



Relationship between Intelligibility and Visual Quality in Czech Sign Language as a Supplement to Television Broadcasting

Petr Zatloukal, Zdenek Svachula, Martin Bernas

Dept. of Radioelectronics, Faculty of Electrical Engineering
Czech Technical University in Prague
Technicka 2, 166 27 Prague 6
Czech Republic

Abstract: This article concerns the determination of intelligibility in relation to the distortion of sign language interpreter. New television technologies are developed to offer more services to the viewers by connecting TV device to the internet. Instead of sign language interpreter as an inseparable part of broadcasted program it can be rather transmitted over an independent channel. We performed a set of experiments investigating comprehension abilities of deaf respondents in relation to the distortion of important parts of the interpreter's body by using minimal pairs in Czech sign language and finger spelling.

Keywords: Intelligibility estimation, sign language, region of interest, visual quality.

I. Introduction

Czech Sign Language (CSL) provides natural communication among members of the Czech Deaf community. Together with the Czech finger alphabet as manual representation of letters according to standard Czech alphabet, they both provide a means of communication of deaf used in common TV content and communication alike. Based on the survey of the Czech Statistical Office [1] almost 26,000 deaf or hard of hearing live in the Czech Republic. The number of CSL users with CSL as their preferred language can be at about 10,000. These are approximate numbers but state a need for additional services for deaf and hard of hearing which would be hidden for the rest of viewers such as closed captioning.

A. Motivation

In the Czech Republic only few television programs are accompanied with the sign language interpreter. When available, the sign language interpreter is actually an inseparable part of a picture broadcasted over DVB standard. It usually appears as a small bordered frame in the bottom right corner of a TV screen. The visual quality of the resulting picture is determined by parameters of MPEG encoder on the transmitting side. The interpreter is thus visible to all viewers, including the hearing, who are likely to prefer the picture without the additional interpreter. We assume a CSL interpreter picture to be processed separately from the main program, which opens up new challenges in the encoding and intelligibility issues of such video.

When evaluating compression and other features of CSL video, traditional video quality estimators are inadequate. As in [2], visual quality should be measured as the intelligibility of the signer's performance. As described in [3] or in detail in [4], the most important area of signer's body is the face acting as the region of interest (ROI), where observers' point of view is focused most of the time. The second important ROI are hands or upper body in general depending on the application. The rest of scene is definitely unimportant for understanding. Focusing on the face observers fixate on this area with their foveal vision, which is most accurate in detail and has the highest resolution. At the opposite site, the hands are perceived with parafoveal vision with lower resolution in spatial domain. This means that the observer's eyes perceive the hands slightly blurred even if they are actually quite sharp.

We found interesting to see how much distortion of the important ROIs affects the intelligibility of the video content. The goal of this work in progress is to evaluate the impact of ROIs distortion to comprehension and express the relationship between comprehension and visual quality in television content.

B. Contribution

Previous researches include mainly American Sign Language (ASL), which is actually different from the CSL in signs execution (position of a sign in space, a hand shape, etc.). In general, sign languages relate to a particular state or region, where they are used, and they are considered as communication tools independent of spoken languages, as well as from each other [5]. We focus on the intelligibility of video in standard or high video resolution more than low resolution formats used in cellular networks and widely discussed in most of analogous research papers.

Next a novel methodology for subjective sign language tests is proposed that is not based on the entire narratives, so the respondents cannot discover the meaning of the signs they are unable to recognize from the context.

II. Related Work

Related articles [2] and [6] are concerned more with transmission of a video at extremely low bit rates at less than 30 kbps used primarily in cellular networks. GPRS is preferred prior to G3 standards, which are not ubiquitously accessible. Our research deals with an independent transmission of CSL interpreter over separate channel such as internet [7]. Considering general internet connection we can handle much higher bitrates than 30 kbps. Internet channel is limited only by throughput and server capabilities so we can expect available bitrate in hundreds of kbps.

Even the resolution of such picture is expected to be greater than 320 x 240 pixels used by Ciaramello [2]. A typical frame with CSL interpreter on Czech Television channels has a resolution of approximately 250 x 250 pixels in standard-definition (SD) broadcasting. Height is the only important dimension and a rectangular shape of such frame should be assumed. When broadcasted separately, CSL users would probably prefer a possibility to select the size of the frame/interpreter. Preserving the main program to be still visible on the screen we can suppose the highest resolution of two thirds of the screen height that is 384 pixels in SD or analogously up to 720 pixels in high-definition (HD) format. Today's television displays have seldom a lower resolution than 1920 x 1080 pixels and even if the broadcasting is still mainly in SD, it makes sense to think about the future transfer of an interpreter at a higher resolution. Under these conditions, the apparent effect of compression artifacts of block algorithms will vary from lower resolutions, where the size of blocks proportionally to the size of the interpreter's body is more significant.

Cavender [6] propose that common encoding algorithms used in cellular mobile phones are inadequate and cannot provide sufficient quality for intelligibility. Current approaches in compression of a sign language interpreter video exploit the ROIs. Face and hands are compressed with a lower compression ratio than the rest of the scene to achieve better bit rates than generic among all block-based encoding algorithms. Sign language specific algorithms can be placed into two basic classes: foveated video encoding and ROI encoding. Foveated encoding employs descend in spatial resolution of the human visual system away from the observer's point of fixation and encodes each block with considering its distance from the fixation on the face [3]. According to Svachula [4] the signer's face is assumed as the fixation point. ROI encoding algorithms use various quality parameters among detected regions with face, hands and the background [8]. These systems assume that human visual system perceives objects out of foveal vision with lower resolution and they encode these regions with lower quality. Distorting only the face as the most important ROI or the whole upper body while leaving the background unaffected brings a question how will these changes in viewing conditions affect the comprehension of such content.

Cooperation with the Institute of Czech Language and Theory of Communication (ICLTC), Faculty of Arts, Charles University in Prague has brought a CSL specified methodology, different from previous research. This experiment involved coherent narratives in ASL and the participants were presented a questionnaire to evaluate the comprehension of the content and the intelligibility of the video. Such kind of experiment can be influenced by the ability of participants to guess the story meaning according to the context, even though they did not understand all the signs, or rather they did not understand to the signer continuously. SL accompaniment in general TV content should meet specific conditions. SL interpreter's performance must be intelligible not only as the whole, but also when a brief piece of information is translated. Especially in news or magazines, where are live entries or chats, deaf should be able to understand the context as soon as possible. There is not any possibility to request repetition or guess the meaning as enables the conversation via mobile phone.

III. Experiments

Eleven deaf and one hearing user of CSL were recruited from the Faculty of Arts, Charles University in Prague to participate in the intelligibility experiments. All deaf signers considered ASL as their primary language and hand-alphabet. The hearing participant has been studying CSL for at least 2 years of daily using.

Exploiting the methodology outlined in Section 2 participants were examined with an intelligibility test with 3 different experiments. After each experiment participants were asked to demonstrate their comprehension by repeating presented sign in CSL (experiment 1) or by answering some questions in a questionnaire (experiments 2 and 3).

A. Stimuli

Reference videos were recorded with HDV resolution of 1440 x 1080 pixels in anamorphic 16:9 format with a frame rate of 50 fields per second. Video sequences for all experiments were recorded in studio with CSL interpreter in front of blue static background (Fig. 1). The size of the interpreter in stimuli video would match the size of 400 x 400 pixels if displayed in the corner of the TV screen as described in section II.

In our previous research [4] we investigated the ROIs of deaf in detail. The influence of H.264 encoder at bit rates of 100 and 300 kbps was taken into account and the distortion of important ROIs was performed using a 2D median filter with a radius of 15 pixels either applied on the facial area or the whole figure that demonstrate the changing conditions in the perception of an interpreter large enough to be able to influence the perception. Two levels of distortion were included. In the first case, the filter was applied only to the face of the interpreter, in the latter case was applied to the entire figure (Fig. 1). Average objective parameters of distorted pictures among various signs are as follows: distorted face (PSNR = 21.28 dB) and distorted figure (PSNR = 18.12 dB). The face

region was tracked using the Adobe environment. The blue uniform background of the interpreter does not have to be concerned. Any distortion caused by encoding will not be excessively visible.

Figure 1. Examples of original picture (left) and picture with distorted face (middle) or entire figure (right).



In the video sequence for each experiment a randomly selected subset of the reference or distorted test takes were presented as described in the previous subsections. Special care was devoted to make sure that each video sequence is viewed the same number of times and that the order, in which they are presented to every subject, is different. Such randomization of the signs order in the video seeks to reduce the memory effect in the assessment of SL intelligibility.

All instructions for experiments were recorded in CSL with the same conditions as the test video in order to avoid inaccuracies in the actual translation from the Czech language into CSL. The stimuli were displayed on a 20 inches high-resolution monitor at a distance of 80 cm.

B. Method

Experiment 1 - Recognition of a single sign

In the first experiment, we evaluated comprehension using minimal pairs instead of complete stories. This phonological term denote two signs, which differ only in one component. The sign is the smallest phonological unit of SL and cannot be divided into any smaller units [9]. Such minimal pairs are essentially unique for each SL, because of the originality in signs performance.

The level of a visual quality must be as high as the signs from a pair are distinguishable from each other. The components of CSL as handshape, orientation and motion of hands or facial expression are thoroughgoing described in [10]. Subjects were tested with an intelligibility test composed of 6 isolated CSL signs from a set of minimal pairs, each 1–2 sec in length. In the set only the same sign from a single pair was always used. Sign pairs in the first part of the experiment were adopted from [9]. An example of a pair is shown in Fig. 2.

Figure 2. Typical minimal pair in CSL. (a) Jablko (apple). (b) Ostuda (shame).



Participants' performances were videotaped using a Sony HVR-Z1E HDV camera and assessed by experts from ICLTC what sign was articulated without knowing which sign was presented. If the articulated and the presented signs were the same, the answer was set to correct, otherwise as incorrect, no matter if the sign was the second from the pair or not.

Experiment 2 - Recognition of finger-spelled foreign names

Second experiment consisted of a series of 5 foreign names, whose letters were finger-spelled in CSL two-handed alphabet. As mentioned in Section 1, the finger alphabet is not a component of SL, but provides an intermediate communication tool between SL and related spoken language.

In this part a short-term memory (STM) together with a working memory as its extension was examined in a serial STM span task. In related researches, both [12] and [13] used a set of ASL finger-spelled letters to investigate the span of STM. Results in [12] showed the mean of the native signers' span of 4.4 items. Employing the same score, as the length of the longest correct sequence, the mean scores according to [13] were 5.3. Propounding an alternate scoring method with two sequences of the same length terminating the test when the participant fails on both sequences of a particular length gives the mean score of even 8.3. Included names consist of 11 - 14 letters, but unlike mentioned studies are not meaningless. We supposed an increased STM span

presenting and recalling meaningful names as stimuli, even if they are not familiar. In our case the rate of presentation has mean of 2.33 (standard error, sem. = 0.21) items/s.

The names are as follows:

- Šakríl Afrídí (Pakistani)
- Vittorio Grilli (Italian)
- Daniela Hodrová (Czech)
- Naomi Aldort (French)
- Gary Connery (English)

All names were unknown for observers. They were randomly selected by ICLTC from current newspaper articles to include an exotic, previously unheard name for the Czechs (Šakríl Afrídí), a number of names more or less closer to Czech participants (Vittorio Grilli, Naomi Aldort and Gary Connery) and eventually Czech name for the typical Czech suffix for female names *-ová* (Daniela Hodrová).

Experiment 3 - Sentence comprehension

In the last experiment participants were presented short texts in CSL consisted of one complex sentence (up to one minute in duration in CSL). This method corresponds to [2] at most, but still differs. Each respondent evaluated one undistorted and distorted short narrative and after that was asked to answer two simple questions directed to the fingerspelled parts of the utterance pertaining to the exhibition for the anniversary of the assassination of Deputy Reich-Protector of Bohemia and Moravia (today's Czech Republic) Reinhard Heydrich. Questions were directed to the garden (Answer 1: Valdstejnka), where the exhibition took place and the codename for the assassination attempt on Reinhard Heydrich (Answer 2: Anthropoid). Responses could be estimated only from the context hardly implying that the individual fingerspelled letters had to be intelligible.

Character of experiments 2 and 3 is similar. The STM is investigated in this case again involving more working memory. Given visuo-spatial information has to be stored and manipulated according to an ongoing task. The participants are forced to remember and recall that require real time manipulation of linguistic information rather than maintenance of serial order [14].

C. Data Processing

A total number of seven subjects were asked to view a randomly selected subset of the test sequences as described in the previous subsections. Before the experiments, each participant passed a visual acuity test according to [11] and a short training session in order to get familiarized with the goal of the test, the meaning of each question posed, and the definition of each score, by being given illustrative examples. Examples of content used during the training were different in the training and the actual experiments. Each session has taken less than 20 minutes during which test subjects assessed 3 language test subsequences.

In processing of result were used standard statistics operations such as mean, standard deviation and t-tests. The average error was calculated according to the formula

$$Er_{av} = \frac{\sum_{i=1}^N n_i}{n},$$

where n is total number of signs and n_i is a number of incorrect signs of i^{th} subject.

IV. Results

Experiment 1

The total average error signs recognition is 15 % for both distorted face and entire figure according to the above mentioned formula. There were no errors in recognition of the original quality picture.

Table I. Relative number of errors

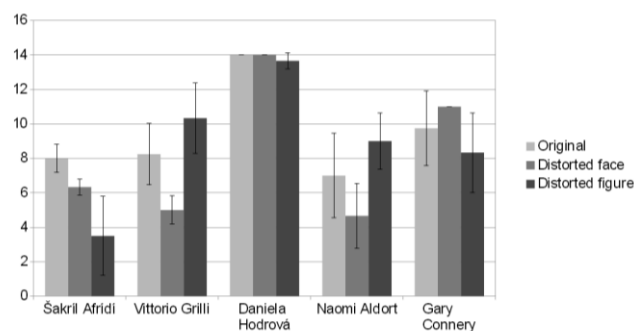
Original image	Distorted face	Distorted figure
0 %	15 %	15 %

In this experiment, a confusion of signs from minimal pairs was expected. However no such case has been detected. In case of error respondents always signed that they do not know or are not sure.

Experiment 2

Figure 3 shows the means and standard deviations of correct letters in answer. As correct were considered all letters at proper position in the name or single shift was allowed. For example in case of Vittorio (8 letters), the answer Victorio was evaluated as 7 correct letters just as the answer of Vittoro with one missing letter.

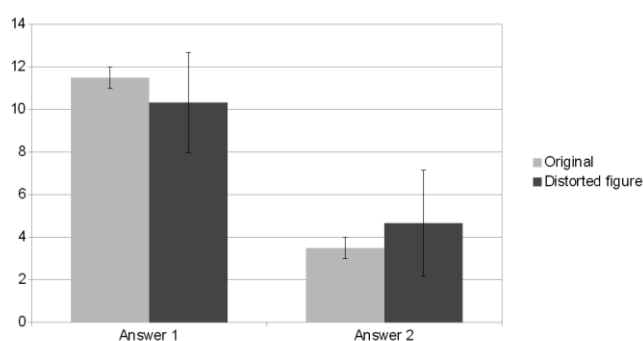
Figure 3. Number of correct letters in answers.



Experiment 3

This experiment applied the same decision criteria as in the second experiment.

Figure 4. Number of correct letters in answers.



V. Conclusion and future work

This paper is a work in progress aiming to develop an extensive methodology for evaluation quality and intelligibility of coding and compression scheme for sign language video.

In experiment 1 was the error rate around 15 % in both cases of distortion, while keeping zero error when watching the signs in original quality. This result corresponds to our former results and confirms that the visual quality of the interpreter's face is most important for understanding of sign language. Distorted face has an apparent impact to inability to identify a sign, while further distortion of the rest of the interpreter's body does not increase the error rate significantly. In both cases, they are different signs in which observers has been mistaken, which means that no matter the particular sign the error rate is higher when distorted than watching the original image. The error rate is not as high as expected considering the distortion level that was chosen at an annoying level and thus higher than assumed to broadcast. The results in experiment 2 seem that answers are similar regardless on the distortion level. In some cases is the span even smaller for original image then for the distorted. The longest memory span is in the case of Czech name Daniela Hodrová with expected Czech suffix "ová", which is typical for the most of female names in the Czech Republic. The participants had to remember only first 4 letters from the surname. The rest of the name could be evident after observing. The Pakistani name has the smallest span due to the unusual composition of vowels and consonants in the name. In neither case cannot be said that the distortion significantly affect the intelligibility of names. Differences among particular names show the influence of working memory and its level by the respondents. It is important to note that the level of short-term memory has not been studied assuming the close value with regard to higher education of all participants. Evaluating the results in experiment 3 indicates similar conclusions as in the second experiment. The results may be influenced by knowledge of the historical event, but we cannot say that the error rate is lower or higher when the picture is distorted.

Considering all proposed experiments implies the conclusion that such a high distortion can moderately affect the intelligibility, but the artifacts in the picture are as distinct as they could be rated as annoying in broadcasting. In future work it is important to find out the level of distortion regarded as acceptable for deaf TV viewers.

References

- [1] The results of the sample survey of disabled people in 2007. Czech statistical office [online]. 30. 5. 2008, [cit. 2013-06-12]. <<http://www.czso.cz/csu/2008edicniplan.nsf/p/3309-08>>.
- [2] F. M. Ciaramello, Jung Ko, and S. S. Hemami, "Quality versus intelligibility: Studying human preferences for american sign language video," in Image Processing Workshop (WNYIPW), 2010 Western New York, Nov., pp. 70–73.

- [3] D. Agrafiotis, N. Canagarajah, D. R. Bull and M. Dye, "Perceptually optimised sign language video coding based on eye tracking analysis". *Electronics Letters*. 2003, vol. 39, no. 24.
- [4] Z. Švachula, M. Bernas, P. Zatloukal, and L. Klabanová, "Investigation of Regions of Interest for Deaf in Czech Sign Language Content," *International Journal of Software and Web Sciences (IJSWS)*, March 2013.
- [5] N. Gordon, "The neurology of sign language: Review". *Brain & Development*. 2004, no. 3, pp. 146-150.
- [6] Anna Cavender, Richard E. Ladner and Eve A. Riskin, "MobileASL: Intelligibility of Sign Language Video as Constrained by Mobile Phone Technology". *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility - Assets '06*. New York, New York, USA: ACM Press, 2006, pp. 71-78.
- [7] HbbTV Roadmap. In: *Hybrid TV in Europe & beyond Symposium* [online]. December 17th & 18th 2012 [cit. 2013-06-12]. <http://hbbtv.org/pages/news_events/docs/hbbtv_roadmap_121217.pdf>.
- [8] D. M. Saxe and R. A. Foulds, "Robust region of interest coding for improved sign language telecommunication," *IEEE Trans. Inf. Technol. Biomed.*, vol. 6, no. 4, pp. 310-316, Dec. 2002.
- [9] Zuzana Silovská, "Minimální páry v českém znakovém jazyce". Praha, 2012. Bachelor thesis. Charles University in Prague, Faculty of Arts, Institute of Czech Language and Theory of Communication.
- [10] Alena Macurová, "Výzkum komunikace neslyšících: český znakový jazyk". *Speciální pedagogika*. 1996, vol. 6, no. 1, pp. 5-19.
- [11] ITU - R Recommendation BT.1683 Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. Geneva : The International Telecommunication Union, 2004.
- [12] Mrim Boutla, Ted Supalla, Elissa L Newport a Daphne Bavelier, "Short-term memory span: insights from sign language". *Nature Neuroscience*. 2004-8-15, vol. 7, issue 9, pp. 997-1002.
- [13] M. Wilson a K. Emmorey, "Comparing Sign Language and Speech Reveals a Universal Limit on Short-Term Memory Capacity". *Psychological Science*. 2006, vol. 17, issue 8, pp. 682-683.
- [14] M. Daneman, P. M. Merilke, "Working memory and language comprehension: a meta-analysis". *Psychon Bull Rev.* 1996, no. 3, pp. 422-433.