World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:5, No:7, 2011

Intelligibility of Cued Speech in Video

P. Heribanová, J. Polec, S. Ondrušová, M. Hosťovecký

***Abstract*—** This paper discusses the cued speech recognition methods in videoconference. Cued speech is a specific gesture language that is used for communication between deaf people. We define the criteria for sentence intelligibility according to answers of testing subjects (deaf people). In our tests we use 30 sample videos coded by H.264 codec with various bit-rates and various speed of cued speech. Additionally, we define the criteria for consonant sign recognizability in single-handed finger alphabet (dactyl) analogically to acoustics. We use another 12 sample videos coded by H.264 codec with various bit-rates in four different video formats. To interpret the results we apply the standard scale for subjective video quality evaluation and the percentual evaluation of intelligibility as in acoustics. From the results we construct the minimum coded bit-rate recommendations for every spatial resolution.

***Keywords*—**cued speech, inteligibility, logatom, video

I. Introduction

EVOLVING technologies and advanced processing techniques in TV, internet, or telecommunications raise their standards of image quality and sound. But high quality video also 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.

Subjective tests show that sound tends to reduce people's ability to recognize video image degradation. Deaf people however are not affected by sound, so their subjective video quality evaluation can differ from hearing people. Actually, the biggest difference of video of cued speech is its purpose - it is the equivalent of sound channel in normal audiovisual recordings. 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.

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].

Our aim is to find criteria for video signal quality encoded in various bit-rates, to achieve full intelligibility of Slovak (or other) cued speech and finger alphabet.

P. Heribanová is with the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava (e-mail: petra.heribanova@gmail.com).

J. Polec is with the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology Bratislava (e-mail: jaroslav.polec@stuba.sk).

S. Ondrušová is with the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology Bratislava (e-mail: sandra.ondrusova@stuba.sk).

M. Hosťovecký is with the Centre of lifelong learning, Trnava University in Trnava, Slovakia (e-mail: marian.hostovecky@truni.sk).

II. Cued Speech and Finger Alphabet

Cued speech is the primary communication tool of hearing impaired or hard of hearing people. It is visual and spatial language with its own grammar and gesture vocabulary. It has visual-motile modality and it is independent of spoken language. But it is not international. It uses three-dimensional space (the gesture space) for communication, which is defined horizontally and vertically. In gesture languages, we have two types of meaning carriers:

- manual = position, shape and movement of hands
- non-manual = facial expression, position of eyes, head, upper body, mouth movement

The basic communication element is gesture. It is given by configuration (shape and placement) of the hands in gesture space, by palm and finger orientation, and also by hand movements themselves. It is quite difficult to learn the gestures 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. It is a system of finger and movement configurations that represent alphabetic characters. The number of characters is related to the number of speech sounds (phonemes) 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 gesture. 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 [2]:

- Single-handed method (also “finger-spelling”) (Fig. 1)
- Double-handed method. (Fig. 2)

Single-handed finger alphabet (dactyl) is used to teach children at schools for students 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.

Double-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 [3].

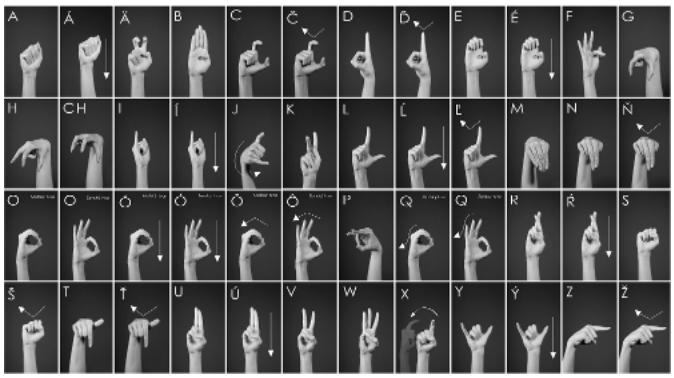

Fig. 1 Example of single-handed finger alphabet

Fig. 2 Example of double-handed finger alphabet

III. The Intelligibility (Recognizability)

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):

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

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 [4].

Similarly, we can explore the intelligibility of video recordings: sentence and word intelligibility using gestures, while logatom and consonant recognizability using the finger alphabet. One sign in finger alphabet represents one speech sound in logatom. Thereby it is possible to create a sort of "sign logatomes" for the deaf.

IV. Subjective and Objective Methods for the Quality and Intellibility

Subjective evaluations are based on comparing the original and processed video signal by a group of hearing impaired volunteers that evaluate (by their subjective feelings) the quality and intelligibility of the stream, according to a defined scale.

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].

A. Sentence intelligibility

In [5] there is a new evaluation methodology of video signal quality in transmissions of gesture language in videoconferencing. This work shows a new objective method for examining the sentence intelligibility (as used in telephonometry for speech sentence articulation) with use of subjective ACR method (full categorical evaluation).

This methodology is based on two criteria - intelligibility according to the variable transmission channel capacity, and the speed of the gesture language. The aim was to determine the video degradation threshold, at which the gesture language sentences are still understood. And, alternatively, how should the speaker adjust his gesture language speed to be still understandable even in worse quality conditions.

Based on this methodology we created the following experiment. We produced 6 video previews with different example phrases in Slovak gesture language (one with a woman, one with a man) in three speed variations - slow, normal and fast. For whole experiment standard video format of 352x288 pixels per frame with 25 frames per second was used. Subsequently, these recordings were encoded by H.264 codec with various bit rates (QP = 30, 35, 40, 45, 50 that corresponds to rates from 87 to 18 kbps respectively).

30 created samples were tested on a group of 15 people. Testing was realized according to subjective ACR method, with the sequences presented one by one. The method was adjusted for evaluating the overall quality of hearing-impaired videoconference. All subjects evaluated the quality of each sequence immediately after its presentation. Time needed for evaluation should not exceed 60 sec, since the demonstration may raise various respondent feelings and thus affect the testing.

a

b

c

Fig. 3 Picture taken from the experiment.: a) original; b) H.264 decoded frame with parameter QP=40; c) H.264 decoded frame with parameter QP=50.

The whole test consists of two parts:

1. Subjective, where the video was evaluated according to given voting options shown in Tab 1.

TABLE I
PROPOSED POTING OPTIONS FOR SENTENCE INTELLIGIBILITY TESTING

1	Completely understandable
2	Partially understood, but understood the content
3	Partially understood, but misunderstood the content
4	Not understandable

2. Objective, where the respondent had to rewrite the sentence he saw in the Slovak gesture language into the Slovak language. The sentence is considered correct also when it is slightly different, but has the correct meaning (because of natural variability in translation from and to the gesture language). Thus it is not necessarily required to understand every word in the sentence, just understand the content.The results show that intelligibility of Slovak gesture language is, of course, dependent on image quality, but not as much as we anticipated. In average, by ensuring the transmission rate of at least 70 kbit/s, one is able to capture the meaning of the whole conversation. Good light conditions, clothing of the speaker and camera settings have also significant impact, as well as gesturing itself that should be expressive enough, to use the whole gesture space.

TABLE II
RESULTS FROM PERFORMED EXPERIMENT

QP	Slow gesture speed			Normal gesture speed			Fast gesture speed			Gesture language speaker
	S	O	Bitrate	S	O	Bitrate	S	O	Bitrate	
30	1	2	126,33	1	1	130,67	2	2	179,5	MAN
35	1	2	68,08	1	2	70,58	1	1	99,58	
40	2	2	39,08	2	2	40,67	1	2	58,75	
45	2	3	24,08	1	2	25,08	3	4	36,58	
50	3	4	17,67	2	3	17,5	4	4	24,83	
30	1	1	87,5	1	1	102	1	1	95,25	WOMAN
35	2	2	48,42	1	2	56,25	1	2	52,83	
40	2	2	29,33	2	2	33,75	1	2	31,75	
45	1	2	19,83	2	3	21,92	2	3	22,08	
50	3	4	18,5	3	3	19,33	3	3	16,5	

S - Subjective evaluation
O - Objective evaluation
Bitrate - Transmission rate [kbit/s]

B. Logatom recognizability

In logatom recognizability evaluation we use artificial monosyllabic words without meaning (logatoms) to mitigate people's tendency to correct the incorrectly understood consonants or words according to the meaning. We create so-called "sign logatoms" – every speech sound in logatom is represented by an appropriate sign from Slovak one-handed or two-handed alphabet. It is a new evaluation methodology of video signal quality in transmissions of gesture language in videoconferencing.

This work shows a new objective method for examining the logatom recognizability (as used in telephonometry for speech sound articulation) with a use of subjective ACR method (full categorical evaluation). 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 alphabet (single-handed and double-

handed) are still correctly understood, the degree of degradation of particular alphabet signs and, alternatively, mutual sign exchangeability.

a

b

c

Fig. 4 Picture taken from the experiment (cut): a) H.264 decoded frame with parameter QP=30 (640x360); b) H.264 decoded frame with parameter QP=40 (640x360); c) H.264 decoded frame with parameter QP=50 (640x360).

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 four different video formats of 1280x720, 640x360, 320x180 and 160x90 pixels per frame 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).

12 created samples were shown to 8 elementary school pupils. Testing was realized according to subjective ACR method. 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. Additionally, there was need for another person (as interpreter), because the children were not able to watch the video and write down the meaning at the same time; so they were just showing (in finger alphabet) what did they see and the interpreter person was writing it into the answer sheet. Then the pupils evaluated the subjective video quality according to Tab 3.

The whole test consists of two parts:

1. Subjective, where the video was evaluated according to given voting options shown in Tab 3.

TABLE III
PROPOSED VOTING OPTIONS FOR CONSONANT INTELLIGIBILITY TESTING

1	Completely understandable
2	Understandable
3	Sporadically inapprehensible
4	Inapprehensible

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 of the intelligibility evaluation of single-handed Slovak finger alphabet fulfilled our anticipation, as they are clearly dependent on image quality. Minimal video transfer speed depends on video format settings. At the bit-rate of 35 kbit/s and video format 640x360 the respondent is able to recognize 50% of consonants, while at the same bit-rate, but in 160x90 the respondent recognizes nearly 90%. 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. Light conditions, camera settings, and background color have big impact on overall intelligibility, as well as on visibility and ability to interpret the signs.

TABLE IV
RESULTS FROM PERFORMED EXPERIMENT

Resolution	QP	30	40	50
160x90	Objective evaluation [%]	90,24	71,95	0
	Subjective evaluation	2	3	4
	Bitrate [kbit/s]	18,3	7,2	4,5
320x180	Objective evaluation [%]	93,90	76,19	45,23
	Subjective evaluation	2	3	4
	Bitrate [kbit/s]	51,6	17	7,5
640x360	Objective evaluation [%]	96,49	83,30	50,00
	Subjective evaluation	1	2	3
	Bitrate [kbit/s]	149,2	42,1	18,4
1280x720	Objective evaluation [%]	95,23	95,12	93,90
	Subjective evaluation	1	1,5	2
	Bitrate [kbit/s]	389,4	126,9	56,2

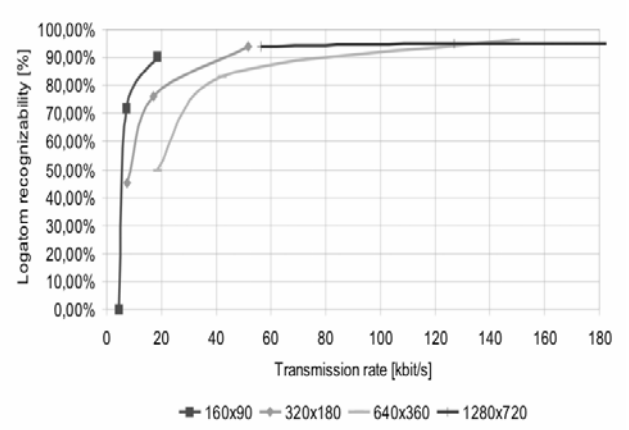

Fig. 5 Dependency logatom recognizability on the transmission rate

V. CONCLUSION

This paper describes the technique of evaluating sentence intelligibility in gesture language and also shows our obtained results. In the first set of results we focus on sentence intelligibility, where (under certain circumstances) it is possible to guess missed words from the context. Conversely, in special cases one needs to use finger alphabet to address the meaning. Therefore the next set describes the methodology of evaluating the quality of video signals based on logatom recognizability using so-called sign logatomes and the result in single-handed finger alphabet used by children. Working with children, however, is more time consuming as help of another person (interpreter) is required.

In our next work we will further investigate the methodology of evaluating the quality of video signals based on logatom recognizability for double-handed finger alphabets.

ACKNOWLEDGMENT

Research described in the paper was financially supported by the Slovak Research Grant Agencies: KEGA under grant No. 119-005TVU-4/2010 and VEGA under grant No. 1/0602/11.

REFERENCES

[1] M. Granat, *Objective methods for evaluation of audio signal quality* (in Slovak), Brno University of Technology. Brno, 2009.
[2] D. Tarcsiová, Pedagogics of hearing-impaired *(in Slovak)*, MABAG spol. s r. o., Bratislava, 2008.
[3] M. Hefty, *Finger alphabet* (in Slovak), Organization Myslím – development of thinking not only for hearing-impaired *(in Slovak)*, 2009, www.zzz.sk
[4] F. Makáň, *Elektroacoustics* (in Slovak), Vydavateľstvo STU Bratislava, 1995.
[5] J. Polec, S. Ondrušová, A. Mordelová, J. Filanová, "New Objective Method of Evaluation Cued Speech Recognition in Videoconferences," *Proceedings Redžúr 2010: 4th International Workshop on Speech and Signal Processing*, Bratislava, Slovak Republic, May, 14, 2010 - Bratislava, STU v Bratislave FEI, 2010, 4 p., CD-Rom.

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.
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.
S. Ondrušová was born in 1983 in Topoľčany, Slovak Republic. She received M.Sc. degree in telecommunication engineering from the Faculty of Electrical and Information Technology, Slovak University of Technology in 2008. She is a PhD. student of telecommunication engineering at the Slovak University of Technology. Her research interests include image coding, reconstruction and filtering.
J. Poctavek was born in 1984 in Hnúšťa, Slovak Republic. Received M.Sc. degree in telecommunication engineering from the Faculty of Electrical and Information Technology, Slovak University of Technology in 2008. He is a PhD. student of telecommunication engineering at the Slovak University of Technology in Bratislava. His research interests include error rate simulation and measuring in transmission channels.