An evaluation of finger alphabet intelligibility using quality assessment of video with masked content

Petra Heribanova[+], Jaroslav Polec[*], Darina Tarcsiova[^]
[+]Department of Algebra, Geometry and Didactics of Mathematics, FMFI, Comenius University
[*]Institute of Telecommunications, FEI, Slovak Univerzity of Technology
[^] Institute of Special Education Studies, Faculty of Education, Comenius University
Bratislava, Slovakia
petra.heribanova@fmph.uniba.sk, polec@ktl.elf.stuba.sk, darina.tarcsiova@fedu.uniba.sk.

Abstract

This paper discusses the finger alphabet recognition and evaluating the requirements for image quality and definition of the criteria of automatical, real-time objective evaluation without respondent involvement for speech intelligibility in the video and electronic communications. Tests with respondent were made under the logatom recognizability, as it is the most precise, because we may be able to identify voices or do not know, we can not infer from their context, so it causing people to avoid the tendency to repair improperly admitted syllables. This methodology is based on the intelligibility according to variable transmission channel capacity for different video formats. The aim is to determine video degradation threshold, at which the signs of one handed alphabet are still correctly understood, the degree of degradation of particular alphabet signs and, alternatively, mutual sign exchangeability. The results obtained were applied a standard scale for subjective evaluation of image quality and percentages evaluation of recognizability as used in acoustics. Based on this results of objective evaluation of logatom recognizability with respondent involvement we search the method, which correlate best with intelligibility. The aim of objective methods for evaluation of video quality is to design algorithms whose quality prediction is in good agreement with the results of objective human evaluation and therefore could represent a method for automatic evaluation of video intelligibility with finger alphabet.

Keywords: *finger alphabet, intelligibility, logatom, video, quality, metrics, geometry.*

1. INTRODUCTION

Today is not problem make the high definition (quality) video or image. High definition (quality) video requires considerable volume of data that needs to be transferred (and paid). Therefore, we always try to find the best compromise between acceptable video quality and cost. Means of determining the compromise are coding and compression, closely related to the quality evaluation criteria. There are many methods and metrics for objective evaluation of video quality, where the main criterion is "lovely" of video regardless of content. However in the evaluating of video with different methods of implementation of augmentative and alternative communication (AAK) or specific method of communication people with hearing impairments [10] we can not ignore content – intelligibility of video. The main difference between the terms quality and intelligibility is that the term "quality" describes the appearance of decoded video signal ("how" the viewer sees it) and the "intelligibility" is just one aspect of quality saying if the received information gives any sense ("what" the viewer sees in it). High-quality video signal is likely to be intelligible. Conversely, of course it may or may not apply. Anyway, unintelligibility is an indicator of poor quality. In the acoustics, intelligibility threshold is defined as a point, after which one does hear, but one does not understand [1].

Subjective tests show that sound tends to reduce people's ability to recognize video image degradation. Hearing-impaired people do not rely that much on video quality, as the most important thing to them is whether they are able to understand the meaning. Their subjective video quality evaluation can differ from hearing people. Actually, the biggest difference of video of sign language is its purpose - it is the equivalent of sound channel in normal audiovisual recordings. Our aim is to find criteria for video signal quality encoded in various bit-rates, to achieve full intelligibility of Slovak (or other) sign language and finger alphabet.

There is no recommendation ITU (International Telecommunication Union) for evaluating the quality and intelligibility of the video containing alternative and augmentative means of communication.

Our purpose is to modify the criteria for objective evaluation of quality and create a method for automatic evaluation of speech intelligibility (one - handed finger alphabet) based on [3]. In order to be made automatic objective evaluation of the inteligibility of sign language or finger alphabet, it is necessary to do testing with human factors – respondents. Any such non-automatic testing is not only challenging in terms of time, but also the needs of a large number of respondents. Is it problem. Sign language and finger alphabet is not international, but it is a speech which is divided by nationality and binds to a specific territory, and thus the community of deaf people speak SPJ is size limited. Is there a maximum number of people's of evaluation team, and he is not big or concentrated in one place.

2. SIGN LANGUAGE, FINGER ALPHABET

Sign language is the primary communication tool of deaf and/or hard of hearing people. It is visual and spatial language with its own grammar and sign vocabulary. It has visual motor modality and it is independent of spoken language. But it is not international, Slovakia used Slovak sign language. All sign languages used three-dimensional space (the sign space) for communication, which is defined horizontally and vertically. In sign languages, we have two types of components (parameters), which we can be analyzed :

- manual parameters = location, handshape and movement
- non-manual parameters = facial expression, position of eyes, head, upper body, mouth movement

The basic communication element is sign. It is given by configuration (shape and placement) of the hands in sign space, by palm and finger orientation, and also by hand movements themselves. It is quite difficult to learn the sign from books or

static images, because even slight difference in movement and location of the hand can change the meaning. Hence, personal demonstration, or understandable video preview is needed.

Finger alphabet was not created naturally and spontaneously by deaf people. It was adapted from monasteries for the purpose of teaching children with hearing impairments. It is a system of finger and movement configurations that represent letters of the alphabet. The number of finger alphabet signs is related to the number of letters (graphemes) of the language. It is commonly used for purposes of clarification, such as unfamiliar words, names of persons, geographical names, or with words, for whose the asking person does not know the appropriate sign. An advantage of the finger alphabet is that its adoption is not difficult or time-consuming. It helps to express the words in correct grammatical form and thus it is the tool for obtaining a richer vocabulary. In the world, there are two widely used systems of the finger alphabet [10].

- One – handed finger alphabet (Figure 1)
- Two –handed finger alphabet

In some countries, using both (for example Slovakia), in some countries only two-handed (for example UK), or one –handed (for example USA).

In Slovakia, the situation is as follows:

One -handed finger alphabet is used to teach pupils at schools for children with hearing impairment. It is more widespread in the world. On international meetings, the only used finger-spelling alphabet is the one approved by The World Federation of The Deaf.

Two-handed finger alphabet tends to be used by older people, because it is slower. Despite its slowness, it is also used at lectures and seminars because of its better intelligibility and visibility [2].

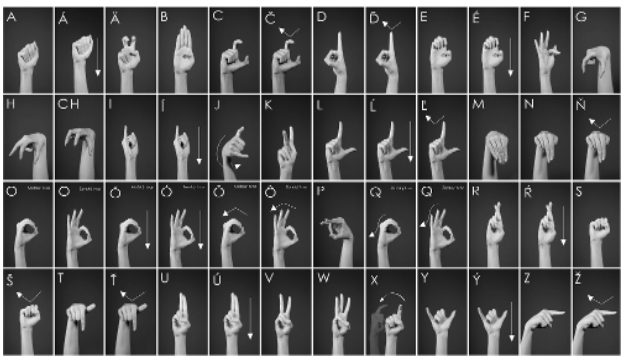


Figure 1: Example of one -handed finger alphabet [2]

3. THE INTELLIGIBILITY

In acoustics, the intelligibility of the language (Z) defines the percentage of correctly received elements or parts of speech (a) divided by their total number (b):

$$Z = \frac{a}{b} * 100 \qquad (1)$$

We distinguish consonant, logatom, word, and sentence based intelligibility. Logatomes are artificial words designed to look alike words of given language, but they do not have the meaning. The term recognizability is used in recognition of speech sounds (phonemes) and logatomes, as one can either recognize or not recognize them, but there is nothing to be understood [5].

Similarly, we can explore the intelligibility of video: sentence and word intelligibility using sign [7], while logatom and consonant recognizability using the finger alphabet. It is possible to create a sort of "sign logatomes" for the deaf, because one sign in finger alphabet represents a one speech sound in logatom.

4. SUBJECTIVE AND OBJECTIVE METHODS FOR THE QUALITY AND INTELLIBILITY AND TESTING

The quality evaluation criteria are closely linked with the encoding and compression, as a means for the intended destination boundaries, as it is possible to reduce the size of the data stream. In basically used to evaluate two groups of methods:

1. Subjective methods of measuring reaction observers pursuing a tested system. These methods are very time consuming and to implement [4].

2. Objective methods are automated methods without the participation of observers and the implementation of any distance gap metrics. The aim is to find appropriate application method with the highest correlation with the results of subjective methods [4, 9,11].

Objectively, intelligibility is measured by statistical methods. In the simplest case, it is the percentage of correctly recognized elements. For sentence intelligibility, recognition is considered successful, when the reproduced sentence has correct context and makes sense. Logatom recognizability is expressed as the percentage of correct consonants and vowels from all speech sounds in transmitted logatomes. Resulting from this, it is clear that logatom based recognizability is much more demanding than sentence or word based one, because the meaning cannot be guessed from the context [1].

4.1 Logatom recognizability

To evaluating the requirements for image quality and speech intelligibility in the video we used the logatom recognizability. Logatom (consonant) recognizability was tested using artificial monosyllabic words without meaning, so called logatom. Using logatoms in our tests, to mitigate people's tendency to correct the incorrectly understood consonants or words according to the meaning. We define the criteria for logatom recognizability in one - handed finger alphabet analogically to acoustics. Sign language and finger alphabet has own character and is incompatible with language for hearing people. We use the finger alphabet signs to create so-called "sign logatoms". Every speech sound in logatom is represented by an appropriate sign from Slovak one-handed alphabet. It is a new evaluation methodology of video signal quality in transmissions of sing language in videoconferencing.

This methodology is based on the intelligibility according to variable transmission channel capacity. The aim is to determine video degradation threshold, at which the signs of one handed alphabet are still correctly understood, the degree of degradation of particular alphabet signs and, alternatively, mutual sign exchangeability.

4.2 Testing

Based on this methodology we created the following experiment. We produced 2 video previews with seven different logatoms in Slovak single-handed finger alphabet (one with 41 consonants, one with 42 consonants). The length of the video previews is about one minute. For the whole experiment we used different video formats with 25 frames per second.

Subsequently, these recordings were encoded by the H.264 codec in various bit rates (QP = 30, 40, 50 that corresponds to rates from 390 kbit/s to 4.5 kbit/s respectively). Testing was realized according to subjective ACR method on groups of hearing impaired volunteers. A random sequence of consonants is quite hard to remember; therefore some sequences were shown multiple times to the same people (in different bit-rate and/or video format) without mentioning it in advance.

a)

b)

Figure 2: Picture taken from the experiment: a) original (704x576); b) "cif" (352x288) H.264 decoded frame with parameter QP=40

The whole test consists of two parts:

1. Subjective, where the respondent that evaluate (by their subjective feelings) the quality and intelligibility of the stream, according to a defined scale.

2. Objective, where the respondent had to rewrite the consonants organized into logatomes to the letters of the Slovak alphabet. While the sentence intelligibility evaluation was based on subjective rating, the logatom recognizability expresses the correctness of all consonants in logatom in percents.

The results obtained were applied a standard scale for subjective evaluation of image quality and percentages evaluation of recognizability as used in acoustics. From the results determine the dependency of recognizability of transmission rate for different video formats. With decreasing recognizability there was an increasing number of consonant interchanges, mostly between 'a' and 's', 'o' and 'f', and there was also higher frequency of missed or extra added consonants.

The test results confirmed that conventional subjective methods to evaluate video quality with markings 1 (nice) - 5 (ugly) are irrelevant and pointless, because do not say anything about intelligibility. In some case subjective feelings of evaluating respondents were contrary to the results of objective evaluating and the results can not be taken as correct evaluation of intelligibility.

The results of objective evaluation of logatom recognizability are the percentage of correctly received signs from all "sign logatoms" in video stream and this results can be taken as correct evaluation of intelligibility.

Based on this results of objective evaluation of logatom recognizability with respondent involvement we test the method from objective methods [4,6,9,11] which correlate best with intelligibility, and therefore could represent a method for automatic evaluation of video intelligibility with finger alphabet.

5. MAIN RESULTS

We work on the basis of full reference method (FR) with differential metrics to evaluate image quality and video between the original and processed video. As the original videos we used the primal videos in format "4-cif" (704x576). The processed video stream in other format "cif" (352x288) and "qcif" (176x144) we resized bilinear interpolation to size of format "4 - cif". Using the software MSU [8] we test the selected metrics, as PSNR, VQM, SSIM, 3SSIM, MSAD and MSE. According to ITU-T recommendations for ratio of objective metrics in regard of the subjective evaluation of a correlation coefficient

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \qquad (2)$$

we find the value of the correlation between existing metrics and intelligibility.

Figure 3: Picture taken from the experiment: Region of interest – elliptical mask

Format	QP	Transmission rate [kbit/s]	Logatom recognizability [%]	MSAD	VQM	SSIM	PSNR	MSE
4cif	30	199,707	94,84	1,651	1,115	0,956	38,414	12,036
4cif	40	74,662	73,40	2,671	1,642	0,905	33,764	29,146
4cif	50	27,172	59,23	4,701	2,589	0,848	29,743	70,780
Correlation				-0,953	-0,961	0,989	0,997	-0,938
cif	30	113,205	91,46	2,431	1,380	0,922	34,958	21,889
cif	40	30,680	73,77	3,528	2,062	0,871	31,497	46,869
cif	50	11,108	39,39	5,809	3,205	0,835	28,151	101,021
Correlation				-1,000	-0,999	0,961	0,981	-1,000
qcif	30	37,923	86,99	3,277	1,866	0,889	32,456	38,191
qcif	40	12,004	64,01	4,559	2,667	0,849	29,595	72,717
qcif	50	5,086	0,00	7,353	3,881	0,824	26,220	158,290
Correlation				-0,999	-0,989	0,921	0,976	-1,000
Total correlation				-0,959	-0,955	0,829	0,874	-0,982

Table 1: The results of correlation without mask

Format	QP	Transmission rate [kbit/s]	Logatom recognizability [%]	MSAD	VQM	SSIM	PSNR	MSE
4cif	30	199,707	94,840	0,618	0,595	0,987	43,729	5,223
4cif	40	74,662	73,400	0,915	0,862	0,977	39,504	9,013
4cif	50	27,172	59,230	1,741	1,503	0,962	34,984	21,682
Correlation				-0,928	-0,939	0,978	0,991	-0,913
cif	30	113,205	91,460	0,783	0,685	0,981	40,683	5,983
cif	40	30,680	73,770	1,153	1,085	0,969	37,382	12,312
cif	50	11,108	39,390	2,117	1,965	0,958	33,250	31,651
Correlation				-0,998	-1,000	0,981	0,993	-0,995
qcif	30	37,923	86,990	1,069	0,986	0,972	37,990	10,935
qcif	40	12,004	64,010	1,552	1,523	0,963	35,112	20,805
qcif	50	5,086	0,000	2,729	2,424	0,956	31,166	51,596
Correlation				-1,000	-0,993	0,936	0,985	-1,000
Total correlation				-0,970	-0,964	0,850	0,893	-0,986

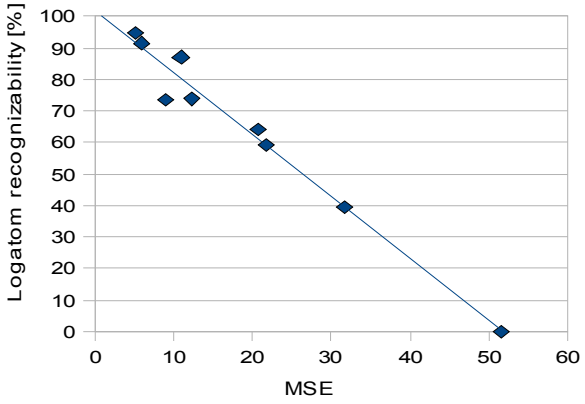
Table 2: The results of correlation with elliptical mask

The results of correlation between logatom recognizability and relevant metrics show Table 1. A comparison between the original and processed video was made for all the pixels with the same value (weight). The best result of correlation for "4 cif" format had the metric PSNR, for "cif" format the metrics MSE and MSAD and for "qcif" format had MSE the best value of correlation. The best results of total correlation had the metric MSE with value of correlation coefficient – 0,982.

Since the video contains areas that are for us in terms of intelligibility finger alphabet irrelevant, we divided the video into regions of interest according to their importance. Figure 3 shows the main region of interest (ROI) for the recognizability of one-handed alphabet. It is an area in the dominant hand (in this case right) showing the signs.

In some case important region can be a mouth, which are especially important to communicate with lip-reading (cued speech) and be used deaf people, too. This region would be subject to the least degradation of the image in coding and compression.The gray part of image contain background (BG). It is not important for intelligibility finger alphabet as wall, clothes, the rest of the face and hair, or second hand and may be of greater degradation by video processing.

By creating masks, we only tested the intelligibility on ROI. The mask have elliptical shape, which is unvaried during testing. In the experiments were used followed settings: ROI = 1 , BG = 0. The results of correlation with the elliptical mask between logatom recognizability and relevant metrics show Tab. 2.



Graph 1: Correlation with mask between MSE annd intelligibility

The best result of correlation for “4 cif” format had the metric PSNR, for format “cif” the metrics VQM and for “qcif” format had MSE and MSAD the best value of correlation. The best results of total correlation had the metric MSE (Graph 1) with value of correlation coefficient - 0,986.

Comparison of the results of correlation with and without the mask is clear to see improvement results using masks for region of interest.

Format	QP	Transmission rate [kbit/s]	Logatom recognizability [%]	MSAD	VQM	SSIM	PSNR	MSE
4cif	30	199,707	94,84	0,301	0,435	0,994	47,061	2,686
4cif	40	74,662	73,40	0,462	0,663	0,988	42,648	5,113
4cif	50	27,172	59,23	0,766	1,216	0,981	38,578	11,645
Correlation				-0,957	-0,939	0,988	0,996	-0,930
cif	30	113,205	91,46	0,400	0,520	0,990	43,906	3,606
cif	40	30,680	73,77	0,592	0,859	0,984	38,676	7,467
cif	50	11,108	39,39	0,971	1,648	0,980	36,817	17,461
Correlation				-1,000	-0,999	0,961	0,900	-0,998
qcif	30	37,923	86,99	0,541	0,772	0,986	41,401	6,299
qcif	40	12,004	64,01	0,778	1,245	0,981	38,448	12,038
qcif	50	5,086	0,00	1,016	2,056	0,978	34,702	28,405
Correlation				-0,965	-0,993	0,916	0,981	-1,000
Total correlation				-0,908	-0,966	0,829	0,863	-0,986

Table 3: The results of correlation with dynamic mask

Figure 4: Picture taken from the experiment: Dynamic mask within ROI (frame 327)

The next table 3. shows the results of correlation with dynamic mask within ROI. On the video, we applied threshold with value about 130 and subsequently elliptical mask. By creating more accurately masks within region of interest, we made mask for each frame in the video (Figure 4). In the same experiments were used settings: the black part of ROI = 1, the white part of ROI and BG = 0. The best result of correlation for "4 cif" format had the metric PSNR, for format "cif" the metrics MSAD and for "qcif" format had MSE the best value of correlation. The best results of total correlation had the metric MSE with value of correlation coefficient – 0,986.

From comparison of the results of correlation for elliptical and dynamic mask is to see improvement results for VQM using dynamic masks for region of interest. The value of MSE remained the same and other results have deteriorated.

6. CONCLUSION

This paper describes the technique of evaluating the quality of video signals based on logatom recognizability using so-called sign logatomes, where it is not possible to guess missed consonants from the context and shows our obtained results in one - handed finger alphabet. The results of logatom recognizability was acquired based on objective evaluation of logatom recognizability with respondent involvement.

Therefore the next part of paper describe evaluating the requirements for video quality and definition of the criteria of automatical objective evaluation without respondent involvement for speech intelligibility (finger alphabet for the deaf) in the video and in the electronic communications. Show the result of correlation between existing relevant metrics and logatom recognizability without mask and with two types of mask use to region of interest in video.

7. AKNOLEDGMENTS

Research described in the paper was financially supported by the Slovak Research Grant Agency (VEGA) under grant No. 1/0602/11 and by Foundation Tatrabanka under projekt No. 11Sds078 and by the Comenius University under project No. UK/106/2012 the Program for support of young researcher for year 2012.

8. REFERENCES

[1] Granat, M. (2009) *Objective methods for evaluation of audio signal quality* (in Slovak), Brno University of Technology, Brno.

[2] Hefty, Michal : Finger alphabet (in Slovak). The organization I think - Development of thinking not only for hearing impaired, 2009. www.zzz.sk

[3] Heribanová, P., Polec, J., Ondrušová, S., Hosōvecký, M.*: Intelligibility of Cued Speech on Video.* In: World Academy of Science, Engineering and Technology. - ISSN 2010-376X. - Iss. 79 (2011), pp. 492-496

[4] ITU-R Recommendation BT.1683: 2004, *Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference.*

[5] Makáň,F.: *Elektroacoustics* (in Slovak), Publisher STU Bratislava, 1995.

[6] Mardiak, M., Polec, J.: *Novel Method for Objectively Measuring Video Quality.* In: Proceedings ELMAR-2010: 52nd International Symposium ELMAR-2010. Zadar, Croatia, 15.-17.9.2010. - Zadar : Croatian Society Electronics in Marine, 2010. - ISBN 978-953-7044-11-4. - pp. 109-112

[7] Mordelová, A. ,Polec, J., Ondrušová, S., Filanová, J. (2010) New Objective Method of Evaluation Cued Speech Recognition in Videoconferences, Proceedings Redžúr 2010, Bratislava, STU v Bratislave FEI. 4 p., CD-Rom.

[8] MSU *Video Quality Measurement Tool.* MSU Graphics & Media Lab (Video Group) Moskva, 2008. [Online] [Dátum: 5.2.2011]

http://compression.ru/video/quality_measure/video_measurement_tool_en.html

[9] Ries, M. et al.: *Video quality estimation for mobile H.264/AVC video streaming.* In: Journal of Communications, vol.3, 2008, no.1, pp. 41-50

[10] Tarcsiová, D.*: The communication system for deaf and ways to overcome their communication barriers. (In Slovak) Sapientia: Bratislava, 2005. ISBN 80-69112-7- 9*

[11] Winkler, S.: *Digital video quality vision model and metrics.* 1. vyd. Chichester : John Wiley & Sons Ltd., 2005. ISBN 0-470-02404-6

About the author

P. Heribanová was born in 1986 in Kremnica, Slovak Republic. She received M.Sc. degree in Geometry from the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava in 2010. She is a PhD. student of Geometry and Topology at the same university. Her research interests include image coding, reconstruction and quality evaluation. Department of Algebra, Geometry and Didactics of Mathematics, FMFI, Comenius University. Her contact email is petra.heribanova@fmph.uniba.sk.

J. Polec was born in 1964 in Trstená, Slovak Republic. He received the M.Sc. and PhD. degrees in telecommunication engineering from the Faculty of Electrical and Information Technology, Slovak University of Technology in 1987 and 1994, respectively. From 2007 he is professor at Department of Telecommunications of the Faculty of Electrical and Information Technology, Slovak University of Technology and at Department of Applied Informatic of Faculty of Mathematics, Physics and Informatic of Comenius University. His research interests include Automatic-Repeat-Request (ARQ), channel modeling, image coding, reconstruction and filtering. Institute of Telecommunications, FEI, Slovak Univerzity of Technology. His contact email is polec@ktl.elf.stuba.sk.

D. Tarcsiova was born in 1963 in Levoča, Slovak republic. She received the M.Sc. and PhD. degrees in Special Education from the Faculty of Education, Comenius University. She is professor at Institute of Special Education Studies of the Faculty of Education, Comenius University. His research interests include special education for deaf people (sign language, finger alphabets, and specific method of education deaf and hard of hearing). Her contact email is darina.tarcsiova@fedu.uniba.sk.