Research Article

Objectives for New Error Criteria for Mobile Broadcasting of Streaming Audiovisual Services

Heidi Himmanen,[1,2] Miska M. Hannuksela,[3] Teppo Kurki,[1] and Jouni Isoaho[1]

[1] *Department of IT, University of Turku, Turku 20014, Finland*
[2] *Turku Centre for Computer Science (TUCS), Turku 20520, Finland*
[3] *Media Laboratory, Nokia Research Center, P.O. Box 1000, Tampere 33721, Finland*

Correspondence should be addressed to Heidi Himmanen, heidi.himmanen@utu.fi

This paper demonstrates the need of and objectives for new error criteria for mobile broadcasting and the problems related to defining numerical error criteria for video services. The current error criterion used in digital video broadcasting to handheld (DVB-H), namely, multiprotocol encapsulation forward error correction (MPE-FEC) frame error ratio (MFER) 5%, was defined to enable instantaneous measurements but is not accurate enough for detailed simulations or postprocessing of measured data. To enable accurate transmission system design, parameter optimization, and performance evaluation, it is necessary to define new practical criteria for measuring the impact of transmission errors. The ambiguity of the MFER criterion is studied, and results for other conventional error criteria are derived from transmission system simulations and objective video quality measurements. The outcomes are compared to results from studies on subjective audiovisual quality. Guidelines are given on the next steps of developing new objective criteria for wireless and mobile video. It is suggested that subjective tests are performed based on the average length and average amount of errors derived from verified mobile radio channel models.

1. INTRODUCTION

Mobile broadcasting is a strong trend in modern telecommunications, and one of the driving forces is real-time television (TV) services to mobile terminals. One of the most popular mobile broadcasting standards is digital video broadcasting-handheld (DVB-H) [1] with two main services defined: broadcasting of streaming video applications and file delivery. These two service categories are of very different nature and have different system requirements. Streaming video services, such as TV programs, are real-time services with hard latency constraints. In video applications, some residual errors can be accepted, without sacrificing the subjective audiovisual quality. File delivery applications, on the other hand, require that the file is received or reconstructed correctly before it can be used, while delays are not as serious a matter as for streaming video.

In this article, we consider streaming video services and their error criteria on the transmission system. We take DVB-H as a case study. What brings more complexity to analyzing audiovisual quality is the lack of good objective measures. Further, subjective quality and the importance of audio or video elements are content-dependent. In DVB-H, the multiprotocol encapsulation-forward error correction (MPE-FEC) frame error ratio (MFER) criterion does not give an unambiguous measure of the quality of an audiovisual stream transmitted over the wireless network. Thus, the transmission system designers lack one sufficient tool for optimizing the system performance, as fair comparisons of different solutions cannot be carried out. Inaccurate error criteria can even lead to wrong conclusions about the optimal solutions and parameters. The baseline for this article is that the technical requirements and criteria for designing and optimizing communication systems should be defined based on the requirements set by the services and applications, but should be easily measurable using common existing tools.

The scope is to demonstrate the shortcomings of the current criterion and show the way forward in designing new criteria. The paper gives the transmission system perspective of streaming audiovisual services, video quality,

and objective error criteria. We explain the requirements on the joint effort between transmission system designers, audio, and video codec experts, and researchers of usability and human-centred technology. The development of the new error criteria will require a huge amount of additional tests and measurements on channel and transmission error statistics and subjective tests to find threshold values for subjectively perceived acceptability. The paper explains what information and further testing are required from the application and subjective testing in order to design measures that meet the requirements for the transmission system criteria.

The article is arranged as follows. First, an overview of the audio and video compression for DVB-H is given in Section 2. DVB-H as a transmission system is presented in Section 3, and current obstacles in system optimization are illustrated in Section 4 using DVB-H simulation results. In Section 5, comparisons to available subjective quality test results are made. Section 6 gives some background and proposes objectives and test cases for transmission system testing, video codec parameter selection, and subjective testing. Finally, we conclude the article.

2. AUDIO AND VIDEO COMPRESSION FOR DVB-H

The IP data casting specifications of DVB-H recommend the use of the high efficiency advanced audio coding version 2 (HE AAC v2) [2] for audio compression and advanced video coding (H.264/AVC) [3] for video compression. Elementary units for transmission of HE AAC v2 and H.264/AVC bit streams are called an access unit and a network abstraction layer (NAL) unit, respectively. An integer number of access units or NAL units are typically encapsulated into one transmission packet. An access unit of HE AAC v2 contains a coded representation of a frame of audio samples. NAL units can be categorized to video coding layer (VCL) NAL units and non-VCL NAL units. VCL NAL units are typically-coded slices of a picture, covering a certain spatial area of the decoded picture. Non-VCL NAL units are used to convey information that is only indirectly related to the decoding process of the coded pictures. Primary-coded pictures of H.264/AVC can be categorized to three types: instantaneous decoding refresh (IDR) pictures, other reference pictures, and nonreference pictures. An IDR picture contains only intra-coded slices and causes marking of all previous reference pictures to be no longer used as references for subsequent pictures. An IDR picture can, therefore, be used as a random access point for starting of decoding or joining a session and it also provides a resynchronization point for decoding after transmission errors have occurred. A reference picture is stored and maintained as a prediction reference for interprediction until it is marked no longer used for reference according to the reference picture marking process of H.264/AVC. A nonreference picture is not used for reference in interprediction and can, therefore, be removed from a bit stream without consequences to any other pictures.

There are no widely accepted objective methods for measuring subjective audiovisual quality. Certain methods, such as the peak signal-to-noise ratio (PSNR), can be used in controlled conditions for pairwise comparison but are not generally suitable for quality measurement, for example, when there are more than one source for quality degradation, such as coding impairments and transmission errors [4]. Moreover, the subjective expectation of the quality, the compression efficiency, and the relative importance of audio and video depend on the type of audiovisual content [5]. Hence, large-scale subjective testing is ultimately the only accurate mean for audiovisual quality measurement.

3. DVB-H AS A TRANSMISSION SYSTEM

3.1. Link layer operations

DVB-H is based on the terrestrial DVB-T standard and was ratified by the European telecommunications standards institute (ETSI) in December 2004. The link layer of DVB-H is an amendment to the physical layer of DVB-T to enable better mobile reception and low-power consumption for handheld devices. A good overview of DVB-H can be found in [6].

The link layer operations are presented in Figure 1. The audiovisual content is passed to the link layer in internet protocol (IP) datagrams. The datagrams are encapsulated columnwise into an MPE-FEC frame, the size of which can be selected flexibly. The number of rows of an MPE-FEC frame can be 256, 512, 768, or 1024. The encoding of the MPE-FEC frame using a Reed-Solomon (RS) (255,191) code [1] is performed rowwise, which results in an interleaving scheme referred to as virtual time-interleaving. By varying the amount of application data columns (1–191) and RS data columns (0–64), different code rates can be achieved. If all application and RS data columns are used, the MPE-FEC code rate is 3/4. MPE-FEC code rates are not fixed by the standard, but commonly considered options are 1/2, 2/3, 3/4, 5/6, 7/8, and 1, which represent uncoded link layer. The Reed-Solomon code can correct as many erasures on each row as there are redundancy columns. Thus, with code rate 3/4 up to 64, erasures can be corrected per row.

For transmission, the MPE-FEC frame is divided into sections. An IP datagram forms the payload of an MPE section, and an RS redundancy column forms the payload of an MPE-FEC section. The MPE sections are transmitted first, followed by the MPE-FEC sections. Both are transmitted in a moving picture experts group-2 (MPEG-2) transport stream (TS) format [7].

Time-slicing is applied to enable power saving, so that one MPE-FEC frame is transmitted in one time-slice burst. The TS bitrate during the burst is significantly higher than the service bitrate, and the receiver can turn off its radio parts between the bursts to save power. The frame size, transmission bitrate, and offtime between bursts are parameters that affect the video bitrate, service switching time, and power saving. That is, with an IP bitrate of 384 kilobits per second (Kb/s), one 512-row frame contains 1.8 seconds and a 1024-row frame 3.6 seconds of video.

DVB-H contains a large set of network and service-inde-pendent parameters. In addition to the link layer operation
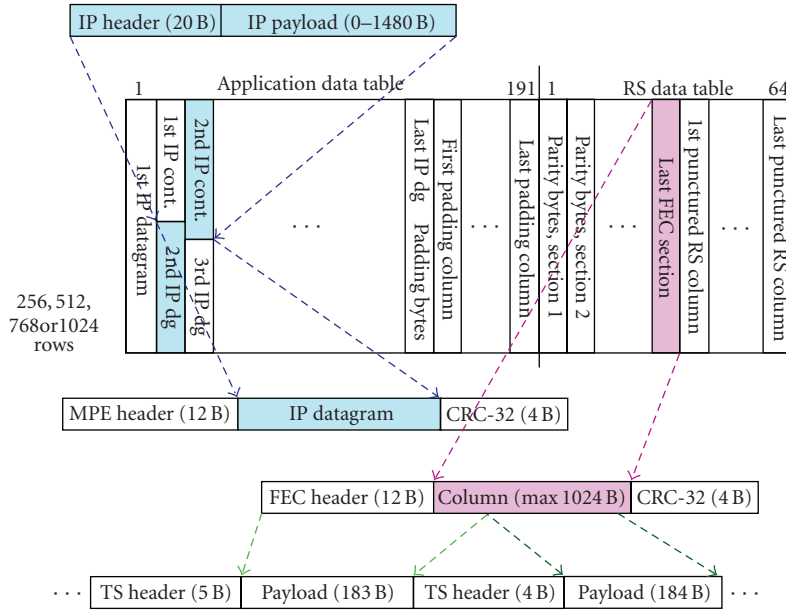
FIGURE 1: The DVB-H link layer operations.

described here, there are a set of physical layer parameters, such as modulation, code rate, guard interval length, and orthogonal frequency-division multiplexing (OFDM) mode. With such a large set of options, simulations are usually the most efficient way to find the optimal parameter combinations.

3.2. Current DVB-H error criteria

The DVB-T standard specifies the *C/N* threshold needed to reach the quasierror-free (QEF) reception criterion, which means one uncorrected error event per hour. Due to the high variations occurring in a mobile channel, the QEF criterion is not suitable for instantaneous measurements for mobile broadcasting. Also, in mobile broadcasting, looser error criteria have been accepted than for fixed reception. The common error criterion for DVB-H has been defined as MPE-FEC frame error ratio (MFER), and the quality of restitution (QoR) limit has been set to MFER 5% [6]. In addition to MFER, the erroneous seconds ratio (ESR) criterion has been occasionally used in some measurements. ESR is defined as seconds with errors over the observation period [6].

The MFER error criterion enables instantaneous laboratory measurements. The length of one measurement has usually been 100 frames, of which 5 can be erroneous. Further, the service bitrate has been increased, that is, the offperiod has been shortened, to enable faster measurements. Still, it is a highly time consuming project to perform extensive DVB-H measurements, including all possible combinations of constellations, fast fourier transform (FFT) sizes, guard intervals, code rates, and burst lengths covering pedestrian and vehicular use cases. According to [6], the observation period for field trials has been reduced to one time interval,

corresponding to one time-slice burst, as the QoR assessment should be instantaneous.

The MPE-FEC frame error criterion is too inexact to evaluate the impact of the channel and system parameters on subjective audiovisual quality. Optimizing the system parameters using only the MFER, 5% criterion might even be misleading and result in incorrect conclusions about the system performance. As systems are also designed, optimized, and verified using simulations or postprocessing of recorded traces from laboratory measurements or field trials, particular IP packet or even on byte level information can be received. There is definitely a need for more accurate error criteria than frame error-based measures.

3.3. Selection of DVB-H transmission parameters

The DVB-H implementation guidelines [8] give recommendations for parameter selections in DVB-H networks. For the physical layer modulation and code rates quadrature phase-shift keying (QPSK) or quadrature amplitude modulation (16-QAM) with code rates 1/2 or 2/3 are recommended. The choice is a compromise between robustness to transmission errors and throughput bitrate. QPSK 1/2 gives a bitrate of 5 Mbps, whereas 16-QAM 1/2 gives a bitrate of 10 Mbps, using guard interval 1/4 of the OFDM symbol duration. [8] recommends the use of 16-QAM 1/2 or 16-QAM 2/3 for mobile and portable reception.

The selection of FFT mode is based on the expected maximum velocity of the receiver. The 8K FFT mode, which is used in most DVB-T networks, gives the largest coverage area, but provides the lowest receiver velocities compared to 2 K and 4 K. Based on [8], when MPE-FEC is used and DVB-H physical layer parameters are selected properly, the use of the 8K mode is feasible at speeds up to 120 km/h.

Transport
stream

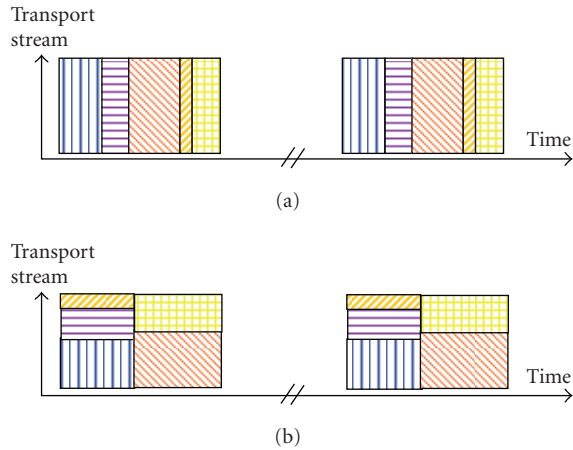(a)

Transport
stream

(b)

FIGURE 2: Consecutive (a) and parallel (b) transmission of different DVB-H services.

The selection of guard interval is based on network topology. For the 8K mode, guard intervals 1/4 or 1/8 are recommended, of which 1/4 tolerates longer single-frequency network (SFN) delays.

Simulations in [9] used several different channel models for DVB-H and showed that, for networks intended primarily for vehicular use, the preferable combinations of modulation, convolutional code rate, and MPE-FEC code rate would, respectively, be QPSK 1/2 3/4, QPSK 1/2 5/6, QPSK 2/3 5/6, 16-QAM 1/2 3/4, or 16-QAM 1/2 5/6. Based on the recommendations and results in [8, 9], the parameters used for evaluating the performance at IP level in Sections 4 and 5 were chosen to be 16-QAM 1/2 3/4, FFT size 8K, and guard interval 1/4. Additionally, in some presented comparisons MPE-FEC is not used, that is, the MPE-FEC code rate is then 1.

When the transmission network is optimized properly, the transmission parameters do not have a direct impact on the video quality but on the size of the coverage area and the capacity of the network. On the other hand, transmission parameters, multiplexing scheme, environment, and movement of the receiver will affect the length and amount of error bursts. In general, when the receiver moves slowly, that is, the channel changes slowly, the error bursts are longer, as the receiver stays in the area with bad reception for a longer time compared to a fast changing channel.

3.4. Multiplexing of services in DVB-H systems

DVB-H services may be transmitted consecutively or in parallel. Consecutive transmission means that only one MPE-FEC frame carrying one service is on air at a time. [8] does not present parallel transmission of services as the main but suggests that IP encapsulators and receivers should support this mode of transmission. Examples of consecutive and parallel transmission of DVB-H services are depicted in Figure 2, where each fill pattern represents one MPE-FEC frame carrying one service.

FIGURE 3: Measurements for evaluating video quality [10].

Parallel transmission can be useful if the service bitrates are very low. Using consecutive transmission in short bursts leads to degradation in time diversity. In mobile transmission, a good choice of burst length would be more than 100 milliseconds. Consecutive transmission, on the other hard, is the main source for the power saving in receivers achieved in DVB-H when compared to continuous parallel transmission of all services.

A special case of transmission would be to transmit several services in every MPE-FEC frame. This could be preferred, for example, if the services are statistically multiplexed together, so that the total capacity of these services is constant. This scheme was utilized in [5] and thus in the results presented in Section 6. With this transmission format, the MFER error criterion becomes even less accurate. An MPE-FEC frame might contain errors after decoding that do not occur in the MPE-sections carrying the data from the wanted service. Thus, the received data could be error-free even if the errors in the MPE-FEC frame cannot be corrected.

4. VISIBILITY OF PACKET LOSS IN MPEG-2 AND H.264/AVC VIDEO

Reibman and Kanumuri et al. have studied the visibility of packet loss in MPEG-2 and H.246/AVC in many papers, for example, in [10–13]. In [10], the need for accurate video quality measures is explained in detail. The approach is similar as in this paper. Figure 3 illustrates three measurement points discussed in [10]. Measurement C corresponds to the transmitted bitstream itself and could be taken either at the input to the decoder or inside the network. Measurements in C assume the use of nonreference methods, as the original video is not available for comparison. The new error criteria for mobile broadcasting of streaming audiovisual services considered in this paper should similarly be nonreference video quality measures in point C. However, assumptions about video coding parameters and used concealment algorithms have a significant impact on the perceived quality.

When measuring network performance and error behavior, it is usually preferred to measure over the whole multiplex, that is, over all service. This is the conventional use of the MFER criterion in laboratory and field measurements. However, the subjectively perceived quality can only be measured over one service. This problem has also been recognized by Reibman. The goal in [11] was to have a method to predict the quality of individual videos with low-enough complexity that it can be easily applied to many different video streams being sent across the network. Similarly, when designing the new criteria for mobile broadcasting, we need to move away from the approach of error measures for the whole multiplex. Measuring service specific quality is especially important in time division multiplex (TDM) systems, such as DVB-H, as the packet loss in

mobile channels is strongly time variant. Thus, the different services might experience very different error behavior. This is discussed further in Section 5.

The previous work on visibility of packet loss can partly be used for designing new criteria for mobile broadcasting. Still, the approach in [10–13] has been different from the assumptions that have to be made for mobile broadcasting. In the mobile environment, errors will always exist. More important than finding the limit for visibility of packet loss or errors is to find the limit for acceptability of errors. Further, we must make the assumptions of using the simplest receiver, which is described in the implementation guidelines [8], and the simplest decoder. This also includes the assumption that concealment algorithms are not used, and the length of an error in the video cumulates to the next nonpredicted frame (IDR frame in the case of H.264/AVC).

5. DIVERSE ANALYSES OF MFER AND OTHER CONVENTIONAL ERROR CRITERIA

In this section, the MFER criterion is analyzed both from the transmission system and video codec perspectives. The ambiguous character of the MFER measure is demonstrated by analyzing it together with two transmission error criteria, namely, IP packet error ratio and byte error ratio, and two objective video quality metrics, namely, peak signal-to-noise ratio (PSNR) and the national telecommunications and information administration (NTIA) video quality metric. In Section 5.1, the transmission system simulation setup is described, and the results are presented in Section 5.2. In Section 5.3, the IP error statistics are analyzed close to the limit for subjectively acceptable quality. The objective video quality analyses are presented in Section 5.4 and, the shortcomings of the MFER criterion are analyzed in Section 5.5.

5.1. Simulations on different MPE-FEC decoding strategies

Different MPE-FEC decoding strategies for DVB-H were presented and analyzed by the author in [14, 15]. The decoding method suggested in the DVB-H standard is referred to as section erasure (SE) decoding. An MPE section or an MPE-FEC section is marked as an erasure, if it contains an error, and discarded in the decoding process. SE decoding provides neither efficient MPE-FEC decoding nor video decoding, as a lot of correct data is dropped at the link layer. However, using SE decoding is optional, and the final decision on the decoding strategy is left to the receiver designer. The most efficient of the suggested decoding methods is hierarchical transport stream decoding (HTS), which uses three levels of erasure information: correctly received TS packets, erroneous TS packets, and lost TS packets. HTS provides very good byte-level error performance.

To evaluate the performance of the different decoding strategies, simulations were carried out in the channel models developed for DVB-H [16] similarly as in [17]. The used models are pedestrian outdoor (PO), vehicular urban
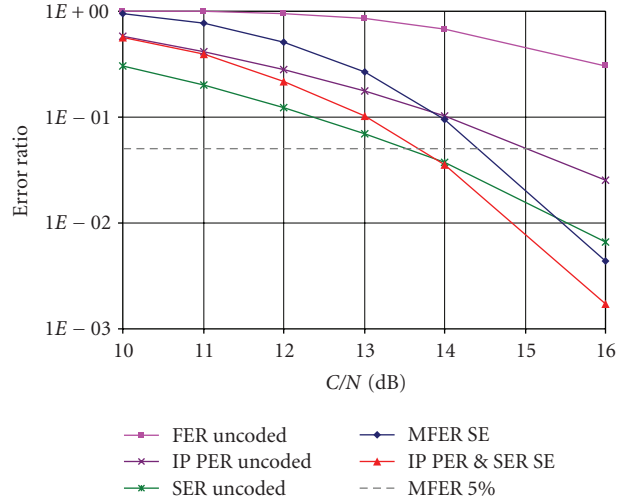

FIGURE 4: MPE-FEC frame error rates (MFER), IP packet error rates (IP PER) and byte error rates (SER) after coded, and uncoded data link layer for the Vehicular Urban channel.

(VU) and motorway mural (MR), corresponding to the velocities of 3 km/h, 30 km/h, and 100 km/h, respectively. The physical layer parameters were 16-QAM modulation with convolutional code rate 1/2, 8K OFDM mode, and guard interval duration 1/4 of the OFDM symbol duration. Error traces from the physical layer were established to allow fast simulations at transport stream packet or byte levels. Error traces are series of binary indicators expressing whether a data block contains errors, in this case after the physical layer error correction decoding. The simulated link layer parameters were as follows: MPE-FEC code rate was 3/4 or 1, 512 rows were present in MPE-FEC frames, and an IP packet of length was 512 bytes. The error rates were measured over all services, that is, over the whole transport stream. The services were multiplexed so that one service always uses the whole bandwidth for transmitting the time-slicing bursts. The results are presented in Section 5.2.

5.2. Frame, packet and byte error ratios

Figure 4 illustrates different error ratios using SE decoding or uncoded DVB-H link layer (for which MPE-FEC code rate is equal to 1) in the Vehicular Urban channel, corresponding to a velocity of 30 km/h. The frame error ratio for uncoded link layer data (FER uncoded) is above 30% for all simulated carrier-to-noise ratios (C/N). Yet, when studying IP packet error ratio (IP PER) and byte or symbol error ratio (SER) for uncoded data, it is seen that there is much more correct data than the frame error ratio implies. When comparing IP PER for SE and uncoded, the difference of C/N yielding the same IP PER is only 1.3 dB. When designing the system for the presented C/N values based on frame error ratio, MPE-FEC code rate 1 could have been discarded from list of good parameter options.

However, when defining the system parameters based on another error criterion, uncoded link layer could be

TABLE 1: Carrier-to-noise ratios, IP packet error ratios, and byte error ratios at MFER 5% in the different channels.

		Pedestrian outdoor	Vehicular urban	Motorway rural
C/N at MFER 5%	SE	13.1 dB	14.3 dB	14.6 dB
	HTS	12.8 dB	13.9 dB	13.2 dB
IP PER & SER	SE	4.0%	2.2%	1.5%
IP PER	HTS	4.8%	2.2%	2.0%
SER	HTS	1.6%	0.6%	0.2%

a possible choice, as less redundancy is needed. Previous work has shown that good transmission modes also can be found among those not using MPE-FEC coding. In [17], different modulation and code rates were compared based on the IP PER 1% criterion, using SE decoding for all link layer code rates in the PO, VU, and MR channels. When also considering the different service bitrates achieved using different code rates, uncoded link layer was included in the list of good modes. For the PO channel, the uncoded mode was even recommended. If MFER 5% had been used in this comparison, the conclusions would have been very different.

Table 1 demonstrates the ambiguity of the MFER 5% criterion. The *C/N* required for achieving the MFER 5% point is given for SE and HTS decoding with MPE-FEC code rate 3/4. Other simulation parameters were similar as for the simulations in Figure 4. The IP packet error ratios and byte error ratios were measured at the MFER 5% point. For SE decoding, IP PER and SER give the same results, as with SE decoding all bytes of an erroneous IP packet are erased, which is not the case with HTS decoding. As HTS decoding provides low-byte error ratios, the SER at MFER 5% is very low compared to SE decoding, especially in the Vehicular Urban and Motorway Rural channels. The error ratios also demonstrate the effect of the receiver velocity. At high velocities, an erroneous frame contains less erroneous data than at low velocities. This is mainly due to the fact that error bursts are shorter at high velocities, as the channel changes faster. At high velocities, the amount of errors at the MFER 5% point is different from the error amounts at low velocities. The same also applies to the length and frequency of the error bursts.

The amounts of erasures occurring in the MPE-FEC frames are illustrated for the different channel models in Figure 5, where the distribution of instantaneous IP PER values for each frame is given. The curves represent the situation, where average IP PER is 10%, when MPE-FEC coding is not utilized (uncoded). The figure shows significant differences in error distributions between the different channel models. The curve of the pedestrian model is very steep, whereas for vehicular speeds, there is a large amount of frames with less than 25% of the IP packets erased. Using MPE-FEC code rate 3/4, all frames with IP PER less than 25% would be corrected. The different distribution of errors leads to different MPE-FEC decoding performance even though the average IP PER over all frames is equal.
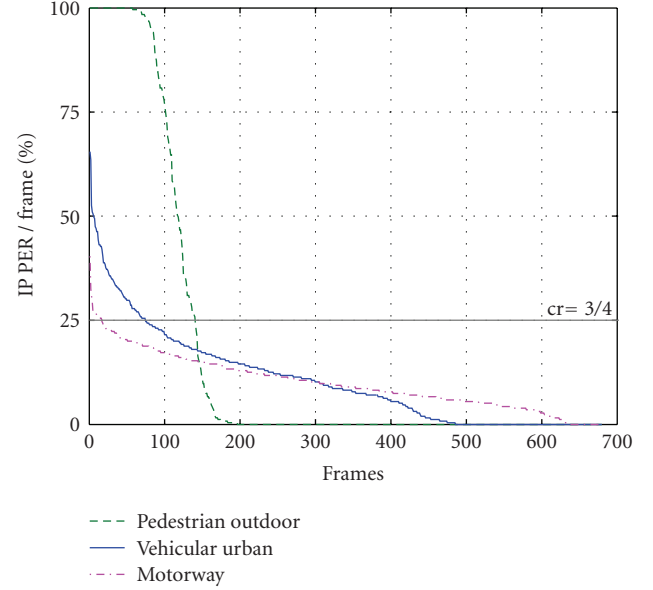

FIGURE 5: IP packet error ratio for each MPE-FEC frame in different channel conditions [17].

5.3. IP error statistics in three different channels at the limits for subjective quality

Some results for subjective audiovisual assessment in DVB-H are available in [5], aiming to discover the approximate value of MFER that is the threshold between subjectively acceptable and unacceptable audiovisual quality. Extensive subjective testing was carried out with four clips of different content types coded according to the lowest interoperability point specified for IP data casting over DVB-H at time-slice interval of about 1.5 seconds [5]. It was concluded that with the tested clips, the boundary of acceptability and unacceptability lies between 6.9% and 13.8% in terms of MFER.

Let us now compare the IP error statistics for the simulated channels with the results from the subjective tests [5]. In Table 2, the IP PERs for MFER 6.9% and 13.8% are presented for MPE-FEC code rate 3/4. As above, the IP packet length was constant 512 bytes. Compared to the VU channel, the MR channel has only slightly lower IP PERs at these MFERs, whereas the PO channel has double the amount of errors.

When measuring the performance of a transmission system, the measurements are performed over the whole transport stream, whereas in subjective quality measurements, the results are gathered for a single service. To enable comparison to subjective tests results in Section 6, a 60-second measurement over the whole multiplex is performed. With the used modulation and coding, this corresponds to transmitting 58 video services of capability class A at 128 Kbps or 29 video services of capability class B at 384 Kbps (see Table 4).

In Table 3, comparisons of the IP packet error characteristics of the channels are presented with MPE-FEC code rate 3/4. It is found that the MR channel has shorter error bursts

Table 2: IP PER at MFER 6.9% and 13.8% in three different channels.

MFER	IP PER in pedestrian outdoor	IP PER in vehicular urban	IP PER in motorway rural
6.9%	5.5%	2.5%	2.0%
13.8%	11.5%	5.4%	4.4%

Table 3: IP error statistics for three different channels measured over 60 seconds.

	Pedestrian outdoor		Vehicular urban	Motorway rural
C/N	12 dB	13 dB	14 dB	14 dB
MFER	19.1%	5.7%	10.9%	13.3%
IP PER	18.5%	5.3%	4.0%	4.4%
Amount of IP errors	21830	6236	4704	5134
Average error burst (AEBL)	23.05	19.01	23.76	15.1
StDev of error bursts	115.76	85.18	29.24	10.94
Max error burst	1262	716	147	71
Amount of error bursts	947	328	198	340
Error bursts > 80 packets	24	11	11	0

Table 4: Capability classes for DVB-H [18].

Capability	Frame size [pixel]	Frame rate [Hz]	Max bitrate [Kb/s]
A	QCIF: 176×144	15	128
B	QCIF: 176×144	30	
	QVGA: 320×240	15	384
C	QVGA: 320×240	30	768

than VU at higher MFERs and IP PERs. This indicates that in the MR channel there are more but shorter error bursts. Also, in the PO channel, the average error burst is shorter for a higher error rate than in the VU channel. However, in the PO channel, there are also longer errors than in the VU channel. The variation in length of the error bursts is much larger in the PO channel, whereas for the VU channel, the error lengths are closer to the average. The comparison shows that the error characteristics are very different in different channels, when studying error rates close to the limit for subjectively accepted video quality.

5.4. Objective video quality measurements

The MFER 5%, as an error criterion, can introduce errors of very different lengths and severity to the video stream. To understand and measure these errors better, a set of simulations and objective measurements was performed. The video used was a 180-second clip, corresponding to 100 MPE-FEC frames, recorded from a TV news broadcast. The content was comparable to a typical news broadcast, including low or no motion scenes showing the newsman or generated graphics and high-motion material from different reporting locations. Resolution, frame rate, and bitrate were chosen to be 320×240, 15 Hz, and 384 Kb/s, respectively. The bitrate for the video stream included header overhead, the actual VCL bitrate being 353 Kb/s. No audio track was used for the content.

Video encoding was performed using Nokia H.264 encoder [19] with default settings, except for resolution, frame rate, and bitrate control. Error concealment was not used, as it is an optional feature for DVB-H services. IDR frames were inserted every 1.8 seconds, corresponding to at least one IDR frame in each MPE-FEC frame. The resulting NAL units were encapsulated to IP packets, achieving an average IP packet length of 512 bytes. These IP packets were then inserted into 100 MPE-FEC frames, using 191 application data columns and 512 rows. Corruption was introduced into 5 of the 100 frames using section erasure with IP PER values of 0.026%, 1.7%, and 5.0%, corresponding to the loss of 1, 65 and 191 IP packets per each erroneous MPE-FEC frame. The MPE-FEC frames were decoded using SE decoding. When using code rate 3/4 and the IP packet lengths being equal to the amount of rows in the frame, these amounts represent some extreme cases of residual errors in the MPE-FEC frame. 191 erased IP packets correspond to one completely corrupted MPE-FEC frame. 65 erased IP packets corresponds to the smallest amount of erasures that cannot be corrected with code rate 3/4, when all erased sections are carrying application data. The loss of one IP packet occurs, if all 64 RS redundancy columns are erased and one application column.

Video quality was assessed using three metrics. Despite its drawbacks, PSNR was used as a primary comparison metric due to its ability to provide results for individual video frames. Secondary metric used was the NTIA VQM [20], which is far more complex than PSNR. NTIA VQM tries to account for, for example, jerky motion, blocking, blurring, and other impairments typical to digital video and has been shown to correlate with subjective measurements very well. The third metric used was erroneous seconds ratio (ESR). A second (15 frames) of video was considered to be erroneous if it contained more than 3 successive visibly erroneous frames, corresponding to 200 milliseconds detection threshold [21]. A PSNR difference of 1 dB was considered as error visibility threshold in error assessment.

Average results obtained from the PSNR metric seem to degrade linearly as the IP PER rises. However, profound conclusions should not be drawn from the PSNR scores due to the drawbacks mentioned in Section 2. The NTIA VQM scores seem to indicate that on average, the video quality is
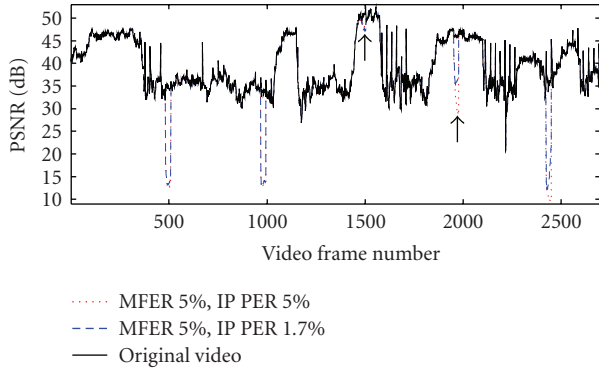
FIGURE 6: PSNR video quality results in MFER 5% with 1.7% and 5.0% IP packet error ratios.

TABLE 5: Video quality measurement results at MFER 5% at different IP packet error ratios.

IP PER	PSNR [dB]	NTIA VQM	ESR	Average lost picture frames per error burst	Average error length per error burst
0.0%	36.61	0.206	0.0%	—	—
0.026%	35.60	0.208	2.2%	0	0.8 s
1.7%	29.00	0.224	5.0%	6.8	1.8 s
5.0%	27.34	0.227	5.0%	27.0	1.8 s

acceptable in all test cases in Table 2. Acceptability threshold for NTIA VQM is around 0.5 [20], corresponding to the border of "fair" and "poor" quality (lower score is better). Despite the good VQM results, the erroneous seconds ratio (ESR) for both the 1.7% and 5.0% IP PER exactly meet the ESR 5% criterion, which is considered to be the limit for acceptable quality in [6]. This can be explained by the ESR metric not accounting for severity of the errors. Errors are clearly longer than the amounts of dropped frames indicate, mostly due to error propagation and the rather sparse placement of the IDR frames. In any case, it seems evident that MFER 5% does not provide an unambiguous error criterion compared to other metrics.

Detailed PSNR results for the 1.7% and 5.0% IP PER simulations are depicted in Figure 6. In addition, the PSNR curve of the error-free video is provided for comparison. These results are derived from the same simulations as the average values in Table 5. Five error bursts and their corresponding drops in terms of PSNR are clearly visible in the figure. Error bursts that occurred during low or no motion scenes, pointed out with arrows, have a significantly smaller quality drop. The result is logical, since losing frames from relatively static content produces only barely, if at all, visible errors. The remaining three error bursts coincide with a high-motion scene, resulting in extremely low-PSNR values, typically 10–15 dB. Such low values result from the dropped frames and do not provide basis for a meaningful comparison as such. Regardless, it is evident that loss of frames in a high-motion scene is critical for the perceived video quality. Due to the low similarity of successive frames in this type of content, a significant amount of information is lost in each burst. It is also notable that with 1.7% IP PER, the PSNR value has a tendency to rise after the initial drop at the start of each error burst. However, error propagation will continue impairing the video until the next nonerroneous IDR frame is encountered, and the video quality returns to optimal levels.

5.5. The shortcomings of the MFER criterion

MFER fails to express many characteristics that would be important for DVB-H system design, some of which are described in the following. First, MFER does not indicate the relation between the frequency of the errors and their duration. For example, MFER equal to 5% corresponds to one and six erroneous time-slice bursts per minute in streams with 3-second and half-a-second time-slice intervals, respectively. It is not obvious how the frequency and duration of clearly perceivable audiovisual errors impact the subjective quality. Second, MFER does not indicate the residual error rate affecting the content of the erroneous frames. For example, the same value of MFER can result from two different error conditions of very different symbol error rates due to different code rate in MPE-FEC. Audio and video decoders may be able to conceal a relatively small residual error rate satisfactorily, but when it exceeds a threshold, most viewers consider the audiovisual quality as unacceptable regardless of the residual error rate. Third, the distribution of residual errors may play a role in subjective quality. For example, an error burst may not affect the entire time-slice, but the start or the end of the time-slice may be intact. Moreover, the method for transmission can affect the distribution of residual errors. One example is provided in [22], where unequal error protection has been proposed to protect audio, video IDR pictures, and other reference pictures more strongly compared to nonreference pictures. Fourth, the operation of the protocol stack and source decoders may be optimized differently in receiver operations when it comes to handling of transmission errors. For example, some DVB-H receivers may implement the HTS method, while others use the SE decoding. Furthermore, error concealment algorithms have not been specified in audio and video codec specifications, hence resulting into different implementations in source decoders.

In broadcasting, error criteria have been conventionally defined as accepted error events during a certain time. In DVB-T, the accepted limit for quasierror-free reception is one erroneous event per hour. Due to low-transmission error rate and common structures for groups of pictures in which intra-coded pictures are periodically and frequently included, the measure of error events per time is sufficient enough in DVB-T. In mobile broadcasting, varying reception conditions and wider range of possibilities for error protection code rates, time-slicing intervals, and group of picture structures make the measure of error events during a certain time unsatisfactory.

In the third generation partnership project (3GPP), some objective quality of experience metrics have been specified [23]. Burst errors are measured using a corruption duration metric, indicating the amount of successive corrupted pictures and successive loss of IP packets. However, the relation of these metrics to subjective quality has not been quantified. Moreover, no numerical limits for these quality metrics have been defined in 3GPP.

6. COMPARISON TO SUBJECTIVE ACCEPTANCE OF AUDIOVISUAL QUALITY

The subjectively perceived audiovisual quality of TV services over DVB-H has been studied in [5, 24]. In these studies, the error patterns used to simulate errors caused by the wireless channel were achieved by using channel characteristics from field measurements in a Gilbert-Elliot model. The results can be compared to the vehicular urban (VU) channel model used in this paper, as the field tests were carried out in a similar environment with a car rooftop antenna. The MPE-FEC code rate was 3/4. QCIF videos were coded with an H.264/AVC encoder at bitrate 128 Kbps and at a frame rate of 12.5 Hz. One IDR picture was encoded per each time-slicing burst. Monaural audio at 32 Kbps and 16 Hz sampling frequency was used. No error concealment was used in the tests. The limit for acceptable and unacceptable audiovisual quality was found to be between MFER 6.9% and 13.8% [24]. There were 30 evaluators in the tests, and each clip was played three times, varying the error locations in the audiovisual stream. The length of the clip was around 60 seconds.

Figures 7 and 8 present the average error length and amount of error bursts in the video and audio streams for the tests in [5] for all tested content types: news, sports, music video, and animation. Each point corresponds to one test case, a combination of the content type, and error trace, rated by all evaluators. The filled (solid) points for MFER 1.7% and 6.9% represent acceptable quality, and the unfilled (hollow) points for MFER 13.8% and 20.8% represent unacceptable quality. It seems that the acceptability is more based on the amount of errors than the duration of these. The limit for acceptability of video is between 4 and 6 errors, and for audio between 5 and 7 errors on the average with the used content and parameters.

As explained in Section 3.4, each service should be carried in its own MPE-FEC frame to achieve maximum power saving in receivers rather than transmitting several services in each MPE-FEC frame as in [5]. This means that the used service specific error traces should not be considered to represent conventional DVB-H services. The used multiplexing has probably also caused the surprising error lengths, where the lowest MFER gives the longest errors. What can be used are the ratings and classification into acceptable and unacceptable quality of the different contents with the different amount and duration of errors, as in Figures 7 and 8. Still, new subjective tests are required to fully understand the acceptability of typical error behavior in mobile and portable channels with different encoding parameters, bitrates, and content types. The requirements for the future subjective tests are described in Section 7.3.
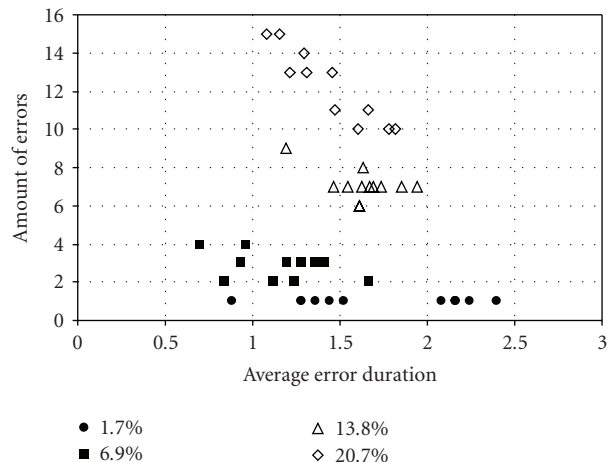

FIGURE 7: Video errors for MFER 1.7%, 6.9%, 13.8%, and 20.7% for the test performed in [5].
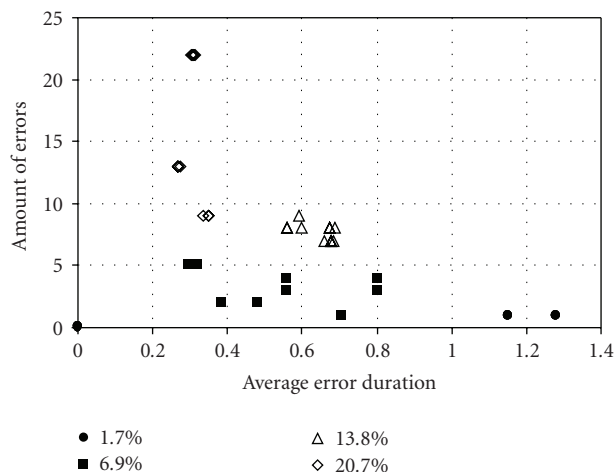

FIGURE 8: Audio errors for MFER 1.7%, 6.9%, 13.8%, and 20.7% for the test performed in [5].

7. DESIGNING THE NEW ERROR CRITERIA

As described in the previous sections, the MPE-FEC frame error ratio criterion does not provide sufficient means for system design and optimization of DVB-H. There is a need for more appropriate error criteria that would represent the subjective impact of transmission errors on the services and applications. Many challenges in defining such criteria relate to the difficulty to derive an objective measure reflecting the subjective experience of audiovisual content, as the expectation for the experience and the relative weight of audio and video elements depend on the content. Still, the error criteria should be easy to measure, using tools familiar to transmission system designers.

7.1. Transmission system aspects

The performance of DVB-H in different channel models and use cases measured in the laboratory and in the field were

compared in [25]. Five parameters for comparing packet channel characteristics were presented in [26] as follows.

(1) *Packet error ratio (PER).*

(2) *Average error burst length (AEBL).* The AEBL parameter describes the average length of all error bursts. The error burst length is defined as the amount of consecutive erroneous units, that is, amount of erroneous packets between two correctly received packets.

(3) *Variance of error burst lengths (VEBL).*

(4) *Mean time between errors (MTBE).* The MTBE parameter describes the average length of the time between errors. The time between errors is defined as consecutive correctly received units, that is, amount of correctly received packets between two erroneously received packets.

(5) Variance of time between errors (VTBE).

These parameters have shown to successfully model packet error behavior in packet channels with constant length packets. For streaming audiovisual services, the IP packets are usually of variable lengths. In the next comparison, a constant IP packet length of 512 bytes has been assumed to enable IP PER comparisons, as in the previously presented simulations.

To illustrate that the error behavior is service specific, the AEBL, MFER, IP PER, and TS PER are shown for a complete multiplex and for 16 services separately. The laboratory and field measurements are the same as used in [25] with 16-QAM modulation and convolutional code rate 1/2. The MPE-FEC code rate was 3/4. The multiplex of 9.95 Mbps was carrying 16 equally multiplexed services, each with a bitrate of 622 Kbps at TS level. The error behavior was measured over a stream corresponding to transmission time of 10 minutes.

In Figure 9, the AEBL at TS level is shown for the whole multiplex "All," the average AEBL over all 16 services "Mean", and for each service separately. In all cases, the TS PER over all services is 4-5%. The simulations show that the error behavior is service specific and varies most in the field. In Figure 10, the MFER, IP PER, and TS PER are shown similarly for the TU6 15 Hz channel at $C/N = 15$ dB, giving an average MFER closest to the area for acceptability in [5] that is, MFER 6.9–13.8%. Surprisingly, the MFER varies more than the TS PER and IP PER. Also, these measures are service specific, although measuring over the whole multiplex gives a fairly good approximation of the service specific TS PER and IP PER.

It is expected that the new objective criteria from the transmission system point of view should be designed as follows.

(i) The five above mentioned parameters should be used for studying service specific error characteristics at TS level. The values for the parameters should be derived from currently used channel models for mobile broadcasting, such as PO, VU, MR, and TU6.
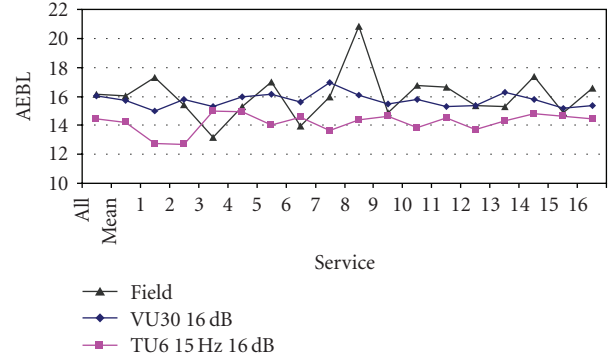

FIGURE 9: Average error burst length at TS level for the VU channel and TU6 15 Hz channels at $C/N = 16$ dB and in the field in a vehicular urban use case.
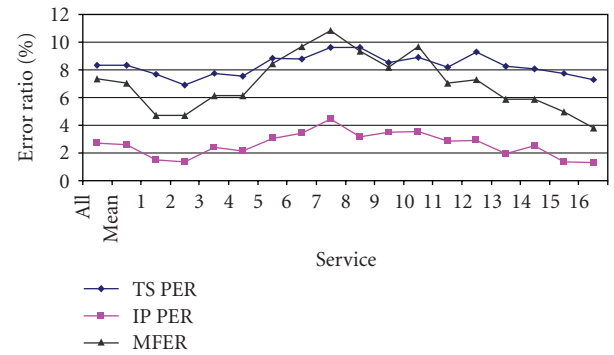

FIGURE 10: MFER, IP PER, and TS PER for TU6 15 Hz channel at $C/N = 15$ dB for 16-QAM 1/2 3/4.

(ii) The effect of the MPE-FEC code rate in different channels should be studied to understand the error behavior at IP level.

(iii) The parameters should be mapped to results from future subjective test described in Section 7.3.

It was concluded in [25] that both VU and TU6 15 Hz are good choices for channel models, when modeling the vehicular use case in an urban environment. If designing subjective test cases based on the laboratory measurements in [25], the C/N values of 14 dB and 15 dB in the VU or TU6 15 Hz channels could be good starting points. The error statistics at TS level with IP PER and MFER are given in Table 6. Based on the results from [5], $C/N = 14$ dB is expected to give unacceptable quality, and $C/N = 15$ dB is expected to give acceptable quality with similar contents as in [5]. C/N points with similar error ratios in the PO channel should also be tested.

7.2. The impact of the decoders

One of the challenges in the task of specifying error criteria is the fact that the same transmission error may be concealed differently by audio and video decoder implementations. In a conservative approach, the simplest error-robust audio and video decoder implementations are considered. It can

TABLE 6: Error statistics for VU and TU6 15 Hz channels at *C/N* 14 dB and 15 dB.

Channel model	VU	VU	TU6 15 Hz	TU6 15 Hz
C/N [dB]	14	15	14	15
TS PER	16.69%	9.18%	15.41%	8.31%
AEBL	23.8	18.0	19.4	18.2
VEBL	2722	1409	1943	1395
MTBE	118.5	178.5	106.2	200.4
VTBE	77827	174824	69045	209196
IP PER	11.53%	3.18%	9.33%	2.71%
MFER	27.96%	9.43%	22.68%	7.35%

be assumed that these error-robust decoders do not crash or halt under any error conditions and are able to receive information on lost packets or detect lost data themselves. The simplest error-robust audio decoder replaces missing audio frames with silent frames. The simplest error-robust video decoder replaces missing or corrupted pictures with the previous correct decoded picture in presentation order. Furthermore, the simplest error-robust video decoder is capable of detecting whether errors occurred in nonreference or reference pictures. If an error occurred only in nonreference pictures, decoding continues from the next correctly received coded picture. If an error occurred in a reference picture, decoding continues from the next correctly received IDR picture.

Error criteria specified according the simplest error-robust decoders above might produce too conservative results for sophisticated decoder implementations, which may be able to conceal errors successfully. For example, an audio frame may be successfully interpolated from temporally adjacent audio frames if those frames are well correlated. However, concluding whether error concealment operates sufficiently well is a challenging problem. One approach is to include auxiliary error concealment information into the audiovisual streams indicating the most efficient error concealment methods and the quality they are able to obtain. For example, the spare picture supplemental enhancement information message of H.264/AVC indicates which colocated areas in the indicated set of pictures are essentially unchanged so that any of those decoded areas can be used for concealing the corresponding area in an erroneously received coded picture.

7.3. Subjective tests

The average length and amount of errors presented in Figures 7 and 8 can be divided into groups, where the difference between points for acceptable and unacceptable quality is clearly distinguishable. Finding the limits for acceptability by means of subjective testing and understanding what kind of transmission errors cause such error behavior is in the focus, when designing new objective error criteria. Also, understanding the length and frequency of errors on the perceived quality is necessary, when translating the subjective quality measures into objective numerical measures. This

will require subjective tests similar to those performed in [5], where the impact of the average amount of errors and average error length are studied.

For consistency, the error information used in the subjective test should be based on error statistics or traces from current mobile radio channel models, as described in Section 7.1. The choice of content and audio and video coding parameters also play significant roles. The encoding and relation between audio and video will represent typical DVB-H service parameters. Probable IP level bitrates with current network parameters in DVB-H are between 300 Kbps and 768 Kbps, corresponding to about 25 and 10 services, respectively, with 16-QAM 1/2 3/4. These correspond to capability classes B and C in Table 4. The quality of the encoded video should be rated acceptable, preferably without visible errors.

The bitrates and contents should be divided into different groups. For example testing the four content types in [5] news, animation, music video, and sports with three different bitrates, for example, 300 Kbps, 500 Kbps, and 700 Kbps, give us 12 different test streams. The content types used in [5] correspond well to findings from user tests on mobile TV content [27]. Applying two *C/N* points for the VU channel and two for the PO channel gives four different error traces. Further, the error streams and the test clips should be matched in different ways so that the errors occur in different parts of the content. In [5], three different ways of matching each test clip and error trace were tested. Alternatively, the error traces could be chosen so that they represent different services in plots such as Figures 9 and 10, including the service corresponding to the maximum, average, and minimum error ratio. If the subjective ratings for these three cases are similar, we can use the service with the average error ratio to represent the whole multiplex.

After encoding and matching the test clips with the error traces, before running the subjective tests, it should be ensured that the test cases represent different points in similar plots as in Figures 7 and 8. Video clips with different bitrates should be treated separately.

8. CONCLUSIONS

Currently, there are no error criteria for mobile broadcasting of streaming audiovisual services that would express all characteristics important for system design and have a verified correlation to subjective perceived quality. Known transmission system error criteria and objective video quality criteria were studied, and the results were compared to results on subjectively audiovisual quality.

In order to find measures for new error criteria to overcome the presented issues, we analyzed the characteristics from the perspective of the transmission system and video codec. We suggest that quality criteria based on average amount and average duration of errors should be defined based on subjective tests of audiovisual content. The error statistics used in the subjective test cases should be derived from conventional mobile radio channel models.

Here, DVB-H, H.264/AVC, and HE AAC v2 were used as an example system and codecs. However, we believe

that the approach can be generalized to other systems and codec designs. Designing new transmission error criteria would be beneficial for developing further understanding of the constraints and degrees of freedom of wireless communication systems for all players in the field.

ACKNOWLEDGMENT

REFERENCES

[1] ETSI EN 302 304, "Digital video broadcasting (DVB); transmission system for handheld terminals (DVB-H)," European Telecommunication Standard, November 2004.

[2] ISO/IEC 14496-3:2005/Amd 2, "Audio Lossless Coding (ALS), new audio profiles and BSAC extensions," 2006.

[3] ITU Rec. H.264/ISO IEC 14996-10 AVC, "Advanced video coding for generic audiovisual services," 2003.

[4] B. Girod and N. Färber, "Wireless video," in *Compressed Video over Networks*, chapter 12, CRC Press, Boca Raton, Fla, USA, 2000.

[5] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and J. Korhonen, "Unacceptability of instantaneous errors in mobile television: from annoying audio to video," in *Proceedings of the 8th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '06)*, pp. 1–8, Helsinki, Finland, September 2006.

[6] G. Faria, J. A. Henriksson, E. Stare, and P. Talmola, "DVB-H: digital broadcast services to handheld devices," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 194–209, 2006.

[7] ISO/IEC 13818-1, "Information technology—generic coding of moving pictures and associated audio information: systems," second edition, December 2000.

[8] DVB-H Implementation Guidelines, v.1.3.1, DVB Document A092 Rev.2, May 2007.

[9] P. Hakala and H. Himmanen, "Evaluation of DVB-H broadcast systems using new radio channel models," Tech. Rep. 817, Turku Center for Computer Science, Turku, Finland, 2007, http://www.tucs.fi/.

[10] A. R. Reibman, S. Kanumuri, V. A. Vaishampayan, and P. C. Cosman, "Visibility of individual packet losses in MPEG-2 video," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, vol. 1, pp. 171–174, Singapore, October 2004.

[11] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 327–334, 2004.

[12] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 341–355, 2006.

[13] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '06)*, pp. 2245–2248, Atlanta, Ga, USA, October 2006.

[14] H. Joki and J. Paavola, "A novel algorithm for decapsulation and decoding of DVB-H link layer forward error correction," in *Proceedings of the IEEE International Conference on Commu-*

nications (ICC '06), vol. 11, pp. 5283–5288, Istanbul, Turkey, June 2006.

[15] J. Paavola, H. Himmanen, T. Jokela, J. Poikonen, and V. Ipatov, "The performance analysis of MPE-FEC decoding methods at the DVB-H link layer for efficient IP packet retrieval," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, part 2, pp. 263–275, 2007.

[16] Wing TV project (EUREKA/Celtic) deliverable D11, "Wing TV Network Issues," August 2006, http://projects.celtic-initiative.org/WING-TV/.

[17] H. Himmanen, T. Jokela, J. Paavola, and V. Ipatov, "Performance analyses of the DVB-H link layer forward error correction," in *Handbook on Mobile Broadcasting*, CRC Press, Boca Raton, Fla, USA, 2008.

[18] ETSI TS 102 005, "Digital video broadcasting (DVB); specification for the use of video and audio coding in DVB service delivered directly over IP protocols," ETSI Technical Specification, April 2006.

[19] Nokia H.264 encoder, ftp://standards.polycom.com/IMTC_Media_Coding_AG/.

[20] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[21] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi, "Sporadic frame dropping impact on quality perception," in *Human Vision and Electronic Imaging IX*, vol. 5292 of *Proceedings of SPIE*, pp. 182–193, San Jose, Calif, USA, January 2004.

[22] V. K. Malamal Vadakital, M. M. Hannuksela, M. Rezaei, and M. Gabbouj, "Method for unequal error protection in DVB-H for mobile television," in *Proceedings of the 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, pp. 1–5, Helsinki, Finland, September 2006.

[23] 3GPP TS 26.234 v6.5.0, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and Codecs (Release 6)," September 2005.

[24] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, M. Liinasuo, and M. M. Hannuksela, "Acceptance of audiovisual quality in erroneous television sequences over a DVB-H channel," in *Proceedings of the 2nd International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM '06)*, pp. 1–5, Scottsdale, Ariz, USA, January 2006.

[25] H. Himmanen, "Studies on channel models and channel characteristics for mobile broadcasting," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–9, Las Vegas, Nev, USA, March-April 2008.

[26] J. Poikonen, "Geometric run length packet channel models applied in DVB-H simulations," in *Proceedings of the 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, pp. 1–5, Helsinki, Finland, September 2006.

[27] D. Schuurman, P. Veevaete, and L. De Marez, "Mobile TV: killer content for the mobile generation," in *Proceedings of the 5th International Conference on Communication and Mass Media*, Athens, Greece, May 2007.