# Attack tuning - Attack Transparency Models and their Impact to Geometric Attacks

Jana Dittmann and Christian Krätzer and Andreas Lang
*(Authors appear in alphabetic order.)*

Research Group Multimedia and Security,
Department of Computer Science,
Otto-von-Guericke-University of Magdeburg, Germany

**Abstract.** Geometric attacks, like the ones performed by the StirMark Benchmark for Audio (SMBA) evaluation suite for audio watermarking, disturb the detector of a watermark signal $w$ used on the attacked sequence $r$. This paper will introduce SMBA and describe how perceptual modeling (in this case psychoacoustic modeling) can be used to improve the transparency of geometrical attacks. Furthermore transparency models and the impact using psychoacoustic methods on the attacked sequence and thereby on the correlation between $r$ and all the cyclically shifted versions of the watermark signal $w$ are discussed. It is shown that an exhaustive search is complicated by the non-linear behavior of the transformation performed by the psychoacoustic model.

## 1 Motivation and Introduction

Today we find a wide variety of geometric watermarking attacks. Most of the attacks are configured based on a standard parameter set. Using the notation of the WaCha[1] we assume here that the attacker cyclically shifts the marked vector $f'$ (watermarked host feature sequence $f$ with watermarking signal $w$ and template signal $s$) by an unknown amount resulting in an attacked sequence $r$. This is a kind of blind manner to modify the watermarked vector $f'$. To apply an attack successfully the overall goal is to keep the signal quality by changing the marked vector $f'$ and producing the attacked sequence $r$. Besides the overall goal to disable the detector of the watermark signal $w$ the perceptual quality of the signal has to be ensured, otherwise the attack cannot be seen as a successful attack. Therefore in accordance to the exhaustive search (ES) the detector needs to compute the correlation between $r$ and all the cyclically shifted versions of $w$ with additionally considering all possibly applied perceptual models of such a geometric attack. The first question of an ES regarding the effectiveness of ES detection can be extended: apart from complexity issues, a wider scope of geometric attacks has to be considered. The question is if ES can handle perceptual tuned geometric attacks? Our idea is to describe the overall attacker model if the attacker uses an attack tuned according to the human perceptual system. In our

[1] 1st Wavila Challenge, Barcelona 2005

paper we discuss therefore transparency models and their impact to geometric attacks to show how geometric attacks can be tuned and the attack strength can be adopted to the demanded signal quality by reducing the attack strength. Our goal is to motivate what kind of tuned geometric attacks has to be handled by an ES. On the example of StirMark Benchmarking for Audio (SMBA) we introduce three approaches of applying perceptual models and their impact on the overall attack parameter. From this discussion we learn that the exhaustive search needs to consider perceptually scaled cyclically shifts. Remark: As we do not consider actual audio watermarking algorithms in this first stage, we also do not evaluate the overall impact on watermark detection in this first discussion.

This paper is organized as follows: Section 2 introduces SMBA, classifies the existing geometric attacks of SMBA and introduces audio perceptual models (psychoacoustic models). Section 3 introduces our approaches to perceptual attack tuning (including the three transparency models) and the psychoacoustic module for SMBA. In section 4 the test scenario used for transparency evaluations of the geometric attacks of SMBA using psychoacoustic modeling is described. In section 5, we discuss our first test results of perceptually tuned attacks by evaluating the original audio quality, the audio quality after attacks without and without psychoacoustic modeling. The section 6 summarizes our approach and impacts to an ES.

## 2 Attacks of SMBA and Perceptual Models

This section introduces briefly the SMBA architecture, the concept of single attacks and the attack classification. Furthermore, this section introduces a perceptual model which is the base to improve the audio attack transparency.

### 2.1 StirMark Benchmark for Audio

This subsection introduces the general SMBA architecture and classifies the single attacks [1]. The architecture of SMBA consists of four different types of modules. First, the attack module StirMark for Audio (*SMFA*) itself, second the *read_write* stream module to convert audio files into and back into files, which is needed for input and output of audio signals. The third module *SM-Bell* is a wrapper for *SMFA* and *read_write* to make it easier to use. The fourth module *SM-Bell_GUI* is a graphical user interface for *SM-Bell*.

From the overall point of view, a digital audio signal depends on different parameters based on the capturing and sampling processes (with the following default values for SMBA): sampling frequency $SampleFrequency = 44.1$ kHz, sampling quantisation 16 bits ($MaxQantisation = 2^{16}$) and $NumberOfChannels = 2$ (stereo).

Based on the digital audio representation, we differ between time and frequency domain. The frequency domain representation can be provided by transforming the time domain audio signal into the frequency domain for example by using a Fourier transformation [16]. The marked vector $f'$, which will be evaluated, depends on the attack itself, the attack parameters and can be the whole audio signal $S_i$ or any particular subset. We describe the audio signal, which is processed by *SMFA* as $S_i = f' + remainder$, where $f'$ is marked vector and $remainder$ is the untouched part of the audio signal. Depending on the attack, it is possible, that $S_i = f'$ and no $reminder$ exists. SMBA evaluates $f'$ without knowledge about the used watermarking algorithm, $f$, $w$ and $s$. As notation for the attacks of *SMFA* working in time domain, we use $S_i(x)$ as input signal for *SMFA* and $S_o(x)$ as output signal from *SMFA* which is the attacked, modified, marked audio signal ($S_o = r + remainder'$). Depending on the attack and attack parameters, the attacked sequence $r$ can be the whole audio output signal ($S_o$) or any particular subset. The $remainder'$ is the unevaluated part of the audio signal and depending on the attack and attack parameters, $remainder'$ can be equal to $remainder$. The value $x$ is the sample value at a discrete point of time $t_i$ in the input and output stream, we use $x = x(t_i)$. As notation for the attacks of SMBA working in frequency domain, we use $F_i(x)$ to signify the frequency input signal and $P_i(x)$ to specify the phase of the signal represented in the frequency domain. Furthermore, we use $F_o(x)$ and $P_o(x)$ as the corresponding output signal in frequency domain.

The motivation for all attacks in SMBA is to destroy or weaken the embedded watermark signal $w$, as Kutter et. all [17] described for geometric attacks. From the signal processing point of view, we can classify the *SMFA* attacks into three attack classes. The first class adds or removes a signal $k$ to or from $S_i(x)$: $S_o(x) = a * S_i(x) + b * k(t_i)$. The value $a$ scales the input audio signal and the value $b$ scales $k(t_i)$ to a specific limit. The second class can be described as filter attacks: $S_o(x) = F_{Attack}(S_i(x))$, where $F_{Attack}$ is the corresponding attack from this attack class. The third attack class can be seen as modification attacks primary against the watermarking template signal $s$, by modifying the overall structure of the signal representation: $S_o(x) = M_{Attack}(S_i(x))$. Table 1 summarizes all current single attacks of SMBA into these three classes by indicating time and frequency domain.

Table 1: Classification of SMBA attacks [1]

| Add/Remove Attacks | Domain | Filter Attacks | Domain | Modification Attacks | Domain |
|---|---|---|---|---|---|
| AddBrumm | Time | Amplify | Time | Invert | Time |
| AddSinus | Time | Normalizer1 | Time | FFT_Invert | Frequency |
| AddNoise | Time | Normalizer2 | Frequency | CopySample | Time |
| AddDynNoise | Time | Compressor | Time | FlippSample | Time |
| AddFFTNoise | Frequency | BassBoost | Time | CutSample | Time |
| NoiseMax | Time | RC-HighPass | Time | ZeroCross | Time |
| Denoise | Time | RC-LowPass | Time | ZeroLength1 | Time |
| LSBZero | Time | FFT_HLPassQuick | Frequency | ZeroLength2 | Time |
| Echo | Time | Stat1 | Time | ZeroRemove | Time |
| | | Stat2 | Time | PitchScale | Frequency |
| | | FFT_Stat1 | Frequency | DynamicPitchScale | Frequency |
| | | Smooth1 | Time | TimeStretch | Frequency |

Continued on next page

**Table 1 – continued from previous page**

| Add/Remove Attacks | Domain | Filter Attacks | Domain | Modification Attacks | Domain |
|---|---|---|---|---|---|
| | | Smooth2 | Time | DynamicTimeStretch | Frequency |
| | | | | Exchange | Time |
| | | | | Resampling | Time |
| | | | | ExtraStereo | Time |
| | | | | VoiceRemove | Time |

## 2.2 Overview of Perceptual Models

When discussing perceptual models this paper is focused on psychoacoustic models like the one introduced by Zwicker et al. ([2], [3], [4], [5]). These models deal with the relation between measured features of sound (sound pressure, frequency) and their subjective counterparts (loudness, pitch of tone). They link the physical properties of sound waves and perception. Psychoacoustic analysis and modeling in combination with compression algorithms is widely used in current audio standards for example: MP3, Ogg Vorbis, and the compression used in SONYs MiniDisc format. Perceptual coding reduces the size of audio data with rates from one fifth to one twelfth [4] by the removal of all features of the audio signal which are considered to be imperceptible to human listeners. Generally the first item defined by any psychoacoustic model is the audible field (also known as the hearing area). It is defined as the range of pressure changes in the air perceptible by the human auditory system and is given by a relation between the pressure level (in dB Sound Pressure Level (dB SPL)) and the frequency (in Hz, usually ranging from 0 to 20,000 Hz).

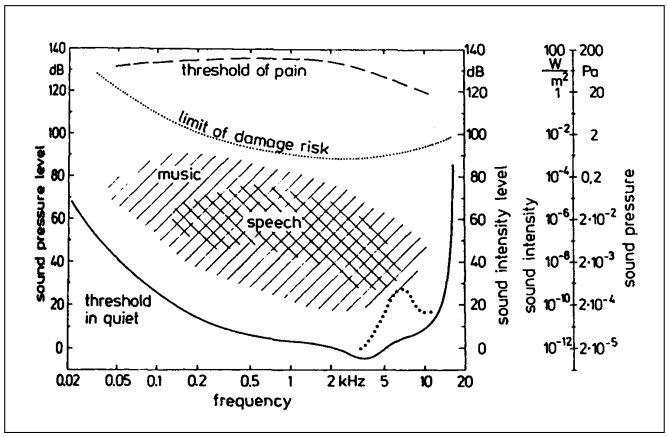


**Fig. 1:** Hearing area. The y-axis is not only expressed in sound pressure level (dB SPL) but also in sound intensity (in Watt per square meter ($W/m^2$)) and sound pressure (in Pascal ($Pa$)). The dotted part of the threshold in quiet stems from subjects who frequently listen to very loud music. (Taken from [5].)

As can be seen in figure 1 the hearing area is limited by a well defined lower bound called the absolute threshold of hearing (ATH). Signals below the ATH are too faint to hear. The ATH changes with increasing age of the subject under test (see [5] and [6]). The upper bound of the hearing area is not as easy to define as the ATH. It is generally described by two curves: the limit of damage risk and the threshold of pain. One of the most important phenomena in human hearing, with respect to processing and measurement, is the occurrence of masking. When two signals are located sufficiently close to each other both in time and frequency, the weaker signal may become inaudible due to the presence of the stronger signal. The signal component that is masked is called maskee and the signal component that masks another one is called masker. The signal level up to which signal components are inaudible due to masking is called the masking threshold or masked threshold, depending on the side from which a masking is looked at. Both terms used are equivalent [5].
Masking results from the limited spectral and temporal resolution of the ear in combination with the non-linear behavior of the human auditory system.

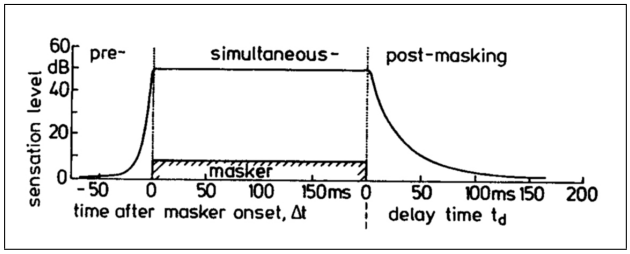


**Fig. 2:** Characterization of the three regions within which masking (pre-, simultaneous- and post-masking) occurs. (Taken form [5].)

In the following paragraphs the categories of masking shown in figure 2 (pre-, simultaneous- and post-masking) are described. As can be seen in figure 2 temporal masking is cut into two separate effects: post-masking and pre-masking. In post-masking (also known as forward masking), signal components are masked after termination of the masker. Apart from the location of masker and maskee in the time-frequency plane, the masking threshold in the case of post-masking also depends on masker duration.
Pre-masking (also known as pre-stimulus masking, backward masking) is usually explained by the assumption that loud signals are processed faster than weak signals and that a masker may therefore overtake the maskee during the processing of the signal, either on the auditory nerve or later on in the higher levels of the auditory system [8]. This covers the phenomenon that a signal can mask another signal before the former one is actually present. Thus in pre-masking signal components are masked before the onset of the masker.
Simultaneous masking is sometimes called frequency masking or parallel mask-

ing. It is the most obvious masking effect. [9] gives a nice example for simultaneous masking displayed in figure 3.

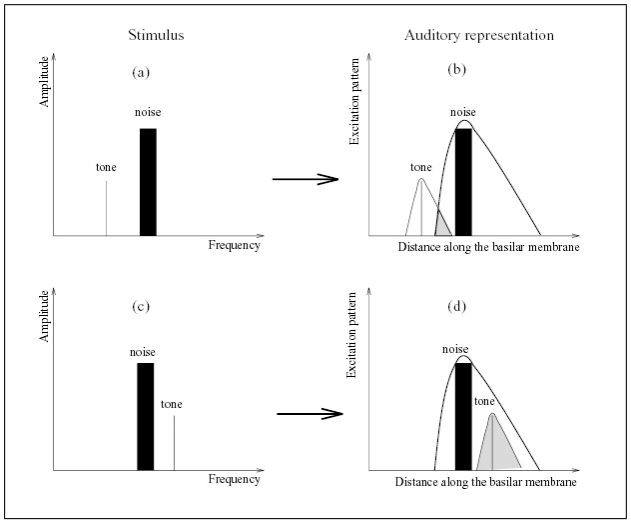


**Fig. 3:** Simultaneous masking. (Taken form [9].)

In the figure 3 a sine wave (a pure tone) and a narrow band of noise are presented simultaneously. The sine wave is at a frequency just below (a) or above (c) that of the noise band. In the first case (b) the sine tone is heard. In the other case (d), excitation pattern of the noise swamps the sine wave and the later is not heard even though their frequency separation remains the same compared to (a).

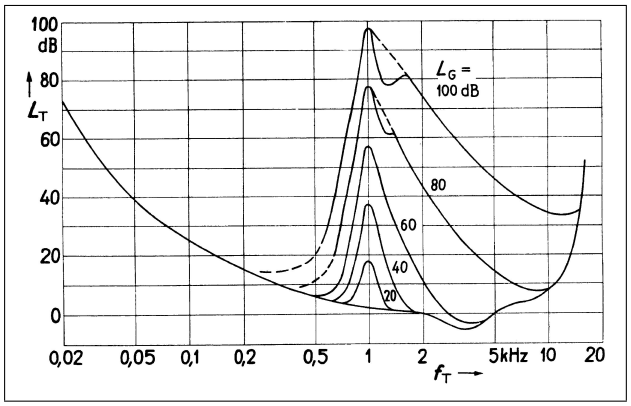


**Fig. 4:** Masking curves for simultaneous masking of tones by narrow band noise ([2]).

Figure 4 shows how dominant the effect of masking in the hearing area is. Masking curves can be approximated by two-sided exponentials when represented as energies over a frequency scale [8]. The low frequency slope (in figure 4 the slope belonging to frequencies below 1,000 Hz) is very steep and depends only slightly on the masker level. The high frequency slope (in figure 4 above 1,000 Hz) is a lot more shallow and strongly depends on masker level. As shown, it is almost as steep as the low frequency slope for low masker levels, whereas it becomes almost flat at very high masker levels.
Among the two categories of masking, simultaneous masking has been examined in more detail than temporal masking effects [8]. Temporal masking seems to be of minor importance for perceptual coding (even though transparent perceptual modeling would be impossible to achieve without the simulation of temporal masking effects).

## 3 Approaches of Perceptual Attack Tuning

This section describes the combination of SMBA and psychoacoustic modeling with the focus on attack transparency improvement. We assume that the correlation between $r$ and all the cyclically shifted versions of $w$ is complicated by the non-linear behavior of the transformation performed by the psychoacoustic model.
Attacks, like the ones performed by SMBA, disturb the detector of the watermark signal $w$ used on the attacked sequence $r$. Strong attacks can, of course, result in strongly modified sequences. Attacks which are perceptible (i.e. lie above a certain perception threshold like the masking threshold of an audio signal) are unsuccessful by definition. Therefore the improvement of the transparency of attacks has to be a main focus to avoid results which are capable of destroying the watermark but are discarded for quality reasons. The attack transparency has to have priority over the attack strength.

In [13] are three different approaches introduced, how a psychoacoustic model can be combined with SMBA. These approaches are also called transparency models and described in the following:

(A) **Pre attack alignment ($T_1$)** Uses a psychoacoustic model to pre-compute the maximal strength and other parameters of the attack, before the attack itself is performed. In this case the psychoacoustic model is a module or stand alone application which runs before SMBA launches the attack.

(B) **Post attack alignment ($T_2$)** After an attack the data is compared by a psychoacoustic model to a copy of the original. The psychoacoustic model is not used directly to calibrate or influence the attack, it only makes quality assessments. The psychoacoustic model in this case is either a module or stand alone application which runs after SMBA has attacked an audio signal.

(C) **Simultaneous/Iterative alignment ($T_3$)** While the attack is running the attack parameters are adjusted (context aware) by the psychoacoustic model to guarantee the quality of the data. The psychoacoustic module used in this case, evaluates the quality after an attack (like in $T_2$ above). If the attack is considered not successful (i.e. audible distortion) it relaunches the attack with the parameters set to a lower level. This process runs in an iteration loop until the psychoacoustic model considers the attack to be successful. This method is, due to the iterations, the most time and computation power consuming of the three.

The transparency model chosen has influence on the resulting attacked sequence $r$. Especially in the case of $T_3$, were the attack strength is optimized incorporating a strict transparency policy, the modifications based on the non-linear methods of a psychoacoustic model will increase the complexity of the ES. Of course, this is bought with an increased computation power necessary to perform the attacks using this transparency model.

The FFT-based psychoacoustic module for SMBA introduced in [12] combines features of $T_1$ and $T_3$. It employs the idea of simultaneousness from $T_3$ without the iteration, and the pre-computing approach from $T_1$ for an extremely limited range (the actual window). In general the model used here computes the parameters of an attack while the attack is running, but not in an iterative way, like suggested in [13]. Instead the attack processes the audio signal window by window and the attack parameters are modified depending on the characteristics of the actual segment.
Two important areas of application for the attacks with psychoacoustic methods were identified:

- Find optimal parameters for "normal" attacks (where only one or two parameters are relevant), like in the examples of AddBrumm and AddSinus.
- Multi-parameterize attacks like BassBoost (where every frequency could be considered independently or context aware).

While the first area of application can be seen as a mere transparency enhancement technique, the second area aims on the safe (under the cover of the masking threshold and therefore by definition imperceptible) increase of the attack strength. Given a fixed quality threshold which has to be maintained, from this second approach a maximal distorted $r$ will result, complicating the correlation between $r$ and all the cyclically shifted versions of $w$ (and thereby the ES).

The psychoacoustic model utilized in SMBA is capable of simultaneous masking techniques but so far does not feature any temporal masking methods. The most important feature of this model is its computation of a simultaneous masking threshold. The mathematical function of this masking is given with:

$$M(x_{Hz}) = max(ath(x_{Hz}), simmask(x_{Hz})) \tag{1}$$

Where $M(x_{Hz})$ is the masking curve at frequency $x_{Hz}$, $max(a, b)$ $(a, b \in \mathbb{R})$ is the maximum function returning the largest of its input values, $ath(x_{Hz})$ is the sound pressure level (in dB SPL) of the ATH (computed by a formula derived by Terhardt [10] from statistical material given by Zwicker et al. [2]) at frequency $x_{Hz}$, and $simmask(x_{Hz})$ is the masking threshold (in dB SPL) at frequency $x_{Hz}$ provided by simultaneous masking.
For the determination of $simmask(x_{Hz})$ a function for the computation of the masking threshold provided by pure tones given by [11] was used. It has to be stated that especially the determination of $simmask(x_{Hz})$ is very computation power consuming. To get the masking threshold at a certain frequency $x_{Hz}$, the masking curve of each tone in the spectrum has to be computed. From the results of these computations the masking threshold for the complete spectrum has to be derived. Then the masking at the frequency $x_{Hz}$ can be read off. Thus the algorithm used has a complexity of $\theta(N^2)$ (with $N$ =20,000; this number is derived from the audible frequency range 0 to 20,000 Hz).
By tuning the SMBA attacks introduced in section 2.1 with psychoacoustic modeling the resulting attacks should become by definition imperceptible.

## 4 Test Scenario

In this section the complete test environment, used to corroborate the hypothesis formed in the preceding section, will be introduced. It consists of the test files, the calibration equipment used, the attacks chosen for a prototypical modification, and the measure principles utilized. For more details of the testing procedure see [12].

**Test Files**
A subset of the SQAM[2] files ([14], [15]) is used for evaluation purposes. These files are tracks from the EBU[3] SQAM disc, with the following characteristics: 44.1 kHz sample rate, 16 bit quantization, stereo. The files have been made available by the EBU on the basis that they are used only for the testing and evaluation of sound systems. The files used for testing are listed in table 2.

Table 2: SQAM files.

| file name | duration | Description |
|---|---|---|
| frer07_1.wav | 34.99s | Electronic tune (Frère Jacques) |
| vioo10_2.wav | 30.07s | Violoncello |
| trpt21_2.wav | 17.86s | Trumpet |
| horn23_2.wav | 12.11s | Horn |
| gspi35_1.wav | 25.92s | Glockenspiel |

Continued on next page

[2] Sound Quality Assessment Material
[3] European Broadcasting Union.

**Table 2 – continued from previous page**

| file name | duration | Description |
|---|---|---|
| gspi35_2.wav | 19.03s | Glockenspiel |
| harp40_1.wav | 16.39s | Harpsichord |
| sopr44_1.wav | 23.66s | Soprano |
| bass47_1.wav | 24.87s | Bass |
| quar48_1.wav | 22.66s | Quartet |
| spfe49_1.wav | 19.02s | Female speech English |
| spme50_1.wav | 17.97s | Male speech English |
| spff51_1.wav | 16.88s | Female speech French |
| spmf52_1.wav | 20.02s | Male speech French |
| spfg53_1.wav | 16.56s | Female speech German |
| spmg54_1.wav | 16.72s | Male speech German |

**Calibration of the Model**
For calibration purposes the AMSL (Advanced Multimedia and Security Laboratory[4]) was used. The core of this testing facility is an anechoic chamber with corresponding audio equipment. In this testing facility typical sound pressure levels for loudspeakers were measured. The knowledge gathered from these measurements was used for calibration of the psychoacoustic model.

**Modified Attacks**
Out of the attack set of SMBA three attacks (AddBrumm, AddSinus and BassBoost) were selected for a prototypical modification. To cover the two fields of application (adjusting single attack parameters and multi-parameterization of attacks) the AddBrumm and BassBoost attacks were chosen. The AddSinus attack was selected for its similarity to the AddBrumm attack to allow for quality considerations.

**Evaluation of Transparency**
Two common approaches for the evaluation of modifications on audio material (like the attacks of SMBA) exist. The first one is the evaluation with listening tests. This method is very time consuming and requires many human test subjects. The second approach is the use of so called objective perceptual measurement techniques. As the measure of choice the Objective Difference Grade (ODG) was chosen, because it is considered to be the only measure directly verifiable against listening test data [8]. The objective perceptual measurement techniques do not have the restrictions of the subjective tests with an audience, but on the other hand still lack acceptance. The reason for this fact is simple: until a model is found which is capable of simulating all the phenomena of the human hearing satisfactorily, objective measures will be considered error-prone.

[4] Research Group Multimedia and Security, Department of Computer Science, Institute of Technical and Business Information Systems, Otto-von-Guericke-University Magdeburg, Germany.

Nevertheless they are a good indicator, which has to be supported by tests with a human auditory, if necessary.
The values for the ODG range from 0 (imperceptible) to -4.0 (very annoying).

## 5 Test Results

This section introduces our test results and discusses them in detail. In [12] all SMBA attacks are discussed in detail whether or not they can be improved by using psychoacoustic methods. As mentioned in the preceding section, three of the SMBA attacks (AddBrumm, AddSinus and BassBoost) were modified in a prototypical implementation using the simple psychoacoustic model introduced in [12].
In the case of the AddBrumm and AddSinus attacks the psychoacoustic model is used to determine the maximum value of one out of two attack parameters (in both cases the parameter *strength* is modified and the attack parameter *frequency* remains unchanged).

Table 3: Evaluation of the modification of AddBrumm and AddSinus.

| **file name** | $strength_B$ | $ODGp_B$ | $ODGn_B$ | $strength_S$ | $ODGp_S$ | $ODGn_S$ |
|---|---|---|---|---|---|---|
| frer07_1.wav | 1424.65 | -3.15 | -1.70 | 1525.88 | -3.89 | -3.76 |
| vioo10_2.wav | 403.7 | -0.88 | -0.97 | 3.31 | -0.04 | -2.27 |
| trpt21_2.wav | 403.7 | -2.24 | -0.95 | 0.959 | 0.03 | -3.51 |
| horn23_2.wav | 1534.97 | -1.80 | -0.70 | 1525.88 | -3.85 | -3.67 |
| gspi35_1.wav | 403.7 | -2.10 | -2.31 | 1525.88 | -3.55 | -3.31 |
| gspi35_2.wav | 403.7 | -1.55 | -1.80 | 727.747 | -2.57 | -2.57 |
| harp40_1.wav | 403.7 | -1.67 | -0.61 | 6.54 | -0.01 | -1.94 |
| sopr44_1.wav | 403.7 | -1.48 | -1.11 | 1525.88 | -0.99 | -2.06 |
| bass47_1.wav | 403.7 | -0.49 | -0.80 | 7.18 | -0.78 | -1.90 |
| quar48_1.wav | 403.7 | -0.38 | -0.38 | 938.827 | -0.44 | -1.76 |
| spfe49_1.wav | 403.7 | -0.64 | -0.88 | 7.143 | -0.56 | -2.14 |
| spme50_1.wav | 403.7 | -0.49 | -0.82 | 3.366 | -0.27 | -2.03 |
| spff51_1.wav | 403.7 | -0.71 | 0 | 6.397 | -0.85 | -2.03 |
| spmf52_1.wav | 403.7 | -0.84 | -1.02 | 7.33 | -0.85 | -2.31 |
| spfg53_1.wav | 403.7 | -0.77 | -0.02 | 14.585 | -0.22 | -2.22 |
| spmg54_1.wav | 403.7 | -0.66 | -0.03 | 5.822 | 0.00 | -2.03 |

In table 3 all files under test are listed. The column $strength_B$ contains the maximum attack value computed by the psychoacoustic model for the attack parameter *strength* (the default value for this parameter is 2500) of the AddBrumm attack, $ODGp_B$ and $ODGn_B$ are the ODG values computed for the files after

performing an AddBrumm attack with ($ODGp_B$) and without ($ODGn_B$) the psychoacoustic module, $strength_S$ is the maximum attack value computed by the psychoacoustic model for the attack parameter *strength* (the default value for this parameter is 120) of the AddSinus attack and $ODGp_S$ and $ODGn_S$ are the ODG values computed for the files after performing an AddSinus attack with ($ODGp_S$) and without ($ODGn_S$) the psychoacoustic module.
As can be seen in table 3 the results vary, while in the case of the AddSinus attack the ODG values, computed on the test files attacked with ($ODGp_S$) and without using psychoacoustic methods ($ODGn_S$), show definitely better results for attack transparency for the attacks with psychoacoustic methods, the resulting ODG values in the case of the AddBrumm attack show only on 50% of the test files under consideration an improvement by the use of psychoacoustic methods (although here strong differences between original and the attacked files can be seen in visualizations of the audio files - see figure 5), which shows the the waveform of a file (spfe49_1.wav, chosen by random) before and after the attacks.

(a) (b) (c)

**Fig. 5:** AddBrumm: Visualization of the time domain of spfe49_1.wav. Subfigure (a) is the waveform of the file after attack with the psychoacoustic module enabled; subfigure (b) is the waveform of the attacked file without the psychoacoustic module, and subfigure (c) is the original file.

In subfigure (b) (the audio signal after the attack with the psychoacoustics module disabled) of figure 5 the distortion by the attack is clearly visible around the centerline, the distortion in subfigure (a) (the audio signal after the attack with the psychoacoustics module enabled) is also visible but much weaker.

In the case of the BassBoost attack the modification approach for this attack was the usage of the psychoacoustic module for the substitution of the $BoostDB$ parameter. Instead of an equal increasing of all frequencies by $BoostDB$ dB, the signal was raised to the masking threshold for each frequency. This approach failed. 15 out of 16 ODG values (all files except trpt21_2.wav) were worse than the value computed for the attack without the use of psychoacoustic methods. This is due to high energy values introduced to the signal by raising all frequencies simultaneously (and by a large amount - at very low frequencies the ATH

is, of course, the dominating effect in the masking threshold and the ATH values for those frequencies lie at 140 dB SPL and higher). After the inverse Fourier transform, the signal in the time domain suffered clipping and overmodulation problems resulting from such large energy values in the frequency domain.
A different approach was searched for and found. Experimentally the psychoacoustic module was used only in a pre attack alignment ($T_1$) (see section 3). The maximum for the attack parameter *BoostDB* is computed by the psychoacoustic module. Then *SMFA* is run with BassBoost and the *BoostDB* determined beforehand. As can be seen in table 4 the results from this test are quite astonishing. The worst of the ODG values for all SQAM files under test for this approach is $-0.35$ which is on the scale between 0.0 (imperceptible) and $-1.0$ (perceptible, but not annoying) and the strongest attack value (*BoostDB*) used is 5.29 dB, which is only short of the default parameter ($BoostDB = 6.123$ dB) for this attack.

Table 4: Evaluation of the modification of BassBoost.

| **file name** | **attack value** ($dB$) | $ODG$ |
|---|---|---|
| frer07_1.wav | 0 | 0.04 |
| vioo10_2.wav | 1.0405 | -0.00 |
| trpt21_2.wav | 0.298 | 0.04 |
| horn23_2.wav | 2.594 | -0.27 |
| gspi35_1.wav | 0.51 | 0.03 |
| gspi35_2.wav | 1.24 | -0.02 |
| harp40_1.wav | 0 | 0.00 |
| sopr44_1.wav | 1.29 | -0.00 |
| bass47_1.wav | 0.007 | -0.00 |
| quar48_1.wav | 0.363 | -0.00 |
| spfe49_1.wav | 2.767 | -0.00 |
| spme50_1.wav | 5.29 | -0.35 |
| spff51_1.wav | 0.69 | -0.01 |
| spmf52_1.wav | 0.81 | 0.02 |
| spfg53_1.wav | 0.36 | -0.02 |
| spmg54_1.wav | 0.89 | 0.01 |

From the facts presented here it can be concluded that by tuning the SMBA attacks with psychoacoustic modeling the resulting attacks will become imperceptible, if the underlying psychoacoustic model performs well enough. Negative results described by [12] can be explained with the limitations of the model used. As can be seen in tables 3 and 4 an improvement of the transparency results normally in a lower attack strength. The reduction of the attack strength depends on the audio data processed, the transparency model and the psychoacoustic model used. More complex psychoacoustic models will result in more transparent and stronger attacks, but will consume much more computation power.

Some objective measures obviously get other results than a subjective evaluation would return. It becomes obvious that subjective testing with an human auditory is necessary to verify the output of research results in this context. Nevertheless objective measures are very useful indicators for evaluation purposes.

## 6 Summary

By tuning geometric attacks like described in this paper with the non-linear methods found in psychoacoustic models, the modification itself becomes context aware, this results in attacked sequences $r$ which are the output of a transparent modification. As a second benefit of the use of psychoacoustic models, attacks could be maximized by refraining from the use of single attack parameters and instead using functions like the masking threshold to parameterize the attack. But the focus of this paper was on the transparency improving results of psychoacoustic methods.
Concluding could be stated, that the goal of this paper is reached by showing the usefulness of psychoacoustic modeling in geometric attacks used in an audio watermark benchmarking environment like SMBA. The exhaustive search needed by the detector to compute the correlation between $r$ and all cyclically shifted versions of $m$ becomes far more complicated. A much wider scope of geometric attacks (resulting from the perceptually scaled cyclical shifts) has to be considered for the exhaustive search.

An open question arises from this paper: How will a test against a real watermarking algorithm perform? Which impact will the use of perceptual (transparency) models have on the watermark detection process?
In a next step the results of this paper have to be evaluated by using a real watermarking algorithm to determine the degree of complexity improvement caused by the perceptual tuned attacks.
A second open question is based on the results of table 3. This table shows similar ODG values for all speech files (spfe49_1.wav, spme50_1.wav, spff51_1.wav, spmf52_1.wav, spfg53_1.wav and spmg54_1.wav). All $ODGp_S$ values are in the range 0 to -0.85 and all $ODGn_S$ values are in the range -2.14 to -2.31. The question which rises from these facts is: Can a classification of audio signals and an adjustment of the geometric attacks based on such a classification improve the results of these attacks?

## Acknowledgements

The work about single SMBA attacks described in this paper has been supported in part by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Effort for transparency evaluation of the audio attacks is sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number FA8655-04-1-3010. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

## References

1. Lang, Andreas; Dittmann, Jana; Spring, Ryan; Vielhauer, Claus: *Audio Watermark Attacks: From Single to Profile Attacks.* To appear in Proc. of ACM Multimedia and Security Workshop 2005, New York, August 1-2, 2005
2. E. Zwicker; H. Feldtkeller: *Das Ohr als Nachrichtenempfänger*, Second Edition, Hirzel Verlag Stuttgart, 1967
3. Eberhard Zwicker, *Psychoakustik*, Springer-Verlag Berlin, 1982
4. Manfred Zollner; Eberhard Zwicker: *Elektroakustik*, Third Edition, Springer Berlin, 1993, ISBN 3-540-56600-7
5. Hugo Fastl; Eberhard Zwicker: *Psychoacoustics. Facts and Models.* Second edition, Springer Berlin, march 1999, ISBN 3-540-65063-6
6. ISO 7029: *Acoustics – Statistical distribution of hearing thresholds as a function of age*, 2nd edition, 2000
7. Th. Thiede; E. Kabot *A New Perceptual Quality Measure for Bit Rate Reduced Audio*, Proc. of the 100th AES Convention, Copenhagen, Denmark 1996
8. Thilo Thiede: *Perceptual Audio Quality Assessment using a Non-Linear Filter Bank*, Technische Universität Berlin, PhD Dissertation, Fachbereich Elektrotechnik, April 1999
9. Daniel Pressnitzer; Stephen McAdams: *Acoustics, psychoacoustics and spectral music*, Contemporary Music Review, volume 19, 2000, pages 33-60
10. E. Terhardt: *Calculating Virtual Pitch*, Hearing Research 1 1979, pages 155-182
11. Kees van den Doel; Dave Knott; Dinesh K. Pai: *Interactive Simulation of Complex Audio-Visual Scenes*, Teleoperators and Virtual Environments number 13, MIT Press, February 2004, pages 99-111
12. Christian Krätzer *Improving Attack Transparency of Audio Watermarks by Using Psychoacoustic Methods*, Diploma Thesis, Research Group Multimedia and Security, Department of Computer Science, Otto-von-Guericke-Universität Magdeburg, P.O. Box 4120, 39016 Magdeburg, Germany, April 30th, 2005

13. Andreas Lang; Jana Dittmann; Martin Steinebach", *Psycho-akustische Modelle für StirMark Bechmark - Modelle zur Transparenzevaluierung* Rüdiger Grimm; Hubert B. Keller; Kai Rannenberg (eds.), Sicherheit - Schutz und Zuverlässigkeit, Informatik 2003 - Mit Sicherheit Informatik, pages 399–410, october 2003, Frankfurt/Main, ISBN 3-88579-365-2
14. *SQAM - Sound Quality Assessment Material*, http://sound.media.mit.edu/mpeg4/audio/sqam/
15. George T. Waters, *Sound Quality Assessment Material Recordings for Subjective Tests*, European Broadcasting Union, Users' handbook for the EBU - SQAM Compact Disc, Avenue Albert Lancaster 32, 1180 Bruxelles (Belgique), April 1988
16. Emmanuel C. Ifeachor, Barrie W. Jervis, *Digital Signal Processing*, Prentice Hall, ISBN 0201 59619 9, 2002
17. M. Kutter, S. Voloshynovskiy and A. Herrigel, *Watermark copy attack*, In Ping Wah Wong and Edward J. Delp eds., IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II, Vol. 3971 of SPIE Proceedings, San Jose, California USA, 23-28 January 2000