Synchronized Audio Redundancy Coding for Improved Error Resilience in Streaming over DVB-H

Miska M. Hannuksela
Nokia Research Center
Tampere, Finland
miska.hannuksela@nokia .com

Vinod Kumar Malamal Vadakital
Tampere University of Technology
Tampere, Finland
vinod.malamalvadakital@tut.fi

Satu Jumisko-Pyykkö
Tampere University of Technology
Tampere, Finland
satu.jumisko-pyykko@tut.fi

ABSTRACT
Digital Video Broadcasting – Handheld (DVB-H) specification suite enables reception of point-to-multipoint multimedia transmission with mobile devices. It provides good forward error correction (FEC) capabilities to combat transmission errors. However, transmission errors can be occasionally so severe that FEC decoding fails. The residual transmission errors can degrade the audio-visual quality dramatically. This paper presents an error control method, called synchronized audio redundancy coding. It aims to ensure audio continuity under severe channel error conditions. The presented method was compared to conventional FEC-based error protection of DVB-H in a large-scale subjective experiment. It was found that synchronized audio redundancy coding outperformed the conventional error protection in most test cases.

Categories and Subject Descriptors
H.5.1 **[Multimedia Information Systems]**: *Audio input/output.*

H.1.2 **[User/Machine Systems]**: *Human information processing.*

C.2.1 **[Network Architecture and Design]**: *Human information processing.*

General Terms
Algorithms, Measurement, Performance, Experimentation, Human Factors,

Keywords
DVB-H, Audio Redundancy, Error Resiliency.

1. Introduction
Digital video broadcasting-Handhelds (DVB-H) [1] extends the Digital video broadcasting-Terrestrial (DVB-T) [2] standard with methods for improved data delivery to hand-held mobile terminals. The most important improvements include a time-slicing procedure and an optional link-layer forward error correction code (FEC). Time-slicing reduces the average power consumption of the receiving hand-held terminals because data of a service is sent in bursts. This is because a receiver receiving a service is switched on only during the duration of the burst, when the service is transmitted, and goes into a standby mode otherwise. The link-layer FEC, known as Multi-Protocol Encapsulation (MPE) FEC, uses Reed-Solomon (RS) codes. Internet protocol (IP) data bytes over a time-sliced burst are interleaved for the calculation of FEC repair data, thus providing good protection against burst errors.

The inclusion of time-slicing and MPE-FEC greatly improve the performance of data delivery to hand-held mobile devices. However, transmission error bursts can occasionally be so long and intensive that the data remains uncorrected, when practical MPE-FEC code rates, such as 3/4, are used. An uncorrected MPE-FEC frame results into clearly noticeable audio-visual distortions, reducing consumer satisfaction dramatically.

Our previous studies [3] revealed that in many types of content, the continuity of the audio stream is emphasized as a major factor affecting the overall perceived quality of a multiplexed audio-visual service. It is therefore desirable to better preserve audio continuity rather than to simply use MPE-FEC with any additional error control methods.

Packet loss recovery techniques for streaming audio have been extensively reviewed in [4]. According to our experiences, many of the presented techniques in [4], such as audio frame concealment by interpolation, operate satisfactorily only if packet losses occur relatively randomly. From a small scale subjective study, it was found that the residual packet loss rate after unsuccessful MPE-FEC decoding is so high and error concealment by interpolation was inferior to replacing lost audio frames with silence.

Audio redundancy coding is one of the proposed error control methods in [4] and the technique also has a payload format of Real-Time Transport Protocol (RTP) [5]. The method is based on transmitting redundant audio frames for primary audio frames sent in earlier packets. The RTP payload format enables attaching the redundant and primary audio frames into the same packets. If a packet containing a particular primary audio frame is lost, the corresponding redundant audio frame in a subsequent packet can be used to recover the lost packet. However, since redundant audio frames are sent later in time than their corresponding primary frames, the method incurs a delay in the receiver, the duration of which is proportional to the maximum interval between any primary audio frame and the corresponding redundant frame.

Mobimedia'07, Month 8, 2007, Nafpaktos, Aitolokarnania, Greece.

A simple modification to the previously described audio redundant coding is to allow transmission of the redundant audio frame earlier in time than the corresponding primary frame. This transmission arrangement shifts the delay impact from receivers to the transmitter, which is beneficial in broadcast/multicast streaming applications. This method minimizes start-up and channel-switching delays at the receiver. However, due to the high probability of bursty residual packet losses, conventional audio redundancy coding schemes are not expected to improve error robustness in DVB-H remarkably.

In this paper, a scheme for improving audio recovery from transmission errors in DVB-H is presented. The scheme is based on the use of a redundant audio stream, where the redundant audio frames are transmitted in the previous MPE-FEC frame compared to the corresponding primary audio frames. Due to time-sliced transmission applied in DVB-H, this arrangement provides good recovery capability against burst transmission errors.

The rest of the paper is organised as follows. Section 2 briefly discusses the popular audio and video codecs recommended for DVB-H. The technique of using redundant audio frames in a time-sliced transmission environment called Synchronized Audio Redundancy (SAR) is introduced in Section 3 and Section 4 describes the test setup for subjective analysis used to compare SAR with a method that uses conventional MPE-FEC protection mechanism. The subjective analysis results are discussed in Section 5 followed by the conclusions in Section 6.

2. Audio - Video Codecs and protocols for DVB-H

The Moving Picture Experts Group (MPEG) specified the Advanced Audio Coding (AAC) and later extended AAC with spectral band replication and parametric stereo coding tools, resulting into High Efficiency AAC version 2 (AAC-HE v2) [6], which is recommended for streaming over DVB-H. AAC encoders can compress generic audio providing up to 48 channels and sampling rates from 8 to 96 kHz. Additional coded data generated by an AAC-HE v2 compared to the basic AAC encoder is included as auxiliary information within the AAC raw block structure. This enables all older AAC decoders to decode the AAC part of the AAC-HE v2 streams.

DVB-H specifications also support the use of the Extended Adaptive Multi-Rate Wideband (AMR-WB+) audio codec [7] specified by the Third Generation Partnership Project (3GPP). AMR-WB+ is extended from the AMR-WB speech codec with bandwidth extension and improved stereo encoding tools. AMR-WB+ supports sampling rates from 16 to 48 kHz and a range of operating modes for different bit rates.

H.264/AVC [8], specified by the joint video team of ITU-T and ISO is one of the recommended codecs for video compression. It is designed as a two-layer abstraction. The first layer, Video Coding Layer (VCL), is concerned with the compression efficiency of the input video. The second layer, Network Abstraction Layer (NAL), formats the VCL representation of the video and provides header information in a manner appropriate for a storage media or conveyance by transport layers. In H.264, as in many other hybrid video coders, an input picture frame is divided into smaller units called slices. Slices and slice data partitions are mapped to VCL NAL units, whereas non-VCL NAL units contain additional information data such as supplemental enhancement information (SEI) and parameter sets. Both types of NAL units can then be mapped onto any kind of transport appropriately. For transmission using Real Time Protocol (RTP) over User Datagram Protocol (UDP), each NAL unit can be transmitted individually in an RTP payload or multiple NAL units can be aggregated to one RTP payload.

Streaming media transmitted in DVB-H is encapsulated in RTP packets according to the codec-specific RTP payload format. RTP packets are encapsulated into UDP (User Datagram Protocol) and IP (Internet Protocol) packets and further into MPE sections for transmission over DVB protocols. Each MPE section consists of a header, the IP datagram as a payload, and a 32 bytes cyclic redundancy check (CRC) for the verification of payload integrity. The MPE section header contains addressing data among other things. RS FEC codes are computed by using a matrix structure as described in [9]. The MPE and MPE-FEC sections are mapped onto MPEG-2 Transport Stream (TS) packets. The TS packets are then sent to the DVB-H physical layer in time-sliced bursts to minimize power consumption at the receiver.

3. Synchronized Audio Redundancy Coding

This section describes the proposed technique called synchronized audio redundancy coding (SAR). While the technique is described in the context of DVB-H, it could be equally well used in any communication system with a long time-interleaved FEC, similar to MPE-FEC.

At the transmitter end, the ADT is filled conventionally with audio-visual data. However, as shown in Figure 1, a redundant copy of the audio packets that fills the next MPE-FEC frame is also included in the currently processed MPE-FEC frame. In other words, the transmission of the redundant audio stream is synchronized to the MPE-FEC matrix structures and time-slice intervals. The redundant audio stream can be an independent RTP stream, whose relationship to the primary audio stream can be indicated with extensions to the Session Description Protocol (SDP) or the Electronic Service Guide (ESG) format of DVB-H.

At the receiver end, the redundant audio data is buffered until the corresponding primary audio data is received in the next MPE-FEC frame. Any incorrect primary audio frames can be recovered from the corresponding correct redundant audio frame. If neither the primary audio frame nor the corresponding redundant audio frame is correctly recovered, the frame is replaced with a silent frame or concealed. Due to the relatively large transmission time intervals between the time-sliced bursts, the probability that both the primary audio data and the redundant copy would remain un-correctable is usually low.

Synchronized audio redundancy coding causes additional latency at the DVB-H transmitter because an MPE-FEC frame cannot be constructed before obtaining the audio data for the following MPE-FEC frame. However, no additional channel-switching delay at the receiver is incurred. Furthermore, this method is backward-compatible to the current specifications of DVB-H, i.e. the current receivers simply use the primary audio stream and ignore the redundant audio stream.

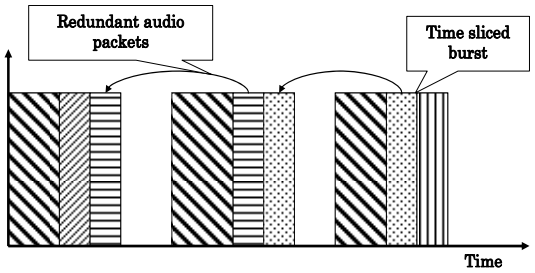

Figure 1: Synchronized Audio Redundancy (SAR).

It is also possible to transmit the redundant audio stream with lower bit rate (lower quality) compared to the primary stream. If the decoder does not support changing the coding mode during the decoding of a stream or different codecs are used for the primary and the redundant streams, both the primary and the redundant streams have to be independently decoded at the receiving end, which naturally increases the computational requirements.

4. Test Setup and Simulation Conditions

Test streams were prepared to compare synchronized audio redundancy coding with conventional error protection of DVB-H. Subjective evaluation was conducted in a controlled laboratory environment to compare the methods. The test setup and procedure are described in section 4.2 and the results are presented and analyzed in section 4.3.

4.1 Test clips preparation for the subjective test

Four semantically complete audio-visual clips of different genres: news, sports (ice-hockey), music video, and animation, with play-time duration of approximately a minute, were chosen for the tests. All the clips except for animation had a high amount of spatial details. The motion content in sports and music video clip was high and the news and animation clips were moderate. All clips except for the music video contained only speech.

The clips were encoded according to the lowest interoperability class specified for DVB-H using H.264/AVC video encoder and AAC audio encoder. The picture size and frame rate of the video clips were QCIF and 12.5 Hz, respectively. One Intra decoder refresh (IDR) picture was inserted for every DVB-H time-sliced burst so as to reduce receiver tuning-in delay as well as to recover from residual transmission errors. Audio was coded in mono at a bit rate of 32 kbps and a sampling rate of 16 kHz. Two sets of clips were encoded. The first of these used conventional transmission with picture freeze (CT-PF), and the second one with SAR with picture freeze (SAR-PF). In the first set of clips, the video bit rate was 128 kbps, whereas for the clips with SAR-PF, the video bit rate was dropped to 96 kbps, as the redundant audio stream occupied 32 kbps and the total audio-visual bit rate in the two sets of clips was kept approximately the same.

As the boundary of acceptability and unacceptability was found to lie between 6.9% and 13.8% in terms of the residual MPE-FEC frame error rate (MFER) in our earlier study [3], the coded audio-visual streams were corrupted with a DVB-H channel model having MFERs 6.9% and 13.8%. A more detailed description of the simulation environment and conditions can be found in [3].

In both SAR-PF and CT-PF methods, when missing video data was detected after MPE-FEC decoding, the video decoder stopped decoding of any subsequent pictures until the subsequent IDR picture and displayed the last uncorrupted decoded picture. In SAR-PF, when an audio frame was lost and could not be recovered from the redundant audio stream, it was replaced with a null frame perceived as silence. In CT-PF, when an audio frame was lost there is no redundant audio frame to recover from. Hence all lost audio frames after MPE-FEC decoding were replaced with null frames.

4.2 Procedure of the subjective test

The test clips were subjectively evaluated by 45 participants. The participants were equally stratified by gender and age groups (18-45 years). The technology-aware and people categorized as innovators and early adopters based on their technology attitude were restricted to 20%. All participants had normal or corrected-to-normal vision and hearing.

The test procedure started with sensorial tests and psycho/demographic data collection. It was followed by a combined anchoring and training where participants are familiarized with the evaluation task, quality range and contents used in the experiment. The actual test procedure followed the Absolute Category Rating (~Single Stimulus) principles [10] [11]. The stimuli were shown one by one with a 5-second interval (for scoring) between them. The participants were asked for a binary (yes/no) acceptance rating and a satisfaction score on an unlabelled 11-point quality scale.

The standards for subjective testing [10] [11] [12] were followed, when the clips were presented in a laboratory environment with Nokia 6630 mobile phone enclosed in a stand. The device was vertically aligned and the viewing distance was set to 44cm. Headphones delivered in the Nokia 6630 sales package were used for audio playback. Audio playback loudness level was adjusted to 75dBA (+ 10 dBA for peaks). The test session consisted of two randomized presentation rounds for all test material and the starting round varied between the subjects.

5. Subjective Analysis Results

In the analysis of acceptance ratings, McNemar's test was used to test the differences between two related and nominal categories. The acceptance evaluations for different error control methods in each content presentation are shown in Figure 2 Both of the error control methods were evaluated being in the same level in music video and animation presentation ($p>.05$) at MFER 6.9%. SAR-PF was found to be significantly better ($p<.05$) for the news clip at MFER 6.9%, but it was inferior ($p<.05$) for the sports clip at MFER 6.9%. Furthermore, SAR-PF was found significantly better for the animation ($p<.001$), news ($p<.05$) and sports ($p<.05$) clips at MFER 13.8%, and there was no significant difference between the tested methods for the music video clip at MFER 13.8%.

Non-parametric Wilcoxon matched pair signed rank test was used to measure the differences between two related ordinal data sets because the data did not reach the normality requirement of parametric tests. In general, the satisfaction ratings, shown in Figure 3, reflected the same trend as the acceptance ratings but the differences can be

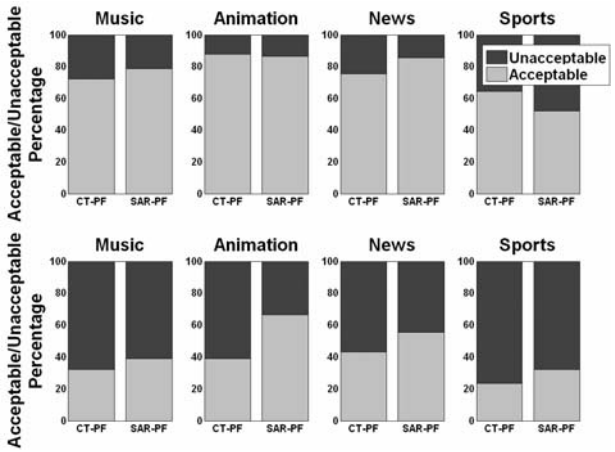

Figure 2: Acceptance graphs for CT-PF and SAR-PF

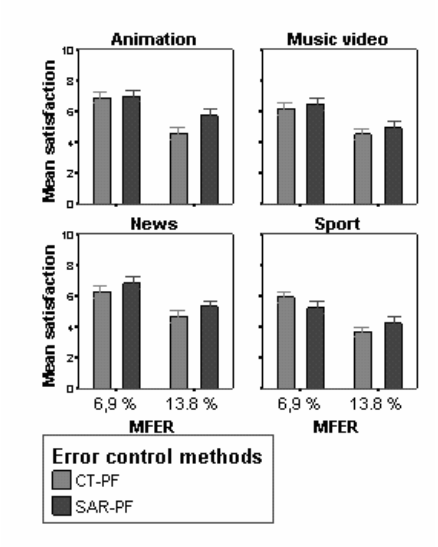

Figure 3: Mean satisfaction scores for CT-PF and SAR at the tested MFERs. The bars indicate mean score and the error bars indicate a 95% confidence interval of the mean.

found more clearly than from the acceptance scores. At MFER 6.9%, there were no significant differences between the methods for music ($Z=-1.784$, $p>.05$) and animation sequences ($Z=-0.213$, $p>.05$). SAR-PF was found to be significantly better for the news clip ($Z=-3.613$, $p<.001$), but it was inferior to the sports clip ($Z=-3.776$, $p<.001$) at MFER 6.9%. SAR-PF was found significantly better in all contents at MFER 13.8% (Animation $Z=-5.753$, $p<.001$; News $Z=-3.274$, $p<.001$; Sport $Z=-3.639$, $p<.001$; Music video $Z=-2.094$, $p<.05$).

From the subjective results, it can be concluded that SAR provides an obvious advantage on the perceived quality of audio-visual sequences, especially at high error rates. This conclusion is evident based on the results for MFER 13.8%, in which SAR consistently outperforms the reference method. However, when the overall bit rate budget is limited, SAR is not recommended for fast moving and visually sensitive sequences at low error rates as illustrated by the results for the sports sequence at MFER 6.9%.

6. Conclusions

A method using synchronized redundant audio coding to combat transmission errors in the DVB-H environment was presented. The method was compared against the conventional error protection of DVB-H in a large-scale subjective evaluation. Synchronized audio redundancy coding (SAR) was found to outperform conventional error protection in most test cases. It seems that SAR provides especially good results in audio-driven content types, such as news, and with error rates that correspond roughly to the boundary between acceptable and unacceptable quality. However, when the overall bit rate budget is limited, SAR is not recommended for fast moving and visually sensitive sequences at low error rates.

7. Acknowledgment

This study was funded by Radio- ja televisotekniikan tutkimus Oy (RTT). S. Jumisko-Pyykkö is funded by UCIT graduate school. We wish to thank Kati Nevalainen for her work in the project.

8. References

[1] ETSI, “Digital video broadcasting (DVB): Transmission system for handheld terminals (DVB-H),” European Standard EN 302 304, version 1.1.1, Nov. 2004.

[2] ETSI, “Digital Video Broadcasting (DVB): Framing structure, channel coding and modulation for digital terrestrial television.” ETSI standard, EN 300 744, 2001.

[3] S. Jumisko-Pyykkö, V.K. Malamal Vadakital, and J. Korhonen, “Unacceptability of instantaneous errors in mobile television: from annoying audio to video,” Proc. of Int. Conf. on Human-Computer Interaction with Mobile Devices and Services (MobileHCI), Sep. 2006.

[4] C. Perkins, O. Hodson, and V. Hardman, “A survey of packet loss recovery techniques for streaming audio,” IEEE Network, vol. 12, no. 5, pp. 40-48, Sep/Oct 1998.

[5] C. Perkins, I. Kouvelas, O. Hodson, V. Hardman, M. Handley, J.C. Bolot, A. Vega-Garcia, and S. Fosse-Parisis, “RTP payload for redundant audio data,” IETF RFC 2198, Sep. 1997.

[6] ISO/IEC 14496-3 (2001), “Information technology - Generic coding of moving picture and associated audio information - Part 3: Audio,” including ISO/IEC 14496-3 Amd-1 (2001), “Bandwidth Extension,” and ISO/IEC 14496-3 Amd-2 (2004), “Parametric Coding for High Quality Audio.”

[7] 3GPP TS 26.290 Rel-6, “Extended AMR wideband codec; transcoding functions.”

[8] ITU-T and ISO/IEC JTC 1, “Advanced video coding for generic Audio visual services,” ITU-T recommendation H.264 – ISO/IEC 14496-10(AVC), 2003

[9] ETSI EN 301 192, “DVB specification for data broadcasting,” version 1.4.1, Nov. 2004.

[10] ITU-R BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” 2002.

[11] ITU-T P.911, “Subjective audiovisual quality assessment methods for multimedia application,” 1998.

[12] ITU-T P.920, “Interactive test methods for audiovisual communications,” 2000.