Preprint No. 4823

Digital Watermarking and its Influence on Audio Quality

C. Neubauer, J. Herre
Fraunhofer Institut for Integrated Circuits IIS
D-91058 Erlangen, Germany

Abstract

Today large amounts of multimedia data are available which can easily distributed, e.g. by the Internet. To track illicit copies or to prove ownership of digital data, watermarking is a well-known technique. Regarding the audio quality, however, concerns about potential degradations are voiced frequently. This paper investigates the audio quality of a presented watermarking scheme including results of objective and subjective measurements. These results are viewed in relation to those of perceptual coding schemes. Along with these data, the corresponding performance of the watermarking detection is given.

1 Introduction

Today large amounts of multimedia data including images, movies and, of course, high quality audio data are available to everyone. The new possibilities in distribution and broadcasting of these data via the Internet lead to problems with illicit copying and proving intellectual property rights or ownership thereof. To deal with these problems, watermarking of digital data is a promising concept. In case of audio data, watermarking means to transmit additional data, e.g. a serial number, within the audio signal. This additional data should be imperceptible to a human listener and resistant against "removal attacks".

Watermarking has been done for years in the domain of still and moving images [1, 2]. Therefrom much is known about methods, algorithms and system requirements of such systems. Watermarking of audio signals, however, is an upcoming field of research which is strongly inspired by the ideas known from image watermarking. Former work done by Boney, Tewfik and Hamdy establishes embedding a watermark [3]. F. Tilki and A. Beex showed how to establish are more general channel, using multiple carrier sinusoids [4].

This paper addresses the general problem of audio watermarking systems to guarantee that no audible distortions are introduced by the watermarking process. Therefore we selected a watermarking system and investigated its influence on the audio quality. The selected system is shortly presented in section 2. It establishes a general purpose hidden data channel using uncompressed audio data as cover data.

The audio quality assessment divides into three parts. First, the system is evaluated using objective audio quality measurement systems. The second part presents the results of a suitable listening test using experienced listeners. Finally a comparison to the quality of perceptual coding is given using a well-known codec.

2 Psychoacoustic Data Embedding System

In this section the data transmission system is described consisting of the encoder and the decoder part. The encoder is responsible for embedding the additional data into the audio signal in a manner, that it is imperceptible to a human listener. The term additional data refers to an arbitrary bit stream to be transmitted. The system is based on a psychoacoustic masking model as well as on spread spectrum transmission techniques.

The task of the decoder is to recover the transmitted data bit stream from the "watermarked" audio track. The overall system is shown in Fig. 1. For convenience we refer to the entire data embedding system including encoder and decoder as watermarking system.

2.1 Encoder

The encoder is shown in Fig. 2. It comprises:

- Modulation
- Signal Conditioning including Masking Model
- Signal I/O

2.1.1 Modulation

The modulator performs a binary phase shift keying (BPSK) spread spectrum modulation [5, 6, 7, 8]. For those skilled in the art of communications theory it is known that spread spectrum systems are typically characterized by their processing gain. The processing gain is used to compensate for a low SNR seen at the receiver input. Using a processing gain of $10 \log_{10}(256) = 24.0$ dB turned out to be a suitable compromise between data transmission rate and processing gain. This yields a data transmission rate of 47 bit/s. A more detailed discussion of the modulation block can be found in [9].

2.1.2 Masking Model

The psychoacoustic model is used to prevent the embedded data signal from becoming perceptible to a human listener. The masking model used in the watermarking scheme is the masking model of ISO/MPEG2 Advanced Audio Coding (AAC) [10, 11]. It operates on a

block basis and divides the audio frequency spectrum into partitions. For each partition a masking threshold is computed giving a measure for the allowed distortion energy within that particular partition which is used in the spectral weighting block.

To avoid preechos within the watermarked signal, masking thresholds are computed both for "short blocks" and "long blocks". Short blocks are based on a 256-point fast Fourier transform, whereas long blocks take into account 2048 samples. The masking model provides information about the recommended block type. This information is used to operate a signal multiplexer behind the spectral weighting block which is discussed next.

2.1.3 Spectral Weighting

Although spectral weighting is computed separately for long and short blocks they share the underlying principle. The threshold provided by the masking model is used by the spectral weighting block to shape the energy distribution of the spread spectrum data signal in the frequency domain. This is accomplished by applying a specific attenuation for each frequency partition to the spread spectrum data signal. After performing this operation the spectral energy distribution of the data signal over frequency follows the masking threshold. This ensures the inaudibility of the added data bearing signal.

2.2 Decoder

The decoder consists mainly of a standard BPSK receiver equipped with special means for measuring the bit error rate of the established channel. The block diagram is shown in Fig. 3.

Note that the decoder is only interested in recovering the hidden additional information using the watermarked audio signal as a carrier signal. The original audio signal itself can be considered as a jammer interfering with the transmitted information bearing signal.

The most important part of the decoder is the matched filter. The filter coefficients match to the spreading sequence used in the encoder. Behind the matched filter, peaks representing the sent information are observed. They occur approximately in symbol timing which is denoted by $T_s \approx 1/47$ s. This timing information is used by the synchronizer in order to estimate the exact symbol timing and sampling time.

Knowing the sampling time, the matched filter output signal can be sampled and a threshold decision recovers the sent information. In the case of BPSK the threshold is zero. Thus a positive sample indicates a binary "0", a negative sample indicates a "1", or vice versa.

The bit error measurement is based on a pseudo random bit sequence which is known to both decoder and encoder. This sequence is sent by the encoder instead of arbitrary data. Assuming fairly low bit error rates the decoder is able to synchronize its internal replica of the sequence with the received sequence. Bit errors are detected by comparing both the internal and the received signal. A more detailed discussion of the decoder can be found in [9].

3 Investigation of Audio Quality

It is well-known from the art of perceptual audio coding that even slight signal modifications of audio signals may result in strong perceptible distortions. In the case of audio coding the distortions arise from the quantization noise due to the coding process. In watermarking systems dealing with uncompressed audio signals, an additional information bearing signal is added to the original audio track. In general this adds some kind of distortion which can be kept to a minimum using a fairly elaborated watermarking system.

This section describes the evaluation of the above system in terms of audio quality. In perceptual audio coding, the quality of codecs often is evaluated by comparing an original signal, called reference, with its coded version. Naturally, this general principle applies to the quality evaluation of a watermarking system as well. Instead of evaluating the coded version (as is the case in codec quality assessment) the watermarked version is analyzed.

First the applied methods are presented. This description includes the objective measurement systems, the listening test setup, mathematical requisites, codec parameters and other information to understand the details of the test procedure. However, the results obtained with the presented tools are shown in section 4.

3.1 Selection of Test Items

The items used for testing were short (8-12s) excerpts of the following tracks:

- "harpsichord", "castanets", testing hard attacks
- "pitch-pipe", "bagpipes", testing tonality
- "triangle", testing attacks + tonality
- "german speech", testing speech signals
- "gershwin", testing general orchestral music

These excerpts are standard items known to be critical for audio coding. Such a choice seems appropriate because a noise-like signal is hidden below the masking threshold by the watermarking system just like in traditional perceptual audio coding. Therefore the same problems and effects known from audio coding are expected to arise with the watermarking system. The selected items provide extreme cases in terms of tonality, attack and spectral distribution. If a watermarking system is able to handle these items successfully one also would hope for a good behavior in the context of more average program material.

3.2 Objective Measurements

To get an estimation of the audio quality, objective measurement methods were performed with the selected audio tracks. The quality measurement systems applied were "Perceptual Audio Quality Measure" (PAQM) [12], the system "Perceptual Evaluation of Audio Quality" (PEAQ) [13] and some selected parameters of the "Noise to Mask Ratio" (NMR) [14] measurement system:

- PAQM derives an estimate of the signals on the cochlea and compares the representation of the reference signal with that of the signal under test. The weighted difference of these representations is mapped to the five-grade impairment scale as used in the testing of speech and audio coders. Tab. 1 shows this Subjective Grades (SG) scale[15].
- The PEAQ system has been developed in order to get a perceptual measurement scheme that estimates the results of real world listening tests as faithfully as possible. In listening tests for very high quality signals, the test subjects sometimes confuse coded and original signal and grade the original signal below a SG of 5.0. Therefore the difference between the grades for the original signal and the signal under test is used as a normalized output value for the result of the listening test. Tab. 1 also lists the corresponding Subjective Diff-Grades (SDG) which are the output values of the PEAQ system. The PEAQ system comprises two different models, the "basic model" and the "advanced model". Results were obtained applying the advanced model.
- The third system used in the evaluations is the NMR. We used the overall NMR_{total} value expressed in dB to indicate the averaged energy ratio of the difference signal with respect to a just masked signal (masking threshold). Usually, at NMR_{total} values below -10 dB there is no audible difference between the processed and the original signal.

SG	Description
5.0	imperceptible
4.0	perceptible, but not annoying
3.0	slightly annoying
2.0	annoying
1.0	very annoying

SDG	Description
0.0	imperceptible
-1.0	perceptible, but not annoying
-2.0	slightly annoying
-3.0	annoying
-4.0	very annoying

Table 1: Five-grade impairment scales used in listening tests

3.3 Subjective Listening Tests

3.3.1 Listening Test Setup

In addition to the objective measurements a listening test was performed. The high quality indicated by the objective measurements (section 4.2) lead to the decision to choose a forced-choice-test, namely the pair-test for subjective tests. Only experienced listeners were selected to take part at the listening test. The listening test was preceded by a training phase which fulfilled the following requirements:

- only one person present in listening room
- unlimited time for training
- unlimited retries to hear each item in the order Reference-Watermarked
- full a-priori knowledge of what stimuli were presented

Following the training which was recommended to be at least 15-20 min. the listening test commenced. During the test 10 pairs of each item were presented to the subjects. Each pair was chosen by a random generator from the set {R-R, R-W, W-R, W-W}, where "R" denotes the reference item and "W" denotes the watermarked item. The subjects were asked if both stimuli of a particular pair were equal or not, e.g. {R-R, W-W} should be rated as equal, whereas {R-W, W-R} should be rated as distinct. Each correct decision about items being equal or distinct is called a "hit". No quality grading in the sense of a scale, e.g. SG, was to be done by the subjects. Each subject and item produced a result of the general form "k hits of 10 trials". If a subject is just guessing one would expect a mean of 5 hits. An exact statistical analysis needs to be done in order to evaluate the test results.

3.3.2 Statistical Requisites

In the following an item that exhibits no audible distortions while being watermarked is denoted to be "transparent". If distortions are perceptible the term "non-transparent" is used.

The pair-test normally tests for the non-transparency hypothesis, e.g. that distortions are perceived by a subject. It is known from the statistics, that failing to reject the null hypothesis H_0 does not imply accepting the alternative hypothesis H_1. However rejecting H_0 allows to assume H_1 at the selected level of significance. In other words failing the test for non-transparency does not imply transparency and vice versa.[1].

Therefore we have to perform two hypothesis tests. The first one tests for non-transparency and for those items that fail, a test for transparency is to be applied. The following two hypothesis tests show the statistical background for viewing the results in section 4.

[1] Just assume a particular listener is in bad condition and does not notice audible differences.

Test for non-transparency is done by assuming the following hypothesis:

H_1 : A subject can perceive distortions in a watermarked track
H_0 : Distortions are not perceptible, e.g. the subject is just guessing, or $p_{detect} = 0$

In order to accept H_1 we have to reject H_0 with a certain level of significance α. We choose, as common in hypothesis testing, $\alpha = 0.05$. Applying

$$P(T \in \mathcal{B} \mid H_0 \text{ true}) \leq \alpha \tag{1}$$

leads to the so-called critical region $\mathcal{B}_{non-transp.} = \{8, 9, 10\}$ hits, while T denotes the actual number of hits the subject achieved per item[2]. In other words:
If a subject shows more or equal than 8 hits for one item, we assume this item with 95% probability to be non-transparent. This decision rule is used later in section 4.2.

Test for transparency is done by assuming that, if any subject is able to distinguish between original and watermarked items he will perform with at least a probability[3] of $p_{detect} > 0$. We decided to choose $p_{detect} \geq 0.7$ because the selected subjects are very experienced listeners which are likely to detect any distortions if there are any. The hypothesis formulates as:

H_1 : No subject can perceive distortions in watermarked items
H_0 : Distortions are perceptible for subjects with probability of detection $p \geq 0.7$

As in the opposite case we have to reject H_0 in order to accept H_1 on a level of significance α. Again $\alpha = 0.05$ is chosen. The probability distribution density in this case is not a binomial distribution but is given as

$$P(n, k, p_{detect}) = \sum_{i=0}^{k} p_{detect}^{k-i}(1 - p_{detect})^{n-k+i} 0.5^{n-k+i} \binom{n}{k-i} \binom{n-k+i}{i} \tag{2}$$

where n denotes the number of presented items and k the number of exact hits. The distribution density (2) for $n = 10$ and $p = 0.7$ is shown in Fig. 4. With (1, 2) the critical region for $n = 10$ and $p_{detect} > 0.7$ results in $\mathcal{B}_{transparent} = \{0...6\}$ hits[4]. This means that if a subject with $p_{detect} \geq 0.7$ shows less than or equal 6 hits, we assume this item with a 95% probability to be transparent. This decision rule is used later in section 4.2.

[2]Assuming H_0, the underlying probability distribution density P is the binomial distribution $B(10, 0.5)$.
[3]Subjects just guessing exhibit $p_{detect} = 0$.
[4]The difficulty with this test is, that H_0 is a compound hypothesis. To satisfy (1) the probability supremum of the infinite number of distributions given by (2) for $p_{detect} \geq 0.7$ is to be considered to meet α.

3.4 Comparison to Perceptual Coding

The quality of certain perceptual coders is well-known from extensive testing. This gives an additional anchor when the quality of the watermarked audio signals is compared to those processed by a codec. A well-known codec has been selected namely the ISO/MPEG-2 Advanced Audio Coding (AAC) codec [10, 11].

All selected original items were processed by an AAC coding/decoding procedure. The bit rate of the AAC codec was 64 kbit/s, mono. The results of the AAC coded items can directly be compared with the results of the watermarked items. This gives an quality estimation of the watermarking system in relation to the known quality of the AAC codec.

4 Results

4.1 Results of Objective Measurements

Tab. 2 lists the results of the applied perceptual measurement techniques for the test items. While there are some differences, the basic tendencies of all three measurements coincide. According to the table, all items are graded in the imperceptible region, between 4.5...5 (PAQM) or -0.5...0 (PEAQ). Also the NMR_{total} is below -14 dB. As a rule of thumb a NMR below -10 dB is considered to indicate "imperceptible" distortions.

These results indicate the transparency of the watermarked items which should further be validated by a listening test in the next section.

Item	PAQM/MOS	PEAQ/ODG	NMR_{total} [dB]
Harpsichord	4.72	-0.19	-17.8
Castanets	4.67	-0.39	-14.6
Pitchpipe	4.72	-0.05	-19.6
German Speech	4.60	-0.30	-14.5
Bagpipes	4.72	-0.12	-20.0
Gershwin	4.72	-0.24	-14.0
Triangle	4.72	-0.43	-14.1

Table 2: Quality of watermarked items evaluated by PAQM, PEAQ and NMR perceptual measurement systems

4.2 Results of Subjective Listening Tests

As usual in audio quality assessment, objective quality measurement systems give an *estimation* for the quality. This estimation needs to be verified by a listening test. The results of the listening test described in section 3.3.1 are presented in Tab. 3a, while the results of applying hypothesis tests onto these results are shown in Tab. 3b.

Item	A	B	C	D	E
Harpsichord	3	4	5	8	1
Castanets	6	6	3	7	5
Pitchpipe	4	8	5	4	8
German Speech	6	7	4	7	4
Bagpipes	8	7	4	7	7
Gershwin	3	5	5	4	6
Triangle	7	3	7	5	5

Item	A	B	C	D	E
Harpsichord	O	O	O	X	O
Castanets	O	O	O	-	O
Pitchpipe	O	X	O	O	X
German Speech	O	-	O	-	O
Bagpipes	X	-	O	-	-
Gershwin	O	O	O	O	O
Triangle	-	O	-	O	O

Table 3: Results of listening test: a) number of hits by item and subject b) results after applying transparency tests. "X" indicates non-transparent items, "O" indicates transparent items and "-" indicates that neither the non-transparency- nor transparency hypothesis can be accepted.

As shown in section 3.3.2 a hit-count of 8 or more hits accepts the non-transparency hypothesis at a level of significance of $\alpha = 0.05$. Looking into Tab. 3a one can see that this is the case for four entries. These four entries are marked with "X" in Tab. 3b and can be considered to be non transparent at a level of significance $\alpha = 0.05$.

On the other hand section 3.3.2 showed, that a hit-count of less than seven hits accepts the transparency hypothesis for a listener with a probability $p_{detect} = 0.7$ at a level of significance of $\alpha = 0.05$. This applies to 23 items out of 31[5]. They are marked with "O" in Tab. 3b. In other words: The transparency hypothesis can be accepted for 23 out of 31 items.

Please note that it is inherent to hypothesis testing, that the remaining 8 items (marked with "-") neither can be considered transparent, nor non-transparent. What follows from these results is, that no statement can be made for these 8 remaining items.

4.3 Results of Comparison to Perceptual Coding

The perceptual measurement and listening test results presented in sections 4.1, 4.2 need some calibration with known artifacts. For this purpose, the original items have been subjected to coding/decoding with MPEG-2 AAC at a bit rate of 64 kbit/s for a mono signal. Tab. 4 lists the results of AAC coding in terms of audio quality. These data can be directly compared to Tab. 2. From this comparison it can be seen that watermarking of the items introduces remarkably less distortion than coding these items with a state of the art codec. The results shown in Tab. 4 are displayed in Fig. 5,6.

[5]Applying a transparency test to the four non-transparent items is senseless.

Item	PAQM/MOS	PEAQ/ODG	NMR_{total} [dB]
Harpsichord	3.89	-1.00	-12.9
Castanets	3.13	-1.17	-10.8
Pitchpipe	4.72	-0.46	-12.2
German Speech	2.90	-2.39	- 8.9
Bagpipes	4.66	-0.76	-14.3
Gershwin	3.89	-1.36	-10.4
Triangle	4.71	-0.38	-17.9

Table 4: Quality of AAC coded items evaluated by PAQM, PEAQ and NMR perceptual measurement systems

4.4 Decoding Results

In the foregoing sections the watermarking system was evaluated by the achieved audio quality. However, this is not sufficient to describe the system performance.

Looking at the functional principle shown in Fig. 1 it is obvious that there must be a tradeoff between reliability of the data transmission and the perceptibility of distortions. This is due to the fact that increasing the data signal energy will improve the SNR of the data transmission, but on the other hand will introduce more distortion energy into the audio signal which will very likely be noticed by the listener.

Therefore the audio quality results should always be viewed in conjunction with the reliability of the decoding process recovering the additional embedded data. To quantify this reliability we selected the raw bit error rate of the transmission channel. The decoding results of the test items are shown in Tab. 5.

It should be noted that these bit error rates show the raw channel bit error rate. At the current stage of development, the system does not employ any error correction. Strong improvements are expected when implementing error correction codes in this system.

The items showing an upper bound for the bit error rate, exploited no bit errors during the length of the item and therefore a fixed bit error rate cannot be specified.

Item	bit error rate
Harpsichord	0.0136
Castanets	0.0059
Pitchpipe	0.0273
German Speech	$< 3 \cdot 10^{-3}$
Bagpipes	0.0411
Gershwin	$< 3 \cdot 10^{-3}$
Triangle	0.0036

Table 5: Decoding results for watermarked items

5 Conclusions

A system providing an additional data channel within audio signals was presented. It combines the advantages of spread spectrum modulation and psychoacoustic hiding of noise signals. The established channel carries arbitrary data and, of course, can be used to transmit watermarking information, as well as licensing information, serial numbers or any other type of information.

This watermarking encoder was evaluated by perceptual measurement systems, a listening test and a comparison to state-of-the-art coding, while the performance of the watermarking decoder was estimated by the achieved bit error rate of the data transmission.

The results of objective measurements and listening tests were in good agreement, showing that the system introduced only a negligible amount of audible distortion into the audio signal. At the same time a good data transmission was achieved with fairly low channel bit error rates.

Based on these evaluations it seems feasible to apply both watermarking *and* preserve a high audio quality when using sophisticated schemes which sufficiently consider the perceptual aspects of watermarking.

6 Acknowledgments

The author would like to thank Thomas Sporer of the Fraunhofer Institut for Integrated Circuits for his support while writing this paper as well as all the subjects who participated in the listening test.

References

[1] I. Cox, J. Kilian: *A Secure Robust Watermark for Multimedia*, Information Hiding, Ed. R. Anderson, Springer LLNCS 1174, 1996

[2] F. Hartung, B. Girod, *Digital Watermarking of Raw and Compressed Video", Proc. European EOS/SPIE Symposium on Advanced Imaging and Network Technologies*, Berlin, Germany, Oct. 1996.

[3] L. Boney, A. Tewfik, K. Hamdy: *Digital Watermarks for Audio Signals*, IEEE Int. Conf. Multimedia, June 17-23, p. 473-480, 1996

[4] J.F. Tilki, A.A. Beex: *Encoding a Hidden Digital Signature onto an Audio Signal Using Psychoacoustic Masking*, 7th Int. Conference on Signal Processing Applications & Technology, Boston MA, pp. 476-480, 7-10

[5] B. Sklar: *Digital Communications*, Prentice Hall, 1988

[6] J.G. Proakis: *Digital Communications*, 3.Aufl., MacGraw-Hill, New York, 1995

[7] K.D. Kammeyer: *Nachrichtenübertragung* , 2.Aufl., Teubner, Stuttgart, 1996

[8] R. Dixon: *Spread Spectrum Systems*, 3rd. Ed., Wiley & Sons Inc., 1994

[9] C. Neubauer, J. Herre, K. Brandenburg: *Continuous Steganographic Data Transmission Using Uncompressed Audio*, 2nd International Workshop on Information Hiding, Portland/Oregon, soon available in Springer LNCS-series

[10] M. Bosi, K. Brandenburg, S. Quackenbush, K. Akagiri, H. Fuchs, J. Herre, L. Fielder, M. Dietz, Y. Oikawa, G. Davidson: *ISO/IEC MPEG-2 Advanced Audio Coding*, JAES, Vol. 45, No. 8

[11] ISO/IEC JTC1/SC29/WG11 International Standard ISO/IEC 13818-7, *Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding*

[12] J. Beerends, J. Stemerdink: *A Perceptual Audio Quality Measurement Based on a Psychoacoustic Sound Representation*, JAES, Vol. 40, No. 12, 1992 December, p. 963-972

[13] ITU-R Draft new Recommendation ITU-R BS.[10/20], *Method for objective measurements of perceived audio quality*, 1998

[14] K. Brandenburg, T. Sporer: *"NMR" and "Masking Flag": Evaluation of Quality using Perceptual Criteria*, Proc. of the 11th International AES Conference on Audio Test and Measurement, Portland 1992, pp 169-179

[15] T. Sporer: *Evaluating Small Impairments with mean Opinion Scale–Reliable or just a guess*, AES 101st Convention, Los Angeles, 1996

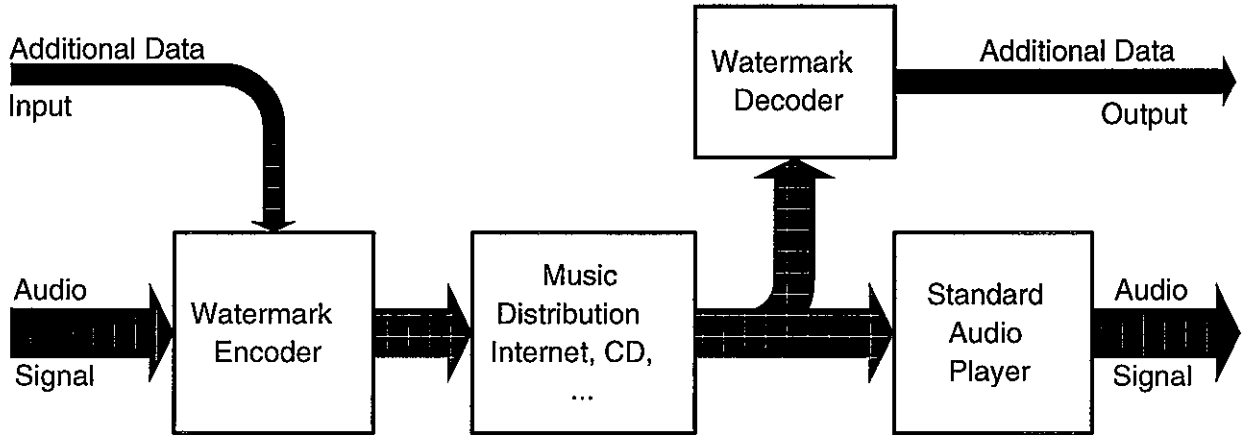

Figure 1: Principle of the psychoacoustic data embedding system

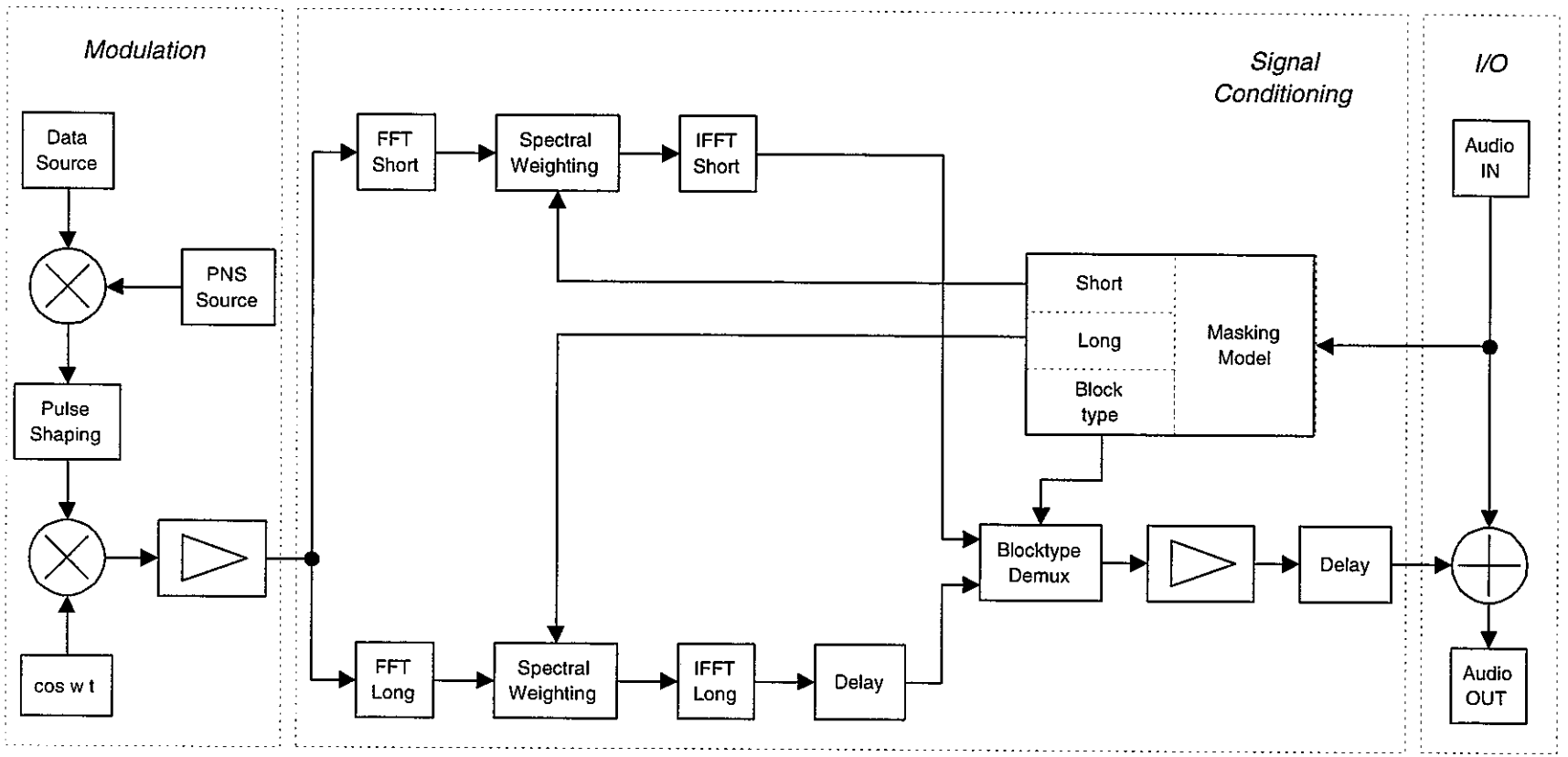

Figure 2: Block diagram of the encoder

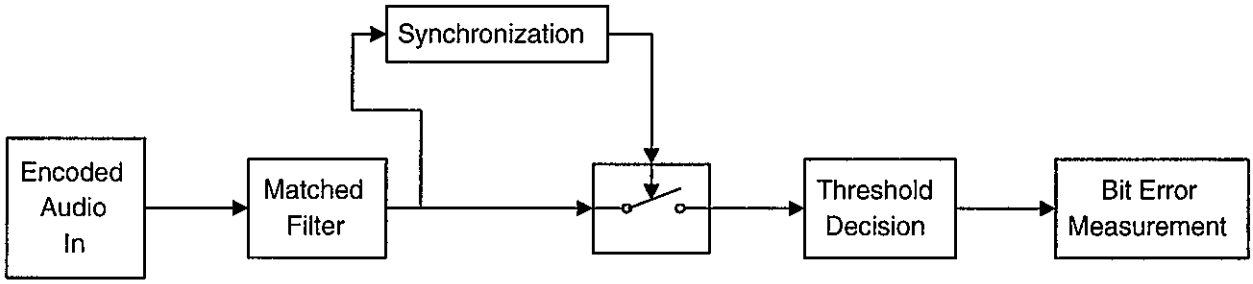

Figure 3: Block diagram of the decoder

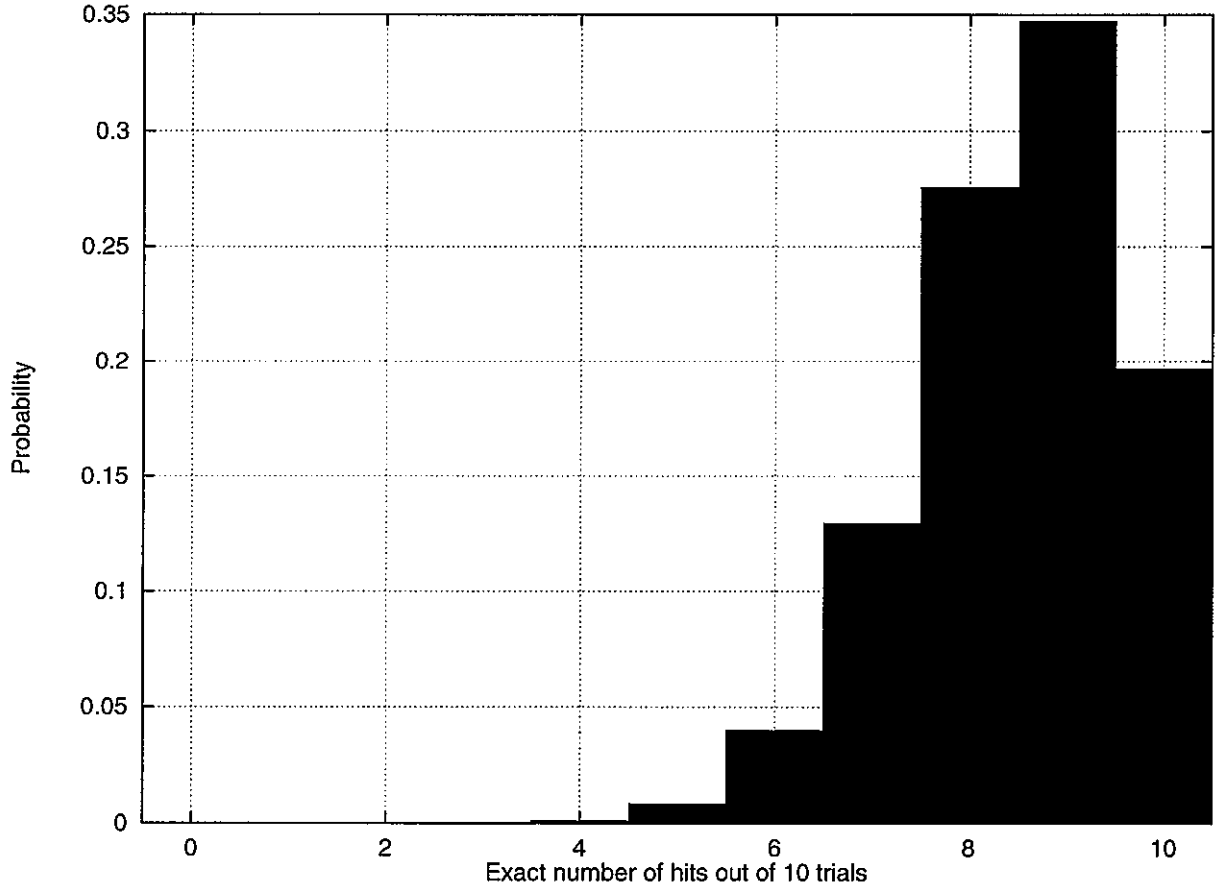

Figure 4: Probability density distribution of the hit-count for a subject with $p_{detect} = 0.7$. Figure shows that 9 hits are most likely for this subject.

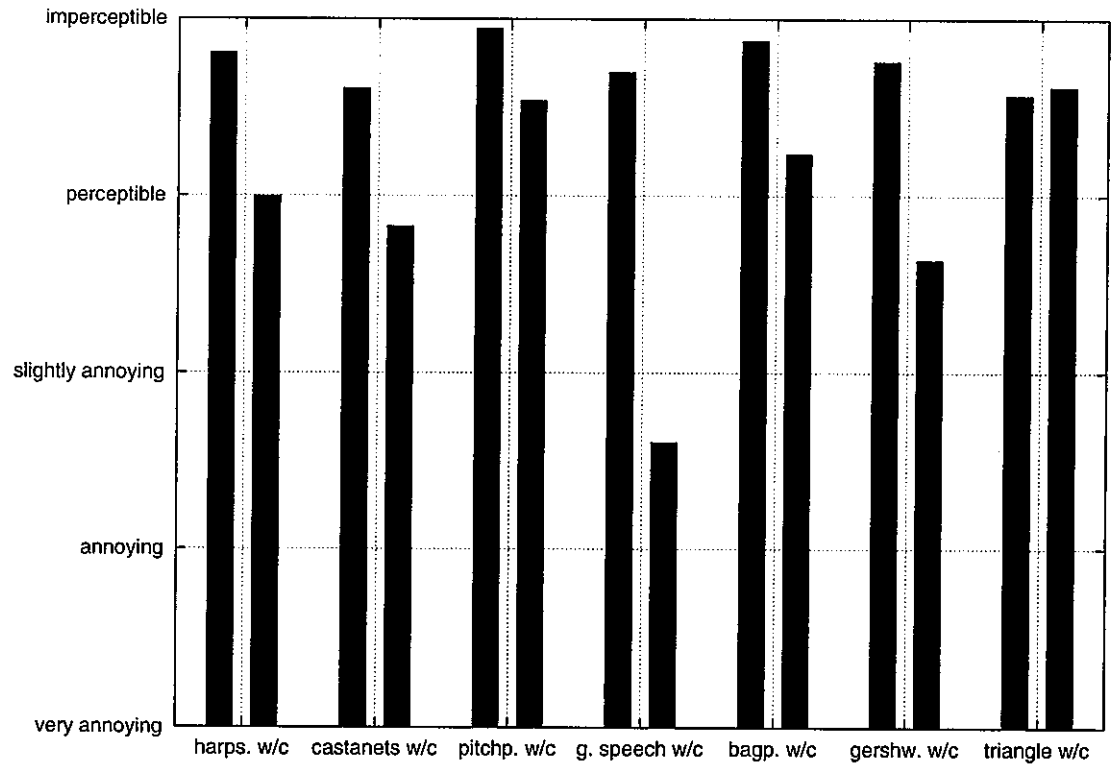

Figure 5: Quality comparison of watermarked items with AAC coded items using the PEAQ system. "w" denotes the watermarked item, "c" denotes the AAC coded item.

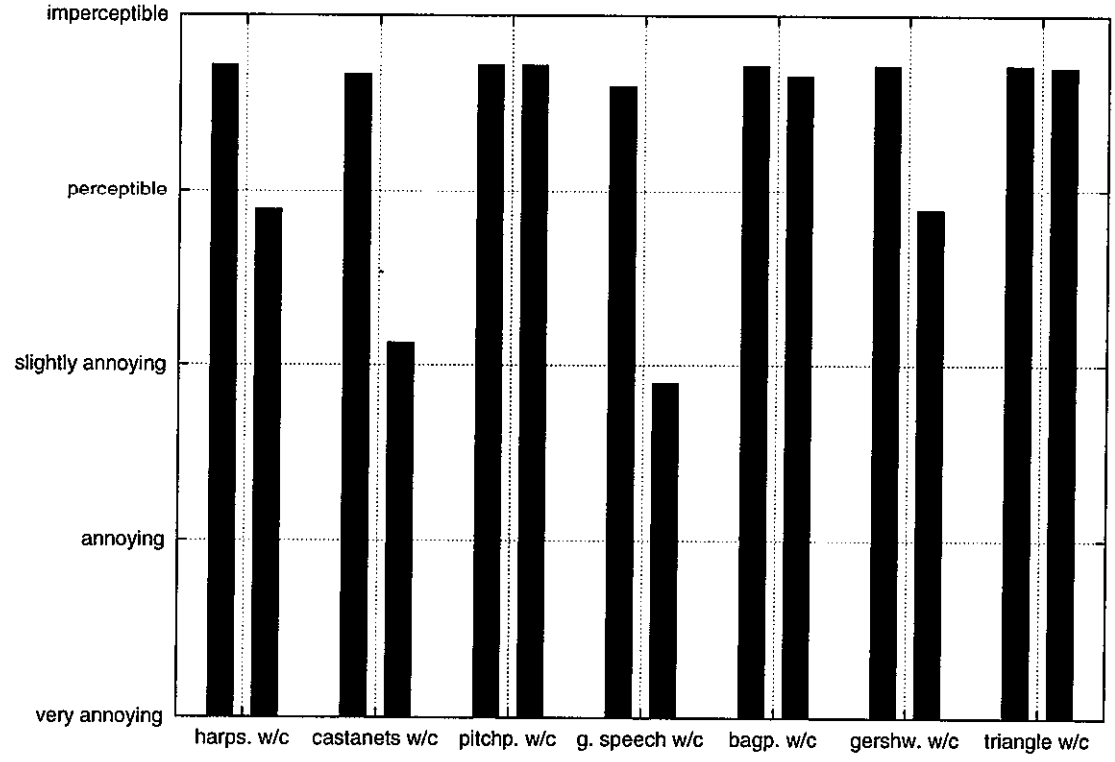

Figure 6: Quality comparison of watermarked items with AAC coded items using the PAQM system. "w" denotes the watermarked item, "c" denotes the AAC coded item.