Subjective Evaluation of Audiovisual Signals

F. Fikejz

Abstract

This paper deals with subjective evaluation of audiovisual signals, with emphasis on the interaction between acoustic and visual quality. The subjective test is realized by a simple rating method. The audiovisual signal used in this test is a combination of images compressed by JPEG compression codec and sound samples compressed by MPEG-1 Layer III. Images and sounds have various contents. It simulates a real situation when the subject listens to compressed music and watches compressed pictures without the access to original, i.e. uncompressed signals.

Keywords: subjective test, audiovisual signal quality, compression, JPEG, MP3.

1 Introduction

We come into contact with compressed images and sounds every day. Compression is used to save space on a disc or to maximize the speed of data transfers. The compromise between compression and subjective quality is still under investigation. A subjective test is one way to find it.

2 Methodology

Good guidelines on how to prepare subjective tests can be found in [1] and [2]. A method without a reference is good if the author of the test wants to model a real situation. If he wants to compare something (such as different types of compression), methods with a reference are better. For more on methods, see Chapter 2.2 in this paper. The choice of suitable subjects is as important as a calibration in a physical measurement. According to [1], subjects should have stable emotions and physical condition, they should be well motivated, they should present the typical reactions of a focus group, they should be independent of the author, and they should not have information that could influence them (about the compression formats, the types of displays, the loud-speakers, etc.)

2.1 Psychometry

Psychometry is field of psychology. It concerns measurement of psychological effects. Psychometric tests evaluate the influence of sound and image stimuli on humans. As subjective test is a type of psychometric test in which the subjects evaluate, in our case, sound and image quality in a defined way. Based on subjective test results, new compression algorithms can be programmed and new audiovisual technology can be developed.

It is important to train subjects correctly on what they are evaluating and how they should do it. The instructions, including the degree scale, should be comprehensible for people with no experience of subjective tests. One of the most important considerations for subjective evaluation is duplication. The same test can be performed several times with the same subject, or the same test can be performed with several subjects. The second version is more frequently used. This method was implemented in the test described in this paper. Opponents of subjective tests say that it is not possible to achieve the same conditions all the time, due to the influence of time and atmosphere, and in addition each subject is an individual. Supporters say that quite similar conditions can lead to good validity (see [1]).

2.2 Methods

Subjective tests can be implemented using various types of methods. Two methods are widely employed. The first type comprises methods without a reference (simple rating methods, Single Stimulus Continuous Quality Evaluation – SSCQE, etc.), while the second type consists of methods with pair comparison (Double Blind Triple Stimulus with Hidden Reference – DBTS, Double Stimulus Continuous Quality Scale – DSCQS, Double Stimulus Impairment Scale – DSIS, etc.)

When a simple rating method is used, the subject has to classify each sample on a metric scale. This is most often a numeric or graphic scale, for more details on types of scales see [2]. The degrees of the scale can be characterized by words, but more frequently a maximum and a minimum are defined.

SSCQE is a relatively new method for dynamic rating of video sequences, and it is suitable for an objective evaluation. The subject evaluates the quality of a video sequence in time periods. For more on this method, see [3], and for a comparison with DSCQS, see [4].

In the DBTS method, the subjects listen to a reference signal and then two samples of an audio signal,

one of which is the reference and the other is a compressed signal. The subjects try to identify the compressed signal and evaluate it. The reference signal is evaluated by the best degree of the scale. A similar method for visual signals is DSCQS. The sequence of two samples is played twice. One is the reference and the other is the compressed image. For more, see [5]. The DSIS method is quite similar, but the subjects know that the first image is the reference and second image is compressed. This may be easier for the subjects than classic DSCQS. For a comparison of the two methods, see [6].

2.3 Compression formats

An audiovisual signal of a subjective test was realized by JPEG pictures accompanied by MP3 compressed sound. JPEG (JPG) is a compression standard which has become the most widely used format for storing digital photos. The principle of the whole algorithm is described in [7]. MPEG-1 Layer III (MP3) is a very popular sound format. For more about MP3, see [8] and [9].

3 Subjective test

The main goal of this subjective test was to model the real situation where people watch compressed images on TV and listen to compressed music. Therefore the subjects did not compare the samples of the audiovisual signal with a reference. They evaluated each sample. The simple rating method used a metric scale from 1 (the worst) to 10 (the best).

3.1 Preparation and realization

All basic image and sound samples have different contents. It is important to eliminate situations where some subject has a previous positive emotional bias towards the signal contents, unlike another subject, and automatically evaluates it more positively. The inverse situation, when a subject has a negative bias towards some content and automatically rates it negatively should also be avoided. The test described in this paper consisted of five basic samples, as listed in Tab. 1a.

All photos were taken with a Nikon E8700 Coolpix camera with $3\,264 \times 2\,448$ pixel picture resolution. The quality was set up to 100 percent. The images were compressed in Irfan View 4.23 software to JPEG with qualities of 80, 60, 40 and 20 percent. All sounds were taken from original CDs. Each original sound was a WAV file with bit rate 1 411 kb/s, quantization 16 bits, stereo signal with sample frequency 44.1 kHz modulated by PCM. The samples were edited in Sony Sound Forge 9.0, and in this software they were compressed to MP3 format with bitrates 256 kb/s, 128 kb/s, 96 kb/s and 64 kb/s. The subjects did not know which formats and standards were being used in the test, because this could have influenced their rating.

Table 1: a) List of samples, b) Authors of music

	Image	Music
1	Decoration	Snow (Hey Oh)
2	Troubadours	The Handsome Cabin Boy
3	Butterfly	Jeux de vagues
4	Church	Invitatorium Hodie Exultandum
5	Monte Carlo	VROOOM

a)

Author (Album, Track)
1 Red Hot Chilli Peppers (Stadium Arcadium, 2 – CD1)
2 Sweeney's Men (The Irish Folk Collection, 1)
3 Claude Debussy (Jean Fournet conducts Debussy, 5)
4 Schola Gregoriana Pragensis (Bohemorum Sancti, 1)
5 King Crimson (Thrak, 1)

b)

The main idea of this subjective test was to discover the interaction between acoustic and visual quality. There were ten combinations of compressed sounds and images. Some of them were extreme (better audio quality with worse visual quality, or worse audio quality with better visual quality), while others were compromises with medium quality of both. A complete list of the combinations of compression is shown in Tab. 2.

Table 2: Combination of sound and image compression /average size of file – megabytes/

C.	Image Format /MB/	Sound Format /MB/
1	JPEG (100 %) /1.55/	WAV (1 411 kb/s) /3.2/
2	JPEG (20 %) /0.3/	MP3 (64 kb/s) /0.15/
3	JPEG (60 %) /0.6/	MP3 (128 kb/s) /0.3/
4	JPEG (80 %) /1.03/	MP3 (256 kb/s) /0.6/
5	JPEG (40 %) /0.5/	MP3 (96 kb/s) /0.2/
6	JPEG (100 %) /1.55/	MP3 (64 kb/s) /0.15/
7	JPEG (20 %) /0.3/	WAV (1 411 kb/s) /3.2/
8	JPEG (60 %) /0.6/	MP3 (256 kb/s) /0.6/
9	JPEG (80 %) /1.03/	MP3 (96 kb/s) /0.2/
10	JPEG (40 %) /0.5/	MP3 (128 kb/s) /0.3/

Each basic sample of Tab. 1a was modified to all ten combinations of Tab. 2. The whole test was composed of fifty samples. The length of each sample was between 17 and 20 seconds. All samples were numbered from 1 to 50. After each sample there was a black display with a white text: Rate example number "x", example number "$x + 1$" follows" where "x" was

a number of the sample being played. The subjects had ten seconds for their rating, which they marked on the test form. The length of the whole test was 25 minutes.

The subjective test was created in Sonic Foundry Vegas 4.0 software as an AVI file. The resolution was PAL DV (720×576 px/25 fields per second). The test was conducted at the Department of Radioelectronics, CTU FEE, in Room B3-552 (multimedia studio). The display was realized using a Panasonic TH-50PX8EA plasma monitor with $1\,280 \times 768$ pixel resolution and high colour quality (32 bits). The monitor was connected with the PC by HDMI (High Definition Multimedia Interface). The sound system consisted of Event Electronics TR8 (100 W/220 V/0.5 A) stereophonic loud-speakers. The block scheme of the workspace is shown in Fig. 1.

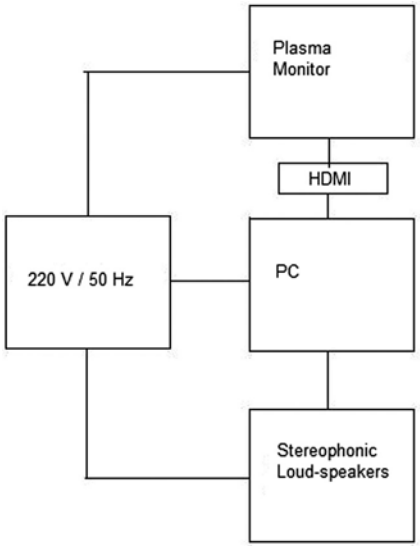

Fig. 1: Block scheme of workspace

The subjects evaluated all samples with regard to subjective quality using the simple rating method. They used a scale from 1 (worst quality) to 10 (best quality). Before the start of the test they were informed that it was not a pair comparison test with a reference, and that they had to evaluate each sample independently. The order of the samples was generated by the Latin squares algorithm (for details, see [12]). The frequency of the same types of samples was not regular, so the subjects did not know which of the five types of samples would come in order.

41 participants (33 male and 8 female) took part in this subjective test. They were aged between 16 and 29 and one person was 36 years old. 21 of them wrote in the answer sheet that they had experience of image and sound processing.

3.2 Analysis and results

The whole statistic evaluation of the subjective test was implemented in MATLAB 7.1. The data was entered into a 50×41 matrix. The rows of the matrix represent the samples, and the columns represent the subjects. The mean value, standard deviation and variance were calculated for all samples of the audiovisual signal. The mean value fell between 4.5 and 7.5, the standard deviation fell between 1.3 and 2.3, and the variance fell between 1.8 and 5.6. In some subjective tests, a few of the last samples can be eliminated, but not in our test, because the variance in the evaluation of the last samples does not show constant growth. This indicates that fatigue of the subjects did not influence our data.

Another element in evaluating this test was calculating the average values for all combinations of compression according to Tab. 2. This was the sum of all five values (one per type of sample according to Tab. 1a) for each combination. This result was divided by 5 (number of types of samples). Fig. 2 shows the average values of all combinations. The values are shown in Tab. 3.

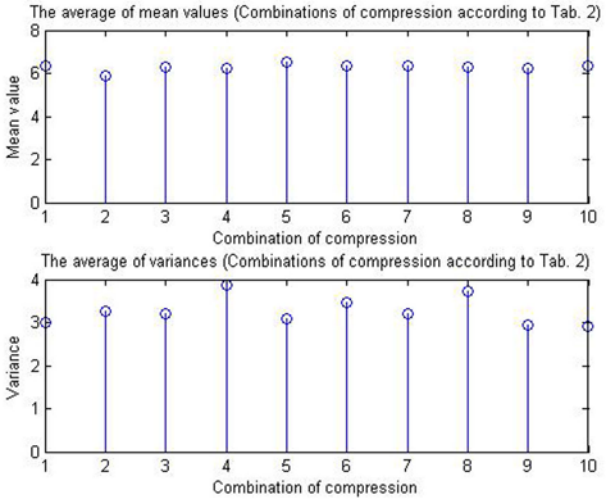

Fig. 2: Average of the mean values and variances of each combination

Table 3: Average of the mean values and variances of each combination

	Combination	μ	σ^y
1	100 % + WAV	6.37	3.01
2	20 % + 64 kb/s	5.90	3.26
3	60 % + 128 kb/s	6.29	3.21
4	80 % + 256 kb/s	6.21	3.87
5	40 % + 96 kb/s	6.50	3.10
6	100 % + 64 kb/s	6.38	3.46
7	20 % + WAV	6.33	3.20
8	60 % + 256 kb/s	6.29	3.71
9	80 % + 96 kb/s	6.25	2.95
10	40 % + 128 kb/s	6.37	2.93

The highest mean value is surprisingly for combination no. 5 (40 % JPEG and MP3 with bit rate

96 kb/s). The reason may be that people are used to listening to compressed records rather than original CDs. The combination of originals was given the third highest mean value. The lowest mean value went to the combination of the worst qualities. Almost all combinations apart from the worst oscillated around 6.35. The highest variance was observed for combination no. 4 (80 % JPEG, MP3 with bit rate 256 kb/s).

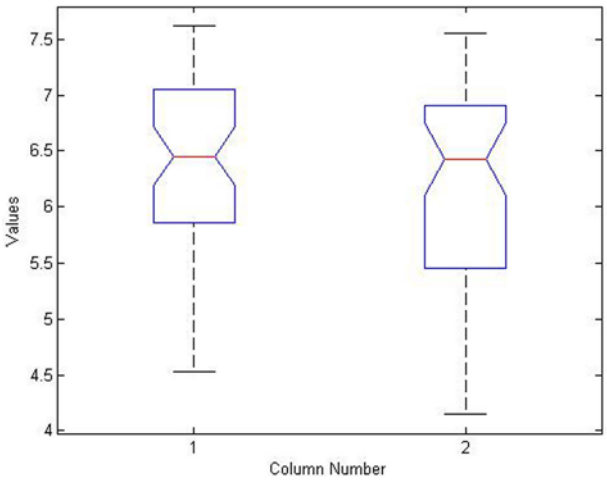

Fig. 3: ANOVA (1 – experienced subjects, 2 – inexperienced subjects)

The evaluation was performed for experienced and inexperienced subjects, respectively. Experienced subjects rated the samples with lower variance than inexperienced subjects. The best rated combinations by experienced subjects were no. 3 (60 % JPEG and MP3 with bit rate 128 kb/s) and no. 5 (40 % JPEG and MP3 with bit rate 96 kb/s) with average mean values of 6.61 and 6.60 respectively. The combination of originals was given a value of 6.48.

Inexperienced subjects rated no. 6 the best combination (original image with the worst sound quality), with average mean values of 6.43, but the rate of both types with original sound (nos. 1 and 7) was also high (6.26 and 6.28). Combination no. 3 which received the highest rating from experienced subjects was rated as the second worst. Both experienced and inexperienced subjects identified the combination of the worst qualities well.

The difference between the mean values awarded by experienced and inexperienced subjects was very small, so there was a hypothesis that the mean values of both groups were almost the same. The results of both groups were analyzed by ANOVA (analysis of variance). For more on ANOVA, see [10]. Fig. 3 and Tab. 4 show the results for ANOVA. SS means the Sum of Squares due to each source, df means the degrees of freedom associated with each source, MS is Mean Squares for each source, which is the ratio SS/df, F shows the statistic, which is the ratio of the MSs. The $Prob$ value decreases as F increases. Because of this result ($F \to 1$) and because of the mean values of both groups were quite similar, the hypothesis can be applied. Fig. 4 and Tab. 5 show that ANOVA was also applied to all subjects, but F was too high and the same hypothesis was rejected.

Table 4: ANOVA (1 – experienced subjects, 2 – inexperienced subjects)

Source	SS	df	MS	F	$Prob > F$
Columns	0.893 5	1	0.893 48	1.28	0.259 9
Error	68.174 7	98	0.695 66		
Total	69.068 2	99			

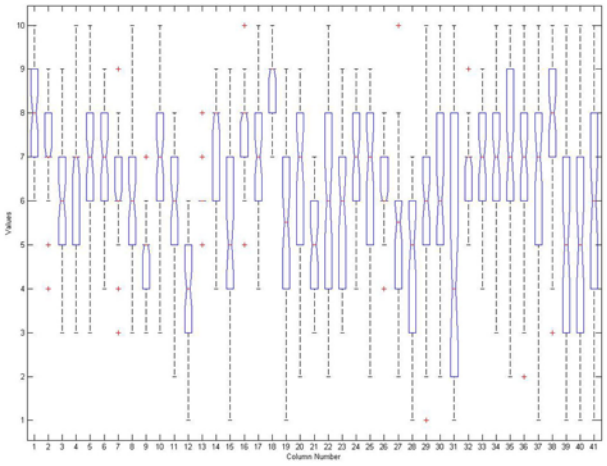

Fig. 4: ANOVA (All subjects)

Table 5: ANOVA (All subjects)

Source	SS	df	MS	F	$Prob > F$
Columns	2 144.82	40	53.620 6	18.62	0
Error	5 784.52	2 009	2.879 3		
Total	7 929.34	2049			

Another part of the evaluation was an analysis of the mean value and variance of all combinations for each type of sample for Tab. 1a. The x axis represents the compression number according to Tab 2., and the y axis shows the mean value or variance. This is like in Fig. 1, but average values of all five types of samples were displayed there). Sample no. 1 (according to Tab. 1a), a photo of decoration and music from Red Hot Chilli Peppers, was rated very well. The best classification was given to combination no. 3 (60 % JPEG and MP3 with bit rate 128 kb/s) and the worst combination was no. 6 (original image and the worst sound quality). In this case, the subjects were influenced by the sound quality.

The second sample (Troubadours and an Irish ballad) was scored very low, and the subjects were more influenced by the sound quality. Sample no. 3 (Butterfly on the flower, with music by impressionist composer Claude Debussy) was scored very high. The variance of the rating was very low. Combination no. 6 (original

image and the worst sound quality) was given the highest mean value. The subjects were more influenced by the image quality. Sample no. 4 (Church and Gregorian chant) was also rated well. The subjects evaluated combination no. 7 (the worst image quality with original sound) as the best. The subjects were probably influenced by the sound quality.

The worst classification was given to sample no. 5 (Monte Carlo panorama and music by King Crimson). The best classification was for compromise combination no. 10 (40 % JPEG with MP3 with bit rate 128 kb/s). In this case, the subjects were influenced by the quality of the images.

Special attention is focused on each combination according to Tab. 2. The values for each type of sample according to Tab. 1a are compared. The results for combination no. 1 (originals) are shown in Fig. 5, where the numbers on the x-axis represent the types of samples, and the numbers on the y-axis represent the mean value and variance.

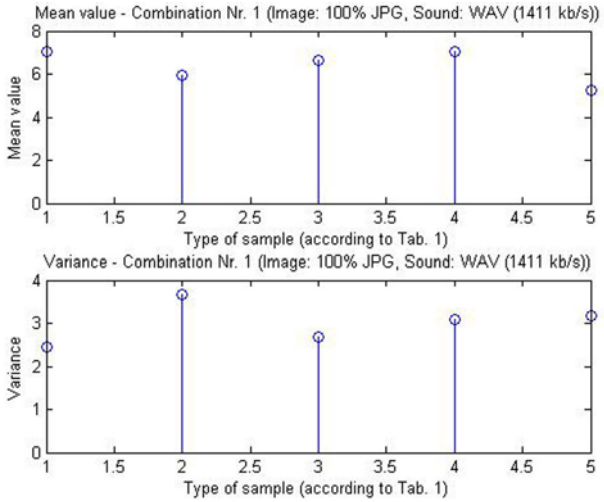

Fig. 5: Mean value and variance – Image 100 %, Sound WAV

The results of the other combinations were quite similar for the mean value. The rating of each combination (according to Tab. 2) was not influenced by the type of sample (according to Tab. 1a). The variances were more varied. Samples no. 1, no. 3 and no. 4 were regularly rated better than samples no. 2 and no. 5. This result may be influenced by the selection of subjects. Younger people mostly give preference to music like sample no. 1, and they do not have much experience with classical music and Gregorian chant. After the test, many of subjects said that they had recognised marks of compression in samples no. 2 and no. 5. On the basis of all the results, the subjects were influenced by the sound element of the sample in 1, 2 and 4, and they were influenced by the image elements of the audiovisual signal in the case of samples no. 3 and no. 5. More details about the whole subjective test are given in [11].

4 Conclusion and discussion

This subjective test evaluated combinations of static images and parts of sound records. The aim of the test was to investigate the interaction between the sound quality and the image quality of audiovisual signals. Forty-one subjects evaluated ten combinations of compressed audio and video signals. Each of ten combinations, see Tab. 2, was applied to five types of samples of different content, according to Tab. 1a. Thus the test was composed of fifty samples in all. A simple rating method was used (with a metric scale from 1 – the worst to 10 – the best). As in real life, the subjects had no reference when they were watching compressed images and listening to compressed music.

It is interesting that a combination of 40 % quality image and an MP3 record with a bit rate of 96 kb/s was the best rated. A combination of 60 % quality of the image and an MP3 record with a bit rate of 256 kb/s, and all combinations with the original sound or the original image were also well rated. A combination of the worst image and the worst sound was the worst rated. More experienced subjects had lower variance in their evaluation, but there was no great difference between the mean values of experienced subjects and inexperienced subjects.

Samples no. 1, no. 3 and no. 4 (according to Tab. 1a) were regularly rated much better than samples no. 2 and no. 5. In the case of no. 1, no. 2 and no. 4, the subjects were particularly influenced by the sound quality, and in the case of no. 3 and no. 5 they were particularly influenced by the image quality. The type of situation determines whether people are more influenced by sound quality or by image quality.

Acknowledgement

This research has been supervised by Libor Husník and supported by research project no. SGS10/265/OHK3/3T/13 Modern modeling and monitoring methods in acoustics.

References

[1] Melka, A.: *Základy experimentální psychoakustiky*, Praha : Akademie múzických umění, 2005.

[2] Guilford, J. P.: *Psychometric Methods*, New York, McGraw-Hill, 1936.

[3] Horita, Y., Miyata, T., Gunawan, I. P., Murai, T., Ghanbari, M.: Evaluation Model Considering Static-Temporal Quality Degradation and Human Memory for SSCQE Video Quality. Visual Communications and Image Processing, *SPIE* Vol. **5 150**, 2003, p. 1 601–1 611.

[4] Pinson, M., Wolf, S.: Comparing subjective video quality testing methodologies. Institute

for Telecommunication Sciences (ITS), National Telecommunications and Information Administration (NTIA), U.S. Department of Commerce.

[5] ITU-R BT.500-11, Methodology for Subjective Assessment of the Quality of Television Pictures.

[6] Van Den Ende, N., Meesters, L. M. J., Haakma, R.: Relation between DSIS and DSCQS for Temporal and Spatial Video Artifacts in a Wireless Home Environment. Human Vision and Electronic Imaging XII, SPIE-IS & T/Vol. **6 492**, 2007, p. 64920lL-1-11.

[7] Furht, B.: *Encyclopedia of Multimedia.* Springer, 2005, p. 372–373.

[8] Spanias, A., Painter, T., Atti, V.: *Audio Signal Processing and Coding.* Wiley Interscience. 2007, p. 120–128.

[9] Kahrs, M., Brandenburg, K.: *Applications of Digital Signal Processing to Audio and Acoustics.* Kluwer Academic Publishers. 2003, p. 75–78.

[10] Sung-Hwan Shin, Jeong-Guon Ih, Hyuk Jeong: *Statistical Processing of Subjective Listening Test Data for PSQ.* Institute of Noise Control Engineering. 2003 July–August, p. 232–238.

[11] Fikejz, F.: *Metodika subjektivního vyhodnocování multimediálního signálu.* Prague, 2009. Diploma thesis, Department of Radioelectronics, CTU FEE in Prague, 2009. Dept. of Radioelectronics. Supervisor Libor Husník.

[12] *Latinské čtverce.*
http://math.feld.cvut.cz/demlova/teaching/avt/pred-a09.pdf [quoted 2009–05–10].

About the author

Filip FIKEJZ was born in Prague, Czech Republic, on December 15th, 1983. He graduated with an Ing. degree from the Czech Technical University in Prague, Faculty of Electrical Engineering in 2009. He is now a PhD. student at the Department of Radioelectronics, CTU FEE. His research interests are in audio signal processing from the psychoacoustic point of view.

Filip Fikejz
E-mail: fikejfil@fel.cvut.cz
Dept. of Radioelectronics
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2, 166 27 Praha 6, Czech Republic