "The Dance Of The Sugar Plum Fairy" by Tchaikovsky (BBC 2009). This technique assumes that a viewer has "heard" and knows that particular music, but this is likely not possible for those who are deaf. Furthermore, displaying a music note only indicates the presence of "music", and not its intended effect or purpose, which if described, still lacks some substance, such as (EERIE MUSIC).

In the case of speaker identification, the standard practise is to display chevrons >> and/or the speaker's name, followed by a colon : before the dialogue, such as >>NAME: or >> (CAB 2008) or using different colour for the captioning, such as white, green, yellow, or cyan (BBC 2009). However, these text-based techniques are ineffective and ambiguous: additional effort is required to associate a name with the character, and using a different colour is ambiguous as the colours are reused and are not consistent between scenes for each character.

Furthermore, displaying just the chevrons (without the name is typical) or using a different colour only indicates a change of the speaker, and not exactly who is speaking. Instead, viewers who are deaf or hard-of-hearing would rely on the movement of a character (e.g. lip or hands gestures) for a better indication of the speaker. However, this is not always even possible, such as when the speaker is off-screen, or if there are multiple characters in a group or lots of movement on-screen.

The focus of this research is to reduce the difficulty of figuring out who is the speaker. This absence of non-speech information may negatively affect the flow of the content and reduce the viewing experience. While there are other limitations with using text descriptions for expressing non-speech information (e.g., music, emotions), speaker identification is a fundamental element that has yet to be adequately addressed. As a result, this “basic” issue of speaker identification was examined further, in order to provide a better experience for people who are deaf or hard-of-hearing.

1.3 Thesis Statement

The use of text-based captioning is not sufficient and further development is needed, in order to provide a similar experience when accessing film and television for people who are deaf or hard-of-hearing. The use of graphics (e.g. colour and image) instead of text descriptions and symbols may be a more efficient method for conveying non-speech information, particularly for speaker identification. There has been little development in captioning, whereas the technology for film and television has advanced enormously, which has resulted in a further “cultural divide” / barrier between people who are hearing and not.

The concept of using image-based identifiers (e.g. avatar) to represent the speaker is a new innovation introduced in this thesis. As such it is important to evaluate this idea with the relevant user communities during the design process. User evaluation strategies commonly employed in human-computer interaction involve collecting qualitative and quantitative data through a variety of methods such as participatory design and formal user studies. In this thesis, participatory design and user studies will be used in the evaluation process.

1.4 Research Questions

The purpose of this research is to explore using a image-based technique for speaker identification, and determine user reactions to this technique. The questionnaires and eye-tracking methodologies will be used to evaluate this enhanced captioning system and compare them to closed captioning, the existing system in North America.

The following research questions will be explored in this thesis:

1. What are the possible designs for speaker identification, which is not limited by the existing captioning technology?

This includes having an understanding of the needs of the deaf and hard-of-hearing communities as well as including them in the solution finding exercise. It is important to include potential users of closed captioning in the process because this is the baseline of the existing solution. This also includes selecting the best designs from this process and evaluating their impact on deaf and hard-of-hearing audiences.

2. What are the differences between closed captioning and enhanced captioning in terms of eye gaze activity for people who are deaf and hard-of-hearing?

3. What are the preferences of identifiers (e.g. image, colour, name) for indicating speakers between people who are deaf and hard-of-hearing?

These questions relate to measuring user attitudes, behaviour and understanding of the new system. Assessing this impact will involve quantitative and qualitative measures that will be presented and discussed in Chapters 4 and 5.

1.5 Scope

In this thesis, a technique for speaker identification of using graphical identifiers was developed using participatory design. The evaluation of this system consisted of a formative user study with people who are deaf and hard-of-hearing to obtain quantitative and qualitative data using questionnaires and eye-tracking methodologies. This enhanced captioning system was compared with closed captioning, which is the existing system being used in North America. People who are hearing are excluded from the system evaluation because they are not the primary users.

The research in this thesis is an exploratory process and the results are unpredictable. As such the viability of this new technique for speaker identification is explored, and not any other aspects with captioning or content, such as the transcription of the

speech, captioning readability, reading speed, or the entertainment value of the content. Furthermore, the captioning for the other non-speech information (e.g. music, sound effects, and speech prosody) is not addressed and remains unchanged as text descriptions and symbols.

The enhanced captioning system was developed as one solution for research in speaker identification, and not for commercial applications (e.g. broadcasting, DVD, film). Only the design of this solution is explored and not other factors, such as industry, market, integration to other software, efficiency, etc. The speaker identification solutions investigated in this thesis are not applicable to a specific distribution technology. However, the interactive features of this enhanced captioning system would not work in an existing analogue broadcasting environment. Other venues such as film, digital television, DVDs or the Internet would be possible.

The issue of captioning that is being addressed is only for post-production content of film, and not live content (e.g. talk shows, news, sporting events). For live content on television, the (typewriter scrolling) captioning is created in real-time by a specially trained captionist, who only has time to focus mostly on the dialogue and less on non-speech information. However, for film and post-production content, the captioning is finished prior to broadcasting, which usually results in a higher quality (pop-on) captioning that includes better non-speech information (e.g., speaker identification).

Chapter 2: Literature Review

The history of captioning, as well its development and current practise will be reviewed in this chapter.

2.1 History of Captioning

In the late 1880s, early films were called *silent films*, as they initially did not contain any sounds. Instead, *intertitles* or *title cards* were shown throughout scenes, which contained descriptive information such as dialogue, sound effects, narrations, and other remarks (see Figure 3). Although there were usually live musical performances, such as a piano, which accompanied these films, intertitles often had elaborate designs to help create the desired mood or atmosphere. Intertitles were the earliest form of captioning, as they provided the necessary information for viewers to consume content of this *silent era*.

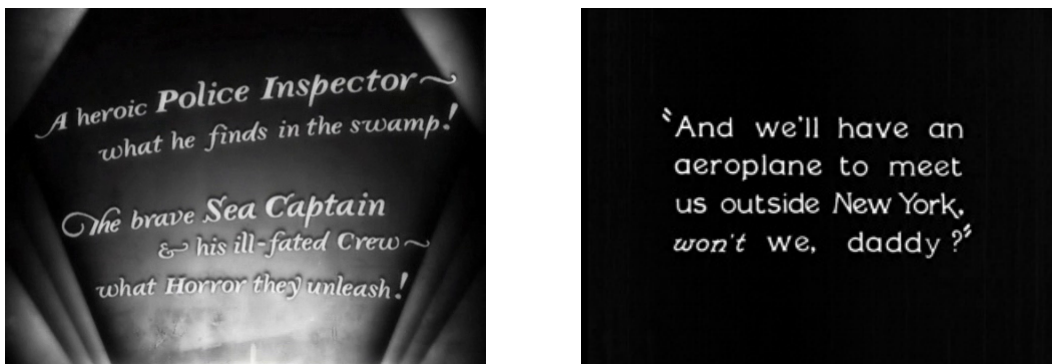


Figure 3 Intertitles in Silent Films

During the late 1920s, the use of intertitles declined as technology for including synchronized sounds directly on film was successfully developed. As such, films were no longer silent and were referred to as "talkies" by people at the time. Nonetheless, *deaf films*, silent films for the deaf, are still a significant part of Deaf culture today. For example, many Deaf film festivals are regularly held in various countries, such as the *Deaf Film and TV Festival* (Deaffest) in the UK, the *WORLDEAF Cinema Festival* (WDCF) in the US, and the *Toronto International Deaf Film and Arts Festival* (TIDFAF) in Canada.

Since being commercially available in the late 1930s, broadcast television has quickly become an "everyday" medium for consuming content (CRTC 2009). However, captioning did not appear on television until the early 1970s, over 30 years later. On February 11, 1972, "The French Chef" (PBS) was the first publicly broadcasted show to be 'open' captioned in the US (WGBH). In 1975, "This is Ceefax" a documentary film was the first to be subtitled in the UK (BBC). By 1980, regularly scheduled programs started to provide closed captioning on the various networks in the US (WGBH). On Sunday, March 16, 1980 among the first of these programmes were "The ABC Sunday Night Movie" (ABC), "Disney's Wonderful World" (NBC), and "Masterpiece Theatre" (PBS).



Figure 4 Family watching television circa 1958

Similarly, *sound films* have also been popularly consumed since being publicly available in the early 1930s. However, captioning for these films did not appear until the early 1990s, about 20 years later than on television. In 1993, WGBH (Great Blue Hill) conducted a user study to determine the most effective solution for providing 'closed' captioning in theatres (Seattle Times 1994). The study results indicated that Rear Window Captioning (RWC) was the most promising and least expensive of the other possibilities: using reflective goggles and back-seat displays. By the late 1990s, 'closed' captioning using RWC was available in theatres for people who were deaf or hard-of-hearing. Among the first films to use this form of captioning were "The Living

Sea" and "Stormchasers" for IMAX theatres in 1995, and "The Jackal" (Universal Pictures) and "Titanic" (Paramount Pictures) for first-run movies in 1997 (WGBH 1998).

2.2 Laws and Regulations

In response to the significance of film and television in terms of cultural expression and advocacy on accessibility rights, laws and regulations have been passed to ensure that publicly broadcasted content is accessible to everyone. This includes people who are deaf or hard-of-hearing. These legislations are regulated and mandated by their respective national organizations, such as the *Federal Communications Commission* (FCC) in the United States (US), the *Canadian Radio-television and Telecommunications Commission* (CRTC) in Canada, and the *Office of Communications* (Ofcom) in the United Kingdom (UK).

In the US, and effectively other countries, users are no longer required to purchase an external decoder device to display captioning on television. Since July 1993, circuits of these captioning decoders are required to be a standard feature on most television displays (*Television Decoder Circuitry Act of 1990*). In recent years, the quantities of captioned content have also increased to reflect this "essential" service (*Telecommunications Act of 1996*, CRTC 2009, *Audiovisual Media Services Directive 2009*). In the near future, these requirements will also apply to other video playback devices, such as smartphones and mobile devices (*Twenty-First Century Communications and Video Accessibility Act of 2010*).

In general, there are other legislations that exist which address accessibility for people with disabilities. For example, the "Americans with Disabilities Act of 1990" (ADA) and "ADA Amendments Act of 2008" (ADAAA) in the US, the "Canadian Human Rights Act" (1985) and "Accessibility for Ontarians with Disabilities Act" (2005) in Canada, and the "Disability Discrimination Act 1995" and "Equality Act 2010" in the UK. In these legislations, assistive technologies (including captioning) are a minimum requirement for public and private services.

2.3 Captioning Technology

Since being developed in the early 1970s, the process for providing captioning on television has more or less remained the same today, more than 40 years later. When captioning is available, it is synchronized with the content and displayed on-screen, usually near the bottom. The term "closed", as in closed captioning but also for subtitles, indicates that the captioning may be turned on/off as needed. Otherwise, if the captioning is embedded into the video and visible to all viewers, this is considered as "open" captioning (see [Glossary](#)).

The captioning data is encoded into the video signal for transmission and is decoded before being displayed on-screen. For 'open' captioning, no special decoding is required as the captioning is embedded onto the video stream. For closed captioning, this data is encoded on Line 21 of the Vertical Blanking Interval (EIA-608) of analog television and in the *picture user data* of the MPEG-2 stream (CEA-708) for digital television. For subtitles, this data is transmitted using *teletext*, an information retrieval service, and on a particular page (e.g., Page 888 using Ceefax in the UK) that varies depending on the country.

The lettering for closed captioning was initially all-uppercase due to difficulties of early decoders in rendering lowercase letters (e.g., **g**, **j**, **p**, and **q**) which have descenders. Although, mixed-case lettering is easier to read (Burt, J.S., & Hutchinson, 2000), all-uppercase lettering was the de facto standard and common practise in closed captioning until recently (CAB 2008). The colour of the text was initially only white on a black background, and typically still is for closed captioning. The most significant change to captioning has been for digital television, with additional functionalities and capabilities: different fonts, coloured text, background transparencies, and an extended set of characters and symbols (see Figure 5).

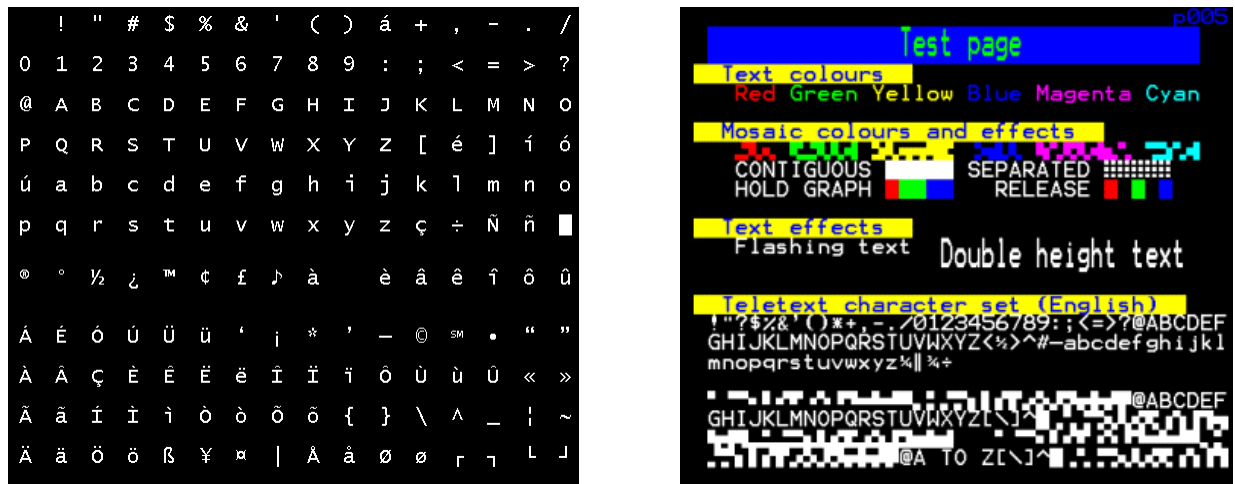


Figure 5 Character Set for Closed Captioning (left) and Subtitles (right)

Teletext, which is used for subtitles, was developed as an information retrieval service and already had various colours and graphics, including mixed-case lettering (see Figure 5). This demonstrates that there are other options that have been successfully used in other countries and there is not just one technique for captioning (see Figure 6).



Figure 6 Example of Teletext using Ceefax on BBC1

2.5 Protocols and Guidelines

There are several protocols and guidelines for captioning provided by organizations from all over the world, such as the Canadian Association of Broadcasters (CAB) in Canada, National Center for Accessible Media (NCAM) at WGBH in the United States, and Office of Communications (Ofcom) in the United Kingdom. Although, these are different organizations, their recommendations mainly differ in representing non-speech information due to the limitations of using text descriptions and symbols for closed captioning and using of colour in subtitles. Nonetheless, these recommendations provide a somewhat effective means despite these limitations as described below.

The placement of captioning near a character is the main indicator for identifying the speaker. However, this is often ambiguous or insufficient when there are multiple speakers in groups, off-screen speakers, or narration. In this case, the speaker's name and/or symbols (e.g., >>, :) may be used to reduce this ambiguity. For example >>ANNE: Good evening everyone. However, this assumes and requires that viewers are able to associate the name to a particular character correctly. This use of the speaker's name not only increases the cognitive load of the viewer, but also requires additional space on-screen. Instead, more commonly only the >> is used in closed captioning, or different colours are used for subtitles. However, this only indicates a different speaker, and not necessarily, who is speaking. Although colour is available in closed captioning, the use of colour is rare (e.g. “special effects” in music videos, or ending credits of the captioning), but is discouraged for speaker identification by itself (CAB 2008).

Similarly, text descriptions and symbols are ineffective for indicating other non-speech information. For music and singing, text descriptions or lyrics are surrounded by symbols: 🎵 for closed captioning, and # for subtitles. In addition, the title of a song is does not indicate the purpose or effect of the music, especially for individuals who were born deaf. Often a text description of the music (e.g., tempo, mood, genre, or style) is shown to convey the mood atmosphere, instead of the title of the music. For

sound effects and background noises, text descriptions or onomatopoeia words inside brackets are used for captioning, such as **(BOOM)**. For subtitles, colour and uppercase lettering are used instead, such as **DOOM SLAM**. Uppercase lettering can be used to indicate shouting or yelling, and for emphasis in closed captioning. For subtitles, a different colour can be used for emphasis, such as **it's the BOOK I want, not the paper**. For subtitles, sarcastic or ironic statements are indicated by using symbols, such as **Charming(!)** and **You're not going to work today, are you(?)** (BBC 2009).

Captioning for Film. In films, another form of transcription, called *subtitles*, is primarily used for translating the language of the content. Although this is commonly misunderstood, subtitles for films are not the same as 'subtitles for the hard of hearing' on television. For simplicity, the term “subtitles” if not mentioned is referring to “subtitles for the hard-of-hearing”, the captioning system on television found in Europe (see [Glossary](#)). The difference is that subtitles for films assume viewers are able to hear, but do not understand the language. As such, these subtitles do not usually contain any non-speech information, and are not as useful for viewers who are deaf or hard-of-hearing. Nonetheless, *same language subtitles* (SLS) may be used, in lieu of closed captioning, but are not as effective for these viewers. These SLS, which are called *subtitles for the deaf and hard-of-hearing* (SDH) in North America, are commonly found on the *Digital Video Disc* (DVD) of films.

On DVDs, subtitles for films are usually 'closed', as the technology to display them is built into the DVD player itself. This allows for multiple different translations, including SLS/SDH, in which the user selects one to be used as the subtitles. In theatres, subtitles for films are usually 'open' as for foreign films, but also for films that have some of its content in another language. However, if available, *Rear Window Captioning* (RWC) may be used instead for providing 'closed' captioning in the theatre. Although, RWC is less distracting and intrusive for non-users, only a few theatres offer this service due the additional hardware and operating cost. Nonetheless, RWC is more effective and contains non-speech information similar to captioning on television.

2.6 Other Purposes and Development

In a relatively short time, captioning has proven to be useful and has been adopted for various applications and purposes. For example, hearing individuals can also benefit from using captioning, such as in noisy environments, learning a new language or improving literacy skills. The latter is popular in the United States, where an annual event since March 2006 called "Read Captions Across America" have been increasing the awareness of using closed captioning in schools.

Captioning can also be found on new media formats, such as on the Internet or video games. However, captioning for these products and services was implemented after general release, and was not considered part of the initially design process. As with television, this delay excluded people who are deaf and hard-of-hearing unnecessarily. Nonetheless, captioning in gaming does show significant improvement of what could be possible with technology and when it is part of the design process. An example is Warcraft III (2003), which is a video game created by Blizzard Entertainment (see Figure 7). In this game, captioning exists in cinematic cut-scenes between characters using animated avatars and highlighting for speaker identification.



Figure 7 Example of a cinematic cut-scene from Warcraft III (2003)

In summary, the captioning systems on television were implemented in the late 1980s just enough, so that people who are deaf or hard-of-hearing are able to access television. However, there is no longer any further research or development for captioning as their implementations were sufficient, although limited to being only text-based due to the technology during that time. There are deficiencies in the current systems where technology not as limited and better solutions could be possible. This thesis will propose a possible approach for a better technique for speaker identification.

Chapter 3: System Design

The *Enhanced Captioning: Speaker Identification* (EC: SID) system was created to explore a potential solution for speaker identification that was primarily image-based compared to the existing text-based method. This captioning system was designed using existing technological options, which addressed the needs and requirements of users, and the limitations and problems with existing captioning systems (e.g. text-based). In particular, the use of an avatar an image-based method for speaker identification, which consists of three identifying components: a portrait image of the speaker, a coloured border that matches their clothing, and the speaker's name. This is different from the text-based method that is currently used for the captioning, either displaying only the speaker's name or using a different colours for the text. The text-based method is often insufficient or ambiguous for common situations, such as with multiple characters on-screen and/or off-screen speakers.

The EC: SID system is part of the *Enhanced Captioning* (EC) research project, which is exploring innovative solutions, using the visual modality, for conveying *non-speech information* (e.g. speaker identification, emotions, music). As a result, this non-speech information can be conveyed more effectively and expressively to viewers, as compared to the current method of using only text descriptions. This would in turn enable viewers, especially those who are deaf or hard-of-hearing to better consume the content as intended, but without access to the sound.

This and other EC systems are not designed for broadcasted television, where the technology is limited, but for computers, as this captioning system is interactive and requires more dynamic computing abilities. Furthermore, the system design of the EC: SID system is based on another EC system, called EnACT (Emotive and Affective Captioning Tool) which was created for representing emotions using kinetic typography (or animated text). Although both EC systems were developed and created by myself, the EC: SID system could be considered as a branch of the EnACT captioning system.

3.1 Participatory Design

The EC: SID system was developed using a *participatory design* (PD) methodology (Bødker, Kensing, & Simonsen, 2004) to include intended users, people who are *deaf* or *hard-of-hearing*, in the design and decision-making process. This PD method called MUST - a Danish acronym for *theories of and methods for design activities* - was inspired by ethnographic approaches (e.g. interviews and observations) and Scandinavian PD approaches (Kensing, Simonsen, & Bødker, 1998).

The MUST method uses a "baseline planning" technique for organizing the project into design phases: initiation, in-line analysis, in-depth analysis, and innovation. These phases consist of various design activities (e.g. planning, analysing, and problem solving) that lead to intermediate and end products. The products from each phase are used for assessing and achieving a baseline, which is a well-defined state within the project and a *decision-making point* between phases (Andersen et al., 1990).

By following this PD approach, a *mutual learning* process is achieved which furthers promotes a co-operative design between users and the designer (Beguin, 2003). For example, a *mapping technique* was conducted to understand the needs of users and explore technological options that are available to address their needs (Lanzara & Mathiassen, 1984).

The initial design of this EC: SID system has been described at conference proceedings (Vy & Fels 2009). I chose to take a flexible approach to this PD method, as this was an individual effort (no team) and for a group of people who are deaf and hard-of-hearing (not an organization). The following sub-sections provide a description of how I applied these activities to my project:

3.2 Investigation and Analysis of Needs

The needs of users were investigated using an activity and interview to obtain their opinion on watching television with closed captioning (CC) – the existing system in North America. This information was then analysed using the mapping technique to establish an overview and understanding of the situations that were problematic

(Bødker et al., 2004). During this participatory design phase, two deaf people were recruited: EH and FF. Although this may seem like a small number of volunteer, this was adequate as there was a lot of time and effort that is requirements for participatory design from both parties.

Activity and Interview. The activity consisted of participants at home watching their favourite content on television with captioning. Participants were asked to take written notes of their experience, particularly of situations that were problematic. The purpose of this activity is for participants to be critical regarding their experience with captioning on television, which would be later analysed. This activity added to their existing experience with closed captioning and ensured that participants had recent/specific examples from which they can draw upon during a follow-up interview with the researcher. The activity was not conducted in a usability lab or in the presence of the researcher as to preserve the natural environment and conditions for watching television – usually in one's living room and possibly with family or friends. The interview was conducted at the research lab, which consisted of exploratory discussions of this activity.

Mapping Technique. From these discussions, a *diagnostic mapping* (see Figure 8) was created to analyse the problematic situations that were identified by participants. Afterwards, a *virtual mapping* was created to evaluate ideas for their solutions (see Figure 9). Overall, this creates a logically step-by-step mapping between the identified problem and idea for solutions.

In Figure 8, there are various instances where captioning does not contain sufficient speech and non-speech information to effectively describe the content. The main causes are lack of space on-screen and amount of time for displaying additional information. The solution that was proposed is to increase the area for captioning and to use graphical elements for indicating speaker identification. Other aspects of non-speech information were not addressed in this research.

Problems	Causes	Consequences	Ideas for Solution
missing words / sentences (amount of captioning does not match amount of the movements of character's mouth)	technical limitation (space and time) - editing / shortening of text by captionists	inaccurate "picture" / representation changes character's perceived personality	increase character limitation / resolution of captioning no editing / maintain verbatim text
identifying speakers (on-screen, off-screen, narration)	technical limitation (space and time) - missing speaker's name (only indicates a different speaker using symbols: >>)	confusion due to not being informed which may lead to misinterpretation of speaker	Avatar (or picture of character) Colour (another identifier element which could be used to associate a particular (group) of characters) user "liked"/suggested: - italic - brackets - text description (e.g. speaker's name, "mailman voice:")
missing timing (start / end of music)	technical limitation (space and time) lack of completeness	Ambience or atmosphere is not conveyed	captioning shown for duration of sound/music - text description (e.g. music starts, music ends)
missing rate of speech	technical limitation (space and time) lack of completeness	flow of content is not well represented	user suggested: - text description (e.g. fast, slow)
missing speech prosody (emotion / mood)	technical limitation (space and time) lack of completeness	confusion / misinterpretation	emoticons kinetic text user liked: - text description (e.g. happy, excited)
missing tone / emphasis	technical limitation (space and time) lack of completeness	confusion / misinterpretation	user liked/suggested: - text styling (e.g. italic) - fonts - colours

Figure 8 Diagnostic Mapping

Some ideas for solutions to these problematic situations are located in Figure 8. These suggestions are both from participants and the researcher. However, the suggestions from participants for narration and emphasis were limited to text formatting (e.g.,

italics). The current practise of using italicized captioning is for indicating emphasis, or off-screen characters or narration. However, the application of italics for these different situations is ambiguous and is not an effective solution in the long-term. Similarly, this is still a problem for subtitles (for deaf and hard-of-hearing), where colour is both used to indicate these non-speech information as well as different speakers. This additional use for colour perhaps would be more distracting than an abrupt visual change in the captioning than having the word just italicized.

Ideas for Solution	Actions	Consequences
Captioning Panel	increase text area / resolution no editing / maintain verbatim text	exact verbatim text translation
Captioning Panel	captioning shown for duration of sound/music - text description (e.g. music starts, music ends)	provides temporal information
Captioning Panel	user suggested: - text description (e.g. fast, slow)	provides rate of speech
Captioning Panel	emoticons kinetic text user liked: - text description (e.g. happy, excited)	provides affective information
Use of Display Capabilities (graphics)	avatar colour user liked/suggested: - italic - brackets - text description (e.g. speaker's name, "mailman voice:") user liked/suggested: - text styling (e.g. italic) - fonts - colours	increase depth of information

Figure 9 Captioning Panel

From the problematic situations that were identified in Figure 8, the solutions and their impacts are “mapped” (or analysed) using a Virtual Mapping activity (see Figure 9). The solution is to create a *captioning panel*, and use graphical elements (e.g. images and colour) for indicating non-speech information. The captioning panel contains all of the captioning information and is a separate layer that is offset and outside of the area for the video. This is different from the existing systems where the captioning is completely in front of the video.

Most of the problems identified are likely because there is not enough space for the captioning, without blocking the video. In general, captioning currently appears in the bottom third of the screen, usually as 2 to 3 lines to avoid blocking a substantial part of the video (see Figure 10 left). Therefore, the solution would be to create a captioning panel, which is a dedicated space on-screen for captioning. Thus, there is a trade-off between the amount of captioning information and amount of video displayed. Another challenge is to display non-speech information as well as dialogue, since speaker identification is usually omitted or shortened. By creating the captioning panel, this trade-off is less of a problem (see Figure 10 right) as there is less overlap between the video and captioning.



Figure 10 Black Bars (left) vs. Single Black Bar (right) for Captioning

In most cases, the content and display does not match, in terms of aspect ratio. The aspect ratio is usually widescreen for film (e.g. 16:9 or 16:10), while most people still have televisions that are full screen, which is a 4:3 ratio. As a result, the video content is scaled down proportionally, resulting in horizontal “black bars” (see Figure 10). In this case, these black bars are not used, which is wasted space that could be used for

captioning. As a result, the SID system takes advantage of this difference by moving the video towards the top edge of the screen, which creates a larger area for the captioning panel. This reduces the overlap that would have occurred between the video and captioning.

Even though, the creation of this additional space for captioning may seem to be a good solution for the problems identified, it does not by any means address the bigger issue, which is representing non-speech information effectively.

The solution for the remaining problems identified in Figure 8 is to use images and colour in addition to (or replacing) the speaker's name for indicating who is speaking. By using these graphical elements, the speaker is more clearly identified with less ambiguity than using just the character's name. Another benefit is that the dialogue and the speaker identification are different modalities, where viewer is able to direct attention to each of them as needed. This is different from the current method where the speaker's name and dialogue are both text descriptions and is difficult to parse without reading both.

By changing the modality of the speaker identification from text to graphical elements, the viewer does not need to parse the name (if present and not needed) just to read the dialogue. In doing so, this would decrease the reading that is already required and thus the cognitive load of the viewer. However, the speaker's name is rarely visible (only the >> symbol is shown) for closed captioning, which is unclear and defeats the purpose of speaker identification all together.

There are other technical issues (e.g. quality and quantity) with captioning that were identified, but were not addressed in this research due to the scope. However, these problems have been previously noted by other researchers (Harkins, Korres, Virvan, & Singer, 1996), which further suggested that captioning is not complete and requires further development and innovation (see Discussion in Chapter 5).

3.3 Prototype Development

A new design for improving speaker identification was conceptualized using graphical elements to represent non-speech information, such as who is speaking, order of dialogue, and sound effects. A paper-based prototype was created using sticky notes on a computer monitor. This allowed participants to see simple examples of multiple speakers and off-screen voices or sounds. The participants viewed a design of a captioning panel that displayed the captions for the dialogue of two characters appearing on screen simultaneously and an avatar of characters adjacent to their respective captions.

Avatars were used, in lieu of text descriptions, to correspond to a character who speaking. Placement of each captioning panel was associated with the location of characters on screen. Each avatar was located next to its corresponding text and each captioning panel was located in close proximity to its respective character on screen. This is to group the speaker identification components together, as well as with the captioning, according to the Law of Proximity Gestalt psychology.

Only two speaker panels were used in this version of the design, as it seems to be most suitable and adaptable for various numbers of speakers. For example, in the case of a dialogue among more than two speakers, the two speaker panels would alternate showing only the last two speakers. This solution seemed to work with the content that was used. An investigation on the scalability of this concept of alternating speaker panels for more than two speakers would be a future research objective.

Two speaker panels indicating the characters and their conversation appear on screen. Avatars were placed on the outside edge of the screen, and the dialogue of the characters next to their respective avatars. The captions were vertically staggered to indicate order (top-level first) and alternation of dialogue that may occur between speakers. The pencil drawing of this figure was shown separately to the participants for feedback on issues that they may identify at this early stage of the design process.

The most surprising comments from the deaf participants were that they were excited and thought that this design was “different”, “great”, and “helpful”. They liked the avatar idea, and suggested it would help them know “who was talking” and “helped their understanding”. EH like the staggered dialogue because it “visually indicates timing” within and between speakers. Participants seemed to be able to conceptualize this design using this simple and primitive form of representation (e.g., paper and pencil). One participant (FF) even asked if this system was already available in movies theatres.

The participants were also able to provide suggestions for improvements. For example, EH suggested that moving the right avatar to the left of the text would be preferred and suggested that it might be “harder to read, but was better for understanding”. This was suggested to FF and agreed as well.

The next step was to create image-based prototypes using visual content from a movie, *Transformers* (2007). This particular movie was selected because it contained off-screen narration, multiple speakers overlapping and lengthy dialogue. A movie was selected instead of a television program because the quality of existing captioning tended to be higher, in terms of standard conformance and completeness. This may be due to greater amount of resources and funding available for movie production. Another advantage of using a DVD movie was that the participants were able to control the playback for the movie (e.g., pause and rewind), if necessary, to take notes. Even though this is a limited setting for studying viewer reactions in a real-time setting (e.g., in a movie theatre or watching television), it allowed a rich set of comments to be generated without the time pressure of continuous movie play that exists in movie theatres and for television shows.

In this phase of the prototype development, graphical and coloured elements were introduced to the caption display to represent the speaker. This provided redundancy that would further distinguish between speakers. For example, a coloured border matching the primary colour of the character’s wardrobe surrounded its corresponding avatar. EH found them to provide a “greater separation” between the characters. Both

participants were positive about the prototypes and there was some consensus regarding some of the new elements. There was agreement that the captions should not appear within the content screen and that the layout for the captioning panel, with the avatar on the left of the respective text captioning was the preferred design.

However, there was also considerable divergence in user preferences for the other elements. The fact that the deaf participants had their own preferences was expected, but the extent of those differences was not. For example, EH found it easier to read captioning located at the bottom of the screen, while FF found reading at the top of the screen easier.

The feedback and suggestions from the deaf participants led to the inclusion of a user preference feature in the final prototype. Preference functions included the ability to move the caption panels anywhere on the entire screen including into the content area, changing the left-right order between avatars and text, changing the size of the avatars and font used, and changing background transparency. Making preferences customizable to each user may increase even more the effectiveness of new design for captioning. As a result of the feedback and commentary from the PD process an enhanced captioning (EC) system has been developed to explore new concepts for speaker identification.

3.4 System Description

As with other captioning systems, the Enhanced Captioning: Speaker Identification (EC: SID) system is synchronized with the content and displayed on the top-most layer of the screen. However, instead of appearing just anywhere on-screen, the captioning information is organized and contained within a pre-defined, yet customizable section called the *captioning panel*. This captioning panel is a semi-transparent layer that is approximately one-third of the screen's height and located at the bottom of the screen by default (see Figure 11). The captioning information remains more or less the same, consisting of verbatim text transcriptions for speech and text descriptions for sound and music. However, instead of using only text descriptions, speakers are represented

in the form of *avatars*, which includes an image of that character along with their name. This form of speaker identification is perhaps more effective for expressing this non-speech information, as it does not requiring remembering names, which is often not clearly indicate who is speaking.

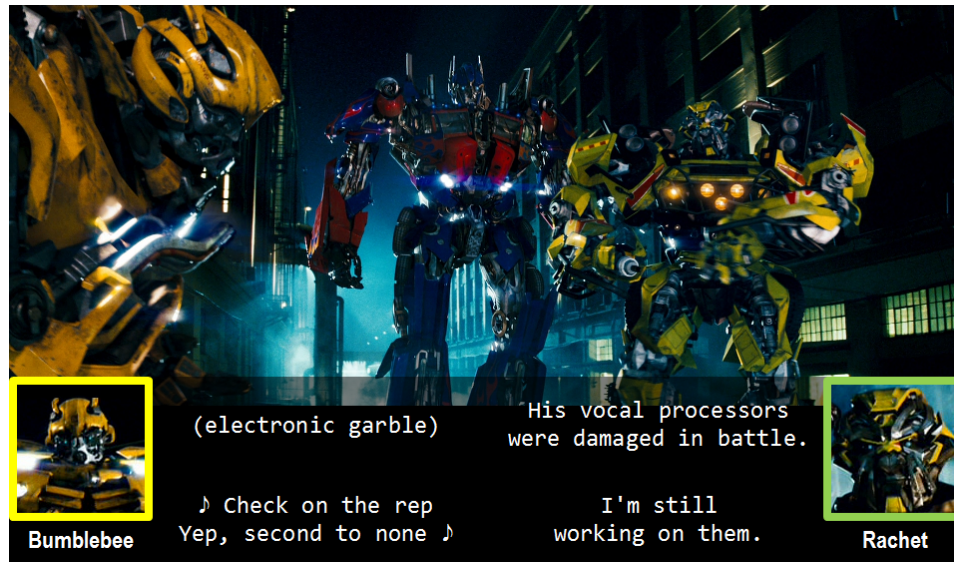


Figure 11 Captioning Panel

The captioning panel is designed to function by displaying conversations between the last two characters, and is divided into left and right sides, where each side contains the dialogue and speaker identification for a particular character. Ideally, the side that is used corresponds to the location of the speaker as they appear on-screen. The dialogue for each speaker is further divided into two levels, where the top level is used first and any subsequent dialogue is shown at the bottom level, and if necessary repeating back at the top level. When the captioning panel is functioning, this division of the sides and levels visually mimics the behaviour of the conversation itself. This again differs from existing captioning systems where this behaviour is less noticeable as the captioning information of those systems are not as structured. Although this behaviour was unintentional, the design structure of the captioning panel was necessary in order to maintain a reliable proximity association between the dialogue and speaker identification.

In order to reduce blocking a significant portion of the video, the captioning panel is located (by default) at the bottom of the screen and the video is shifted up to the top edge of the screen. By doing so, a large area of empty space is created below the video and serves as an ideal location for the captioning panel. This unused space would otherwise be known as horizontal mattes or “black bars” that appear when the aspect ratios of the video and screen are different (e.g. a 16:9 video displayed on a 4:3 or 16:10 screen). In addition to its background being translucent, the captioning panel may move up-and-down along the vertical axis of the screen. Furthermore, each of the components within the captioning panel is also configurable in terms of layout and size (see *User Preferences* below).

3.4.1 Speaker Identification

The method of speaker identification of the EC: SID system consists of three identifying elements:

Image. Ideally, this is a screenshot of the character when they are facing the camera (head shot or portrait). This image should be updated throughout each scene to match changes in the lighting, clothing, etc.

Coloured Border. This border surrounds the image and the colour usually matches the characters wardrobe. Another purpose for using colour was suggested for displaying their emotion.

Label. This is usually the name of the speaker, or if that is unknown their role (e.g. mail carrier) or at the very least a description of their voice (e.g., gender).

In order to accommodate the different needs of users, these identifiers may be toggled on/off separately as required. However, the coloured border was considered too ambiguous and unreliable by itself and was grouped with the portrait image into one entity called the avatar. As such, there are three (3) configurations (or styles) for identifying the speaker: EC1 which is only the name of the speaker, EC2 which is the image and coloured border (e.g. avatar), and EC3 which includes all elements (see

Figure 12). The EC1 style is similar to closed captioning, however the name of the speaker is always displayed.

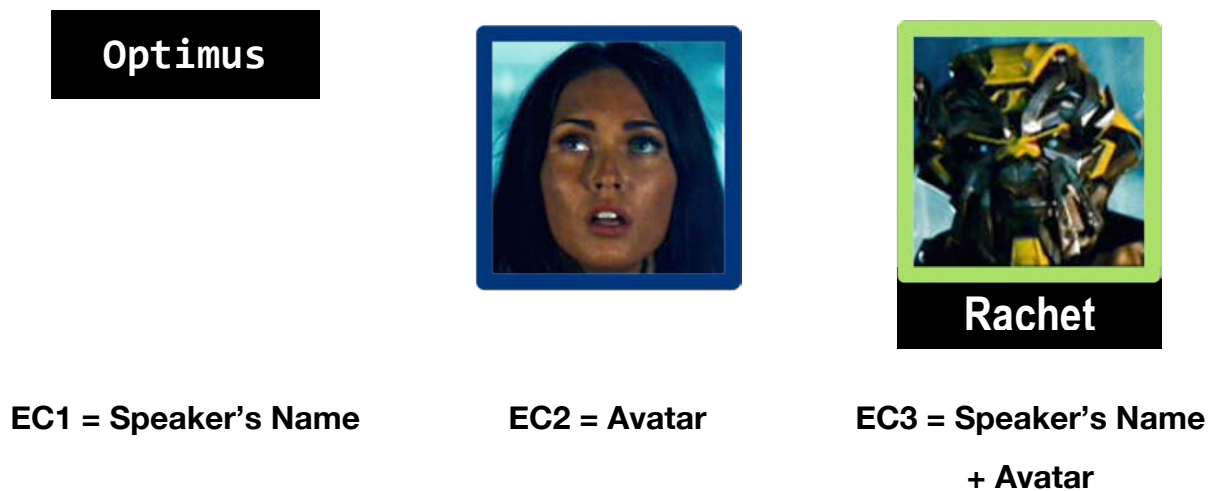


Figure 12 Enhanced Captioning: Speaker Identification Styles

3.4.2 User Preferences

The EC: SID system is very flexible in terms of displaying and configuring the various components within the captioning panel. As such, users are able to customize this system to address their specific and individual needs. The following is the list of preferences that are available to the user:

Avatar. The location of the avatar with respect to the dialogue can be changed using the space bar: left, right, inside outside, and off. The individual elements of the speaker identification components (e.g. image, label, and coloured border) may be toggled on or off.

Captioning Panel. The captioning window can be dragged anywhere vertically on the screen, between the top and bottom edge of the screen.

Chapter 4: System Evaluation

4.1 Methodology

The system evaluation consisted of using questionnaires and eye tracking methodologies to obtain quantitative and qualitative data. The control or baseline was closed captioning, the existing captioning system for television in North America.

4.1.1 Content

A live-action film, *Transformers* (2007) was used for the content, where the particular scenes were selected as follows:

The film was divided into scenes (fading to and from a black screen), which were marked by the start and end times. For each of these scenes, a brief description was given which consisted of the setting and/or the subject/topic of that particular scene.

The content of each scene was analysed to determine comprehension questions, which could only be answered from the captioning and could not be found from only in the visuals. This would ensure that the viewer is paying attention and that they are reading the captioning.

The final scenes used were based the difference in the number of speakers and characters on-screen as well the different “rate of speech” to ensure variety of possible cases (see Table 1).

Captioning Style. The captioning styles consisted of closed captioning as the control condition and three combinations of the speaker identification components: EC1 used the speaker’s name, EC2 used the avatar with a coloured border, and EC3 that used the speaker's name and the avatar with the coloured border.

Scene. The following are the scenes that were selected as the content for the system evaluation:



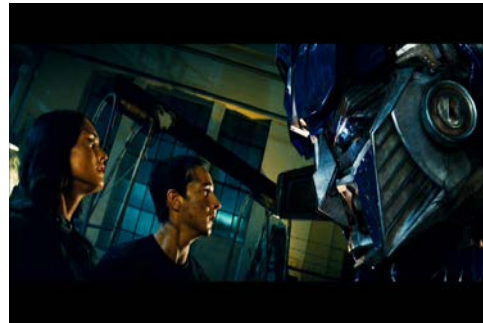
Scene A



Scene B



Scene C



Scene D

Figure 13 Scenes from Transformers (2007) for System Evaluation

Conditions. The captioning styles (CC, EC1, EC2, EC3) and scenes (A, B, C, D) were combined to create 16 conditions. The viewing order of these conditions was randomly presented to minimize any learning or ordering effects that may occur.

Scene	Duration	Speakers	Captioning		
			Length	Words	WPM
A	01:12	1	64s	127	120
B	00:51	4	48s	151	188
C	01:36	3	92s	298	194
D	01:23	7	70s	167	142

Table 1 Properties of Scenes

4.1.2 Data Collection

In order to evaluate the EC: SID system, qualitative and quantitative data were collected using questionnaires and eye tracking methodologies. The user study consisted of three phases: pre-study questionnaire, trial (questionnaires + eye tracking), and post-study questionnaire. (see [Appendix B](#) for the documents for this user study, including the questionnaires).

4.1.3 Study Participants

A total of 19 people (twelve deaf and seven hard-of-hearing) were recruited to participate in this user study. These participants were between 20 and 89 years old ($M = 40.68$, $SD = 17.54$) and watched ($M = 9.55$, $SD = 4.61$) hours of television per week, of which the majority (16) always use captioning and the remaining (3) hard-of-hearing have never used captioning. These participants were recruited from deaf social events (e.g. MAYFEST), the Canadian Hearing Society, and an emailing listing for people who are deaf or hard-of-hearing.

Deaf Group. The deaf group consisted of nine females and three males, between the ages of 20 and 59 years old ($M = 34.92$, $SD = 10.70$). The highest levels of education completed for this group is two (1 female, 1 male) high schools, seven (6 females, 1 college) colleges, and three (2 females, 1 male) graduate schools. This group watched ($M = 9.33$, $SD = 5.19$) hours of television per week and always used captioning while doing so.

Hard-of-Hearing Group. The hard-of-hearing group consisted of three females and four males, between the ages of 20 and 89 years old ($M = 50.57$, $SD = 21.99$). The highest levels of education completed for this group is 1 high school, 1 college, 3 universities and 2 graduate schools. This group watched ($M = 9.93$, $SD = 3.33$) hours of television per week, while the majority (3 females, 1 male) always used captioning, and the remaining (3 males) have never used captioning.

Opinion of Closed Captioning

Participants were asked for their opinion of closed captioning, in terms of the speed and placement of the captioning and the use of text descriptions, symbols, and colour. In general, the deaf and hard-of-hearing groups liked most aspects, except the placement of the captioning for the hard-of-hearing group (see Table 2). Due to the low $N = 19$, these five-point Likert scale were compressed to three-point (1 = dislike, 2 = neutral, 3 = like) for analysis.

	hoh			deaf		
	n	Mean	SD	n	Mean	SD
Speed	5	2.60	0.89	12	3.00	0.00
Placement	6	1.33	0.82	12	2.42	0.79
Text Descriptions	6	2.67	0.52	12	2.92	0.29
Symbols	6	2.17	0.75	12	2.92	0.29
Colour	6	2.00	0.63	9	2.44	0.73

Table 2 Descriptives for Preference of Closed Captioning by Hearing Status

Preference of Speaker Identification for Closed Captioning.

Participants were asked for their opinion of the speaker identification for closed captioning, in terms of the combinations of the speaker's name and symbols. In general, the deaf and hard-of-hearing groups liked most methods, except for the deaf group where the speaker's name is either in brackets or after the chevrons (see Table 3). Similarly, these five-point Likert scale were compressed to three-point (1 = dislike, 2 = neutral, 3 = like) for analysis.

	hoh			deaf		
	n	Mean	SD	n	Mean	SD
>>	4	2.00	1.16	10	2.10	0.99
SPEAKER:	4	2.50	1.00	11	2.27	0.91
>>SPEAKER:	5	2.00	1.00	7	1.86	0.90
(SPEAKER)	5	2.40	0.89	8	1.88	0.99
Dialogue Only	6	2.50	0.84	10	2.40	0.70

Table 3 Descriptives for Preference of Speaker Identification by Hearing Status

4.2 Results

There were two sets of data that were analysed from the eye-tracking and questionnaires methodologies. The initial analysis of the eye tracking data has been published in a journal manuscript (Vy & Fels 2011), where the Copyright Permission for including its content is located in Appendix C.

4.2.1 Eye Tracking

Due to the large amount of data collected, only Scenes A and D were analysed, as they represented the extremes, in terms of complexity: number of speakers. Scene A was the simplest, with one speaker off-screen who was narrating, and Scene D was the most complex, with seven speakers, but only 2 or 3 at a time. Scene D was the most dynamic with several camera angle changes, and required more focus and attention. In addition, there were many complaints (18) regarding the captioning being “too fast” (18), whereas Scene B (188 WPM) and Scene C (194 WPM) were the highest, compared to Scene A (120 WPM) and Scene D (142 WPM).

Gaze Fixation. The basic information that is obtained from eye-tracking is *gaze point* (where on-screen) and *fixation duration* (for how long). A *gaze fixation* is determined by a pre-defined radius size and duration of a *gaze point* – the x-y co-ordinates of where the participant is looking on the screen. For the purpose of this analysis, a fixation filter of a 30 pixels *Velocity Threshold* (radius size) and a 100ms *Duration Threshold* was used as recommended when using a *stimulus* containing both images and text (Tobii 2006). The analysis of this data is useful for determining differences between attention and focus, called *area of interest* (AOI). This is in contrast with the questionnaire data, which will be later analysed.

Area of Interest (AOI). The regions of the screen that were defined are categorized as follows:

1. Video: the video content
2. Captioning: where captioning may appear
3. SID: speaker identification component

Due to a limitation of ClearView eye-tracking software, defining “dynamic” AOIs are not possible. For closed captioning, there is no SID category as the speaker’s name is not separate from the dialogue, and may varied in length when displayed. For example:

>>ANNE: Good evening everyone . In addition, the speaker’s name does not always display. This is not as critical, since there were only 5 instances where the speaker’s name appears in the closed captioning condition. For Scene A, there was only one at the start of the narration (e.g. OPTIMIUS @ 00:01). For Scene D, there were four instances: OPTIMUS @ 00:22 and 00:49, JAZZ @ 00:26, and BUMBLEBEE @ 01:03.

This is a dynamic AOI, which the analysis software is unable to define. Even if defining dynamic AOIs were possible, the small size of such AOI is a limitation, which is below the “buffer zone” of at least 30 pixels for a fixation to be determined with confidence (no error).

Closed Captioning

For closed captioning, the categories of AOIs are as follows:

Video: CC.Video

Captioning: CC.Lower, and CC.Upper



Figure 14 AOI for Closed Captioning: Scene A (left) and Scene D (right)

Enhanced Captioning

For enhanced captioning, the categories of AOI are as follows:

Captioning: EC.UpperLeft, EC.LowerLeft, EC.UpperRight and EC.LowerRight

SID: Name.Left and Name.Right (Captioning Style: EC1)

Avatar.Left and Avatar.Right (Captioning Style: EC2 and EC3)



Figure 15 AOI for Enhanced Captioning: Scene A (left) and Scene D (right)

Repeated Measures of Fixation Duration

A mixed-factor repeated measures ANOVA (Analysis of Variance) was carried out for the sum of fixation duration, using within-subject factors: *Captioning_Style* (CC, EC), *AOI_Category* (Video, Captioning, SID), *Scene* (A, D), and between-subjects factor: *Hearing_Group* (deaf, hard-of-hearing).

Naming Convention of Variables. The variables used for this analysis are named as follows: *Captioning_Style.Scene.AOI_Category*. For example, EC.A.Captioning would be the fixation duration for the AOI Category: Captioning of Scene A using the Enhanced Captioning style. If a within-subjects factor is not mentioned, all cases are considered.

Data Aggregation. As the number of fixations varied within and between participants for each condition (within-subject factors), the eye-tracking data was aggregated in

order to create a normalized and consistent set of data. A pivot table of this data was created using the participant ID as rows, and the conditions as columns and filters, where the values are the sum of the fixation duration. This allowed the ability to analyze this data using repeated measures ANOVA and paired T-Tests.

Data Analysis. As there is no SID category for closed captioning (not always shown and fixed size/length, too small for a reliable AOI), two separate analyses were conducted:

1. Captioning Style = Enhanced Captioning.

The first analysis is only the Enhanced Captioning style, but contains all of the three categories: *Video*, *Captioning*, *SID*.

2. AOI Category = Video + Captioning.

The second analysis includes Enhanced Captioning and Closed Captioning, but only the categories: *Video* and *Captioning*.

Analysis #1: Captioning Style = Enhanced Captioning

There was a significant interaction effect between *Scene * AOI_Category* [$F(2, 34) = 11.09, p < 0.05$]. There was a significant main effect for *AOI_Category* [$F(1.51, 25.63) = 17.84, p < 0.05$] with a Huynh-Feldt adjustment as the Test of Sphericity was significant, $p < 0.05$. There was no significance for the between-subjects factor: *Hearing_Group*.

Paired T-Tests were then conducted for the *Scene * AOI_Category* interaction and *AOI_Category* main effect.

There was only one significant result for the interaction: *EC.A.Video* and *EC.D.Video* [$t(18) = -4.715, p < 0.05$]. The Video was viewed less in *Scene A* ($M = 328.60, SD = 219.62$) than in *Scene D* ($M = 526.62, SD = 235.15$). There were no other significant for the interactions between *EC.A.Captioning* and *EC.D.Captioning*, or *EC.A.SID* and *EC.D.SID*. The means and standard deviations for these variables are shown in Figure 16.

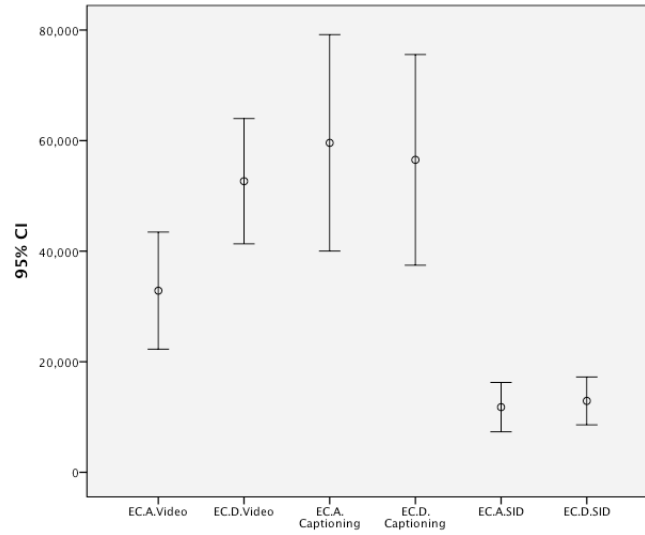


Figure 16 Sum of Fixation Duration: Scene * AOI_Category (Style = EC)

There were two significant results for the main effect: *EC.Video* and *EC.SID*: [$t(18) = 7.099, p < 0.05$], *EC.Captioning* and *EC.SID*: [$t(18) = 6.140, p < 0.05$]. The *Video* ($M = 855.22, SD = 416.59$) and *Captioning* ($M = 1161.14, SD = 757.89$) were viewed more compared to the *SID* ($M = 247.01, SD = 166.01$). There was no significant main effect for *EC.Video* and *EC.Captioning*. The means and standard deviations for these variables are shown in Figure 17.

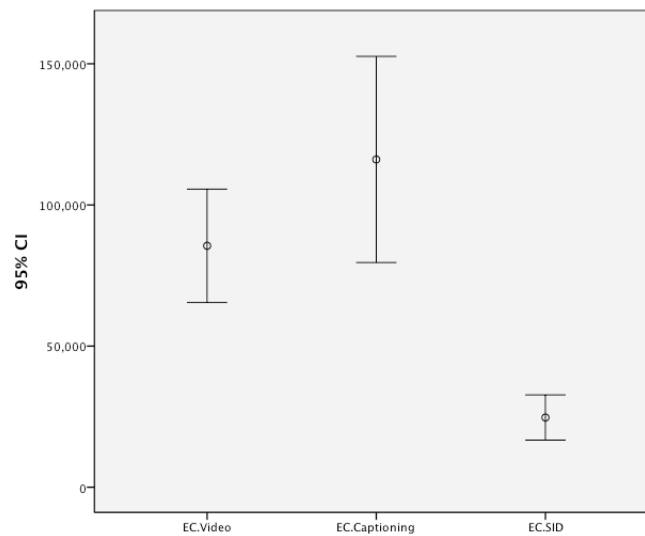


Figure 17 Sum of Fixation Duration: AOI_Category (Style = EC)

Analysis #2: AOI Category = Video + Captioning

There was a significant interaction effect between *Scene * AOI_Category* [$F(1, 17) = 19.55, p < 0.05$]. There was no significant difference for the between-subjects factor: *Hearing_Group*.

Paired T-Tests were then conducted for the *Scene * AOI_Category* interaction.

There was one significant result for the interaction: *A.Video* and *D.Video*: [$t(18) = -5.519, p < 0.05$]. The Video was viewed more in Scene D ($M = 728.47, SD = 290.13$) than in Scene A ($M = 433.92, SD = 289.17$). The means and standard deviations for these variables are shown in Figure 18.

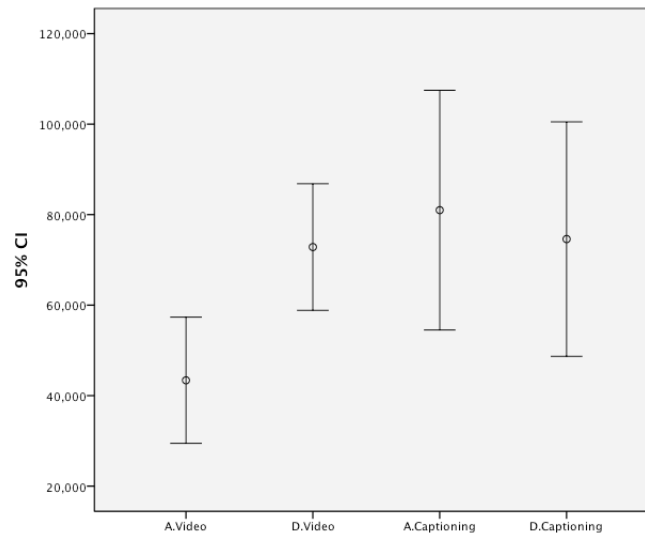


Figure 18 Sum of Fixation Duration: Scene * AOI_Category (Video, Captioning)

No Main Effect Significance. There was no main effect for any of the within-subjects factors: *Scene*, *Captioning_Style*, or *AOI_Category*.

4.2.2 Repeated Measures for Understanding, Distraction, and Preferences

A repeated measures ANOVA was carried out for the understanding, distraction and preference of the speaker identification components using within-subjects variables: *Captioning_Style* (CC, EC1, EC2, EC3), *Scene* (A, B, C, D), and between-subjects factor: *Hearing_Group* (deaf, hard-of-hearing).

The 5-point Likert-scales were reduced to 3-point Likert-scales due to the low number of participants ($N = 19$). For the 3-point Likert-scale, ratings of 1 = positive (e.g., liked, not distracting, helpful), 2 = neutral or no opinion, and 3 = negative (e.g., disliked, distracting, not helpful).

Understanding of Speaker's Name. There was a significant interaction effect between *Scene * Hearing_Group* [$F(3, 15) = 3.07, p < 0.05$]. Paired T-Tests showed significance between Scenes A and B, $p < 0.05$. For Scene A, the hard-of-hearing group found the speaker's name less helpful than the deaf group. For Scene B, this was the opposite, where the deaf group found the speaker's name less helpful than the hard-of-hearing group (see Table 4).

Style	Scene	hoh			deaf		
		Mean	SD	n	Mean	SD	n
EC1	A	2.86	0.38	7	2.55	0.82	11
	B	2.14	1.07	7	2.67	0.78	12
	C	2.33	1.03	6	2.30	0.95	10
	D	2.33	1.03	6	2.73	0.65	11
EC3	A	2.86	0.38	7	2.80	0.63	10
	B	2.17	0.98	6	2.82	0.60	11
	C	2.60	0.89	5	2.55	0.82	11
	D	2.50	0.84	6	2.80	0.42	10

Table 4 Descriptives for Understanding of Speaker's Name

Distraction of Speaker's Name. There was a significant interaction between *Scene * Hearing_Group* [$F(3, 14) = 5.31, p < 0.05$], and a main effect for *Scene* [$F(3, 14) = 3.65, p < 0.05$]. Paired T-Tests showed significance between Scene A and D, $p < 0.05$. In general, the hard-of-hearing group found the speaker's name more distracting than the

deaf group. The exception is for Scene A for EC1 and EC3, and Scene D for EC3, where the hard-of-hearing group found the speaker's name significantly less distracting. Overall, the speaker's name was less distracting in Scene A than Scene D (see Table 5).

Style	Scene	hoh			deaf		
		Mean	SD	n	Mean	SD	n
EC1	A	1.14	0.38	7	1.82	0.98	11
	B	1.86	1.07	7	1.67	0.98	12
	C	2.00	1.10	6	1.80	1.03	10
	D	1.83	0.98	6	1.73	1.01	11
EC3	A	1.00	0.00	7	1.90	0.88	10
	B	2.17	0.98	6	1.73	0.79	11
	C	1.80	1.10	5	1.64	0.92	11
	D	1.67	1.03	6	2.30	0.95	10

Table 5 Descriptives for Distraction of Speaker's Name

Preference of Speaker's Name. There was a main effect for *Scene* [$F(3, 15) = 3.25, p < 0.05$]. Paired T-Tests showed significance between Scene A and C, $p < 0.10$. In generally, the speaker's name was least preferred in Scene A than Scene C (see Table 6).

Style	Scene	hoh			deaf		
		Mean	SD	n	Mean	SD	n
EC1	A	2.86	0.38	7	2.64	0.81	11
	B	2.14	1.07	7	2.58	0.79	12
	C	2.00	1.10	6	2.20	1.03	10
	D	2.33	1.03	6	2.73	0.65	11
EC3	A	2.71	0.76	7	2.60	0.70	10
	B	2.17	0.98	6	2.64	0.67	11
	C	2.60	0.89	5	2.55	0.82	11
	D	2.33	1.03	6	2.70	0.67	10

Table 6 Descriptives for Preference of Speaker's Name

No Significance for Other Components. There was no other significant difference for the other SID components: avatar, and coloured border.

Comprehension Questions. The majority of participants were able to correctly answer the multi-choice questions, and/or describe the basic content (dialogue) of each scene. The exception was a few participants because they could do not remember (1 deaf, 1 hard-of-hearing), or found the captioning being “too fast” (1 deaf, 1 hard-of-hearing) or “too distracting” (1 deaf) to recall.

4.2.3 Preference of EC Components (Crosstabs)

Crosstabs were performed for the preference of the EC: SID components: image, label, coloured border, and dialogue. There was no significance for any of the preferences of the individual components and the hearing status.

There were two participants who did not complete all of the questions: 1 deaf did not complete any of the post-study questionnaire, and 1 hard-of-hearing who did not rate their preference of the coloured border. The n for each group is adjusted accordingly and displayed along with the mean and standard deviations.

A Likert-scale of preference (1 = disliked, 2 = neutral, 3 = liked) was carried for the components of the EC: SID system, which includes the dialogue and each of the identifying elements of the avatar (see Table 7). In general, the hard-of-hearing group liked the image and the label (name) of the speaker the most, and disliked the coloured border and dialogue. The deaf group liked only the label (name) of the speaker, and disliked the other components.

	hoh			deaf		
	n	Mean	SD	n	Mean	SD
Image	7	2.43	0.98	11	1.55	0.93
Coloured Border	6	1.33	0.82	11	1.91	1.04
Label	7	2.43	0.98	11	3.00	0.00
Dialogue	7	1.00	0.00	11	1.64	0.92

Table 7 Preference of EC: SID components

Preference for Image. The majority of the deaf group (73%) disliked, while the majority of the hard-of-hearing group (71%) liked the image of the speaker (see Table 8).

Although this relationship is not significant [$\chi^2 (1, N = 18) = 3.38, p = 0.07$], the effect size (Cramer's $V = 0.43, p = 0.07$) is moderate (Cohen 1988).

	Frequency			n	Percent		
	dislike	neutral	liked		dislike	neutral	liked
hoh	2	0	5	7	29%	0%	71%
deaf	8	0	3	11	73%	0%	27%

Table 8 Preference of Image for SID by Hearing Status

Preference for Coloured Border. The majority of the deaf group (55%) and hard-of-hearing group (83%) disliked the coloured border (see Table 9). This relationship is not significant [$\chi^2 (1, N = 17) = 1.41, p = 0.24$], and the effect size (Cramer's $V = 0.29, p = 0.24$) is weak (Cohen 1988).

	Frequency			n	Percent		
	dislike	neutral	liked		dislike	neutral	liked
hoh	5	0	1	6	83%	0%	17%
deaf	6	0	5	11	55%	0%	45%

Table 9 Preference of Image for SID by Hearing Status

Preference for Label. The deaf group (100%) and the majority of the hard-of-hearing group (71%) liked the label for the speaker's name (see Table 10). Although this relationship is not significant [$\chi^2 (1, N = 18) = 3.54, p = 0.06$], the effect size (Cramer's $V = 0.44, p = 0.06$) is moderate (Cohen 1988).

	Frequency			n	Percent		
	dislike	neutral	liked		dislike	neutral	liked
hoh	2	0	5	7	29%	0%	71%
deaf	0	0	11	11	0%	0%	100%

Table 10 Preference of Label for SID by Hearing Status

Preference for Staggered Dialogue. The hard-of-hearing group (100%) and the majority of deaf group (64%) disliked the dialogue (see Table 11). Although this relationship is not significant [$\chi^2 (2, N = 18) = 3.27, p = 0.20$], the effect size (Cramer's $V = 0.43, p = 0.20$) is moderate (Cohen 1988).

	Frequency			n	Percent		
	dislike	neutral	liked		dislike	neutral	liked
hoh	7	0	0	7	100%	0%	0%
deaf	7	1	3	11	64%	9%	27%

Table 11 Preference of Staggered Dialogue by Hearing Status

4.2.4 Purpose of EC Components (Crosstabs)

Crosstabs were performed for the purpose of the EC: SID components: image, label, coloured border, and dialogue.

There was no significance for any correlations between the deaf and hard-of-hearing group, and the purpose of the components.

Purpose of Avatar (Image + Coloured Border). The deaf group and majority of the hard-of-hearing group thought that the avatar was for speaker identification (see Table 12). The second majority of each group thought that that avatar was also distraction at the same time as being used for speaker identification. Although this relationship is not significant [$\chi^2 (3, N = 18) = 4.33, p = 0.22$], the effect size (Cramer's $V = 0.49, p = 0.23$) is moderate (Cohen 1988).

	hoh		deaf	
	n = 7	%	n = 11	%
id	3	43%	9	82%
distraction	1	14%		
id + distraction	2	29%	2	18%
id + emotion	1	14%		

Table 12 Purpose of Avatar by Hearing Status

Purpose of Label. The hard-of-hearing group and the majority of the deaf group thought that the purpose of the label was for speaker identification (see Table 13). Although this relationship is not significant [$\chi^2 (3, N = 18) = 2.29, p = 0.51$], the effect size (Cramer's $V = 0.36, p = 0.51$) is moderate (Cohen 1988).

	hoh		deaf	
	n = 7	%	n = 11	%
id	7	100%	8	73%
id + distraction			1	9%
emotion			1	9%
id + distraction + emotion			1	9%

Table 13 Purpose of Label by Hearing Status

Purpose of Staggered Dialogue. The majority of the hard-of-hearing group found the dialogue distracting, while the majority of the deaf group found the dialogue was for speaker identification (see Table 14). The second majority of each group thought the opposite, where the hard-of-hearing group found the dialogue was for speaker identification and the deaf group found the dialogue distracting. Although this relationship is not significant [$\chi^2 (5, N = 18) = 6.78, p = 0.24$], the effect size (Cramer's $V = 0.61, p = 0.24$) is strong (Cohen 1988).

	hoh		deaf	
	n = 7	%	n = 11	%
id	2	29%	4	36%
distraction	4	57%	2	18%
id + distraction			3	27%
timing			1	9%
emotion			1	9%
id + timing	1	14%		

Table 14 Purpose of Dialogue by Hearing Status

4.3 Discussion

The EC: SID system allows for the use of graphical and text-based identifiers to adapt to the different needs and preferences of its users. The differences between deaf and hard-of-hearing users of the results provide some initial evidence to support changing existing captioning practices, guidelines and tool design. Given that all of the enhanced caption options are feasible for either digital television or online video content, a 'one-size-fits-all' approach used in conventional television may be replaced with an approach that allows for user customisation options. Furthermore, there seems to be a learning curve for the EC: SID, which is expected, since captioning has not changed much since the early 1970s, over 40 years.

4.3.1 Participatory Design for EC: SID

In order to design a captioning system that is more effective, the researcher is required to understand needs of its users and perhaps including them in the design phase. It is important to include users of the existing system (e.g. closed captioning) because this is the baseline of the existing solution. In this case, people from the deaf and hard-of-hearing communities were recruited in order to select the best design for speaker identification, as well as evaluating its impact on viewers who are deaf or hard-of-hearing.

During the participatory design, the results from the mapping activity indicate that in some situations, captioning is often inadequate or perhaps ineffective for conveying non-speech information. The issues that were identified were related to non-verbatim captioning, speaker identification, non-speech information, sound and music, and other technical issues. These issues are similar to those identified by other researchers (Harkins, Korres, Virvan, & Singer, 1996), where their recommendations for non-speech information were to use explicit text descriptions. However, this is more of a workaround than a solution to the limitation of a text-based captioning system.

As technology, particularly computing, has progressed there are new technological options that could be explored and considered to address these issues and take

advantage of these new functionality. There have been attempts to address some of these issues. For example, emotions can be conveyed using kinetic text (Rashid, Vy, Hunt & Fels, 2008; Vy & Fels 2008). However, this solution has only been for research purposes and is not commercially available or ready to be implemented for public consumption. The EC: SID system is also for research purposes in attempt to solve one of the several problems with the problems, in this case speaker identification.

The findings from the mapping technique seem to indicate that deaf users rely on the quality/accuracy and completeness of captioning in order to understand the content. Without complete and good quality of captioning, deaf users are unable to have an equivalent entertaining experience similar to their hearing counterparts. However, during the brainstorming of ideas for solutions for the virtual mapping, the deaf volunteers could not think “outside of the box”, which is the television or captioning decoder. The deaf volunteers only suggest using text formatting (e.g. italic and colour) for differentiating non-speech information. Instead, the researcher suggested the use of an image of the speaker, where the deaf volunteer suggested adding back the speaker’s name. The coloured border was added to further clarify who was the speaker.

This is a shortcoming of participatory design where users are not necessary capable of being designers and their suggestions may be limited to their knowledge. However, this should not deter that working along with users has the potential of creating a system that is practical and useful for them. Furthermore, some effort was require from both parties in order for suggestions to be critiqued and not just pleasing each other. This is usually would be the first time for both users and the researchers, however these issues should be considered, in order to allow for a successful collaboration for participatory design.

4.3.2 Eye Tracking Behaviour of Captioning Systems

The results from the eye-tracking analysis indicate that there was a significant interaction between the AOIs and Scenes, and a main effect in the enhanced captioning system for SID with the other categories: Video and Captioning.

In general, the Video category was viewed significantly less in Scene A than Scene D. This was expected, as Scene D was more complex: where up to 3 of the 7 speakers would be conversing at a time during the 17 changes in the camera angles. On the other hand, Scene A was less complicated, with only one speaker narrating off-screen and the camera slowly tracking the Cube (the object on screen) with no changes to the viewing angle.

For EC only, the SID category was viewed significantly less than Video or Captioning. This finding was hoped for since, this may indicate that less is required to identify the speaker, or that the SID was ignored (not useful or distracting). However, the majority of responses from each group (45% hard-of-hearing, 82% deaf) determined that the purpose of the SID was for identifying the speaker. This would support the idea that using an image (e.g., avatar) may perhaps require less time to process compared to only using a text-based identifier (e.g., speaker's name).

4.3.3 Use of Identifiers

Purpose and Distraction. In general, the components were determined by the groups to be used for speaker identification. However, the avatar and placement of the dialogue were distracting as well for some participants. This is perhaps due to the switching between the left and right sides of the captioning panel, which was considered unnecessary especially for the same speaker (e.g. Scene A). This would support the preference of subtle changes of images compared to moving images (Cooper et al. 2006).

Below are some of the comments from participants that would support this theory:

deaf: "avatar does not slow me down as I like to read captions before spoken words. Staggered method takes my eyes time away from video."

deaf: "left-right bad, top-bottom okay"

deaf: "I'm getting understanding how it works, but wow that's crazy dialogues!"

hard-of-hearing: "use of avatar (especially two for same speaker) is downright distracting, colour border somewhat helpful, but can be distracting"

hard-of-hearing: "avatar helpful and comforting except when at both sides (two speakers) in which case was distracting"

Figure 19 Comments from regarding understanding and distracting SID components

The left-right position changes in the dialogue seem to be distracting, and likely affecting the effectiveness of the avatar. Therefore, changes to the placement of the captioning should perhaps be minimized, regardless of any change in the location of the speaker on-screen, which may not be as important for speaker identification.

Preference. In general, the hard-of-hearing group liked the label and image of the speaker, and disliked the coloured border and dialogue, while the deaf group liked only the label, but disliked the other components. In common, both groups liked the label, and disliked the dialogue and coloured border. However, only the hard-of-hearing liked the image of the speaker. This was unexpected, as people who are deaf are usually more visual, in terms of their thinking (e.g. sign language). The negative ratings of the deaf group are likely because they are perhaps not accustomed to this dramatic change in the captioning system from closed captioning to EC: SID.

Below are some of the comments from participants that would support this theory:

deaf: *"would have to get used to it [avatars], don't mind speaker's name or the coloured border or staggered dialogues. interesting experience"*

deaf: *"dislike absence of identification in 'busy' dialogue"*

hard-of-hearing: *"most important are avatar/image + dialogue for best understanding of what is going on"*

hard-of-hearing: *"icons [avatars] very effective, be careful not to reveal plot twists with names/icons, otherwise great! Use of colour also helpful"*

Figure 20 Comments from regarding preference of SID components

4.3.4 Using Label for Speaker Identification

As there was only significance for the label (e.g. speaker's name) from the questionnaire results, these findings will be discussed:

Understanding of the Label. The results indicated that there was a significant interaction effect between Scene and *Hearing_Group*. The hard-of-hearing group found the speaker's name less helpful for Scene A, and more helpful for Scene B than the deaf group. The main difference between these scenes is the words per minute (WPM) and the number of speakers. Scene A was a slow narration (120 WPM) of one speaker off-screen, while Scene B was four speakers who were speaking faster (188 WPM).

This may indicate that when there is no change in the speaker (Scene A), the speaker's name is not helpful. This is expected, as if there is not change, the speaker's name is not necessary. As such, the speaker's name should perhaps only be shown when there is a change of the speaker, and not always visible.

Distractibility of the Label. The results indicated that there was a significant interaction effect between Scene and *Hearing_Group*, and a main effect for Scene. The hard-of-hearing group found the speaker's name less distracting than the deaf group for Scenes A and D. In general, the speaker's name was less distracting in Scene A than in Scene D. The main difference between these scenes is the number of speakers:

Scene A with one speaker off-screen, and Scene D with seven speakers in total, where up to 3 of them are speaking at a time.

This may indicate that when lots of changes in the speaker, the speaker's name is likely ignored by the hard-of-hearing group, but is distracting for the deaf group. This is expected, as unlike the hard-of-hearing group, the deaf group is unable to detect any changes in the voice, but only visually through movement (e.g., lips or hand gestures) on-screen. This may indicate that the speaker's name perhaps not as helpful and would be distracting in these cases, especially for viewers who are deaf.

Preference of the Label. The results indicated that there was a main effect for *Scene*. In general, the speaker's name was disliked more in Scene A than Scene C. The main difference between these scenes is the words per minute (WPM) and the number of speakers. Scene A was a slow narration (120 WPM) of one speaker off-screen, while Scene C was three speaker's who were speaking faster (194 WPM).

This may indicate that when there are no changes in the speaker, the speaker's name is least preferred. This is similar to the understanding of the speaker's name, where if there is not change in the speaker, the speaker's name is perhaps not required. Therefore, the speaker's name should perhaps be shown only when there is a change of the speaker, and not always visible.

4.3.5 Limitations

Content. The content for the system evaluation was only from one particular film, which was a live action movie with humans and computer-generated characters. This may have influenced the results because participants liked or disliked the movie. Although their preference of the content is unavoidable, various genres could be used for the content of the system evaluation in order to mitigate this effect and minimize any possible learning effect. The presence of the computer-animated character was also a factor, as they do not have any of the facial expressions that a human character would normally exhibit.

Low N. Despite numerous efforts and venues, there were a low number of participants who were recruited for the system design and evaluation. This was especially the case for the participant design phase when a lot of time and effort is required from both parties. In addition, the hearing groups were unbalanced between deaf and hard-of-hearing people for the system evaluation. This was unavoidable, as the number of respondents was unpredictable. As there was eye-tracking involved in the system evaluation, participants had to be physically present in the research lab, which prevented a few people who lived far away from participating.

This low number of participants, which is typically in Human-Computer Interaction, does not represent the normal distribution of the population. As such, the findings of this research are limited to this sample, and may not necessarily reflect the entire population. Furthermore, the number of people in each group is unbalanced especially for the hard-of-hearing group. Therefore, further research with a larger sample of the deaf and hard-of-hearing population would be required, before these findings could be generalized.

Communication Barriers. It is important to note that sign language has no written form and users tend to have reduced abilities to express themselves as clearly in a written language as they do in sign language. Interpretation of written comments expressed by sign language users may require particular care and clarification by the researcher, who is hearing. For example, one participant expressed frustration with typing on the keyboard, as this was a slow process: “English is not my tongue! and eyes. ASL is my language :(”. Sign language interpreters are often employed to assist in the communication process between people who are hearing and deaf. There also has been an inclusive methodology such as the Gestural Talk Aloud method that has been developed with interpreters (Roberts & Fels 2005). However, the constant and frequent need to involve deaf users for participatory design makes having a dedicated interpreter expensive and not practical. Also the interpreter needs to be very knowledgeable in this technical domain, which is difficult to obtain as well.

Testing Environment. The system evaluation was conducted in an empty research lab with a computer screen. This is not the natural environment, in which the content is usually consumed such as at the movie theatre or the living room, with a large screen in a comfortable and dark setting, and sitting further away. Obtaining such of a testing environment was not possible due to the limited resource available at the university.

Findings from Eye-Tracking. This is a limitation of any eye-tracking methodology, as eye behaviour (e.g. gaze fixation, and scan path) does not necessarily indicate cognitive workload or attention (Rayner 1998). A combination of eye-tracking and Electroencephalography (EEG) may perhaps obtain better measures of attention and cognitive load. Therefore, the findings from the analysis of eye-tracking data are only an approximation of changes in the attention or cognitive load.

Eye-Tracking Software. Although this is not really a limitation, but is more of problems that occur while using the ClearView software. This is being mentioned to warn others that using eye-tracking may be more effort than its worth. If anything, this is a limitation of patience and determination from bad customer support. These technical problems made using eye-tracking difficult, but after a long and unnecessary struggle, possible for this research.

The eye-tracking software (ClearView) quickly became discontinued, as there was a long delay (over 15 months) between ordering and obtaining the equipment through a third party: Noldus Information Technology Inc. This resulted in wasted hours sorting out the technical difficulties that occurred with little help from the only one “Support Engineer” from Tobii who was knowledgeable about this software. For example, the researcher had to instead the software (probably violating the license agreement) just to analyse the code in order to determine the expected input of a file, which was not documented in the user manual. The software allowed importing a file for pre-defined scenes, but there was no export feature, after defining them using the software. Otherwise, defining every scene for each video recording would have been inefficient use of time, and would have been exhausting if this was done manually.

The ClearView software that was provided was limited in that it was not able to record a “live screen”, but instead only images or videos as the stimulus. The license key that was initially given was restrictive in enabling this feature. Instead, additional preparation was required to create a video file of each condition using a screen captioning software in real time. It was only after running the experiment and complains to Noldus and Tobii that the license key was “upgraded” with one that was not as restricted and “free-of-charge”. However, the license key, which was unlimited in the features, was time-limited and expired after 1 month. Another request was sent again, which asked for another license key that was unrestricted in features and time.

Chapter 5: Conclusion

5.1 Summary of Findings

The current method for speaker identification, which is text-based (e.g. name, chevrons, and colours), is often not sufficient to determine who is speaking for viewers who are deaf or hard-of-hearing. Instead, the viewer often must rely on movement (e.g. lip or hand gestures) on-screen for a better indication, but this is difficult when there are multiple characters or speakers that are off-screen. A captioning system (EC: SID) was developed to address this issue of speaker identification, using participatory design and evaluated with people who are deaf and hard-of-hearing. The system uses an image of the character, surrounded by a coloured border that matches their clothing, as well as their name to indicate the speaker. . The EC: SID captioning system provides the ability for customizing the layout and configurations of its components. In particular, the components used for speaker identification are customizable and can be toggled on or off by the user.

The system was evaluated with a representative sample of people who are deaf or hard of hearing using questionnaires to gather opinions about the system and eye tracking to gather quantitative data regarding where people were looking on the screen. The main findings and observations from this research indicate that there are differences in the needs and requirements of the various users of captioning. This is opposite from the existing captioning system, where the captioning appears the same for all users, whether they are deaf or hard-of-hearing.

There is also a difference in using the graphical and text-based identifiers, depending on the content and the user. When there is no change in the speaker, some participants found the speaker's name distracting and not helpful. As such, the speaker's name should perhaps only appear when there is a change in the speaker, and not always displayed. This is especially the case for when there is a narration of a speaker who is off-screen. When there are many changes in the speaker, some deaf

participants find the speaker's name more distracting, whereas the hard-of-hearing participants may simply ignore the speaker's name. This seems to indicate that the speaker's name is perhaps not always effective, depending on the user as well as the content.

The placement of the captioning near the character was found to be not as important for indicating the speaker. This constant changing in order to match the location of the speaker on-screen was found to be distracting. Furthermore, this was an undesirable artefact that may have negatively affected the effectiveness of speaker identification components. As such, the location of the captioning should perhaps remain constant regardless of whether the speaker has changed location on-screen. This was especially the case when there was constantly one character that was speaking.

The findings of the eye tracking indicated that the speaker identification components were looked at significantly less than either the captioning or the video content. This was expected as video and content was the main visual display of the content, whereas the components were only for reference for indicating who is speaking. There was also significant difference where participants were looking between different scenes that were either simple or complex. This was also expected, as the complicated scene would require more attention to follow along than a simple scene with no changes in the camera angle and only one speaker, narrating off-screen.

The development of the EC: SID captioning system using participatory design was a learning experience for the researcher as well as the users. There were some problems with communicating with people who are deaf, which could be overcome with a dedicated team of interpreters who are knowledgeable with this domain. The researcher could also gain more experience with using this methodology and be more collaborative and creative, since users are not necessarily designers and have ideas for solutions. Nonetheless, participatory design is useful for creating a system that is perhaps effective and useful for the intended users: people who are deaf or hard-of-hearing. This is especially the case, when the problems of captioning have not been addressed for over 40 years.

5.2 Contributions

Designing for Speaker Identification. A graphical technique for speaker identification was developed using participatory design, which uses an image of the speaker, and a coloured border, in addition to the speaker's name. This was different from the current method, which is text-based, using either the speaker's name or a different colour. The layout of the dialogue was also changed, where the left and right side was reserved for a particular speaker. The dialogue was also divided into levels, in order to show the timings and exchange of the conversation.

Evaluation of Captioning System. This graphical speaker identification technique was implemented in a captioning system, and evaluated with people who are deaf and hard-of-hearing, the targeted users of these systems. An eye tracking methodology was also used, in order to obtain gaze fixation behaviours for the possible configurations of the enhanced captioning system. The findings are that the speaker identification components were looked at the least, compared to either the video or the captioning.

5.2 Future Work

System Design Changes. The results from the system evaluation indicate some design changes would be required for the EC: SID captioning system to become more usable. These include keeping the placement of the captioning constant, and the reducing the frequency of displaying the speaker's name. The location of the captioning should not change once displayed, regardless of where the character is located afterwards. The speaker's name should only be displayed when there is a new speaker, and not always be visible, if still speaking.

Addressing Limitations. The influence of the preference of a particular content may be address by using various genres, and not limited to one film. This is to avoid any biased introduced by the preference of participants for that content. Attempts should be made so that the number of participants is increased and there is balanced

between the deaf and hard-of-hearing groups. This may involve expanding the recruiting at other events and locations and increasing the duration for conducting the user study from 3 months to longer. A dedicated team of interpreters, who are knowledgeable in this domain, could be used during the system evaluation to increase the efficiency of obtaining feedback from participants who are deaf. The researcher could also take more ASL classes in order to improve his communication skills. A more natural environment could be used for testing, such as setting up a usability lab, which includes a couch and large television screen. Further longitudinal studies are also required in order to determine the long-term usefulness for and acceptance of this captioning system for film and movies, or post-production television. In addition, other measures such as Electroencephalography (EEG) could be used for measuring the cognitive load and attention of participants to balance the limitations of eye-tracking.

Practical and Commercial Application. The practicality of this speaker identification technique could also be explored. Currently, there is additional work required to caption content using this graphical speaker identification technique and whether this extra work is feasible in the current practise for captioning is unknown. However, it may be possible to use automatic audio and video analysis techniques to help identify the speaker, and track their location on-screen, and assign the location and create live avatars, in place of static images. This research of analysing the video and script has been done already (Everingham et al. 2009). However, collaboration of these research projects could provide a commercial application for the EC: SID captioning system.

Appendix A: Enhanced Captioning: SID

Source Code

The source code may be obtained by sending a request to the following:

qvy@ryerson.ca

List of Files

\Engine.swf

\Settings.xml

\resources\speakers.xml

\resources\Scenes*.xml

\resources\Videos*.mp4

\resources\Avatars*.png

\Engine fla

\Engine.as

\EnACT\Avatar.as

\EnACT\Caption.as

\EnACT\Captions.as

\EnACT\CuePoint.as

\EnACT\Panel.as

\EnACT\Settings.as

\EnACT\Speakers.as

\EnACT\Window.as

Appendix B: System Evaluation

Research Ethics Board



To: Quoc Vy
Computer Science
Re: REB 2009-128: Enhanced Captioning - Using Avatars for Improving Speaker Identification
Date: May 25, 2009

Dear Quoc Vy,

The review of your protocol REB File REB 2009-128 is now complete. The project has been approved for a one year period. Please note that before proceeding with your project, compliance with other required University approvals/certifications, institutional requirements, or governmental authorizations may be required.

This approval may be extended after one year upon request. Please be advised that if the project is not renewed, approval will expire and no more research involving humans may take place. If this is a funded project, access to research funds may also be affected.

Please note that REB approval policies require that you adhere strictly to the protocol as last reviewed by the REB and that any modifications must be approved by the Board before they can be implemented. Adverse or unexpected events must be reported to the REB as soon as possible with an indication from the Principal Investigator as to how, in the view of the Principal Investigator, these events affect the continuation of the protocol.

Finally, if research subjects are in the care of a health facility, at a school, or other institution or community organization, it is the responsibility of the Principal Investigator to ensure that the ethical guidelines and approvals of those facilities or institutions are obtained and filed with the REB prior to the initiation of any research.

Please quote your REB file number (REB 2009-128) on future correspondence.

Congratulations and best of luck in conducting your research.

A handwritten signature in black ink, appearing to read "Nancy Walton".

Nancy Walton, Ph.D.
Chair, Research Ethics Board

Research Title: Enhanced Captioning - Speaker Identification

Principal Investigators: Deborah Fels, Ph.D, P.Eng
Ted Rogers School of Information Technology Management, Ryerson University

Quoc Vy, BSc
Department of Computer Science, Ryerson University

Study - Information Sheet

Instructions. You are being asked to participate in a research study. Before you give your consent to be a volunteer, it is important that you read the following information and ask as many questions as necessary to be sure you understand what you will be asked to do.

Purpose of Study. The purpose of this study is to evaluate a prototype of a captioning system which uses graphics (e.g. images and colour) along with text descriptions. This graphical captioning system will be compared with the conventional style of closed captioning which only uses text descriptions.

Description of Study. You will be asked to watch several video clips from a movie (rated PG-13) with various styles of captioning from these two captioning systems. An eye-tracking device which is non-intrusive (e.g. infrared light) will be used to observe the viewer's eye patterns and behaviours.

Duration of Study. The expected duration of the study is about 55 - 60 minutes. You will be allowed to take breaks throughout the study and light refreshments will be provided. You may also withdraw from the study at any time.

<u>List of Tasks for Study</u>	<u>Estimated Time</u>
<input type="checkbox"/> <i>Complete pre-study questionnaire</i> - to obtain demographic information and experience with closed captioning	~7 minutes
<input type="checkbox"/> <i>Conduct study with eye-tracking device</i> - view 4 video clips with 4 different captioning styles and configurations - complete questionnaire for each trial (to obtain feedback and comments)	~50 minutes
<input type="checkbox"/> <i>Complete post-study questionnaire/interview</i> - to obtain final comments and opinions of graphical captioning system	~3 minutes

Recording. You may be video recorded to capture your responses and reactions. If an interpreter is present, their voice translation may be recorded instead.

Location of Study. The study will be conducted in a room with a computer screen and eye-tracking device. The room is located in the *Ted Rogers School of Management* at Ryerson University.

Expected Risks. You may experience some discomfort or fatigue while performing these various tasks as they are repetitive. To minimize this risk, you will be given adequate time to complete each task and may take breaks between the trials.

Potential Benefits. Your participation will help us develop a new and improved captioning system that may eventually replace the current captioning system. There are no expected benefits to you from participating in this study.

Confidentiality. Although you may be identified in the video recording, your identity will be kept confidential. Any data collected, including any video recordings, will be only accessible by the researchers for the purpose of data analysis.

Incentives to Participate. You will be offered a payment of \$30 to cover the cost of travel and your time for participating in this study. You will be given this payment at the end of the study.

Voluntary Nature of Participation. Participation in this study is voluntary. Your choice of whether or not to participate will not influence your future relations with Ryerson University. If you decide to participate, you are free to withdraw your consent and to stop your participation at any time without penalty or loss of benefits to which you are allowed. At any particular point in the study, you may refuse to answer any particular question or stop participation altogether.

Questions About Study. If you have any questions about the research now, please ask. If you have questions later about the research, you may contact:

Dr. Deborah Fels

Phone: 416 979 5000 ext. 7619

Email: dfels@ryerson.ca

Research Ethics Board. This study has been approved by the Ryerson Research Ethics Board (REB). If you have questions regarding your rights as a participant in this study, you may contact the Ryerson University Research Ethics Board for information:

Research Ethics Board

Phone: 416 979 5042

Mailing: c/o Office of the Vice President, Research and Innovation
Ryerson University
350 Victoria Street
Toronto, ON M5B 2K3
CANADA

Research Title: Enhanced Captioning - Speaker Identification

Principal Investigators: Deborah Fels, Ph.D, P.Eng
Ted Rogers School of Information Technology Management, Ryerson University

Quoc Vy, BSc
Department of Computer Science, Ryerson University

Study - Consent Agreement

Your signature below indicates all of the following:

- You have read the information in this agreement and have had a chance to ask any questions you have about the study
- You agree to be in the study and have been told that you can change your mind and withdraw your consent to participate at any time
- You give permission for yourself and/or the interpretation of your signing to be videotaped
- You have been given a copy of this agreement
- By signing this consent agreement, you are not giving up any of your legal right

Name of Participant (please print)

Signature of Participant

≈Date: ____ / ____ / 2009
DD MM

Signature of Investigator (Quoc Vy)

Date: ____ / ____ / 2009
DD MM

Honorarium - Receipt:

By signing below, I acknowledge that I have received **\$ 30.00** as an honorarium for participating in the **Enhanced Captioning - Speaker Identification** study.

Signature of Participant

Date: ____ / ____ / 2009
DD MM

Questionnaires

Pre-Study Questionnaire

Demographics

Q1. What is your **hearing status**?

- ☐ hearing
- ☐ cochlear implant
- ☐ hard of hearing
- ☐ deafened
- ☐ deaf

Q2. What is your **gender**?

- ☐ male
- ☐ female

Q3. What is your **age**?

- | | | |
|----------------------------------|----------------------------------|-------------------------------------|
| <input type="checkbox"/> 18 - 19 | <input type="checkbox"/> 35 - 39 | <input type="checkbox"/> 55 - 59 |
| <input type="checkbox"/> 20 - 24 | <input type="checkbox"/> 40 - 44 | <input type="checkbox"/> 60 - 64 |
| <input type="checkbox"/> 25 - 29 | <input type="checkbox"/> 45 - 49 | <input type="checkbox"/> 65 - 69 |
| <input type="checkbox"/> 30 - 34 | <input type="checkbox"/> 50 - 54 | <input type="checkbox"/> 70 or over |

Q4. What is your highest level of **education completed**?

- ☐ no formal education or did not graduate from high school
- ☐ high school (or equivalent)
- ☐ college (Diploma, 1 - 3 years)
- ☐ university (Bachelor's Degree, 4+ years)
- ☐ graduate school (Masters or PhD)

Television and Closed Captioning - Experience

Q5. How many **hours of television and movies** (e.g. Blu-ray, DVDs, VHS) do you watch **per week**?

- | | | |
|--------------------------------------|--|---|
| <input type="checkbox"/> 0 - 1 hour | <input type="checkbox"/> 8 - 9 hours | <input type="checkbox"/> 16 - 17 hours |
| <input type="checkbox"/> 2 - 3 hours | <input type="checkbox"/> 10 - 11 hours | <input type="checkbox"/> 18 - 19 hours |
| <input type="checkbox"/> 4 - 5 hours | <input type="checkbox"/> 12 - 13 hours | <input type="checkbox"/> 20 hours or more |
| <input type="checkbox"/> 6 - 7 hours | <input type="checkbox"/> 14 - 15 hours | |


Q6. How often do you use **closed captioning** while watching television or movies?

- ☐ always
- ☐ often
- ☐ sometimes

- ☐ seldom
☐ never

Closed Captioning – Opinions

Q7. Please **rate your opinion** on the following **attributes** of **closed captioning**:

	Very Much Like	Somewhat Like	Neutral / No Opinion	Somewhat Dislike	Very Much Dislike	No Opinion
Presentation Rate (speed)	1	2	3	4	5	<input type="checkbox"/>
Placement / Location (e.g. not blocking objects)	1	2	3	4	5	<input type="checkbox"/>
Use of Text Descriptions	1	2	3	4	5	<input type="checkbox"/>
Use of Symbols (e.g. music note: )	1	2	3	4	5	<input type="checkbox"/>
Use of Colour (white text on black)	1	2	3	4	5	<input type="checkbox"/>

Q8. Please **rate your opinion** on the following **methods** for **indicating speakers** found in **closed captioning**:

	Very Effective	Somewhat Effective	Neutral / No Opinion	Somewhat Not Effective	Not Very Effective	Never Seen This Before
>>Hello.	1	2	3	4	5	<input type="checkbox"/>
ANNE: Hello.	1	2	3	4	5	<input type="checkbox"/>
>>ANNE: Hello.	1	2	3	4	5	<input type="checkbox"/>
(ANNE) Hello	1	2	3	4	5	<input type="checkbox"/>
Hello (near the speaker)	1	2	3	4	5	<input type="checkbox"/>

Non-Speech Identification - Representation

Q9. Please **choose** (check-off) the various **methods** (using text and/or graphics) and **properties** (using different font styles, colour, and/or animation) that you think may be used to represent:

	Text Descriptions	Graphics: Icons, Symbols	Underline, Bold, Italic	Colour	Animation
speaker identification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
sound effects	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
music	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
speech prosody (rhythm, stress, intonation)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q10. Based on your answer in the previous question, **sketch** a particular example of how you think **speaker identification** should be indicated

Q11. Do you have any other suggestions or comments in improving closed captioning and/or speaker identification?

Trial Questionnaire

Closed Captioning - Questionnaire

Understanding

Q1. How difficult was it to determine who was speaking using **regular** closed captioning in the video clip that you have just seen?

	Very Difficult	Somewhat Difficult	Neutral / No Opinion	Somewhat Easy	Very Easy
Who's Saying What?	1	2	3	4	5

Q2. Please **rate how effective** was the following styles used for **speaker identification** in the clip that you have just seen in helping you understand the video clip:

	Very Effective	Somewhat Effective	Neutral / No Opinion	Somewhat Not Effective	Not Very Effective	Did Not See This Used
>>	1	2	3	4	5	<input type="checkbox"/>
SPEAKER:	1	2	3	4	5	<input type="checkbox"/>
>>SPEAKER:	1	2	3	4	5	<input type="checkbox"/>
(SPEAKER)	1	2	3	4	5	<input type="checkbox"/>
None (dialogue only)	1	2	3	4	5	<input type="checkbox"/>

Comments (Likes and Dislikes):

Enhanced Cautioning - Trial Questionnaire

Q1. Please rate the level of **comprehension** in the use of each of the following components:

	Very Helpful	Somewhat Helpful	Neutral / No Opinion	Somewhat Not Helpful	Not Very Helpful	Did Not Notice
Avatar / Image on / off	1	2	3	4	5	<input type="checkbox"/>
Speaker's Name on / off	1	2	3	4	5	<input type="checkbox"/>
Staggered Dialogue	1	2	3	4	5	<input type="checkbox"/>

Q2. Please rate the level of **distraction** in the use of each of the following components:

	Very Distracting	Somewhat Distracting	Neutral / No Opinion	Somewhat Not Distracting	Not Very Distracting	Did Not Notice
Avatar / Image on / off	1	2	3	4	5	<input type="checkbox"/>
Speaker's Name on / off	1	2	3	4	5	<input type="checkbox"/>
Staggered Dialogue	1	2	3	4	5	<input type="checkbox"/>

Q3. Please rate the level of **preference** in the use of each of the following components:

	Very Much Like	Somewhat Like	Neutral / No Opinion	Somewhat Dislike	Very Much Dislike	Did Not Notice
Avatar / Image on / off	1	2	3	4	5	<input type="checkbox"/>
Speaker's Name on / off	1	2	3	4	5	<input type="checkbox"/>
Staggered Dialogue	1	2	3	4	5	<input type="checkbox"/>

Comments (Likes and Dislikes):

Comprehension Questionnaire

Scene A – Introduction

Q1. Explain how the Cube was lost in space?

Scene B – Aircraft

Q1. What were some of the things that the soldiers wanted to do when they go home?

Scene C – Car Dealership

Q1. Who said: "Don't go Ricky Ricardo on me"?

- ☐ Samuel (Son)
- ☐ Ron (Father)
- ☐ Bobby (Dealer)
- ☐ I do not know / recall

Scene D – Transformers

Q1. Who said the following dialogue:
"The boy's pheromone level suggests he wants to mate with the female"?

- | | |
|--|-----------------------------------|
| <input type="checkbox"/> Megatron | <input type="checkbox"/> Jazz |
| <input type="checkbox"/> Optimus | <input type="checkbox"/> Ratchet |
| <input type="checkbox"/> Bumblebee | <input type="checkbox"/> Ironhide |
| <input type="checkbox"/> I could not tell because: _____ | |

Q2. Who is Samuel's guardian?

- | | |
|--|-----------------------------------|
| <input type="checkbox"/> Megatron | <input type="checkbox"/> Jazz |
| <input type="checkbox"/> Optimus | <input type="checkbox"/> Ratchet |
| <input type="checkbox"/> Bumblebee | <input type="checkbox"/> Ironhide |
| <input type="checkbox"/> I could not tell because: _____ | |

Q3. Who is a first lieutenant?

- | | |
|--|-----------------------------------|
| <input type="checkbox"/> Megatron | <input type="checkbox"/> Jazz |
| <input type="checkbox"/> Optimus | <input type="checkbox"/> Ratchet |
| <input type="checkbox"/> Bumblebee | <input type="checkbox"/> Ironhide |
| <input type="checkbox"/> I could not tell because: _____ | |

Post-Study Questionnaire

Q1. How difficult was it to determine who is speaking using the new graphical captioning system in the video clips that you have just seen?

	Very Difficult	Somewhat Difficult	Neutral / No Opinion	Somewhat Not Difficult	Not Very Difficult
Graphical System	1	2	3	4	5

Q2. What is the **purpose** of the **avatars / image**?
(check all that apply)

- ☐ identification of speaker
- ☐ emotion of speaker
- ☐ distraction
- ☐ other: _____

Q3. What is the **purpose** of the **speaker label**?
(check all that apply)

- ☐ identification of speaker
- ☐ emotion of speaker
- ☐ distraction
- ☐ other: _____

Q4. What is the **purpose** of the **staggered dialogue**?
(check all that apply)

- ☐ identification of speaker
- ☐ emotion of speaker
- ☐ distraction
- ☐ other: _____

Q5. Please rank each of the following **graphical features** from 1 to 4:
(1 being the most preferred and 4 being the least preferred)

_____ avatar / image	_____ speaker label
_____ coloured border	_____ staggered dialogue

Comments (Likes and Dislikes) and improvements:

Captioning of Scenes

Scene A - Narration

Start	Duration	Captioning
0:00:52	05.200	<i>OPTIMUS: Before time began, there was the Cube.
0:00:58	02.170	<i>We know not where it comes from,
0:01:00	03.770	<i>only that it holds the power to create worlds
0:01:04	02.840	<i>and fill them with life.
0:01:07	03.130	<i>That is how our race was born.
0:01:10	02.500	<i>For a time, we lived in harmony,
0:01:12	06.970	<i>but like all great power, some wanted it for good, others for evil.
0:01:19	02.800	<i>And so began the war,
0:01:22	05.500	<i>a war that ravaged our planet until it was consumed by death,
0:01:28	05.310	<i>and the Cube was lost to the far reaches of space.
0:01:33	02.540	<i>We scattered across the galaxy,
0:01:36	03.170	<i>hoping to find it and rebuild our home,
0:01:39	04.970	<i>searching every star, every world.
0:01:44	03.200	<i>And just when all hope seemed lost,
0:01:47	02.700	<i>message of a new discovery drew us
0:01:50	04.470	<i>to an unknown planet called Earth.
0:02:01	02.440	<i>But we were already too late.

Scene B - Aircraft

Start	Duration	Captioning
0:02:18	01.890	Oh, God, five months of this.
0:02:20	02.540	I can't wait to get a little taste of home.
0:02:22	02.060	<i>A plate of mama's alligators etouffee.
0:02:24	02.220	You've been talking about barbecued 'gators and crickets for the last two weeks.
0:02:27	03.430	I'm never going to your mama's house, Fig. I promise.
0:02:30	03.190	But Bobby, Bobby, 'gators are known to have the most succulent meat.
0:02:33	01.050	I understand.
0:02:34	01.740	(SPEAKING IN SPANISH)
0:02:36	01.640	(TALKING IN MOCK SPANISH)
0:02:38	02.300	English, please. English.
0:02:40	01.470	I mean, how many times have we...
0:02:42	02.300	We don't speak Spanish. I told you that.
0:02:44	02.100	Why you got to ruin it for me, man? That's my heritage.
0:02:46	01.560	(SPEAKING IN SPANISH)
0:02:49	01.230	Go with the Spanish. Whatever.
0:02:50	03.270	Hey, you guys remember weekends? Huh?
0:02:54	01.740	The Sox at Fenway.
0:02:56	03.650	Cold hotdog and a flat beer. Perfect day.
0:02:59	03.270	What about you, Captain? You got a perfect day?
0:03:03	02.840	I just can't wait to hold my baby girl for the first time.
0:03:06	01.530	FIG: He's adorable. EPPS: That's too...
0:03:07	01.160	Shut up!

Scene C - Car Dealership

Start	Duration	Captioning
0:12:37	02.670	Here? No, no, no, what is this? You said...
0:12:40	01.700	You said half a car, not half a piece of crap, Dad.
0:12:42	02.290	When I was your age, I'd have been happy with four wheels and an engine.
0:12:44	01.520	Okay, let me explain something to you. Okay?
0:12:46	01.530	<i>You ever see 40-Year-Old Virgin? Yeah.
0:12:47	01.520	Okay, that's what this is.
0:12:49	01.630	And this is 50-year-old virgin.
0:12:51	01.450	Okay. You want me to live that life?
0:12:52	02.330	No sacrifice, no victory. Yeah, no victory. You know, I got it.
0:12:54	01.350	The old Witwicky motto, Dad. Right.
0:12:56	02.000	Gentlemen.
0:12:58	04.040	Bobby Bolivia, like the country, except without the runs.
0:13:02	01.220	How can I help you?
0:13:03	02.970	Well, my son here, looking to buy his first car.
0:13:07	02.300	You come to see me? I had to.
0:13:09	01.600	That practically makes us family.
0:13:11	03.030	Uncle Bobby B, baby. Uncle Bobby B.
0:13:14	02.640	Sam. Sam, let me talk to you.
0:13:16	02.670	Sam, your first enchilada of freedom
0:13:19	02.350	awaits underneath one of those hoods.
0:13:21	01.330	Let me tell you something, son.
0:13:23	02.200	(ENGINE REVVING) A driver don't pick the car.
0:13:25	01.720	The car'll pick the driver.
0:13:27	02.840	It's a mystical bond between man and machine.
0:13:30	02.950	Son, I'm a lot of things, but a liar's not one of them.

Start	Duration	Captioning
0:13:33	02.210	Especially not in front of my mammy.
0:13:35	03.520	That's my mammy. Hey, Mammy!
0:13:39	03.450	Oh, don't be like that. If I had a rock, I'd bust your head, bitch.
0:13:42	02.130	I tell you, man, she deaf, you know?
0:13:44	02.500	(LAUGHING)
0:13:47	05.000	Well, over here, every piece of car a man might want or need.
0:13:52	01.470	This ain't bad.
0:13:55	02.070	This one's got racing stripes. Yeah.
0:13:57	01.900	It got racing...
0:13:59	01.870	Yeah, what's this? What the heck is this?
0:14:01	01.640	I don't know nothing about this car.
0:14:03	01.120	Manny! What?
0:14:04	02.630	What is this? This car! Check it out!
0:14:07	02.520	<i>I don't know, boss! I've never seen it! That's loco!
0:14:09	02.550	Don't go Ricky Ricardo on me, Manny! Find out!
0:14:12	01.560	Feels good.

Scene D - Transformers

Start	Duration	Captioning
1:02:54	02.170	Are you Samuel James Witwicky,
1:02:56	02.000	descendent of Archibald Witwicky?
1:02:59	02.100	They know your name.
1:03:02	01.020	Yeah.
1:03:03	02.470	My name is Optimus Prime.
1:03:05	04.300	We are autonomous robotic organisms from the planet Cybertron.
1:03:09	02.280	But you can call us Autobots for short.
1:03:12	01.150	Autobots.
1:03:13	01.320	What's cracking, little bitches?
1:03:15	03.680	OPTIMUS: My first lieutenant. Designation, Jazz.
1:03:18	04.120	JAZZ: This looks like a cool place to kick it.
1:03:23	01.770	What is that? How did he learn to talk like that?
1:03:24	03.590	We've learned Earth's languages through the World Wide Web.
1:03:30	02.500	My weapons specialist, Ironhide.
1:03:33	02.130	You feeling lucky, punk?
1:03:35	01.650	Easy, Ironhide.
1:03:37	04.150	Just kidding. I just wanted to show him my cannons.
1:03:41	02.200	(SNIFFING) OPTIMUS: Our medical officer, Ratchet.
1:03:43	04.340	The boy's pheromone level suggests he wants to mate with the female.
1:03:50	03.460	You already know your guardian, Bumblebee.
1:03:53	02.100	(RAP MUSIC PLAYING) Bumblebee, right?
1:03:56	01.330	<i>BUMBLEBEE: ♪ Check on the rep Yep, second to none ♪
1:03:57	02.400	So you're my guardian, huh?
1:03:59	05.240	His vocal processors were damaged in battle. I'm still working on them.
1:04:10	01.520	Why are you here?

Start	Duration	Captioning
1:04:11	02.520	We are here looking for the All Spark.
1:04:14	02.810	And we must find it before Megatron.

Appendix C: Copyright Permission

***Telecommunications Journal of Australia* Contributor agreement (single author)**

Date 15 of January 2011

Parties

- | | | |
|---|--|--------------------------------------|
| 1 | The Australian Computer Society
Level 3, 160 Clarence St, Sydney, NSW, 2000, Australia | ("the TSA") |
| 2 | Quoc Vy
c/o Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada | ("the Author")
(Author's address) |

Operative provisions

1 The Work

The Author has created a Work entitled
Enhanced Captioning - Speaker Identification: Text vs. Images

("the Work")

2 Grant of copyright

- 2.1 The Author grants the ACS the non-exclusive licence to publish and distribute the Work in electronic, digital and print form worldwide for the full period of the copyright, to exploit the subsidiary rights listed and defined in Schedules 1A and 1B (and to licence others to do any of these things).
- 2.2 The Author grants the ACS the exclusive licence to publish and distribute the Work in electronic, digital and print form worldwide and to exploit the subsidiary rights listed and defined in Schedules 1A and 1B (and to license others to do any of these things). This licence shall persist for twelve months, commencing the month the Work is published in the *Telecommunications Journal of Australia*.
- 2.3 It is intended that the Work be published in a journal entitled *Telecommunications Journal of Australia*.

3 Promotion

The Author acknowledges that in order to promote the Work, it may be necessary for the ACS to produce and distribute marketing material in print and/or electronic formats that includes adaptations or abridgements of the Work, or to authorise others to do the same, and the Author consents to the TSA, or others authorised by the TSA, doing any of these acts. The ACS agrees that it will not modify, or authorise modification of, the Work for any purpose other than promotion without prior written permission from the Author. The Author agrees that permission will not be unreasonably withheld.

4 Author's warranties and indemnity

- 4.1 The author warrants that:
 - a. the Work has not been published before in a refereed journal and is not being considered for publication in a refereed journal, in either print or electronic form;
 - b. the source of any copyright materials in the Work has been acknowledged;

- c. the Work does not infringe any copyright or other rights held by third parties and does not breach any duty of confidentiality or obligation of privacy and does not contain any material which is libellous, obscene or otherwise of an unlawful nature;
 - d. the Author who has signed this Agreement has full right, power and authority to enter into this agreement.
- 4.2 The Author will obtain, at the Author's expense, consents releases or permissions in relation to illustrations, photographs and other third party material used in the Work. The Author confirms that the benefit of all consents, releases or permissions obtained by the Author will extend to the benefit of the ACS throughout the Territory for the term for which rights are granted to the ACS pursuant to this Agreement.
- 4.3 The author undertakes to indemnify the ACS against all actions, proceedings, claims, demands and costs arising directly or indirectly from any breach of the warranties contained in 4.1.

5 Rights granted back to the author

- 5.1 The ACS grants back to the Author the following rights:
- a. The right to photocopy or make single electronic copies of the Work for the Author's own personal use, including the Author's own lecture or classroom use (excluding the preparation of course-pack material for onward sale by libraries and institutions), provided those copies are not offered for sale on a for-profit basis and are not distributed in any systematic way outside of the employing institution (for example, via an email list or public file server).
 - b. The right to post the Work on a secure network (not accessible to the public) within the institution that employs the Author.
 - c. The right to retain a Preprint, as defined in Schedule 1B, of the Work on a public electronic server such as the World Wide Web.
 - d. The right, subsequent to publication by the ACS or by other/s authorised by the TSA, to use the Work, or any part thereof, free of charge in a printed compilation of works of the Author's own, such as collected writings or lecture notes.
 - e. The right to include the Work in a thesis or dissertation provided that this is not to be published commercially.
 - f. The right to present the Work at a meeting or conference and to provide copies of the paper to the delegates.
- 5.2 Should the Author wish to reuse the Work, in whole or in part, in another publication in any way other than those listed in clause 5.1(a)-(f), the Author will make the request to the TSA.

6 Copyright notice

- 6.1 Where the Work has been published by the ePress (as defined in Schedule 1B), the Author undertakes to:
- g. add the following notice to the documents listed in subclause 6.2

This material has been published by Monash University ePress in *Telecommunications Journal of Australia* [insert volume number, issue number, date of publication, and paper number (including DOI) as it appears at the end of the published article].

Monash University ePress is the definitive repository of the content that has been certified and accepted after peer review.

You may copy the material onto a single computer and make a print copy for your personal use only. For any other use, prior written permission must be obtained from Monash University ePress, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.
 - h. where technically possible, add—to any electronic versions of the documents listed in subclauses 6.2(a), (b), (d), and (e)—a link to the definitive version of the Work as made available online by the ePress. The Author should contact the ePress for instructions on the text required to create the link.
- 6.2 The following documents must carry the copyright notice in subclause 6.1:
- i. any copies (photocopied or electronic) of the Work created and distributed by the Author for the Author's own lecture or classroom use under the terms of clause 5.1(a)

- j. the Work as posted by the Author on a secure network under the terms of clause 5.1(b)
 - k. the Work, or any part thereof, included by the Author in a printed compilation of works of the Author's own under the terms of clause 5.1(d)
 - l. the Work, included by the Author in an article in a thesis or dissertation under the terms of clause 5.1(e)
 - m. the Work, as presented by the Author at a meeting or conference, or handed out by the Author to the delegates attending the meeting or conference, under the terms of clause 5.1(f)
- 6.3 Where the Work has been published by the ePress and a Preprint existed prior to such publication, the Author undertakes to add the following notice to any copies (photocopied or electronic) of the Preprint made and distributed by the Author after publication of the Work by the ePress:

This is the author's version of the work. The definitive version has been published by Monash University ePress in *Telecommunications Journal of Australia* [insert volume number, issue number, date of publication, and paper number (including DOI) as it appears at the end of the published article].

7 Permissions

Subject to clause 5.1, all requests to reuse the Work, in whole or in part, in another publication (including in all commercially published edited Works) will be handled by the ACS. The TSA, or others authorised by the TSA, reserves the right to charge a fee for the grant of permission. The Author acknowledges that such fees may be retained by the TSA. All requests to use or include substantial parts of the Work in another publication (including publications of the ePress, as defined in Schedule 1B) will be subject to the Author's approval, which is deemed to be given if the ACS has not heard from the Author within 4 weeks from the date of the ACS writing to the Author at the Author's last notified address.

8 Copyright Agency Limited (CAL) payments

- 8.1 The Author acknowledges that proceeds of payments made pursuant to CAL-administered licensing schemes authorising the copying of works are to be paid to the TSA. These schemes include, but are not limited to Statutory Licences administered by CAL and Voluntary Licences Administered by CAL. The Author acknowledges that CAL may enter into arrangements with the ePress to share these payments. The ACS undertakes to use its share of such payments, after any distribution to the ePress, to further the activities of the TSA.

9 Entire agreement

This Agreement constitutes the entire agreement between the parties and supersedes all prior representations, statements and understandings whether verbal or in writing.

10 Governing law

This agreement shall be governed by the laws of the state of Victoria.

SIGNED for and on behalf of the ACS by



(signature)

Peter Gerrand

(print name)

15 January 2011

(date)

SIGNED by the Author



15 January 2011

(signature)

(date)

Quoc Vy

(print name)

Schedule 1: Definitions

A: Subsidiary rights

“Anthology and Quotation Rights” means the exclusive right to authorise the reproduction of extracts and quotations from the Work (including illustrations diagrams and maps contained in the Work) in other publications in all forms, formats and media whether now known or hereafter developed (including, without limitation in print, digital and electronic form) throughout the world;

“First Serial Rights” means the exclusive right to authorise the publication of extracts of the Work in one issue or more than one successive or non-successive issue of a publication in all forms, formats and media whether now known or hereafter developed (including, without limitation in print, digital and electronic form) throughout the world;

“New Technology Rights” means the exclusive right to use, store, reproduce all or any part of the Work, with other works or subject matter (including products of other publishers) by means of CDR, Cdi, computer technology or any other technology now known or subsequently discovered in any format and delivered to a user by any means whatsoever;

“Online Rights” means the exclusive right to communicate the Work to the public on any service for carrying or transmitting data and/or communications by means of guided or unguided electromagnetic energy or both, including the Internet, accessible by any receiver;

“Promotion Rights” means the right to use and to authorise others to use, by any means in all media including but not limited to an Online Service or Website, excerpts from the Work in connection with the publicity and promotion of the Work;

“Reprographic Rights” means rights to copy governed by Australian statutory license schemes.

“Sound Recording Rights” means the exclusive right to authorise the making of a recording of the Work and the provision of multiple copies in CD-Rom, minidisk or other digital formats and transmission via the Internet;

“Translation Rights” means the exclusive right to authorise the making and exploitation of the Work in foreign languages, in all forms, formats and media whether now known or hereafter developed (including, without limitation in print, digital and electronic form) throughout the world;

“Website Rights” means the exclusive right to create a website relating to the Work or include the work in a Website created by another publisher.

B: Other definitions

“the ePress” means Monash University ePress, of Wellington Road, Clayton, Victoria 3800, Australia

“Preprint” means an unrefereed draft of a Work.

“Territory” means worldwide;

“Website” means a page or pages on the world wide web and includes the Internet.

Bibliography

1. Andersen, N.E., Kensing, F., Lundin, J., Mathiassen, L., Munk-Madsen, A., Rasbech, M. & Sørkaar, P. (1990). Professional Systems Development: Experience, Ideas and Action. Prentice-Hall, Inc., Upper Saddle River, New Jersey, USA.
2. Attorney-General's Department. (1992). Broadcast Services Act. Office of Legislative Drafting and Publishing. Australia.
http://www.austlii.edu.au/au/legis/cth/consol_act/bsa1992214/
3. BBC. (1975). This is Ceefax. Great Britain: British Broadcasting Corporation.
4. BBC. (2009). Subtitling Editorial Guidelines. British Broadcasting Corporation.
http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf
5. Beguin, P. (2003). Design as a mutual learning process between users and designers. *Interacting with Computers*. 15 (5), 709-730.
<http://www.sciencedirect.com/science/article/B6V0D-49CT644-1/2/050cf72e89baf24493595cffcba2def>
6. Bødker, K., Kensing F., & Simonsen, J. (2004). Participatory IT design: Designing for business and workplace realities. Cambridge: MIT Press.
7. Burt, J.S., & Hutchinson, B.J. (2000). Case-Mixing Effects on Spelling Recognition: The Importance of Test Format. *Journal of Psycholinguistic Research*. 29 (4), 433-451.
8. Canadian Association of Broadcasters. (2008). Closed captioning standards and protocol for Canadian English language television programming services.
9. <http://www.cab-acr.ca/english/social/captioning/captioning.pdf>
10. Canadian Association of the Deaf. (2007). Statistics on Deaf Canadians.
http://www.cad.ca/statistics_on_deaf_canadians.php
11. Canadian Radio-television and Telecommunications Commission. (2009). Access to TV for people who are deaf or hard of hearing: closed captioning.
http://www.crtc.gc.ca/eng/info_sht/b321.htm

12. Canadian Radio-television and Telecommunications Commission. (2011). CRTC report shows more Canadians are adopting broadband Internet and wireless services. <http://www.crtc.gc.ca/eng/com100/2011/r110728.htm>
13. Canadian Radio-television and Telecommunications Commission. Glossary. <http://www.crtc.gc.ca/eng/glossary-glossaire.htm>
14. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.
15. Cooper, L., de Bruijn, O., Spence, R., and Witkowski, M. (2006). A comparison of static and moving presentation modes for image collections. In *Proceedings of the working conference on Advanced visual interfaces (AVI '06)*. ACM, New York, NY, USA, 381-388.
<http://doi.acm.org/10.1145/1133265.1133345>
16. European Commission. (2009). *Audiovisual Media Services Directive*.
<http://ec.europa.eu/avpolicy/reg/avms/>
17. Everingham, M., Sivic, J., Zisserman, A. (2009). Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing*. 27 (5), 545-559.
<http://dx.doi.org/10.1016/j.imavis.2008.04.018>
18. Julia Child. (1972). *The French Chef* [Television series]. Boston, Massachusetts: WGBH.
19. Harkins, J.E., Korres, E., Virvan, B., Singer, B. (1996). *Caption Features for Indicating Non-Speech Information: Guidelines for the Captioning Industry*. Gallaudet Research Institute.
http://tap.gallaudet.edu/Captions/nsi_prj.asp
20. Kensing, F., Simonsen, J. & Bødker, K. (1998). MUST: A Method for Participatory Design. *Human-Computer Interaction*, 13(2), 167-198.
http://dx.doi.org/10.1207/s15327051hci1302_3
21. Lanzara, G.F. & Mathiassen, L. (1985). Mapping situations within a system development project. *Journal Information and Management*. 8 (1), 3-20.
[http://dx.doi.org/10.1016/0378-7206\(85\)90065-5](http://dx.doi.org/10.1016/0378-7206(85)90065-5)

22. Lanzara, G. & Mathiassen, L. (1988). Intervening into System Development Area Projects: Tools for Mapping Situations. Working with Computers: Theory versus Outcome. Academic Press, London, 177-213.
<http://www.informaworld.com/smpp/content~content=a714917638>
23. Lee, D.G., Fels, D.I., & Udo, J.P. (2007). Emotive captioning. Computers in Entertainment. 5 (2). <http://dx.doi.org/10.1145/1279540.1279551>
24. Rashid, R., Vy, Q. V., Hunt, R. G., and Fels, D. I. (2008). Dancing with Words: Using Animated Text for Captioning. International Journal of Human-Computer Interaction, 24 (5), 505-519.
25. Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. Psychological Bulletin. 124 (3), 372-422.
<http://dx.doi.org/10.1037/0033-2909.124.3.372>
26. Roberts, V.L., Fels, D.I. (2006). Methods for inclusion: Employing think aloud protocols in software usability studies with individuals who are deaf. International Journal of Human-Computer Studies, 64 (6). 489-501.
<http://dx.doi.org/10.1016/j.ijhcs.2005.11.001>
27. Soegaard, Mads (2010). Gestalt principles of form perception.
http://www.interaction-design.org/encyclopedia/gestalt_principles_of_form_perception.html
28. Statistics Canada. (2011). Population by year, by province and territory. Canada.
<http://www40.statcan.ca/l01/cst01/demo02a-eng.htm>
29. The Nielsen Company. (2012). How Americans are Spending their Media Time... and Money.
http://blog.nielsen.com/nielsenwire/online_mobile/report-how-americans-are-spending-their-media-time-and-money/
30. Transformers. (2007). Directed by Michael Bay. Los Angeles, California: Paramount Pictures.
31. The Library of Congress. (1990). Television Decoder Circuitry Act. United States.
32. The Library of Congress. (1996). Telecommunications Act. New York: Practising Law Institute. United States. <http://lccn.loc.gov/96157120>

33. The Library of Congress. (2010). Twenty-First Century Communications and Video Accessibility Act. United States.
34. Tobii Technology AB. (2006). User Manual. *Tobii Eye Tracker ClearView Analysis Software*. Sweden.
35. United Nations. (2006). Convention of the Rights of Persons with Disabilities.
<http://www.un.org/disabilities/default.asp?navid=12&pid=150>
36. United Nations, Department of Economic and Social Affairs, Population Division (2011). World Population Prospects: The 2010 Revision.
<http://esa.un.org/unpd/wpp/>
37. Up. (2007). Directed by Pete Docter. Los Angeles, California: Walt Disney Pictures.
38. Vy, Q. V., and Fels, D. I. (2008). EnACT: A Software Tool for Creating Animated Text. ICCHP 2008. Linz. In: Miesenberger, Klaus et al (eds): Computers helping people with special needs: 11th International Conference. Lecture Notes in Computer Science. Springer, 609-616.
http://dx.doi.org/10.1007/978-3-540-70540-6_87
39. Vy, Q. V., and Fels, D. I. (2009). Using Avatars for Improving Speaker Identification in Captioning. Human-Computer Interaction. INTERACT 2009. Lecture Notes in Computer Science. Springer, 916-919.
http://dx.doi.org/10.1007/978-3-642-03658-3_110
40. Vy, Q. V., and Fels, D. I. (2010). Using Placement and Name for Speaker Identification in Captioning. ICCHP 2010. Vienna. In: Miesenberger, Klaus et al (eds): Computers helping people with special needs: 12th International Conference. Lecture Notes in Computer Science. Springer, 247-254.
http://dx.doi.org/10.1007/978-3-642-14097-6_40
41. Vy, Q. V., and Fels, D. I. (2011). Enhanced Captioning - Speaker Identification: Text vs. Images. Telecommunications Journal of Australia. 61 (2): 29.1-29.13.
42. Webster, K. (1994, Jan 10). Screen Test: Captioning Devices Will Enable Deaf To Read Movies At Theaters. Seattle Times.
<http://community.seattletimes.nwsources.com/archive/?date=19940110&slug=1888903>

43. Zdenek, S. (2011). Which sounds are significant? Towards a Rhetoric of Closed Captioning. *Disability Studies Quarterly*. 31 (3).
<http://dsq-sds.org/article/view/1667/1604>
44. WFD. About us. World Federation of the Deaf.
<http://www.wfdeaf.org/about>

Glossary

Acronyms and Abbreviations

ASL – American Sign Language

BSL – British Sign Language

hoh – *hard-of-hearing*

CC - *closed captioning*

TV – *television (includes analog, digital, high-definition and 3D)*

DVD - Digital Versatile/Video Disc

NSI - non-speech information

RWC - Rear Window Captioning

SDH - subtitles for the deaf and hard-of-hearing

Terminologies

captioning - refers to *closed captioning* (North America) and *subtitles for the hard-of-hearing* (Europe) which were developed to be displayed on television, either broadcasted or local media disc.

subtitles (*for the hard of hearing*) – captioning found on television throughout Europe. These “subtitles” are not the same as for film (see below).

subtitles - captioning for film that is usually of a different language from the content, assumes the viewers can hear, but not understand the language of the content (e.g foreign films).

same language subtitles (SLS) - subtitles that are in the same language as the content, which are popular on DVDs.

Subtitles for the Deaf and Hard-of-Hearing (SDH) - the term used by the film industry which are same language subtitles (SLS) that are primarily used by people who are deaf or hard-of-hearing.

Rear Window Captioning (RWC) – a closed captioning technology developed for theatres, using a reflective and smoked sheet of plastic

“open” captioning – captioning or subtitles that are visible to all viewers, usually cannot be turned off by users (e.g., embedded into video stream)

“closed” captioning – captioning or subtitles that may be visible, when needed by the user

deaf – includes deafened and Deaf

deafened - complete loss of hearing (later in life), if only partial: hard-of-hearing

(culturally) Deaf – born without hearing, does not consider as a medical condition, but a culture

Linguistically Deaf – is Deaf and knows sign language (e.g. ASL, BSL)

hard of hearing (hoh) - scientific term, limited ability of hearing, with or without cochlear implants