

Audio Watermarking

— summary results of EBU tests

Andrew Mason

EBU Project Group N/WTM

After putting out a call for audio watermarking systems, EBU Project Group N/WTM conducted extensive tests between June 2001 and July 2003 on the only two suitable systems that were received – one from Philips, the other from Fraunhofer-IIS. This article reports on the results of these tests.

The EBU has been actively involved in watermarking, for video and audio, for several years. A technical project group, N/WTM, was formed to evaluate the application of watermarking signals transmitted over the Eurovision network for purposes of copyright protection [1][2].

Thorough tests were conducted on video watermarking systems in 2000 [3]. Audio watermarking tests followed and have recently been completed. The process was lengthy, starting in June 2001 and finishing in July 2003. In summary, a list of technical requirements was made ... *Calls For Systems* were issued ... systems were received ... and then they were tested. Several members of EBU Project Group B/AIM (Audio In Multimedia) were involved in the test process; in particular, the IRT, Polish TV and France Télécom, as well as the BBC.

Calls For Systems

The EBU issued *Calls For Systems* via a press release in February 2002 (see the text panel on the next page). The desirable technical requirements of an audio watermarking system were specified in an annex of that press release and an extract from it is reproduced here in *Appendix A*. In short, the systems should be inaudible, should convey 48 bits of data in a 5s segment of audio, and should do that even after a wide variety of processes have been applied to the audio. Audio interfaces should be AES/EBU or S/PDIF, data input via RS-232, output via RS-232 or computer data file.

An unforeseen ambiguity in the data capacity, apparent only when testing was well under way, necessitated a *Call For Modified Systems* from the system proponents remaining at that time.

Systems submitted

Three companies submitted systems. **Fraunhofer-IIS** submitted a system based on a Linux PC, **Philips** a system based on a Windows 2000 PC, and **Communications S.A.** a DSP-based system in 19" rack-mount form.

In preliminary tests, the Communications S.A. system was found to have some problems with false positive detections, and perceptibility on some signals. As a result, it was not subjected to the full set of tests.

EBU Press Release – February 2002

Geneva, February 2002 – The European Broadcasting Union today called on industry from around the world to collaborate in identifying the best system of electronic “watermarking” to protect the copyright of digital broadcasts.

The need for a secure means of marking video and audio material is increasingly urgent at a time when digital technology enables perfect pirate copies to be made and distributed through a growing number of communications channels.

A special EBU project group has concluded that a new “watermarking” technology, which has been developed in cooperation with industrial partners, could be of great use in copy protection and intellectual property and rights management (IPRM) throughout the audiovisual chain – from the production phase through transmission to the viewer. The system involves the imperceptible marking of video/audio sequences (watermarking) and a watermarking reader able to recognise watermarked data in real time from a digital signal.

Extensive video watermarking testing has been made by EBU through the year 2000 and test results were published in the EBU Technical Review no. 286 (March 2001).

Today's Call For Systems (CFS) is designed to open up the work already undertaken to all companies working on audio watermarking, to broaden agreement on the theory behind the method proposed by the EBU, and to compare and assess existing systems by tests using the same methodology. Technical details of the CFS are annexed.

The provisional schedule for the tests has been proposed as follows:

- 18th February 2002 Issue of the press release;
- 18th March 2002 Replies from industry, letter of interest;
- 1st June 2002 Delivery of the equipment at BBC;
- ~ October 2002 Results of the tests (may depend on the number of units of equipment to be tested).

The EBU serves 69 national broadcasters from 50 countries in the European area, and has 49 associate members further afield. It operates the Eurovision and Euroradio networks, coordinates the exchange of news and sports programming, conducts technical research, stimulates co-productions, and defends public service broadcasting.

Robustness test method

To measure the *system robustness*, a series of pseudo-random payloads was embedded – using each watermark system – into a 20-minute selection of audio. The watermarked audio from each system was then segmented, and unrelated audio – with a fixed payload – was inserted between the segments. Software to generate *Cool Edit Pro* session files was written to automate this task. The segmented sequences were then played into the watermark detectors.

The robustness to a particular process is expressed as the number of payloads recovered as a percentage of the number of segments.

The unsegmented sequences were also played into the detectors. This makes the test less severe in some cases. The results are expressed in the same way as for the segmented signals; that is, as the number of payloads recovered as a percentage of the same number of segments.

A payload being recovered that was not embedded is called a *false positive* detection. The probability of these should be very small, so as not to cause too many false accusations of misuse.

Two different interpretations were made of the payload capacity requirement by the system proponents. This meant that two series of robustness tests had to be run:

- the first series ... with a segment length of 5s;
- the second series ... with a segment length of 10s.

Robustness results

In the first phase of robustness tests, the FhG-IIS system had an effective watermark *minimum segment length* of 10s and the Philips system one of 5s. In the second phase the FhG-IIS system was modified to have a watermark minimum segment length of 5s and the Philips system one of 10s. *Table 1* shows the results of the robustness tests using a 10s watermark minimum segment length (FhG-IIS first phase “original” and Philips second phase “bis”). *Table 2* shows the results of the robustness tests using a 5s watermark minimum segment length (FhG-IIS second phase “bis”, and Philips first phase “original”).

Because of restrictions on the time available to conduct the second phase of tests, not all processes were tried in the second phase. Results from the first phase are shown, but entries in the tables for the second phase are

Table 1

Percentage of marks recovered with 10s watermark segments:

FhG system from the first phase of tests, Philips system “bis” from the second phase

Attack	Continuous replay			Segmented replay		
	FhG A	FhG B	Philips bis	FhG A 10s	FhG B 10s	Philips bis 10s
No attack	95.2	94.8	97.6	-	91.2	96.0
Minidisc (BBC)	93.5	94.4	96.0	84.6	91.2	94.4
Minidisc (TVP)	93.5	94.0		-	88.8	
Dolby AC-3, 128 kbit/s, stereo	92.7	93.7		-	88.0	
MPEG Layer II, 128 kbit/s, joint stereo	94.4	93.1	96.8	-	89.6	91.9
MP3, 96 kbit/s, stereo	91.5	93.5	62.1	58.8	89.6	57.3
MP3, 64 kbit/s, stereo	73.8	75.0	6.5	-	63.8	8.1
AAC 32 kbit/s, stereo	7.7	0.8	0.0	-	0.0	0.0
MPEG Layer II, 32 kbit/s, mono	0	0.4		-	0.0	
Linear time stretch, (10%)	0	84.6	95.2	0	0.0	43.5
Pitch-corrected time-stretch, 5%	0.4	0.8	0.0	1.6	0.0	0.0
Voice-over, +15 dB	0.8	0.0		-	0.8	
Added white noise, -30 dB	0.8	0.0		0.8	0.8	
Dynamic range compression	95.2	95.2	96.0	60.4	92.0	94.4
Analogue conversion	92.3	91.5		91.2	88.8	
Combined audio processing	94.8	94.8		-	92.8	
Broadcast chain 1, FM	87.5	86.3	57.3	83.0	80.6	58.9
Broadcast chain 2, NICAM	91.5	90.3		78.2	90.4	
Broadcast chain 3, NICAM + MPEG	90.7	90.7	91.1	61.2	69.4	71.8
Broadcast chain 4, Dolby E+AC3	93.1	89.5		-	86.2	
First watermark detection after application of second watermark	95.2	94.4	93.5	-	88.8	96.8
Second mark detection	94.0	95.6	96.8	-	95.2	95.2

Colour code:

100% - 80%

79% - 60%

59% - 40%

39% - 20%

19% - 0%

left blank in this case. These table rows are shown with a white background, to avoid giving a misleading impression of comparative performance between the systems ... where no comparable results are available.

Some tests caused erroneous behaviour by the detectors. In these cases, it was not possible to calculate a valid result: these are shown as “-” in the tables. This behaviour was subsequently explained by the system proponent to be the consequence of the design of a sampling frequency scaling search strategy. The strategy was confounded by the segmentation. The system proponent has reported that this problem has now been solved.

Table 2

Percentage of marks recovered with 5s watermark segments:

FhG system “bis” from the second phase of tests, Philips system from the first phase

Attack	Continuous replay			Segmented replay		
	FhG A bis	FhG B bis	Philips	FhG A bis 5s	FhG B bis 5s	Philips 5s
No attack	98.0	98.0	73.4	- ^a	90.7	89.1
Minidisc (BBC)	95.6	97.6	7.7	-	85.9	14.5
Minidisc (TVP)			0.0			0.0
Dolby AC-3, 128 kbit/s, stereo			0.0			0.0
MPEG Layer II, 128 kbit/s, joint stereo	97.2	96.4	0.0	-	86.3	0.0
MP3, 96 kbit/s, stereo	90.3	91.9	0.0	-	68.1	0.0
MP3, 64 kbit/s, stereo	56.5	42.7	0.0	-	15.3	0.0
AAC 32 kbit/s, stereo	4.8	0.0	0.0	0.0	0.0	0.0
MPEG Layer II, 32 kbit/s, mono			0.0			0.0
Linear time stretch, (10%)	0.0	97.2	0.4	0.0	0.0	2.6
Pitch-corrected time-stretch, 5%	4.0	1.6	18.1	2.8	2.0	1.2
Voice-over, +15 dB			0.0			0.0
Added white noise, -30 dB			0.0			0.0
Dynamic range compression	98.0	98.0	75.0	-	92.7	60.1
Analogue conversion			0.0			0.0
Combined audio processing			0.0			0.0
Broadcast chain 1, FM	79.4	63.7	0.0	32.7	52.0	0.0
Broadcast chain 2, NICAM			0.0			0.0
Broadcast chain 3, NICAM + MPEG	96.0	96.0	0.0	64.1	82.3	0.0
Broadcast chain 4, Dolby E+AC3			7.3			19.4
First watermark detection after application of second watermark	98.0	98.0	39.9	-	90.3	63.3
Second mark detection	98.0	98.0	41.1	- ^b	93.5	39.5

a. When the interstitial audio was digital “0”, 60.5% correct was obtained

b. When the interstitial audio was digital “0”, 79.8% correct was obtained

colour code:

100% - 80%

79% - 60%

59% - 40%

39% - 20%

19% - 0%

False positive detections

The false positive results – where a watermark is detected that was not inserted – can be summarised as follows. In the first phase of testing (Philips 5s segments, Fraunhofer-IIS 10s segments), there were no false positive detections. In the second phase of testing (Philips 10s segments, Fraunhofer-IIS 5s segments), the Philips system again produced no false positive detections, but the Fraunhofer-IIS system produced 0.4% false positive detections (as a percentage of the theoretical maximum number of true detections) in 5 of the 26 robustness tests. The requirement was for a false positive probability of less than 1 in 10^8 detections.

Subjective test method

The audio watermarks were required to be inaudible in studio listening conditions. This would be assessed by subjective tests according to ITU-R Recommendation BS.1116 [4]. In this method, subjects are presented with three stimuli from which to choose. One stimulus is the unprocessed reference (identified to the subject). The other two stimuli (not identified to the subject) are randomly assigned to the processed signal and to a copy of the reference. The subject has to give grades to the two unidentified signals, according to the perceived impairment relative to the known reference, using the 5-point impairment scale:

- 5 – imperceptible;
- 4 – perceptible, but not annoying;
- 3 – slightly annoying;
- 2 – annoying;
- 1 – very annoying.

At least one of the signals should receive a grade of “5” ... because it is identical to the known reference.

However, because it was hoped that the standard of systems submitted would be high (such that any impairments to the audio signal would not be revealed by BS.1116 tests), an addition was made to the test method used. Subjects were required to indicate which of the two signals to be graded (the hidden reference and the watermarked signal) they thought was watermarked — *even if they would have indicated that the effect was imperceptible*. This is called a *forced choice*.

The normal BS.1116 selection-panel process was used to find the set of items to be used in the tests. These are listed in *Table 3*.

Table 3
Signals used in the test sessions, with watermark at normal level

Test item number	Item type (and source)
4	xylophone (EBU SQAM, track 41)
8	flute (EBU SQAM, track 13)
9	glockenspiel (EBU SQAM, track 35)
13	triangle (EBU SQAM, track 32)
16	violin (EBU SQAM, track 8)
27	harpsichord (EBU SQAM, track 40)
31	English male speech (EBU SQAM, track 49)
34	German male speech (EBU SQAM, track 54)
36	wind ensemble (EBU SQAM, track 36)
39	Tennis (BBC)

Some other signals were used, with artificially increased watermark levels, in order to train the listeners – and to provide something in the test that would be detectable by ear.

Subjective results

The results of the subjective tests are in three categories:

- ITU-R BS.1116 difference-grades (diff-grades) and the 95% confidence intervals;
- a Wilcoxon rank sum test on the BS.1116 diff-grades for each subjective test item;
- recognition rate from forced choice as a function of test item and of listener.

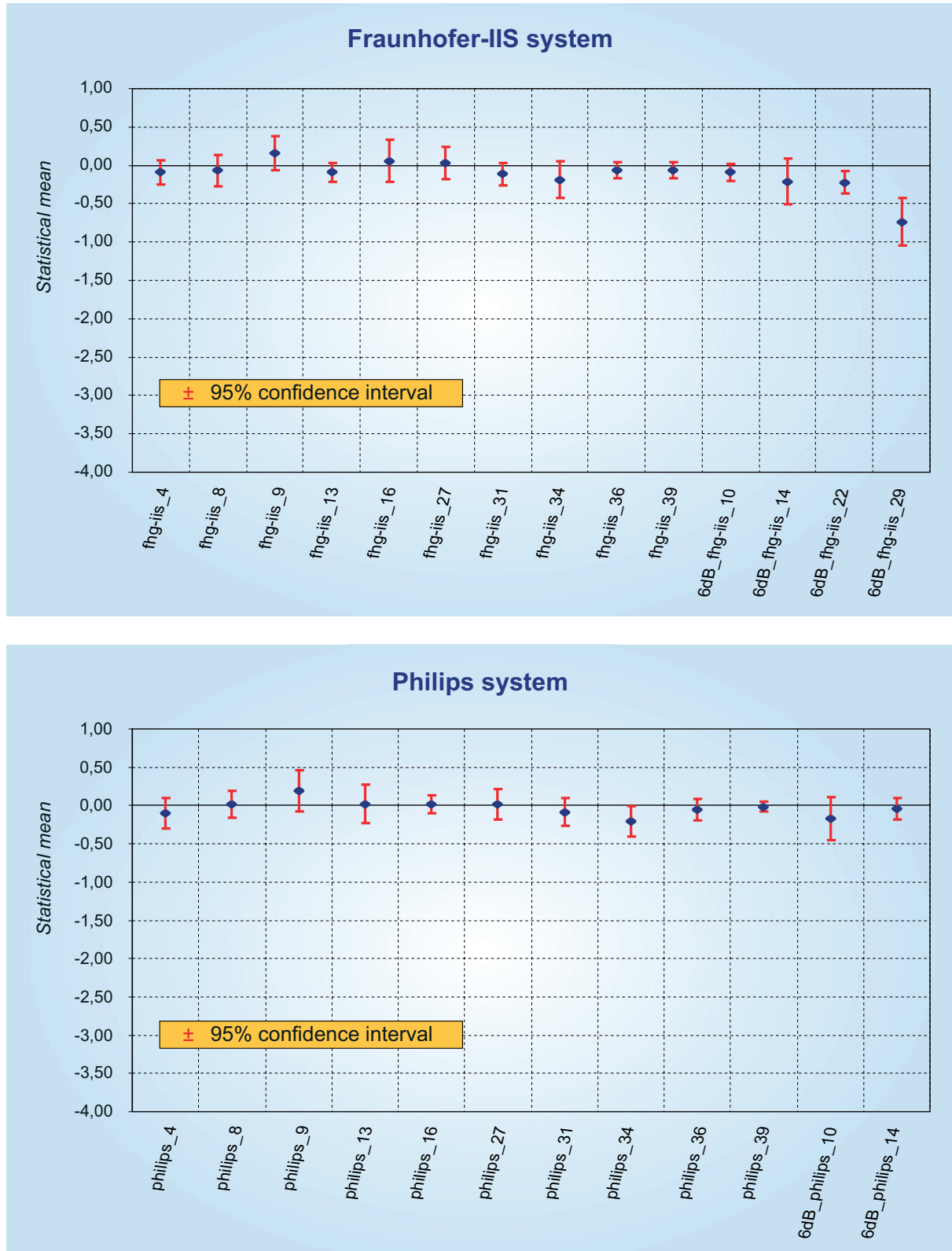


Figure 1
Mean grade and 95% confidence interval for (*upper*) the Fraunhofer-IIS system and (*lower*) the Philips system

As was noted earlier, the robustness tests were conducted in two phases because of different interpretations of the 5s watermark *minimum segment length* requirement. The formal subjective tests were conducted on the systems supplied for the first phase of robustness tests. Assurances were given by the system proponents that the alternative configurations supplied for the second phase would not differ in perceptibility. Informal, but expert, listening confirmed this.

BS.1116 difference grades and confidence intervals

The diff-grades, together with 95% confidence intervals for the tests on the FhG-IIS and Philips systems, are shown in *Fig. 1*. Item labels starting with “6dB_” indicate signals with the watermark inserted at an artificially boosted level.

Only one of the confidence intervals, for a normal level of watermark, does not cross the 0 diff-grade axis: item 34 (German male speech) with the Philips watermark. As is apparent in the figure, it come very close to the axis, but it does not quite cross it.

None of the confidence intervals, for the signals watermarked at a normal level, extend below a diff-grade of – 0.5.

Wilcoxon rank sum test

The results of the *Wilcoxon rank sum* test on the subjective test grades were consistent with the confidence intervals. Only item 34, with the Philips watermark, showed a significant difference (with a 95% probability) from the zero distribution.

Forced choice

The forced choices made by the subjects were analysed and the percentages of correct identifications calculated. Percentages were calculated as a function of test item and subject for each system, and overall for each system.

The recognition rates, as a function of subject, show quite a wide variation. One subject scored a 90% recognition rate for the Fraunhofer-IIS system, another scored 90% for the Philips system. However, several listeners scored 20% recognition for one system or the other.

The recognition rates, as a function of test item, showed a much smaller variation with no item having a recognition rate more than 68% for the Fraunhofer-IIS system, or 65% for the Philips system. The lowest recognition rates for items were about 48% and 32% respectively.

The overall recognition rates (for all items and all subjects) were 56.0% for the Fraunhofer-IIS system and 46.8% for the Philips system.



Andrew Mason received a BSc in physics from the Imperial College of Science and Technology in London in 1986, and then joined the BBC's Research and Development department at Kingswood Warren. There he has worked mostly in digital audio, applying digital signal processing techniques to broadcast operations. Several years were spent working on audio coding techniques, particularly tandem coding, including helping to develop the AES41 standard for audio coder control data, and BS.1534 – the MUSHRA subjective test method.

For the last 4 years or so, he has been working in watermarking for both digital video and digital audio, developing and evaluating systems with potential applications in broadcasting.

Mr Mason is a member of EBU project groups B/AIM (Audio in Multimedia) and N/DRM-T (Digital Rights Management – Technology), of AES Standards Committee working groups SC-02-02 (input-output interfaces) and SC-06-04 (internet audio delivery systems), and of DVB CPT (Copy Protection Technology).

Abbreviations

AES	Audio Engineering Society	MUSHRA	(EBU) MULTI Stimulus test with Hidden Reference and Anchors
CFS	Call For Systems	NDA	Non-Disclosure Agreement
DSP	Digital Signal Processor / Processing	S/PDIF	Sony/Philips Digital InterFace
IPRM	Intellectual Property and Rights Management		

Conclusions

A call by the EBU for audio watermarking systems resulted in a number of systems being submitted for testing. The technical requirements listed in the call included desired *robustness*, *data capacity* and *perceptibility*.

Of the three systems submitted, two – from Fraunhofer-IIS and Philips – were found to be suitable for thorough evaluation. The third – from Communications S.A. – was found to be unacceptably audible on several test items, and had a very high false positive detection rate.

The robustness tests showed that the watermarks of the Philips and Fraunhofer-IIS systems were fairly robust to processes that did not significantly reduce the quality of the audio signal. In general, it was processes that resulted in poor audio quality that produced low detection rates. Overall, as expected, the 5s segment systems were less robust than the 10s segment systems.

It is not possible to conclude that any one of the systems is always more robust than the other: after one process, one system might detect more watermarks than the other but, after a different process, the results would be the opposite.

The two systems put through the BS.1116 subjective testing process showed high audio quality. None of the mean diff-grades for normal watermark-level signals were less than -0.25 and none of their 95% confidence intervals extended below -0.5 . Only one watermarked test item, with the Philips system, showed a distribution of grades that differed, in a statistically significant way, from the original.

The analysis of the forced choice data merits more study. Some individuals scored very high recognition rates, others very low. At the moment, the data has not proved that the systems are audible. However, further study is required to establish the recognition rate required to differentiate between zero and close-to-zero probabilities of audibility.

Hitherto, audio watermarking has not seen widespread adoption in the many areas where it could be used. The decisions have been influenced by valid concerns about perceptibility and reliability of detection. To conclude, the study has shown that audio watermarking technology is now available, which is practically inaudible, has a useful data capacity and is usefully robust. It remains to be seen whether it will be adopted.

References

- [1] Jean Barda and Louis Cheveau: [Eurovision network security through access control and watermarking](#)
EBU Technical Review No. 281, Autumn 1999.
- [2] Jean Barda and Louis Cheveau: [Access control and watermarking](#)
EBU Technical Review No. 282, March 2000.
- [3] Louis Cheveau, Eddy Goray and Richard Salmon: [Watermarking – summary results of EBU tests](#)
EBU Technical Review No. 286, March 2001.
- [4] ITU-R Recommendation BS.1116-1: **Methods for the subjective assessment of small impairment in audio systems including multichannel sound systems**
International Telecommunication Union

Appendix A:

Extract from EBU doc. BPN 058

CFS technical details and requirements

Two sets of tests will be considered:

- 1) A subjective quality test following ITU-R BS.1116 methodology with a “two alternative, forced choice”.
- 2) A second test, related to robustness of the watermark, in accordance with constraints listed in the table of technical requirements.

Results will be released first to individual companies on a confidential basis.

Companies interested in taking part in the tests are expected to confirm their intention of submitting equipment by mid-March 2002 and the equipment for testing should be delivered to the BBC in London before June 1st 2002. They must provide:

- 1) A watermark embedder able to watermark a digital audio signal in real time;
- 2) A real time watermarking reader able to read watermarked data from a digital audio signal.

Companies will be expected to provide information on the dimensions and weight of the proposed equipment, and estimates of a date for industrial production and of cost. Some information about the algorithm to be used must also be provided, to an independent expert under NDA, to enable calculation of error probability and particularly false positive errors (see table of Technical Requirements).

The embedder should be equipped with a switch allowing two working modes: normal (all parameters of the system optimized) and boosted (non-optimized parameters) for listener training purposes.

The possibility of orthogonal watermark payload channels will be considered.

The delay introduced by the embedder must be as short as possible (maximum 80 ms).

The following mandatory systems parameters have been set to facilitate comparative testing:

Embedder:

Audio interfaces – Preferably AES/EBU on 3-pin XLR, or on 75 ohm BNC, or IEC 60958 on RCA/phono/cinch.

Data – The data to be embedded (48 bit payload) will be provided to the system through an RS-232 Interface: SubD 9 connector, Baud rate: 19200, 1 start bit, 1 stop bit, odd parity Data 48 bit in 12 Hexadecimal ASCII characters terminated by a line feed.

Data Reader:

Audio interface – Preferably AES/EBU on 3-pin XLR, or on 75 ohm BNC, or IEC 60958 on RCA/phono/cinch.

Data – The output will be in the form of a log file (1 line per decision – payload recovered – with time stamp) or available in a serial form at the output of a RS-232 data interface with the same characteristics as defined (EBU audio watermarking technical requirements (February 2002)).

Requirements

√ = mandatory
 # = should not have
 □ = recommended

W1 = production level
 W2 = contribution level¹

1. Audibility Of The Watermark	
1.1 Not audible in comparison with the original under studio listening conditions	√
1.2 Not audible under domestic/consumer listening conditions	√
2. Payload: (Net Bit Rate)	
2.1 Watermark Minimum Segment (WMS) duration	5 sec.
2.2 Data capacity	48 bits / WMS
2.3 Detection probability = complement of the false negative probability / WMS	95%
2.4 False positive probability / WMS^a	10 ⁻⁸
2.5 Probability for (bit) error-free payload / WMS	1-10 ⁻⁸
3. Purpose Of The Watermark	
3.1 Identification	√
3.2 Authentication	#
4. Security - Secret Watermarking-Key^b	
4.1 Difficult-to-predict, cryptographic strong	√
4.2 Number of available watermarking-keys	As large as possible
4.3 Watermarking-key management	√
5. Watermark Detection and Payload Extraction	
5.1 Single-ended watermark detection and payload extraction	√
5.2 Option for double-ended watermark detection and payload extraction e.g. in order to increase the level of reliability	□
6. Format Of Original Unwatermarked And Watermarked Signal	
6.1 PCM-Format, Mono and Stereo, Sampling frequencies: 24.0, 32.0, 44.1,48.0 kHz	√
6.2 PCM-Format, amplitude quantization: 24 bits	√
6.3 ISO/MPEG-Layer-II bit stream (transcoding operation)	□
7. Robustness	
7.1 Data Compression	
7.1.1 ISO/MPEG-Layer-II stereo: mono:	≥ 128 kbit/s ≥ 32 kbit/s
7.1.2 ISO/MPEG-Layer-III, stereo	≥ 64 kbit/s
7.1.3 ISO/MPEG-AAC, stereo	≥ 32 kbit/s
7.1.4 Dolby/AC3	√
7.1.5 NICAM	√
7.1.6 Sony/MiniDisc	√

1. The tests are aimed at the W1 and W2 requirement and do not address the W3 requirement.

7.2 Digital and Analog Filtering	
7.2.1 Re-sampling, e.g. D/A -> A/D conversion	√
7.2.2 Inaudible distortion of frequency response, e.g. notch filter guided by psycho-acoustic model	√
7.2.3 Sampling frequency conversion, up-conversion ratios: e.g. 24.0 kHz -> 32.0 kHz -> 44.1 kHz -> 48.0 kHz down-conversion ratios: e.g. 48.0 kHz -> 44.1 kHz -> 32.0 kHz -> 24.0 kHz	√
7.2.4 Stereo to mono conversion	√
7.2.5 Add white noise at 30dB below peak signal level	√
7.2.6 Pitch-corrected time-scaling up to +5%	√
7.2.7 Linear speed change without pitch correction + 10%	√
7.2.8 All pass filtering	√
7.2.9 General multi-band equaliser with up to +6dB change of level	√
7.3 Special Effects	
7.3.1 Add inaudible echo with random delay and amplitude	√
7.3.2 Multi-band non-linear amplitude compression/gain (Optimod like)	√
7.3.3 Add voice-over at 15dB above signal level	√
7.4 Collusion and Collusion-like Attacks	√
8. Watermark-Editing For Users With Secret WTM-Key	
8.1 Render watermark undetectable	#
8.2 Replace = overwrite the payload bits	#
9. Cascaded Watermarking/ Compatibility of W1/W2	
9.1 Two non-interfering watermarks in the same signal are inaudible	√
9.2 Three non-interfering watermarks in the same signal are inaudible	☐
10. Processing Time And Delay	
10.1 Real-time (on-line) embedding, hardware	√
10.2 Non-real-time (off-line) embedding, software-tool	☐
10.3 Real-time (on-line) detection and payload extraction, hardware	√
10.4 Non-real-time (off-line) detection and payload extraction, software-tool	☐
10.5 Delay for the real-time embedder	< 80 ms

- a. Verification of this figure will require the disclosure of details of the system under NDA to an independent expert.
b. The security should rely on the secret key use, not on the secrecy of the algorithm.