The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Video quality requirements for South African Sign Language communications over mobile phones

A dissertation submitted to the Department Of Computer Science, Faculty of Science at the University Of Cape Town in partial fulfilment of the requirements for the degree of Master of Science (in Information Technology).

By
Daniel Erasmus
February 2012

Supervised by
Prof Edwin H. Blake

University of Cape Town

Department of
Computer Science

University of
Cape Town

Acknowledgements

This dissertation could not be completed without the help and support of the following people. I would like to take this opportunity to thank them for their help with this dissertation.

Thank you to my supervisor Prof. Edwin Blake for his guidance, support and patience throughout this research.

My thanks to my mother, father, sister and Marchelle for their continued support and encouragement, especially when it felt as if this dissertation will never get finished.

I would also like to thank everyone at DCCT, especially Meryl Glaser, as well as all the participants that gave me their time and valuable feedback. Lastly thank you to Michelle Lombard and Betty Mokoena for their assistance as Sign Language interpreters.

University of Cape Town

Abstract

The Deaf community in South Africa currently can make use of the mobile communications networks through text-based means only, using services such as SMS, MXit and email. But this robs the Deaf from the opportunity to communicate in their first language, Sign Language. Sign Language, being a visual language, does not translate well to the text-based communications available to them. To enable the Deaf community to also share in the use of the mobile communications infrastructure means mobile video communications. With increasing network speeds, more affordable bandwidth and more capable and affordable mobile phones this is becoming a reality. This project aims to find the minimum video resolution and frame rate that supports intelligible cell phone based video communications in South African Sign Language.

University of Cape Town

Table of Contents

University of Cape Town

University of Cape Town

List of Tables

University of Cape Town

List of Figures

University of Cape Town

Glossary

3G	Third generation (Third generation of mobile telecommunications technology)
3GP	Multimedia container format used on third generation (3G) mobile phones.
AAC	Advanced Audio Coding
ADSL	Asymmetric Digital Subscriber Line
AMR	Adaptive Multi-Rate
API	Application programming interface
ASL	American Sign Language
AVC	Advanced Video Coding
AVI	Audio Video Interleave
CBR	Constant bit rate
CDMA	Code Division Multiple Access
CIF	Common Intermediate Format (352 x 288 pixels)
codec	Coder/decoder
DCCT	Deaf Community of Cape Town
DVD	Digital Versatile Disc
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
JPEG	Joint Photographic Experts Group
LAN	Local Area Network
Mbps	Megabits per second
M-JPEG	Motion JPEG
MMS	Multimedia Messaging Service
MP4	File extension used for MPEG-4 Part 14 multimedia container format files.
MPEG	Moving Picture Experts Group
MSE	Mean square error
NGO	Non-governmental welfare organisation
PAL	Phase Alternating Line
PC	Personal Computer
PSNR	Peak signal-to-noise ratio
PSTN	Public Switched Telephone Network
QCIF	Quarter Common Intermediate Format (176 x 144 pixels)
QVGA	Quarter Video Graphics Array (320 x 240 pixels)
SASL	South African Sign Language
SMS	Short Message Service
SQCIF	Sub Quarter Common Intermediate Format (112 x 96 pixels)
TVML	Television Mark-Up Language
VBR	Variable Bit Rate
VRS	Voice Relay Service

University of Cape Town

1 Introduction

1.1 Background

According to the National Institute for the Deaf there are just over 400 000 profoundly deaf people and just over 1 200 000 extremely hard-of-hearing people in South Africa [20][21]. Sign Language is the first language for people who were born deaf or became deaf before acquiring language, and as such is the language wherein they can communicate best. The Deaf sees themselves as a cultural group with their own language. Bilingualism is encouraged, especially for the Deaf to become part of the wider community. The second language, such as Afrikaans, English or Xhosa is learned mainly as a reading and writing language, while basic speech is learned to complement signs in communicating with hearing persons.

Sign Language is a visual form of communication, conveying meaning through a combination of hand shapes, movement of the hands and arms, in addition to facial expressions. The majority of the signs in sign language are formed in a "signing space", which includes the signer's head and chest, extending down to the hips. The grammar of Sign Language is markedly different from that of spoken languages and hence the written text of many Deaf users is often not grammatically correct [38].

The visual nature of Sign Language is not well supported by modern mobile communication, which is based primarily on voice communication and in increasing amounts on text based (written language) communication through services such as Short Message Service (SMS) and email.

Mobile text based communications are an option for the Deaf community. It is already implemented and supported by even the cheapest cell phone on the market. But for a Deaf person to communicate with another Deaf person through text would be the equivalent of two Afrikaans first language speakers being forced to speak English to each other when using a cell phone. Why must a person be forced to communicate in a second language?

The third generation cell phone networks support video calls, but these calls are limited in resolution and frame rate, and are primarily designed to support spoken communications, and not video as a primary communications channel.

This research work aims to assist in bringing mobile communications to the Deaf community by determining the minimum video quality, frame rate and resolution needed for South African Sign Language (SASL) video material playback on a cell phone to be still intelligible in a conversational context.

Throughout this research real users were used. The experimental participants were all native signers and have used SASL as their principal mode of communications most, if not all, of their lives and had English, Afrikaans or Xhosa as their language of literacy, regardless of what their hearing families used. The experimental work was completed with the assistance of The Deaf Community of Cape Town (DCCT), a grassroots non-governmental welfare organization (NGO) founded in 1987 and run by Deaf people to serve the needs of the disadvantaged Deaf community in Cape Town. They are based at the Bastion of the Deaf in Newlands, Cape Town.

Multiple studies were conducted with the help of the Deaf community to evaluate sign language videos, viewed on a cell phone, for intelligibility. Various SASL video sequences were shown to the long time SASL users at different video resolutions and frame rates with each clip being evaluated for intelligibility.

1.2 Aims and Expected Outcomes

In giving the South African Deaf community access to the telecommunications infrastructure and helping members of the community communicate in their first language, cell phones could play a

University of Cape Town

very important role providing affordable access to video based communication. In reaching this objective of affordable first language telecommunications for the Deaf community affordability and practicality is of the essence. The lower the quality of video, while still supporting an intelligible Sign Language conversation, the lower the cost of the bandwidth and the lower the required specification of the cell phone and thus cost of the cell phone.

The main question that was asked by this research was:

> *What is the lowest video resolution and frame rate that would provide intelligible SASL video on a cell phone?*

With the secondary question:

> *How does one measure intelligibility of Sign Language video material?*

The collected information could be used in the future development of video communications over mobile phones for the Deaf community using SASL. The ultimate goal is the development of a usable video communications application on low end smart phones, bringing affordable telecommunications to the South African Deaf community.

1.3 Dissertation Outline

The text based telecommunications options available to the Deaf community are described in Sections 2.2 and 2.3. This is followed by a basic introduction to digital video, including current cell phone support for digital video capture, and cell phone specific compression techniques in Section 2.4, before looking at the video based telecommunications options that are available to the Deaf community in Sections 2.6 and 2.7. Section 2.8 finishes off with an overview of Sign Language video quality requirements, and a review of related work.

Chapters 3 and 4 describe the subjective evaluations of SASL video clips at different resolutions and frame rates. The results from these pilot studies are incorporated into the development of the final experiment, described in Chapter 5.

The dissertation is concluded in Chapter 6 by reviewing these discussions and considering future work.

University of Cape Town

2 Background and related work

Telephones are by definition designed for audio communication, spoken words, whereas the Deaf communicate visually through Sign Language, making the telephone inappropriate for the use of the Deaf community without adding to the installed telephone infrastructure.

This chapter looks at the current telecommunications options, both deaf-to-hearing as well as deaf-to-deaf, available to the Deaf community, including the work still in research phase, which is not widely available yet.

2.1 Relay Services

For the Deaf to use the Public Switched Telephone Network (PSTN) there needs to be a visual to audio translation phase added. One way is a Voice Relay Service (VRS), which adds a live operator to assist through translation. This can take the form of a text relay, where the Deaf user types a text message that is received by an operator, who reads out loud the message to the hearing caller on behalf of the Deaf caller, and then types out the hearing caller's response enabling the Deaf caller to read the response [9].

This basic idea of an operator translating between hearing and Deaf caller can be extended to sign language through the use of a video link, for example using a webcam connected to a personal computer (PC). However these services require advanced infrastructure and qualified translators to be available 24 hours a day, an expensive proposition, resulting in these services not being universally available, and even where available being cancelled because of financial constraints [3].

A third option is captioned telephony in which a computer based gateway is set up to handle the translation, alleviating the need for the 24 hour availability of a live translator. The captioned telephony system uses text to speech technology to translate what the Deaf user typed to speech that is relayed through to the hearing recipient, and then uses speech recognition technology to translate the spoken response back into text to be read by the Deaf user. An example of a captioned telephony system is the South African developed Telgo323 [29], although the Telgo323 only worked in one direction, from text to speech, and not in the reverse direction.

There has also been research into automated sign language translation. But this is a wide ranging problem combining knowledge and technology from multiple fields including computer vision, neural networks, sign recognition methods, 3D animation and natural language processing. Not only does the system need to translate between spoken, text based language to a sign language with a very different grammatical structure, but also generate the equivalent 3D avatar animation of the gestures. In addition there is the need to recognize gestures, including facial expressions, and translate those back to spoken language [27].

One such project [12] looked at enabling communication between hearing and Deaf by sending avatar based animations obtained through automatic interpretation of text to sign language, using Multimedia Messaging Service (MMS). The usefulness of the system though is limited by it enabling communications in only one direction, from text to sign language. There are no bi-directional communications available.

One of the latest systems was developed by the Science and Technology Research Laboratories of the NHK (Japan Broadcasting Corporation) [13]. The work focused on adapting Television Mark-Up Language (TVML) to produce Japanese Sign Language animation. TVML is a text-based computer language that enables the production of computer graphics animated video content by simply writing a script in TVML. The user is able to specify in the TVML script the words spoken, the movements and even the facial expressions.

The researchers extended the existing TVML facilities with the aim of generating sign language animation by developing high-quality computer graphics models and an improved TVML player that can render the manual movements of sign language. In addition a Japanese-to-Japanese Sign

University of Cape Town

Language dictionary was developed, and in the latest work they are focusing on a way to combine optical motion captured data to generate sign language sentences. They are now able to translate a set of texts automatically into a string of sign language words. The range of sentences that can be translated is currently still limited, and the generated animation still lacks fluency as the automatic transitions between different signs is not as smooth as what would be expected from a human signer.

2.2 Deaf-to-Hearing Text Based Telecommunications

Relay services prevent the Deaf user from being in direct communications with the hearing person in the conversation. There is always an intermediate translation step, be it via a live operator or an automated system.

Email has been around for a long time and is widely used, including being accepted for official and business communications. Email enables the distribution of electronic documents, as well as audio and video through attachments. But email does not enable interactive, conversational communications.

Even the cheapest cell phone supports the SMS providing easy access to affordable mobile, text-based communications. With the deep penetration of cell phones into the South African population, SMS provides the Deaf community with easy access to a large section of the community without the need for any special intervention. Yet SMS is not an effective channel for conversational communications. It is possible to receive delivery receipts, showing that the message was delivered to the recipient's phone, but there is no way of knowing if the recipient has read the message or is busy replying to the message.

Instant messaging overcomes some of the shortfalls of using SMS and email, enabling near synchronous text based communications. Tucker [40] describes the unique advantages of the instant messaging system. With good connectivity and both users actively involved instant messaging can appear synchronous, while at the same time allowing the communications to be temporarily or even extensively interrupted. Delays are more tolerated in an instant messaging environment where true synchronous communications are not expected.

2.3 Deaf-to-Deaf Text Based Telecommunications

From a technical perspective, the purely text based telecommunications options are the simplest to implement within the Deaf community.

In South Africa the Teldem device [8], designed especially for people with hearing difficulties, was available from Telkom. The latest Telkom tariff list (1 August 2011), lists the Teldem service as no longer available, with rental of the device only available to existing customers [37]. This device is a portable text telephone, with a QWERTY keyboard and alphanumeric display, which can be connected to any telephone and can communicate point-to-point with any other Teldem or TTY terminal. The major drawback of the Teldem is the fact that the Teldem can only exchange text with another Teldem device.

In addition, there are a wide range of generic internet and cell phone based text communications solutions used widely by the hearing community, as was shown earlier in this document, also available to the Deaf community. These solutions, such as email, SMS and instant messaging services, such as Skype and MXit, are as usable for Deaf-to-Deaf communications as for Deaf-to-Hearing communications and have the same advantages and disadvantages.

For the Deaf to communicate through text is forcing them to communicate in a second language, putting them at a disadvantage.

University of Cape Town

2.4 Digital Video

The digital video camera, like the one inside a cell phone, consists of a lens that focuses an image of the world onto a light sensitive electronic chip. This is the same setup as in any digital stills camera. To enable the capture of movement a sequence of still images, or frames, are captured one after the other in rapid succession. If this sequence of still images are then displayed on a screen, one after the other and the number of frames per second is not too low, the brain perceives smooth, realistic motion.

To store this sequence of still images that were captured, a wide variety of digital video formats have been developed. A video format refers to, among other things, how many pixels form an image, how may frames were recorded per second, how colour was recorded and how the video information was compressed. These formats will now be discussed further, as well as definitions of video terms.

2.4.1 Video Resolution

A digital video consists of multiple images or frames. Each frame formed by a rectangular grid of pixels, or picture element, each representing the colour of that specific part of the image. Phase alternating line (PAL) standard definition digital versatile disc (DVD) would have images consisting of 576 horizontal lines of 720 pixels each, giving the image, often written as 720 x 576. High definition video for example has a resolution of 1920 x 1080 (Each image being 1920 pixels wide by 1080 pixels high) [1].

The more pixels there are in the image, the higher the resolution of the image, and the clearer and sharper the picture. As the resolution is reduced the fewer details are captured, and the overall fuzziness of the image will increase.

2.4.2 Video Frame Rate

The video frame rate is the number of still images, or frames, captured, stored and displayed per second. The number of frames recorded each second affects how motion appears on the screen.

At lower frame rates motion artefacts are introduced into the video. If an object moves across the screen quickly it will be blurred while it is in motion. The motion is not perfectly continuous, and can seem to jump or stutter from one position to the next, as the object moves a bigger distance between frames than at a higher frame rate.

2.4.3 Colour Depth

Every colour that the human eye sees is a mix of red, green and blue light in different proportions. The sensor inside the digital camera also measures the relative amounts of red, green and blue light in the image. In single-chip colour cameras this is accomplished through tiny red, green and blue filters over individual pixels.

Colour depth refers to the number of bits used to represent the colour of a single pixel. The more bits used the broader the range of distinct colours that can be represented and stored, and the more accurately the image is represented.

2.4.4 Data rate or Bit rate

For a video file the bit rate refers to the number of bits used per unit of playback time after data compression, if any. Standard resolution DVD video contents for example have an average bit rate of 3.8 megabits per second (Mbps). That is 3800 kilobits of data stored per second of video. Values range from heavy Motion Picture Experts Group (MPEG) MPEG-2 compression of 2 Mbps to high-quality compression of 6 Mbps [1].

University of Cape Town

The actual data rate of digital video contents depends in part on the size of the frame, frame rate, as well how much the video is compressed (if at all) before it is recorded. The higher the video resolution and frame rate, the more data is captured per second and the higher the bit rate of the corresponding video data stream. If the data rate is limited, video quality has to be sacrificed at higher resolutions and frame rates, to keep to the specified data rate. Either the compression ratio has to be increased, adding compression artefacts, or frames will have to be dropped, either way the visual quality of the video will drop to try and keep within the limitations.

2.4.5 Video Compression

The higher the resolution and frame rate, the more digital data has to be captured, stored and transmitted. This increases the cost of working with the video because that requires big storage devices, and high-speed connections for distribution. To balance the need for high quality video on the one side, and cost of high bandwidth data on the other we have digital compression.

Digital compression aims to shrink the video data down to a smaller size while maintaining picture quality. This means the same video material takes up less storage space, and can be transmitted over the same connection in less time, and thus at lower cost. To watch the video it has to be decompressed, and the objective is to have to decompressed video look as closely to the original uncompressed video material as possible. Compression schemes are called codecs (coder/decoder).

With lossless compression the decompressed video frames are identical to the original frames before compression. Lossy compression, on the other hand, throws information away during the compression process, and it is impossible ever to restore the original frames as they were before compression. By taking into account human perception, the requirements for exact reconstruction can be relaxed. A picture, or in this case one frame of the video, may contain more detail than the human eye can perceive, and by dispensing with this extraneous data the picture can be degraded without the user noticing and in the process less storage is needed for the picture. Almost all codecs make use of lossy compression. It is possible with some codecs to adjust the amount of compression, but usually the heavier the compression the worse the compressed video looks [28].

All video codecs start by compressing each individual video frame. This is called intraframe or spatial compression. Each frame is compressed/decompressed on its own, independent from the frames before and after it, speeding up the compression/decompression process. Intraframe compression becomes less efficient the more complex and detailed the picture becomes. Motion-JPEG (M-JPEG) is an example of an intraframe video codec, adapting the Joint Photographic Experts Group (JPEG) algorithm used for lossy compression of still images, for compressing motion video. Each frame of the M-JPEG compressed video sequence is a self-contained compressed picture, achieving compression ratios ranging from about 2:1 to about 20:1.

Some codecs also take into account the fact that video frames are interrelated in time. Interframe or temporal compression looks a set of frames over time and finds ways to remove repetitive information that is similar between consecutive video frames. Often very little changes from frame to frame. Interframe compression works by looking at a group of frames, the first frame, or key frame, is stored in full, but for the subsequent frames the codec only stores the differences between the frame and its predecessors. Interframe compression becomes less efficient the more motion is present in the video. Relatively static video sequences have the best temporal compression efficiency. In for example the MPEG standard, an initial, self-contained picture provides the starting point from which following frames can be encoded by looking at pixel differences between successive frames. The MPEG standard includes the original MPEG-1 standard, which was superseded by the MPEG-2 standard used in DVD discs, as well as the MPEG-4 standard used in Blu-ray discs [28].

Returning to bit rates for a moment, for some codecs the same amount of data is stored for every frame, regardless of motion and details. This is constant bit rate (CBR) compression. Other codecs

University of Cape Town

allow the bit rate to adjust depending on the shot. These are variable bit rate (VBR) codecs, and allows for higher bit rates during shots that are complex or active, and reduce the bit rate for static less complex shots. VBR compression, though, requires more processing power during the compression of the video.

A wide variety of proprietary and standardised video compression algorithms have been developed over the years; the most important of these are published by recognised standardisation bodies, such as the International Telecommunication Union (ITU) and the Motion Picture Expert Group.

The ITU H.263 video codec was developed by the ITU Telecommunication Standardization Sector (ITU-T) Video Coding Experts Group for use as low-bit rate compressed format for video conferencing. The standard was further improved by the H.263+ and H.263++ standards approved in 1998 and 2000 respectively. H.263+ added optional features to improve compression efficiency and allow for quality, bitrate, and complexity scalability [10].

MPEG-4 Part 2 is H.263 compatible, and partially based on ITU-T H.263. It is similar to previous standards such as MPEG-1 and MPEG-2. DivX is an example an implementation of this standard. Most often reference to MPEG-4 refers to MPEG-4 Part 2 Simple Profile [11].

MPEG-4 Part 10, also known as MPEG-4 AVC (Advanced Video Coding) or H.264, is widely used in such applications as Blu-ray discs and direct broadcast satellite television services. It was designed as a standard to provide good video quality at substantially lower bit rates than previous standards, such as MPEG-2, H.263, or MPEG-4 Part 2.

2.4.6 Video container formats

A video file, for example an Audio Video Interleave (AVI) file or MP4 file is just a container format. The container format only defines how to store information inside them, and not what kinds of data are stored. A video file usually contains multiple tracks, a video track without audio, one or more audio tracks (without video), and multiple subtitle tracks and so on. The tracks are usually interrelated enabling the synchronisation of the different media tracks.

3GP is a multimedia container format used on third generation (3G) mobile phones and stores video streams as MPEG-4 Part 2 or H.263 or MPEG-4 Part 10 (AVC/H.264), and audio as Advanced Audio Codec (AAC) or Adaptive Multi-Rate (AMR). Most 3G capable phones support the recording and playback of video in 3GP format. The file extension is either .3gp for GSM-based phones or .3g2 for Code Division Multiple Access (CDMA) based phones [39].

The MPEG-4 Part 14 file format is a multimedia container format designed as part of the MPEG-4. It is based on the QuickTime format specification, and in addition to audio and video streams can store other data such as subtitles and still images. The file extension used is .mp4 [26].

2.4.7 Cell phone video capture support

Cell phones have various operating systems. The most common of these will now be discussed, specifically how they relate to video.

Symbian

Some Nokia phones support hardware based video encoding; enabling high quality video compression even on battery powered computing platforms such as cell phones. However, not all codecs are supported. The Nokia N96, used in this project, for example supports hardware encoding for H.263 and MPEG-4 video, but only software encoding for H.264.

Software for Symbian phones is developed using Symbian C++ and the accompanying application programming interfaces (APIs). To ease development of Symbian software Nokia moved over to Qt as their de-facto development framework [31]. Qt is a cross-platform application and UI framework with APIs for C++, providing support for the development of applications for

University of Cape Town

Symbian and Maemo/Meego in addition to desktop platforms, such as Microsoft Windows, Mac OS X, and Linux [30].

In November 2010 Qt Mobility 1.1.0 was released which included the Camera API extending the Multimedia API to provide access to the camera and video encoding functionality of the cell phone [33]. The QCamera object is used in conjunction with a QMediaRecorder object to record video. Through the QVideoEncoderSettings class the developer can specify the video codec used, bit rate, resolution, frame rate and quality settings used for capturing and compressing the video from the camera [34][32]. The available codec and recording settings supported by Qt on Nokia phones is shown in *Table 2-1* [22].

Encoding quality can be set to constant quality encoding, constant bit rate encoding, average bit rate encoding or two pass encoding. If constant quality encoding is selected, the quality encoding parameter is used and bit rate is ignored, otherwise the bit rate is used.

Setting the video quality setting allows backend to choose the balanced set of encoding parameters to achieve the desired quality level. The quality settings parameter is only used in the constant quality encoding mode.

In February 2011 Nokia announced that it would move to Windows Phone as its primary smartphone platform, with development of Symbian based phones coming to an end after the transition has been completed [31].

Codec	Possible resolutions	Possible frame rates (dependant on resolution)	Possible bitrates (Kbps) (dependant on resolution and frame rate)
Primary Camera			
H.263	176 x 144 352 x 288	15 30	64 - 2 048
H.264 (only Nokia N8)	176 x 144 352 x 288 640 x 480 1280 x 720	7.5 15 16.9 30 33.8	64 – 14 000
MPEG-4 Visual Part 2	176 x 144 352 x 288 640 x 352 or 640 x 480 720 x 576 or 720 x 480 1280 x 720	15 25 30	64 – 12 000
Secondary Camera			
H.263	176 x 144 352 x 144	15 30	64 – 20 048
MPEG-4 Visual Part 2	176 x 144 352 x 288 640 x 480	15 30	64 - 4 000
H.264 (only Nokia N8)	176 x 144 352 x 288 640 x 480	15 16.9 30 33.8	64 – 10 000

Table 2-1: Video settings supported by the QtMultimediaKit library on Nokia phones.

Android

Video capture support is dependent on the hardware of the specific Android phone being used. The Android operating system supports, as of Android v2.2, H.263, H.264 and MPEG-4-SP video capture, with output support for MPEG4, as well as 3GPP files. As can be seen in *Table 2-2 [16]*, as of Android 2.3.3 the Google developed VP8 codec and WebM container files are also supported.

Format/Codec	Encoder	Decoder	Details	Supported File Type(s)/Container Formats
H.263	Yes	Yes		3GPP (.3gp) and MPEG-4 (.mp4)
H.264 AVC	Yes (Android 3.0+)	Yes	Baseline Profile (BP)	3GPP (.3gp) and MPEG-4 (.mp4)
MPEG-4 SP	No	Yes		3GPP (.3gp)
VP8	No	Yes (Android 2.3.3+)		WebM (.webm)

Table 2-2: Core video and codec support of the Android platform.

The listed codecs and container formats are those provided by the Android platform, in addition to these, any Android powered device may provide device-specific media codecs. It is best practice though to use media encoding profiles, such as those listed in *Table 2-3* [16], that are device-agnostic.

	Lower Quality	Higher Quality
Video codec	H.264 Baseline Profile	H.264 Baseline Profile
Video resolution	176 x 144 pixels	480 x 360 pixels
Video frame rate	12 frames per second	30 frames per second
Video bitrate	56 Kbps	500 Kbps

Table 2-3: Examples of supported encoding profiles and parameters on the Android platform.

The MediaRecorder class is used to record audio and video. The developer has control over the bit rate, video frame rate and the video resolution. On devices that have auto-frame rate the specified frame rate will be taken as the maximum frame rate and not a constant frame rate. The specified bit rate might be clipped to ensure that video recording can proceed smoothly based on the capabilities of the platform [17].

iOS

The iPhone operating system provides the least flexibility. A predefined collection of presets are made available, shown in *Table 2-4*.

The resolution and bit rate for the output depend on the capture session's preset. The video encoding is typically H.264 and audio encoding AAC. The actual values vary by device, as illustrated in the following table.

In iOS 4.0 and later, you can record from a device's camera and display the incoming data live on screen. You use AVCaptureSession to manage data flow from inputs represented by AVCaptureInput objects (which mediate input from an AVCaptureDevice) to outputs represented by AVCaptureOutput [2].

University of Cape Town

Preset	iPhone 3G	iPhone 3GS	iPhone 4 (Back)	iPhone 4 (Front)
High	No video Apple Lossless	640x480 3.5 Mbps	1280x720 10.5 Mbps	640x480 3.5 Mbps
Medium	No video Apple Lossless	480x360 700 Kbps	480x360 700 Kbps	480x360 700 Kbps
Low	No video Apple Lossless	192x144 128 Kbps	192x144 128 Kbps	192x144 128 Kbps
640x480	No video Apple Lossless	640x480 3.5 Mbps	640x480 3.5 Mbps	640x480 3.5 Mbps
1280x720	No video Apple Lossless	No video 64 Kbps AAC	No video 64 Kbps AAC	No video 64 Kbps AAC
Photo	Not supported for video output	Not supported for video output	Not supported for video output	Not supported for video output

Table 2-4: iOS capture session presets

2.4.8 Real-time mobile video communications challenges

To provide real-time Sign Language video communications using mobile phones one needs to overcome three main challenges, namely low bandwidth, low processing speed and limited battery life.

For a mobile video conversation to happen, the video data has to be sent from one phone to the other, transmitted over the cellular network. The quality of the video and the time to transmit the data is thus limited by the speed at which the required data can be sent and received by the cell phone. Mobile bandwidth capacity is improving, and in South Africa we are in the privileged situation that our networks are still relatively young and based on modern technology, with continued aggressive expansion and upgrading of the network infrastructure. The networks generally support speeds of 7.2 Mbps and 14.4 Mbps, with speeds of 42 Mbps possible in select areas. But a high speed network does not guarantee bandwidth. The two phones used in this research, the Nokia N96 and Vodafone 858 Smart for example only support 3.6 Mbps communications, that is 3.6 Mbps while receiving data, and only 384 kbps when transmitting. Meanwhile depending on network, location and signal strength the user might be limited to only GPRS (32 – 48 kbps) or EDGE (maximum 384 kbps) speeds [43].

The need for portability in a cell phone means limits in available battery capacity to power the processor used in the cell phone. The Nokia N96 cell phone is powered by a dual core 264 MHz processor and 128 MB of RAM [23], while the more recent but entry level Vodafone 858 Smart is powered by a 528 MHz processor [41]. Neither is anywhere near as powerful as the processors used in current laptop computers.

The low processing power available on a cell phone limits the use of the cell phone as a video communications device in two ways. First limited processing power limits the video resolution and amount of compression that can be handled by the processor before the processing of the video starts introducing delays and affecting the intelligibility of the video. Secondly recording, compressing and decompressing video requires intensive use of the processor in the mobile phone. In addition to the processor the energy stored in the battery is further drained by the backlit screen as well as the data connection to the cellular network. All of this adds up to a very negative scenario for battery life.

The Nokia N96 cell phone for example has a stand-by time on 3G of 200 hours, a talk time on 3G of 160 minutes (2 hours and 40 minutes), and lists an offline video playback battery life of 5 hours [23]. For sign language video communications we do not want to only playback video, we want to record, compress, transmit, receive and play back video all in real time. In the end the battery is the most limiting constraint in mobile video communications.

University of Cape Town

2.4.9 Sign language specific video compression techniques

Some studies have been done to attempt to overcome the above limitations. Sperling et al. [36] provides a good overview of early attempts at compressing American Sign Language (ASL) images, and evaluates three basic image transformations for intelligibility, namely gray-scale subsample transformations, two-level intensity transformations converting the grey scale images to black and white images, and lastly taking it even further by converting the images into black and white outline drawings.

The goal of the MobileASL [4] project running at the University of Washington is to enable Deaf people to use mobile phones for communicating in Sign Language in real-time. Several H.264 compliant encoders were developed in an effort to lower the required resources while at the same time maintaining ASL intelligibility.

Variable frame rate was used to save processing cycles and battery life. By automatically detecting when the user is signing, the frame rate is adjusted on the fly, from the highest possible frame rate when the user is signing, down to 1 frame per second while the user is not signing.

Earlier research [18], found through the use of eye-movement tracking experiments, that Deaf people fixate mostly on the facial region of the signer to pick up the small movements and details in the facial expression and lip shapes of the signer. Peripheral vision is used to process the larger body and hand movements of the signer. The research concluded that increasing compression quality in the important regions of the video image, may improve bandwidth usage, as well as the quality of the sign language video as perceived by the Deaf.

Using these findings the MobileASL project approached the limited bandwidth problem by using dynamic skin region-of interest encoding. This meant that the visible skin areas of the video image was compressed at a higher quality at the expense of the remainder of the frame.

2.5 Synchronous and asynchronous communication

Synchronous communication is direct, live communication where all participants involved in the communication are present at the same time and respond to each other in real time. Examples of synchronous communication are telephone conversations and instant messaging. On the other hand, asynchronous communication does not require all participants to be present at the same time, such as email messages, discussions boards and text messaging over cell phones, and there can be a delay between when a person receives a message and a response is sent.

2.5.1 Synchronous Video over Internet

There are various freely available synchronous video solutions accessible to anyone with a PC, webcam and network connection, tools such as Skype and CamFrog. Because of network constraints video quality is limited. Skype video quality for example was found to be sufficient when used over a local area network (LAN), but not satisfactory over a 512kbps Asymmetric Digital Subscriber Line (ADSL) link [15].

Third generation cell phone network does support video calls, but these calls are limited in resolution and frame rate, and are primarily designed to support spoken communications, and not as a primary communications channel.

2.5.2 Asynchronous Video over Internet

As seen above, in low bandwidth environments synchronous video communications is only possible by reducing video quality or specialised compression. An ITU application profile [35] details the requirements for sign language video communications at a minimum common intermediate format (CIF) resolution (352 x 288 pixels) and a frame rate of at least 25 frames per second. In addition the video needs to cover enough area to include the detailed movements of the signer. These requirements can be met by modern video codecs, but only at high bit rates. When the bit rate falls

University of Cape Town

below 200 kilobits per second, the picture quality needs to be sacrificed, video size reduced and frame rate dropped. This leads to the Deaf user of the system to compensate for these problems, by for example slowing down signing and exaggerating movements. Even the improved efficiency of the H.264 codec may not provide acceptable video quality for sign language communications at low bit rates [18].

As mentioned earlier, delays are more tolerated in an instant messaging environment while at the same time allowing the communications to appear synchronous when both users are actively involved [29]. This provides the opportunity to use asynchronous video at higher quality, lessening the limitation in video quality at limited bit rates.

Ma and Tucker (2007) [14] found that asynchronous video over internet protocol (IP) was a valid solution; offering improved video quality regardless of bandwidth constraints. Although there still were issues to consider, such as reducing the inherent transmission delays, as well as improving the user interface to provide feedback to the user to alleviate these delays in the conversation.

Follow up research [15] focused on improving these issues, settling on the x264 video codec to provide low latency, high quality video for asynchronous video telephony. The research focused on finding the optimal settings to be used with the x264 codec to provide fast compression, small resulting file size to minimize transmission time and high quality playback with less complicated calculations. Further the user interface was simplified, as well as incorporating better notification of events to the user.

2.6 Sign Language video quality requirements

Which brings us to the question: What are the quality requirements for the capture and transmission of Sign Language?

2.6.1 ITU specifications

For the successful use of video for telecommunications via a visual language, such as Sign Language, certain quality requirements must be met. Currently the minimum quality requirements for Sign Language and lip-reading video material are specified in the *ITU-T Series H Supplement 1 (05/99)* [35] document, released by the ITU.

The ITU is the United Nations Specialized Agency in the field of telecommunications. The ITU-T on its part is a permanent organ of the ITU, responsible for studying technical, operating and tariff questions issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

ITU-T Series H Supplement 1 describes the factors to be taken into account when low bit-rate video is used for Sign Language and lip-reading telecommunications. The document sets out performance requirements that should be met to ensure a successful person-to-person conversation using a video communication system. In setting the requirements, video compression is ignored and the focus is on the resolution and frame rate. The stated requirements though should not be taken as fixed and absolute, but depending on the situation may need to be more stringent, or more relaxed.

The document shows that 20 frames per second provide good usability for both sign language and lip-reading, while still understandable at 12 frames per second. Between 8 and 12 frames per second usability becomes very limited, with no practical usefulness below 8 frames per second.

When looking at resolution for person-to-person sign language video communication, Supplement 1 concludes that it is possible to use Quarter Common Intermediate Format (QCIF) (176 x 144 pixels) resolution, with an increase to CIF (352 x 288 pixels) giving better language perception. Sub Quarter Common Intermediate Format (SQCIF) (112 x 96 pixels) is too coarse for reliable perception, with some signs occasionally perceivable.

University of Cape Town

The application profile concludes with the basic performance goal of aiming for 25-30 frames per second at CIF (352 x 288 pixels) resolution, while if needed in very low bit-rate environments dropping the frame rate to 12-15 frames per second at a resolution of 176 x 144 pixels.

2.6.2 Subjective and Objective evaluation of video quality

When looking at Sign Language communications over limited bandwidth communications channels such as the cellular telephone network an appropriate quality measurement is needed to compare different video parameters. In a subjective evaluation video sequences are shown to a group of viewers The viewers opinion of the video material is then captured, assigned a numeric value and averaged to provide a quality measurement for the video sequence. The details of the testing can vary depending on the objective of the testing and the aspect of the video that is being evaluated.

Objective video quality metrics are mathematical models that approximate results of subjective quality assessments as closely as possible. Video quality metrics such as mean square error (MSE) and peak signal-to-noise ratio (PSNR) are the most widely used objective measures for evaluating video. These measurement techniques though are focused on traditional quality in terms of aesthetics. Recent objective quality measures, modelling the human visual system, have shown substantial improvements over MSE and PSNR in predicting aesthetic quality. But as Ciaramello et al. [6] state, sign language video is a communications tool, and quality must be judged in terms of intelligibility.

Ciaramello et al. [6] demonstrated that PSNR is not a good measure of intelligibility in Sign Language video material, and proceeded to propose and evaluate a metric based on the spatial structure of ASL and as a function of MSE in both the hands and the face. The proposed metric gave a substantial improvement over PSNR.

The user experience of MobileASL was evaluated in a laboratory setting, with both subjective and objective measures [5]. The subjective measurements were done in a conversational setting, with two participants conversing in Sign Language using cellphones. The quality of the video was measured subjectively by how hard or easy it was to understand. This was done through a 5 question questionnaire. The survey questions were the following:

1. *During the video, how often did you have to guess what the signer was saying (where 1 is never and 5 is all the time)?*
2. *How difficult would you say it was to comprehend the video (where 1 is very easy and 5 is very difficult)?*
3. *Changing the frame rate of the video can be distracting. How would you rate the annoyance level of the video (where 1 is not annoying at all and 5 is extremely annoying)?*
4. *The video quality over a cell phone is not as good as video quality when communicating via the Internet (e.g., by using a web cam) or over a set top box. However, cell phones are convenient since they are mobile. Given the quality of conversation you just experienced, how often would you use the mobile phone for making video calls versus just using your regular version of communication (e.g., go home to use the Internet or set top box, or just text)?*
5. *If video of this quality were available on the cell phone, would you use it?*

The objective measure of the video quality was made through a count of the number of repair requests, for each repair request the number of times the requester asked for a repeat was counted, as well as a count of conversational breakdowns. This was all calculated from the videotaped user study sessions, during which participants were having conversations using phones set on a table in front of them.

As Nakazono et al. [19] state, in evaluating Sign Language video we must evaluate how well the linguistic information is transmitted and should be careful not to be swayed by impression of the

University of Cape Town

appearance of the video. They used two kinds of evaluations, the intelligibility test and the opinion test.

In the intelligibility test a short video sequence of sign language is presented to subjects, subjects are instructed to write down the contents of the sentences, dictated sentences are then evaluated from 0 to 3, keeping in mind to be careful not to be affected by the difference in subjects' ability in written language.

In the opinion test a short video sequence of sign language is presented to subjects, subjects were asked to evaluate the intelligibility of the sign language at five levels, from 1 to 5, and the mean value of the score is used for the evaluated value of the data. In the above study subjects were asked to evaluate the intelligibility of the sign video, and not to evaluate the preference of picture quality.

Ciaramello et al. [7] used a four-question, multiple-choice survey given on a computer at the end of each video in their subjective sign language video evaluation. The first question, "*What was the name of the main character in the story?*" was asked to encourage the participants to pay close attention to the contents of the video, and was not used in any statistical tabulation. The second question was "*How difficult would you say it was to comprehend the video?"* with five possible answers: very easy (1.00), easy (0.75), neither easy nor difficult (0.50), difficult (0.25) and very difficult (0.00). The third question asked "*How would you rate the annoyance level of the video?"* this time with four possible answers: not at all annoying (1.00), a little annoying (0.66), somewhat annoying (0.33) and lastly extremely annoying (0.00). The fourth question asked of the participant would use a video cell phone at this video quality. The subjective intelligibility and annoyance ratings for each video were calculated for each video by averaging each participant's answers to the two questions.

2.7 Summary

Sign Language being a visual language, conveying meaning through a combination of hand shapes, movement of hands and arms, in addition to facial expressions, requires a visual telecommunication channel, making video the only appropriate means of first language telecommunications for the Deaf community.

Video quality can be evaluated either subjectively, capturing viewers opinion of video material, or objectively, using mathematical analysis of the video. The objective evaluations, although good at predicting perceived quality in terms of aesthetics, are not as applicable to quantifying the intelligibility of video material as a lot more is involved than purely if it looks good. In addition Sign Language is not a single language but has many variations across the world, as well as different dialects within the same sign language, such as SASL.

Video communications using mobile phones provides three main challenges, low bandwidth, low processing speed and limited battery life. In an attempt to overcome these challenges Sign Language specific video compression techniques have been investigated, but these techniques rely on modified versions of the standard video encoders to provide better compression, and this is not possible to implement on all phones, especially at the lower end of the market (the target audience of this research).

This research is not focused on video compression schemes, but on the effect of the reduction of video resolution and frame rate on the intelligibility of video containing SASL. The objective is to evaluate the intelligibility of the sign language video, not the picture quality of the video.

University of Cape Town

3 Pilot user study (Experiment 1)

Based on the ITU requirements and limitations (see Section 2.8.1), and the aim of subjective evaluation of Sign Language video on a cell phone a pilot study was conducted to validate the questionnaire with the Deaf participants for evaluating the intelligibility of SASL video on a cell phone (see *Appendix A*).

3.1 Aim

The pilot user study aimed to validate the questionnaire with the Deaf participants for evaluating the intelligibility of SASL video on a cell phone, to uncover any problems with the planned experimental setup. Reducing the video resolution and frame rate is the simplest way to reduce video file size, and thus the required amount of data to transfer over the cell phone network. This experiment only looked at the impact of video resolution and frame rate, keeping compression constrained to 256 kbps in all of the test videos.

3.2 Background

The size of a video file is determined by three basic settings: the video resolution (spatial resolution), video frame rate (temporal resolution) and how the video has been compressed.

3.2.1 Video Resolution

Video resolution is the size (width and height) of the frames in the video. The lower the resolution the less detail in the video content and the less storage is needed per video frame.

This experiment will be looking at two resolutions, namely:

- 320 x 240 (Quarter Video Graphics Array (QVGA))
- 174 x 144 (3GP)

The resolution of 352 x 288, although an industry standard resolution for video compression and used for capturing video on cell phones, is a higher resolution than the physical screen on the cell phones used can display and was for this reason dropped from the study. It would have been nice to go above 320 x 240, but the standards for cheaper cell phones meant this was not feasible.

3.2.2 Video Frame Rate

The video frame rate is the number of frames of video stored and displayed per second. The lower the frame rate the less storage is needed per second of video, but at lower frame rates less detail is visible of objects in motion and blurring of the image starts occurring, which can become a problem especially in Sign Language.

This experiment will be looking at the following three frame rate values:

- 30 frames per second
- 15 frames per second
- 10 frames per second

3.2.3 Video Compression

Video compression is used to process the frames of the video, at the given resolution and frame rate, to further reduce the amount of storage required by the video. The size reduction and resulting quality of the final video is dependent on not only which video compression algorithm was used, but also which compression and quality settings were used. But in general the more the video is compressed the lower the quality and the smaller the file size.

University of Cape Town

In this experiment video compression was kept to a minimum and consistent throughout the twelve video clips, to be able to see only the impact that resolution and frame rate has on the size and the intelligibility of the video.

3.3 Procedure

3.3.1 Participants

Five adult members of the Deaf community (five men, no woman) ranging in age from 33 to 46 (mean = 36) participated in this study. All were native signers and have used SASL as their principal mode of communications all their lives. The five participants were all staff members of DCCT, and had English as their language of literacy, regardless of what their hearing families used.

All participants were introduced to the experiment and each signed a consent form to confirm that they fully understand the project, agree to participate and understand that all information provided would be kept confidential.

3.3.2 Experimental setup

The participants were gathered in high ceilinged, open room with fluorescent lighting and windows on one side. They were seated at desks arranged in a half circle, two participants to a desk, with a pack of 12 questionnaires each numbered with A1-A12, B1-B12, and so forth, a pen, as well as a Nokia N96 cell phone preloaded with the correspondingly numbered video clips in front of each participant.

All communications between the researcher and participants were interpreted by a certified SASL interpreter who was known to the participants. Although the questionnaires were explained in SASL and all queries were answered through the SASL interpreter, the questionnaires were provided in written English and answered in written English.

The participants were introduced to the experiment with the help of the SASL interpreter. It was made clear during the introduction that the focus of the experiment was on evaluating the quality of the video clips and the intelligibility of the SASL in the video clips at different quality settings, and not to evaluate the participants' proficiency in SASL.

Seeing that written/spoken English is not the participants' first language, and the questionnaire required the participants to write down what they understood the Sign Language video clip contained, all participants were asked if they are comfortable writing their answers out. They were given the option of giving their responses to the questionnaire through the interpreter. None of the participants took this option, and indicated that they were comfortable with reading the questionnaire and writing down their responses in English.

The participants were asked to view each video clip only once and then finish the questionnaire for that clip, without reviewing the clip, rating the intelligibility of that video clip. This was done to get the participants initial response to the video clip, and not allow the participant to try and review sections of the clip that were unclear. If any sections were unclear that should be reflected in the answers for that clip.

3.3.3 Cell phones

The Nokia N96 cell phones used in this experiment, as shown in *Figure 3-1*, has a screen size of 2.8" (71 mm) diagonally and a resolution of 240 x 320 pixels displaying up to 16 million colours. The N96 cell phone runs Symbian OS 9.3 (S60 rel. 3.2) on a dual ARM 9 264 MHz processor with 128 MB of RAM [25].

It was left up to the participants to decide how the cell phone would be held while viewing the video clips. All participants used the cell phone in the default portrait orientation, at a distance comfortable to each individual.

University of Cape Town

Figure 3-1: A Nokia N96 cell phone.
The cell phone model used in the first two experiments.

3.3.4 Video clips

Twelve video clips were used, each showing the same sign language user in the same environment, with consistent lighting, background and distance from camera, as shown in *Figure 3-2*, signing in SASL.

Figure 3-2: Example frame from sign language video clip.
All video clips showed the same sign language user in the same environment, with consistent lighting, background and distance from camera as seen in this example frame from one of the video clips

These twelve clips were acquired from the Sign Language Education and Development (SLED) SASL dictionary DVD as MPEG-4 files at full resolution and frame rate, and at best possible quality, after which each of the clips were recompressed to the required resolution and frame rate, using MPEG-4 compression, at a fixed bit rate of 256 kbit/s. The aspect ratio was preserved through letterboxing, a technique to fit widescreen video material onto lower aspect ratio screens by adding black bars at the top and the bottom of the video material (see *Figure 3-2*). This was accomplished using *QuickTime Pro 7.6.6 (1720)*.

The basic details of the twelve video clips are listed in *Table 3-1*. The full details of the video clips, including the data rate, file size and duration of each of the video clips are available in *Table A-6*, in *Appendix A*.

Five sets of clips, one set per participant, were then created from the twelve prepared clips. Each set contained the same twelve clips but in a different random order. The randomizing was done using Microsoft Excel.

The reason for the randomizing of the order in which the clips was viewed was twofold. If the participant viewed the clips in the original order, she might assume after a few clips that the next clip will be of better quality than the previous clip. By viewing the clips in a random order of quality this possibility is removed. The order of the clips was also random between participants to ensure there was no accidental influence between participants on the quality evaluation of the clips.

Video No	Resolution (w x h)	Frames per second	Signed phrase
1	320 x 240	30	Could you please fetch me that cup over there?
2	320 x 240	15	Father stands and waits for the taxi.
3	320 x 240	10	After you've played all day, you bath at night.
4	176 x 144	30	Before you go to sleep, brush your teeth.
5	176 x 144	15	Yesterday I tripped and fell.
6	176 x 144	10	Next month I will buy new clothes.
7	320 x 240	30	You put the fork on the left and the knife on the right.
8	320 x 240	15	The boy washed the window. Now it is clean.
9	320 x 240	10	On the plate was a small loaf of bread.
10	176 x 144	30	When you eat pap your tummy feels good.
11	176 x 144	15	Put on your trousers because we are going to church.
12	176 x 144	10	I scatter the seeds and the chickens eat them.

Table 3-1: Experiment 1 – Video clip specification.

These five sets of twelve randomly ordered video clips of differing quality were then copied one set per cell phone to the five Nokia N96 cell phones. Other than the filenames of the twelve files, there was no difference between the phones, the files or how the videos were viewed by the users.

3.3.5 Questionnaire

Each set of questionnaires, as shown in *Appendix A*, contained a cover page explaining the purpose of the experiment and provided a summary of the experimental procedure. For each video clip to be evaluated a questionnaire was attached consisting of seven questions divided into two freeform questions and five scale questions. All answers were captured. The answers to the freeform questions were not assigned a numeric value, while the answers to the five Likert scale questions were assigned a numeric value. The more acceptable the video, the higher the value assigned to the answer.

Question 1

What was said in this video?

Following the questioning technique used by Ciaramello et al. [7] this question served two purposes, the first to encourage the participant to pay attention to what was being said in the video, and concentrate on understanding what was said in the video, and secondly to get an idea of how close to the original phrase the participant understood the message.

The answer to this question was captured, but no numeric value was assigned to the answer.

University of Cape Town

Question 2

How sure are you of your answer to Question 1 above?

Possible answer	completely sure	sure	so-so	not sure	not sure at all

The second question aims to provide a numeric value to the comprehensibility of the sign language in the video clip. This question functions in conjunction with question 1, and provides an opportunity to check the participants answers. If the participant correctly wrote down the signed phrase in question 1, the answer to this question should show the participant sure of his answer.

This question was assigned a numeric value, with *completely sure* given a value of 5, down to 1 for *not sure at all*.

Question 3

How easy or how difficult was it to understand what was said in this video?

Possible answer	very difficult	difficult	average	easy	very easy

Question 3 was also derived from the work done by Ciaramello et al. [7] and was included as a further check of comprehensibility, this time changing the wording as well as order of values, to help to confirm the participant's ability to understand the contents of the video clip. The first three questions should correlate closely and if all three point in the same direction it is a good indication of the comprehensibility of the sign language contents at the given resolution and frame rate.

This question was assigned a numeric value, with *very easy* given a value of 5, down to 1 for *very difficult*.

Question 4

How easy or how difficult was it to follow the facial expressions in this video?

Possible answer	very difficult	difficult	average	easy	very easy

Sign Language uses two main parts of the body for communications, the face and the hands of the speaker. Question 4 and 5 focuses on these two areas and attempts to quantify the impact lowering the frame rate and resolution has on the comprehension of these areas separately. Question 4 focuses on the face of the speaker.

This question was assigned a numeric value, with *very easy* given a value of 5, down to 1 for *very difficult*.

Question 5

How easy or how difficult was it to follow the hand gestures in this video?

Possible answer	very difficult	difficult	average	easy	very easy

Sign Language uses two main parts of the body for communications, the face and the hands of the speaker. Question 4 and 5 focuses on these two areas and attempts to quantify the impact lowering the frame rate and resolution has on the comprehension of these areas separately. Question 5 focuses on the hands of the speaker.

This question was assigned a numeric value, with *very easy* given a value of 5, down to 1 for *very difficult*.

University of Cape Town

Question 6

If you could chat using a cell phone with video this easy/difficult to understand, would you use it?

Possible answer	definitely yes	yes	maybe	no	definitely no

The last question used in the analysis, question 6, was added to the questionnaire as a final summary question, to get an overall view of the participants' opinion of the clip, the intelligibility of the clip and the clip's usability in cell phone based SASL video communications.

This question was assigned a numeric value, with *definitely yes* given a value of 5, down to 1 for *definitely no*.

Question 7 *(Numbered 4, by error, on the handed out questionnaire)*

Any other comments on this video?

Question 7 provided the participant the opportunity to give any general comments on the just viewed and evaluated video clip.

As with question 1, the answer to this question was captured, but no numeric value was assigned to the answer.

3.4 Observations

With the participants' willingness and aptitude to write down their responses in English, and not having to rely on the SASL interpreter for answering each question, the experiment ran smoothly and efficiently. The spelling and grammar of the responses of what was said in the each video might seem peculiar to a first language English speaker, but this is because of the distinct difference in grammar between English and SASL, as well as English not being the participants' first language.

There were few hiccups and misunderstandings. All the participants had no problem selecting a video file to play, moving between video files and playing a video, but the numbering of the files and the order the phones listed the files in made it problematic for the participants to find the specific video file they were looking for. The files were named A1, A2 …. A11, A12 and because the phone listed the files alphabetically they were listed as A1, A10, A11, A12, A2, A3 … A9. The only other misunderstanding was one of the participants understood the instruction to view the clip only once before answering the full questionnaire as view the clip once before every question in the questionnaire. A quick explanation cleared up the misunderstanding.

Other than the confusing file order no further help was needed by any participants in selecting and playing the video files. All participants were clearly familiar and comfortable using the cell phones. It took the participants roughly an hour to view all twelve clips and finish the questionnaires.

3.5 Results

Subjective intelligibility ratings were calculated for each video from the participants' answers to the questionnaire. These average participant ratings were calculated by averaging the participants' answers to each question for each of the videos. An overall rating was also calculated for each video frame rate and resolution combination by averaging all participants' answers to the five questions for each of the combinations.

A one-way ANOVA analysis of variance was completed to determine if any of the six video clips were preferred over the other video clips. The one-way ANOVA compares the means between the groups and determines whether any of those means are significantly different from each other. It tests the null hypothesis that all the means of the groups are the same, in this case that all the video clips had the same mean participant rating, irrespective of the video resolution or frame rate. If the one-way ANOVA returns a significant result, a significance value $p < 0.05$ then we accept the

University of Cape Town

alternative hypothesis, which is that there are at least two video clips rating means that are significantly different from each other.

Question	Mean						ANOVA
	320 x 240 pixels			*176 x 144 pixels*			
	30 fps	*15 fps*	*10 fps*	*30 fps*	*15 fps*	*10 fps*	*p*
How sure are you?	2.90	3.30	2.90	3.20	3.50	3.10	.856
How easy to understand?	3.20	3.80	3.50	3.50	3.80	3.20	.564
Face	3.30	3.90	3.50	3.30	3.80	3.40	.628
Hand gestures	3.30	3.80	3.30	3.70	3.90	3.20	.420
Use on cell phone	3.40	3.40	3.40	3.40	3.40	3.20	.991
Average rating	3.22	3.64	3.32	3.42	3.68	3.22	.674

Table 3-2 : Statistical analysis for the intelligibility measures of Experiment 1.
None of the questions yielded statistically significant results.

Table 3-2 contains the mean participant rating for each video clip, as well as the ANOVA significance value for the five questions. As can be seen in the table all of the questions returned a significance level of greater than 0.05 ($p > 0.05$) and, therefore, there is no statistically significant difference in the mean participant rating for each of the video clips. No combination of frame rate and video resolution, either high or low, was preferred significantly more or less than any other combination of frame rate and resolution.

Figure 3-3 to *Figure 3-7* show the average participant rating for the each of the questions answered by the participants in the questionnaire, with *Figure 3-8* showing the overall average participant rating across all questions.

The average response is plotted on the vertical axis, with 5 = very easy to understand (high intelligibility) and 1 = very difficult to understand (low intelligibility).

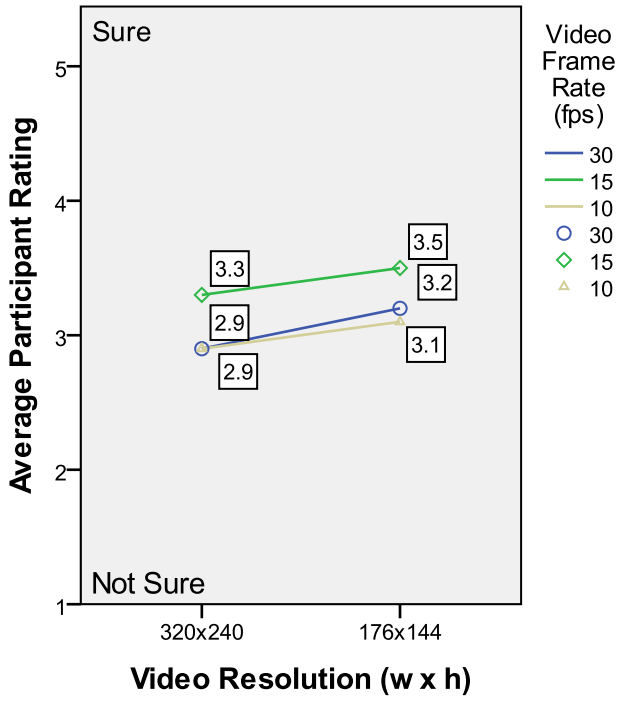

Figure 3-3: Qualitative results for Question 2.
The qualitative results for the question "*How sure are you of your answers to Question 1 above?*" for each of the three frame rates and two resolutions. With a significance level of 0.856 ($p = .856$) there was no statistically significant difference in the average participant rating for each of the video clips.

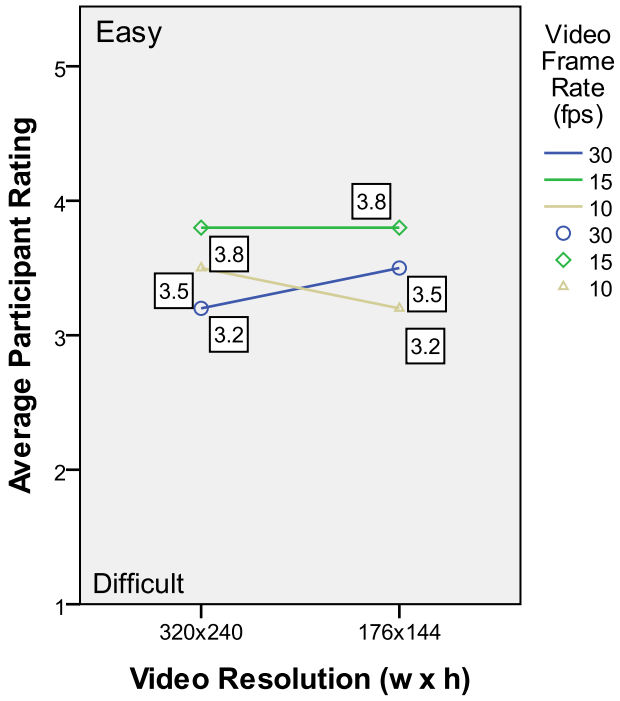

Figure 3-4: Qualitative results for Question 3.
The qualitative results for the question "*How easy or difficult was it to understand what was said in this video?*" for each of the three frame rates and two resolutions. With a significance level of 0.564 ($p = .564$) there was no statistically significant difference in the average participant rating for each of the video clips.

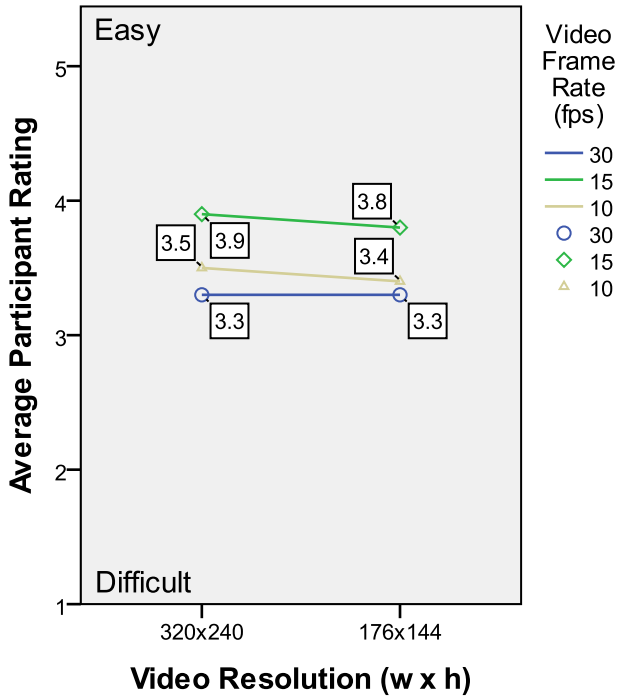

Figure 3-5: Qualitative results for Question 4.
The qualitative results for the question "*How easy or difficult was it to follow the facial expressions in this video?*" for each of the three frame rates and two resolutions. With a significance level of 0.628 (p = .628) there was no statistically significant difference in the average participant rating for each of the video clips.

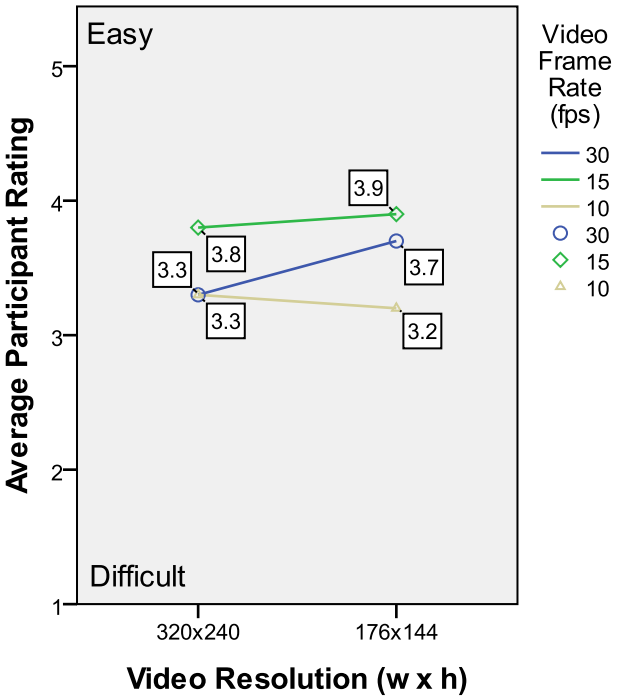

Figure 3-6: Qualitative results for Question 5.
The qualitative results for the question "*How easy or how difficult was it to follow the hand gestures in this video?*" for each of the three frame rates and two resolutions. With a significance level of 0.420 (p = .420) there was no statistically significant difference in the average participant rating for each of the video clips.

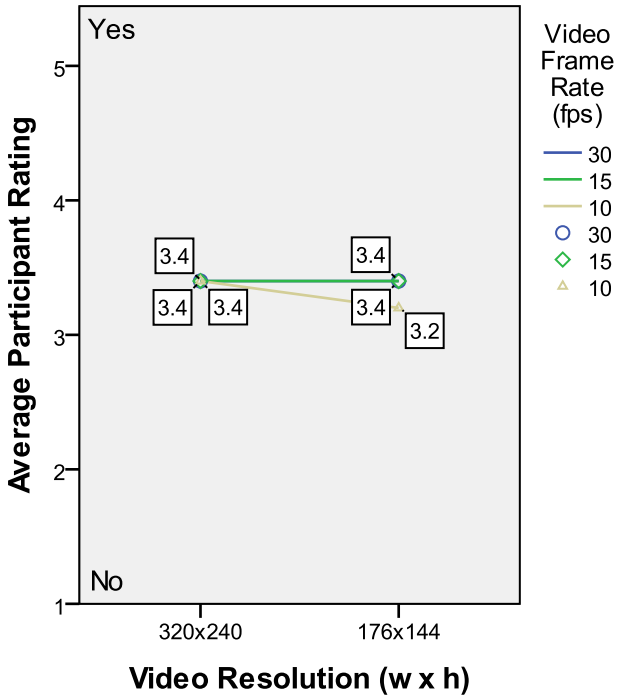

Figure 3-7: Qualitative results for Question 6.
The qualitative results for the question "*If you could chat using a cell phone with video this easy/difficult to understand, would you use it?*" for each of the three frame rates and two resolutions. With a significance level of 0.991 (p = .991) there was no statistically significant difference in the average participant rating for each of the video clips. ***(Note: The 30 fps and 15 fps lines are on top of each other.)***

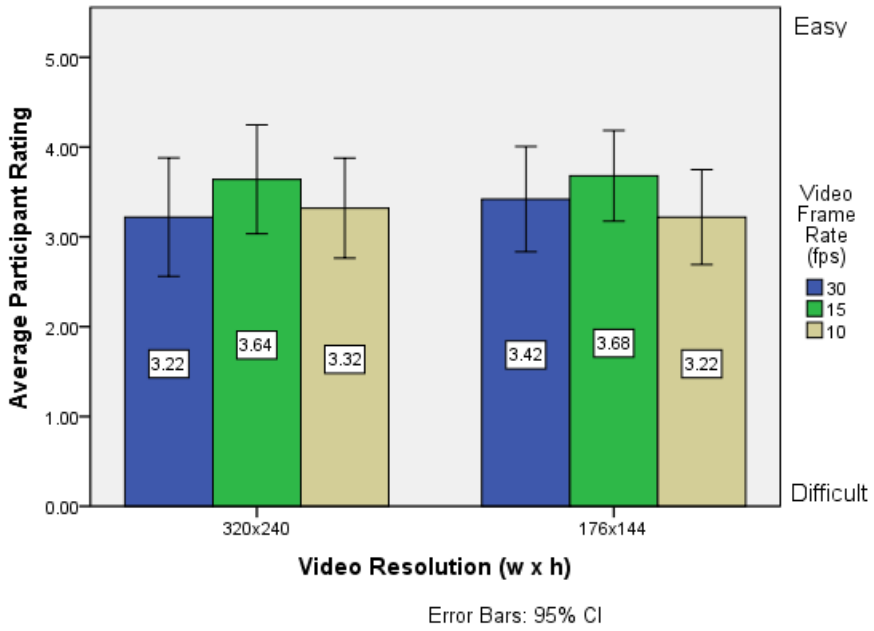

Figure 3-8: Overall mean participant response and across all questions.
The overall mean participant response and standard deviation across all questions for each of the three frame rates and two resolutions. The y-axis is the average participant response. Each group on the x-axis is a particular video resolution, with each colour representing a particular video frame rate. With a significance level of 0.674 (p = .674) there was no statistically significant difference in the average participant rating for each of the video clips.

Despite there being no statistically significant results, a general trend is evident. Across all questions, as well as in the overall qualitative results, the participants preferred a frame rate of 15 frames per second.

It is interesting to note that a frame rate of 15 fps is consistently preferred over the higher frame rate of 30 fps. Also video resolution seems to play less of a role in the participant's evaluation of the video clips' intelligibility.

A few comments were vague as to what was meant and would need further investigation:

"nothing, not clear"

"not clear"

It is not obvious if "not clear" refers to the quality of the video, or the meaning of the message signed in the video.

The comments also pointed towards influences on the intelligibility of the video clip, other than video quality:

"That was easy and the way she do, understandly"

"Easy body clear picture"

"Easy slow sign language"

"sign difficult"

"easy clear, but not clear not word"

"I should have no problem. She is good"

"Should have no problem, she is not good. Because she must clear sign language"

The way the person in the video clip signed, as well as the signs used have an impact on the intelligibility and the opinion of the viewer on the video clip.

University of Cape Town

In addition to the signing technique and the actual signs used, three factors over and above the video clip frame rate and resolution could have influenced the participants' rating of the video clips.

The first possible factor was that video clips used in this experiment were all compressed at a fixed data rate of 256 kbits/s. Looking at the experimental results the video clips shown at 30 frames per second scored lower overall, at all resolutions, compared to 15 and 10 frames per second. When looking at resolution one would expect the highest resolution to be scored consistently high in intelligibility, but this was not the case. Both these observations could be explained by the fixed data rate limiting video quality at higher frame rates and resolutions.

As discussed earlier at low resolutions and frame rates the amount of bandwidth needed is low, but as the resolution and frame rate is increased the amount of information per video frame increases and correspondingly the bandwidth requirement increases as well. The cell phone does not support the playback of completely uncompressed video material. Because the codec was forced to keep the required bandwidth limited to 256 kbits/s video quality had to be sacrificed at higher resolutions and frame rates. The focus of these experiments is to the evaluate the impact of frame rate and resolution on the intelligibility of sign language video, not the quality of video. At this bit rate the video compressor has to reduce video quality to keep the bandwidth limitation. In future experiments the allowed bit rate will be taken far beyond that required at these resolutions and frame rates to reduce the impact of video compression to an absolute minimum.

Figure 3-9: Letterboxed video frame, as used in Experiment 1.
The video aspect ratio is kept constant by adding black bars to the top and bottom of each video frame. Relatively large sections of the background are visible in the final video clip.

Figure 3-10: Cropped video frame, as should have been used in Experiment 1.
The video aspect ratio is that of the final clip, not the source material. None of the video frame is used for unnecessary background or black bars.

Secondly the video clips taken from the DVD were letterboxed (*Figure 3-9*) down to the test resolutions, instead of being cropped (*Figure 3-10*) to the desired resolution. The DVD material was shot at a wide screen aspect ratio of 16 x 9, while the cell phone screen has of aspect ratio of 4 x 3. In addition to the screen real estate wasted on black bars, the wide screen format included extraneous background that was never used by the signer in the video. A better technique would have been to crop the DVD material to the required resolutions resulting in the full screen of the cell phone used to show the signer with no space wasted on black bars or background, better fitting the available screen area to the signing space. This would have also simulated the video captured by the cell phone more accurately.

The last factor that could have influenced the responses from the participants is unrelated to the quality of the video clips, and is the actual content of the video clips. To evaluate the intelligibility of the video clips, questions were asked about the content of the video clips and if the participants

understood what was communicated in the clip through SASL. But what if the participant could clearly distinguish the face and hand movements in a video clip but actual signs used in the video clip were unfamiliar to the participant? This would have had a negative effect on the participant's rating of intelligibility of that video clip.

4 Follow-up Pilot User Study (Experiment 2)

4.1 Aim

The aim of the follow-up pilot study was to check the compression issue, eliminating the impact of limited bit rate on higher resolutions and frame rate video clips, as well as validate the new questionnaire incorporating the lessons learned in the initial pilot study (see Appendix B).

Through the experience and findings of the pilot study the following changes were made to the experimental setup in the follow-up study:

- To remove as much video quality degradation due to video compression artefacts and get as close as possible to uncompressed video all clips were compressed at a data rate of 5 000 kbits/s.
- To more accurately resemble the video that would have been captured on the cell phone itself when using the cell phone as a video based Sign Language communications medium, all clips were cropped (*Figure 3-10*) instead of being letterboxed (*Figure 3-9*). This removed extraneous background area, better fitting the available screen area to the signing space of the signer, making full use of the cell phone's screen area and keeping the resultant video clips at the same aspect ratio as resolutions being tested.
- To encourage the participants to focus more and give more detailed feedback throughout, the number of clips were reduced by including only one clip from each resolution-frame rate combination. In addition the questions were simplified.
- In the pilot study a few comments mentioned the video being "*not clear*". To attempt to expand on these comments the two questions in the pilot study covering the details of the video were extended to five questions, to include motion blurring, the speed of the video in addition to the facial and hand detail visibility.
- And lastly in an attempt to factor in the possible unfamiliarity of the actual signs used in the video clips the participants were allowed to view the clip as many times as desired before and while finishing the questionnaire. The number of times a clip was viewed was captured on the questionnaire and a question was added to specifically investigate this factor.

4.2 Procedure

4.2.1 Participants

Six adult members of the Deaf community (three women, three men) ranging in age from 33 to 64 (mean = 38) participated in this study. All were native signers and have used SASL as their principle mode of communications all their lives. The six participants were all staff members of DCCT, and had English as their language of literacy, regardless of what their hearing families used. Three had taken part in the first pilot study.

All participants were introduced to the experiment and each signed a consent form to confirm that they fully understand the project, agree to participate and understand that all information provided would be kept confidential.

4.2.2 Experimental Setup

The follow-up experiment was conducted in the same high ceilinged, open venue as the pilot study. The six participants were seated at similar desks arranged in a half circle, two participants to a desk, with an individually numbered pack of six questionnaires, a pen, as well as a Nokia N96 cell phone, preloaded with the corresponding video clips in front of each participant.

University of Cape Town

All communications between the researcher and participants were interpreted by a certified SASL interpreter who was known to the participants. Although the questionnaires were explained in SASL and all queries were answered through the SASL interpreter, the questionnaires were provided in written English and answered in written English.

The participants were introduced to the experiment with the help of the SASL interpreter. It was made clear during the introduction that the focus of the experiment was on evaluating the quality of the video clips and the intelligibility of the SASL in the video clips at different quality settings, and not to evaluate the participants' proficiency in SASL.

Seeing that written/spoken language is not the participants' first language, and the questionnaire required the participants to write down what they understood the Sign Language video clip contained, all participants were asked if they are comfortable writing their answers out. They were given the option of giving their responses to the questionnaire through the interpreter. None of the participants took this option, and indicated that they were comfortable with writing down their responses in English.

In the first experiment some of the questions had to be explained while the participants were answering the questionnaires, in addition there were some confusion in finding clips as well as how many times a clip were to be viewed. In an attempt to alleviate these problems and make sure each completed questionnaire were completely reliable, a practice video clip, similar in look and difficulty to those used in the experiment, and practice questionnaire, identical to the questionnaire used in the experiment, were added to the experiment. This was done to enable the participants to familiarise themselves with the phone and questionnaire. This provided them with an opportunity to ask for clarification on any of the questions, as well as the use of the cell phone as they worked through the practice questionnaire. After an introduction to and demonstration of the cell phone, the participants were asked to view the practice clip on the cell phone and answer the separate loose practice questionnaire. The answers to practice questionnaire was not captured or used in the experiment.

After all the participants finished the practice evaluation in their own time, it was confirmed with each participant individually that they were comfortable with the questionnaire and could select and play any of the video clips, and move to the next video clip without problems. They were then given the go ahead to start filling in the questionnaires evaluating the six video clips. The participants were allowed to view any clip as many times as they wanted to, with a count of the views noted on the questionnaire.

University of Cape Town

4.2.3 Cell phones

The same set of Nokia N96 cell phones that were used in the pilot study were used in this experiment and again it was left up to the participants to decide how the cell phones would be held while viewing the clips. As in the pilot study all participants held the phone in the default portrait orientation, at a distance comfortable to each individual participant. Same held the phone in their hand while some preferred the phone lying flat on the table while viewing a video clip.

4.2.4 Video clips

Six video clips were used, each showing the same sign language user in the same environment, with consistent lighting, background and distance from camera, signing in SASL.

The same two resolutions as were used in the pilot study were used in the follow-up study, namely:

- 320 x 240 (QVGA)
- 174 x 144 (3GP)

Similarly the same three frame rates as in the pilot study were used in the follow-up study, namely:

- 30 frames per second
- 15 frames per second
- 10 frames per second

Where in the pilot study the video was taken directly from the widescreen DVD material, resized to the desired resolutions by letterboxing, in this experiment the video clips was cropped before being taken to the desired resolution. This made sure no space on the cell phone screen was wasted with black bands or unused background area, making much better use of the available screen resolution, better fitting the available screen area to the signing space of the signer. And giving an accurate simulation of the screen real estate usage as would be the case when the phone was used for video communication.

These six clips were acquired from a DVD, as MPEG-4 files at full resolution and frame rate, and at best possible quality. Each of the clips were cropped and recompressed to the required resolution and frame rate, using the Export (Using QuickTime conversion) feature of *Final Cut Express (v4.0.1*).

A data rate of 5000 kbits/sec was used to minimise the impact of the video compression on the quality of the resulting video clip.

The basic details of the six video clips are shown in *Table 4-1*. The full details of the video clips, including the data rate, file size and duration of each of the video clips are available in *Table B-7*, in *Appendix B*.

Five sets of clips, one set per participant, were then created from the six prepared clips. Each set contained the same six clips but in a different random order. The randomizing was done using Microsoft Excel.

Video No	Resolution (w x h)	Frames per second	Signed phrase
1	320 x 240	30	The girl rides the horse.
2	320 x 240	15	The man bounces the ball on his head.
3	320 x 240	10	The small boy is dirty all over.
4	176 x 144	30	Tomorrow is my birthday.
5	176 x 144	15	Yesterday I caught a big fish.
6	176 x 144	10	Your T-shirt is too small for you.

Table 4-1: Experiment 2 - Video clip specifications

The order of the clips was randomised to minimise the possibility that the participants could assume the next clip would be of better quality than the previous. In addition the clips were randomised between sets to ensure that there was no accidental influence between participants on the quality evaluation of the clips.

These five sets of six randomly ordered video clips of differing quality was then copied one set per cell phone to five Nokia N96 cell phones. Other than the filenames of the six files, there was no difference between the phones, the files or how the videos were viewed by the users.

As six participants and only five phones were available on the day of the experiment, the fifth phone was shared between participant E and F. Thus the clip order for video clip set E and video clip set F was identical. However, close observation showed that the two participants looked at the clips and answered the questionnaire completely separately.

University of Cape Town

4.2.5 Questionnaire

Each set of questionnaires, as shown in *Appendix B,* contained a cover page explaining the purpose of the experiment and provided a summary of the experimental procedure.

For each video clip to be evaluated a questionnaire was attached. All answers were captured, but the answers to the freeform questions were not assigned a numeric value, while the answers to the five scale questions were assigned a numeric value. The more acceptable the video, the higher the value assigned to the answer.

In addition to the six clips to be evaluated, a practice video clip was added. This clip and a separate loose questionnaire sheet were used to explain and familiarise the participants with playing the video clips, understanding the questionnaire and answering the questionnaire. When all participants felt comfortable with the phone and the questionnaire, they were given to go ahead to evaluate the six video clips.

Question 1

What was said in this video?

As in the pilot user study, this question served two purposes. The first was to encourage the participant to pay attention to what is being said in the video, and concentrate on understanding what is said in the video. The second was to get an idea of how close to the original phrase the participant's understanding of the message was.

The answer to this question was captured, but no numeric value was assigned to the answer.

Question 2

How sure are you of your answer to Question 1 above?

Possible answer	completely sure	sure	so-so	not sure	not sure at all

The second question aims to provide a numeric value to the comprehensibility of the sign language in the video clip. This question functions in conjunction with question 1, and provides an opportunity to check the participants answers. If the participant correctly wrote down the signed phrase in question 1, the answer to this question should show the participant sure of his answer.

This question was assigned a numeric value, with *completely sure* given a value of 5, down to 1 for *not sure at all*.

Question 3

How easy or how difficult was it to understand what was said in this video?

Possible answer	very difficult	difficult	average	easy	very easy

Question 3 was kept as is from the pilot study and is included as a further check of intelligibility, this time changing the wording as well as order of values, to help to confirm the participant's ability to understand the contents of the video clip. The first three questions should correlate closely and if all three point in the same direction give a good indication of the intelligibility of the sign language contents at the given resolution and frame rate.

This question was assigned a numeric value, with *very easy* given a value of 5, down to 1 for *very difficult*.

Question 4

Please select the appropriate choice from the options provided below.

From the results of the pilot study it was decided to simplify, but also broaden the evaluation of the different aspects of the video quality from the perspective of the Deaf user.

University of Cape Town

In the pilot study quite a few comments mentioned blurry motion and the speed of the video. The two questions in the pilot study covering the details of the video were extended to five questions, to include motion blurring, the speed of the video in addition to the facial and hand detail visibility.

A fifth question was added to determine if a low score on intelligibility is purely because of the quality of the video or if unfamiliarity of a Sign Language phrase were impacting on the scoring of the video clips.

Question 4.1

Possible answer	*The movement was clear.*	*The movement was blurry.*

From the comments in the pilot study, blurred video was often a problem in the clips. This question was added in response to these comments, it is focussed on the movement of the hands and arms being blurred, something that is expected to happen at lower frame rates.

The answer to this question was captured as a numeric value, with 5 being given to *the movement was clear* and a value of 1 to *the movement was blurry*.

Question 4.2

Possible answer	*I could clearly see all the details of the face.*	*I had difficulty seeing the details of the face.*

This question was present in the pilot study, but has been simplified in this study to a binary answer.

Sign Language uses two main parts of the body for communications, the face as well as the hands of the speaker. Question 4.2 and 4.3 focuses on these two areas and attempt to evaluate the impact lowering the frame rate and resolution has on the comprehension of these areas separately. Question 4.2 focused on the face of the speaker.

The answer to this question was captured as a numeric value, with 5 being given to *I could clearly see all the details of the face* and a value of 1 to *I had difficulty seeing the details of the face.*

Question 4.3

Possible answer	*I could clearly see the hands.*	*I had difficulty seeing the hands.*

This question was present in the pilot study, but has been simplified in this study to a binary answer.

Sign Language uses two main parts of the body for communications, the face as well as the hands of the speaker. Question 4.2 and 4.3 focuses on these two areas and attempt to evaluate the impact lowering the frame rate and resolution has on the comprehension of these areas separately. Question 4.3 focused on the hands of the speaker.

The answer to this question was captured as a numeric value, with 5 being given to *I could clearly see the hands* and a value of 1 to *I had difficulty seeing the hands.*

Question 4.4

Possible answer	*The video was the right speed.*	*The video was too slow/too fast.*

This question, as is the case with question 4.1, was added in response to the comments participants made during the pilot study. The two main complaints were blurred motion and the speed of the video clip being wrong.

University of Cape Town

This and blurring of motion is a function of the frame rate of the video clip. The lower the video clip's frame rate the lower the rating should be for questions 4.1 and 4.4.

The answer to this question was captured as a numeric value, with 5 being given to t*he video was the right speed* and a value of 1 to *the video as too slow/too fast*.

Question 4.5

Possible answer	*I knew all the signs.*	*Some signs were unknown to me.*

The pilot brought another question to mind. If a participant finds one of the Sign Language phrases unfamiliar, what impact will that have on their evaluation of the intelligibility of the video clip? This question was added in response.

The answer to this question was captured as a numeric value, with 5 being given to *I knew all the signs* and a value of 1 to *some signs were unknown to me*.

Question 5

How many times did you view this clip?

It was decided in this experiment to do away with the single view of video clip constraint and rather provide the participant the opportunity to review the clip as needed, but record the number of views on the questionnaire.

The single view constraint was removed to more closely resemble the conversational use of a Sign Language video clip where the listener could ask the signer to resign the previous phrase.

Question 6

Any other comments on this video?

Question 6 provided the participant the opportunity to give any general comments on the just viewed and evaluated video clip.

As with question 1, the answer to this question was captured, but no numeric value was assigned to the answer.

4.3 Observations

With only six clips used during this experiment, instead of the 12 as in the pilot, there was no problem with oddly ordered clips (A1, A10, A11, A12, A2, A3 … A8, A9) and all video clips were selected and viewed without any problems. Again all participants were clearly familiar and comfortable using the cell phones.

Again, as with the first pilot study, the participants were willing to write down their responses in English.

Where the assistance of the SASL interpreter was needed though, was helping with the correct spelling of words, and in a few cases the English word for a sign. The words in these cases were simply finger spelled out for the participant.

4.4 Results

Subjective intelligibility ratings were calculated for each video from the participants' answers to the questionnaire. These average participant ratings were calculated by averaging the participants' answers to each question for each of the videos. An overall rating was also calculated for each video frame rate and resolution combination by averaging all participants' answers to the five questions for each of the combinations.

A one-way ANOVA analysis of variance was completed to determine if any of the six video clips were preferred over the any of the other video clips. The one-way ANOVA compares the means between the groups and determines whether any of those means are significantly different

University of Cape Town

from each other. It tests the null hypothesis that all the means of the groups are the same, in this case that all the video clips had the same mean participant rating, irrespective of the video resolution or frame rate. If the one-way ANOVA returns a significant result, a significance value $p < 0.05$ then we accept the alternative hypothesis, which is that there are at least two video clips rating means that are significantly different from each other.

Question	Mean						ANOVA
	320 x 240 pixels			*176 x 144 pixels*			
	30 fps	*15 fps*	*10 fps*	*30 fps*	*15 fps*	*10 fps*	*p*
How sure are you?	3.50	3.33	3.33	4.50	3.67	3.50	.732
How easy or difficult to understand?	3.33	3.33	3.17	4.17	3.67	3.33	.840
Movement	5.00	4.33	3.40	5.00	5.00	3.40	.172
Face	4.33	4.33	4.33	4.33	5.00	3.67	.825
Hands	3.67	4.33	4.20	5.00	5.00	3.67	.491
Speed	3.67	4.33	5.00	5.00	5.00	5.00	.208
Signs	4.20	3.67	3.00	5.00	4.33	4.33	.462
Average Rating	3.79	3.95	3.58	4.71	4.52	3.83	.540

Table 4-2 : Statistical analysis for the intelligibility measures of Experiment 2.
None of the questions yielded statistically significant results.

Table 4-2 contains the mean participant rating for each video clip, as well as the ANOVA significance value for each of the seven questions as well as for the average participant rating over all the questions. As can be seen in the table all of the questions returned a significance level of greater than 0.05 ($p > 0.05$) and, therefore, there is no statistically significant difference in the mean participant rating for each of the video clips. No combination of frame rate and video resolution, either high or low, was preferred significantly more or less than any other combination of frame rate and resolution.

Figure 4-1 to *Figure 4-7* show the average participant rating for each of the questions answered by the participants in the questionnaire, with *Figure 4-8* showing the overall average participant rating across all questions.

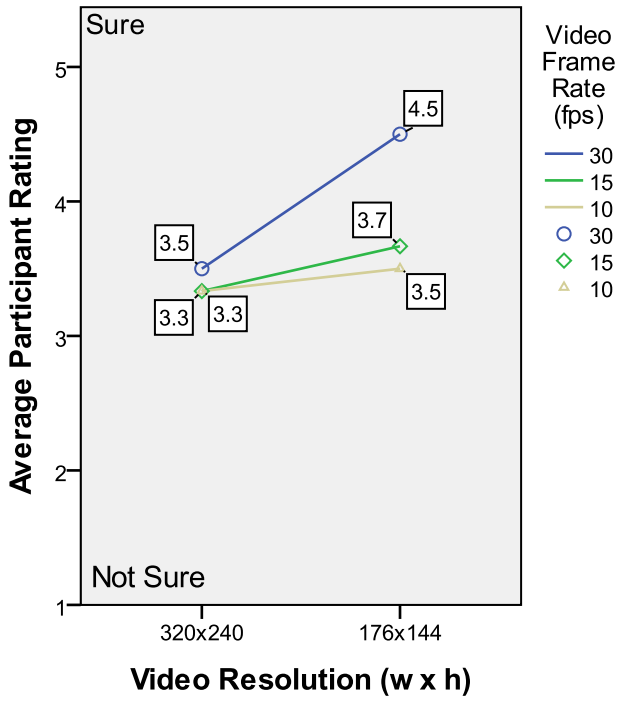

Figure 4-1: Qualitative results for Question 2.
The qualitative results for the question **"***How sure are you of your answers to Question 1 above?"* for each of the three frame rates and two resolutions. With a significance level of 0.732 ($p = .732$) there was no statistically significant difference in the average participant rating for each of the video clips.

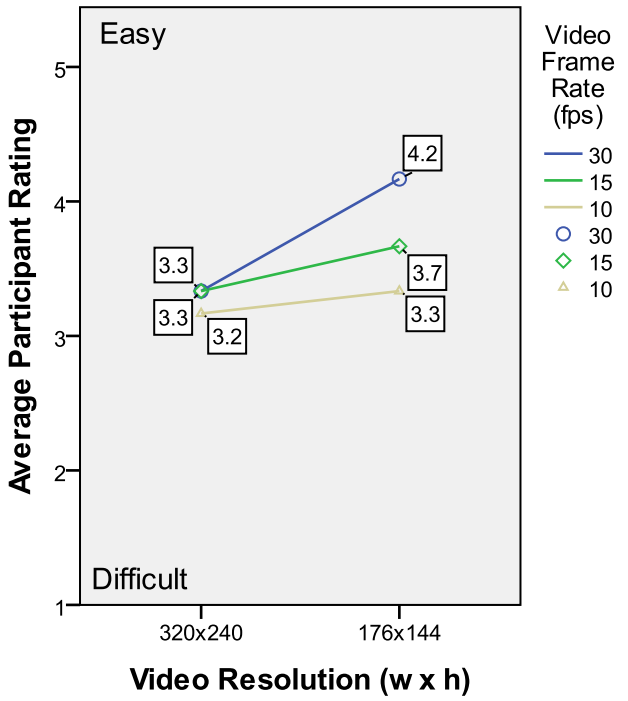

Figure 4-2: Qualitative results for Question 3.
The qualitative results for the question **"***How easy or how difficult was it to understand what was said in this video?"* for each of the three frame rates and two resolutions. With a significance level of 0.840 ($p = .840$) there was no statistically significant difference in the average participant rating for each of the video clips.

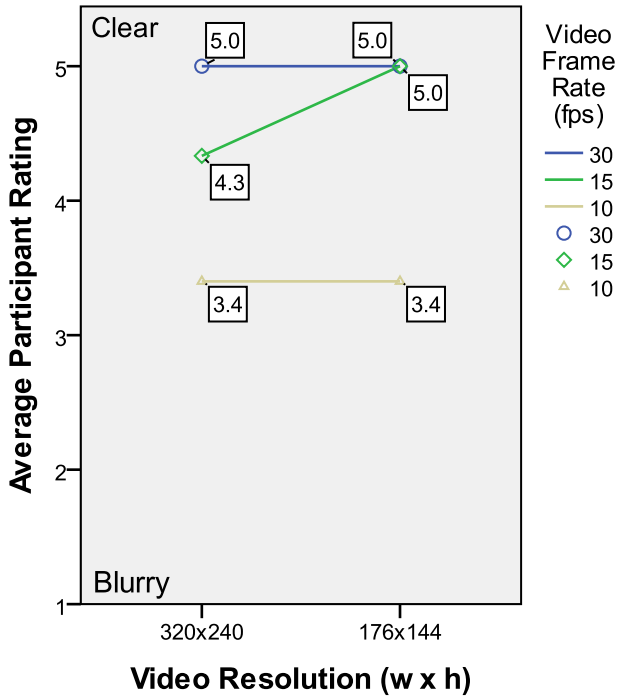

Figure 4-3: Qualitative results for Question 4.1.
The qualitative results for the question **"***The movement was clear/blurry.*" for each of the three frame rates and two resolutions. With a significance level of 0.172 (p = .172) there was no statistically significant difference in the average participant rating for each of the video clips.

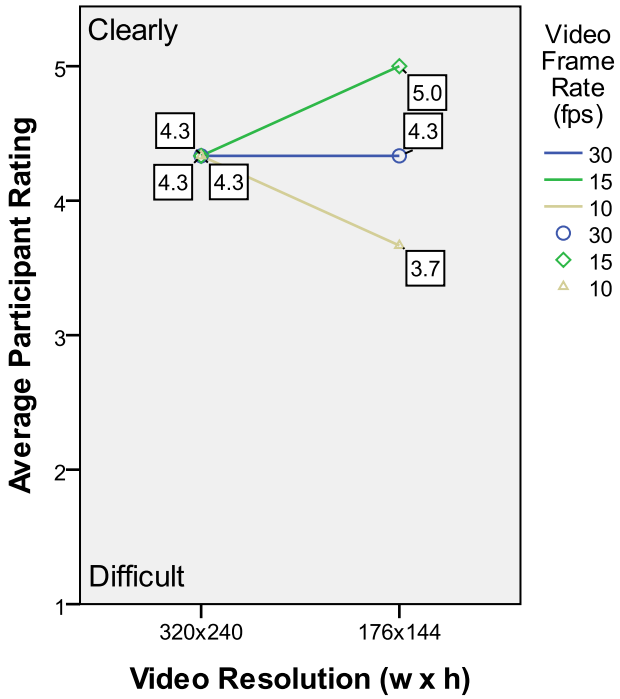

Figure 4-4: Qualitative results for Question 4.2.
The qualitative results for the question **"***I could clearly see all the details of the face/I had difficulty seeing the details of the face."* for each of the three frame rates and two resolutions. With a significance level of 0.825 (p = .825) there was no statistically significant difference in the average participant rating for each of the video clips.

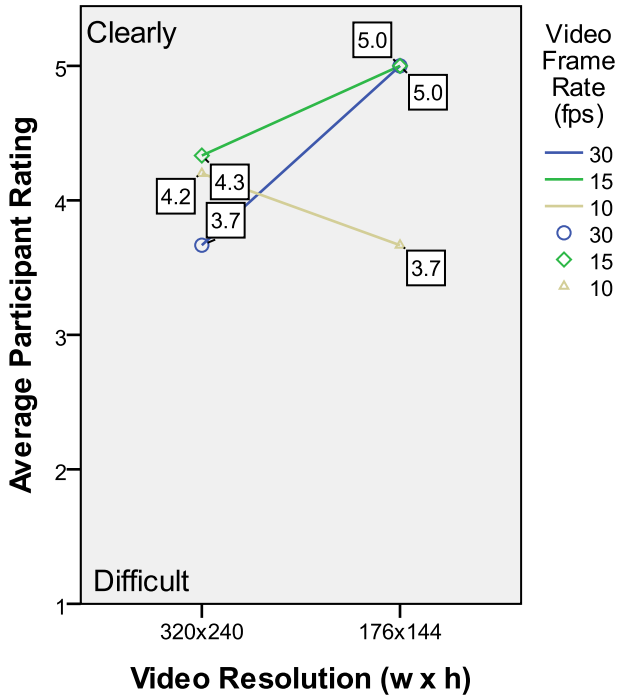

Figure 4-5: Qualitative results for Question 4.3.
The qualitative results for the question **"***I could clearly see the hands/I had difficulty seeing the hands."* for each of the three frame rates and two resolutions. With a significance level of 0.491 (p = .491) there was no statistically significant difference in the average participant rating for each of the video clips.

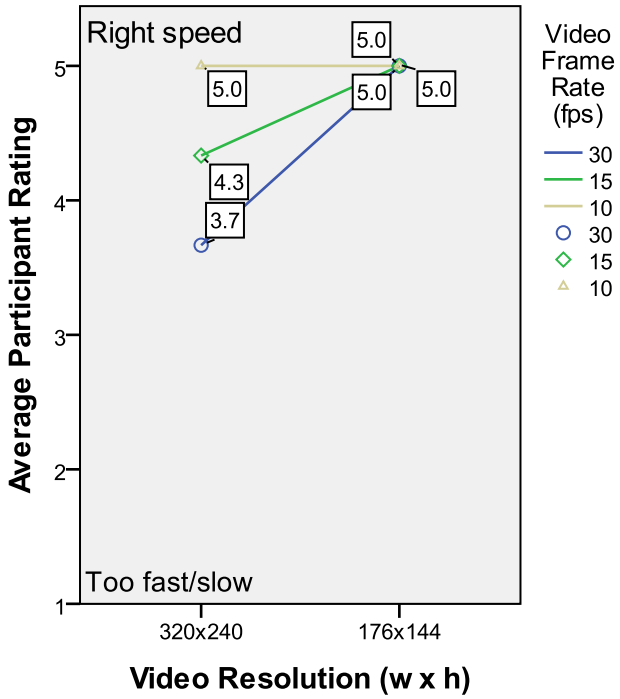

Figure 4-6: Qualitative results for Question 4.4.
The qualitative results for the question **"***The video was the right speed/The video was too slow/too fast."* for each of the three frame rates and two resolutions. With a significance level of 0.208 (p = .208) there was no statistically significant difference in the average participant rating for each of the video clips.

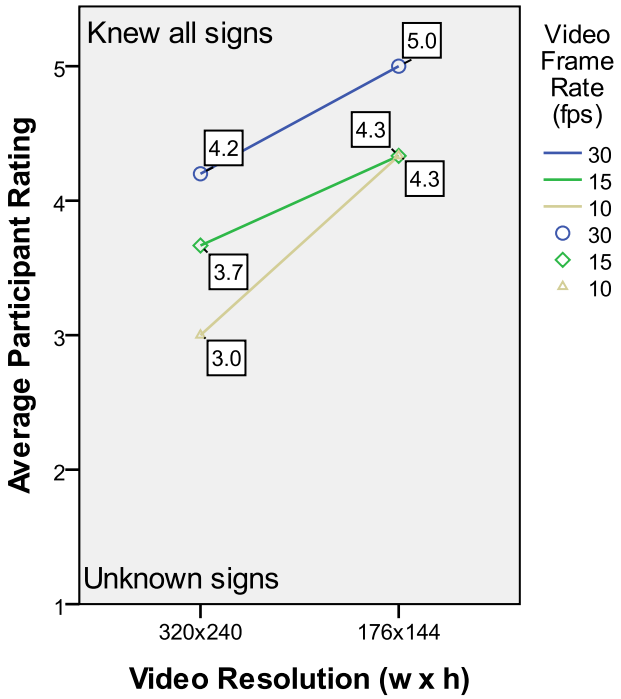

Figure 4-7: Qualitative results for Question 4.5.
The qualitative results for the question **"***I knew all the signs/Some signs were unknown to me."* for each of the three frame rates and two resolutions. With a significance level of 0.462 (p = .462) there was no statistically significant difference in the average participant rating for each of the video clips..

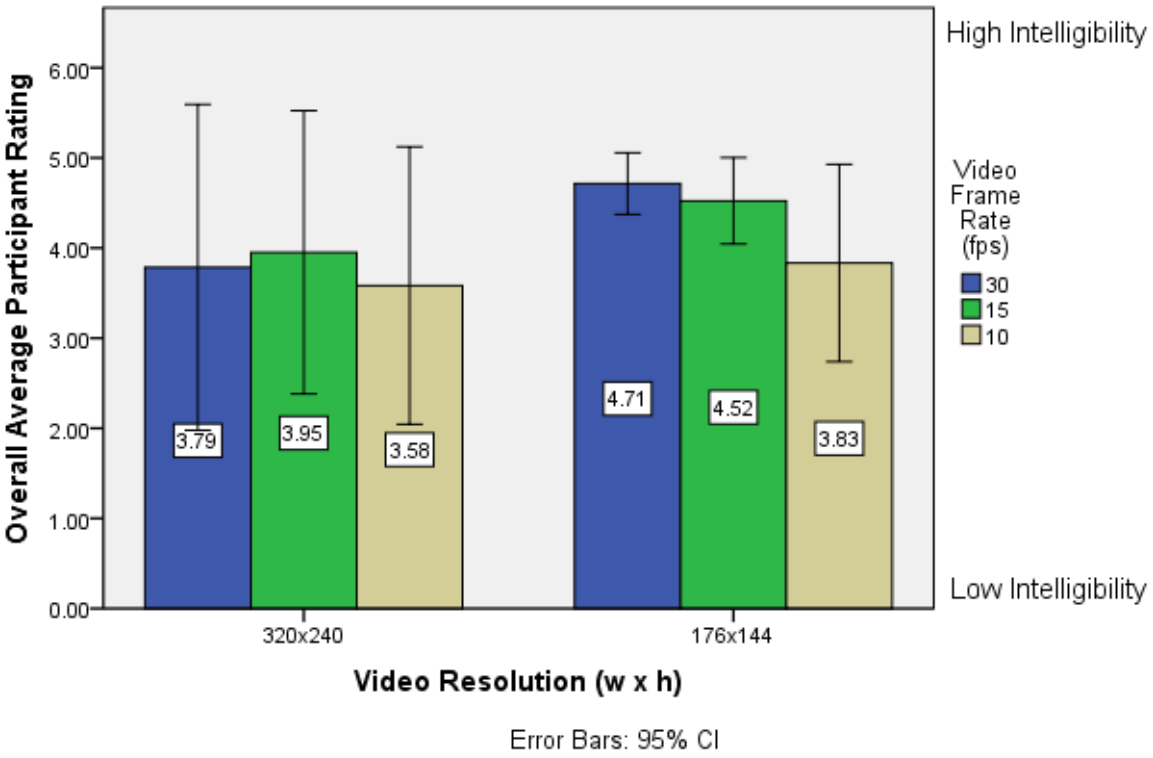

Figure 4-8: Overall mean participant response across all questions.
The overall mean participant response and standard deviation across all questions for each of the three frame rates and two resolutions. The y-axis is the average participant response. Each group on the x-axis is a particular video resolution, with each colour representing a particular video frame rate. With a significance level of 0.540 (p = .540) there was no statistically significant difference in the average participant rating for each of the video clips.

Again, as with the pilot study, there were no statistically significant results. The video clip at a resolution 176 x 144, at 15 frames per second, were the only clip from the six clips that all six participants agreed that the movement was clear, could see clearly all the details in the face, could clearly see the hands, and the video was at the right speed.

The impact of compression was eliminated from this experiment by compressing the video clips with a bit rate specification far above the required bit rate for the given resolutions and frame rates. Looking at the final bit rate values listed in *Table 4-1*, all final bit rates are all well below the specified bit rate of 5000 kbits/sec, with the highest being only 2663.28 kbit/sec.

Where the comments made by the participants in the first pilot study hinted at problems with unknown signs, the follow-up study's comments confirmed that the actual signs used in the test video clips are impacting on the intelligibility of the videos:

"some sign language confuse"

"boy – different sign language"

"sign language bit confuse"

"problem with sign play"

"problem with sign boy"

"No, it was difficult about sign language"

Only one clip (176 x 144 at 30 frames per second) was confirmed to contain only known signs by the participants. Whether the rest of the clips truly contained signs unknown to the viewer or it was simply a case of not being able to identify the sign because the sign was not clear enough in the video, is not clear from the results.

This experiment added a question to the questionnaire to test the possibility of unknown signs impacting on the subjective evaluation of the intelligibility of the video clips. The response to Question 4.5 (shown in *Figure 4-7*), as well as the comments from participants points to the different SASL dialects leading to a single sign making perfect sense to some participants, while being completely unknown to other participants.

Again, as in the first pilot study, no statistically significant results were recorded. Two possible reasons may be the small sample group, and the fact that the same participant evaluated multiple clips of different resolutions and frame rates. The same participant viewing and evaluating multiple clips at different specifications could impact the participant evaluation of subsequent video clips.

University of Cape Town

5 Intelligibility Study (Experiment 3)

Based on the results and gathered feedback of the two pilot studies the final intelligibility study was conducted (see *Appendix C*).

5.1 Aim

The final intelligibility study aimed to incorporate all the lessons learnt through the two preceding pilot studies to answer the main research question: *What is the lowest video resolution and frame rate that would provide intelligible SASL video on a cell phone?*

Through the experience and findings of the pilot studies the following changes were made to the experimental setup in the final intelligibility study:

- Both pilot studies gave no statistically significant results. In an effort to increase the chances of the final experiment giving a statistically significant result the number of participants was greatly increased. In addition, to remove the possibility that a participant's response to a specific video clip could be impacted by a previously viewed video clip, each participant only watched and gave feedback on a single video clip. With multiple participants evaluating the same video clip.
- The binary questions of the follow-up study could not give a clear enough picture of the participants' opinions and were replaced with a consistent set of five-level Likert items all using the typical Likert scale, this not only simplified the explanation of the questionnaire to the participants, but also the analysis of the answers. It was now possible to directly compare answers and form a true overview value.
- The question statements in this study were based on the statements used in the binary questions of the second pilot study. Each set of binary statements used in the follow-up study became two Likert scale statements in this study.
- The two pilot studies both gave statistically insignificant results, with no clear preference or rejection of any of the resolution-frame rate combinations. In an effort to attain statistical significance the number of resolution-frame rate combinations being compared was reduced. With the number of resolutions being evaluated already at only two, the number of different frame rates was reduced. With the objective of finding the lowest usable frame rate and resolution supporting intelligible SASL video communications the 30 frames per second frame rate was dropped focusing on the lower end of the frame rate scale. In a further effort to attain statistical significance 15 frames per second was replaced by 20 frames per second to have a more distinct difference in frame rate between the video clips.
- To minimise the possibility of the actual signed contents of the clips having an impact on the evaluation of the video clip, only signed phrases with no known dialectic differences were used.
- Lastly following the results of the original pilot study and the subsequent changes to the preparation of the video clips, the compression and cropping of the video clips were kept the same as for the second pilot study.

5.2 Procedure

5.2.1 Participants

Twenty four adult members of the Deaf community (twelve women, twelve men) ranging in age from 20 to 64 (mean = 37) participated in this study. All the participants were native signers and have used SASL as their principle mode of communications for most of their lives, with years of SASL experience ranging from 10 to 60 years (mean = 32). Six of the participants were staff

University of Cape Town

members of DCCT, with the remaining eighteen participants being visitors to The Bastion. Five of the participants participated in one or both of the pilot studies. Of the twenty four, sixteen had English as their language of literacy, two Afrikaans, one Xhosa, three English and Afrikaans and two participants used both English and Xhosa as their reading and writing language. A further five participants completed the experiment, but they were removed from the study because of too many unanswered or wrongly answered questions.

All participants were introduced to the experiment and each signed a consent form to confirm that they fully understand the project, agree to participate and understand that all information provided would be kept confidential.

5.2.2 Experimental Setup

The final intelligibility experiment was conducted across two separate days at The Bastion in Newlands, Cape Town. Because of the number of participants involved in the final experiment, they were handled in groups of between four and eight participants at a time, depending on availability.

Each participant was seated at a desk with a pen and a copy of the questionnaire.

All communications between the researcher and participants were interpreted by a certified SASL interpreter. Although the questionnaires were explained in SASL and all queries were answered through the SASL interpreter, the questionnaires were provided in written English and answered in written English.

The participants were introduced to the experiment with the help of the SASL interpreter. It was made clear during the introduction that the focus of the experiment was on evaluating the quality of the video clips and the intelligibility of the SASL in the video clips at different quality settings, and not to evaluate the participants' proficiency in SASL.

Seeing that written/spoken language is not the participants' first language, and the questionnaire required the participants to write down what they understood the Sign Language video clip contained, all participants were asked if they are comfortable writing their answers out. They were given the option of giving their responses to the questionnaire through the interpreter.

The questionnaire and how to answer the questions were explained to the group of participants after which they were given the opportunity to read through the questionnaire at their own pace, asking for clarification on any of the questions. With only one video clip to be viewed by each participant and no clear advantage provided by the practice video clip and questionnaire in the second experiment, no practice questionnaire was used.

When all participants in the group were ready, the researcher moved from one participant to the next showing one of the four video clips to each participant using the Vodafone 858 Smart cell phone. Each participant could watch their video clip once, after which they were given the go ahead to complete the questionnaire on the clip they were shown.

The same clip was never shown to two adjacent participants to make sure that no two participants could influence each other's answers.

5.2.3 Cell phones

The Nokia N96 cell phones that were used in the two initial experiments were no longer available by the time the third experiment was conducted. In the final experiment the Nokia N96 was replaced with the Vodafone 858 Smart cell phone, as shown in *Figure 5-1* [42]. The Vodafone 858 Smart has a screen size of 2.8" (71 mm) diagonally and a resolution of 240 x 320 pixels, similar to the Nokia N96 in both physical screen size as well as resolution, but where the Nokia is capable of displaying up to 16 million colours, the Vodafone 858 can only display 256K colours. Because of the similar physical screen sizes as well as similar resolutions between the two phones the Vodafone 858 was deemed an equivalent replacement for the Nokia N96 in the third experiment. The

University of Cape Town

Vodafone 858 Smart cell phone runs Android OS, v2.2.1 (Froyo) on a 528 MHz ARM 11 processor with dynamic underclocking and an Adreno 200 GPU. It is equipped with 256 MB RAM, of which 180 MB is accessible to applications [41].

5.2.4 Video clips

Four video clips were used, each showing the same sign language user in the same environment, with consistent lighting, background and distance from camera, signing in SASL.

To simplify the experiment and limit the study to four groups it was decided to focus on only two frame rates, namely 20 and 10 frames per second.

Figure 5-1: A Vodafone 858 Smart.
The cell phone model used in the final experiment.

Four clips were acquired from a DVD, as MPEG-4 files at full resolution and frame rate, and at best possible quality. Each of the clips were then recompressed to the required resolution and frame rate, using the Export (Using QuickTime conversion) feature of *Final Cut Express (v4.0.1*).

As was done in the follow-up pilot study (Experiment 2) the source video clips were resized to the desired resolutions by cropping the frames. This made sure no space on the cell phone screen was wasted with black bands or unused background area, making much better use of the available screen resolution. And giving an accurate simulation of the screen real estate usage as would be the case when the phone was used for video communication.

A data rate of 5000 kbits/sec was used to minimise the impact of the video compression on the quality of the resulting video clip.

One clip was created for each of the four possible combinations of resolution and frame rate. The basic details of these four video clips are shown in *Table 5-1*. The full details of the video clips, including the data rate, file size and duration of each of the video clips are available in *Table C-2*, in Appendix C.

Video No	Resolution (w x h)	Frames per second	Signed phrase
1	320 x 240	20	He is a short man.
2	320 x 240	10	The family is home.
3	176 x 144	20	I read a book.
4	176 x 144	10	I want that apple.

Table 5-1: Experiment 3 - Video clip specifications

5.2.5 Questionnaire

Each set of questionnaires, as shown in *Appendix C*, contained a cover page explaining the purpose of the experiment and provided a summary of the experimental procedure, as well as consent form to be signed by each participant to confirm that they understand the project, they agree to participate and that all information provided will be kept confidential.

On the back of this page was a short form to gather background information about each participant, including gender, age, preferred reading and writing language, as well as number of years the participant has been speaking SASL.

The second page contained the questionnaire to be completed by the participant to evaluate the sign language video clip.

Question 1

What was said in this video?

As in both pilot user studies, this question served two purposes. The first was to encourage the participant to pay attention to what is being said in the video, and concentrate on understanding what is said in the video, and secondly to get an idea of how close to the original phrase the participant understood the message.

No numeric value was assigned to the answer.

Question 2

I am sure of my answer to Question 1?

Possible answer	strongly disagree	disagree	neither agree nor disagree	agree	strongly agree
	1	2	3	4	5

This question functions in conjunction with question 1, and provides an opportunity to check the participants answers. If the participant correctly wrote down the signed phrase in question 1, the answer to this question should show the participant sure of his answer.

Remaining questions

The remainder of the questionnaire consisted of seventeen five-level Likert items all using the typical Likert scale, as was used in question 2.

strongly disagree	disagree	neither agree nor disagree	agree	strongly agree
1	2	3	4	5

These questions were grouped into sets, the statements in each set testing the same feature of the video, but one in a confirmative and the other in a negative phrasing. The order of the questions was randomised to limit the answers of one question influencing the other question in the pair.

Sign Language uses two main parts of the body for communications, the face as well as the hands of the speaker. The first four groups of questions focuses on these two areas and attempt to

University of Cape Town

evaluate the impact lowering the frame rate and resolution has on the comprehension of these areas separately.

Hands

14. I could clearly see the hands.

18. It was difficult to see the hands.

Hand Gestures

3. It was difficult to follow the hand gestures in this video.

8. I could clearly see all the hand gestures in this video.

Face

10. I had difficulty seeing the details of the face.

12. I could clearly see the details of the face.

Facial Expressions

5. I had no problems seeing the facial expressions in this video.

9. It was difficult to follow the facial expressions in this video.

The movement and video speed is focussed on the movement of the hands and arms being blurred, something that is expected to happen at lower frame rates. It evaluates the general feel of the video clip, separate from the specifics of the face and the hands.

Movement

4. The movement was blurry.

7. The movement was clear.

Video speed

6. The video was the right speed.

15. The video was too slow.

19. The video was too fast.

The last two groups of the questionnaire focuses purely on the intelligibility of the video clip, and not on the quality of the video clip. Because of the different dialects in SASL, a sign used in the video clip might be a known sign to one participant, while completely senseless or out of context to another participant speaking a different dialect of SASL.

Signs

16. I knew all the signs used in this video.

17. Some signs used in this video were unknown to me.

Understanding

11. I had difficulty to understand what was said in this video.

13. It was easy to understand what was said in this video.

5.3 Observations

The experiment ran smoothly, with only a few misunderstandings and recurring questions.

The Sign Language interpreter's assistance was needed a few times answering the first question of the questionnaire to help the participants with spelling or finding the written word for a specific sign. This occurred more often than was the case in the two preceding pilot studies because of the wider range of literacy of the participants, compared to the initial groups consisting of all DCCT staff members.

University of Cape Town

Two questions needed regular explanation. The first being the general information question: *Number of years using South African Sign Language*. This question was most problematic to the participants that grew up using Sign Language, and could also have been stated more clearly by asking since what year the participant has been using Sign Language. Question four of the questionnaire was the second recurring problem question, requiring the term "*blurry*" to be explained often.

Despite looking at the phrases used in the video clips to minimise the chances of using a phrase that might have more than one sign, depending on Sign Language dialect, the phrase "*short*" as used *"He is a short man*" was pointed out as an unknown sign by a number of participants, with most of the participants knowing the sign.

5.4 Results

All the above questions were assigned a numeric value, as marked by the participant on the questionnaire. For analysis values for the negative statement were inverted e.g. 1 became 5, 5 became 1, and an overall score was calculated for each questionnaire by summing all the answers.

A one-way ANOVA analysis of variance was completed to determine if any of the four video clips were preferred over the any of the other video clips. The one-way ANOVA compares the means between the groups and determines whether any of those means are significantly different from each other. It tests the null hypothesis that all the means of the groups are the same, in this case that all the video clips had the same mean participant rating, irrespective of the video resolution or frame rate. If the one-way ANOVA returns a significant result, a significance value $p < 0.05$ then we accept the alternative hypothesis, which is that there are at least two video clips rating means that are significantly different from each other.

	Mean				ANOVA
	320 x 240 pixels		*176 x 144 pixels*		
Question	***20 fps***	***10 fps***	***20 fps***	***10 fps***	***p***
Hands					
14. Clearly	4.33	4.33	4.60	4.67	.896
18. Difficult	2.83	2.33	3.50	3.17	.644
Gestures					
3. Difficult	2.00	3.00	3.20	2.50	.527
8. Clearly	3.33	3.67	4.60	4.67	.228
Face					
10. Difficult	2.33	2.17	2.50	2.67	.945
12. Clearly	3.67	4.00	4.40	4.00	.743
Expressions					
5. No problems	3.67	4.67	3.40	3.60	.330
9. Difficult	2.50	3.00	2.83	3.33	.792
Movement					
4. Blurry	3.20	3.00	3.60	2.17	.552
7. Clear	4.17	4.17	5.00	4.33	.416
Video Speed					
6. Right	4.00	3.50	3.83	4.00	.913
15. Too slow	3.20	3.50	1.83	2.00	.133
19. Too fast	2.00	3.50	2.83	2.40	.386
Signs					
16. Known	4.67	3.67	4.17	4.20	.651
17. Unknown	3.33	3.50	3.50	4.17	.670
Understanding					
11. Difficulty	3.67	3.83	2.83	3.00	.540
13. Easy	3.20	4.00	4.60	4.00	.502
2. Sure	4.33	5.00	4.50	3.83	.227
Average Rating	58.83	61.33	61.00	60.33	.990

Table 5-2: Statistical analysis for the intelligibility measures of Experiment 3.
The table shows the significance and mean square values for the eighteen questions, as well as for the overall participant intelligibility rating.

Table 5-2 contains the mean participant rating for each video clip, as well as the ANOVA significance value for each of the eighteen questions as well as for the average participant rating over all the questions. As can be seen in the table all of the questions returned a significance level of

greater than 0.05 ($p > 0.05$) and, therefore, there is no statistically significant difference in the mean participant rating for each of the video clips. No combination of frame rate and video resolution, either high or low, was preferred significantly more or less than any other combination of frame rate and resolution.

Figure 5-2 to *Figure 5-19* show the average participant rating for the each of the questions answered by the participants in the questionnaire, with *Figure 5-20* showing the overall average participant rating across all questions.

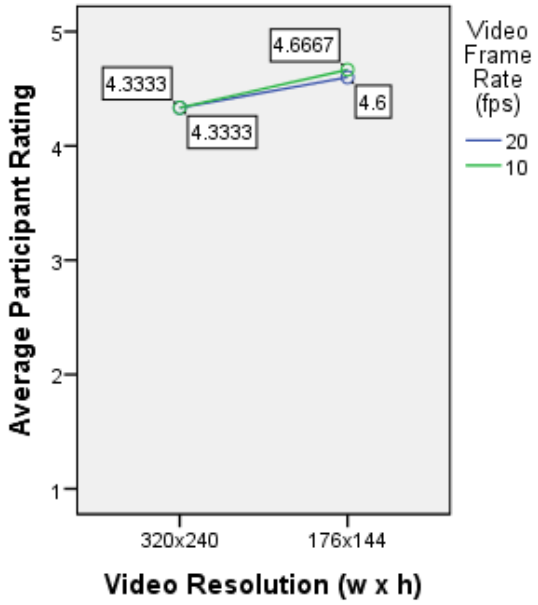

Figure 5-2: The qualitative results for Question 14 "*I could clearly see the hands.*" for each of the three frame rates and two resolutions. With a significance level of 0.896 ($p = .896$) there was no statistically significant difference in the average participant rating for each of the video clips.

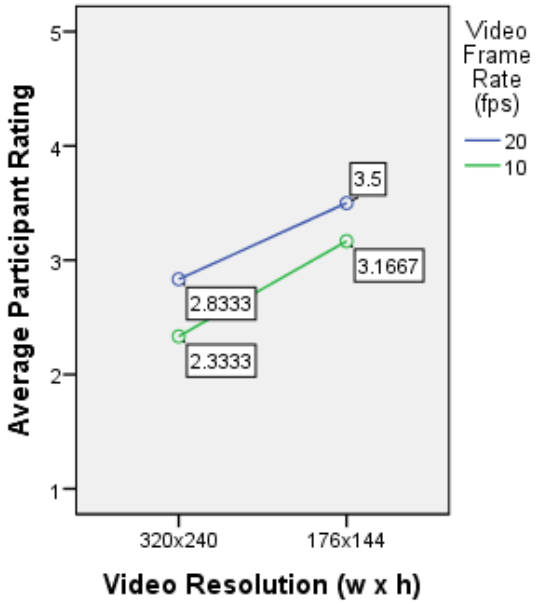

Figure 5-3: The qualitative results for Question 18 "*It was difficult to see the hands.*" for each of the three frame rates and two resolutions. With a significance level of 0.644 ($p = .644$) there was no statistically significant difference in the average participant rating for each of the video clips.

University of

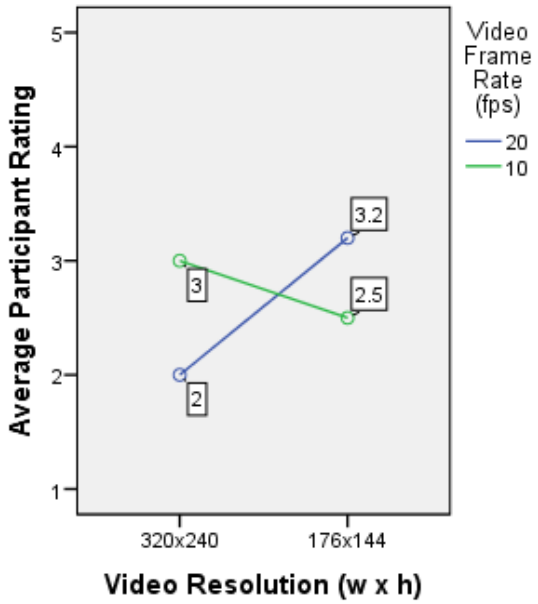

Figure 5-4: The qualitative results for Question 3
"*It was difficult to follow the hand gestures in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.527 (p = .527) there was no statistically significant difference in the average participant rating for each of the video clips.

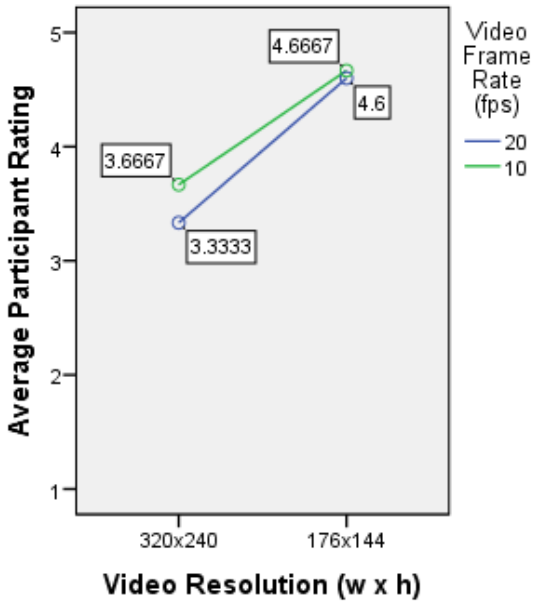

Figure 5-5: The qualitative results for Question 8
"*I could clearly see all the hand gestures in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.228 (p = .228) there was no statistically significant difference in the average participant rating for each of the video clips.

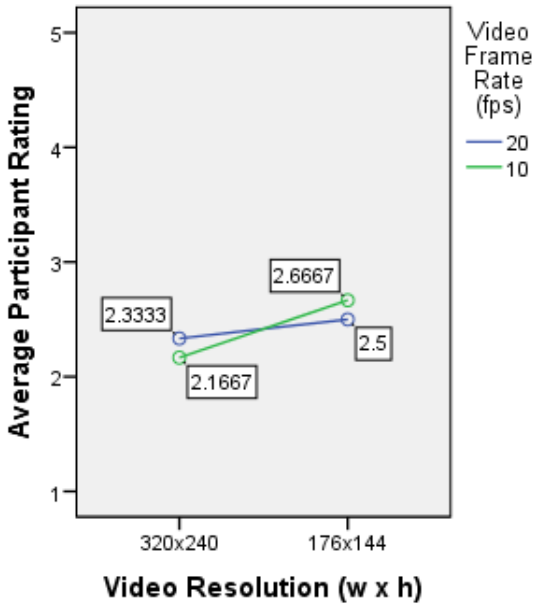

Figure 5-6: The qualitative results for Question 10
"*I had difficulty seeing the details of the face.*" for each of the three frame rates and two resolutions. With a significance level of 0.945 (p = .945) there was no statistically significant difference in the average participant rating for each of the video clips.

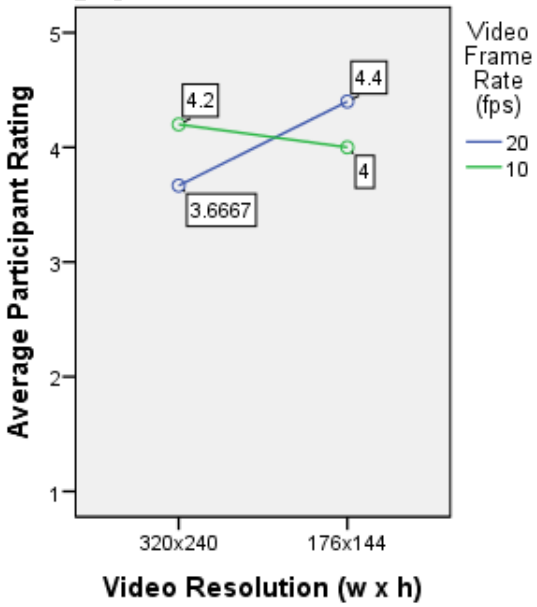

Figure 5-7: The qualitative results for Question 12
"*I could clearly see the details of the face*." for each of the three frame rates and two resolutions. With a significance level of 0.743 (p = .743) there was no statistically significant difference in the average participant rating for each of the video clips.

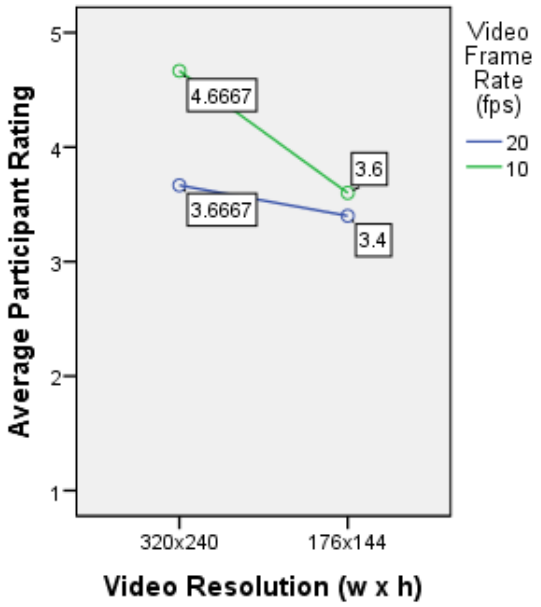

Figure 5-8: The qualitative results for Question 5
"*I had no problem seeing the facial expressions in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.330 (p = .330) there was no statistically significant difference in the average participant rating for each of the video clips.

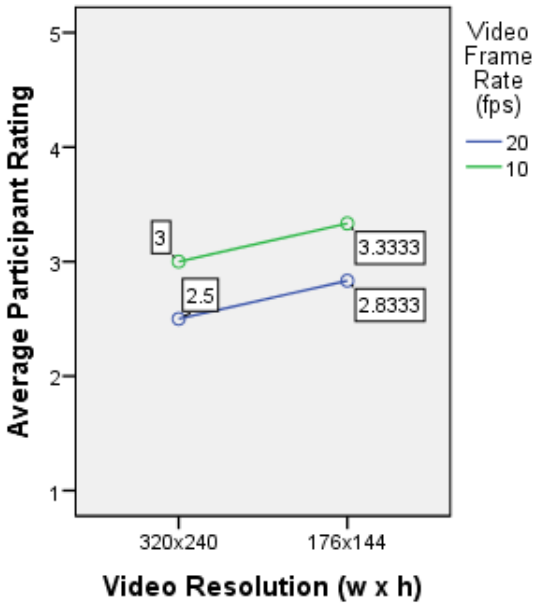

Figure 5-9: The qualitative results for Question 9
"*It was difficult to follow the facial expressions in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.792 (p = .792) there was no statistically significant difference in the average participant rating for each of the video clips.

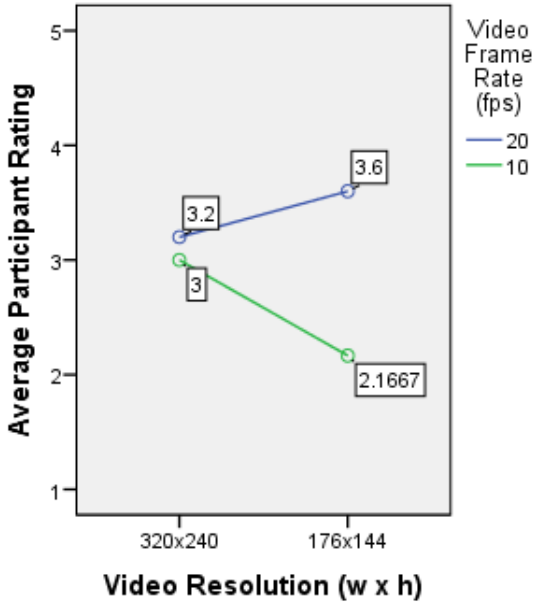

Figure 5-10: The qualitative results for Question 4
"*The movement was blurry.*" for each of the three frame rates and two resolutions. With a significance level of 0.552 (p = .552) there was no statistically significant difference in the average participant rating for each of the video clips.

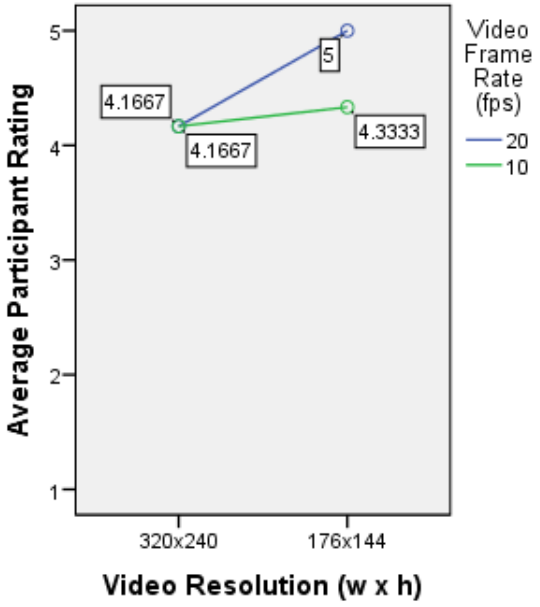

Figure 5-11: The qualitative results for Question 7
"*The movement was clear.*" for each of the three frame rates and two resolutions. With a significance level of 0.416 (p = .416) there was no statistically significant difference in the average participant rating for each of the video clips.

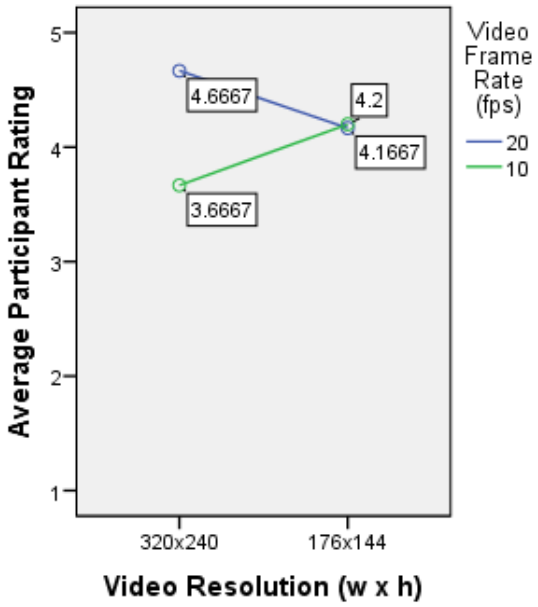

Figure 5-12: The qualitative results for Question 16 "*I knew all the signs used in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.651 (p = .651) there was no statistically significant difference in the average participant rating for each of the video clips.

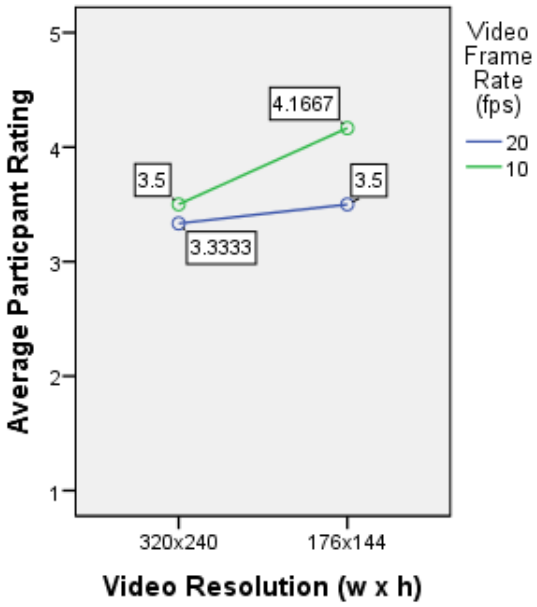

Figure 5-13: The qualitative results for Question 17 "*Some signs used in this video were unknown to me.*" for each of the three frame rates and two resolutions. With a significance level of 0.670 (p = .670) there was no statistically significant difference in the average participant rating for each of the video clips.

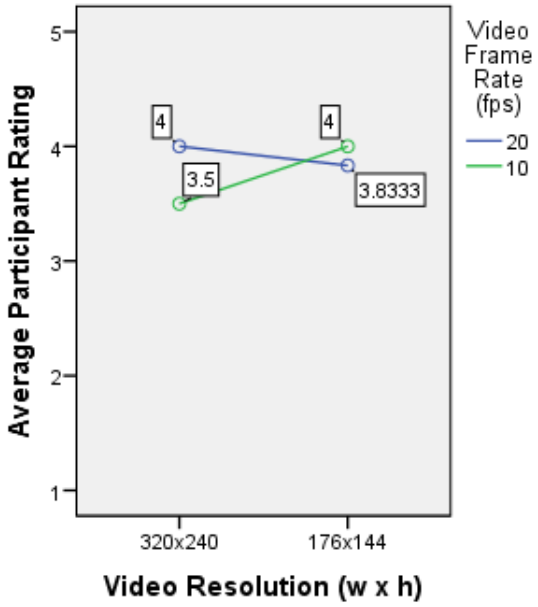

Figure 5-14: The qualitative results for Question 6 "*The video was the right speed.*" for each of the three frame rates and two resolutions. With a significance level of 0.913 (p = .913) there was no statistically significant difference in the average participant rating for each of the video clips.

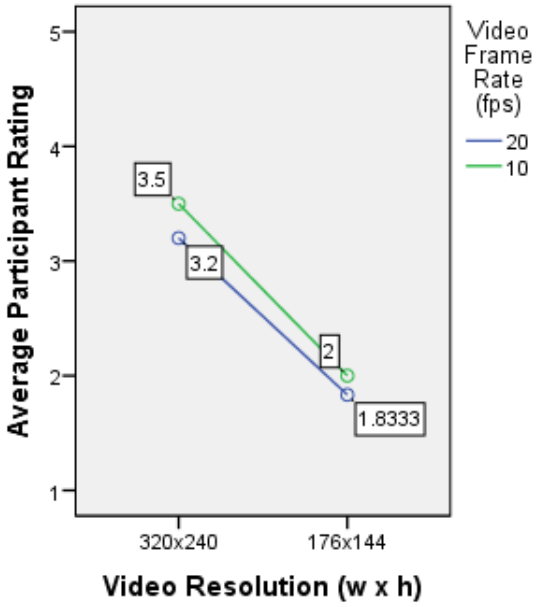

Figure 5-15: The qualitative results for Question 15 "*The video was too slow.*" for each of the three frame rates and two resolutions. With a significance level of 0.133 (p = .133) there was no statistically significant difference in the average participant rating for each of the video clips.

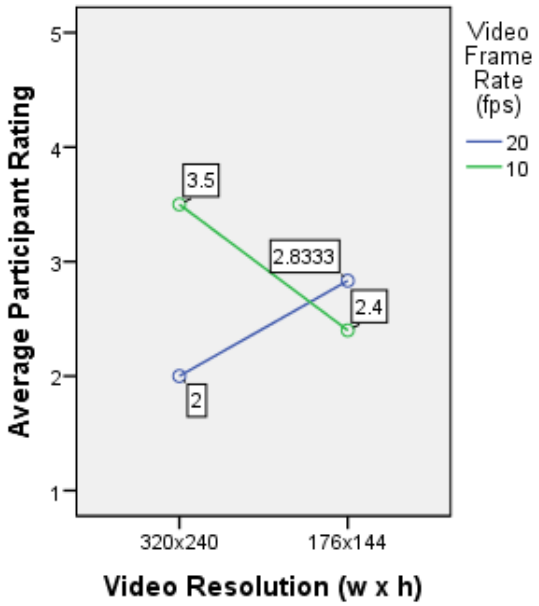

Figure 5-16: The qualitative results for Question 19 "*The video was too fast."* for each of the three frame rates and two resolutions. With a significance level of 0.386 (p = .386) there was no statistically significant difference in the average participant rating for each of the video clips.

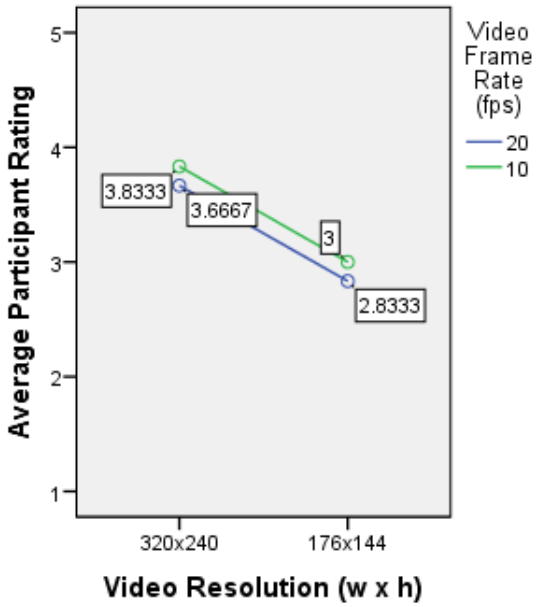

Figure 5-17: The qualitative results for Question 11 "*I had difficulty to understand what was said in this video."* for each of the three frame rates and two resolutions. With a significance level of 0.540 (p = .540) there was no statistically significant difference in the average participant rating for each of the video clips.

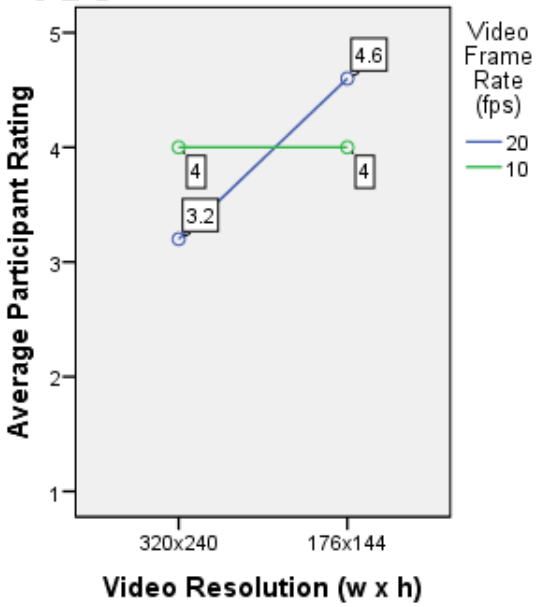

Figure 5-18: The qualitative results for Question 13 "*It was easy to understand what was said in this video.*" for each of the three frame rates and two resolutions. With a significance level of 0.502 (p = .502) there was no statistically significant difference in the average participant rating for each of the video clips.

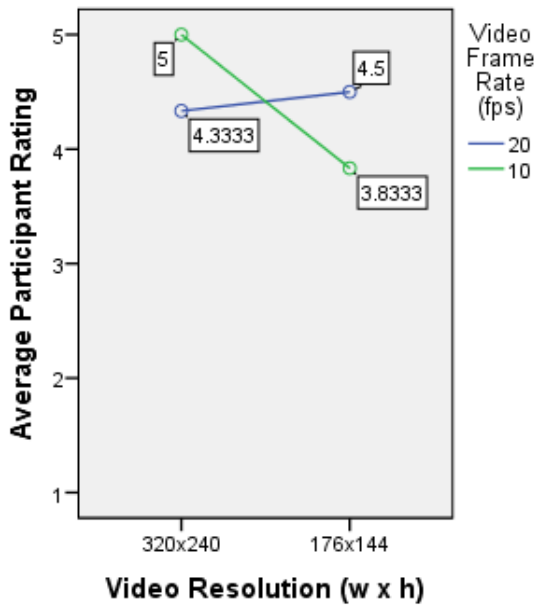

Figure 5-19: The qualitative results for Question 2
"*I am sure of my answer to Question 1.*" for each of the three frame rates and two resolutions. With a significance level of 0.227 (p = .227) there was no statistically significant difference in the average participant rating for each of the video clips.

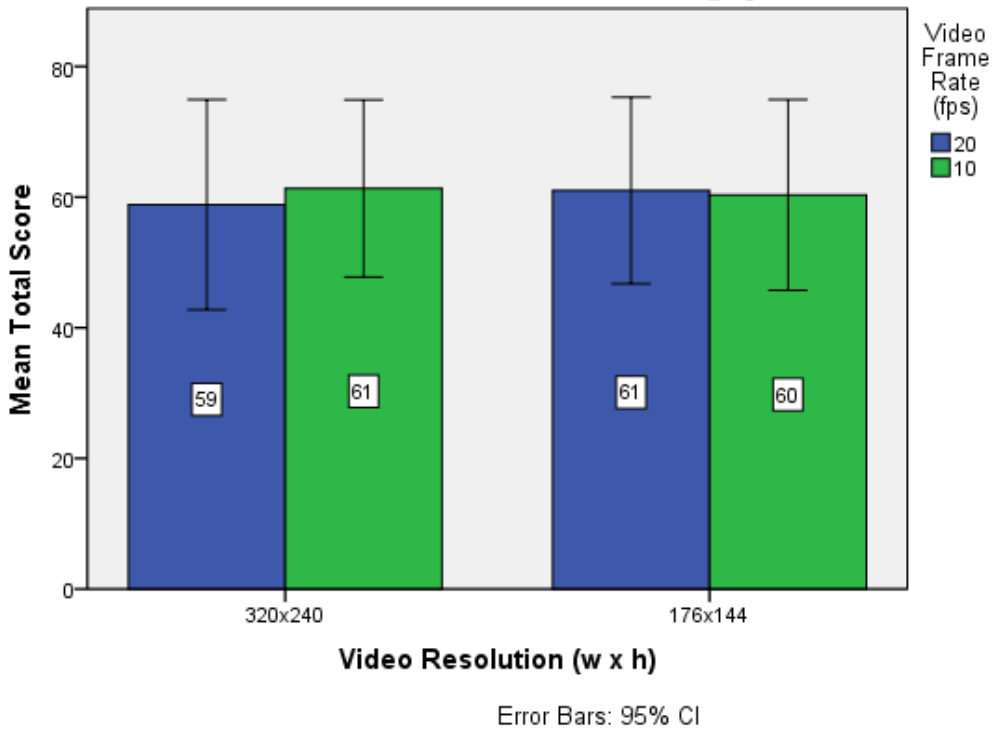

Figure 5-20: Estimated marginal means across all questions.
The estimated marginal means of the participant responses across all questions for each of the two frame rates and two resolutions. The y-axis is mean total score. On the x-axis is a particular video resolution, with each colour representing a particular video frame rates. With a significance level of 0.990 (p = .990) there was no statistically significant difference in the average participant rating for each of the video clips.

No clear preference by the participants was found for any particular combination of frame rate and resolution, as can be seen from the estimated marginal means of the participant responses

across all questions for each of the two frame rates and two resolutions, and is confirmed by the ANOVA analysis.

Despite getting feedback about possible Sign Language dialect problems and attempting to only use signed phrases without dialect problems, one of the four phrases ("*He is a short man*") still was not caught before the experiment.

Where the two pilot studies were done with the help of DCCT staff members, all fully literate, the final experiment only included six DCCT staff members, with the remainder of varying literacy level. This could have impacted on the quality of the captured reponses.

No specific questions pertaining to the video intelligibility and quality gave any stand out problems. Of the four sign language video clips the only one that had problems contents wise was the "*He is a short man*" clip, and after pointing out the problem sign, participants continued to evaluate the video clip as per the questionnaire.

All lessons learnt through the pilot studies were applied, yet there were no frame rate and resolution combination that were clearly judged inadequate or below par.

University of Cape Town

6 Conclusion

6.1 Conclusion

This dissertation studied the effects of reducing the frame rate and resolution of SASL video played back and viewed on a lower-end, more affordable cell phone. The subjective intelligibility of Sign Language videos at the different frame rates and resolutions were evaluated through user studies with members of the South African Deaf community.

When this research started, looking at similar research it seemed a relatively simple question: *What is the lowest video resolution and frame rate that would provide intelligible South African Sign Language video on a cell phone?* But after two pilot studies and a final intelligibility study, what has become abundantly clear was that measuring the intelligibility of Sign Language video is a multifaceted problem, providing many obstacles, amongst others the difficulty with written language by the participants, making the use of written questionnaires problematic. In addition to this is the fact that SASL has different dialects, meaning a simple sign for one Deaf person could be an unknown sign to another. Each of these facets adds a layer of possible miscommunication and misunderstanding between the researcher and the Deaf participant that could impact on the evaluation of intelligibility.

Based on the results there does not seem to be a preferred frame rate or a clear drop in subjective intelligibility at low frame rate or low resolutions. The frame rate can be reduced to 10 frames per second, with the resolution reduced to 176 x 144 pixels while still providing intelligible SASL reproduction when viewed on a handheld cell phone, and being acceptable and comfortable for day to day use.

The final conclusion of this study is that, using long questionnaires and simple signed phrases, there is no clearly discernible difference between Deaf participants' opinion of the intelligibility of sign language video clips based purely on frame rate and resolution.

6.2 Limitations

Further work is needed in the subjective evaluation of the intelligibility of Sign Language video on a cell phone. While Nakazono et al. [19] used sign language video sequences of similar length as this research, about 7 to 8 seconds; other subjective assessments used longer video sequences, such as Cavender et al. [4] using clips with durations from 0:58 to 2:57 minutes and Ciaramello et al. [7] with video sequences ranging from 7.2 seconds to 150.9 seconds.

Short signed phrases might be good for ease of testing, but from the results of this research does not seem to be appropriate for the evaluation of intelligibility of Sign Language video sequences. A normal conversation consists of multiple longer sentences and provides the opportunity to clear up a missed or unknown word or sign by continuing to listen and possibly collect the missing information from its context. If the listener were to fail in this, it is simple to ask the speaker to repeat or explain the missed or unfamiliar word or sign. This conversational context is missing in the short phrases used in this research. A better approach might be to use a longer video clip showing a signer signing a short story about what happened that morning on the way to work, a part of life familiar to most people.

Looking at the questionnaires used in the related research they consisted of similar but fewer and simpler questions. The subjective questionnaire used by Ciaramello et al. [7] consisted of 12 videos in total with a four-question, multiple-choice survey focussing how difficult it was to understand what was said in the video. In the case of Cherniavsky et al. [5] the participants used the phone to hold a sign language conversation at different video settings with a five question subjective measurement after a five minute conversation. A five level scale was used to evaluate the video intelligibility.

University of Cape Town

In this research paper a similar questioning technique was followed with the participants being asked to subjectively evaluate sign language videos of differing video quality on a five level scale (except in experiment two where a binary answering technique was used). But in hindsight the questionnaire became too in-depth too early. Increasing the number of questions and the complexity of the questions increased the burden on the Deaf participants, introducing the additional complication of user fatigue, especially on the last experiment where English literacy was even less prevalent.

Looking at participant numbers in the related research these varied from 11 participants used by Ciaramello et al. [7] to 15 participants recruited by Cherniavsky et al. [5], compared to five, six and twenty four participants respectively in the three experiments in this research paper.

In the findings of both Ciaramello et al. [7] and Nakazono et al. [19] a clear progression was found from low intelligibility at low resolutions and frame rates to increased intelligibility scores at the higher quality video clips. This was not clearly evident in the results of this study. A clearer comparison between the results of this study and the related studies would have been possible if similar questions (in number and kind) were used with a larger number of Deaf users.

The literacy level of the participants and the grammatical differences between Sign Language and written language has an impact on the use of questionnaires in this research. The use of fewer, simpler questions could ease the execution of the experiments as well as improve the usefulness of the results.

To improve the results and attain statistical significance the sample size needs to be increased, the questionnaire shortened and simplified, and the experiments should make use of longer video clips. All related work had consistently longer video clips, more closely simulating a conversational use of the phone. By shortening and simplifying the questionnaire participant fatigue would be decreased and a more honest evaluation of the intelligibility would be captured.

6.3 Future work

Open questions are: Are simple, short signed phrases evaluated through a written questionnaire a valid test of intelligibility and Sign Language communications over video? What is the impact of Sign Language dialects and sign execution by the signer on the evaluation of the clip? Can the impact of differences in Sign Language usage be negated or quantified during intelligibility evaluations?

These results indicate that we do not need to use the latest smart phone with high resolution video capabilities to provide the Deaf community the opportunity to converse in their first language wherever and whenever they want, bringing us closer to providing Deaf people affordable, full access to the mobile telecommunications network.

This research looked at the frame rate and resolution requirements using Sign Language video clips pre-recorded in controlled conditions, with good lighting and an even background. Further research is needed to confirm that the findings of this research hold up under non-ideal conditions, especially when the video to be viewed is recorded using the built-in camera of the cell phone. In addition there is the question of how much bandwidth, and thus cost, is truly saved by the drop in frame rate and resolution. Taking into consideration bandwidth cost as well as phone cost do these findings lead to widely affordable mobile video communication for the South African Deaf community?

Another avenue for future research would be to remove the written questionnaire and conduct the experiment completely in SASL instead of text. Would minimising the impact of participant literacy level and keeping the whole experimental setup as natural as possible for the participants improve the significance of the findings?

University of Cape Town

Bibliography

1. Ascher, S., Pincus, E. *The Filmmaker's Handbook – A comprehensive guide for the digital age.* Third Edition (2007), Plume, New York, USA.

2. *AV Foundation Programming Guide.* 2010. [online]. [Accessed 2 December 2010]. Available from World Wide Web: <https://developer.apple.com/library/ios/#documentation/AudioVideo/Conceptual/AVFoundationPG/Articles/03_MediaCapture.html%23//apple_ref/doc/uid/TP40010188-CH5-SW2>

3. BBC NEWS | Technology | *Deaf people lobby MPs over phones.* 2008. [online]. [Accessed 29 March 2009]. Available from World Wide Web: <http://news.bbc.co.uk/1/hi/technology/7670175.stm>

4. Cavender, A., Ladner, R., Riskin, E. 2006. *MobileASL: Intelligibility of Sign Language Video as Constrained by Mobile Phone Technology*, ASSETS 2006: 8th int SIGACCESS Conf. on Computers & accessibility, 71-78.

5. Cherniavsky, N., Chon, J., Wobbrock, J., Ladner, R. and Riskin, E. 2009. *Activity Analysis Enabling Real-Time Video Communication on Mobile Phones for Deaf Users.* Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '09), Victoria, British Columbia.

6. Ciaramello, F.M., and Hemami, S.S. 2007. *'Can you see me now?' An Objective Metric for Predicting Intelligibility of Compressed American Sign Language Video.* Proc. Human Vision and Electronic Imaging (HVEI) 2007, San Jose, CA, January 2007.

7. Ciaramello F.M., Cavender, A., Hemami, S.S., Riskin, E.A., and Ladner, R.E. *Predicting Intelligibility of Compressed American Sign Language Video With Objective Quality Metrics* Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, January 2006.

8. Glaser, M. 2000. *A field trial and evaluation of Telkom's Teldem terminal in a Deaf community in the Western Cape.* Proc. South African Telecommunication Networks and Applications Conference, (SATNAC 2000), Stellenbosch, South Africa, (CD-ROM publication).

9. Glaser, M. and Tucker, W.D. 2004. *Telecommunications bridging between Deaf and hearing users in South Africa.* Proc. Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI 2004), Granada, Spain. (CD-ROM publication).

10. Hanzo, L. L., Cherriman, P. J., Streit, J. *Video Compression and Communications: From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers*. Second Edition (2007), Wiley-IEEE Press, England.

11. Harrington, R., Weiser, M. *Professional Web Video: Plan, Produce, Distribute, Promote, and Monetize Quality Video*. First Edition (2010), Focal Press, Burlington, USA.

12. Jemni, M., Ghoul, O.E., Yahia, N.B., Boulares, M. 2007. *Sign language MMS to make cell phones accessible to the deaf and hard-of-hearing community.* Conference and Workshop on Assistive Technology for People with Vision and Hearing Impairments (CVHI '07). Granada, Spain. August 2007.

University of Cape Town

13. Kaneko, H., Hamaguchi, N., Doke, M., Inoue, S. 2010. *Sign language animation using TVML*, VRCAI '10 Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry, 289-292.

14. Ma, Z.Y. and Tucker, W.D. 2007. *Asynchronous video telephony for the Deaf.* Proc. South African Telecommunications Networks and Applications Conference, (SATNAC 2007), Mauritius, 134-139.

15. Ma, Z.Y. and Tucker, W.D. 2008. *Adapting x264 for asynchronous video telephony for the Deaf.* Proc. South African Telecommunications Networks and Applications Conference, (SATNAC 2008), Wild Coast Sun.

16. *MediaRecorder | Android Developers.* 2011. [online]. [Accessed 4 September 2011]. Available from World Wide Web: <http://developer.android.com/reference/android/media/MediaRecorder.html>

17. *MediaRecorder | Android Developers.* 2010. [online]. [Accessed 2 December 2010]. Available from World Wide Web: <http://developer.android.com/reference/android/media/MediaRecorder.html#setVideoEncoder(int)>

18. Muir, L.J., Richardson, I.E.G. 2005. *Perception of Sign Language and Its Application to Visual Communications for Deaf People.* The Journal of Deaf Studies and Deaf Education 2005 10(4):390-401.

19. Nakazono, K., Nagashima, Y, Ichikawa, A. 2006. *Digital Encoding Applied to Sign Language Video.* IEICE-Transactions on Info and Systems, Volume E89-D, No. 6 June 2006: 1893-1900.

20. *National Institute for the Deaf: FAQs.* 2009. [online]. [Accessed 29 March 2009]. Available from World Wide Web: <http://www.deafnet.co.za/institute/1faq.html>

21. *National Institute for the Deaf: Myths and misconceptions that hearing people have about the Deaf.* 2009. [online]. [Accessed 29 March 2009]. Available from World Wide Web: <http://www.deafnet.co.za/institute/1myths.html>

22. Nokia Corporation. 2011. *Guide for Qt Multimedia Developers*. [online]. [Accessed 29 October 2011]. Available from the World Wide Web: <http://www.developer.nokia.com/info/sw.nokia.com/id/4abf12e7-72d8-45ef-b1a2-46184abe18ba/Guide_for_Qt_Multimedia_Developers.html>

23. *Nokia Europe - Technical specifications - Nokia N96 support.* 2011. [online]. [Accessed 5 November 2011]. Available from World Wide Web: <http://europe.nokia.com/support/product-support/n96/specifications>

24. *Nokia N96 Designed for Video and TV entertainment.* 2010. [online]. [Accessed 5 November 2011]. Available from World Wide Web: <http://www.businessele.com/product/pro936.html>

25. *Nokia N96 - Full phone specifications* 2011. [online]. [Accessed 12 November 2011]. Available from World Wide Web: <http://www.gsmarena.com/nokia_n96-2253.php>

26. Parekh, R. *Principles of Multimedia*. First Edition (2008), Tata McGraw-Hill, New Delhi, India.

University of Cape Town

27. Parton, B.S. 2005. *Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence.* The Journal of Deaf Studies and Deaf Education 2006 11(1):94-101.

28. Paynton, C. *Digital Video and HDTV Algorithms and Interfaces.* First Edition (2003), Morgan Kaufmann Publishers, San Francisco, USA.

29. Penton, J., Tucker, W.D. and Glaser, M. 2002. *Telgo323: An H.323 bridge for Deaf telephony*. *Proc. South African Telecommunications Networks & Applications Conference,* (*SATNAC 2002*), Drakensberg, South Africa, 309-313.

30. *Qt — Qt - A cross-platform application and UI framework.* 2011. [online]. [Accessed 4 September 2011]. Available from World Wide Web: <http://qt.nokia.com/products>

31. *Qt for Nokia — Qt - A cross-platform application and UI framework.* 2011. [online]. [Accessed 4 September 2011]. Available from World Wide Web: <http://qt.nokia.com/products/qt-for-mobile-platform>

32. *Qt Mobility 1.1 Multimedia.* 2010. [online]. [Accessed 4 September 2011]. Available from World Wide Web: <http://doc.qt.nokia.com/qtmobility-1.1.0/multimedia.html#camera-support>

33. *Qt Mobility 1.1.0 Released.* 2010. [online]. [Accessed 4 September 2011]. Available from World Wide Web: <http://labs.qt.nokia.com/2010/11/09/qt-mobility-1-1-0-released>

34. *QVideoEncoderSettings Class Reference.* 2010. [online]. [Accessed 2 December 2010]. Available from World Wide Web: <http://doc.qt.nokia.com/qtmobility-1.1.0-beta/qvideoencodersettings.html#setCodec>

35. ITU. Series H Supplement 1: *Application Profile – Sign Language and lip-reading real-time conversation using low-bit rate video communication* (1999). Geneva: International Telecommunications Union.

36. Sperling, M., Landy, M., Cohen, Y., and Pavel, M. 1985. *Intelligible encoding of ASL image sequences at extremely low information rates.* Computer vision, graphics, and image processing, vol. 31, no. 3: 335-391. September 1985.

37. *Telkom SA Limited - TELKOM TARIFF LIST 1 August 2011*. 2011. [online]. [A*ccessed 5 November 2011]*. *Available from World Wide We*b: <http://www.telkom.co.za/general/pricelist/downloads/tarifflist_Aug11.pdf>

38. *Thibologa Sign Language Institute - DEAF and SASL*. 2007. [online]. [Accessed 15 October 2012]. Available from World Wide Web: < http://www.thibologa.co.za/sasl.html>

39. Tiwari, S., Elro, E. *AdvancED Flex 4*. First Edition (2010), APress, USA

40. Tucker, W.D. 2003. Social amelioration of bridged communication delay. *8th European Conference of Computer-supported Cooperative Work,* (*ECSCW 2003 Doctoral Colloquium*), Helsinki, Finland.

41. *Vodafone 858 Smart - Full phone specifications.*2011. [online]. [Accessed 5 November 2011]. Available from World Wide Web: <http://www.gsmarena.com/vodafone_858_smart-3955.php>

University of Cape Town

42. *Vodafone 858 Smart Test e Video on-line Notizie Telefonino.net.* 2011. [online]. [Accessed 5 November 2011]. Available from World Wide Web: <http://www.telefonino.net/Vodafone/Notizie/n27310/test-videoprova-vodafone-858-smart-huawei.html>

43. *What is GPRS (General Packet Radio Service).* 2011. [online]. [Accessed 5 November 2011]. Available from World Wide Web: <http://www.mobile-phones-uk.org.uk/gprs.htm>

Appendix A Experiment 1

A.1 Questionnaire

Introduction

Please help us find the best video compression settings for sending South African Sign Language between cellphones. We are looking for the best settings that would give us the smallest videos, but still is easily and comfortably understandable on a cellphone.

There are no right or wrong answers in this experiment, we are simply looking for your honest opinion.

You will view 18 clips on the cellphone. After viewing a clip, please complete a questionnaire for that clip, and then continue to the next clip. Filling in one questionnaire per video clip.

Thank you for your time and help today.

Consent Form

I, _____, fully understand the project and agree to participate. I understand that all information that I provide will be kept confidential, and that my identity will not be revealed in any publication resulting from the research unless I choose to give permission. Furthermore, all recorded interview media and transcripts will be destroyed after the project is completed. I am also free to withdraw from the project at any time.

I understand that a South African Sign Language interpreter will provide sign language translation. That person is bound by a code of ethics that does not allow him/her to repeat any information that is given during the session. This means that my identity will remain confidential.

For further information, please do not hesitate to contact:

Daniel Erasmus
Department of Computer Science
University of the Cape Town
Email: derasmus@uct.co.za

Name: _____

Signature: _____

Date: _____

University of Cape Town

Video Clip No: ____ :____

1. What was said in this video?

_____

_____

_____

2. How sure are you of your answer to Question 1 above?

completely sure	sure	so-so	not sure	not sure at all

3. How easy or how difficult was it to understand what was said in this video?

very difficult	difficult	average	easy	very easy

4. How easy or how difficult was it to follow the facial expressions in this video?

very difficult	difficult	average	easy	very easy

5. How easy or how difficult was it to follow the hand gestures in this video?

very difficult	difficult	average	easy	very easy

6. If you could chat using a cell phone with video this easy/difficult to understand, would you use it?

definitely yes	yes	maybe	no	definitely no

7. Any other comments on this video?

_____

_____

_____

University of Cape Town

A.2 Experiment 1 Questionnaire captures

A	Clip #	What was said in this video?	2	3	4	5	6	Comments
10	1	Pass me this cup, please	4	4	4	4	4	
1	2	Father waiting, why taxi come	5	5	5	5	4	This video was very clear to see
11	3	I played a whole day then go to bath	4	4	4	4	4	
9	4	I do not understand clear, what she said	1	3	3	3	4	
4	5	last day, I was tell	4	4	4	4	4	
12	6	Next month I will buy a new cloth	4	4	4	4	4	
8	7	The fork is on the left of the plate and the knife is on the right of the plate	4	4	4	4	4	
13	8	The boy cleaned a cardboard	4	4	4	4	4	
2	9	Small bread is on the plate	4	4	4	4	5	
3	10	Eat pap making my stomach cool	4	4	3	4	4	link the 4 question due to poor picture but not much
17	11	I wear a pant, because I am going to church	4	4	4	4	4	
18	12	I feed a mealies to the hens	3	4	4	4	4	

Table A-1: Experiment 1 – Captured questionnaire A

University of

B	Clip #	What was said in this video?	2	3	4	5	6	Comments
7	1	Please can I get a cup	5	5	5	5	4	That was easy and the way she do, understandly
1	2	(Father is waiting for taxi) Father waiting and waiting, why? waiting for the taxi	4	4	5	4	4	I think it proving well and good understanding on the video when you chat
18	3	all-day I play and enjoying	2	4	4	4	3	
3	4	wash your teeth before you sleep	5	5	4	5	4	
17	5	Yesterday I walk and fall off	5	4	5	5	4	
9	6	This month I pay my clase	2	3	4	2	3	
5	7	fack and nifis	2	1	2	2	5	
16	8	Bay wash....? (boad)	2	2	4	2	3	
6	9	Miss out	1	1	1	1	4	poor expression and hand gestures
14	10	Pap is to eat and get full	5	5	4	5	4	
13	11	bart is for church	4	4	5	5	4	
4	12	Giving food to all the chicken	4	1	2	1	3	The facial expression need more expression

Table A-2: Experiment 1 – Captured questionnaire B

C	Clip #	What was said in this video?	2	3	4	5	6	Comments
4	1	Please	1	2	2	2	3	sign difficult
3	2	Father wait for the taxi	4	5	4	4	4	Easy slow sign language
15	3	All day play and nigh wash	2	3	2	2	3	yes I though so more clear
6	4	Please help	1	3	3	3	3	Not clear
10	5	Yesterday I walk and fall	2	3	3	3	3	Face expression
8	6	Month pay new cloths	1	3	2	2	2	not clear
18	7	knife and fork there	2	3	2	3	3	normal
5	8	Man wash a window clean	3	4	4	3	4	Normal clear
17	9	plate small have bread	4	4	4	4	3	Easy show hand clear
16	10	Pamp eat full	2	2	1	3	3	short
2	11	Why pants use go to church	4	4	4	4	4	Easy body clear picture
11	12	Throw weed maize hen eat the maize	2	3	4	4	3	Easy clear but not clear not word

Table A-3: Experiment 1 - Captured questionnaire C

D	Clip #	What was said in this video?	2	3	4	5	6	Comments
16	1		2	3	3	3	2	
15	2		2	3	2	3	2	
4	3		3	3	3	4	3	
3	4		4	2	3	4	3	
1	5		2	3	3	3	2	
9	6		4	3	3	4	4	
17	7		2	3	3	3	2	
18	8		2	3	4	4	2	
2	9		3	4	5	3	3	
12	10		3	3	4	3	2	
5	11		3	4	2	3	2	
14	12		3	3	3	3	2	

Table A-4: Experiment 1 - Captured questionnaire D

E	Clip #	What was said in this video?	2	3	4	5	6	Comments
9	1	I want to cap, Please!	4	4	4	4	4	That's fine.
1	2		5	5	4	5	4	That's fine. It should have no problem.
11	3	all day, play then so I have go to bath nite	3	4	4	4	3	difficult about this sign language.
3	4	Sleep, before I go.	4	4	4	4	4	Maybe, it is right
7	5	Yesterday, I was fell.	4	4	4	4	4	I understand it is good
15	6	This is month, pay for clothing.	4	4	4	4	4	fine.
17	7		3	3	4	3	3	I was not think so, she was said that.
14	8	the boy is clean wash for window	2	3	3	4	3	No right.
10	9		3	4	4	3	3	I don't understand she said.
5	10		3	4	4	3	3	No right, about talk
16	11	Why I go to church.	3	4	4	4	3	not at all, fine.
18	12		4	4	4	4	3	Should have no Problem, she is not good. Because she must clear sign language.

Table A-5: Experiment 1 - Captured questionnaire E

	Clip Details							
Video No	**Format**	**Resolution (w x h)**	**Colours**	**Frames per second**	**Data Rate [kbit/s]**	**File Size [KB]**	**Duration [s]**	**Signed phrase**
1	MPEG-4	320 x 240	Millions	30	247.93	204593	6.54	Could you please fetch me that cup over there.
2	MPEG-4	320 x 240	Millions	15	251.48	310650	9.82	Father stands and waits for the taxi.
3	MPEG-4	320 x 240	Millions	10	260.44	317907	9.72	After you've played all day, you bath at night.
4	MPEG-4	176 x 144	Millions	30	243.30	216341	7.04	Before you go to sleep, brush your teeth.
5	MPEG-4	176 x 144	Millions	15	243.87	213802	6.96	Yesterday I tripped and fell.
6	MPEG-4	176 x 144	Millions	10	238.54	258605	8.63	Next month I will buy new clothes.
7	MPEG-4	320 x 240	Millions	30	247.80	250840	8.01	You put the fork on the left and the knife on the right.
8	MPEG-4	320 x 240	Millions	15	269.10	267771	7.91	The boy washed the window. Now it is clean.
9	MPEG-4	320 x 240	Millions	10	263.77	238840	7.20	On the plate was a small loaf of bread.
10	MPEG-4	176 x 144	Millions	30	249.48	260345	8.27	When you eat pap your tummy feels good.
11	MPEG-4	176 x 144	Millions	15	234.94	198422	6.70	Put on your trousers because we are going to church.
12	MPEG-4	176 x 144	Millions	10	238.57	247585	8.26	I scatter the seeds and the chickens eat them.

Table A-6: Experiment 1 - Video clip details

University of Cape Town

Appendix B Experiment 2

B.1 Questionnaire

Introduction

Please help us find the best video compression settings for sending South African Sign Language between cell phones. We are looking for the best settings that would give us the smallest videos, but still is easily and comfortably understandable on a cell phone.

There are no right or wrong answers in this experiment; we are simply looking for your honest opinion.

You will view 9 clips on the cell phone. After viewing a clip, please complete a questionnaire for that clip, and then continue to the next clip, filling in one questionnaire per video clip.

Please make sure to answer all questions, and make sure to provide comments.

Thank you for your time and help today.

Consent Form

I, _____, fully understand the project and agree to participate. I understand that all information that I provide will be kept confidential, and that my identity will not be revealed in any publication resulting from the research unless I choose to give permission. Furthermore, all recorded interview media and transcripts will be destroyed after the project is completed. I am also free to withdraw from the project at any time.

I understand that a South African Sign Language interpreter will provide sign language translation. That person is bound by a code of ethics that does not allow him/her to repeat any information that is given during the session. This means that my identity will remain confidential.

For further information, please do not hesitate to contact:
Daniel Erasmus
Department of Computer Science
University of the Cape Town
Email: derasmus@uct.co.za

Name: _____

Signature: _____

Date: _____

University of Cape Town

Video Clip No: ____ :____

1. What was said in this video?

_____

_____

2. How sure are you of your answer to Question 1 above?

completely sure	sure	so-so	not sure	not sure at all

3. How easy or how difficult was it to understand what was said in this video?

very difficult	difficult	average	easy	very easy

4. Please select the appropriate choice from the options provided below:

4.1 ☐ **The** movement **was** clear**.** **or** ☐ **The** movement **was** blurry**.**

4.2 ☐ **I could** clearly **see all the details of the** face**.** **or** ☐ **I had** difficulty **seeing the details of the** face

4.3 ☐ **I could** clearly **see the** hands**.** **or** ☐ **I had** difficulty **seeing the** hands**.**

4.4 ☐ **The** video **was the** right speed**.** **or** ☐ **The** video **was** too slow / too fast**.**

4.5 ☐ **I** knew all **the** signs**.** **or** ☐ Some signs **were** unknown **to me.**

5. How many times did you view this clip?

_____

6. Any other comments on this video?

_____

_____

_____

University of Cape Town

B.2 Experiment 2 Questionnaire captures

A	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
7	1	The gril ride horse	5	5	5	5	5	5	5	1	
2	2	Men he ball	3	3	5	5	5	5	5	4	
8	3	child is dirty	5	5	5	5	5	5	5	1	
4	4	More birthday	5	5	5	5	5	5	5	1	
1	5	I very happy my uncle vitsit	3	3	5	5	5	5	5	4	Stop smile
6	6		3	3		1	1	5	5	5	

Table B-1: Experiment 2 - Captured questionnaire A

B	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
1	1	Me Happy why uncle visit	3	3	5	5	5	5	5	2	
5	2	Man ball bump	3	3	5	5	5	5	1	2	Bit confuse
3	3	Boy small body dirty	2	3	5	5	5	5	1	3	Boy - different sign language
8	4	Tomorrow birthday mine	3	3	5	5	5	5	5	2	OK
6	5	yesterday fish big me hook	2	3	5	5	5	5	1	3	Sign language bit confuse
7	6	short your small top	2	3	5	5	5	5	1	4	Confuse

Table B-2: Experiment 2 - Captured questionnaire B

C	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
1	1	She said the girl ride the horse	5	5	5	5	5	5	5	1	Fine nothing problem
4	2	She said man play ball on he's head	4	4	5	5	5	5	5	1	The video very clear
9	3	She said that little boy, he was very dirty	5	5	5	5	5	5	5	1	Problem with sign boy
2	4	She said tomorrow will her's birthday	5	5	5	5	5	5	5	1	The video is very clear
7	5	She said yesterday saw fish very big the take it	5	5	5	5	5	5	5	1	Very clear
8	6	She said her top is very tight and short	5	5	5	5	5	5	5	1	Very clear

Table B-3: Experiment 2 - Captured questionnaire C

D	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
3	1	The girl is ride on the horse	5	5	5	5	5	5	5	1	Absolutely!
9	2	The man is playing with the ball on his forehead	5	5	5	5	5	5	5	2	The view of sign was clear
2	3	The boy's outfit is dirty	5	3	1	5	5	5	5	1	The movement was not quality
1	4	My birthday is tomorrow	5	4	5	1	5	5	5	2	Screen problem and unclear view
8	5	last day, the fish was big so I caught them	5	4	5	5	5	5	5	1	The quality of the screen was little poor
6	6	Your short is small and tight	5	4	1	5	5	5	5	3	Screen of the view was poor becoz the movement was little poor

Table B-4: Experiment 2 - Captured questionnaire D

E	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
9	1		2	1	5	5	1	1	1		
6	2	(Tomorrow port is my) Man bob with ball on the hard	4	4	5	5	5	5	5	2	
2	3	Not	1	1	1	5	1	5	1	2	
3	4	Child goes to play	5	5	5	5	5	5	5	1	
1	5	Me happy why uncel will visit	4	4	5	5	5	5	5	1	
5	6	Your T-shirt is small	4	4	5	5	5	5	5	2	

Table B-5: Experiment 2 - Captured questionnaire E

F	Clip #	What was said in this video?	2	3	4.1	4.2	4.3	4.4	4.5	5	Comments
9	1		1	1		1	1	1		3	no comments!!!
6	2	The bal	1	1	1	1	1	1	1	2	yes, I'm not understand that what said
2	3	The boy is mess	2	2		1			1	1	..no not yet. It is not right hand for sign
3	4	Tomorro cake!	4	3	5	5	5	5	5	1	I understand, no comment
1	5	Fish is big	3	3	5	5	5	5	5	1	it is ok, no comments
5	6	T-shirt - sort - It is cold	2	1	1	1	1	5	5	2	No, it was difficult about sign language

Table B-6: Experiment 2 - Captured questionnaire F

	Clip Details							
Video No	**Format**	**Resolution (w x h)**	**Colours**	**Frames per second**	**Data Rate [kbit/s]**	**File Size [KB]**	**Duration [s]**	**Signed phrase**
1	MPEG-4	320 x 240	Millions	30	2663.28	3098	9.52	The girl rides the horse.
2	MPEG-4	320 x 240	Millions	15	1979.15	1654	6.84	The man bounces the ball on his head.
3	MPEG-4	320 x 240	Millions	10	1455.42	1281	7.20	The small boy is dirty all over.
4	MPEG-4	176 x 144	Millions	30	1250.55	869	5.68	Tomorrow is my birthday.
5	MPEG-4	176 x 144	Millions	15	791.42	717	7.40	Yesterday I caught a big fish.
6	MPEG-4	176 x 144	Millions	10	567.82	570	8.12	Your T-shirt is too small for you.

Table B-7: Experiment 2 – Video clip details

University of Cape Town

Appendix C Experiment 3

C.1 Questionnaire

Introduction

Please help us find the best video compression settings for sending South African Sign Language between cell phones. We are looking for the best settings that would give us the smallest videos, but still is easily and comfortably understandable on a cell phone.

There are no right or wrong answers in this experiment; we are simply looking for your honest opinion.

You will be viewing one video clip on the cell phone. Watch the video clip only once, and then complete the questionnaire. Please make sure to answer all questions.

Thank you for your time and help today.

University of Cape Town

Consent Form

I, _____, fully understand the project and agree to participate. I understand that all information that I provide will be kept confidential, and that my identity will not be revealed in any publication resulting from the research unless I choose to give permission. Furthermore, all recorded interview media and transcripts will be destroyed after the project is completed. I am also free to withdraw from the project at any time.

I understand that a South African Sign Language interpreter will provide sign language translation. That person is bound by a code of ethics that does not allow him/her to repeat any information that is given during the session. This means that my identity will remain confidential.

For further information, please do not hesitate to contact:
Daniel Erasmus
Department of Computer Science
University of the Cape Town
Email: derasmus@uct.co.za

Name: _____

Signature: _____

Date: _____

General Information

Gender: ☐ Male ☐ Female

Age: _____ years

Preferred reading and writing language:
☐ English
☐ Afrikaans
☐ Xhosa
☐ Other
If other, please specify:

_____

Number of years using South African Sign Language: _____ years

University of Cape Town

1. What was said in this video?

_____

_____

_____

For the following questions please circle the appropriate response.

		strongly disagree	**disagree**	**neither agree nor disagree**	**agree**	**strongly agree**
2.	I am sure of my answer to Question 1.	1	2	3	4	5
3.	It was difficult to follow the hand gestures in this video.	1	2	3	4	5
4.	The movement was blurry.	1	2	3	4	5
5.	I had no problems seeing the facial expressions in this video.	1	2	3	4	5
6.	The video was the right speed.	1	2	3	4	5
7.	The movement was clear.	1	2	3	4	5
8.	I could clearly see all the hand gestures in this video.	1	2	3	4	5
9.	It was difficult to follow the facial expressions in this video.	1	2	3	4	5
10.	I had difficulty seeing the details of the face.	1	2	3	4	5
11.	I had difficulty to understand what was said in this video.	1	2	3	4	5
12.	I could clearly see the details of the face.	1	2	3	4	5
13.	It was easy to understand what was said in this video.	1	2	3	4	5
14.	I could clearly see the hands.	1	2	3	4	5
15.	The video was too slow.	1	2	3	4	5
16.	I knew all the signs used in this video.	1	2	3	4	5
17.	Some signs used in this video were unknown to me.	1	2	3	4	5
18.	It was difficult to see the hands.	1	2	3	4	5
19.	The video was too fast.	1	2	3	4	5

University of Cape Town

C.2 Experiment 3 Questionnaire captures

#	Clip #	Gender	Age	Language of literacy	SASL experience [years]	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	male	47	English	47	4	5	4	3	4	4	2	4	3	2	2	4	4	999	5	2	4	4
2	1	male	21	Xhosa	21	5	5	999	3	5	5	4	4	5	3	4	3	5	5	5	4	3	5
3	1	male	35	English	35	4	4	3	2	3	4	3	3	5	3	4	999	4	3	5	2	4	4
4	1	male	27	English	27	1	1	3	2	888	999	1	3	5	5	2	999	2	999	2	5	5	3
5	1	female	29	English	29	3	5	1	4	2	3	1	5	4	2	3	1	4	2	5	3	2	5
6	1	female	42	English & Xhosa	42	999	999	999	2	3	999	999	999	999	999	999	2	5	999	999	999	3	999
7	4	female	30	English	30	1	5	3	2	5	2	5	1	5	2	3	1	5	5	1	1	5	5
8	2	female	34	English	34	5	1	1	4	5	5	5	4	5	1	1	999	5	1	5	1	5	1
9	2	male	47	English	47	5	999	3	5	5	5	5	3	4	3	5	999	5	5	5	3	3	3
10	2	female	64	English	59	5	999	5	5	2	2	2	5	5	2	5	2	5	2	2	4	5	4
11	2	male	37	English	37	5	2	4	5	1	3	1	3	4	1	999	999	1	1	1	2	4	3
12	2	female	31	English	31	2	999	4	888	4	999	999	999	888	999	999	4	999	999	4	999	2	999
13	3	male	23	Afrikaans	23	5	3	5	999	2	5	5	2	5	2	5	999	5	5	4	2	3	5
14	3	female	39	Xhosa	29	999	5	999	4	999	999	999	999	4	999	3	999	999	999	5	3	5	5
15	3	female	35	English & Xhosa	35	4	999	999	2	5	5	4	3	5	4	999	4	5	5	4	2	4	2
16	3	male	53	English	12	3	4	1	2	5	5	4	3	4	5	3	5	4	2	5	5	4	5
17	3	male	37	English & Afrikaans	37	5	2	2	3	2	5	5	5	2	2	4	4	4	3	4	3	2	2
18	3	female	60	English	60	5	4	3	5	4	5	999	5	4	5	5	5	999	5	3	2	1	4
19	3	female	36	English & Xhosa	32	5	1	1	5	5	5	5	1	1	1	5	5	5	5	5	1	1	1
20	4	male	20	English	20	5	4	5	999	4	5	4	3	3	5	4	5	5	4	888	3	5	999
21	4	female	55	English	55	5	3	5	5	2	5	5	5	5	3	4	999	4	5	5	1	1	5
22	4	female	26	English	10	4	3	4	2	5	5	5	2	2	2	4	5	5	4	5	1	1	3
23	4	female	31	English & Xhosa	31	1	888	999	1	888	888	999	999	888	999	1	888	888	2	999	999	888	2
24	4	female	35	English	26	4	2	1	4	4	5	5	2	2	2	5	5	5	2	5	1	1	1
25	4	male	35	English & Afrikaans	28	4	4	5	5	4	4	4	3	3	4	4	4	4	4	5	4	4	4
26	1	female	37	Afrikaans	19	5	4	5	5	5	4	5	4	4	3	4	3	4	3	3	4	5	5
27	2	male	35	English & Afrikaans	25	5	4	1	4	4	5	4	2	4	5	5	5	5	4	4	3	4	2
28	1	male	32	English	25	5	1	1	5	5	5	5	1	1	1	5	5	5	1	5	1	1	1
29	2	female	34	English	24	5	5	4	5	4	5	5	1	1	1	5	5	5	2	5	2	1	2

Table C-1: Experiment 3 - Captured questionnaires.
A value of 888 signifies the participant marks more than one response, while a value of 999 signifies the participant marked none of the possible responses. The five greyed out questionnaires (questionnaire number 4, 6, 12, 14 and 23) was dropped from the analysis because of too many unusable responses.

University of Cape Town

	Clip Details							
Video No	**Format**	**Resolution (w x h)**	**Colours**	**Frames per second**	**Data Rate [kbit/s]**	**File Size [KB]**	**Duration [s]**	**Signed phrase**
1	MPEG-4	320 x 240	Millions	20	2000.35	1397	5.58	He is a short man.
2	MPEG-4	320 x 240	Millions	10	1384.82	954	5.50	The family is home.
3	MPEG-4	176 x 144	Millions	20	910.28	745	6.51	I read a book.
4	MPEG-4	176 x 144	Millions	10	584.24	459	6.21	I want that apple.

Table C-2: Experiment 3 – Video clip details

University of Cape Town