Digital Watermarking for Verification of Perception-based Integrity of Audio Data

Digitale Wasserzeichen zur Verifikation der wahrnehmungsbasierten Integrität von Audiodaten

Zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation von Sascha Zmudzinski, Diplom-Physiker, geb. in Berlin (West)
Tag der Einreichung: 20. April 2017, Tag der Prüfung: 9. Juni 2017
Darmstadt 2017
D 17

1. Gutachten: Prof. Dr. rer. nat. Michael Waidner
2. Gutachten: Prof. Dr.-Ing. Martin Steinebach

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik

Lehrstuhl für Sicherheit in der Informationstechnik

Digital Watermarking for Verification of Perception-based Integrity of Audio Data

Digitale Wasserzeichen zur Verifikation der wahrnehmungsbasierten Integrität von Audiodaten

Genehmigte Dissertation von Sascha Zmudzinski, Diplom-Physiker, geb. in Berlin (West)

– gekürzte Fassung –

1. Gutachten: Prof. Dr. rer. nat. Michael Waidner
2. Gutachten: Prof. Dr.-Ing. Martin Steinebach

Tag der Einreichung: 20. April 2017
Tag der Prüfung: 9. Juni 2017

Darmstadt 2017
D 17

Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den

(Sascha Zmudzinski)

Kurzdarstellung

(Abstract in German)

In bestimmten Anwendungsfeldern enthalten digitale Tonaufzeichnungen zuweilen sensible Inhalte. Beispiele hierfür sind historisches Archivmaterial, das in öffentlichen Archiven unser kulturelles Erbe bewahrt; oder Tonaufzeichnungen als Indizien im Kontext von Strafverfolgung und Zivilstreitigkeiten. Angesichts der Leistungsfähigkeit moderner Bearbeitungstools für Multimedia ist dieses Material anfällig gegenüber Manipulationen des Inhaltes und Fälschung des Ursprungs in böwilliger Absicht. Unbeabsichtigte Veränderungen aufgrund von technischem oder menschlichem Versagen können in der Praxis ebenfalls vorkommen. Die Vertrauenswürdigkeit dieses Materials im Sinne von Unverfälschtheit des Inhaltes und Echtheit der ursprünglichen Quelle sind daher kritische Faktoren.

Für die Behandlung dieses Problems stellt diese Dissertation daher ein Verfahren zur Verifikation der Integrität und der Authentizität digitaler Tonaufzeichnungen vor. Es ist unempfindlich gegenüber gängigen Bearbeitungsschritten im Lebenszyklus der Audiodaten, die die subjektive Klangwahrnehmung nur unwesentlich oder gar nicht beeinflussen. Beispiele hierzu sind verlustbehaftete Kompression bei hoher Klangqualität oder verlustfreie Formatwandlung. Es ist das Ziel, *de facto* falsch-positive Detektionen in der Gegenwart dieser – legitimen – Bearbeitungsschritte zu vermeiden, wie sie bei Standardverfahren für kryptografiebasierte Echtheitsprüfungen zu erwarten wären. Um dieses Ziel zu erreichen, wird eine geeignete Kombination der Technologien der ”Digitalen Wasserzeichen” mit audio-spezifischen Hashfunktionen untersucht.

Hierzu wird im ersten Schritt ein geeignetes schlüsselabhängiges Audio-Hashverfahren entwickelt. Es nutzt und erweitert Erkenntnisse des sog. ”Audio-Fingerprintings” aus dem Bereich der inhaltsbasierten Audio-Identifizierungsverfahren. Der vorgestellte Algorithmus (im Folgenden bezeichnet als ”*rMAC*”-Authentifizierungscode) erlaubt die sog. ”wahrnehmungs-basierte” Verifikation der Integrität. Dies bedeutet, Integritätsverletzungen als solche zu klassifizieren (erst) sobald sie hörbar werden.

Als weiteres Ziel werden diese Authentifizierungscodes mittels Audio-Wasserzeichen-Technologie unhörbar innerhalb der Audiodaten eingebettet und gespeichert. Dies erlaubt die Authentifizierungscodes auch über die oben erwähnten, zulässigen Bearbeitungsschritte hinweg mitzuführen und für eine spätere Integritätsprüfung verfügbar zu halten. In dieser Arbeit wird dazu ein existierendes Audiowasserzeichen-Verfahren gezielt weiterentwickelt.

Die Schlüsselabhängigkeit der *rMAC*- und Wasserzeichen-Algorithmen erlaubt es auch, in eingeschränktem Maße, die *Authentizität* der geschützten Audiodaten zu prüfen. Hierzu analysiert diese Dissertation auch den Aufwand für Bruteforce-Angriffe auf das vorgestellte, schlüsselabhängige Verfahren der Echtheitsprüfung.

Die experimentellen Ergebnisse zeigen, dass das entwickelte Verfahren eine gute Trennschärfe bei der Klassifizierung zwischen echten und gefälschten Audioinhalten bietet. Es erlaubt weiterhin eine zeitliche Lokalisierung der Datenveränderungen innerhalb einer Datei. Die experimentelle Evaluation liefert schließlich auch Empfehlungen über geeignete technische Feineinstellungen des Verfahrens.

Über die Frage der wahrnehmungsbasierten Echtheitsprüfung für Audio hinaus, liefert diese Dissertation auch neue allgemeine Erkenntnisse in den Bereichen des Audio Fingerprintings und der digitalen Wasserzeichen für sich.

Die Hauptbeiträge dieser Arbeit wurden auf Fachkonferenzen der Multimedia-Sicherheit vorgestellt. Diese Publikationen wurden anschließend von einer Reihe anderer Autoren zitiert und haben daher weiterführende Arbeiten zum Thema mitbeeinflusst.

Abstract

In certain application fields digital audio recordings contain sensitive content. Examples are historical archival material in public archives that preserve our cultural heritage, or digital evidence in the context of law enforcement and civil proceedings. Because of the powerful capabilities of modern editing tools for multimedia such material is vulnerable to doctoring of the content and forgery of its origin with malicious intent. Also inadvertent data modification and mistaken origin can be caused by human error. Hence, the credibility and provenience in terms of an unadulterated and genuine state of such audio content and the confidence about its origin are critical factors.

To address this issue, this PhD thesis proposes a mechanism for verifying the integrity and authenticity of digital sound recordings. It is designed and implemented to be insensitive to common post-processing operations of the audio data that influence the subjective acoustic perception only marginally (if at all). Examples of such operations include lossy compression that maintains a high sound quality of the audio media, or lossless format conversions. It is the objective to avoid *de facto* false alarms that would be expectedly observable in standard crypto-based authentication protocols in the presence of these legitimate post-processing. For achieving this, a feasible combination of the techniques of *digital watermarking* and *audio-specific hashing* is investigated.

At first, a suitable secret-key dependent audio hashing algorithm is developed. It incorporates and enhances so-called *audio fingerprinting* technology from the state of the art in content-based audio identification. The presented algorithm (denoted as *"rMAC"* message authentication code) allows "perception-based" verification of integrity. This means classifying integrity breaches as such not before they become audible.

As another objective, this *rMAC* is embedded and stored silently inside the audio media by means of audio watermarking technology. This approach allows maintaining the authentication code across the above-mentioned admissible post-processing operations and making it available for integrity verification at a later date. For this, an existent secret-key dependent audio watermarking algorithm is used and enhanced in this thesis work.

To some extent, the dependency of the *rMAC* and of the watermarking processing from a secret key also allows *authenticating* the origin of a protected audio. To elaborate on this security aspect, this work also estimates the brute-force efforts of an adversary attacking this combined *rMAC*-watermarking approach.

The experimental results show that the proposed method provides a good distinction and classification performance of *authentic* versus *doctored* audio content. It also allows the temporal

localization of audible data modification within a protected audio file. The experimental evaluation finally provides recommendations about technical configuration settings of the combined watermarking-hashing approach.

Beyond the main topic of perception-based data integrity and data authenticity for audio, this PhD work provides new general findings in the fields of audio fingerprinting and digital watermarking.

The main contributions of this PhD were published and presented mainly at conferences about multimedia security. These publications were cited by a number of other authors and hence had some impact on their works.

Acknowledgment

I wish to thank various people for their contribution to the research work presented in this thesis.

At first I would like to express my deepest gratitude to my Primary PhD Advisor *Prof. Dr. Michael Waidner* for his supervision of my research work.

My grateful thanks also have to be dedicated to my Secondary Advisor and Mentor, *Prof. Dr. Martin Steinebach* for his constructive advice and patient guidance during the progress of my PhD research. The same is true for his inspiration during the project work on multimedia security as a member of his team at the *Fraunhofer SIT* institute.

My gratitude also has to be expressed to my Mentor Prof. Dr. Rüdiger Grimm (University Koblenz-Landau) for his exhaustive reviewing of this thesis.

My appreciation has to be extended to the *Fraunhofer Gesellschaft* and the *Center for Advanced Security Research Darmstadt (CASED)* for providing the inspiring working environment and for having promoted the research on multimedia security for many years. Additionally, this thesis work had been supported by the German Federal State of Hesse under the *LOEWE* research programme for many years, which I greatly appreciate.

I also would like to thank my colleagues in *Martin*'s team for all the fruitful discussions and feedback on different technical aspects of this research, especially to *Christian Winter*, *Waldemar Berchtold, Marcel Schäfer* and *Daniel Trick*.

I also would like to express my gratitude to *Matthias Schwab* for his editorial remarks and thorough proof-reading of this thesis.

Finally, I wish to thank my family & friends who have contributed to this thesis by their constant encouragement and infinite patience all along this thesis work.

Frankfurt, June 22, 2017

Sascha Zmudzinski

”This is a journey into sound.”

(Geoffrey Sumner, 1958)

The quotation on the previous page was taken from the introduction of the LP album *"A journey into stereo sound … an introduction into ffss"* (*London Records, 1958*). It is one of the earliest vinyl records produced in *stereo*. The album's introduction was narrated by the actor and commentator *Geoffrey Sumner*.

Contents

List of Figures

List of Tables

Chapter 1

Motivation and Introduction

Nowadays, computer hardware and software provide many ways of producing, recording, editing, distributing, and archiving digital multimedia content. Especially modern editing software for digital images, video, and audio data allow applying modifications to that content very effectively and efficiently. As a consequence, such content can be modified inadvertently or even tampered with easily and with leaving little traces, if at all. Examples of significant content that is vulnerable to forgery are petabytes of data from historical archival records preserving the cultural heritage of a civilization or user created content on the Internet. Digital exhibits could also be subject to tampering before being used as evidence by law enforcement agencies or at court. Nowadays content manipulations can be applied easier to multimedia content in digital storage formats than it used to be the Analog Age.

In the public, the challenge of forging digital multimedia content is very often discussed for *pictures* in the first place. And it is common knowledge that photographs have been tampered with as an act of censorship or propaganda since the early days of analog photography in the 19th century. Sometimes, press photos are edited in order to beautify the picture or dramatize the picture. Also the current discussion about so-called ”fake news” and indications of systematic desi-nformation by different entities remind us about the credibility of media data circulating in the press. Hence in the Digital Era, the statement

> *”...seeing is not believing”.*

as by *Farid* or *Zhu* [ZST2004, Far2009] reminds us of this vulnerability.

1.1 Credibility of Digital Audio Data

Like photography, the history of sound recordings goes back to the 19th century, namely to the invention of the *phonautograph* device by *de Martinville* in 1860 or the *phonograph* by *Edison* 1877. Since then an uncountable number of analog and digital audio media have been created. And in certain scenarios, the relevant content is actually contained in pieces of *audio* information, instead of pictures (in motion). This is obviously true for *pure* audio content like voice recording / transmissions or radio broadcast. But it also applies for sound tracks contained in various kind of ”video” data.

Insertion, splicing, deletions, muting, or trimming of audio media can be applied easily and could mean a significant change of what a human user hears and eventually *understands*. For example in a voice recording even an added, modified or deleted statement of short duration (for example: by the words *"yes"* or *"no"*) or a prefix syllable (for example: e.g. by the words *"agreed"* versus *"disagreed"*) can totally reverse its meaning. Even worse, complete audio recordings can be made up from scratch which can spoof a forged origin of a certain piece of audio evidence. And like for digital images, such forgery can be applied and concealed to an average listener easily using audio editing software and other measures: Accordingly, *hearing is not believing*!

Hence, the credibility and provenience for of these media in terms of verifying its genuine and unadulterated state are vital. Examples of such significant content are historical archival material preserving the cultural heritage and digital evidence are outlined in the following.

1.1.1 Preserving the Cultural Heritage in Public Archives

Sensitive audio content worth protecting is present in *historical archival records* that preserve the cultural heritage. The most relevant classes and a few motivating examples of such content are:

- **Historical Speeches:** Examples of famous quotations across different centuries that contribute to cultural memory are part of common knowledge. Recordings of such events have been available since the end of the 19th century. Providing their availability is the natural mandate of public archives and similar cultural institutions.
- **Oral History Interviews:** In addition, public archives provide millions of testimonies of contemporary witnesses from different eras and culture circles. In contrast to the previous example of well-known public speeches, these testimonies preserve the memory and experiences of an uncountable number of "ordinary" people.
- **The News:** News about current events and affairs in the public contribute to contemporary history too. The same is true for collected news content even if it is not broadcast eventually but remains hosted by news agencies. Here, verifying authenticity and integrity of the sources (inquiries, reporters, informers, photographers, camera crew etc.) has always been a major obligation for journalists and writers.

Figure 1.1.: Examples of cultural heritage (from left to right, sources given in parentheses): Inauguration speech *Nelson Mandela* 1994 (unknown South African TV station), Audio tape archive at *Yad Vashem Archive*, Jerusalem, Israel (Baz Ratner/*Reuters*), German TV newscast *"Tagesschau"* (*ARD/NDR*), *Berlin Philharmonics/Sir Simon Rattle* 2012 (*Michael Trippel*)

- **User Created Content:** *Any* kind of user-created content that circulates on the Internet or in private circles expresses the interests and concerns of a society (and can be, in turn, be published in The News again).
- **Radio/TV Programs:** Also trivial everyday TV and radio programs and even commercial advertisement in TV/radio broadcast reflect the cultural variety of a modern civilization.
- **Music:** Serious music and pop music are important elements of the cultural heritage too. Music from various genre and created/performed by various artists, composers or conductors is available.

All the types of audio content mentioned above are available to the public as analog or digital/digitized recordings from archives, museums or other cultural organizations, or are available as commercial products, or are circulating the Internet.

1.1.2 Verifying Digital Evidence

Another important class of significant audio content is *evidence* that is contained in digital audio. Examples are given in the following.

- **Police Interrogations and Court Trials:** The testimony of a witness or victim or the confession of a suspect during police interrogations, court trials (including video hearings) are often recorded using electronic video equipment or simple dictaphones. This is done *exhaustively* in many countries (but not in Germany).
- **Lawful Interception / Electronic Bugging:** Eavesdropping on criminal suspects is carried out by law enforcement in the course of criminal investigation. Here, audio surveillance by electronic bugging of crime scenes or wiretapping a suspect's phone communication or computer in the course of lawful (and also strategic) interception is a common measure for collecting evidence.
- **Military Combat Operations:** In military combat, the combat units and their equipment are often provided with live video transmission and recording devices. Voices, environmental noise (e.g. gun shots), or the radio communication can be used for investigating incidents or even war crimes.
- **Police Operations in the Public:** Often, police operations at major public events involving large crowds (e.g. demonstrations, sports events) are documented by police squads on the

Figure 1.2.: Examples of digital evidence (from left to right, sources given in parentheses): Police interrogation, *Russell Williams* murder case (*Ottawa Police Service*), Leaked footage of US combat operations, *Bradley Manning* case (*US Army*), Body-worn camera at Birmingham Police Dep., UK (*West Midlands Police/Flickr*), Cockpit voice recorder, *Germanwings* 4U9525 airplane disaster (*BEA*)

spot with video cameras, with ”body-cams” worn on the body or with dash-cams in the police patrol cars. Also the protesters sometimes take recordings of the police actions with cell phone cameras etc. In case of illegal conduct on either side this footage can be published or brought into criminal or civil proceedings.

- **Dash Cams:** On-board ”dash cams” in vehicles of private users seem to become more and more popular in some countries. Again, the audio sound track in such user-created video footage can support to substantiate civil claims in case of traffic accidents.
- **Video Surveillance / CCTV:** Finally, video surveillance systems of public areas and private premises have become more and more in use. Under certain legal conditions, such *closed circuit TV* equipment (CCTV) also captures the sound which can be used as evidence in criminal or civil cases.
- **Telephone Communication:** Business-to-customer communication over the phone has been common for decades in retail or home banking services. Sometimes, the call centers make a recording of the phone conversation to provide evidence in case of customer complaints. Further examples of such kind of significant phone calls are recordings of emergency calls to rescue services or the police.
- **Civil Aviation and Shipping:** Commercial airplanes and vessels must be equipped with recording devices (e.g. *cockpit voice recorder / data voyage recorder*). In case of an accident or a disaster, the final minutes of the crew's conversation, announcements to passengers and other sounds/noises have to be held available for later investigation. This also applies to the radio communication recorded at the spot of the air/naval traffic controllers.

The examples on digital evidence given previously have in common that a complete chain of proof is vital. As a consequence, the proposed protection mechanisms should be integrated in a (trusted) recording device like a voice recorder, camera or into the monitoring center.

Examples of audio media corresponding to cultural heritage and digital evidence are visualized in Figure 1.1 and Figure 1.2. In light of these examples, the notion of credibility of audio data will be discussed in more detail in the following.

Finally note that the set of audio test data used in the experimental evaluation in this thesis reflects most of the fields listed previously. For an exhaustive list of the audio test samples the reader is referred to Appendix B.

1.2 Challenges and Objectives in Audio Data Authentication

Before discussing and selecting suitable technical approaches for closer investigation, the most important challenges of the investigated protection system shall be explained.

1.2.1 Integrity and Authenticity in the Audio Domain

The previous overview demonstrates that audio data by itself or as the sound track of video content can represent sensitive content. The general credibility objective of this thesis work comprises that the content is not maliciously tampered with and the confidence about its origin is provided. This directly corresponds to the security objective of *integrity* and *authenticity* which are recalled in the following:

Integrity of Audio Data

It is recapped that the term *integrity* itself has its origin in the Latin expression "*integritās*" which literally means

> *"integritās (Latin): completeness, soundness, 'the undiminished or unimpaired condition of a thing' "*

and figuratively corresponds to *"correctness"* or "*inviolability*". It also refers to the opposite notion of "*tangere (Latin): to touch*" [LS1879]. Well-known standard mechanisms for integrity verification like hash functions or checksums can verify said *"correctness"* and *"completeness"* of its input data very exactly and sensitively. But especially for *audio* data there exist many post-processing operations that, in fact, *do* change the binary representation but leave the audio content in its said *"undiminished"* state from the perspective of an average human listener. For example, after lossless format conversions or marginally lossy compression in the course of post-processing of audio data, it still provides said *"soundness"* [sic!] with regards to what a human user hears or understands.

Hence the notion of "integrity" and the correspondent concept of "maliciousness" have to be defined appropriately and more precisely in practice. This will be elaborated upon in the following Section.

Authenticity of Audio Data:

About the notion of "authenticity" it should first be noted that this term is derived from the expression

> *"autenticus (Latin): original (document), genuine ... 'that comes from the author' "*

from ancient times [Gla1982]. In the context of audio data said *genuineness* shall be imperative for the trusted identity of the individual person, entity or device that *originally* created a particular audio medium. According to the examples above, this might cover from which archive or press agency a medium originates, which voice recorder or video camera device was used, which interviewer or music artist created a certain work etc. But it also covers that the medium itself has to be *genuine* in that it is not completely made-up as a malicious act or that it has not been mixed-up with the actually authentic medium by mistake.

Please note: in the research community on multimedia security the objectives of integrity and authenticity are *both* subsumed under the term *"data authentication"* as a very common denotation.

1.2.2 Modeling of "Perceptual Identity" of Audio

For digital audio recordings not *every* detectable act of data modification during the life-cycle of a media can be regarded as a malicious violation of integrity, as can be seen from the following examples:

1. Some common modifications are "transparent" with regards to sound sensation and remain inaudible to average listeners. Examples are
   - transferring recordings from a *MP3* voice recorder to an Audio CD
   - "ripping" an Audio CD to lossless formats like *linear PCM-WAVE, FLAC*, or *MPEG-4 ALS*.
   - upsampling to 48 or 96 *kHz*
   - lossy encoding at high quality settings e.g. *MP3* at bitrates of 256 kbit/s or higher
   - lossless conversion to the *BWF* or *RF64* format in the course of ingesting such data into audio archives.

   Please see [Ros2009] for detail on the *"preservation of archival sound recordings"* and usage of the common archiving file format *BWF* [Cha1997] and its extension *RF64* [EBU2009].

   Interestingly, even lossy encoding at good sound quality settings is not even considered as *"editing a copyright protected work"* in terms of the strict German copyright laws [Bru2004, Loh2008].

   Such acts usually do not have to be regarded as malicious.

2. Other modifications are in fact audible and can even diminish the sound quality – but do not change what listeners "understand" intellectually about the semantics. Examples are the transcoding of a voice recording to an *MP3* file at medium or lower sound quality settings for the sake of saving storage space or bandwidth. Another example is down-mixing of 5.1 channel sound tracks to stereo.

   The comprehension of the content's meaning would still be "correct" so that the legitimate transcoding cannot be regarded as a malicious act either.

3. Finally, for other audible non-malicious modifications it is indifferent if they mean a change of the "correctness" or not. An example is a noise cancellation filter that attenuates background noises in a voice recording or telephone conversation. Depending on the context such background noise can be both irrelevant or vital: On the one hand, such noise cancellation can improve the intelligibility of the voice communication and indeed *increases / improves* what a listener understands. This is a desired feature for example in everyday hands-free or outdoor telephone conversation. On the other hand, the background noise can reveal important information about the context or the physical environment of the phone conversation. This might be relevant in investigation of criminal cases or other incidents: in this case such noise filtering can inadvertently or even maliciously mask incriminating (or exonerating, resp.) evidence.

The reader should bear in mind that the previous examples are very specific for audio or other kinds of multimedia. The notion of "lossy compression", "analog transmission" or "background/foreground noise" do not apply for textual data (like e.g. in documents), numerical data (like e.g. in financial figures or bank transfers) or binary data (e.g. executable object code or encrypted data). For these kinds of non-multimedia data, even small changes of the content can almost never be tolerated. Hence, the notion of *integrity* will have to be revised and adapted to the case of protecting multimedia from forgery and tampering.

The examples show that in a multimedia context, also the notions of ”semantics” or ”content” by themselves are ambiguous and often context-dependent. And it would take large efforts to extract, process and assess the identity of data semantically by technical means: for example current technologies for semantic-based *information extraction* or *knowledge extraction* are *extremely* wide fields of research that would go far beyond the scope of this thesis after all. In addition, most semantic technologies are focused on *textual* input instead of voice data or even general audio (music, noises etc.).

On the other hand, the ”perceived identity” of general audio data is somewhat easier to model and to measure from a technical point of view: albeit not trivial, human auditory perception and perceptual quality have been deeply analyzed in lossy audio coding, audio analysis, and in cognitive sciences for decades. Mathematical models on *psychophysics* were developed that can predict the influence of *objective* sound modifications on the *subjective* sensation by a listener. An overview of most important findings in psychophysics of human auditory perception, denoted as *psychoacoustics* in the remainder, can be found in the Appendix A.2. For avoiding the need to bridge the ”semantic gap”, data integrity verification will rather be conducted on a *perceptual level* in this thesis work as proposed in the following.

Proposal: The following hierarchical model of different levels of integrity for multimedia is introduced:

1. **Binary Integrity:** This describes the identity of data on the level of its digital representation. Two pieces of audio data are regarded as identical on a *binary* level if all data words are identical, i.e. if they are identical ”bit by bit”.

2. **Perceptual Integrity:** This describes the identity of what an average human user *perceives* from a multimedia data set. Here, the sensation of what he/she *hears* of the data is the relevant criterion. Two pieces of audio data are regarded as *perceptually identical* if they sound identical. This means that average users could hardly tell these audio data sets apart by listening to the data.

3. **Semantic Integrity:** This reflects the *meaning* (i.e. the *semantics*) that is associated with the data. Semantic identity is maintained when a user *understands* the same from the two compared data sets or, more generally spoken, when they correspond to the same *notion*.

For simplification purposes, the different levels are regarded in *ascending* order (see Figure 1.3): for example, perceptual identity is a sufficient condition for semantic integrity. This definition is inspired by the model of so-called *soft authentication* described by *Zhu, Tewfik* and *Swanson* [ZST2004]: the authors distinguish manipulations that preserve the perceptual quality (*denoted by the authors as ”quality-based”*) from manipulations that preserve the semantic meaning (denoted as *”content-based”*).

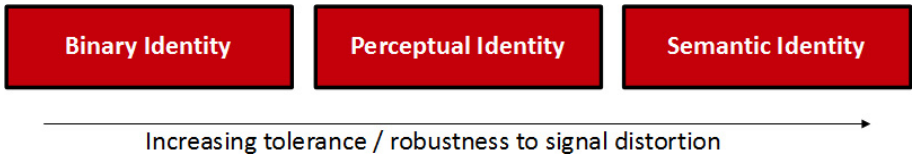


Figure 1.3.: Hierarchical integrity model used in this thesis work

As a convention, the strict requirements of exact integrity verification as it the case for crypto hashes etc. are relaxed for protection of audio and other multimedia data: The investigated approach shall rather provide a certain degree of tolerance to admissible or at least to inaudible modifications of audio. Perceptual Identity can and will be utilized as necessary criterion for verifying/falsifying the semantic integrity of a piece of audio data throughout this thesis work.

1.2.3 Challenges and Objectives

Mechanisms for protecting data integrity from cryptography for arbitrary or data or multimedia data, but also from general communication or computer science have been known for long. But only a few are in line with the relaxed integrity modeling of perceptual integrity.

In light of the real-world examples explained so far, this thesis work will use, extend and develop algorithms from multimedia security research to verify the perceptual integrity of digital audio recordings. To achieving this, the following objectives are defined:

- **Sensitivity to audible modification:** For the detection of malicious tampering it is obviously desired that the system utilizes authentication codes that indicate that two audio signals (or parts of it) are perceptually different. That shall allow detecting malicious doctoring.
- **Invariance to inaudible modification:** The system shall be invariant or, at least, only little sensitive to common transformations of audio media that leave the audio perceptually similar. Unlike for cryptographic integrity verification techniques, admissible and common post-processing steps shall be tolerated in light of perceptual integrity defined above.
- **Localization:** The motivating examples from Section 1.1 show that in some scenarios even doctoring of short section of a few seconds duration or less can cause significant changes of the semantics in a voice recording. Thus, the design of the system should provide a sufficient temporal localization of an attack. Localizing a potential attack on the data could even help identifying the attacker's intent.
- **Assessment:** In addition, the proposed approach should give a graduated estimate on the severity or likelihood of data modification caused by the attack. This is in contrast to plain binary "accept/reject" verification results as for example in common crypto primitives or protocols.
- **Availability / Embedding:** As explained before, in large amounts of multimedia data it can be difficult to retrieve the correct original recording or its separated authentication codes (e.g. as in hash databases). This applies especially when large amounts of multimedia data have to be handled for example in stock photo/video agencies or in public audio archives. Another challenge is that for long-term archiving applications it can be expected that the data will be transcoded to other (i.e. newer) file formats from time to time. Hence

it is desirable to keep the verification codes ”inside” the protected media data instead of maintaining them as separate meta data as overhead in a proprietary file format or the like.

- **Source origin authentication:** As a minor thesis objective, it is desirable to verify the origin and/or the original (*sic!*) creator of a sound recording.
- **Security:** The system should withstand attacks aimed directly at the respective algorithms. Efforts for deliberate doctoring by exploiting weaknesses of the protection mechanisms shall be too high to carry out.

1.3 Structure of the Thesis

The structure of this PhD thesis is given as follows. This first Chapter has given the overall motivation for the authentication challenge for this thesis' subject. Special focus was put on the modeling of the term 'integrity" in light of protection audio data. This allowed identifying objectives for a suitable authentication protection mechanism by means of digital watermarking and perceptual audio hashing.

The following Chapter 2 recalls at first existing technical solutions for integrity protection. This will show that and why a suitable combination of digital watermarking with perceptual hashing is a suitable approach to address the thesis challenges. Subsequently, the state of the art in the techniques of audio hashing, watermarking and perceptual modeling is discussed as their respective combination is the technical basis for this thesis work. Finally, the related work in the research area of authentication mechanisms based on audio hashing/watermarking is recalled and analyzed for need for further research. For the sake of readability of that Chapter, correspondent technical background that is utilized in this thesis work is presented in the correspondent Appendix A for the interested reader.

In Chapter 3 the proposed approach is explained. This includes proposed enhancements to existing audio hashing and audio watermarking approaches and their eventual integration into an authentication watermarking system.

Security aspects of the audio hashing and watermarking components with regards to brute-force attacks are evaluated theoretically in Chapter 4.

The empirical evaluation of the proposed approach is given in Chapter 5. Note that the list of all audio test samples that were used in the experiments can be found in the correspondent Appendix B. For the sake of readability of the Chapter, additional Figures showing plots of exhaustive test results are given in Appendix C.

Finally, in Chapter 6 the results of this thesis work are summarized with regards to major contributions and shortcomings as well. This allows identifying potential for future research and the overall usefulness of the proposed solution as a protection mechanism.

Chapter 2

Literature Review

This Chapter at first gives an overview and a comparison of general integrity verification mechanisms. This shows why the investigated approach is an interesting and relevant research field of investigation for audio authentication. Then the definition and the ”state of the art” in the technologies of audio hashing, digital watermarking and psychoacoustic modeling are recalled as they are the technical basis for this thesis work. Finally, the related work in authentication watermarking is reviewed. This allows identifying the need for further research and classifying the proposed approach.

2.1 Existing Integrity Protection Mechanisms for Digital Data

The following Section gives an abstract categorization and a concrete overview of existent protection mechanisms in light of integrity protection for audio data. That allows distinguishing the selected approach from the state of the art.

2.1.1 General Categorization Criteria

The existent mechanisms can be characterized according to the following criteria:

- **Integrity Model (Binary versus Perceptual Identity):** Most mechanisms aim at verifying or maintaining the integrity exactly on a binary level. Other approaches are more ”tolerable” as they immanently tolerate certain deviance in the data from its original state.

  Note: The tolerance / invariance to admissible distortions is commonly denoted as *robustness* which will be used in the remainder.

- **Modification of the Cover Content (Active versus Passive Protection):** *Active* mechanisms cause a modification of the data at the point in time from which the protection becomes effective. The modification can be reversible or irreversible. In *passive* approaches the data remains unchanged at protection time.

- **Workflow (Pre-Processing versus Post-Processing):** Some mechanisms require a preparation or *pre-processing* stage before the protection becomes effective. Other approaches allow being applied on the respective data at any arbitrary point in time without previous preparation. Obviously, such *post-processing* is only possible for passive mechanisms.

- **Protection Objective (Verification versus Prevention versus Reconstruction):** A number of mechanisms do not technically prevent data modification but allow *verifying* the integrity at a later point in time. Other mechanisms maintain the integrity by actively *preventing* data modifications or, at least, by allowing a *reconstruction* of the original state of the data.

2.1.2 Overview on Integrity Protection Mechanisms

The most common algorithm classes for integrity protection suitable for audio data are discussed and compared in the following. Pure hardware-based solutions like WORM storage media (*"write once, read many [times]"* like optical CD-R media), write blocking equipment (like in forensic hardware), or quantum key distribution are not discussed in detail for the sake of simplicity.

2.1.2.1 Cryptographic Hashes

Cryptographic hashes ("crypto hashes") are standard approaches in cryptology for integrity verification. They are defined as algorithms that map a binary input data set c_0 of arbitrary length to a compact, binary message digest $H(c_0)$ of *fixed* length, typically 128 to 512 *bit*. The crypto hashes are – by design – extremely sensitive: even the slightest change in the input data causes significant changes of the output hash due to an intended algorithmic "avalanche effect". According to *Preneel*'s work on the *"Analysis and Design of Cryptographic Hash Functions"* [Pre1993], the first formal definition was given by *Damgård* not earlier than 1988 in [Dam1988].

Especially *collision-secure one-way hash functions* feature a number of properties that are vital to meet cryptographic security objectives. As explained by *Preneel* or by *Menezes et al.* [MVO1996], the most important are

- **First Pre-image resistance / *one-way* property:** The hash function must be *one-way*, i.e. for any arbitrary value h_0 it must be "hard" to find a matching data set as pre-image c_0 so that $h_0 = H(c_0)$.
- **Second pre-image resistance:** For a known given data set c_0 with its hash $h_0 = H(c_0)$, it must be "hard" to find another data set $c_0' \neq c_0$ as pre-image so that $H(c_0') = H(c_0)$. Such pair is denoted as a *hash collision*.
- **Strong collision resistance:** It must be "hard" to find/create a pair of two distinct data set (c_0, c_1) that hash to an (arbitrary) identical result $H(c_0) = H(c_1)$.

The term "hard"' reflects that finding a 1st pre-image or creating collisions (the "birthday attack") is computationally demanding and the task cannot be solved with reasonable efforts in light of the particular application or desired security level.

Widely used crypto hash approaches are *MD5* [Riv1992], the *Secure Hash Algorithm SHA-1* [EJ2001] (still) and its more secure successors *SHA-2* and *SHA-3*, or the family of the *RACE Integrity Primitives Evaluation Message Digest (RIPEMD)* [DBP1996].

Beyond the context of cryptographic verification of integrity, hash functions are a widely used auxiliary algorithms for the general purpose of data identification (i.e. indicating the *identity* of data sets). For example, they are commonly used applied in sorting and matching algorithms, for processing database queries or in data de-duplication.

Categorization: Crypto hashing is a passive verification mechanism which requires pre-processing at protection time. By design, it does not provide any robustness due to the intended algorithmic "avalanche effect". Security of crypto hashes with regards to pre-image and collision resistance is heavily investigated in cryptanalysis. For example, successful collision attacks for MD5 and SHA-1 were found in the meanwhile [WYY2005, SSA+2008, SBK+2017].

2.1.2.2 Message Authentication Code (MAC)

In extension to crypto hashes, the construction algorithm of some hash algorithms allow for including a dependency from a secret key K. Such kind of key dependent hash $H(c_0, K)$ is commonly denoted as (keyed) *message authentication code (MAC)*. An example is the *CMAC* which is derived from block-based encryption functions [Dwo2005]. Another example is the family of *HMAC* (like *HMAC-SHA1, HMAC-MD5*) which is algorithmically derived from cryptographic hash functions as explained above [KBC1997]. They are being used in the *TLS* protocol for integrity verification.

Note: MACs also allow verifying the *authenticity* of the data origin to some extent: if the MAC of a given data set is verified as "correct", the creator of the MAC'ed data must be in the set of users or entities that have access to the *symmetric* MAC key.

Categorization: MACs can be categorized identically to crypto hashing.

2.1.2.3 Fuzzy Hashing

A new family of hash functions allows identifying data sets that are only *almost* identical. This so-called *fuzzy hashing* is realized by piecewise/partial hashing of elaborately selected data subsets. Unlike crypto hashes, the fuzzy hash values of almost identical data sets are *almost* the same (e.g. in terms of *Hamming* distance).

The approach and the terminology was introduced by *Kornblum* [Kor2006a, Kor2006b]. Fuzzy hashing is well suited for searching, matching, and forensic applications like matching executable binary code or human-readable text strings. Frequently-cited algorithms are *ssdeep*[1,2] [Kor2006b], the *F2S2* algorithm [WSY2013] by *Winter, Schneider* and *Yannikos* or *saHash* [BZLB2014] by *Breitinger* and *Baier*.

> For more detail and references to other fuzzy hashing algorithms, please refer to the introduction given by the thesis author in co-authorship in [SYZW2015][3].

Categorization: Similar to crypto hashing, fuzzy hashing is a passive verification mechanism which requires pre-processing at protection time. It provides a certain level of robustness to minor data modifications in the input data. However, because of their algorithm design, fuzzy hashing approaches are not well suited for multimedia applications.

[1] The *ssdeep* project: *http://ssdeep.sourceforge.net*

[2] All URLs in this thesis lastly retrieved and verified in February 2017 (if not specified otherwise)

[3] References to scientific publications by the author of this thesis are indicated in the remainder with a bar mark on the left page margin. The exhaustive *"List of Publications"* published by the thesis author can be found on pp. 157.

2.1.2.4 Perceptual Hashing

The technology of *perceptual hashing* is an identification technique for digital multimedia. As in every hash algorithm, an input data set of arbitrary length is mapped to a compact, content-based descriptor or message digest of very short length. In the case of *perceptual* hashing, the hash extraction process is insensitive or even invariant to inaudible/admissible transformations of the input signal. Equivalent denotations are *robust hashing* and also *audio fingerprinting*.

This is achieved by extracting and post-processing of perceptually relevant audio features, so-called *robust*[4] features. Robust audio hashing is sensitive to moderate distortion like, for example, caused by lossy *MP3/AAC* or *AVS* compression [Gao2005] or DA/AD conversion in the course of analog transmission (loudspeakers, microphone etc.).

Perceptual hashing allows identifying audio data that *sounds* similar to each other. Conversely, it allows identifying data modifications as soon as they become audible. Thus, perceptual hashing will be used in the context of this thesis to identify audible modification as an indicator for doctoring of the audio.

A more detailed explanation of perceptual (audio) hashing will be given in Section 2.2.

Categorization: According to the categorization explained before, perceptual hashing requires pre-processing at the point of time from which the protection shall become effective It is a passive mechanism that allows future integrity verification at a later date. The algorithms provide a certain level of robustness to data modifications in the audible multimedia content.

2.1.2.5 Digital Audio Watermarking

Digital watermarking is a technique for hiding a secret binary message silently into multimedia data. The embedding is achieved by deliberately applying minor modifications to the content that *represent* the symbols of the message to be hidden. The message embedding is done without changing the file format or file size. In audio watermarking the perceived sound quality is maintained as well. To achieve this the degree of embedding modification is controlled and minimized by using skillful mathematical psychophysical models (”psychoacoustics”) that reflect human perception.

If carefully carried out, the embedded audio watermark message is inaudible *and* it can tolerate moderate distortion of the marked content – at the same time. The latter property is denoted as *robustness* also in the watermarking research area: detection of the embedded message symbols at a later date is possible even if the marked content undergoes certain post-processing for example lossy *MP3/AAC* compression or analog transmission.

Integrity protection based on watermarking can be realized as follows: The embedded message represents an inaudible ”digital seal” that breaks if malicious doctoring is applied to parts of the watermark protected media. This elaborated upon in Section 2.3.3.

Categorization: Watermarking is an *active* protection mechanism that allows integrity verification. Note that with a few approaches even a reconstruction of the original state is possible to some extent if the spatial resolution of verification is very high.

See Section 2.3 of this thesis for more detail on watermarking in general.

[4] Note that the term *robustness* shall denote the tolerance (or even invariance) to minor data modification in the remainder of this thesis.

2.1.2.6 Checksums

Checksums are well-known error detecting codes in computing and communication technologies which allow indicating transmission errors. Like in hashes and MACs, an input message of arbitrary length is mapped to a short message digest of very short length.

Very common is the *cyclic redundancy check (CRC)* proposed by *Peterson* and *Brown* [PB1961] in the 1960s based on the modulo/remainder of the binary polynomial divisions. The typical length of CRC codes is 32 *bit* as in the *CRC-32* standard. Other common checksum types are parity checks, *Fletcher* codes, speed-optimized *Adler* codes, or (weighted) *cross summing* like in *ISBN* book identifiers, serial numbers of bank notes, ID cards/passports etc.

Categorization: Like hashes and MACs, checksums are passive mechanisms that provide a separated verification code. The CRC output is very sensitive to any kind of input data modification. It can be shown that it provides no security against malicious tampering of the data. For example, even if the construction of the polynomial division in CRC-32 codes was extended to 128 *bit* or greater, collisions can be created easily with little computational efforts due to the insecure algorithm design of CRC codes.

2.1.2.7 Digital (Multimedia) Forensics

Digital Forensics in general is the discipline of acquiring, analyzing and reporting of/about evidence in digital data. The objective is collecting information or traces thereof from hard drives, mobile devices, remote storage clouds etc.

Forensics especially for *multimedia* data provides powerful approaches for

- reconstructing the history of modifications,
- identifying meta data about the data's origin (like location and time of recording),
- identifying the recording conditions (the microphone, encoder device/software etc.), or
- recovering deleted fragmented files by exploiting the file syntax (so-called *file carving*).

This is usually based on analyzing skillfully selected data features. Tampering is indicated by finding discontinuities or other unexpected statistical anomalies in the suspicious data set. For example, forensic for audio data allows

- identifying editing of audio data by analyzing the reverberation/echo characteristics in a microphone recording caused by the room acoustics as proposed by *Mailk* and *Farid* [MF2010],
- identifying date and time of a recording based on the 50 Hz/60 Hz sound caused by the mains hum of the *electric network frequency (ENF)* [GS2012, CGW2012],
- identifying the *MP3* encoder software version as published by *Böhme* and *Westfeld* [BW2004] which allows identification or repudiation of a certain recording device as data source,
- detecting the presence of double compression in *MP3* or *AMR*-compressed audio file which can indicate that an authentic file was opened, then modified and eventually saved (hence compressed) again [BDRF+2013, LYH2014],
- file carving for deleted and fragmented *MP3/AAC* audio data, as for example published by the author beyond this thesis work [ZTS2012, SYZW2015].

Categorization: According to the categorization explained above, forensics obviously is a passive technique. Its robustness characteristics are very varying across different subcategories of multimedia forensics. In addition, most publications lack a thorough discussion of their vulnerability to targeted circumvention as was shown for example by *Nguyen* and *Katzenbeisser* [NK2011] for image forensics. Hence, traces be removed or "implanted" into digital data so that a lot of forensic algorithms can be fooled. A virtual "arms race" between forensic, counter forensic, anti-counter forensic (and so on) can be observed in the literature after all.

2.1.2.8 Forward Error Correction

One discipline for maintaining the data integrity in noisy transmission channels is *forward error correction* (FEC). This *channel coding* technique from communication theory allows overcoming transmission errors by encoding the binary input message to a new representation with many redundancies. This encoded message is transmitted instead of the original message. On the receiver side, a correspondent decoder can cope with transmission errors to some extent and can retrieve the correct original message again. For example, FEC is utilized in a vast number of communication standards (like Ethernet, GSM, Digital Video Broadcast (DVB), satellite links, Bluetooth, WLAN), for storage media (like CDs/DVDs, Solid State Disks), or in optical QR codes.

Well known examples of such channel coding schemes are *Hamming codes* [Ham1950], *RS codes* invented by *Reed/Solomon* [RS1960], *BCH codes* developed independently by *Bose/Chaudhurim* and by *Hocquenghem* [BRC1960], or the *Turbo codes* by *Berrou* [BGT1993] for general purpose. The interested reader is referred to the textbook by *Moon* [Moo2005] which is *the* standard reference for FEC techniques.

Categorization: FEC is an active mechanism that intends to reconstruct the original state of data instead of identifying potential data modifications explicitly. Hence FEC is not suitable for achieving the main integrity objectives of thesis work because in the case of intentional breaches of data integrity they would be "repaired" instead of being indicated explicitly. Another drawback is that the FEC-proprietary encoded representation is not compliant with the input file format. Even more, it increases the data payload by several times of the original size in.

Note: Although it might appear *abstruse* discussing FEC in light of this thesis work, this technique is however useful as an auxiliary means: *Turbo* coding is used for increasing the robustness of the proposed watermarking approach, see Section 3.4.2.

> Own research on *Turbo* coding was conducted as a separate research activity by the thesis author. It was investigated and successfully improved in the course of a supervised student thesis by *Berchtold* [Ber2008]. This is not further elaborated upon in this thesis.

2.1.3 Comparison and Conclusion

Among the mechanisms described above, only perceptual hashing, digital watermarking, and some multimedia forensic algorithms allow integrity verification even if the content undergoes legitimate modifications due to their robustness. Interestingly, robust watermarking and perceptual hashing can be tuned so that they even provide the same level of insensitivity to post-processing operations, for example down-mixing from stereo to mono, lossy transcoding (*MP3*, *AAC* etc.) and even analog microphone recording.

That is, while common cryptography mechanisms fail when content is transcoded or taken from the digital to the analog domain, watermarks and perceptual hashes can persist and keep the content protected here. Hence other solutions cannot be considered further for being investigated in this thesis work.

The elaboration on the different technologies for integrity protection from the previous Section is summarized in Tables 2.1 and 2.2.

Mechanism	Protection Objective	Robustness	Security	Localization
Digital Watermarking	verification (reconstruction)	high	medium	yes
Checksums	verification (reconstruction)	no	low	yes
Forward Error Corr.	reconstruction	high	–	no
Crypto Hashes/MACs	verification	no	high	yes
Perceptual Hashes	verification	high	medium	yes
Media Forensics	verification	varying	medium	yes
WORM Media	prevention	no	high	–
Write Blocking	prevention	no	low	–

Table 2.1.: Categorization of different integrity protection mechanisms according to objectives and robustness (Dashes indicate that property does not apply)

Looking at the overview of the techniques described above, a very promising approach is given by combining perceptual hashing and robust watermarking. This addresses the challenges of this thesis work as defined above in the Introduction in Section 1.2 and shows a number of advantages:

- Perceptual hashing allows recognizing acoustic events that sound similar to an average listener. Conversely, it is very likely that perceptual hashing can also detect malicious tampering. Hence the content integrity can successful be verified on a *perceptual* level. Standard cryptography solutions like crypto hashing or MACs would raise false alarms in light of perceptual identity.
- Watermarking allows for making the perceptual hash available as an authentication code embedded *inside* the protected audio. Watermarking as an active mechanism does not rely on the availability of an existing security infrastructure that provides the authentic verification codes. The advantage is that the watermark message can easily be made available again.
- Because of watermarking robustness capabilities, the embedded perceptual hash/authentication code is available even if the protected media data undergoes moderate lossy compression or the like.

	Active	Passive
Pre-Processing	Digital Watermarking Forward Error Correction	Checksums Crypto Hashes/MACs Fuzzy Hashes Perceptual Hashes WORM Media Write Blocking/Locking Quantum Key Distribution
Post-Processing	-	Media Forensics

Table 2.2.: Categorization of different integrity protection mechanisms according to processing modes

It must be admitted that watermarking does apply modifications to the audio. Skillful utilization of psychophysical models on human perception provide that the data modification in the course of watermarking will not become audible and will not interfere with the desired degree of integrity.

The details of combining watermarking with perceptual hashing will be explained in Section 2.3.3 under the term *"content-fragile watermarking"*.

2.2 Perceptual Audio Hashing / Audio Fingerprinting

This Section recalls the state of the art in perceptual audio hashing. This allows comparing the proposed hashing approach with the state of the art in audio hashing.

Note: The following explanation assumes the reader to be familiar with the technical basics of digital audio data, spectral transforms, windowing etc. as recalled in Appendix A.1 for the interested reader.

2.2.1 Perceptual Hashing Model and Terminology

Perceptual Audio Hashing is an identification technique specially designed for digital audio data. Technically, perceptual audio hashing is defined as a mechanism for mapping an audio input signal of arbitrary length to a compact, content-based descriptor or message digest of very short length as output. Formally spoken, an audio data set c_0 is mapped to a binary sequence $H = H(c_0)$. Unlike in *cryptographic* hashing, the perceptual hash value is robust to minor transformations of the input signal. An audio data set will be regarded as "similar" in terms of perceptual hashing even if the data undergoes e.g. moderate lossy *MP3/AAC* compression or DA/AD conversion in the course of microphone recording.

Confusingly, this technology is synonymously known under the keywords

- *"audio fingerprinting"*,
- *"robust (audio) hashing"* or
- *"content-based signatures"*

in numerous publications in the state of the art.

For achieving its robustness, perceptual hashing algorithms extract acoustic features from the data that are relevant with respect to the human perception and can be exploited for identification. Hence, *Doets* and *Lagendijk* in their publication on *"Theoretical Modeling of a Robust Audio Fingerprinting System"* [DL2004] comprehensibly define the extracted audio *"fingerprint"* as

> *"...a compact representation of the perceptually relevant parts of audio content, which can be used to identify an audio file, even if it is severely degraded due to compression or other types of signal processing operations".*

This expresses the technical aspects of robustness in terms of insensitivity to lossy compression etc. which is desired in many audio applications.

For example, the recording of a particular symphonic concert on Audio CD and its *MP3* copy shall have the same, or at least similar, fingerprint. Conversely, different performances of the same concert by different orchestras/conductors shall have different fingerprints. Also *cover songs* or *remixes* of the same original pop song shall have audio fingerprints distinct from one another. Research for identifying such new versions of a song is being conducted (for example in [CS2007a, CDXZ2015]) but is out of this thesis scope.

A different definition by *Haitsma, Oostveen*, and *Kalker* [HOK2001b] emphasizes more the *subjective* perspective of the human listener by describing it as

> *"...a function that associates to every basic time-unit of audio content a short semi-unique bit-sequence that is continuous with respect to content similarity"*

This also expresses that it is common to extract hashes not only from *complete* audio files but also from sections thereof. The *"semi-uniqueness"* is defined by the authors as the property

that for two signals that represent different content, their perceptual hashes shall allow to distinguish these signals. To elaborate upon this point, the term *"continuous"* suggests that human perception and assessment of "acoustic similarity" can be modeled and measured by mathematical functions with a continuous co-domain.

According to the integrity modeling discussed in Section 1.2.2, perceptual hashing provides and assessment of identity on a level of *perception* or sensation, instead of semantics or even true understanding.

Perceptual Hashing Model

As explained by *Lancini* [LMP2004], the typical perceptual hash extraction is carried out in the following processing steps (see Figure 2.1):

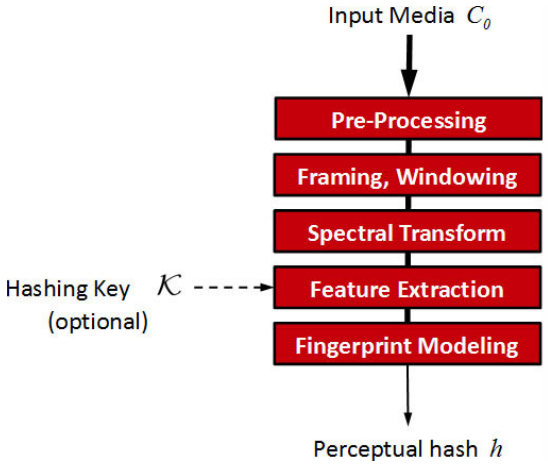


Figure 2.1.: Processing steps for extraction of fingerprint/perceptual hash

1. **Pre-processing**: At first, the input data is prepared to improve the efficiency or the robustness. For the example of this thesis work the PCM input data can be subject to re-sampling to a standard sample rate and to down-mixing to mono.
2. **Framing**: Typical frame lengths in this thesis work are 1024 or 2048 which represents 23 or 46 milliseconds for which the stationarity is sufficiently provided for most music and voice content.
3. **Windowing**: If a spectral transform is carried out (in the following step) the framed audio data is often multiplied sample-by-sample with an appropriate windowing function, for example *Hamming* or *von Hann* window. It avoids undesired effects due to *spectral leakage* (as explained in A.1.3.3) in the spectral domain.
4. **Spectral Transform**: In most approaches, an appropriate spectral transform is applied on the windowed frames. Often, the spectrum is more revealing with regards to sound recognition than the temporal representation is. Typical spectral representations are using DFT coefficients, the (modified) discrete cosine coefficients (DCT/MDCT), or the so-called *Mel-frequency cepstral coefficients* (MFCC). This thesis work will utilize performance optimized implementations of the DFT.
5. **Feature Extraction**: From the representation in the transformed spectral basis, suitable features are selected for extraction. The *skillful* selection of features in this processing step is *vital* for the robustness and distinction performance of the overall algorithm. In

some algorithms the feature extraction stage is dependent on a secret key $\mathscr{K}$ as additional input.

6. **Output Modeling**: The selected features are analyzed according to a predefined algorithm. Usually, the extracted robust hash is a *binary* sequence. However, in some approaches it is an *n-tuple* with real-valued, i.e. continuous elements, instead.

Ambiguity of the Term *"Fingerprinting"* in the Literature

The reader is reminded that the term "fingerprinting" is a commonly used synonym for "perceptual audio hashing". Unfortunately, the term is ambiguously used. It appears in other fields in multimedia security research that are *not* investigated in the scope of this PhD thesis, namely

- as a synonym for "transaction watermarking" that can identify different copies from the same original work
- in the denotation of "collusion-secure fingerprinting" which is a coding technique used in transactional watermarking (as discussed in Section 2.3.4.3),
- in the denotation of "camera sensor fingerprints" for image authentication in forensics and
- in biometric fingerprints (i.e. dactylograms from *minutiae* features on human fingers).

For the avoidance of confusion[5], for the remainder of this thesis the term *fingerprinting* shall denote robust identification methods for audio data as defined in this Section. The terms *perceptual hash*, *robust/audio hash* are also common and will be synonymously used throughout this thesis.

Perceptual Hash Characteristics

The robust hash must withstand attacks that are intended to fool the hashing algorithm for example by tampering or manipulating the content without being indicated by the hash. For this, *Cano, Gomes et al.* [CGB+2002] or *Mıçak* [MV2001] define a number of characteristics that the perceptual audio hash should provide:

Randomness: As every hash calculation, the perceptual hashing implies a significant "compression" of the data because a hash of short length is assigned to an input data set of arbitrary length. Here, the probability for a *hash collision* shall be as small as possible. For this, the robust hash values should be equally distributed among all possible pieces of audio data that sound different. More precisely: if one interprets the M-bit hash as a M-digit dual number, the values of this number shall be equally distributed i.e. shall occur with a probability of $p = 2^{-M}$. A necessary condition for this is that each (binary) digit is equally distributed on $\{0, 1\}$.

Discrimination performance: For the detection of perceptually relevant transformations of the signal, it is required that the hash values H should be different for two perceptually distinct audio signals:

$$c_0 \neq c_0' \quad \Longleftrightarrow \quad p(H(c_0) = H(c_0')) \leq 1/2^M \quad . \tag{2.1}$$

[5] For example, this ambiguity caused an unfortunate *rejection* of a research paper submitted by the thesis author. The paper actually addressed a research topic (beyond this thesis work) in the field of "collusion-secure fingerprinting". But it was regarded by one of the reviewers as if "perceptual hashing" was addressed. Under this unfortunate misconception the rejected paper appeared as *"difficult to read"* to the reviewer because of *"a lot of confusion"*.

The upper bound reflects that the hash values can be identical *by coincidence*. This requirement shall allow the detection of clearly noticeable modifications or even malicious tampering of the audio data.

Even more, it is required that the hash values of two hash protected audio signals c_0 and c_0' that are known to be perceptually distinct, should be mutually independent:

$$p(H(c_0) = h_1 | H(c_0') = h_2) \approx p(H(c_0) = h_1) \quad \forall\ h_{1,2} \in \{0,1\}^M \quad . \tag{2.2}$$

This means that an attacker must not be able to predict the hash value for a given piece of audio (or parts of it) from previously calculated hash values. This should make systematic attacks as difficult as possible. Like in crypto hashing these two requirements should discourage from an collision attack by keeping the effort for an adversary as "hard" as possible.

Robustness: The robust hash must be invariant to transformations for which the audio data remains perceptually similar. The following condition must be fulfilled:

$$c_0 \approx c_0' \quad \Longleftrightarrow \quad p(H(c_0) = H(c_0')) \approx 1 \quad . \tag{2.3}$$

For the integration into authentication watermarking as investigated in this PhD thesis, the hash must be robust against the distortion introduced by watermarking, too.

It is obvious that the discrimination and robustness properties are somewhat mutually exclusive: In practice, perceptual hashes provide a "gray area" of distinction performance when increasing levels of distortion and/or tampering are applied to the protected input data.

Compactness: The extracted features must provide a very compact representation of the content. In the context of this PhD thesis, the robust hash will be embedded as a robust watermark message. In the construction of the hash, it should be reflected that robust audio watermarking schemes usually provide a very limited embedding capacity. Hence, the hash should be as compact as possible (see Section 2.3 for an elaboration on watermarking and its "capacity" properties).

Granularity: The design of the system should allow to extract a meaningful perceptual hash from *sections* of the input data. That permits an identification even if a part of the analyzed content is missing. In the context of this thesis, this shall allow the *temporal* localization of malicious tampering inside a file, for example by its time-code.

If utilized for authentication purposes, the perceptual hash must meet a number of security requirements. Creating hash collisions deliberately must be computationally difficult. The randomness and discrimination property explained before are important conditions for that.

In addition, *Fridrich et al.* pointed out that for achieving this, the "Feature Extraction" step must be key-dependent [FG2000]. Other approaches for achieving a more secure audio hash can be realized by a key-dependent quantization of the features (see *Nouri* [NZAF2012]) or the projection of the feature vectors on a key-dependent basis (see *Radhakrishnan* [RM2002, RXM2003]) .

In addition, *efficiency/computational demands* and *scalability* are usually named as important characteristics, too. In the context of this PhD thesis, this is of minor relevance and not analyzed much further. See Section 3.4.3 for some information on runtime performance of this thesis work.

2.2.2 Applications of Perceptual Hashing

Beyond this thesis work perceptual hashing for audio by itself is the technical basis for a number of other application fields too. The most important are outlined in the following [HOK2001a, AHH+2001b, dGCG+2003]:

Music Recognition and Sales: Perceptual audio hashing is used in commercial systems for recognizing music songs and providing meta data about it. This might be helpful for a consumer to recognize and retrieve the title of an unknown song that he/she listens to *on the air* or in a public environment. Current commercial recommendation services for this are provided by *Shazam Entertainment Ltd.*[6], *Audible Magic Corp.*[7].

Content Filtering: Perceptual hashing is also used for identifying and banning copyright infringements in company networks or on platforms for user-created content. For example, *Youtube*'s *Content ID* system[8] (based on technology licensed from *Audible Magic Corp.*) prevents uploading banned content to the *Youtube* platform again (even in modified versions).

Automatic Music Library Organization: Many users have a somewhat *unsorted* collection of music files on their computers or mobile devices. Perceptual hashing allows tagging all songs, i.e. by artist name, song title etc. and de-duplication in a music collection. Examples for such services or products are the *mufin player*[9], the sound recording tools by *Audials AG*[10], the Audio CD database by *Gracenote Inc.*[11], or *MusicBrainz*[12].

Broadcast Monitoring / Audience Measurement: Perceptual hashing enables music artists or music labels to monitor the popularity of a work across mass media broadcasting. Copyright collecting societies (like e.g. the German *GEMA*) can verify if royalties are paid accordingly. Finally, marketing agencies can monitor if commercial advertisement is distributed according to airtime agreements with broadcasters. One commercial provider is *Musictrace GmbH*, Erlangen, Germany[13].

User Interaction, "Second Screen" services: Perceptual hashing can enrich services for user entertainment. It can trigger that value added content (like background information, advertisement) is displayed on a mobile phone or tablet computer in parallel to consuming a TV programme or movie. Here, the mobile device virtually serves as a "second screen".

Watermarking support: Perceptual hashing can be a *supporting mechanism* to watermarking in various ways: For example, perceptual hashing can help deciding if given audio content is known as having been watermarked earlier, in the first place. This can help saving computational efforts by avoiding (time-consuming) watermark detection when the content is unmarked anyway.

Different authors propose using perceptual hashes to increase the efficiency of temporal synchronization during watermark detection (e.g. by *Hauer* for video watermarking [HS2006]).

---

6 *Shazam Entertainment Ltd.*, London, UK: *http://www.shazam.com*

7 *Audible Magic Corp.*, Los Gatos, USA: *http://www.audiblemagic.com*

8 *Content ID* by *Youtube/Google Inc.: https://support.google.com/youtube/answer/2797370*

9 *mufin* software by *mufin GmbH*, download source: *http://www.heise.de/download/mufin-player-1145769.html*

10 *Audials One/Tunebite* software by *Audials AG, http://audials.com*

11 *Gracenote Inc.*: *http://www.gracenote.com* (formerly known as *CDDB*)

12 *Musicbrainz by MetaBrainz Foundation, San Luis Obispo, USA*: *https://musicbrainz.org*. The service uses the *"AcoustID"* fingerprinting technology by Lukas Lalinsky: *https://oxygene.sk/category/acoustid*

13 *Musictrace GmbH: http://www.musictrace.de*

As another example, *Mıçak* uses an approach in which a perceptual hash serves as the secret watermark key for embedding and detection [MV2001]. This can prevent copy attacks or information leakage about the key by analyzing watermarks with same message across many different watermarked media. The interested reader is referred to Sections 2.3.1 and 2.3.4.3 about the definition of the terms ”watermark keys” and ”copy attacks”, resp.

> Own publications on this topic were published by the thesis author [ZSN2006, SZB2010, ZSB2012]. These works are beyond the scope of this thesis and are not elaborated upon.

Some fingerprinting algorithms are technically suitable for several purposes at the same time. One example is *YouTube's Content ID* algorithm: it allows strict content filtering (denoted as *”blocking”*) but also broadcast monitoring for *”monetizing”* using the same audio hashing system[14].

Finally, it has to be emphasized that *integrity verification* as in an information security context usually is not among the applications of perceptual hashing.

2.2.3 Example: Fourier-based Audio Hash by *Haitsma, Kalker et al.* (”*Phillips* Hash”)

The following explanation again assumes the reader to be familiar with the basics of the Fourier transform as recalled in Appendix A.1.3.

One of the most frequently cited and discussed research contribution in audio hashing was presented by *Haitsma, Kalker*, and *Oostveen* [HOK2001a, HOK2001b], conducted at *Phillips Research*, Eindhoven, The Netherlands. This ”*Phillips* Hash” is explained first because many other publications in the state of the art – and this thesis work – were influenced by this work. The design impresses by its simplicity and its desired robustness properties.

The robust feature extraction uses a spectro-temporal analysis of the absolute value of Fourier coefficients of the audio signal. For extracting a so-called *sub-fingerprint* from each audio frame, the following processing is carried out:

- In the original algorithm, the audio signal is digitally represented by PCM samples and analyzed in rather large frames of $L = 2^{14}$ =16384 samples that overlap by 31/32. Large frames provide a fine frequency resolution but poor temporal resolution. The latter is compensated for by the large frame overlap.
- The frame data is weighted with a Hamming window [S$^+$1997] first for avoiding spectral leakage effects. Then the Fourier transformation is computed.
- Subsequently, the Fourier magnitude coefficients $\{r_k\}$ are resampled according to a logarithmically scaled frequency axis: the linear frequency axis with indices k is mapped onto a decimated scale according to 33 musical semi-tones with indices i correspondent to 300-2000 Hz. The decimated spectrum is expressed in terms of the quantity $\{e_{i,t}\}$ which is the partial signal energy in the i-th semi-tone at the t-th time-step. This takes into account that the human perception of *pitch* can be modeled mostly on a logarithmic frequency scale like it is commonly known for musical tones.

It was found experimentally, that the mutual spectro-temporal differences between these energy coefficients are a quite robust feature for arbitrary music and voice audio material. A robust feature is derived by comparing the differences of neighboring energy coefficients (at frequency

[14] *Youtube Content ID*, user manual at *https://support.google.com/youtube/answer/2797370*

indices i and $i+1$) at a given time index t to those in the following time-step $t+1$ according to the following definition:

$$d_{i,t} := e_{i,t} - e_{i,t+1} - \left[e_{i+1,t} - e_{i+1,t+1} \right] \quad , \quad i = 1...32 \quad . \tag{2.4}$$

Audio hash bits $H_{i,t}$ are defined by the authors as the *sign* of the quantity $d_{i,t}$:

$$H_{i,t} = \begin{cases} 1 & \text{if } d_{i,t} \geq 0 \\ 0 & \text{if } d_{i,t} < 0 \end{cases} \quad .$$

According to Equation (2.4) from the original algorithm, FFT coefficients are "picked" in *regularly* increasing order across time and frequency indices: consecutive hash bit indices correspond to consecutive/adjacent frequency indices.

A visualization of the feature selection is given in Figure 2.2. Note that the plot only gives a *symbolic* visualization of a small range from a domain of 12 by 10 time-frequency indices. In the actual implementation of the algorithm the following settings are used:

- A 32-bit so-called *sub-fingerprint* is extracted at time index t' by evaluating the 33 energy coefficients per frame according to Equation (2.4).
- This is carried out repeatedly 256 times, every 11.6 milliseconds with an overlap of 31/32. A complete *fingerprint block* $\{H_{i',t'}\}$ is defined as the concatenation of those 256 consecutive sub-fingerprints, i.e. $t'=1...32$, $i'=1...256$.
- Hence the payload of the fingerprint block is $256 \cdot 32\ bit = 8\ kbit$. It covers approximately $256 \cdot 11.6$ milliseconds ≈ 3 seconds of audio. Hence the fingerprint data rate is approximately $8\ kbit/3\,s \approx 2.6\ kbit/s$. This rather high data rate is mainly caused by the 31/32 overlap which compensates for the limited time-resolution of the single Fourier transforms.

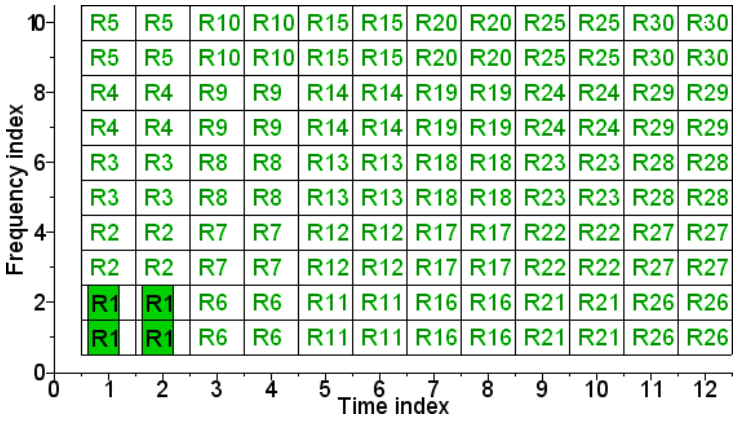


Figure 2.2.: Feature selection scheme in *Phillips* audio hash – Symbolic representation of DFT indices in the time-frequency domain. Every set of four correspondent coefficients denoted as "Rm" contributes to the m-th hash bit. Coefficients R1 which are correspondent to the first hash bit are highlighted for demonstration purpose

Note that a correspondent approach for *video* data was introduced by the same authors [OKH2001] based on average block luminance values instead of Fourier coefficients.

More experimental results, thorough theoretical analysis and extensions on this audio hashing algorithm can be found in the original works or in a large number of sec-

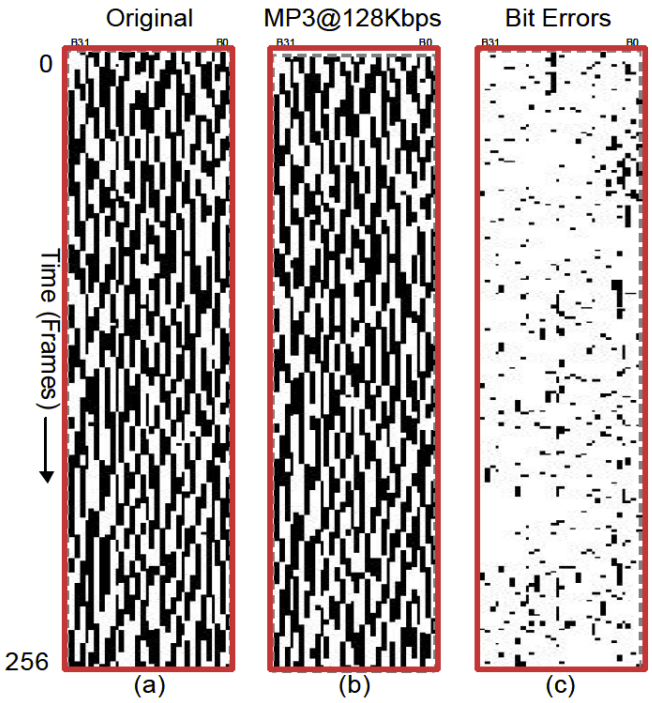


Figure 2.3.: "*Phillips* Hash" feature extraction example; black dots = '1', white dots = '0', each row contains one sub-fingerprint;
Plot (a): Audio hash of input audio section, Plot (b): Audio hash from *MP3* compressed copy of input audio section, Plot (c): XOR difference between Plots (a) and (b); bit error rate = 0.078;
Source: [HOK2001a]

ondary sources [DL2004, HBMS2007, BHMS2007a, BHMS2007b, JYLN2007, JLLN2008, LYK2009, GSM+2012, CH2014, Seo2014, YWN2015].

Robustness Aspects

Although simple by its definition in Equation (2.4), the quantity d provides a high level of robustness: Obviously, energy *differences* between coefficients are robust to changes of the *total* volume. The feature is also robust against moderate levels of lossy compression, re-sampling, filtering, dynamics compression, noise addition, and even DA/AD conversion using a tape recorder. Unlike for crypto hashes, the bit error rates caused by such moderate transforms are 5 to 10% as reported by the authors.

Example: An example from [HOK2001a] is shown in Figure 2.3: Plot (a) visualizes a fingerprint block, i.e. the "height" represents a duration of approximately three seconds. Note that the temporal transitions of the sub-fingerprints over time in Plot (a) and Plot (b) are very smooth along the time-axis (from top to bottom). This is caused by the high correlation between neighboring rows because of the 31/32 overlap. The fingerprint difference between Plots (a) and (b) is shown in Plot (c): it demonstrates that the perceptual hash is rather insensitive to moderate *MP3* compression of the audio: the bit error rate between the fingerprints of the original audio and its *MP3* copy in 128 *kbit/s* is 7.8% only, as claimed by the authors.

Security Aspects

Own research (as published in co-authorship [TNSZ2009]) showed that the overall algorithm can be fooled by targeted attacks.

In [TNSZ2009] it was shown that inaudible modifications can be applied that can nevertheless change some of the fingerprint bits. This is favored by two circumstances:

- The value of the extracted feature $d_{i,t}$ shows a continuous distribution. In some cases its absolute value $\left|d_{i,t}\right|$ is near zero. In this case, even slight modifications or deliberate attacks to the energy coefficients that contribute to $d_{i,t}$ can flip its sign and hence flip the value of the fingerprint bit. *Coover et al.* denote such bits as *"weak bits"* [CH2014] which is a concept that does not exist in crypto hashing. This behavior can be expected from the definition of the quantity $d_{i,t}$: As the four energy coefficients in Equation (2.4) will have similar expectation values, the expectation value of $d_{i,t}$ will actually be near zero.
- The selection of energy coefficients that contribute to $d_{i,t}$ is not key-dependent. That allows studying systematically which of the hash bits are "weak" as defined above with minimum efforts. This is why introducing a pseudo-random feature selection is a condition for a secure perceptual hash construction model as was explained in Section 2.2.1.

This allows protocol attacks by creating *false positives* which can make the perceptual hash appear useless. It could also be used for circumventing audio hash based content filtering mechanisms. Hence an additional randomization step is investigated in the literature on audio hashing and in this thesis work.

2.2.4 State of the Art in Perceptual Audio Hashing Algorithms

This Section gives an overview on publications on perceptual audio hashing.

2.2.4.1 History of Audio Data Hiding Research

The topic of perceptual audio hashing became increasingly investigated from the year 2000 which is many years after digital production, archiving, and distribution of music had become common on a large scale. Until nowadays, the research progress in this field has been steadily increasing, as can be seen from Figure 2.4.

As can be expected from the overview on applications in Section 2.2.2 above, most works in the field consider retrieval mechanisms e.g. for recognition of music content, discriminating speech from music content, advanced music genre categorization or even recommendation services. Apart from content filtering applications, audio hashing has not been thoroughly investigated in the information security research community.

2.2.4.2 Audio Hashing Algorithms

The approaches in the literature are roughly categorized according to the following criteria:

Feature domain: Perceptual audio hashing analyzes audio features in different data domains. All approaches that are regarded as relevant for this thesis work analyze audio features mainly in a spectral domain, e.g. Fourier/DFT, MFCC, MFCC or Wavelets.

[15] *IEEE Xplore Digital Library: http://ieeexplore.ieee.org*

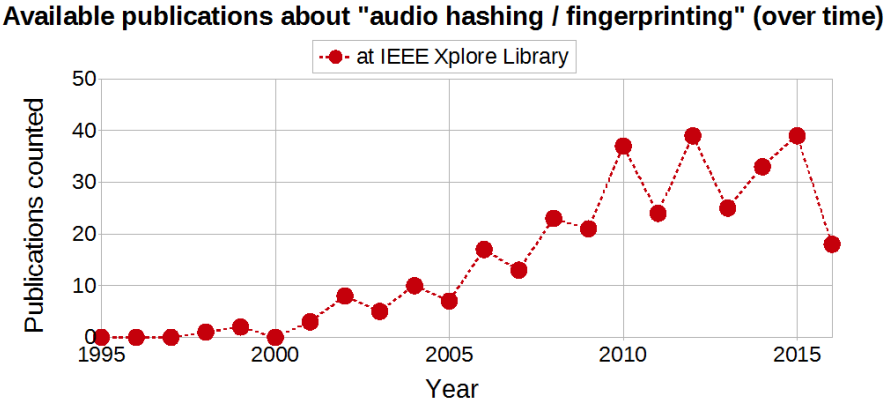


Figure 2.4.: Number of available publications about "audio hashing/fingerprinting" (over time) that are available at/via *IEEE Xplore Digital Library*[15]

Internal randomization: A few approaches discuss or even propose internal randomization during feature extraction for providing a more secure hash calculation. That is, a dependency from as secret key can be implemented in some of the algorithms.

It can already be mentioned here that the robustness is not a significant criterion to distinguish algorithms as *all* selected audio hashes show a sufficient robustness to common distortions on the hashed audio (as claimed by the authors). Notable publications apart from the *Phillips* Hash approach as explained above are outlined in the following:

The "*Shazam* Hash"

An approach by *Avery Li-Chun Wang* [Wan2003, Wan2006] from *Shazam Entertainment Ltd.*, London, UK, derives low-level features in the Fourier domain: It heuristically identifies dominant peaks in the Fourier spectrum which are denoted by the author as *"anchor"* points. Then, the displacement vectors between such anchors, i.e. their mutual difference of the time and frequency indices are evaluated and utilized as characteristic feature, see Figure 2.5. In terms of the MPEG psychoacoustic modeling as explained in Section 2.4 the anchor points are somewhat similar to *tonal components* (but not identical).

The *Wang* algorithm is integrated in the commercial *Shazam* system for music identification. The interested reader can investigate its outstanding robustness by trying the *Shazam* by himself/herself[16].

The "*AudioID*" algorithm

It was developed by *Allamanche, Herre, Cremer et al.* [AHH+2001a, AHH+2001b] at the *Fraunhofer IIS* institute. It is based on an analysis over time of the volume/loudness in the time-domain and the so-called *spectrum flatness measure (SFM)* and *spectral crest factor (SCF)* in the Fourier domain which describe the tonality of the spectrum as characteristic features. These quantities are rather basic statistical characteristics as simple as, for example, ratios of geometric means versus arithmetic means, or maximums over frequency coefficients which are easy to compute. The exact definitions can be found in the MPEG-7 specification on *"low level audio descriptors" (LLD)* [ISO2002].

[16] *Shazam* smartphone app for *Android* or *iOS*, *http://www.shazam.com*

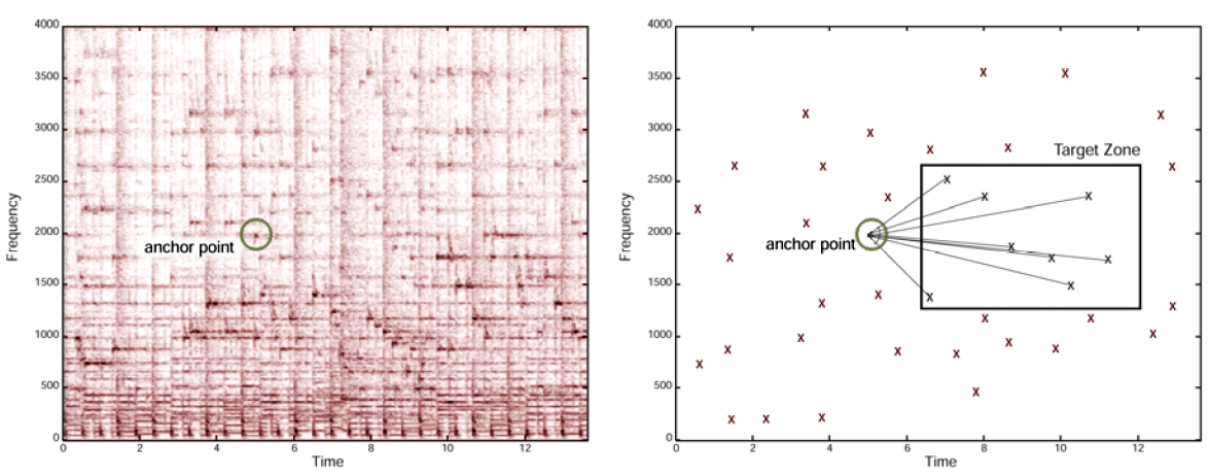


Figure 2.5.: "Shazam Hash" feature extraction example;
Left Figure: Fourier spectrogram as false color plot (higher Fourier magnitudes indicated by darker color);
Right Figure: position of isolated peaks in spectrogram (crosses), displacement vectors between *"anchor"* peak versus other peaks in *"target zone"*
Source: [Wan2003]

The same main features are being utilized in a different way in the work by *Lancini* [LMP2004].

This *AudioID* is included in the consumer software *mufin player*[17] for automated re-organization of music collections and is utilized for commercial broadcast monitoring services by the *Music-trace GmbH*, Erlangen, Germany[18].

Miscellaneous Fourier Domain-Based approaches

Like the examples explained above, the majority of approaches in the literature on audio hashing utilize features derived from the Fourier magnitudes (while Fourier phases are discarded). Significant works are outlined in the following:

- One DFT-based approach was published by *Seo* [SJL+2005]. Interestingly it reflects human perception by temporarily dividing the high-resolution spectrum (2048 FFT coefficients) in 16 critical bands in the range of 300-5300 Hz. Then, the set of indices of the *centroids* in each of the critical bands is used as a robust feature. *Jang et al.* proposes [JYL+2009] an extension thereof by skillful filtering and quantizing the centroid identification result followed by mutual comparison of those quantized values (*"pairwise boosting"*).
- *Baluja* and *Covell et al.* from *Google Inc.*, Mountain View, USA, proposed their *Waveprint* algorithm [CB2007, BC2007]. It extracts characteristic features by applying a Wavelet transform on the DFT spectrogram as post processing operation for improving the robustness and discrimination performance.
- In an approach by *Yu Liu* [LYK2009] the Fourier spectrum is evaluated similar to the Phillips hash approach described above. Then the energy spectrum is subject to post-processing by the DCT which is a standard approach for de-correlating the spectra. This is a reasonable approach for abstraction: because of the large overlap of the FFT frames

[17] *mufin* player published by *MAGIX Software GmbH*, *http://www.mufin.com*
[18] *loc. cit*

the spectra are slowly varying over time which introduces many correlations between the spectra.

- *Malekesmaeili* proposed an interesting approach based on the so-called *chroma* feature over time [MW2012, MW2014]. This is a post-processing operation that evaluates the spectrum according to *octave series*: for a given frequency f_0 the octave series is defined as all octaves at frequencies $(2^n \cdot f_0)$ with $n \in \mathbb{Z}$. The chroma feature accumulates frequency components irrespective in which octave they are present. This is well suited for (but also limited to) analyzing *music* content. It is also a starting point for developing an audio feature that is invariant to pitch shifting. As claimed by the authors it outperforms the *Shazam* hash.
- A work by *Shibuya et al.* describes an approach that extracts so-called pseudo-sinusoidal components [SAN2013]. These are defined as peaks in the Fourier spectrum that are constantly dominant over a short period of time. Hence it shares some of its core ideas with the *Shazam* hash approach.
- *Ouali et al.* presented a simple approach that compares the Fourier coefficients in the spectrogram against their arithmetic mean across one second of duration [ODG2014].
- Finally, *Coover* and *Han* from *Gracenote Inc.*, Emeryville, USA, propose an approach that develops the *Phillips* Hash further [CH2014]. As explained in Section 2.2.3 closer analysis shows that not all bits in the original audio hash of a certain audio segment are equally robust to distortions. The authors distinguish such *"strong bits"* from *"weak bits"* for defining a so-called *power mask* for increasing the robustness. This idea is adapted by *Seo et al.* [Seo2014] for improving the performance of the hash matching process.

Various MFCC-based Approaches

A number of algorithms analyze the so-called *Mel-frequency cepstrum coefficients (MFCC)* (as explained in Section A.1.4.1) along with the plain Fourier spectrum. The MFCC spectrum is commonly utilized in (but not limited to) tasks of speech recognition and speaker recognition. Significant MFCC-based works in audio hashing are listed in the following:

- *Grutzek* and *Knospe et al.* presented works on applying the feature extraction principle of the Phillips Hash on MFCCs [GSM$^+$2012, Kno2013]. Subsequently, in an internal processing step an intermediate result of feature extraction is subject to a cryptographic message authentication coding, namely the CMAC [Dwo2005]. This is claimed to preserve the privacy as the CMAC is a secure one-way function which prevents reconstructing the audio sound from the final hash result. The authors themselves state that the privacy protecting CMAC post processing is not sufficient to provide security in authentication applications as the MFCC extraction itself is not secure.
- *Cheng et al.* [CCW2003] proposes a general audio hashing algorithm using Hidden Markov Modeling (HMM) of feature vectors consisting of low-level audio features like the loudness/volume and the zero-crossing-rate in the time-domain, the bandwidth in the Fourier domain and the MFCC spectrum.
- Another MFCC and HMM based approach by *Gomez, Cano et al.* [GCdG$^+$2002] solely for integration in authentication watermarking is outlined in the Section 2.5 on Related Work.
- *Özer* and *Memon* propose a post-processing on the MFCCs by utilizing the singular value decomposition (SVD) as "mathematical tool" *et al.* [OSMA2005]. See Appendix A.1.4.2 for a brief explanation of the SVD. In simple words, the SVD provides a very abstracted

”summary” of the principle axes of the MFCC spectra that allows extracting robust features for audio identification.

MCLT-based Approach

Mıçak et al. presented a perceptual hashing approach in [MV2001] based on signal statistics in the spectral domain. Here, the *complex modulated lapped transform (MCLT)* is used. This transform from the family of *discrete cosine transforms (DCT)* are common in lossy compression, for example in *Dolby AC-3* for audio or *JPEG* for images.

The publication is relevant and inspiring in the context of this thesis for the following reasons:

- It gives a formal and comprehensible definition of requirements on perceptual hashes in general, as cited in Section 2.2.1.
- It proposes to use *adaptive quantization* in the feature extraction step of the hash calculation. The goal is improving the robustness of the scheme.
- It reflects the human auditory perception in a simple way to some extent by ignoring frequency components that are below the absolute hearing threshold in silence (as explained in Section A.2.4. For this, a binary feature is extracted by comparing the MCLT coefficient with the hearing threshold: the so-called *significance map* indicates if a spectral coefficient is greater or smaller than the correspondent hearing threshold.

Experimental results show a good robustness against many attacks that are inaudible or just little audible. However, targeted attacks on the collision security were not evaluated.

Interestingly, the author discusses to utilize the MCLT-based audio hash output as the secret key input for consecutive watermarking. Using such content-dependent key increases the security of any watermarking algorithms with regards to copy-attacks. The interested reader is referred to Section 2.3.1 and Section 2.3.4.3 about the definition of the terms ”watermark keys” and ”copy attacks”.

Key-dependent Bitstream Hashing for *MP3/AAC* data

An interesting aspect is addressed by the works of *Jiao et al.* in [JYLN2007, JLLN2008]. It offers the opportunity of hashing for compressed *MP3* and *AAC* input with lower computational efforts: these bitstreams already provide the audio spectrum in the so-called in MDCT representation. This reduces computational demands because only *marginal* parsing efforts for obtaining the spectra are required.

In the context of this thesis, this means a certain limitation because it is also desirable to protect plain PCM audio data and other audio formats too. Nevertheless this work is somewhat inspiring because the authors propose a *key-dependent* pseudo-random selection of MDCT coefficients in the feature extraction and hash modeling steps. Hash values of the same audio but with different keys are shown to be independent in these works.

Audio Hashing based on *Zernike* Transform

An audio hashing approach by *Ning Chen* and *Hai-dong Xiao* [CX2013] makes a virtual ”detour” by applying techniques from digital image processing and optics. The (presumably time domain) audio signal is divided into frames and the data is temporarily projected / reordered into a two-dimensional representation. Then the dataset is developed in terms of 2D-*Zernike* polynomials which were originally introduced [Zer1934] for calculations in (analog) phase contrast microscopy.

The motivation for this particular method can be understood from the property of *Zernike* moments providing an image data representation that is invariant to translation, rotation and scaling [Tea1980, KH1990]. Authors show that this is, in turn, a suitable robustness property for audio hashing too. Eventually the robustness was successfully evaluated by the authors for the presence of content-preserving audio transforms. Nevertheless, detection of malicious tampering was not the objective of the work by *Chen*.

Key-dependent Wavelet based Approach (by *Nouri, Abdolmaleki*)

For the sake of completeness, the approach by *Nouri, Abdolmaleki et al.* shall be mentioned [NZAF2012]. Although somewhat imprecise in its technical detail, the work is significant as it discusses a *key-dependency* in the Wavelet feature selection.

2.2.5 Summary, Discussion and Current Research Trends

Perceptual hashing is nowadays a well investigated and developed technology. Research progress is being constantly published in major multimedia processing conferences and journals from time to time. Technically, most algorithms analyze the audio data in a spectral domain, especially the Fourier domain. Doing so, most approaches achieve a sufficient robustness suitable for the main applications of audio hashing like music recognition or second screen services. Low bit error rates are observed for the most important admissible transformations of the audio content like lossy compression or DA/AD conversion in the course of loudspeaker playback/microphone recording. Most of the algorithms are only little robust to time stretching or pitch shifting which is, however, of minor relevance for tampering detection.

It is notable that a significant portion of works in the state of the art consider the "*Phillips* hash" modeling: many authors (still) discuss certain aspects of this algorithm in more detail [DL2004, HBMS2007, BHMS2007a, BHMS2007b] or apply its overall idea to their own works [JYLN2007, JLLN2008, LYK2009, GSM$^+$2012, CH2014, Seo2014, YWN2015].

About the selection of features it is also notable that for the sake of robustness many algorithms focus on identifying salient features, mostly dominant peaks in whatever spectrum. This would be insufficient with regards to the thesis objectives because also the audio spectrum *off the peaks* can contain audible and significant content.

Example: A synthetic but nonetheless instructive audio data example can be seen in the "Numerical Example" in Section 2.4.1: a white noise signal is mixed with pure sine tone signals. This sound s (loosely spoken) as "beeps with background noise" (see Figure 2.13). But only the sine tones would contribute to the audio hashes proposed by *Avery Li-Chun Wang* or *Shibuya* [Wan2003, Wan2006, SAN2013] which could be insufficient with regards to tampering detection.

On the other side, utilization for authentication purposes is rarely discussed in perceptual audio hashing. Thus, security aspects with regards to ressillience against circumventing the tampering detection have been out of scope in the state of the art. However, a few publications *do* introduce a randomization process in the feature extraction and hash modeling as required for secure authentication. According to *Jiao* or *Nouri* this can be realized by a key-dependent (not: content-dependent) quantization of the features or the projection of the feature vectors on a key-dependent set of basis vectors [JLLN2008, NZAF2012]. The latter is also stressed in a work by *Radhakrishnan* [RM2002] or in an early work by *Fridrich* on secure image hashing [FG2000].

Publication	Time	DFT	MDCT	MFCC	Wavelets	other	key-dependent
[HOK2001a]		x					
[AHH+2001a]	x	x					
[MV2001]			x				
[CCW2003]	x	x	x				
[SJL+2005]		x					
[OSMA2005]				x			
[Wan2006]		x					
[BC2007]		x					
[JLLN2008]			x				x
[JYL+2009]		x					
[LYK2009]		x					
[GSM+2012]		x	x	x			
[NZAF2012]					x		x
[CX2013]						x	
[MW2014]		x					
[SAN2013]		x					
[ODG2014]		x					
[YWN2015]		x					

Table 2.3.: Categorization of perceptual hashing in the literature (in chronological order); Left columns: First author and reference, middle columns: data domain for feature extraction, right column: key-dependent (*"x" = yes*)

The main properties of the approaches described above with regards to the data domain of feature extraction and the discussion of security aspects are summarized in Table 2.3. Note that the overview so far covers audio hashing given by itself, mostly used for music identification. A few more hashing approaches that are proposed exclusively in the context of authentication watermarking will be explained in the Related Work in Section 2.5.

2.3 Digital Audio Watermarking

This Section recalls more than sixty years of audio watermarking for analog and digital audio signals. This allows comparing the watermarking-related aspects of the investigated approach with state of the art in audio watermarking.

2.3.1 Watermarking Model and Terminology

Digital Watermarking is a technique for hiding information into digital multimedia data. The hiding is done without changing the file format and the perceived quality compared to the original input: the watermark is technically transparent and invisible or inaudible, respectively.

The wording in data *"hiding"* reflects that the technique is a practice from the field of *cryptology* which has been under research and in use since ancient times. But unlike cryptography and cryptanalysis, watermarking and the related field of steganography of digital data are rather young disciplines. For example, the first *ACM Information Hiding Workshop*, a well-respected academic conference on these research subjects, took place not earlier than 1996 [And1996].

A frequently cited definition by *Dittmann* in [Dit2000a, p. 20] introduces the terms of "embedding" and "detection" and stresses the technical nature of the watermark as

> *"...a transparent, imperceptible pattern that is embedded in the data. The watermarking procedure...includes an embedding process (embedding or marking algorithm) and a retrieval process (retrieval algorithm, retrieval of message). The watermark pattern mostly is a pseudo-noise pattern that encodes the watermark message"*[19].

In extension, *Cox, Miller, and Bloom* [CMB+2007, page *XV, Preface*] define watermarking as being

> *"...a practice of hiding a message about an image, audio clip, video clip or other work of media within that work itself".*

This stresses the *purpose* of watermarking: the embedded messages in somehow related to the media content (e.g. representing copyright information, user IDs or authentication codes).

Watermark Embedding

Most embedding approaches are based on adding a pseudo-noise signal, modifying statistical or quantization properties or replacing perceptually irrelevant parts of the cover data by the watermark message. The basic embedding model consists of the following elements (see Figure 2.6):

Cover data: This is the input data C_0 to the embedding process. In focus of this PhD thesis will be *audio* data. The audio content covered can be available on its own or it can be present as the sound track in a "video". Audio cover data can be represented either in the *raw/temporal* domain, for example, PCM audio samples in WAVE files or Audio CDs. Or it can be given in the *spectral* domain: examples are spectral coefficients as present in *MP3* or *AAC* data or (temporarily) as the output of a Fourier transform.

Message: The input message $\mathscr{M}$ mostly is sequence of message symbols of a certain length. Mostly, the alphabet of symbols is binary.

Unlike in the research field of *steganography*, the embedded watermark message is semantically related to the watermarked content, e.g. a copyright notice, an individual ID about

[19] Definition translated literally from German

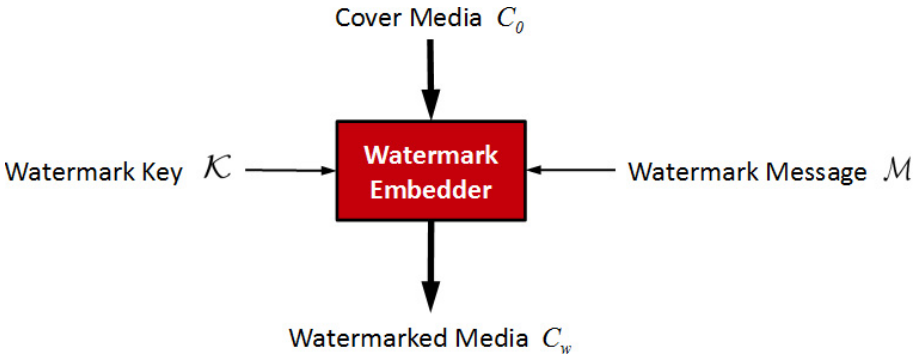


Figure 2.6.: Watermark embedding model

the particular copy of the file, or any kind of metadata about the watermarked content. In the context of this PhD thesis, the embedded meta data will allow verifying the integrity of the cover data.

Secret key: The confidentiality of and the access to the embedded information is provided by a secret watermarking key $\mathscr{K}_1$ as input data. Depending on the watermarking algorithm, the key may be the seed of a random number generator which is used for controlling the embedding positions in the media. Watermark keys in current watermarking approaches are *symmetric*. Hence the key provides a proof of authenticity in that the watermark must have been embedded by an authorized entity from the set of entities that have access to the watermarking key[20]

Watermarked media: The output to the embedding process is the watermarked media C_W. It is usually created in the same data format as the cover data input.

Watermark Detection and Retrieval

Each watermarking embedding algorithm corresponds to a watermark *detection* algorithm. In many approaches, including the one investigated in this thesis, at first the algorithm attempts to detect the pure presence of an embedded message. If this *synchronization* is successful with a sufficient degree of significance, the embedded message sequence is eventually *retrieved* from the marked data (see Figure 2.7).

The message is detected and retrieved by either correlation, evaluating the quantization properties of the data, several statistical methods or comparison of the marked data with the cover data, if available. Common *embedding* and *detection* principles will be explained in detail in the state of the art overview in Section 2.3.4.

One outstanding property of digital watermarking in comparison to other security mechanisms is the inseparably interwoven relationship between embedded information and the protected multimedia data: For most algorithms, the process of embedding is irreversible and the watermark cannot be removed. Even more, while common cryptography mechanisms fail when content is converted to other file formats or taken from the digital to the analog domain, watermarks can persist and keep the content protected.

[20] It should be noted that the term ”key” is frequently used in an ambiguous manner in the watermarking research community. Some authors propose applying the encryption of the watermark message $\mathscr{M}$ using a secret ”key”. But for the remainder, the term ”(watermark) key” shall denote a secret which defines where and how the message is embedded into the cover media and which is shared between embedding and detecting algorithms and parties.

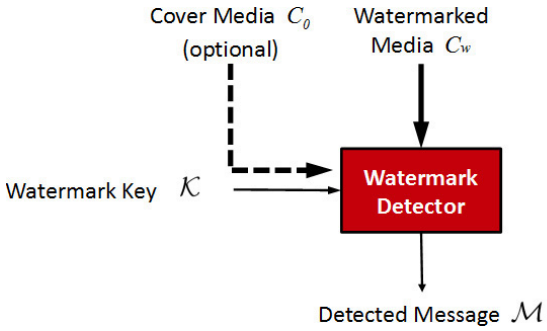


Figure 2.7.: Watermark detection and retrieval model; Dashed line: Utilizing the original cover data is optional (non-bind watermarking)

Watermarking Characteristics

Digital watermarking algorithms can be categorized with respect to different characteristics (see *Cox* [CMB+2007], *Dittmann* [Dit2000a] or *Kalker* [Kal2001]):

Transparency: The watermark should be imperceptible as possible, i.e. the degradation of the acoustic or visual quality caused by the embedding should be minimal. For achieving this, this thesis work uses psychophysical models like the one explained in [ISO1993a] for controlling the embedding modifications. Furthermore, the embedding should be transparent to the digital representation of the signal: the embedding is usually done without changing the file format.

Robustness vs. Fragility: Most approaches are designed as *robust* watermarking algorithms: here, the embedded message can still be detected and retrieved even if the watermarked data is subject to further post-processing or signal transformation or DA/AD conversion.

Robustness usually is determined by using different levels of *redundancy* during embedding. That can be done by embedding each message bit repeatedly, i.e. using larger sections of the cover to embed each message bit. Also on the coding level, redundancy can be exploited by including forward error correction coding or collusion-secure fingerprint coding.

In contrast, the more rare *fragile watermarking* approaches for integrity verification are designed so that the embedded message is destroyed by moderate or even *any* signal transformation. The different classes of fragile watermarks can be used for integrity verification and are hence explained in detail in Section 2.3.3.

Security: The watermark should withstand attacks aimed directly at the embedded information. Unauthorized retrieval or modifying of the watermark message without knowledge of the secret watermark key shall be prevented. Successful deletion of the message shall require transforms or distortions of the marked media that are so strong that they render the media unusable and ”spoil” the sound quality. Following *Kerckhoffs*’ Principle [Ker1883], the security shall not rely on the confidentiality of the embedding/detection algorithm, but only on the confidentiality of the secret key.

Note: Security attacks in general are explained in Section 2.3.4.3 while the security of the approaches extended in this PhD thesis are analyzed in Chapter 4.

Symmetry vs. Asymmetry: Almost all watermarking approaches are *symmetric* security mechanisms so that the keys for embedding and detection are identical, i.e. they are secrets shared between embedder and detector . Approaches for public watermark detection are

under discussion [HG1997, CK2001, HQ2002, YL2005] but they are not comparable to the properties of asymmetric encryption algorithms for the time being.

Invertibility: In most watermarking approaches in research and practice, the embedding process is not reversible. Furthermore, in some approaches an initial embedding process leaves the watermarked content in a state that prevents it from being watermarked a second time using the *same* watermark key but with *different* watermark message. This property is useful with regards to preventing forgery of watermark messages in case that the key is compromised.

> The property was investigated by the author beyond the scope of this thesis in the context of watermarking support for secure key exchange protocols [ZSKR2010, RKSZ2010].

Conversely, there exists a class of (few) watermarking algorithms that allow restoring the original state of the audio data (so-called *invertible watermarking*).

Capacity / Data Rate: The capacity of an algorithm is defined by the amount of data embedded within a certain image area or audio playback time in *bit*. It is often specified as the *embedding rate* in *bit per second* or as the ratio of the message length divided by the file size of the cover data[21]. Current audio watermarking solutions vary from only a few bits per second (in robust watermarking) up to thousands of bits per second of payload (e.g. in semi-fragile or invertible watermarking).

Blind, Informed and Non-blind Detection: This aspect describes whether knowledge about the unmarked cover file is required for the detection or not:

- For *non-blind* watermark detection, the original data is required. Note that non-blind detection is obviously *futile* for integrity verification because tampering could be identified by comparing the suspicious media with their correspondent (assumed available) cover data directly.
- For *informed* watermarking, at least some meta data about the cover needs to be available (apart from the watermark key),
- If only the key but no further meta data is required, the approach is called *blind* watermarking in the literature[22].

In addition, the *efficiency / computational demands*[23] is usually named as important characteristic for practical applications like audio streaming or CD/DVD creation. In the context of this PhD thesis, this is of minor relevance and not analyzed much further. See Section 3.4.3 for some information on runtime performance of this thesis work.

Not all of the requirements on the presented properties can be fully met *at the same time*. Many of them are mutually competitive, especially transparency, capacity, and robustness. Hence, watermarking techniques have to be parametrized appropriately to be suitable their application scenario.

---

[21] Please note that some authors more precisely define the ”capacity” like in communication theory as the *theoretical* upper bound of information that can be embedded into the cover data, or carrier, resp. This is different from the usually *smaller* embedding rate that an algorithm *actually* achieves. However, the latter definition is commonly used in the watermarking literature and will be used in the remainder.

[22] Although the term ”blind” reminds of *visual* impairment, the wording is common for all watermarking schemes, i.e. including audio

[23] Note that this processing time properties at runtime are commonly but confusingly denoted as ”complexity” in the watermarking research community. It shall not be confused with algorithmic complexity in terms of the ”big $O(...)$” notation.

2.3.2 Audio-specific Conditions for Watermarking

Most approaches in watermarking research have been developed and presented for image data. Nevertheless, *audio watermarking* is a significant field of research in watermarking. The domain of audio data technically differs significantly from image and video with regards to opportunities and challenges for watermarking. This includes:

- **The time-dependent nature of audio data**: A raw sound signal (e.g. as PCM data) can be represented as a purely time-dependent scalar amplitude signal x_t. In contrast, raw image data like gray values or RGB values of pixels can be represented as being scalar quantities depending on *two* spatial image coordinates/indices. The same quantities in video data are even dependent on *three* coordinates (namely in the spatial and the temporal dimension).
- **Audio specific attacks**: As a consequence of the one-dimensional nature of raw sound data, serious attacks as in image/video watermarking based on *geometrical distortions* are not applicable to the audio domain (and *vice versa*) in principle or in a meaningful manner. Examples are rotation or geometric distortions (shearing, perspective projection etc.) for images, or down-mixing of stereo content to mono or adding delay effects and echoes in audio.
- **Uncompressed storage formats**: In audio data also uncompressed and lossless compressed storage formats/codecs are common as input format. For input video file data such standards are not used in practice because of huge storage or bandwidth requirements.
- **Human perception of audio:** Obviously, auditory perception greatly differs from visual perception. Psychophysical models for audio (as explained in Appendix A.2) do not apply for visual perception, and vice versa. The same is true for the superior sensitivity of the human ear to sound across many orders of magnitude in frequency and volume and to even small differences thereof.

For completeness, note that watermarking approaches for other audio-related data types are available for example for music description protocols like *MIDI* [Dit2000b], or printed music sheets [MNS2001, BSN2002].

For completeness it should also be mentioned that beyond the field of audio, the largest number of publications in watermarking is available for image and video data. Also approaches for many data types beyond multimedia like plain text [TTA2006, CC2010, HSWZ2013], vector-based drawings and fonts [WM2013], 3D models [BBC+2004, ABHB2009, TBSS2013], databases [Li2006, BSS2013], integrated circuit layouts [CT1999a], software binaries [CT1999b, SSSS2008], DNA sequence data [Lee2014] etc.

2.3.3 Authentication Watermarking Versus Other Watermarking Applications

At first the most common applications of watermarking are recalled:

- **Copyright Marking**: the embedded message gives proof of ownership; used for resolving claims of copyright
- **Transactional Marking**: messages is a*user ID* or the like for tracing individual media copies when distributed to different users; used in tracing illegal internet piracy or document leakage
- **Broadcast Monitoring**: the message allows recognizing music songs *on the air* or online, and in metering devices/apps
- **Copy Control / DRM**: the message triggers external copy protection means in terms of digital rights management (DRM)
- **User interaction/"2nd screen"**: the message triggers interactive features of TV programs or DVDs on a separate mobile phone or tablet
- **Media annotation**: the message represents arbitrary meta data about the cover data.

These applications have in common that a high level of watermarking robustness is required: the embedded message must be invariant to even strong levels of distortion or forgery attacks on the watermarked content. For more detail, the interested reader is referred to the text books by *Cox* [CMB+2007] or *Dittmann* [Dit2000a].

Authentication Watermarking Types

Watermarking can also be applied for integrity verification. In the literature on media security this is commonly (and in the author's opinion a little fuzzily[24]) subsumed under the term of *"authentication watermarking"*. To be more precise, the term can describe watermarking-based protection of either

- data integrity by itself [WM2000, FGD2001, LS2006b], or
- both integrity and data source authenticity [BBF+2000, DKSV2004, CDF2006].

In authentication watermarking methods the embedded message can indicate potential acts of tampering (see *Zhu* and *Swanson* [ZS2003]). This can be realized in different ways:

- **Fragile Authentication Watermarking:** In these approaches the embedded watermark is not robust *at all*. Instead it is being destroyed even by slight modifications of the marked content. Here, the embedded message can be seen as a "digital seal" that breaks if changes are applied to parts of the watermark-protected media. The failure of detection is intended and allows indicating that the data was potentially doctored.
- **Semi-Fragile Authentication Watermarking:** If the embedded integrity watermark message is capable to withstand a tolerable amount of modification of the marked content, the approach is often denoted as *semi-fragile* watermarking in the literature. This term expresses that the watermark here is less vulnerable to distortion than in pure fragile watermarking. For integrity verification, often a static watermark message or a dynamic time code is embedded. Here the objective is that malicious doctoring renders the watermark unreadable while minor data transformations (like lossy encoding) maintain the message.

[24] It has to be admitted that the denotation as "*authentication watermarking*" appears to be misleading because the term ought to be correspondent to the objective "authenticity" in the first place. Nevertheless it has been a *very* common wording in the research community on digital watermarking for more than fifteen years – and will be used in the remainder of this thesis.

- **Content-Fragile Authentication Watermarking** In this thesis work *"content-fragile watermarking"* is investigated. Here, deliberately selected content-dependent meta data about the original state of the cover data is embedded as a robust watermark message. In the context of this thesis work, perpetual hashes will be used as embedded authentication code messages. The initial publication about the overall concept of combining perceptual hashes with digital watermarking in this manner was published by *Dittmann* [Dit2001] already in 2001.

About the latter, the wording in the literature is ambiguous to some degree, unfortunately, as the concept of "content-fragile watermarking" is also denoted as

- *"mixed watermarking-fingerprinting"* [GCdG+2002],
- *"self-embedding"* [FG1999, YH2004, NAK+2016],
- *"content-based authentication watermarking"* [GMS2008a].
- *"semi-fragile (signature) watermarking"* [FKK2004].

Note that the wording as *"content-..."* by *Dittmann* or *Gulbis* is ambiguous to some degree as the authors do not *exactly* describe which level of identity (e.g. in terms of the definition in Section 1.2.2) shall be achieved. Nevertheless, all approaches using the listed wording operate beyond the level of binary identity though.

Comparison

First of all, it is important to notice that the content-fragile watermarking technique in general has much higher capacity requirements than in applications related to protecting copyrights/anti-piracy. In this thesis work a dynamic authentication code and additional metadata of approximately 150 *bit* needs to be embedded every few seconds. Conversely, in copyright transaction watermarking for example it is sufficient to embed a static or dynamic ID of a few dozen bits for a few times somewhere in the file.

On the other hand, the robustness requirements in content fragile watermarking are smaller to some extent: in the case of very strong distortions being applied to a watermarked media (for whatsoever reason) the watermark detection might be faulty or will even fail. In authentication watermarking this can be tolerated as it (nevertheless) indicates breaches of integrity.

Finally, a very good transparency is desired in almost any watermarking applications. Audible quality loss on the recipient side can hardly be acceptable in any application – neither in this thesis work.

An exhaustive overview about the requirements on watermarking in different applications is given in Table 2.4.

2.3.4 State of the Art in Audio Watermarking

This Section gives an overview on publications on the approaches in general audio watermarking. It discusses the history of research on audio watermarking, algorithms from the state of the art, their security and other aspects in the field.

2.3.4.1 History of Audio Data Hiding Research

According to *Cox* and *Miller* [CM2002] the technique of *"electronic watermarking"* has been under research and development since the 1950s. But not until the mid 1990s, academic re-

Watermarking purpose	Robustness	Capacity	Transparency	Security
Fragile authentication	low	various	high	high
Semi-fragile auth.	medium	low	high	high
Content-fragile auth.	medium	high	high	high
Copyright / ownership	high	low	high	high
Transaction tracing	high	low	high	high
Copy Control	high	low	high	high
Broadcast monitoring	high	low	high	high
2nd Screen	high	low	high	low
Annotation	medium	high	high	low

Table 2.4.: Comparison of watermarking characteristics across different applications; dashed lines: watermarking grouped into the application fields of "data authentication" (upper area), "copyright/anti-piracy" activities (middle) and miscellaneous "rich media services" (lower).

search in the field significantly increased. It was sparked when *digital* recording, processing, storage, online distribution, and illegal piracy became more and more apparent. That time, many "pioneer works" were presented at/in special conferences and journal issues on multimedia security or signal processing only. Examples are the *ACM Information Hiding*, *ACM MMSEC*, *ACM IH&MMSEC*, or the *SSWMC* multimedia security conference hosted at the *IS&T SPIE Electronic Imaging*.

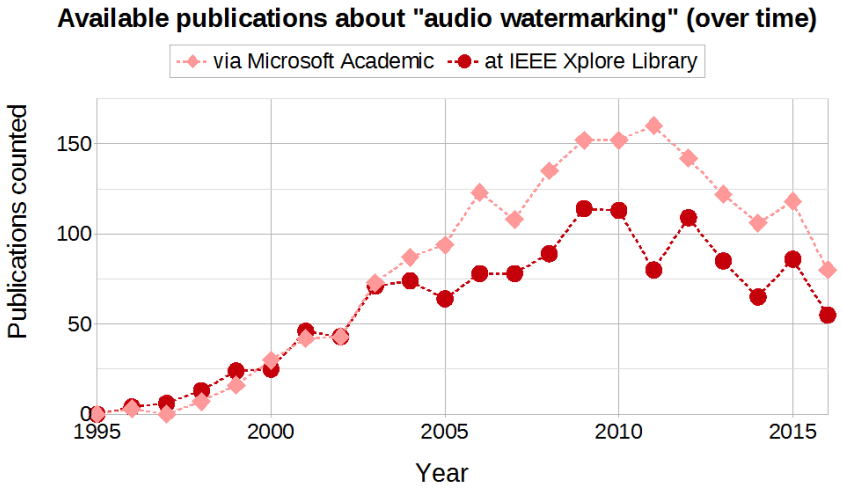


Figure 2.8.: Number of available publications about "audio watermarking" (over time) that are available at/via *IEEE Xplore Digital Library* and *Microsoft Academic Search*[25] (retrieved February 2017)

In the early 2000s, a few more watermarking-focused conferences could establish itself due to increasing attention, for example the *IEEE IIH-MSP* and *IWDW* conferences. From then, the topic has received increasing attention also by related *major* journals and conferences with larger audience and greater attention, for example conferences *ACM Multimedia*, *IEEE ICASSP*,

[25] *IEEE Xplore Digital Library: loc. cit.*,
Microsoft Academic: http://academic.research.microsoft.com

IEEE ICIP, the *IEEE* journals *"Multimedia"* and *"Signal Processing"*, the *"IEEE/ACM Transactions on Audio, Speech, and Language Processing"*, and the journal *"Digital Signal Processing"* published by *Elsevier*, as can be seen from Figure 2.8.

Nevertheless, watermarking and the related field of steganography have remained under-represented at major conferences in general IT security though. As an example, at the five most frequently-cited conferences on general information security[26] in the years 2010-2015, there had been very few (i.e. less than five) publications listed on the term "audio watermarking".

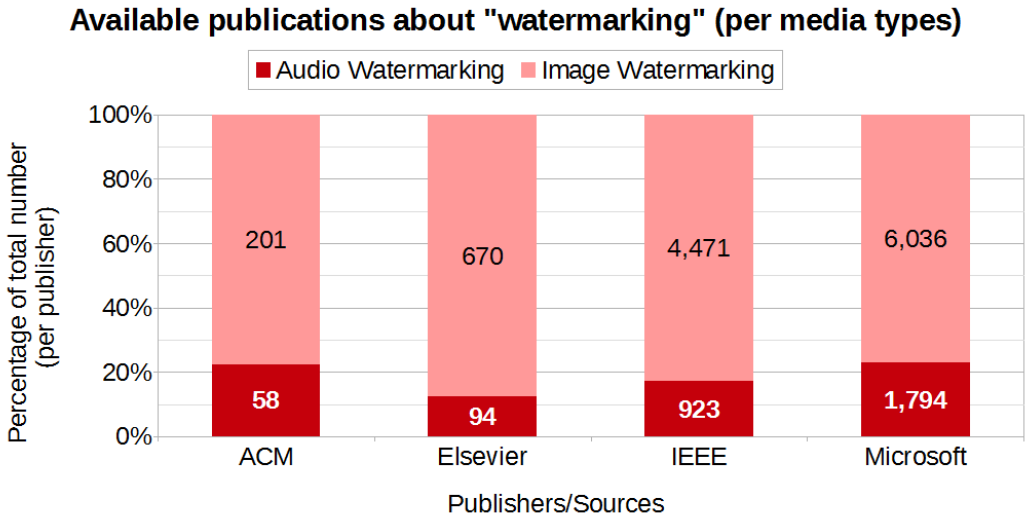


Figure 2.9.: Percentage of available publications about "audio watermarking" versus "image watermarking" by/via *ACM Digital Library, Elsevier/ScienceDirect, IEEE Xplore Digital Library, Microsoft Academic*; total numbers given bar-wise data descriptors (retrieved February 2017)

Finally, resources in watermarking for *audio* has always lagged behind the watermarking research for other media types, especially watermarking for image data. As an example, at the well-respected resources at *ACM Digital Library*[27], *IEEE Xplore Digital Library*[28], *Elsevier/ScienceDirect*[29], and *Microsoft Academic*[30], the total number of publications on "*image AND watermarking*" in their abstracts is at least *four* times greater than for "*audio AND watermarking*". That means that audio watermarking is of moderate interest in the research community as compared to watermarking for image data, unfortunately.

[26] "Top 5" conferences according to the "Impact Rating" provided by the *Microsoft Academic Search* (*loc. cit.*); namely: *CCS*, *EUROCRYPT*, *S&P*, *USENIX*, and *CRYPTO* (Impact Rating values retrieved May 2015; rating service now being decommissioned at the website, see: https://academic.microsoft.com/FAQ)

[27] *ACM* Digital Library: *http://dl.acm.org*

[28] *loc. cit.*

[29] *Elsevier/ScienceDirect: http://www.sciencedirect.com*

[30] *loc. cit.*

2.3.4.2 Audio Watermarking Algorithms

The works of the state of the art can be categorized as follows:

Feature Domain: The data domain for embedding can be either the time domain or an elaborately selected spectral domain (like *Fourier*, DCT, Wavelet, SVD etc.).

Embedding Principle: Several fundamental embedding methods of embedding and detection algorithms are known. Most important are so-called echo based, Spread Spectrum, Patchwork, and QIM approaches etc. which will be explained in the following. Note that these principles can be applied to watermarking for other media types like image, video etc., too.

In the following the most important classes of algorithms and corespondent publications are reviewed in rather chronological order of evolution. The interested reader is referred to the recent work by *Hua et al.* [HHS+2016] for a more detailed review and discussion on a hierarchical categorization of algorithms.

Notch-filtering based approach by *Hembrook*

One of the first known *analog* data hiding methods for copyright protection of audio content was developed by *Muzak Inc.* from Seattle, USA, and eventually patented in the 1960s [Hem1961]. An analog notch filter was applied to the audio spectrum that selectively muted a very narrow sub-band in the audio spectrum at 1.0 *kHz*. Here, the time intervals of the sub-band removal represent the embedded messages in *Morse* code. The deliberate filtering of this sub-band usually can hardly be perceived by a human listener. But the absence of this frequency component can easily be detected using a matching narrow band-pass filter and which reveals the hidden *Morse* code.

Correlation-based Embedding/Detection in the Temporal Domain

Early approaches are based on adding pseudo-noise sequences to the PCM signal by mixing in the time domain. Different pseudo-noise sequences or respective offsets in these sequences correspond to different message symbols. The detection and retrieval of the watermark is conducted using correlation-based analysis of the marked audio. Simplified, the *Nyman-Pearson* cross correlation coefficient indicates the presence of the respective message symbol.

One of the most frequently cited ”pioneer works” following this principle was presented by *Boney, Tewfik and Swanson* [BTH1996, SZTB1998]. Here, a pseudo-random noise sequence is added to the input signal in the time domain. This work meant a significant progress to the state of the art and was one of the first works that exploited frequency masking characteristics: Prior to its adding, the noise sequence is subject to filtering according to a psychoacoustic model. Here, the authors could benefit from research on lossy audio coding that made a lot of progress in parallel, those days. A number of approaches extending these works were presented, for example by *Bassia* [BP1998] or *Lemma et al.* [LAOVdK2003].

Spread Spectrum Watermarking in the Spectral Domain

Adding pseudo-noise sequences in temporal or frequency domain can be regarded as an example of the general *spread spectrum* technique. In watermarking, adding a noise sequence effectively means adding a noisy *wide-band* signal. That is, the watermark message/signal is being ”spread” over a wide frequency band. This is why many watermarking approaches have been proposed explicitly under the name of *spread spectrum watermarking* since the late 1990s.

The most frequently cited spread spectrum based works were published by *Cox et al.* [CKLS1996, CKLS1997] or *Kirovski et al.* [KM2003]. Many works proposed spread spectrum algorithms for the spectral domain instead of the temporal domain, for example *Chung et al.* [WSK2000], *Li et al.* [LY2000] *Haitsma et al.* [HvdVKB2000], *Lu et al.* [LLC2000], and recently by *Hamdouni* [HALT2013] and *Khalil* [KA2015]. It is still subject of theoretical analysis e.g. about the maximum channel embedding capacity [ZXH2015].

Again the presence of the watermark is detected by calculating the correlation coefficient between the (known) watermark pattern and the audio data (which is carried out in the spectral domain instead of the temporal domain). Interestingly, a work by *Nakashima* uses the spread spectrum not only for detecting the watermark but also for estimating the spatial location of the detector in a movie theater at detection time [NTNB2006].

Note that spread spectrum technology is rather a general principle from communication theory than a watermarking-focused algorithm. It was firstly investigated much earlier, at the beginning of the 20th century for example by *Nikolai Tesla* [Tes1903]. It was increasingly investigated in the course of World War I and World War II for secure radio communication [GK1942]. Nowadays, Spread Spectrum technology is still being adapted in military or civil telecommunication like cellphone communication (before *4G*) or in *WLAN*.

"Patchwork" Watermarking

In so-called *Patchwork* watermarking, the embedding of a message bit is based on enforcing that certain statistical signal properties of the data are changed to abnormal values that unmarked cover data usually does not have [BGML1996]. The watermark detector can later identify the presence of such statistical anomalies and can retrieve the watermark.

The detailed description of the Patchwork principle and the origin of its name can be found in the later Section 2.3.5.1. A Patchwork audio watermarking approach in the DFT domain proposed by *Steinebach* [Ste2003] is utilized in this thesis work and is explained in detail in Section 2.3.5.2.

Other publications on audio Patchwork watermarking were presented in the DCT domain by *Yeo et al.* [YK2001], or in the FFT domain by *Arnold et al.* [Arn2000] and by *Tachibana et al.* [Tac2003]. Interestingly, the latter approach was optimized for processing delay especially in *live* processing. More recently, an approach for the cepstrum domain (see definition in Section A.1.4.1) was presented by *Hu* [HC2012] while *Natgunanathan* extends existing DCT-based approaches [NXE+2013].

Note that the Patchwork principle and the previously explained spread spectrum principle are analytically equivalent as will be shown as a corollary in Section 2.3.5.3 below. However, both terms have been commonly used in the literature over the years and hence are given here separately for historical reasons.

SVD-based embedding

Another kind of embedding in the spectral domain is utilizing the mathematical formalism of the *singular value decomposition* (SVD) as recalled in Appendix A.1.4.2. The SVD allows identifying independent (and hence robust) components in the input cover data. In numerous fields of science and engineering the SVD formalism is used as a "mathematical tool" – so it is in watermarking: Embedding a message bit by modifying the spectrum of singular values was found to be an effective starting point for many SVD-based watermarking schemes. They offer high embedding capacity (up to a few hundred bits per second) especially in recent years for

example by *Jian Wang* [WHT2011]: Here, the DCT spectrum in selected sub bands is analyzed in terms of the SVD and then watermarked in the "SVD domain". Other SVD-based works were presented by *Lei* or *Dhar* [LSZ+2012, LST2013, DS2013].

Quantization Index Modulation (QIM)

A few years after the introduction of Spread Spectrum watermarking the so-called *quantization index modulation* (QIM) was proposed. This approach embeds watermark information by enforcing the quantization properties of the cover data to a pre-determined scale. For the two message symbols "one" and "zero", two different quantizers are defined and applied on the input data. On the detector side, the message bit symbol is retrieved from the marked sample by identifying which of the two quantizers is the best match. It was introduced and defined by *Brian Chenet al.* in a very frequently cited work [CW2001].

In the work by *Xiang-Yang Wang* [WNY2009] the watermark bit is embedded into the statistical mean value of low frequency components in the Wavelet domain according to human auditory masking.

A more general overview / meta study of QIM embedding into mean values in *whatever* spectral representation was recently given by [LZ2016].

In the already mentioned work by *Lei* [LSZ+2012] a watermark is inserted in the coefficients of the LWT low frequency sub-band taking advantage of both singular value decomposition (SVD) and QIM.

A frequently cited work by *Xinkai Wang* proposes QIM embedding in the Wavelet domain, too [WWZ+2013]. Although presenting promising results the work lacks a thorough experimental evaluation (as will be discussed in Section 2.3.6 on "Cursory Transparency Evaluation").

Finally, very recent works by *Fallahpour* [FM2012, FM2015] and *Hu* [HHC2014, HH2015] applying the QIM principle in the DFT, DCT, and Wavelet domain, resp., have to be mentioned. Quantization steps are defined by means of *Fibonacci* series. Detailed results can be found in Section 2.3.6 on "High-capacity Embedding". These works were recently picked up by *Neethu et al.* for QIM embedding in the *Haar*-Wavelet domain [NK2016]. However, they are not further investigated as the evaluation in the publications based only a very small number of audio samples and human test subjects for listening tests.

Least Significant Bit Watermarking (LSB Watermarking)

The most simple implementation of QIM is the so called *least significant bit watermarking*. Here, the integer representation of cover data is used as the starting point. The respective least significant bits (LSBs) of a set of cover data samples are *replaced* by the binary message symbols. The two different message symbols *"1"* and *"0"* correspond to the quantizers for *"odd"* and *"even"* quantization which shows that LSB watermarking is a very simple implementation of the QIM principle.

Phase Coding

As an alternative to modifying the absolute value of FFT coefficients, also the *FFT phase* can be exploited. This is taking advantage from the fact that the human ear is insensitive to slow variations of the FFT phase, as explained early by *Bender et al.* [BGML1996]. The principle is exploited in works e.g. by *Dong et al.* [DBI2004] and *Arnold et al.* [ABV2009, ACB+2014]. The latter is also the technical basis of the commercial audio watermarking systems by *Technicolor*

(formerly *Thomson*). Finally, a recent work by *Ngo* and *Unoki* [NU2015] proposes a QIM-like approach on the FFT phase coefficients.

Echo Hiding

An early approach introduces imperceptible echoes for embedding the message as presented by *Gruhl, Lu, and Bender* [GLB1996]. Here, different durations of delay (a few milliseconds) are used for embedding different message symbols. The delay value of the echo can be detected subsequently and the message symbol can be retrieved. Interestingly, embedding and detection are carried out in different domains: embedding is done by applying an echo filter kernel in the *temporal* domain. The detection is usually done in the *cepstrum* domain (see definition in Appendix A.1.4.1).

More recent works on echo hiding were published by *Wen-Chih Wu* [WC2008], *Huiqin Wang* [WNSM2008], *Erfani* [ES2009], and *Guang Hua* [HGT2015].

Muteness-based Approach

Another approach was presented by *Kaabneh et al.* It slightly modifies the duration of silent (muted) sections in an audio. It is well suited for voice recordings in which such speech pauses are usually available for embedding [KY2001]. It can not be well categorized into the notation explained so far and is mentioned for the sake of completeness.

Bitstream Embedding

In many commercial applications, the audio data to be protected is available or will be distributed in compressed file formats. For this, there exist embedding approaches by modifying the encoded bitstream *directly*. Approaches were presented

- for *MP3* data by *Arnold, Schmucker et al.* [ASW2003, pp. 99],
- for *AAC* data by *Allemanche, Siebenhaar, Neubauer* and *Herre* [AH2000, SH2001],
- for MP2 data by *Nahrstedt* [NQ1998], by *Steinebach* and *Dittmann* [SD2003a] or by *Quan* [QZ2006],
- for AC-3 data by *Xiao-Ming Chen, Arnold* [CABD2012], and
- for GSM data by *Yuan* and *Huss* [YH2004].

For a more detailed description of the latter work by *Yuan* see Section 2.5.

Bitstream embedding has the obvious advantage that computationally intensive calculations for spectral transforms, decoding etc. can be avoided if the input audio is available in compressed formats anyway. This allowed providing fast *on-the-fly* watermarking solutions a few dozen times faster than real-time already in the early 2000s.

2.3.4.3 Security of Watermarking

This Section discusses the state of the art in watermarking in light of its security aspects. The term *security* here refers in particular to the aspects of confidentiality, integrity and availability of the embedded watermark message.

This discussion is important because the (rather provocative) statement

> *"watermarking is not cryptography"*

by *Cox, Doërr* and *Furon* in their paper of same name [CDF2006] is true in many ways. For example, the authors distinguish the term *key length* in watermarking from its meaning in cryptanalysis. From a technical point of view the authors explain that the ”key-space” analogy as in cryptography does not fully apply to watermarking: Very often, the watermark key is used as the seed to a pseudo-random number generator. Its output is then used for controlling essential processing steps in the course of embedding/detection.

One example is pseudo-randomly ”picking” a subset of cover data features used for embedding (as explained in the example in Section 2.3.5). Here an attacker does not necessarily require to disclose the secret key completely for recovering the pseudo-random sequence. Instead, recovering the picked subsets directly (or even a sub-subset thereof) is sufficient.

This discussion was resumed by *Bas* and *Furon* [BF2012] by raising the question:

> *”Are 128 bits long keys possible in watermarking?”*

in their publication of same name. The same authors conclude [BF2013, p. 3] that in watermark detection – unlike in symmetric cryptography – the symmetric key as a shared ”secret” is not unique. In other words, although not identical, a number of those shared secrets are in effect *equivalent* as all of them allow detection and retrieval of the watermark message. Note that this concept is different for example from standard encryption technologies in cryptography in which usually *the one and only* private or public key grants access to the plain text. This observation led to *Bas*’ and *Furon*’s formal definition of the *”effective key length”* of a watermarking system. Its size describes the *actual* difficulty for an adversary to get access to the watermarking message. The elaboration in Chapter 4 discusses the consequences for the example of the Patchwork watermarking algorithm as technically proposed this thesis work.

Note that beyond the pure technical aspects, *Cox*’ provocative statement is also true from a more general perspective: It reflects well that the security of watermarking algorithms and protocols has always been of minor interest in the research community on watermarking. It has never been investigated as thoroughly (in terms of man power) as for example in cryptanalysis. By far most watermarking publications present a cursory, and sometimes even negligent, view on the security against targeted attacks. Instead, many authors focus on the robustness and on application scenarios. Occasional exceptions were the security analysis on the core algorithms or protocols in different works over time, as for example by *Craver/Felten, Cayre, Fridrich* and *Venturini* [PAK1998, CMYY1998, CK2001, CWL$^+$2001, Fri2002, Ven2004, CFF2005]; or the work by Petitcolas [PAK1998] which is recalled in the following Section.

What adds to the security aspects of watermarking is that the term ”key” is frequently and ambiguously used in a different manner in the watermarking research community. Some authors propose applying whatever encryption of the watermark message $\mathscr{M}$ using a secret ”key” [DDPP2013, FM2015, NK2016, TLMA2016]. This shall not be confused with the true watermarking key that defines where and how the message is embedded into the cover media.

General Security Attacks

In many application scenarios the watermarked media and the embedded message can be subject to the following security attacks by an adversary [PAK1998, VPP$^+$2001].

Removal / Replacement Attack: One obvious attack on the availability of the watermark message is given by applying strong global distortions on the watermarked media. This could allow removing the embedded watermark message before detection time. Attacks can aim at information leakage about the watermark key $\mathscr{K}_1$ for localizing / estimating the data subsets that are actually watermarked inside the media. Then, targeted distortions can be

applied there to remove the message. Even more, the message can be retrieved or even forged which means an attack on confidentiality and integrity of the message as well.

De-Synchronization Attack: Many watermarking algorithms use synchronization approaches that at first indicate to the detector the presence and the start of an embedded watermark message. Thus, an attacker can try to prevent watermark detection and retrieval by attacking the synchronization pattern rather than attacking the message itself. This attack is mainly aimed at the availability of the message detection.

Oracle Attack / Sensitivity Attack: An adversary can implement attacks systematically on the embedded message by introducing distortions, adding noise, re-encoding etc. If the detector of a watermarking algorithm is available to the public, it can serve as an *oracle*: the detector then indicates the success or the failure of the attack and hence the sensitivity of the algorithm to certain attacks [CL1997]. Oracle attacks can assist in estimating the data subsets that are actually watermarked. An early example for successfully attacking copy control watermarks was published by *Craver, Felten et al.* in 2001 [CWL+2001] and was under *heavy* debate those days in the course of the so-called *SDMI* competition.

Copy Attack: Security also includes the behavior of the algorithm when the watermark is somehow separated and copied to a different media file. This is especially relevant if the unmarked cover data is available to an adversary. In this case, the *difference signal* (obtained from subtracting the PCM audio data sets of the cover data minus the marked data, sample by sample) often allows separating the ”watermark” and copying it to another file.

Collusion Attack: Collusion attacks exploit vulnerabilities of transaction watermarking. Here usually different watermarked copies of the same cover with different embedded message are available. Such attacks can be implemented by mixing the different copies. Collusion attacks are only little relevant on pure integrity watermarking as it requires many differently watermarked copies as in the transaction watermarking scenario. It is not further elaborated upon in the remainder of this document.

> Beyond the scope of this thesis own publications on collusion resistance [SZ2006a, SBH+2010, SBZS2010, BZSS2011, SBZS2012] were published in co-authorship, see commented publication list at pp. 157.

Ambiguity Attack/Invertibility Attacks: This protocol attacks aim at finding the presence of an ostensible watermark in the media that appears as a valid watermark message but that had never been actually embedded before [CMYY1997, CMYY1998]. Such ”virtual” watermarks could cause confusion by *ambiguous* claims of ownership. The confusion can eventually render a watermarking based protection useless. Hence, no apparently correct watermark message should be detectable from actually *unmarked* media data. This can be rephrased into the requirement that the detection needs to feature a very low false positive rate [AKS2003].

These security challenges exist in watermarking for *any* media type. The mentioned attacks also mean a threat to watermarking research and technology in general: proven successful attacks have the potential to make watermarking appear as futile or even compromising. When watermarks can be deleted easily, innocent users are accused by forged watermark messages, or false detection are triggered too often, the trust in watermarking-based security mechanisms as a whole can get lost.

Security challenges often overlap with robustness challenges from a technical point of view. For example, lossy *MP3* transcoding of an audio can be applied by a user both in the course of everyday *fair use* and as a *malicious* act.

Attacks on Integrity Watermarking

Specific kind of attacks apply to watermarking systems for integrity protection. Adversaries can exploit weaknesses in both the watermark embedding/detection and (if applicable) in the perceptual hashing involved. The most important attack models were explained by *Fridrich* [Fri2002] and are recalled in the following:

Undetected Modification: Obviously, an adversary can attempt to apply tampering to the protected media that – by itself – is little enough to remain undetected in the verification stage. Examples are deletions of content that are too short in an audio or too small in an image file. As in any feature based classification mechanism, a reasonable trade-off between missed detections and false alarms must be identified. If a metrics for the degree of modification can be defined, a suitable numerical decision threshold for the detector's response can be chosen.

In addition, attacks are defined that can fool the detector by causing misinterpretation of the detection result. For example, *Holliman* and *Memon* discussed [HM2000] an attack by exchanging independently protected units inside or between protected media. Without further ado, such exchanging of units might remain undetected.

Oracle Attack/Sensitivity Attack: One example of allowing undetected integrity breaches are oracle attacks (as described in the previous Section for watermark embedding). If unlimited access to the verification stage is available, the sensitivity of the approach can be studied by systematic and gradual attacks. Hence vulnerabilities could be identified.

Information Leakage: Another kind of attacks attempt to obtain information about the secret authentication key (e.g. if a key-dependent perceptual hash is used) or the placement of the perceptual hash extraction. The same applies for the security of the watermark key. This information leakage can assist attempts to apply undetected modifications in a very targeted manner and/or subsequently overwriting the original watermark message.

Protocol Weakness: In combination with oracle/sensitivity attacks, authentication watermarking can be circumvented on a protocol level as well. For example, an adversary can attempt to modify a protected audio media so that it sounds similar but its correspondent hash becomes significantly different. This intended violation of the robustness requirement will introduce many false alarms which eventually renders the complete scheme unreliable and useless.

> One such attack on an existent audio hashing algorithm was developed as a minor research activity and published by the thesis author in co-authorship [SZN2011].

Research activities in the course of this PhD thesis address some of the above security aspects, as explained in Chapter 4.

2.3.4.4 Other Works in the Field

Publications on Watermarking Benchmarking (by *Lang / Petitcolas / IHC Committee*)

Beyond the description of concrete audio watermarking algorithms and their security, research was conducted on *benchmarking* of audio watermarking. One research project on this was the *Stirmark Benchmark Audio (SMBA)* project[31] driven by *Lang* [SDS+2001, LDS2003, DSLZ2004]. It was inspired by *Stirmark* for image watermarking by *Petitcolas* [Pet2000]. It offers a set of

[31] *Stirmark Benchmark Audio (SMBA) website: http://omen.cs.uni-magdeburg.de/alang/smba.php#smba_get*

standardized robustness and security attacks on marked audio content. Although very promising and respected in peer-reviewed publications, *SMBA* never got widely accepted as a *standard tool* in the research community or industry. If being used by other authors at all, only subsets of *SMBA* attacks are tested in publications. The project is not pursued anymore[32].

Another benchmarking activity has been promoted in recent years by the *Information Hiding Criteria (IHC) Committee*[33]. Suggested evaluation criteria are published by the IHC Comittee as an *"Evaluation Procedure"* [IHC2016].

Overview Publications, Textbooks, Non-Academic White Papers (edited by *Cvejic* or *He*)

Apart from detailed algorithms in journals and conferences, a few textbooks or elaborate overview/introductory publications on watermarking were published by *Bender* [BGML1996], *Hartung and Kutter* [HK1999], *Cox, Miller, and Bloom* [CMBM2001], Katzenbeisser [KP2000], *Dittmann* [Dit2000a], or *Arnold et al.* [ASW2003]. These early works (and their respective later editions like [CMB+2007]) can be regarded as valuable standard references for this research subject.

Also a small number of reviews especially on *audio* watermarking were published in the past years, for example by *Cvejic et al.* [CS2007b], by *He* [He2008] or recently by *Hua et al.* on *"twenty years of digital audio watermarking"* [HHS+2016].

For completeness it should be mentioned that many use cases are presented and compared in White Papers or meta studies by non-academic business organizations, for example by the *European Broadcasting Unit* [EBU2004] or repeatedly by the *Digital Watermarking Alliance*[34], the *2nd Screen Society*[35] or the *Movie Labs*[36].

2.3.5 Example: Fourier-based Patchwork Audio Watermarking by *Steinebach*

A well-known embedding principle in watermarking is the *Patchwork* embedding which will be used in the course of this PhD thesis. It was originally introduced by *Bender et al.* as a statistical embedding and detection approach [BGML1996]. Here, the embedding of a message bit relies on enforcing statistical signal properties of the data to "abnormal" values that unmarked cover data usually does not show.

2.3.5.1 Patchwork Embedding and Detection Principle

Generally spoken, the Patchwork embedding/detection is performed in the following processing steps: At first, for embedding each single message bit two disjoint subsets

$$A := \{a_i\} \quad B := \{b_j\} \quad , \quad A \cap B = \{\} \quad , \quad i, j = 1 \ldots N$$

[32] The *SMBA* project website states that it was *"last modified"* in Dec. 2007 and *"free download of this tool is closed"*. Access to the original *SMBA* audio software is discontinued and was denied to the thesis author.

[33] *Information Hiding Criteria (IHC) Committee*, chaired by Tokyo University of Science.
IHC Comittee website: *http://www.ieice.org/iss/emm/ihc/en*.
Note: *IHC Comittee* not to be confused with *Information Hiding conference (IH)*.

[34] *The Digital Watermarking Alliance: http://www.digitalwatermarkingalliance.org*

[35] *2nd Screen Society: http://www.2ndscreensociety.com*

[36] *Movie Labs: http://www.movielabs.com*

of data samples are pseudo-randomly selected from the given cover data. It can be picked from the Fourier magnitude spectrum of an audio frame, as in this thesis approach.

For the following, these sample values are regarded as *random variables*. Under the reasonable assumption that the N data sample values are identically distributed, sufficiently independent and that the subsets contain sufficiently many data samples the respective mean values

$$\bar{a} := 1/N \sum_{i=1}^{N} a_i \qquad \bar{b} := 1/N \sum_{j=1}^{N} b_j$$

will be *approximately identical* which can be derived from the *Central Limit Theorem* by *Lindeberg-Lévy*, see [Stö1995, p. 743]. Hence, the random variable

$$s := \bar{a} - \bar{b}$$

shows an expectation value close to zero. In practice, for a skillful selection of the cover data representation, this assumption is sufficiently true.

Then, for embedding a message bit symbol $m =$ "1" each sample value in A is slightly increased while each sample value in B is decreased slightly; and *vice versa* if $m =$ "0", respectively. As a result, the mean values will differ and quantity s is enforced to deviate from its zero expectation. This property robustly withstands many kinds of signal transformations for multimedia data.

The modified coefficients are finally written back to the cover data by replacing the original values in-place.

The name of the algorithm was inspired by the original design of the approach by *Bender et al.* [BGML1996] for digital images. There, the areas of modified pixel subsets A and B represent the shape of a *patchwork* pattern.

Patchwork Detection

For the detection, the same subsets $A' = \{a'_i\}$ and $B' = \{b'_j\}$ (now being watermarked) are selected from the marked data again. Then the correspondent mean value among those subsets $\bar{a}'$ and $\bar{b}'$ are compared to get an estimator of the embedded message bit. For this, the quantity

$$s' := \bar{a}' - \bar{b}' = 1/N \sum_{i} a'_i - 1/N \sum_{j} b'_j \tag{2.5}$$

is evaluated against a decision threshold d_0, again, which allows separating marked from unmarked content:

$$\begin{aligned} s' > d_0 \quad &\Rightarrow \quad \hat{m} = \text{"1"} \\ s' < -d_0 \quad &\Rightarrow \quad \hat{m} = \text{"0"} \quad . \end{aligned} \tag{2.6}$$

The comparison in the previous equation is done based on data from the watermarked content only. As no cover data is involved here, the Patchwork approach allows *blind* detection. However, a non-blind extension of the algorithm can be constructed as well.

2.3.5.2 Patchwork Embedding/Detection Algorithm in the DFT Domain

An audio Patchwork approach for watermarking absolute values of FFT coefficients was originally proposed by *Steinebach* [Ste2003]. It will be applied on the approach in this PhD thesis as presented in Chapter 3.

> It was investigated [SZSL2003] and further developed with regards to robustness and transparency as a research activity of this thesis and results were eventually published in co-authorship [SZ2008a, SZ2008b].

Simplified, the embedding and detection is carried out as follows:

1. **Input:** The input audio data is expected as PCM samples in the time domain. The PCM file or stream is divided in frames of $L=2048$ samples. At a sample rate of 44.1 *kHz* this represents 46 milliseconds of duration. In practice, the total number of modified frames T is a few hundred (which represents a few seconds).
2. **Windowing:** The frame samples are multiplied element-wise with a *Hamming* window for reducing spectral leakage effects, as explained in Appendix A.1.3.3.
3. **Fourier Transform:** The windowed frame data is transformed to the spectral domain using the Fast Fourier Transform (FFT), obtaining $L/2$ absolute value coefficients (r_k) and phase coefficients (ϕ_k).
4. **Selection of Patchwork Pattern:** Then, in each frame's spectrum two disjoint subsets $A=\{a_i\}:=\{a_{i_n}\}$ and $B=\{b_j\}:=\{b_{j_n}\}$ of N Fourier absolute coefficients each are "picked" pseudo-randomly from the spectrum of the $\{r_k\}$. Formally, such picking of coefficients is done by pseudo-randomly picking the respective N *indices* i_n and j_n from the possible range $i_n, j_n \in \{1...L/2\} \in \mathbb{N}$ and assigning them to the sets[37] of indices $\mathscr{A}$ and $\mathscr{B}$:

$$\mathscr{A} := \{i_n\} \quad \mathscr{B} := \{j_n\} \quad , \quad |\mathscr{A}| = |\mathscr{B}| = N \quad , \quad \mathscr{A} \cap \mathscr{B} = \emptyset \quad , \quad n = 1 \ldots N \quad .$$

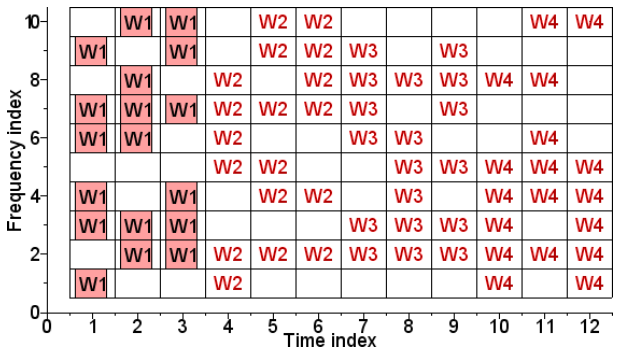


Figure 2.10.: Example: Simplified Patchwork embedding example in time-frequency domain: every 9+9 correspondent coefficients *Wn* contribute to the *n*-th message bit according to Equation (2.6). Coefficients for the first watermarking bit *W1* are highlighted for demonstration purpose. Note: in practice, the number available frequencies (the "height" of the above area) reaches a few hundred indices instead of 10.

[37] The notation $|S|$ about a set S shall denote its *cardinality* i.e. the total number of elements in S.

Figure 2.10 gives a simplified visualization of the embedding approach: For increasing the robustness furthermore, a number of consecutive frames is used to embed the n-th message bit repeatedly. That is, the set of picked coefficients defines a pseudo random pattern across different time steps. The sequence of picked frequency indices varies over time, i.e. from frame to frame. This increases the transparency and the efforts for an adversary for unauthorized access to the watermark message. The pseudo-random selection is usually done dependent on the watermark secret key $\mathscr{K}$ which is used as a seed.

5. **Patchwork Embedding:** For embedding a message bit, the Fourier coefficients are modified according to the Patchwork principle explained above in Section 2.3.5.1. The in-

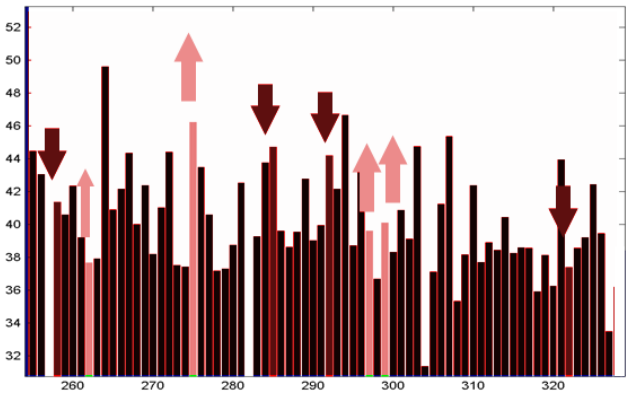


Figure 2.11.: Example: Patchwork embedding principle in Fourier spectrum; horizontal axis: frequency index, vertical axis: power (in *dB*); arrows: increase/decrease of selected spectral coefficients

crease/decrease is analytically carried out in an *exponential* representation for technical reasons:

$$m = \text{"1"} \quad \Rightarrow \quad a'_{i_n} = a_{i_n}^{1+d} \quad b'_{j_m} = b_{j_m}^{1-d} \qquad \forall\, n = 1 \ldots N, \quad d \geq 0 \tag{2.7}$$

$$m = \text{"0"} \quad \Rightarrow \quad a'_{i_n} = a_{i_n}^{1-d} \quad b'_{j_m} = b_{j_m}^{1+d} \qquad \forall\, n = 1 \ldots N, \quad d \geq 0 \tag{2.8}$$

The modified coefficients are written back to the magnitude spectrum by replacing the original values in-place and the watermarked spectrum (r'_n) is obtained.

6. **Perceptual Modeling:** The degree of increase/decrease of coefficients in the previous step is controlled by an implementation of the *ISO MPEG-1* psycho acoustic model [ISO1993b, Appendix D], see explanation in Section 2.4.1. That is, the exponents are actually individually calculated for the n-th spectral coefficient, i.e. $d = d_n$. In practice, the range of the exponents $(1 \pm d_n)$ is typically in the range of 0.9 to 1.1 for reasonable embedding distortion. The left plot in Figure 3.11 in the next Chapter shows the distribution of $|d|$ of an empirical example.

7. **Inverse Fourier Transform:** Finally, the modified audio spectrum (r'_n, ϕ_n) is transformed back to the spectral domain using the inverse FFT. Recall that the phase information (ϕ_n) needs to be available for this. Unlike in "phase coding" embedding schemes [DBI2004, ABV2009, ACB+2014] (see Section 2.3.4.2), the Fourier phase is not modified. It is just copied from the FFT output in Step 3.

8. **Fading Watermarked Data with Cover Data:** Recall that the cover data was multiplied with a windowing function for avoiding spectral leakage effects. For avoiding ”ripple” artifacts because of the windowing, the watermarked data is cross-faded with the cover data at the frame bounds.

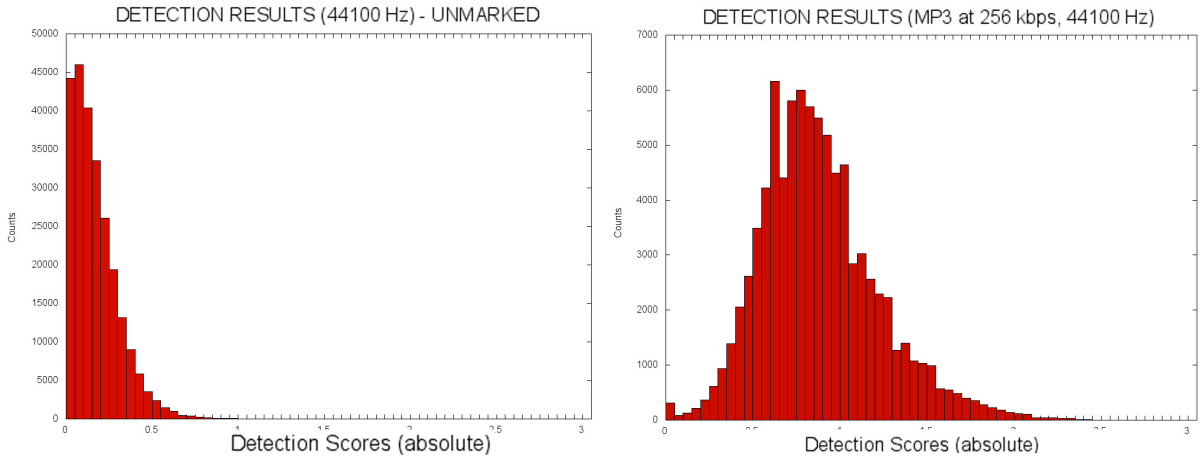


Figure 2.12.: Example: Patchwork detection scores: distribution of (apparent) detection scores in example content. For simplification purposes, the absolute values $|s|$ are displayed;
Left Figure: result in actually unmarked content;
Right Figure: result in marked content after *MP3* compression at 256 *kbit/s* in joint-stereo

On the *detector* side, the steps 1 to 4 will be carried out again. Then the quantity

$$s := \bar{a}' - \bar{b}' := 1/N \sum_{i_n \in \mathscr{A}} a'_{i_n} - 1/N \sum_{j_m \in \mathscr{B}} b'_{j_m}$$

will be evaluated as a test statistic according to the Patchwork detection principle. The quantity s can be regarded as the *detector's response* or *detection score*. Positive values of the score indicate a detected bit symbol m=”1”. Negative scores indicate a bit symbol m=”0”.

As an example, the detection score s for typical audio content (e.g. music, voice) is shown in Figure 2.12. For typical technical settings of the approach described in [SZ2008a], the distribution of $|s|$ is shown *prior* and *after* embedding a watermark and subsequent lossy *MP3* encoding. For unmarked content the detection score is close to zero. Significantly higher absolute values are enforced in the course of watermark embedding. As will be shown in the experimental results in Chapter 5, this approach provides an acceptable trade-off between robustness and transparency of the watermark.

2.3.5.3 Corollary – Analytical Interpretations of Patchwork Embedding

Patchwork watermarking in the state of the art of watermarking is one of the often-cited basic embedding principles. In the following it will be shown that it is equivalent to the fundamental principle of *correlation*-based detection or spread-spectrum watermarking.

Correlation-based Interpretation

The detector response/score can formally be expressed as a correlation of the cover data with a *suitable selection* pattern. For this, the notation with so-called *indicator functions* is used. The indicator function, denoted as "$\mathbf{1}_{\mathscr{S}}(o)$" is a common auxiliary function in stochastic for *indicating* if an object o is element of a set $\mathscr{S}$ or not[38]:

$$\mathbf{1}_{\mathscr{S}}(o) = \begin{cases} 1 & \forall o \in \mathscr{S} \\ 0 & \text{else} \end{cases} \quad . \tag{2.9}$$

In the context of Patchwork detection, the objects o are the frequency indices $bk \in \mathbb{N}$. For the following explanation, a *modified* indicator function $\mathbf{1}'_{\mathscr{S}_\infty,\mathscr{S}_\in}(o)$ for the frequency indices k is defined as:

$$\mathbf{1}'_{\mathscr{A},\mathscr{B}}(k) = \begin{cases} 1 & \forall k \in \mathscr{A} \\ -1 & \forall k \in \mathscr{B} \\ 0 & \text{else} \end{cases}$$

to indicate if the respective index is in $\mathscr{A}$ or $\mathscr{B}$. Then, the calculation of the *detection score* can be expressed as

$$s := 1/N \sum_{i_n \in \mathscr{A}} a_{i_n} - 1/N \sum_{j_m \in \mathscr{B}} b_{j_m} = 1/N \sum_{k \in \mathbb{N}} r_k \cdot 1'_{A,B}(k) \quad .$$

Regardless of the scaling factor $1/N$, this is formally the cross-correlation between the spectral coefficients (r_k) and the watermark "selection pattern" $\mathbf{1}'_{A,B}(k)$. The score s describes if the watermark is correlating with the union of all pattern for the message bit "0" or with the inverse pattern for message bit "1" or with neither of the two ("unmarked"). That means that the Patchwork approach can be seen as an example of the general correlation-based watermarking principle.

Spread-spectrum Interpretation

In addition, the approach described in the previous Section can also be seen as a kind of an spread-spectrum approach (see Section 2.3.4): The indices of the watermarked frequency indices are picked pseudo-randomly and do vary over time. The detector needs to know the pseudo-random sequence in order to retrieve the watermark, Thus, the transmission is performed similarly to so-called *frequency hopping* approaches as one standard implementation of spread-spectrum techniques. This also means that some watermarking techniques as listed in Section 2.3.4 are not strictly disjoint.

To summarize, the wording in the research community on this field sometimes distinguishes algorithm classes which are actually more or less equivalent, as was also discussed by *Hua* [HHS+2016]. As was shown Spread spectrum and Patchwork watermarking share common principles and show analytical equivalence. Distinct denotation is still common for historical reasons.

[38] For the avoidance of confusion: here, the symbol "**1**" is a common convention which denotes the *function name* like the letter "f" in "$f(x) = ...$". This function symbol shall not be confused with the numerical value 1.0

2.3.6 Summary, Discussion and Current Research Trends

Summarizing the state of the art in audio watermarking a number of observations is becoming apparent.

Technical Maturity

Audio watermarking can be considered as a *mature* technology nowadays. Here a suitable trade-off between watermarking robustness, transparency, and capacity is available for different scenarios. Even *commercial* solutions for copyright protection/anti-piracy activities and miscellaneous multimedia interaction services are available. This is also reflected by the observation that watermarking algorithms from academic research (and commercial products too) in recent years still follow the same fundamental principles like correlation-based approaches, QIM or phase coding as introduced in the 1990s. Published works in recent years differ in the technical detail of implementing. For example, most works currently vary in which spectral transforms and coding techniques are used and (very often) combined rather than in proposing new fundamental principles or presenting major breakthroughs.

Of course, watermarking still is a topic of academic research. The most important observations on technical and general aspects from this watermarking research are outlined in the following[39].

Cursory Transparency Evaluation

Note: the following remarks expect the reader to be familiar with objective sound quality assessment as explained in Section 2.4.2).

Other works show a lack of thorough experimental evaluation e.g. on the transparency. Examples of frequently cited works from recent years are the following:

- Works by *Khaldi* [KB2013] or *Neethu* [NK2016] conduct the evaluation on very few audio examples and/or very few human test persons as test basis.
- In a work by *Xinkai Wang* [WWZ+2013], the transparency was even evaluated on a *single* example audio file of 10 seconds length only.
- In a work by *Lei* [LST2013], the transparency was expressed in terms of (nevertheless promising) PSNR values instead of more significant assessments using the objective *ODG* scale (as explained in Section 2.4.2).

Although these works appear to be notable contributions, the lack of technical detail or experimental evidence make it difficult to understand if and how the major watermarking challenges are successfully solved by these works, unfortunately.

Commercial Exploitation

Current research has become conducted and driven by *commercial* suppliers of watermarking solutions. Beyond *academic* research, watermarking algorithms are being developed and provided *commercially* by suppliers of watermarking solutions like *Civolution B.V., Digimarc Corp., Fraunhofer Gesellschaft e. V., Microsoft Inc., Musictrace GmbH, Technicolor S.A., Verance Corp.* or *Verimatrix Inc.*, to name a few examples. Preliminary results are often treated as trade secrets instead of being made available to the public in scientific publications. Sometimes the *patents*

[39] Note that common features about the *security* aspects of watermarking publications were already discussed in Section 2.3.4.3

are more informative about new approaches than the scientific papers by the same engineers, if available at all.

Declining Research Activities

As was explained before, research on audio watermarking sparked at the beginning of the 2000s and reached its peak of published research publications around the year 2010 (see Figure 2.8 before). Since then the total attention for watermarking technology has declined in the academic community. One reason for the declining number of research publications may be the increased *commercial* exploitation of the technology.

Furthermore, watermarking research has become more intense in Far Eastern research communities, mainly in China but also in Korea and Japan to some extent. What has also become apparent in this context is a significant *separation* of the research communities in the "Western" world (i.e. Europe and the USA) from those in Far East countries. Especially Far Eastern conferences and journals present mainly works by Eastern authors who, in turn; almost only cite other Eastern authors in their references. A significant number of works were published in the local language and hence could not be analyzed in this thesis work.

To summarize, progress in watermarking has become more and more an engineering challenge instead of an academic research challenge over time. In addition, the techniques of *steganalysis* and *multimedia forensics* have come into focus in media security research.

Recent High-capacity Embedding Algorithms

Despite the declining intensity of research in recent years, some progress is being published especially with respect to embedding *capacity*: While commercial robust watermarking solutions offer an embedding capacity up to a few bits per seconds, some of the publications listed above introduce algorithms that permit embedding robustly up to few *hundred* bits per seconds as claimed by the authors [UM2011, WHT2011, LSZ$^+$2012, DS2013, HHC2014, NU2015, KA2015].

Especially the very recent works by *Hwai-Tsu Hu et al.* and by *Fallahpour et al.* appear to be outstanding progress on achieving a feasible compromise between high embedding capacitym transparency and robustness:

- *Fallahpour* proposes a high-capacity QIM approach in the Fourier domain [FM2012, FM2015]. Authors claim that an embedding rate of at least 600 bit/s can be achieved while, at the same time, the watermark is robust against moderate lossy compression (MP3 128 kbit/s) at a bit error rate smaller 0.0-0.1 (which can be coped with by modern error correction techniques to a large extent). A "good" sound quality (*ODG* range -1.0 to -0.3, see Table 2.5 for definition on *ODG*'s) can be achieved which might be applicable in a number of scenarios.
- The latter work even appears to be outperformed by a work by *Hu*. Authors claim that an embedding capacity of 300-600 kbit/s can be achieved. At the same time, the bit error rate is as low as 0.0002 in the presence of an MP3 attack at 128 kbit/s (0.03 for MP3 at 64 kbit/s, resp.) and a good transparency of at least *ODG*=-0.15 can be achieved.
- Another work by *Hu* proposes a QIM approach in the DCT domain [HH2015]. Authors claim that an embedding capacity of 80-500 kbit/s can be achieved. At the same time, the bit error rate is as low as 0.0003 in the presence of an MP3 attack at 128 kbit/s. The transparency remains "better" than *ODG*=-0.12 which makes the embedding scheme suitable for a number of applications, including authentication watermarking.

Kindly note: Unfortunately, most of these high capacity embedding schemes above were not available as inspiration during the main working period for this thesis work (2007-2012). In order to meet the increased capacity requirements for authentication watermarking, own research on this became advisable.

2.4 Perceptual Modeling for Audio

In the course of this research work making use of psycho-acoustic modeling became advisable. Hence this Section explains mathematical models for human audio perception. This facilitates the description of the proposed enhancements to the audio watermarking algorithm as proposed in this thesis and the sound quality metrics used in its experimental evaluation.

Note: The following explanation relies on the technical basics of audio data (like PCM coding, spectral transforms, sound pressure levels etc.) and the fundamental properties of psychoacoustics (like masking or critical bands) as recalled in Appendix A.1 and Appendix A.2 for the interested reader.

2.4.1 MPEG Psychoacoustic Modeling and Audio Coding

In the late 1980s, it became obvious that personal computers and multimedia entertainment devices and services will become more and more powerful and popular. However, storage space and network bandwidth were very limited at that time. For overcoming this, lossy compression algorithms for audio, video and image data were developed.

The most popular compression approaches for audio are the coding standards of *MPEG Audio*. For more than twenty years, they have been in use in the *MP3* and *MP4/AAC* formats for consumer music devices and services or in the *MP2* standard for soundtracks on Video DVDs [ISO1993b, ISO1993b, ISO2000]. Detailed explanations of the MPEG psychoacoustic model can be found in the comprehensible report by *Lanciani* [Lan1995] and, of course, in the *ISO MPEG* specifications. The latter [ISO1993b, Appendix D, pp. 5] describes it as follows:

1. **Calculation of the audio spectrum:** The digital PCM input signal is divided in frames of $L = 1024$ samples. To reduce spectral leakage effects the PCM signal is multiplied sample by sample with a sampled *Hanning* window function. Then, the Fast Fourier Transform (FFT) is calculated from the PCM signal. The linear FFT absolute values are converted on the logarithmic *dB*-scale obtaining the *sound pressure levels* denoted as X_k.

   Note: As a convention, frequency indices will be denoted as i, j or k with $i, j, k \in \{1, ..., 512\} \subset \mathbb{N}$.

2. **Identifying the tone-like components**: Then *tonal maskers* are searched by identifying local maxima in the instantaneous audio spectrum. Only sufficiently dominant peaks are ”picked” (7 *dB* higher sound pressure than its neighbors).

3. **Identifying the noise-like components**: The *noise maskers* are calculated by summing up the sound pressure levels of the *remaining* spectral components among the same critical band. Distinguishing ”tonals” from noise maskers is important because human perception of sounds is different for the two cases.

4. **Decimating the tone-like and noise-like masking components**: Only the most relevant components shall be used for estimation of the global masking threshold. For this, all tone-like and noise-like maskers below the absolute hearing threshold are discarded at first. Then, the critical band characteristic is made use of again: in every critical band, all tone and noise maskers except for the one with *highest* energy are discarded too.

   Note: As a convention, critical band indices are counted by the quantity $z(k)$ in the following: it denotes the critical band value (in *Bark*) that corresponds to the (integer) FFT frequency index k.

5. **Calculation of the individual masking thresholds**: Note that, as a convention, the index j denotes the *fixed* frequency index of one of the maskers while i denotes the independent, continuous index for modeling the shape of the masking curve. Then, the exact shape of the masking curve of each individual masker is denoted as $LT(i,j)$ and calculated as follows [40]:

$$LT_{\text{tonal}}(i,j) = X_j + av_{\text{tonal}}(j) + vf(i,j) \quad ,$$

$$LT_{\text{noise}}(i,j) = X_j + av_{\text{noise}}(j) + vf(i,j) \quad .$$

These three terms have the following meaning:

- *Masker, X*: The quantity X_j is the sound pressure level of the tonal or noise masker at fixed frequency index j.
- *Masking index, av*: In the MPEG specification the quantity av is defined as a slightly declining linear function in $z(j)$ of the form

$$av(j) = -p\ z(j) - q \quad .$$

Detail on the values of p, q can be found in [ISO1993b, Appendix D, pp. 5] (values omitted for simplicity). The masking index essentially means some sort of "fine-tuning" by adding a little skewness to the curve. It is of little relevance for this thesis work and it is mentioned only briefly for completeness.

- *Masking function, vf*: The quantity vf dominantly defines the shape of the masking curves. It is expressed as being dependent on the quantity $z'(i,j) := z(j) - z(i) \in \mathbb{R}\,[in\ Bark]$ which is the distance of the independent, continuous index i from the masker at index j measured in *Bark*. Then the masking function is modeled as a piece-wise linear function in $z'(i,j)$, i.e. relatively to $z(j)$, of the respective masker as

$$vf(i,j) = \begin{cases} 17\,(z'(i,j)+1) - (0.4\,X_j + 6) & \text{for } -3 <= z'(i,j) < -1 \\ (0.4\,X_j + 6)\,z'(i,j) & \text{for } -1 <= z'(i,j) < 0 \\ -17\,z'(i,j) & \text{for } 0 <= z'(i,j) < 1 \\ -(z'(i,j)-1)\,(17 - 0.15\,X_j) - 17 & \text{for } 1 <= z'(i,j) < 8 \end{cases} \tag{2.10}$$

Roughly spoken, the shape of the plots of the partial functions is a steep peak with "shoulders" in the domain of the critical band scale (as can be seen from the plot in Figure 2.13 from the numerical example below). Here, the partial function definition assures that discontinuities do not occur. As can be seen from this partial definition, the influence of each masker can be up to 8 *Bark* towards higher frequencies.

All values of the "magic numbers" in Equation (2.10) were obtained by the *MPEG* consortium from elaborate listening tests with human test subjects in the 1990s.

6. **Determination of the global masking threshold**: All maskers are combined by summing up the content-dependent masking thresholds from the previous step and adding the static threshold in quiet LT_{silence}. The adding is done according to the *Decibel* scale:

$$LT_{\text{global}}(i) = 10\lg(LT_{\text{tonal}}(i)/10) + 10\lg(LT_{\text{noise}}(i)/10) + 10\lg(LT_{\text{silence}}(i)/10) \tag{2.11}$$

[40] The denotation of the original MPEG specification is used in the remainder of this Section

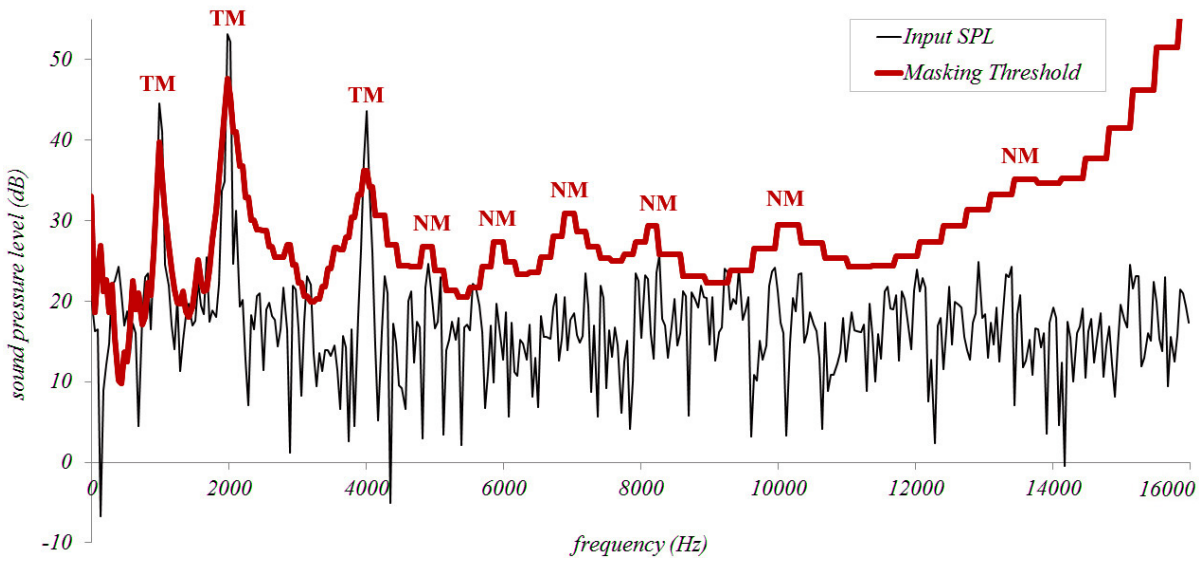


Figure 2.13.: Example: Psychoacoustic modeling; Thin line: input audio spectrum; Thick line: estimated masking curve; *"TM"*: tonal maskers; *"NM"*: noise maskers

In software implementations of the MPEG model, the quantity $LT_{\text{silence}}(i)$ is usually available as a static look-up table.

In this thesis work, this MPEG psychoacoustic model will be enhanced for controlling the embedding strength of the watermark distortion and improving the watermark detection.

Numerical Example

The psychoacoustic modeling as described before can be studied from the following descriptive example: an input audio signal was defined in which

- white noise at average sound pressure level of 50 *dB* is mixed with
- three sine-wave tones at 1 *kHz*, 2 *kHz* and 4 *kHz* at 50 or 60 *dB*.

Figure 2.13 shows the estimated masking curve obtained from the psychoacoustic model described above:

- On the left side, the three slopes of maskers can clearly be noticed at 1 *kHz*, 2 *kHz* and 4 *kHz*. These maskers are obviously caused by the sine components in the input signal, i.e. they are tonal maskers (denoted as "*TM*").
- Six local maxima at frequencies beyond 4 *kHz* in the masking curve can be explained by the noise maskers identified ("*NM*").
- The steep slope of the masking curve at high frequencies (beyond 12 *kHz*) is dominated by the shape of the absolute hearing threshold in silence.

2.4.2 Perceptive Audio Quality Measures (PEAQ)

For sound quality comparison of audio content, researchers and engineers use automated perceptual analysis systems. It helps reducing efforts on elaborate (and costly) listening tests with human test subjects and it provides reproducible results. Such assessment tools are commonly used in evaluation of audio processing software or hardware equipment – so it is in this thesis too: in the experimental evaluation such tools will allow

- the transparency evaluation of the proposed watermarking approach,
- measuring the effect of a breach of integrity with regards to the human perception.

For hi-fi audio content like music or audio books an accepted standard for quality assessment is the *perceptive audio quality (PEAQ)* model by the *International Telecommunication Union (ITU)*, see [ITU1998, TTB+1998]. It expresses sound quality differences between the input *reference file* and the correspondent input *test file* in terms of so-called *objective difference grades (ODG)*. The *ODG* scale is defined from -4.0 to 0.0 (see Table 2.5) and, in simple words, it expresses how much "worse" a test signal sounds to an average listener than the respective reference signal.

Example: According to the test results in Chapter 5, *MP3* compression at declining bitrates of 256, 160 and 128 *kbit/s* corresponds to declining *ODG* values of -0.05, -0.60, and -1.20, resp. (median value across various music, audio book and voice reference stereo files).

This thesis work uses the commercial *OPERA* assessment tool[41] which implements the PEAQ model. Other, free, implementations were developed in the *PEAQB* project[42], the *PQevalAudio* project[43] and the *EAQUAL* project[44].

ODG	Perceived difference
0	*"inaudible"*
-1	*"audible but not annoying"*
-2	*"slightly annoying"*
-3	*"annoying"*
-4	*"very annoying"*

Table 2.5.: Definition of objective difference grades (*ODG*) describing sound quality difference between reference and test signal

For the remainder of this work such tools are regarded as "black boxes" that objectively measure the influence of watermark embedding or arbitrary attacks on watermarked audio on the sound quality. They are used in the empirical evaluation of this thesis work.

[41] *OPERA* Audio Quality Analysis software by *OPTICOM GmbH*, Erlangen, Germany, *http://www.opticom.de*

[42] *PEAQB* project by *G. Gottardi*: *http://sourceforge.net/projects/peaqb* (2003)

[43] *PQevalAudio* project by *P. Kabal*, TSP Lab, McGill University, Montreal, Canada: *http://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp* (2010)

[44] *EAQUAL* toolbox by *Alexander Lerch*, *Heinrich-Hertz-Institut*, Berlin, Germany (2002). Note: Project discontinued, no official sources available

2.5 Related Work in Authentication Watermarking

The following overview presents and discusses the Related Work in the field of authentication watermarking. It motivates the technical approach that is investigated in this thesis work by listing valuable sources of inspiration.

2.5.1 Categorization of Approaches

The existent mechanisms can be characterized according to the following criteria:

- **Content of the Embedded Watermark Message:** Different kinds of embedded authentication codes can be observed in audio authentication watermarking literature, namely:
  - Perceptual hashes derived from quantities that are closely related to human perception like loudness, masking etc.,
  - content dependent meta data which are derived from otherwise signal depending quantities,
  - time codes,
  - static watermark messages which are time and content *independent*,

  and concatenations thereof. Of course, only the first two categories can be regarded as truly content-fragile approaches as defined above. The latter two are nevertheless worth discussing because the majority of the related works on authentication watermarking belong to these categories.
- **Robustness of the Embedded Authentication Code:** The related work covers all kinds of integrity watermarking approaches as explained in the introduction of watermarking (see Section 2.3.3). Hence, all levels of robustness of the embedded messages itself can be observed, i.e. fragile, semi-fragile, and robust embedding.

2.5.2 Publications on Authentication Watermarking Algorithms

In the following overview, the respective headings reflect briefly

- the technical basis for deriving the authentication code, and
- the technical approach for consecutive embedding.

Note that not only works on content-fragile watermarking but also other authentication watermarking approaches for audio are listed. This allows placing the proposed approach(es) in the narrow field of watermarking-based integrity protection research.

Time Domain Volume Feature & LSB Watermarking (by *Fan Chen* / *Cvejic and Seppänen*)

A fragile audio watermarking approach was described by *Fan Chen* and *He* [CHW2008]. At first, the input data is expected as 16 *bit* PCM samples i.e. as *short integers*. For the following description, the very least significant bit (LSB) of these short integers is denoted with bit index ”0” while the very most significant bit then has index ”15”. The PCM samples are first grouped in so-called *”segments”* of four samples each. As content-describing feature of each segment, the sum over the squares of those four PCM sample values is calculated. Here, bit indices 0 to 3 are ”ignored” during feature extraction by temporarily setting these bits to zero. Finally, this 16 *bit* audio volume descriptor is embedded by replacing the 4·4=16 LSBs with the 16 *bit* feature.

This embedding and feature extraction approach is rather simple and can easily be circumvented because no key-dependent feature extraction or watermark embedding is implemented. The embedded message can easily be forged and replaced. The publication is interesting for some reasons, though:

- It proposes a *linked chain* of feature extraction and feature embedding: the audio features of one segment are embedded into the following segment, and so on.
- Even more, before linking as described before, the sequence of the segments is permuted dependent on a secret key.
- By design it offers a very high temporal resolution of tampering (if successful) of 4 samples duration, i.e. less than 1/10,000 second.

Because of the permutation it is obscure to an adversary in the first place which audio segments are linked: for example, without access to the detector software as an oracle, the adversary cannot reconstruct which audio descriptor is kept in which frame. Linking of frames is one way to achieve the security with respect to security attacks by removing or flipping of segments as proposed in the *Holiman Attack* [HM2000].

Time Code Pattern & Patchwork Embedding (by *Park et al.*)

Park, Thapa and *Gi-Nam Wang* describe an approach that embeds a simplified time code by means of spread spectrum watermarking in the Fourier domain [PTW2007]. A pseudo-random noise sequence (which depends on the secret watermark key) is added to the DFT spectrum. The algorithm allows embedding one out of five different message symbols from the ”alphabet” from *”0”* to *”4”*. This is carried out by using the same noise sequence with five different settings of cyclical time shift/offset. The authors intend to identify tampering of audio data by detecting offsets or missing parts in the cyclic *time code pattern*. The temporal resolution of the time code is 0.6 *sec.* It is not discussed if the approach is secure against *Holiman* Attacks by deliberate replacement of an audio section that contains the same time code state or a complete time code sequence. The interested reader of the original work should note that the terminology in the title of the publication is somewhat misleading: the term *”pattern recovery”* shall not be confused with *pattern recognition / matching* techniques from machine learning or computer vision.

Time and Spectral Domain *”Feature Checksum”* & Patchwork Embedding (by *Steinebach* and *Dittmann*)

In a work by *Steinebach* [SD2003b, pp. 1005], an approach based on the extraction of checksums derived from the FFT-power spectrum, the root-mean-square (RMS), and the zero-crossing-rate (ZCR) from each protected audio frame was presented for general audio data (music, speech etc.).

- The DFT magnitude spectrum is evaluated in different sub bands in the low to mid frequencies e.g. 0.5-4.0 or 2.0-6.0 *kHz*. Authors argue that this covers most of the relevant sub band to protect integrity of voice and music data. Higher frequencies are discarded.
- The RMS is basically a numerical feature derived from the sum of the squared PCM sample values and corresponds to the loudness/volume of the audio.
- The ZCR counts the number of sign changes of the PCM signal in a given frame. This correlates with the amount of high frequency components contained in the audio because high frequency signal obviously show more sign changes per time unit than low frequency signals.

In colloquial words, RMS and ZCR describe in a simplified manner how ”loud” and how ”bright” an audio section appears. From these features, a so-called *feature checksum* is derived by first (coarsely) quantizing the ZCR, RMS, and DFT values and then taking the quantized features as input to a crypto hash (like SHA-1 or MD5), or to an CRC or to simple XOR checksums.

The feature checksum is finally embedded as a robust watermark message using the Patchwork embedding technique in the Fourier domain as described in Section 2.3.5.2. An actual fingerprint data rate (i.e. embedding rate) of approximately 10 *bit/s* is achieved which reflects the capacity limitation of audio watermarking at that time. An interference of the feature checksum with its embedding as a watermark message is avoided by using different frequency ranges for extracting the feature checksum and the consecutive embedding.

Note that no apparent key-dependent feature selection was described by the authors which limits the security of the approach with regards to collision or second pre-image attacks: The length of the feature checksum can be set as small as 4 *bit* which causes a false negative rate of at least $2^{-4} \approx 0.06$ and allows birthday attack extremely easily.

Note also that SHA-1 and MD5 algorithms utilized in the approach by themselves were found to have certain vulnerabilities against collision attacks in the meanwhile (see *Xiaoyun Wang* or *Hoffman, Schneier* [WYY2005, HS2005, SBK+2017] for an overview on *”attacks on cryptographic hashes in Internet protocols”*).

Masking Threshold Feature & Spread Spectrum (by *Radhakrishnan et al.*)

In an inspiring publication by *Radhakrishnan* and *Memon* [RM2002], a robust hashing approach based on psycho acoustic modeling from MPEG-1 audio coding (see Section 2.4) is used. It uses the shape of the instantaneous masking threshold as a robust feature. It also reflects the *dynamics* of the audio signal in terms of its respective masking characteristics over time as explained in Appendix A.2.5.

The work also applies an interesting randomization approach by *Fridrich* and *Goljan* [FG2000] to audio hashing in order to reduce the hash length, i.e. the required watermarking capacity: At first, all 43 measurements of masking curves (in frequency resolution of 32 sub bands) in one second are interpreted as one measurement in a $32 \cdot 43$-dimensional space $\mathcal{V}$. Then, a set of 32 pseudo-random vectors in $\mathcal{V}$ is defined dependent on a secret key. Finally, the measurement is projected onto this pseudo-random basis of the 32-dimensional subspace. The *signs* of the 32 projections are used as the hash bits. Here, the projection is an elegant approach of mapping the *continuous* masking curve feature to a *binary* feature. The extracted audio hash is eventually embedded as a robust watermark using a spread spectrum embedding algorithm in the spectral domain developed earlier by *Li et al.* [LY2000].

Note that the audio hash is 32 *bit* long which means that it is very vulnerable to *birthday attacks*. The approach is interesting though as it is one of the few that utilizes psychoacoustic modeling. The (very few) empirical test results show that the audio hash allows a significant assessment of the degree of modification of the audio. For example, drastic modifications like cropping can be separated from minor sound modifications like moderate *MP3* compression.

Loudness Feature & QIM Embedding in hybrid DCT/Wavelet Domain (by *Hong-Xia Wang, Ming-Quan Fan et al.*)

In a work by *Hong-Xia Wang* and *Ming-Quan Fan*, the centroid (i.e. center of mass) among the loudness of a sequence of short audio sections is used as a major authentication feature [WF2010]. It is utilized for deriving a robust authentication feature for determining

watermark embedding positions in the first place. The following explanation is rather simplified and shall outline the overall idea(s) of the algorithm and the main contribution.

At first, a longer section of PCM audio input data of L PCM samples is processed in a block-based manner. In the experimental section of the publication, such section represents for example approximately 95 s of duration. It is divided in M so-called *frames* of shorter length L_1 (e.g. approx. 100 ms) without overlapping. Each frame is further split into N much shorter sub-frames[45] of length $L_2 = L_1/N$ (e.g. approx. 3 $msec$). Then, the n-th sub frame's sample data $x_t^{(n)}$ $(t = 1....L_2)$ is transformed to the Fourier domain $(x_t^{(n)}) \longrightarrow (X_k^{(n)})$. The DFT representation of the n-th sub frame allows deriving some sort of weighted "sub frame loudness" measure $D_{m,n}$. for the m-th sub-frame inf the n-th frame. The mathematical detail are skipped for simplicity.

This allows calculating the time index C_m of the particular sub frame that represents the *centroid* with respect to sub frame loudness D. Subsequently, the embedded watermark messages are constructed by post-processing the sequence of C_m by means of crypto hashing, concatenation and encryption by an XOR operation with a key-dependent pseudo-random sequence (the detail can be found in the original publication). For skill full selection of L, L_1, L_2, the centroids can be regarded as robust to many post-processing operations. Conversely, tampering is expected to be detectable as insertion or deletions of audio will likely change the time dependent characteristics of the total loudness (in terms of the D measure defined above) and, hence, the positions of the centroids.

The message embedding is eventually done in the sub frames by means of QIM embedding in the hybrid DCT and Wavelet domain. Interestingly, only sub-frames that represent the frame centroid of the respective frame are used for embedding. This is motivated by the fact that significant tampering will significantly change the positions of the centroids. This will cause the watermark detector to fail which is a strong indicator for tampering.

In the experimental chapter of the publication, a watermark message of only 4 *bit* only is constructed from the centroid feature(s) and eventually embedded as authentication code. It is quite doubtful that this is secure with respect to deliberate attacks.

Fourier Spectrum Feature & Infrasound Embedding (by *Ching-Te Wang* and *Ming-Quan Fan*

A different Fourier-based approach domain was presented by *Ching-Te Wang* and *Ming-Quan Fan* [WLC2007]. Unfortunately the relevance of this work cannot be assessed completely because the technical description in that work is not fully precise: authors define the extracted authentication feature as *"the maximal frequency"* in the spectrum obtained from audio sections of five sections. In contrast, according to a secondary source by *Gulbis* [Gul2013] *"the frequency with highest amplitude"* is said to be actually intended (which in the thesis author's opinion appears more feasible in light of other works). No matter which of the "interpretations" is true, any of these mean a very imprecise simplification of the audio content. It is doubtful if minor modifications to the protected audio become detectable. Test results with regards to *sensitivity* to malicious attacks were not given in that work.

Another issue is the following: According to the original authors the embedding of the (RSA-encrypted) audio feature shall be carried out in the very low frequency range of 1-20 Hz. For this, the original spectrum shall be erased first and then be deliberately replaced by sinusoids in that frequency range which represent the watermark. The authors claim that this is a transparent operation because the human ear is insensitive to sound in this *infrasound* frequency range

[45] The interested reader who intends to reproduce the calculations in the original publication should note that the sub-frames are denoted as *"sub-bands"* by the authors. This is a rather misleading denotation as the term *"bands"* commonly reminds of a *spectral* data representation, not a temporal/spatial one.

anyway – which is correct (as explained in Appendix A.2.4). But following the authors' argument, the watermark can easily be destroyed without quality loss by erasing the low frequency range *again*. Hence the approach cannot be regarded as reliable or secure to targeted attacks on the watermark.

On the other hand, the approach is notable at least *in principle* as it proposes to use an *asymmetric* encryption of the audio authentication feature. The proposed usage of RSA encryption allows for the *public* verification of the embedded (but actually *insecure*) authentication code at a later date.

HMM Modeling of MFCCs & Spread Spectrum Embedding (by *Gomez, Cano et al.*)

Gomez, Cano et al. present an extraction of an alphabet of representative sounds that are present in the audio [GCdG+2002]. The alphabet of so-called *"audio descriptor units (ADU)"* is modeled by means of MFCCs and hidden Markov modeling. That is, the approach is analogous to common automated speech recognition techniques (ASR) in which the *phonemes* are identified as basic units. Here, the presented fingerprint approach covers *general* sound data like music, noises and voice, as well. The consecutive embedding of the ADU feature is only vaguely described. Apparently this is done by a spread spectrum approach in a spectral domain which was not specified in detail by the authors.

In the test setup, the representative sounds are classified into 16 different ADUs represented as 4 *bit* integers. The output rate is approximately 100 ADUs per minute. Thus, the total fingerprint bitrate is $4 \cdot 100/60 \approx 7$ *bit*/s which is compliant with the capacity limitations of robust audio watermarking at that time. Experimental results (based on little test material) show that the approach is capable of indicating malicious inserting and deletion attacks on the audio.

Although an interesting approach, also this approach lacks an analysis of the security to targeted attacks. For example, it is doubtful how fingerprint collisions can be avoided as each ADU is only 4 *bit* long, i.e. the "alphabet" has only 16 different possible states (similar to the approach by *Hong-Xia Wang* explained in this overview [WF2010]).

Time Code Feature & Spread-Spectrum Embedding (by *Petrovic et al.*)

Petrovic proposes using spread-spectrum embedding in the time domain [Pet2005]. An algorithm from an earlier (patent) publication [LMW1998] is used for embedding a fixed *"Embedder ID"* (20 *bit*) and a dynamic time-code (20 *bit*) into the audio data every eight seconds. On the detector side, inconsistent time-code data or low detection response (correlation coefficient) are used for indicating the tampering with the audio data. Embedding such time code robustly as authentication meta data is a very promising approach for localizing tampering and avoiding targeted attacks. Time codes are an alternative to key-dependent permutation of protected audio section as proposed by *Fan Chen*, see above [CHW2008].

DCT and Critical Band Features & Patchwork Embedding (by *Gulbis et al.*)

The research work by *Gulbis* focuses on content-fragile watermarking in the spectral domain [GMS2008a, Gul2013]. A robust feature is extracted from the DCT transform: at first, the DCT spectrum is non-uniformly mapped on the critical band scale (see Appendix A.2.6) to reflect the human auditory perception. Then, a binary feature is extracted by analyzing the sign of the difference between DCT coefficients in consecutive time-steps. This feature is used as a robust hash and finally embedded by means of Patchwork watermarking.

It is shown that the proposed audio hash can well distinguish modifications causing little sound quality degradation from stronger modifications. Compared to other peer-reviewed publications or theses on this topic, the set of test data (more than approximately two hours) is rather large. Hence the empirical evaluation provides rather significant results.

Static Semi-Fragile Message Approaches (various authors)

All approaches explained in this paragraph are typical semi-fragile approaches. Hence, the embedded watermark message is static and not content-dependent. Nevertheless, because of the significant number of publications (even in recent years) they are listed here for completeness.

An approach for tamper detection was presented by *Ning Chen* which embeds both, a robust copyright watermark and a fragile integrity watermark [CZ2008, CZL2010]. The embedding is done in the domain of detail Wavelet coefficients (*Haar* wavelet basis). A pseudo-random *"chaotic sequence"* is used as a fixed watermark message. Confusingly, the authors denote the embedded watermark message as a *"key"* that is later used for identifying tampering. The message shall not be confused with watermark or encryption keys. The security of the scheme is doubtful because no dependency from a watermarking key (as defined in Section 2.3.1) is implemented in the algorithm. That is, it is likely that tampering can be obscured by creating a seemingly correct watermark message again and overwriting the pre-existent fragile watermark with the forged watermark.

Zhao and *Shen* propose similar QIM approaches in the Wavelet domain. Here, a fixed bit pattern, namely a binary image of a rose is used the watermark message. Embedding is done either on the coarse Wavelet coefficients directly [ZS2009a] or on the DCT coefficients of the coarse Wavelet coefficients [ZS2010]. Again, it is not clear in this publication how the fixed watermark message is protected against unauthorized write access by an adversary.

Another approach for the Wavelet domain was presented by *Cvejic, Seppänen* [CS2004]. From every frame of 512 PCM samples, the so-called *Haar* wavelet spectrum of 512 samples is calculated. A larger subset of 384 samples is used for embedding a fragile LSB authentication watermark. Unfortunately, it is not fully clear from the technical description if the authentication watermark is content-dependent or a fixed message.

A similar work by *Ma* uses, again, a fixed watermark message. Before embedding in the Wavelet domain, the message which is encrypted by XOR composition with a pseudo-random sequence [MLL2008].

Unoki proposes [UM2012] to adapt an own earlier copyright watermarking approach [UM2011]. The approach is an interesting echo hiding algorithm in extension to the explanation in Section 2.3.4. It exploits certain *"cochlear delay characteristics"* of the human perception that allow to use different echo delays at different frequencies. Similar to the approach by *Zhao* mentioned before, a fixed watermark message sequence is embedded for data authentication.

Ming-Quan Fan proposed an approach [FLWL2013] based on an earlier work (which was explained above [WF2010]). It uses the hybrid DCT/Wavelet transform for embedding a semi-fragile watermark in the low-frequency sub-band.

Dhavale proposes a semi-fragile watermarking scheme in the concatenated Wavelet-DCT domain [DDPP2013]. For integrity verification, a fixed message is embedded, namely an alternating bit series of 1024 bit length. Unfortunately, it is not clear how the approach is made secure against replacement attacks as no true watermark key is introduced. Interestingly, the authors propose using "synchronization codes" along with the watermarking payload for overcoming

de-synchronization attacks. That is, this approach is to some extent similar to the contribution in this thesis work. The approach appears to be well capable for detecting significant amounts of cropping caused by deletion of audio content.

In a recent work by *Jinquan Zhang* [Zha2015], a pseudo-random sequence is embedded as a semi-fragile watermark in the DCT domain by means of QIM embedding. It allows for temporal localization of malicious tampering and appears to be invariant under admissible signal transformations like resampling, re-quantization, adding moderate noise etc. Unfortunately, lossy compression was not tested and the effectiveness was evaluated on the basis of only *three* music file examples.

Finally, the work by *Haiyan Liu* [LZ2016] does not introduce new audio authentication algorithms by itself. But it reviews a number of existent QIM approaches from the state of art in (seim-) fragile watermarking. The focus is exclusively on journals and proceedings from the Chinese and Japanese research community from the years 2008-2015.

LSB feature in GSM Data & Bitstream Embedding (by *Yuan* and *Huss*)

A different approach for speech data by *Yuan et al.* is based on GSM 610 codecs for the protection of voice-over-IP transmission [YH2004]. It uses the least-significant bit (LSB) of the sum over so-called *log area ratio (LAR)* coefficients as content-based feature. The LAR models the reflection coefficients of the shape of the vocal tract. One LSB per 20 milliseconds is extracted and embedded into that frame as a bitstream watermark (see Section 2.3.4 for bitstream watermarking) in the GSM stream.

Note that the authentication code is defined on GSM coded data only. In the authors opinion, it is very unlikely that the independent extraction of a sub-fingerprint of only 1 *bit* length per frame can provide security against targeted collision attacks. It is also unclear if LAR coefficients as principal speech features are secure in general: the authors themselves state that LARs are very commonly-used in *speaker recognition* techniques. That is, LARs are not only specific about the particular speech content (phonemes, words etc) but also about the speaker's voice – no matter *what* he/she is saying. It should be further investigated if the latter property increases the vulnerability to collision attacks even more.

G.723.1 Speech Codec Features & Header Attaching (by *Wu et al.*)

Different approaches for audio data in *lossy* compression formats for speech data are presented, as well. An publication by *Chung-Ping Wu et al.* is based on the extraction of speech coding coefficients [WK2001]. Here, the G.723.1 speech codec from VoIP communication is used to derive a representative audio descriptor from modeling parameters of the shape of the vocal tract are extracted. The volume and the pitch of the speech signal are also analyzed. Finally, one 27 *bit* fingerprint is extracted per frame of 30 milliseconds. That is, the fingerprint bitrate is 900 *bit/s* (which is approximately 14% of the speech signal's payload).

Note that the authentication code is defined on G.723.1 coded data only. Even more, the fingerprint is not embedded as a truly fragile watermark message but as meta data *"attached at the header information"*, instead. The attached message will not have any robustness to any format conversion, obviously. Hence, this work is beyond the scope of the research of this thesis and is only mentioned for completeness.

Fragile QIM Watermarking for Audio-Visual Content (by *Rigoni*)

In a recent work by *Rigoni et al.* a watermark is embedded by means of QIM embedding [RFF2016]. Although not stated explicitly by the authors, the embedding appears to be applied on the PCM signal in the time domain. Robustness can not be expected by this approach (and is not evaluated by the authors, anyway). The approach is interesting, though, as it discusses protection of audio-visual content featuring a video part and an audio part. As a general proposal, using both media types can increase the redundancy of the approach.

Kindly note that the opportunities of joint audio-video authentication will be further discussed in the outlook in Section 6.4.4.

Reversible Embedding of Crypto Hashes in the DCT Domain (by *Huang*)

A family of approaches of fragile audio authentication algorithms was presented by *Huang, Echizen et al.* [HEN2010, HEN2011, HOEN2014]. An MD5 or SHA-1 hash (128 bit) serves as authentication code for the integrity verification. It is embedded in the DCT domain by means of LSB embedding. Interestingly, the authors propose a simple but effective method for providing a *reversible* scheme. Hence the embedding process can be fully/lossless inverted on the detector side. This avoids false positives on the very sensitive crypto hashes on the detector side which might be caused by the embedding process.

It should be noted that the hashing functions in the intermediate processing step were found to be insecure in the meanwhile [WYY2005, SSA+2008, SBK+2017].

Image Authentication with Localization (by *Liu)*

A notable source of inspiration can be seen in a work on authentication watermarking by *Huajian Liu* [LS2006a, LS2006b]. Although being an algorithm for the *image* domain, it is mentioned because its general approach of identifying potentially doctored locations is independent of the protected media type.

It is a semi-fragile QIM watermark algorithm in the Wavelet domain. The watermark message is a pseudo random sequence as a (content-independent) authentication code. Due to an intermediate permutation step, each of these watermark message bits is embedded into Wavelet coefficients from image areas that are spatially *scattered* all over the image. The message bits are virtually scattered all over the image. In reverse, a particular fixed area in an image can contain Wavelet coefficients that contribute to a number of different watermark message bits. Hence, if such area was subject to integrity breaches, several message bits would be affected.

Figure 2.14.: Example: Image authentication watermarking algorithm by *Liu*: Upper left: original image, upper right: doctored image (circle marker: object removed), lower left: raw authentication result (speckled white markers: detected "candidates" of doctoring), lower right: filtered authentication result
Source: [Liu2008, pp. 62-70]

The (spatial) localization is carried out in a two-step process. On the detector side, the identification of a localized doctoring can only be carried out on the basis of these "flagging" message bits. Because of their scattered nature, many sub areas are in question about where the doctoring took place. This is visualized in Figure 2.14: in the lower left figure, a large number of areas are marked as "candidates" for being unauthentic. But only in a certain area in the middle of the image, these markers *accumulate* to a notable area. This idea will be applied to audio data investigated in this thesis work.

2.5.3 Summary and Discussion

Note that the overview given so far is not an *exhaustive* collection of audio authentication watermarking, perceptual hashing and general audio watermarking. Instead it shall give an overview on publications that are frequenty cited and are representative with respect to the technical aspects. Note that some more sources from the Related Work could not be included in the discussion due to a lack of clarity in the technical description provided by the authors for example in [YWZ2009, GBYK2013].

An overview on the technical properties of the existent Related Work is given in Table 2.6. From this overview a number of interesting observations become apparent. For example, most approaches belong either to the class of content-fragile watermarking (upper right corner in Table 2.6) or to semi-fragile watermarking with static watermark messages (lower middle column, resp.). Both classes of algorithms are still being investigated to some extent. Fully fragile audio watermarking has remained of only of little interest in the research community due to its limited advantages over crypto-based solutions.

Scientific works on semi-fragile watermarking (i.e. using static watermark messages) has even outnumbered the works on content-fragile watermarking (i.e. content-dependent authentication codes) over the years with respect to the number of publications and of active authors/research teams. The resources in research on authentication watermarking for *audio* has always lagged behind the research on copyright and transaction watermarking and supporting mechanisms. It has even slightly declined presumably when the technique audio forensics became more and more into research focus in the field of audio authentication.

	fragile embedding	semi-fragile embedding	robust embedding
perception-based message	[CHW2008]	[WK2001]	[RM2002] [SD2003b] [WF2010] [YZLX2013] [GMS2008a] [Gul2013]
signal-dependent message	[HEN2010] [HEN2011] [HOEN2014]	-	[GCdG+2002] [WLC2007]
time code message	-	-	[Pet2005] [PTW2007]
static message	[YH2004] [RFF2016]	[CS2004] ([LS2006b]) [MLL2008] [ZS2009a] [ZS2010] [CZL2010] [UM2012] [FLWL2013] [DDPP2013] [Zha2015]	

Table 2.6.: Categorization of authentication watermarking approaches in the literature

What becomes apparent from the works on content-fragile watermarking is that some important *security* aspects of the approaches are not analyzed explicitly.

- A lot of the works use a rather short audio authentication of 4 to 32 *bit* only [RM2002, SD2003b, GMS2008a, WF2010]. It is very unclear how severe the challenge of collusion attacks or 2nd pre-image attacks on the authentication code is prevalent.
- In addition, most approaches listed above do not meet the security requirements on the feature extraction in the course of audio hashing as defined by *Fridrich et al.* or discussed by *Nouri* [FG2000] [NZAF2012]. Here no key-dependent selection of feature subsets is implemented during the modeling of the output audio hash values.
- The same is true for the watermarking algorithm used for embedding the authentication codes into the cover data. As already explained in this thesis above, this fact reflects that watermarking security is often not thoroughly investigated either.

Exceptions to this are embedding time-codes into each protected audio segment as proposed by *Park* [PTW2007] or *Petrovic* [Pet2005] or the key-dependent permutation and linking of

protected audio segments as presented by *Fan Chen* [CHW2008]. This is an effective measure to overcome attacks aiming at *"undetected modification"* as defined in Section 2.3.4.3 above or as discussed by *Holliman* and *Memon* [HM2000]. Such time indexing can and will also be implemented in a simple but effective in this thesis work.

What can also be regarded as missing in a lot of works is a thorough / exhaustive analysis using a larger test data set. Many works rather *demonstrate* than evaluate the notable performance on a few audio examples of a few minutes duration only.

Most related works *do* investigate if admissible modifications can be separated from malicious tampering. But most works do not explicitly investigate if such modifications are conducted *in sequence*. With regards to a potentially long-lasting life-cycle, it can be expected that the (rare) case of tampering will be conducted *after* the watermark-protected audio media was legitimately trans-coded to a new lossy coded format in the meanwhile. Especially in long-term archiving this could become relevant. The empirical investigation in this thesis will reflect this.

In extension to this, from the research on human hearing and lossy audio coding (as explained in Section A.2 on psychoacoustics) it is well investigated which components of an instantaneous audio spectrum are contributing to human perception by a listener and which are not. Interesting works on this were already published by *Mıçak* or *Radhakrishnan* analyzing the masking threshold [MV2001, RM2002]. Hence, it is feasible to investigate if such background can be utilized in the audio hash extraction.

2.6 Conclusion on Need for Further Research

As explained in the literature overview in this Chapter, the state of the art in audio watermarking, perceptual hashing and their respective combination for audio authentication provides valuable sources of inspiration for developing or improving technical approaches. Based on the summary on the Related Work and the available hashing and watermarking technologies it is concluded:

- In this thesis work, perceptual hashing approaches using the works by *Kalker et al.* [HOK2001a, HOK2001b] and successors thereof are developed using audio fingerprinting algorithms as starting point which are known to have good robustness and distinction performance
- Psychoacoustic modeling might be included in order to improve the robustness of the audio hashes and for providing a truly perception-based identification of audio data identity.
- In order to improve the security against targeted attacks a key-dependent feature extraction processing step shall be implemented in the audio hashing. A sufficiently large effective key length shall be realized (desirably at least 64 *bit*).
- As an additional metadata for verifying the integrity, time-odes or the like shall be used.
- The Patchwork audio watermarking core algorithm by *Steinebach* [Ste2003, SZSL2003] shall be used as a starting point for embedding the audio hashes. Its good robustness and transparency has been proven in many commercial applications for many years. Earlier works on adapting it for integrity watermarking and its full availability to the thesis author mean a good starting point for this thesis research.
- Nevertheless, the embedding capacity of the involved Patchwork audio watermarking shall be increased to be able to embed authentication code and additional meta data.
- A theoretical analysis of the security of the involved Patchwork watermarking should be conducted.
- Empirical results should be obtained from experiments on a sufficiently large and representative test set of audio data.

An approach in light of the objectives and requirements listed in Section 1.2 and Section 3.1.2 that reflects this conclusion will be introduced in the following Chapter.

Chapter 3

Proposed Content-based Integrity Watermarking System

In this Chapter the proposed approach is explained. It follows the general content-fragile watermarking scheme as defined in Section 2.3.3 for verification of audio data integrity and also authenticity.

3.1 Outline of Content-fragile Watermarking Approach

This Section explains the protection model in the proposed content-fragile watermarking system and specific technical requirements.

3.1.1 Protection Model

The protection and verification is carried out in two stages (recall Figures 3.1 and 3.2 for a visualization):

1. Embedding Stage: It is assumed that the input audio data can be trusted with regards to integrity and authenticity. Now, the input audio is divided in shorter segments, typically 10 to 20 seconds. From every audio segment its perceptual hash is computed. The extracted hash is then embedded as a robust digital watermark message into the audio data, see Figure 3.1.

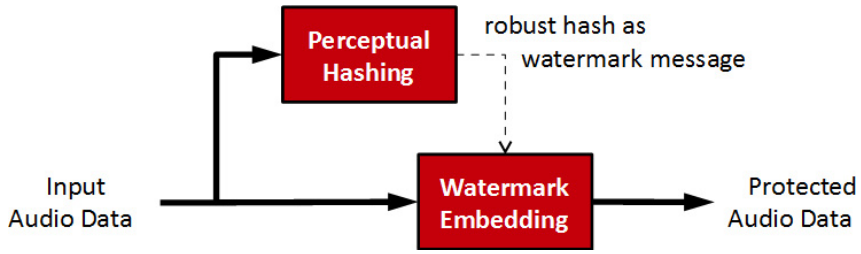


Figure 3.1.: Content-fragile watermarking model – protection/embedding stage

2. Detection Stage: At any future point in time, the watermark message, i.e. the perceptual hash of the previous *authentic* state of the audio data, can be detected and retrieved from the protected audio sections. Comparing the previous with the *current* state of the perceptual hash allows indicating integrity breaches, see Figure 3.2.

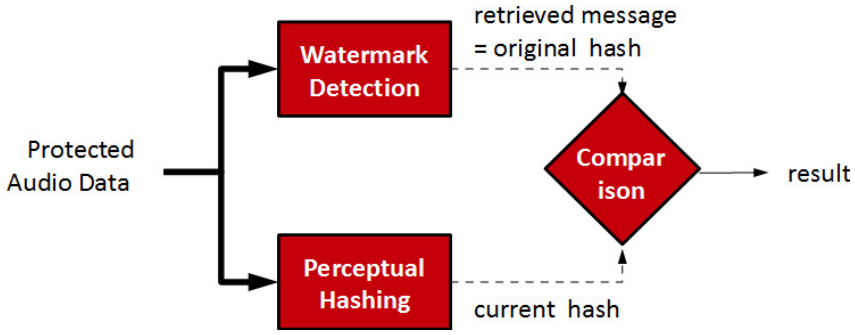


Figure 3.2.: Content-fragile watermarking model – verification/detection stage

For this, new or enhanced approaches for perceptual hashing, robust audio watermarking, perceptual modeling and their respective combination thereof are proposed. Prior to this, a number of technical requirements have to be introduced first.

3.1.2 Technical Requirements

In addition to the general objectives of this thesis work as defined in Section 1.2, a number of rather technical requirements on perceptual hashing and digital watermarking by themselves and on their combination in a content-fragile watermarking system have to be met:

1. **Acceptable sound quality of the watermarked content:** The watermark embedding should be as imperceptible as possible, i.e. the degradation of acoustic quality caused by embedding should be minimal. Furthermore, the embedding should be transparent for the digital representation of the signal, i.e. the embedding has to be carried out without changing the file format.
2. **Compliance of perceptual hashing with watermarking (and *vice versa*):** In a content-fragile watermarking approach, the perceptual audio hashing has to reflect the technical properties and limitations of the involved watermarking algorithm, and *vice versa* as follows:
   - **Invariance of audio hash against watermarking:** As the robust hash is part of an authentication watermarking system, obviously it must be robust also against the distortions introduced by the watermarking to avoid immediate false alarms. An appropriate *alignment* will be proposed in this thesis work so that watermarking and perceptual hashing do not interfere with each other.
   - **Watermarking Capacity:** Audio hashes, as any hash algorithm, necessarily need to provide a sufficient length (in *bits*) in order to avoid false alarms and to be secure against brute-force attacks or targeted birthday attacks. Most current robust watermarking approaches typically offer very low embedding rates of a few message bits approximately a dozen *fdefinitionbit* per second. On the other hand a payload of minimum 150 *bits* must be embedded approximately every 10 seconds. Hence moderate increase in embedding capabilities will be developed in the course of this thesis work.

- **Audio Hash Compactness:** Given by the mentioned capacity limitations the audio hash must provide a content descriptor that, albeit separating admissible from malicious acts of data modification, is compact enough. For comparison: 10.0 seconds of raw audio in PCM format (stereo, 16 *bit* per sample, sample rate 44.1 kHz) represent roughly 1.5 MB of input data. Nevertheless, this amount of data needs to be hashed into a compact hash output of, say, 128 *bit* (i.e. 100,000 times shorter than the input) that still allows separating admissible from malicious acts of data modification.

3. **Availability:** A meaningful requirement is that the watermark protection shall be provided for *all* possible input audio data if possible. In the context of watermarking this can be challenging as some kinds of audio material are not well suited for carrying an embedded watermark in the first place. This is frequently the case when the input audio has only little signal energy in some sections of its duration or in its spectrum. Examples are given by voice content with very ”clean/dry” voice sound (like in audio books), content with limited frequency bandwidth (like in telephony, radio communication), and some examples of modern electronic and/or experimental music. Here, the (partial) absence of signal energy makes it difficult to hide an audio watermark without becoming noticeable or even annoying to the listener. At least limitations with respect to applicability shall be identified.

   Example: An obviously extreme example of critical audio content is the composition *4’ 33”* by the avantgarde experimental composer *John Cage*: it ”contains” pure silence for a duration of 4:33 minutes. Recordings of live performances (by soloists or even symphonic orchestras) feature complete silence except for the ambient acoustic ”atmosphere” of the concert hall and the virtual tension of the audience[46].

4. **Security:** The watermark and the perceptual hash should withstand attacks aimed directly at the respective algorithm. With respect to watermarking, it must be impossible for an unauthorized user to read, change or delete the embedded message without rendering the marked cover unusable. With respect to the perceptual hash, it is required that no malicious tampering is possible that can circumvent the hash-based verification.

5. **Efficiency:** Finally, the proposed solution should provide sufficiently fast processing time. The real-time capability may be an essential requirement especially in live recording devices, for example a video camera or a PC-based recording unit.

[46] *John Milton Cage Jr.*, *4’33”* (pronounced: *”Four thirty-three”*), Composition in Three Movements, 1952

3.2 Perceptual Hashing Approach (*rMAC*)

This Section describes the perceptual hashing approach which is developed for achieving the objectives in this thesis. It incorporates ideas from an existing and often-cited audio fingerprinting scheme that was explained above in Section 2.2.3 [HOK2001b, HOK2001a].

> Own work-in-progress and results on the hashing approach were published at several conferences in multimedia security [ZS2007, ZS2008a, ZS2008b, ZS2008c].

3.2.1 Key-dependent DFT Feature Selection

> The initial publication on the contribution explained in this Section can be found in [ZS2007].

For a better understanding of the following explanation, the hash bit calculation definition from the original "*Phillips* Hash" approach by *Haitsma, Kalker et al.* shall be recalled from Equation (2.4):

$$d_{k,t} := e_{k,t} - e_{k,t+1} - \left[e_{k+1,t} - e_{k+1,t+1} \right] \quad .$$

That is, for every hash bit four energy coefficients from *adjoining* frequency indices k and $k+1$, and *consecutive* time-steps t and $t+1$ are evaluated[47].

Example: See Figure 3.3 (left Figure) for a visualization: Each "cell" in the Figure corresponds to one time-frequency index (k, t) in the virtual "area" in the time-frequency domain. The regular assignment of "picked" indices[48] along the time and frequency axis can clearly be seen. Note that the Figure only gives a *symbolic* representation of the time-frequency domain. The very limited frequency range of 10 indices is a simplified representation for demonstration. In this thesis work the actual "height" of the "area" of possible frequency indices can be up to 1024 which is given by (half of) the DFT frame length.

A security analysis of the "*Phillips* Hash" algorithm under knowledge of the algorithm, was not given by the original authors as the security is not a relevant requirement for audio retrieval purposes in many scenarios. Even more, in an earlier research work being supervised and co-authored by author of this thesis, it could be demonstrated that targeted attacks actually *are* possible:

> In [TNSZ2009], a targeted violation of the robustness requirement was implemented. It was shown that given pieces of audio data can be deliberately modified so that the modified data still sounds similar to the original data but their robust hash changes significantly. This targeted attack indeed could not be utilized for masking targeted doctoring in terms of a *2nd pre-image attack* or *collision attack*. Instead the attack is suited for *protocol attacks* as explained in Section 2.3.4.3 by evoking false alarms: Such false alarms eventually can render a content-fragile watermarking system useless.

For integrity protection, security must be provided with respect to the requirements given in section 2.2.1 in a way that an adversary cannot generate a hash collision or violate the robustness requirement as explained in the previous paragraph. On this *Swaminathan* or *Fridrich* pointed out that this should be addressed by a *key dependent* feature selection [FG2000, SMW2005] instead of a regular selection of coefficients as in the original work.

[47] Note that, albeit not very elegant, in the previous Equation the the original authors' notation is used in the following. This is done for the sake of clarity when comparing this thesis work with the original works.

[48] Note for the avoidance of confusion that the terms "index" or "'indices", resp., correspond to the *pair* (k, t) consisting of a frequency index k and time index t

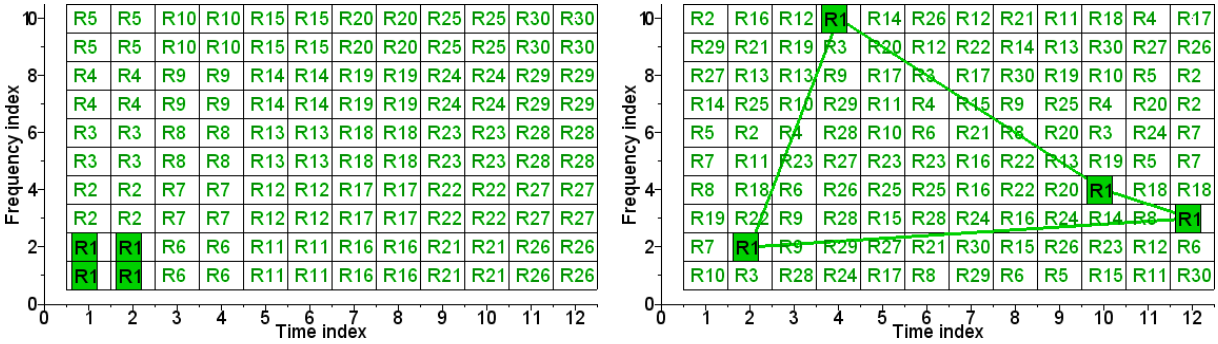


Figure 3.3.: Proposed *rMAC* feature selection scheme – Symbolic representation of DFT indices in the time-frequency domain. Every set of four correspondent coefficients denoted as "Rm" contributes to the m-th hash bit. Coefficients R1 which are correspondent to the first hash bit are highlighted for demonstration purpose;
Left Figure: picking of FFT coefficients in *regular* order according to Equation (2.4);
Right Figure: picking in *pseudo-random* order according to Equation (3.1).

Thus, it is introduced that each "sub-fingerprint" (using the original denotation) will now be extracted from an audio segment consisting of M consecutive frames from a few hundred times steps, i.e. from a few of seconds audio. Stereo or multi-channel input PCM signals are down-mixed to mono before the FFT spectrum is calculated.

Then it is proposed to derive the hash bits from four FFT coefficients at *arbitrary* time-steps $t_1, t_2, t_3, t_4 = 1...L$, and band indices k_1, k_2, k_3, k_4 in the time-frequency domain:

$$d'_{k,t} := e_{k_1,t_1} - e_{k_2,t_2} - \left[e_{k_3,t_3} - e_{k_4,t_4}\right] \quad . \tag{3.1}$$

The selection of time and frequency indices shall be performed *pseudo-randomly* dependent on a secret key $\mathcal{K}_2$, instead of using adjoining time and frequency indices. This approach conceals it to an adversary which frequency components contribute to a certain hash bit (and which are not). See Figures 3.3 (right Figure) or Figure 3.6 for a visualization: the pseudo-random picking of time-frequency indices can clearly be seen there.

Like in the original work, the value of the hash bits $H'_{i,t}$ is defined as the sign of d':

$$H'_{k,t} = \begin{cases} 1 & \text{if } d_{k,t} \geq 0 \\ 0 & \text{if } d_{k,t} < 0 \end{cases} \quad .$$

The reader should note that the definition in Equation (3.1) shows a certain formal similarity to the definition of the Patchwork embedding scheme described in Section 2.3.5.1: The definition in Equation (3.1) can obviously be rewritten as

$$d'_{k,t} := e_{k_1,t_1} - e_{k_2,t_2} - \left[e_{k_3,t_3} - e_{k_4,t_4}\right] = \underbrace{e_{k_1,t_1} + e_{k_4,t_4}}_{\alpha} - \underbrace{\left[e_{k_2,t_2} + e_{k_3,t_3}\right]}_{\beta} \quad . \tag{3.2}$$

Comparing this expression with Equation (2.5), the quantities α, β can be regarded as the partial sums $\alpha \simeq N\bar{a} = \sum_{i=1}^{N/2} a'_i$ and $\beta \simeq N\bar{b} = \sum_{i=1}^{N/2} b'_i$, respectively. That is, the analyzed/watermarked audio features d' and s' show a formal similarity.

Because the proposed audio hash is dependent on a shared secret $\mathcal{K}_2$ it can be regarded as a *message authentication code* (MAC) rather than a hash. As will be shown in the following Chapter, it withstands a number of signal transformations. To express its *robustness* property the following terminology will be used in the remainder:

> **Proposal:** The proposed algorithm for a *robust* message authentication code will be denoted as "*rMAC*" in the remainder of this thesis.

For completeness it should be noted that this denotation was later adopted by a different author as e.g. in [YZLX2013].

In contrast to the algorithm by the original authors, the following significant modifications are proposed:

- For the frequency band indices k_i no sub-sampling on a logarithmic frequency scale is carried out. In simple words: frequency bands are counted on a linear scale (with indices k_i from 1 to 512) instead of only 33 musical semi-tones. This provides a much larger basic population of indices that can be pseudo-randomly picked. The larger size of the population makes brute-force attacks much more difficult.
- No overlapping of the FFT framing was carried out. This shall avoid that the hash extraction process is affected by the subsequent watermark embedding. This simplification is a prerequisite for the alignment scheme as proposed in Section 3.4. Nevertheless, the loss of temporal resolution is well compensated for by the fine synchronization in the watermarking detector.
- The reader is also reminded that in the original approach by *Haitsma et al.* and in the previous explanation, *four* coefficients contribute to each fingerprint/*rMAC* bit. But the number of coefficients can be changed from four to any other *even* total number accordingly for improving improve the hash coverage across the audio spectrum. This increases the sensitivity and the computational efforts for a brute-force attack as will be shown in the numerical examples in Section 4.2. Hence, a total number of eight coefficients per hash bit will be used in the experiments to this thesis work.

The contribution of this Section is evaluated below in Sections 5.4.2, 5.4.3, 5.4.4 about the classification performance and in 5.4.8 about the security.

3.2.2 Standardization of DFT Spectrum

> The work in this Section extends earlier works as published in [ZS2007]. Results were published and used in [ZS2008a, ZS2008b, ZS2008c].

Closer analysis shows that for real-world audio data like music or speech recordings, the FFT magnitude coefficients are neither equally distributed nor uncorrelated. For example, in most speech recordings, the coefficients corresponding to lower frequencies (e.g. below 1000 Hz) usually have a much greater mean energy than coefficients corresponding to higher frequencies. Two examples are shown in Figure 3.4. Thus, the key dependent selection of bands for fingerprint extraction raises problems with respect to the security requirements listed in Section 2.2.1: If such low frequency coefficient is selected by the *rMAC* key as the first summand in Equation (3.1) that coefficient will dominate the sum. Thus, the respective hash bit will be more likely a "1" than a "0" which decreases the sensitivity to *for any kind* of integrity breaches of the input voice data. The *rMAC* as a whole would not be equally distributed, and *rMAC*s

from different music segments would not be independent, allowing systematic security attacks on the *rMAC*.

Example: In internal/unpublished experiments, an *rMAC* of length $M = 128$ *bit* was extracted from audio data units of 10 seconds (6.5 hours of total duration). The *rMAC*s of the original files were mutually compared to a time-shifted copy of the same file: For the example of speech data this simulates malicious tampering by replacing parts of the audio with other audio data that has the same voice or background noise.

One would expect that the *rMAC* bits should be identical only by chance, and the average Hamming distance should be close to 64 *bit* (i.e. at a bit error rate of 0.5). In fact, the average Hamming distance was only approximately 39 *bit*, as can be seen from the histogram in Figure 3.5. Closer analysis showed that the *rMAC* bits were not equally distributed, and some were even almost constantly one or zero. This reduces the Hamming distance when comparing any audio data in term of the *rMAC*.

Therefore, a *standardization* of the FFT spectrum is proposed as follows: all FFT coefficients $e'_{k,t}$ at *fixed* band index k at all time-steps $t = \{1, ..., L\}$ in an audio segment are regarded as a random variable with empirical mean m_k and variance s_k^2. Then the *standardized* values are defined as:

$$e'_{k,t} := \frac{e_{k,t} - m_k}{s_k} \quad . \tag{3.3}$$

with

$$m_k := \frac{1}{L} \sum_{t \in \mathcal{T}} e_{k,t} \qquad s_k := \sqrt{\frac{1}{L} \sum_{t \in \mathcal{T}} \left(e_{k,t} - m_k\right)^2} \quad .$$

Note that the mean and variance are not calculated across all time-steps. Instead, the set $\mathcal{T} \subset \{1, ..., L\}$ denotes the subset of time indices that will *not* be used for watermark embedding. This is done in anticipation of the subsequent watermark embedding step. Skipping the watermarked time indices avoids an impact of the watermarking distortions on the *rMAC* values. This is explained in detail in Section 3.4.1 about the feature *alignment*.

Because of the basic properties of the propagation of mean and the variance, all transformed quantities $e'_{k,t}$ of a given band index k have zero mean and variance 1.0 and thus contribute

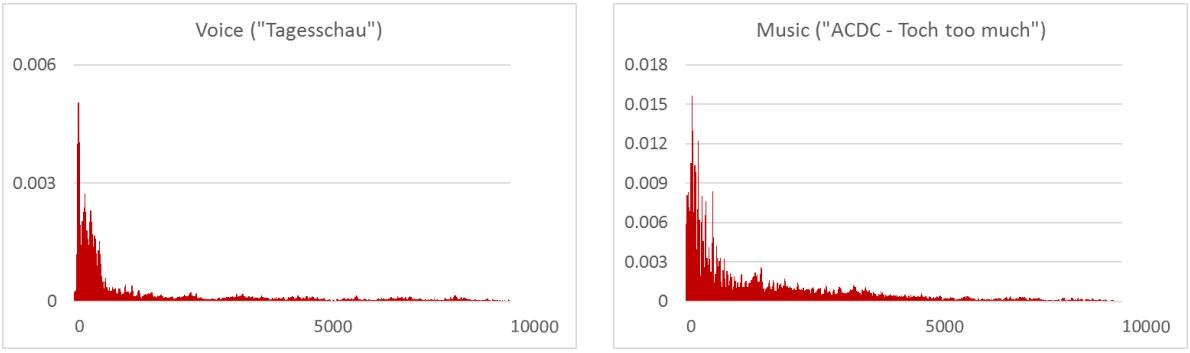


Figure 3.4.: Example: Distribution of average signal energy across frequencies from two audio example files
Left plot: news anchorman in German TV newscast *"Tagesschau"*, male voice;
Right plot: Rock song *"AC/DC: Touch too much"* (duration 15 s each

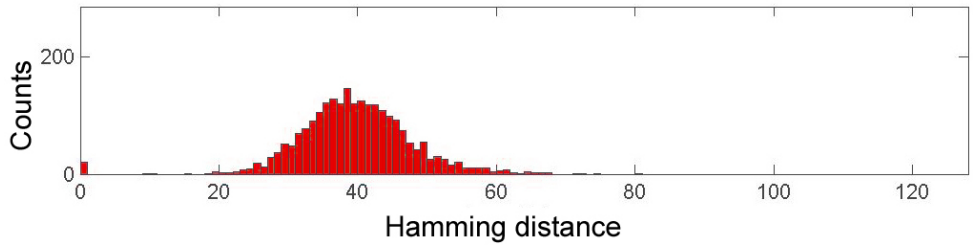


Figure 3.5.: Example: Histogram of *rMAC* Hamming hamming distance for time-shifted audio data versus original audio data; Source: own experiments (unpublished)

equally to the hash bit calculation. Then, the sign comparison is done out on the *standardized* FFT coefficients selected as described above:

$$d'' := e'_{k_1,t_1} - e'_{k_2,t_2} - \left[e'_{k_3,t_3} - e'_{k_4,t_4}\right] \quad . \tag{3.4}$$

From the basic properties of the mean and the variance, it can be seen that this quantity d''

- also has zero mean,
- it has a variance of $(s'')^2 = N$ as the standardized coefficients $e''_{k,t}$ of different time-steps and frequency indices across many seconds of duration can be regarded as sufficiently uncorrelated.

As will be shown in the experimental results in Chapter 5 the related *rMAC* bits

$$H''_i = \begin{cases} 1 & \text{if } d''_i \geq 0 \\ 0 & \text{if } d''_i < 0 \end{cases} \quad i = 1...M \tag{3.5}$$

will be equally distributed on {0,1} as required in Section 2.2.1 with the additional property of key-dependence. In order to provide a sufficient degree of security against brute force attacks, a minimum of $M = 128$ *rMAC* bits will be extracted. Finally, the full *rMAC* H is constructed straight-forward by concatenation of all *rMAC* bits as an M-tuple, i.e.

$$H = (H''_1, H''_2, ..., H''_M) \quad .$$

Note again that more than four coefficients might be used for calculating a hash bit; any *even* value N is permitted. For this, the Equation (3.4) would have to be extended accordingly.

The experimental evaluation of the standardization step can be found in Section 5.4.5.

3.2.3 Temporal Localization of Tampering – "Scattered" Approach

The contributions in this Section were published in [ZS2008a, ZS2008b, ZS2008c].

One of the objectives of this thesis is the *temporal localization* of tampering. In order to facilitate this, two approaches are proposed in this and the following Section.

For describing the first approach the pseudo-random assignment scheme for the *rMAC* time-frequency indices proposed above is recalled: according to the description in Section 3.2.1 the N time-frequency indices that contribute to a fixed *rMAC* bit index are *scattered* across the *full*

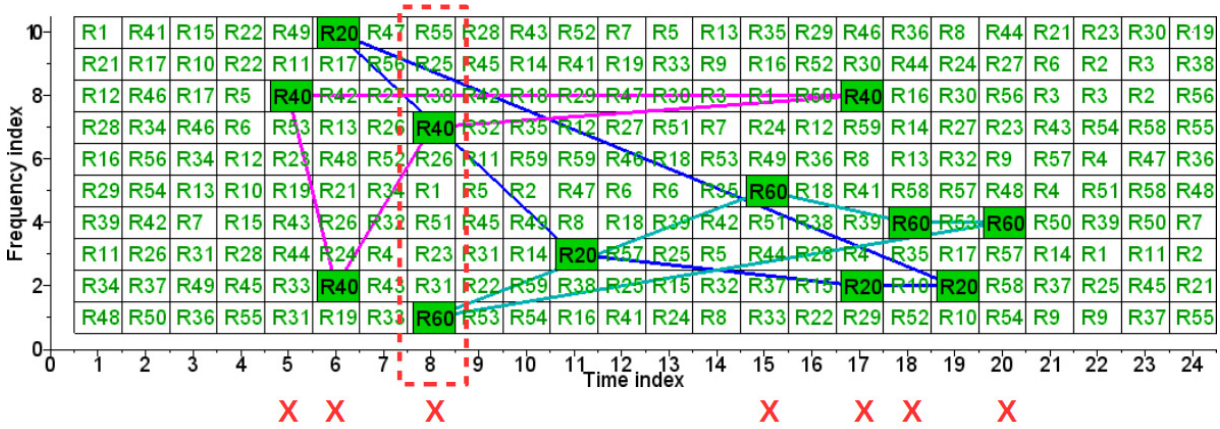


Figure 3.6.: Example: *rMAC* coverage across time steps ("scattered" mode): As example, the quadruples of coefficients which correspond to the 20th, 40th, and 60th hash bit (denoted as "R20", "R40", and "R60") are highlighted for demonstration purpose; dashed box: an attack at time index 8 is assumed. "X" markers: time indices that correspond to the attacked bits R40 and R60

area in the time-frequency domain (denoted as *"scattered mode"* in the remainder). The exact pseudo-random sequence is assumed to be known (given by the known secret key). It is very likely that coefficients from different *rMAC* bits will fall into a certain time index. Hence it can be expected that a breach of data integrity even with a short duration can have an impact on more than one *rMAC* bit.

If on the detector side one of the audio hash bits is later found to "flag" an attack, this has to be explained by data modifications in *any* of the N correspondent coefficients. Their correspondent time-indices in question are known. However, it is obscure *which* of the N coefficients were modified. A best guess it is proposed by identifying those time indices for which the above "*rMAC* flags" *accumulate*.

Example: A symbolic visualization is given in Figure 3.6: As an example it is assumed that a very short attack is carried out at time index $t_0=8$. This attack has a 50% chance to be indicated by the *rMAC* bits with index 40 and 60 (the other indices "31, 23, 51, ... etc." are omitted for simplicity). These two bits cover seven time indices (flagged with an "X" marker in the Figure) which are "candidates" for the correct location of the attack. But the audio frame at time index $t_0=8$ is the only in question which is flagged more than once. It can be seen as a best guess for the correct temporal position of the integrity breach.

Note that in Figure 3.6 each of the 24 time steps is covered by the same number of *rMAC* bits (namely 10) which is a simplified visualization. In this thesis experiments the total number counter will vary from time step to time step by a few times because of the pseudo-random scattering. Realistic figures about this can be seen in the following...

...Example: In the experiments in Chapter 5 an *rMAC* of $M=128$ *bit* is extracted. $N=8$ different FFT coefficients at a time contribute to each of the *rMAC* bits. With the technical settings that were used for carrying out those experiments, the duration of each hash is 255 frames (total duration: 11.8 seconds). Figure 3.7 shows how many *rMAC* bits "in question" cover

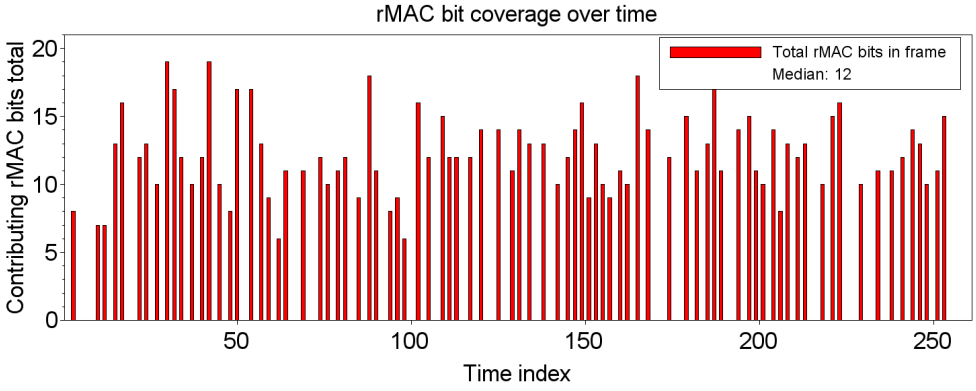


Figure 3.7.: Example: "*rMAC* bit coverage" counter versus time

the respective frames (no matter if attacked or not). As can be seen, the total number varies from 6 to 19. The respective median value across all non-zero entries is 12. For audible data modifications it can be expected that a sufficiently large portion of these will indicate and accumulate "flagged" bits.

Note: In Figure 3.7 it can be seen that a majority of frames is not covered *at all*. This is caused by the *alignment* of the *rMAC* feature extraction with watermark embedding as explained below in Section 3.4.1. Only 33% of available audio frames are used for *rMAC* extraction while the remaining are spared. Their total *rMAC* bit coverage is zero.

In order to reflect the varying total number of potentially indicating *rMAC* flags, the quantitative analysis is carried out as follows:

1. For each frame, i.e. for each time index, it is counted how many *rMAC* bits in total correspond to the *rMAC* coverage of that time index (*"rMAC bit coverage per frame"*). Formally this can be expressed as follows: At first, let $\mathscr{R}$ be the union set of *all* picked indices $\mathscr{R} := \{(k_{i,n}, t_{i,n})\} \subset \mathbb{N} \times \mathbb{N}$ that correspond to all M *rMAC* bits with each N spectral coefficients ($i = 1...M, n = 1...N$). Then, let $\mathscr{R}_{t_0}$ be the subset of those indices that have a given/fixed time index t_0, i.e.

$$\mathscr{R}_{t_0} = \left\{(k_{i,n}, t_{i,n}) : t_{i,n} = t_0\right\} \subset \mathscr{R} \subset \mathbb{N} \times \mathbb{N} \quad \forall (i,n) \quad ,$$

no matter if the i-th *rMAC* bit flags an attack or not.

2. Then it is counted how many "*rMAC* flags" indicate a particular frame as being presumably attacked. For this, let $\mathscr{A} = \{a_1, a_2, ..., a_h\}$ denote the set of attacked *rMAC* bit indices that are flagging an attack on the audio. For avoidance of confusion: its cardinality $|\mathscr{A}|$ is equivalent to the Hamming distance $h \leq M$ between the authentic *reference* rMAC and the *test rMAC*. Then, it is considered the subset of correspondent indices $\mathscr{F}_{t_0}$ that flag an attack at a given time index t_0. This subset of *"rMAC flags per frame"* can be written as

$$\mathscr{F}_{t_0} = \left\{(k_{i,n}, t_{i,n}) \; \forall (i,n) : (t_{i,n} = t_0 \wedge i \in \mathscr{A})\right\} \subset \mathscr{R}_{t_0} \subset \mathbb{N} \times \mathbb{N} \quad .$$

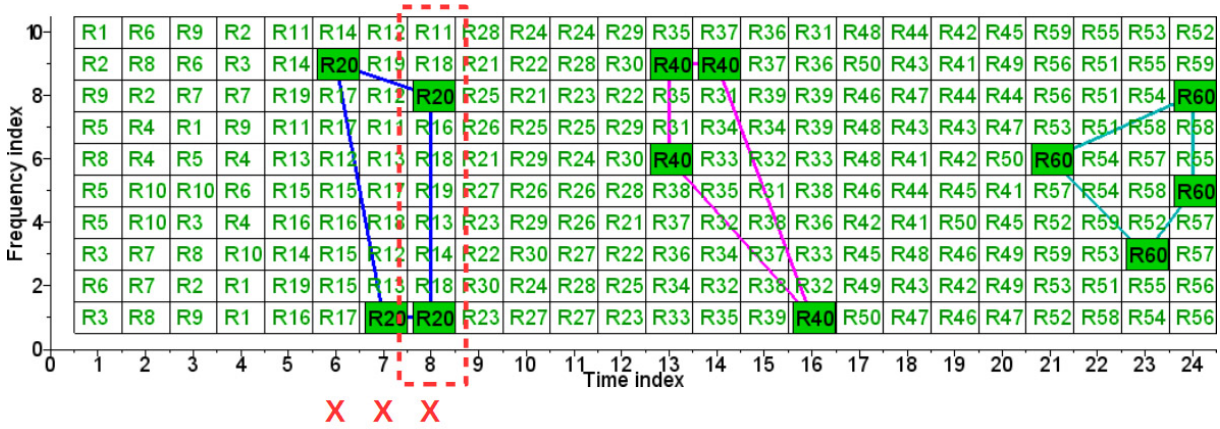


Figure 3.8.: Example: *rMAC* coverage across time steps ("serial" mode)

3. Finally it is proposed to consider the following quantity as a measure for the localization: In each time step t, the ratio of the total number of *rMAC* flags per frame is divided by the *rMAC* bit coverage of that frame, i.e. the quantity

$$\alpha_t := \frac{|\mathscr{F}_t|}{|\mathscr{R}_t|}$$

is used as criterion (denoted as *"rMAC flag ratio"*).

Especially if non-zero *rMAC* flag ratios can be observed for a number of consecutive times indices *in sequence* this should indicate a localized attack on the protected audio data.

This approach is somewhat in analogy to the image watermarking approach by *Huajian Liu* [LS2006b, LS2006a] as outlined in Section 2.5.2. Independent of the media type (image versus audio), the detection is carried out by analyzing for accumulations of "flags" on a whatsoever authentication code in a "soft decision" manner. By contrast, the proposed approach uses a content-dependent robust MAC instead of a pseudo-random but fixed but one.

3.2.4 Temporal Localization of Tampering – "Serial" Approach

> The thesis work described this Section was partly conducted in the course of the Master's thesis by *Munir* [Mun2011] (under supervision of the thesis author) and eventually published [ZMS2012].

On the downside of the approach explained in the previous Section, a significant number of false positive *rMAC* flags are identified in the first step which have to be "filtered out" by setting reasonable thresholds about the *rMAC* flag ratio. It is worth investigating if such false positives can be avoided in the first place to further increase the sensitivity and specificity. To achieve this, it is proposed:

Proposal: The set of the N coefficients contributing to an *rMAC* bit shall by placed only in a limited period of time. For simplicity, *rMAC* bits are extracted in serial order over time (denoted as *"serial mode")*

In simple words: the hash bit with index zero (the ”first” bit) shall be placed rather at the *beginning* of the hashed audio section (with a certain degree of freedom), the ”last” hash bit shall be placed near the *end* of the hashed audio section. This should provide that the *rMAC* flags accumulate quickly and in a limited range which avoids false detections *off* the true attack positions.

Example: The Figure 3.8 visualizes this assignment principle symbolically: The ”degree of freedom” for the time indices of is now limited. Lower *rMAC* bit indices are placed near the beginning while higher bit indices are placed near the end (the Figure shows it for the 20th, 40th, and 60th bit). That is, the bit index directly correlates to the time index. For an assumed attack on the audio data at time step $t_0=8$ all affected *rMAC* bits are scattered only in a close temporal neighborhood of this time index.

Experimental Example

> The empirical results as cited in the following were published in [ZMS2012] as ”work in progress”.

Here, the modified algorithm was tested on audio files that contained speech and music with approximately one hour total duration. The audio files were divided into segments of 10 sec. playing time (360 total). From each segment a 128 *bit* audio hash was extracted according to the explanation in this Section. The audio hash was embedded as a robust watermark and later verified. A secret key was chosen for creating pseudo-randomly picked time-frequency indices for feature extraction and watermark embedding.

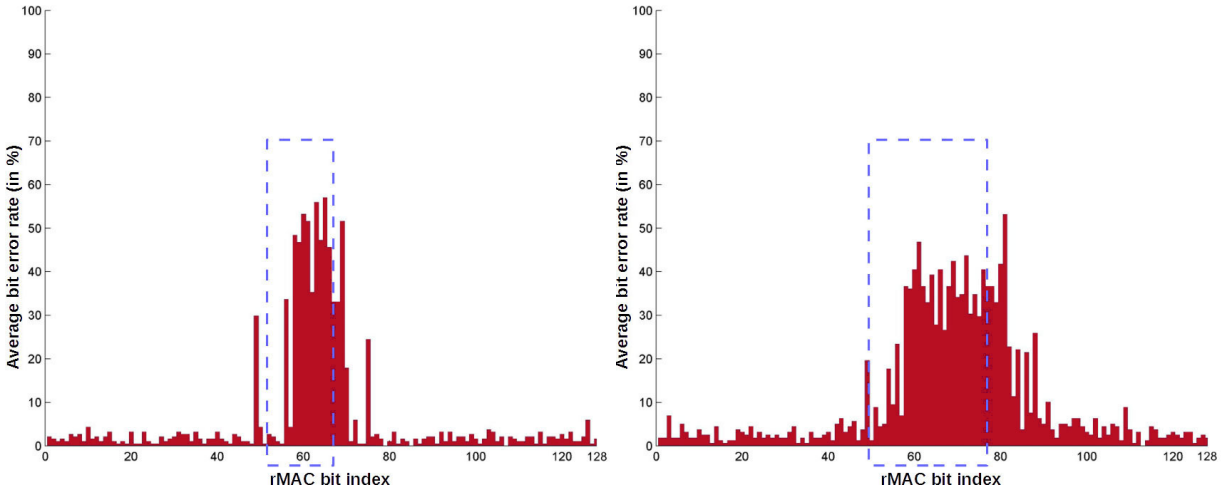


Figure 3.9.: Test result: *rMAC* with temporal localization;
left plot: Copy and move attack from time code 0:05 min to 0:06 min
right plot: Noise adding attack from time code 0:05 min to 0:07 min
dashed boxes: the actual duration xof the audio attacks
Source: own publication [ZMS2012]

After embedding but prior to the *rMAC* verification the following attacks were applied to the audio content (and hence the embedded watermark):

- **Copy-and-move attack**: In each audio segment, one second of audio starting at time code 0:05 *min.* was replaced with one second of audio from the same file at time code 0:08 *min.*

This should simulate that audio content is maliciously ”removed” by replacing it with other content from the same context.

- **Noise adding attack**: Into each audio segment, a 2-second white noise file was mixed into the audio data starting at time code 0:05 min. The noise power level was tuned so that the original content in the middle became rarely intelligible. This should simulate that audio content is maliciously ”removed” by masking with noise instead of deleting/erasing it.

That is, the attack on the audio affects the respective *middle* section of the hashed audio segment in an audible way.

The earlier empirical results in Figures 3.9 show the average bit error rate (BER) for each hash bit index between the originally embedded audio hashes *before* the attacks versus *after* the attacks. As can be expected the BERs are significantly higher in the ”middle” sections of the audio hashes than in the other sections. That is, modifications can be correctly indicated and also localized implicitly by the bit index. It must be admitted that the BERs in the actually *unmodified* audio sections do not vanish as ought to be expected, see Figures 3.9. This is caused by general shortcomings in the watermark detection as discussed in Section 5.4.1.

Results from more elaborate tests of the major contribution of this Section can be found in Section 5.4.7.

3.2.5 Discontinued approaches

A number of further ideas for enhancements of the feature extraction of this audio hash had been investigated. But preliminary experimental evaluation eventually showed that they were not capable to improve the distinction performance any further. The main ideas of those discontinued approaches are listed only briefly in the following for completeness:

1. Psychoacoustic Model-based Adaptive Quantization / Feature Extraction It was also investigated if the robustness can be improved by *quantization* of the DFT coefficients prior to hash modeling. Special attention was paid to both uniform quantization and non-uniform quantization based on psycho-acoustic modeling. The former means to represent the FFT absolute values or the respective sound pressure levels in a resolution of ± a few *Decibel* for example. About the latter case the idea was investigated to represent inaudible spectral components in lower resolution than audible components or the like. Preliminary result showed that in total the overall robustness does not benefit from either quantization processing.

   On the one hand, more coarsely quantized coefficients *do* benefit from their lower sensitivity to noise. But on the downside, occasional *quantization errors* cannot be avoided in principle and would have an even greater effect now. Eventually, both effects annihilate and error rates and accuracy did not improve even further, at best. Hence, a uniform and fine quantization of ± 0.01 Decibel was used in the experiments in Chapter 5. For the same reason, key-dependent feature quantization as proposed by Nouri [NZAF2012] could not be adapted either.

   Thus, quantization approaches as published by the author as part of [ZS2009b, ZS2008b] have not been not pursued in this thesis work. The same is true for an approach of using the frequency indices of tone-like or noise-like components of the audio spectrum. Preliminary research results were published in [ZSN2005].

2. Avalanche Effect Another idea was to introduce an algorithmic *avalanche effect* for the *rMAC*: The classification accuracy was investigated when a particular DFT coefficient contributed

to more than one *rMAC* bit. In simple words, the edges of the ”polygon” in the symbolic visualization in Figure 3.3 should not be fully disjoint.

Closer analysis showed that the true positives (i.e. the sensitivity) to actual tampering would successfully increase. But, simplified, the false positives would increase even more which in total would provide a lower accuracy.

Hence research activities in these directions were discontinued in the meanwhile.

3.3 Enhancements to Patchwork Audio Watermarking

This Section describes the proposed adaptations to the audio watermarking algorithm that became necessary for this thesis work in order to achieve better watermarking capacity and robustness.

3.3.1 Pseudo-random Patchwork Assignment in Time-Frequency Domain

> The following contribution was outlined in a joint publication in [BZSS2011]. The publication explains the overall algorithm idea and applies it in a context different from this thesis objectives, namely *collusion-security* in transaction watermarking.

For embedding the perceptual audio hash as a watermark message, an existent blind spread spectrum watermarking approach [Ste2003, SZSL2003] as explained in detail in Section 2.3.5.2 is used as the technical basis. Here, the embedding is basically done in the Fourier domain according to the Patchwork watermarking approach.

In the original approach, each frame carries one bit of the watermark message. Typically a few dozen out of a few hundred total possible FFT coefficients are pseudo-randomly picked per frame. In order to increase the robustness, each message bit is embedded redundantly in a group of consecutive frames (typically two to ten frames). This increases the population of the Patchwork detection statistics in Equation (2.5). It allows detecting the message bits more significantly, even in the presence of distortions of the marked signal. Finally, subsequent bits of the watermark message are embedded into subsequent frame groups in sequential order.

Example: Figure 3.10, left, demonstrates the embedding of message bits in sequential order. For simplification purposes in the visualization, the total message length is only $M=4$ *bit*, and only $2N=6$ frequency indices out of $W=10$ total (the band*width*) are picked per frame. Each message bit is embedded redundantly three times i.e. in three consecutive frames. That is, the total duration is $3 \cdot 4$ Frames $= 12$ Frames, and for each message bit $3 \cdot 6$ coefficients $= 18$ coefficients are used. The virtual area in the time-frequency domain covers $12 \cdot 10 = 120$ possible coefficients total.

For a sufficiently fine-granular detection of doctoring, protected units of a few seconds are desired. Thus, a sufficient embedding capacity of the audio watermarking is vital with regards to the requirements listed in Section 3.1.2. So far, the existent approach as explained previously has been insufficient due to its rather inflexible assignment of watermarked Fourier coefficients to message bit indices. Closer analysis shows that a capacity improvement by a factor of 5 to 10 is required. More detail can be found in the numerical example given below.

Hence, it is proposed to implement a more flexible mapping:

> **Proposal:** All coefficients correspondent to a certain message bit index shall be pseudo-randomly distributed across the *complete* available virtual area in the time-frequency domain, instead of a fixed mapping.

The fixed mapping is given up for the sake of higher flexibility and coverage as it is visualized in Figure 3.10, right. The ”cloud-like” scattering of the embedding location easily allows increasing the embedding capacity compared to the fixed mapping by increasing the width of the watermarked sub-band (vertical axis).

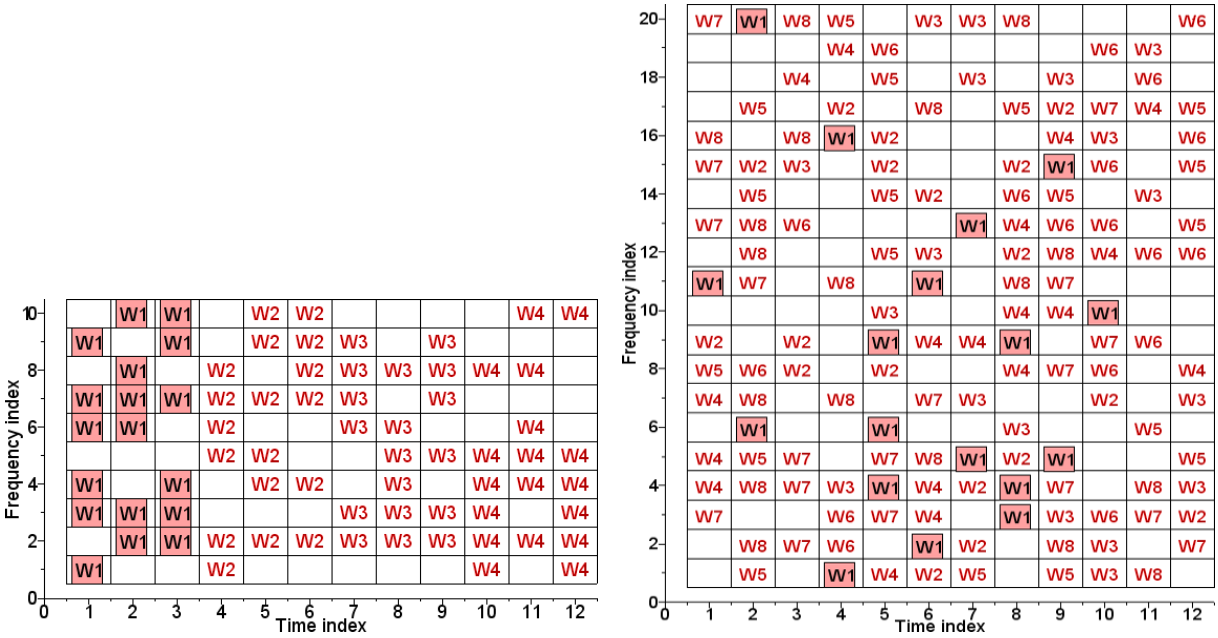


Figure 3.10.: Proposed watermark embedding schemes – Message bits embedded in *sequential* temporal order (left Figure) and in *pseudo-random* temporal order (right Figure). Every 9+9 correspondent coefficients Wn contribute to the n-th message bit according to Equation (2.6). The respective coefficients of the first message bit W1 are highlighted for demonstration purpose.

This contribution is experimentally evaluated in Section 5.3.1 and affects all results in Section 5.4.

Numerical Example

As a realistic figure, it can be assumed that the actual gross message length is roughly 450 *bit*. This includes the *rMAC* itself, additional meta data, a short CRC code and that forward error correction is used. For achieving a reasonable robustness a total number of 180 coefficients shall be used for embedding each of the bits into the spectrum. The frequency range of 1 kHz to 9 kHz (i.e. bandwidth 8 kHz) shall be used for embedding. This bandwidth corresponds to $W = 8\,kHz \cdot 2048/44.1\,kHz = 371$ frequency indices. This is visualized as the ”height” of the virtual ”area” in the time-frequency domain as visualized in Figure 3.10.

The area of all 450 message bits requires to hold $450 \cdot 180 = 81000$ coefficients. Hence, the ”length” of the area along the time axis should be $81000/371 \approx 219$ frames. This represents approximately $219 \cdot 2048/44.1kHz \approx 10$ seconds. That is, the flexible implementation allows for shortening the duration by using more coefficients per frame i.e. increasing the frequency bandwidth. Note that with a *fixed* mapping of time indices to watermark such embedding capacity can *not* be increased by just using more coefficients per frame.

The contribution of this Section is evaluated in Section 5.3.1 and contributes to all test results in Section 5.4.

3.3.2 Psychoacoustic-enhanced Watermarking Detection

The work described in this Section was conducted in a student thesis by *Merlé* [Mer2007] under supervision of the thesis author and eventually published in co-authorship [SZ2008b].

For a better understanding of the following explanation, the Patchwork detection algorithm investigated closely in this thesis shall be recalled. According to the "detection model" in Equation (2.5), the detection and retrieval of a watermark message bit is performed by evaluating the quantity $S' := \bar{a}' - \bar{b}' = 1/N \sum_i a'_i - 1/N \sum_i b'_i$. Remember that each summand a'_i and b'_i is a Fourier coefficient from a spectrum that potentially was being modified before according to the definition in the "embedding model" in Equation (2.7):

$$\begin{aligned} m = \text{"0"} \quad &\Rightarrow \quad a'_i = {a_i}^{1-d_i} \quad b'_i = {b_i}^{1+d_i} \quad , \qquad 0 \le d \le 1 \quad \forall\, i = 1 \ldots N, \\ m = \text{"1"} \quad &\Rightarrow \quad a'_i = {a_i}^{1+d_i} \quad b'_i = {b_i}^{1-d_i} \quad , \qquad 0 \le d \le 1 \quad \forall\, i = 1 \ldots N, \quad . \end{aligned}$$

Obviously, the greater the value of $|d_i|$, the greater the increase/decrease of the Fourier original coefficients a_i, b_i.

For completeness note that in the software implementation in this thesis work, coefficient values a and b are smaller than 1.0 due to the normalization coefficients during the FFT spectrum calculation. Hence taking a and b to the power of $1 + d \ge 1.0$ in fact *decreases* the coefficient. However this is irrelevant as long as embedder and detector software both correspond to this convention.

For now, all summands contribute *equally* to the quantity S in the above detection equation. But this does not reflect well that the d_i were controlled by a psychoacoustic model and are content-adaptive and frequency-dependent: for real-world audio content some coefficients probably underwent stronger modifications and contribute more to the quantity S than others. This can be utilized *on the detector side* as described in the following.

Effects of Multiple Embedding on Psychoacoustic Characteristics

At first it was investigated if and how the embedding process changes the psychoacoustic properties of the audio. For this the behavior of the exponents in the embedding model was evaluated. This was done by simulating that the watermarked audio was watermarked *again*. Then the correspondent exponents for the first embedding step $(1 \pm d)$ and the virtual second embedding step $(1 \pm d')$ were compared.

Close analysis in [Mer2007] showed that the psychoacoustic characteristics are maintained in the course of embedding. This means that the psychoacoustic properties of the original cover data can be reconstructed from the watermarked content at detection time. Based on this result, several approaches were implemented that utilize this obtained knowledge. The two most important methods are:

Method 1 – Static Prioritization of Sub-bands

It was found that Fourier coefficients that correspond to frequencies smaller than 2 to 3 kHz in practice undergo much less modifications than at higher frequencies for many kinds of input audio data. That was verified for one hundred test files with different music data, voice data, and few other synthetic test files (noise, seesaw signals etc.). It was learned that under 2.0 kHz the values of d_i typically are in the range $\pm 10^{-2}$. The range at higher frequencies is at least ten times greater, see Figure 3.11 (tested for the range 1.0-5.0 kHz).

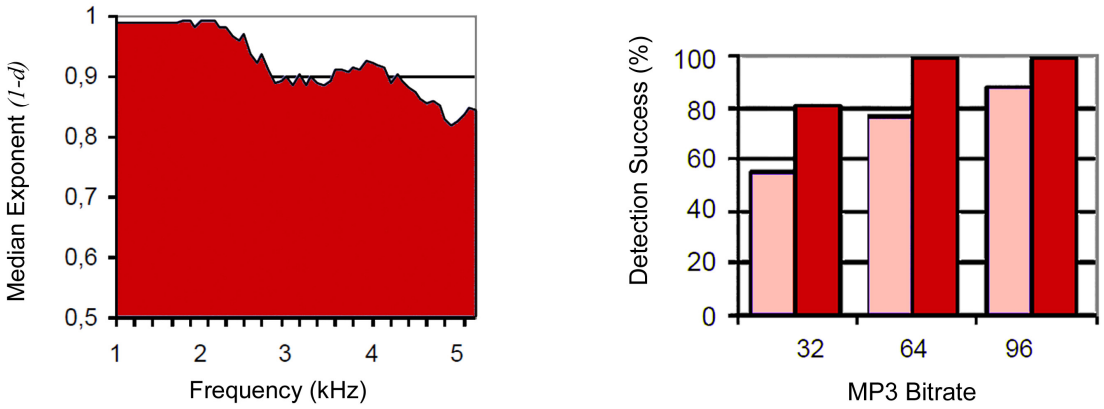


Figure 3.11.: Left figure: test results of median value of "embedding exponents" $(1-|d|)$ across 100 test files; right figure: test results of robustness test at different *MP3* bit rates, bright bars: detection success *without* psychoacoustic modeling; dark bars: psychoacoustic modeling enabled in detector
Source: own works [SZ2008b]

.

Thus it is proposed that Fourier coefficients correspondent to frequencies greater than 2 kHz are picked with a higher priority than those at smaller frequencies. Different prioritization can be realized by ignoring less relevant coefficients or using them with a lower weight. Note that this prioritization is done *statically* for any kind of input data in every time-step and is not content-dependent.

Method 2 – Dynamic Frequency-dependent Weighting

Another approach utilizes the instantaneous psychoacoustic properties in a watermarked audio frame. As explained above, the unknown psychoacoustic properties at embedding time in terms of exponents $(1 \pm d_i)$ can be approximated by those $(1 \pm d'_i)$ available at detection time. Then, those reconstructed exponents are used to implement an adequate *weighting* of the Fourier coefficients in the detection model in Equation (2.5).

For this, it is defined that the measured Fourier coefficients are taken to the power of the estimated exponents *again*:

$$a''_i := (a'_i)^{1-d'_i} \quad , \quad b''_i := (b'_i)^{1-d'_i} \quad . \tag{3.6}$$

Finally, the detection model is applied to these re-transformed spectral coefficients

$$S'' := \bar{a}'' - \bar{b}'' = 1/N \sum_i a''_i - 1/N \sum_i b''_i = 1/N \sum_i (a'_i)^{1-d'_i} - 1/N \sum_i (b'_i)^{1-d'_i} \quad . \tag{3.7}$$

This non-linear weighting reflects exactly the non-linear behavior of the earlier embedding process. Each coefficient is weighted according to the modification it probably underwent at embedding time, earlier. Note that this method is truly dynamic as the time-varying, content-dependent psychoacoustic model parameters of the respective audio frame are used.

Note that this approach is somewhat similar to one of the (rare) related works published by *Kirovski* [KM2003] that reflect perceptual modeling in the detection. The authors propose that the "detector should correlate [i.e. evaluate] only the audible frequency magnitudes". In

contrast, the approach investigated in this thesis *does* evaluate such inaudible frequency components, but downgraded by weighted by their distance to the masking threshold.

Implementation in the Detector

Both methods were implemented and integrated in the watermarking algorithm used in this thesis. Test results showed that the proposed approach can improve the watermarking robustness for any kind of watermark message. An example of test results for *MP3* robustness as published in [Mer2007, SZ2008b] is shown in Figure 3.11. It shows that using the psychoacoustic-based detector can improve the detection success rate.

More elaborate test results can be found in Section 5.3.5.

3.4 Combined Perceptual Hashing/Watermarking Approach

3.4.1 Alignment of Hashing and Watermark Embedding

Own work-in-progress and results about the contribution in this Section contribute to the following own publications at conferences in multimedia security: [ZS2008c, ZS2009b, ZMS2012].

This Section explains how the proposed audio hashing and the enhanced audio watermarking approaches as explained in the previous Sections 3.2 and 3.3.1. are combined so that the thesis objectives can be met.

As pointed out in Section 2.2.1, in a content-fragile watermarking system it is essential that the audio data distortions caused by embedding whatever watermark message do not significantly change the values of the *rMAC*. For an *rMAC* algorithm that is perfectly insensitive to consecutive watermark embedding, one would expect the Hamming distance of the *rMAC* before and after applying the protection to be *zero*. In fact, at an earlier stage of the research as published in [ZS2008c, ZS2009b], the Hamming distances did not vanish completely[49]. This can be seen from the following...

...Example: For the experiments published in [ZS2008c] an *rMAC* of length $M=128$ *bit* was extracted from audio data units of 10 seconds (6.5 hours of total duration). At reasonable embedding strength, the distortions of the watermark embedding process *did* caused some of the *rMAC* bits to flag a change of the bit value, as can be seen from the histogram in Figure 3.12.

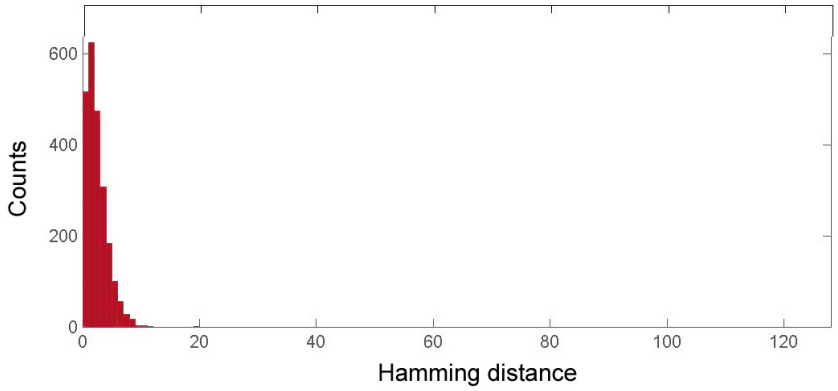


Figure 3.12.: Example: Histogram of *rMAC* Hamming distance for watermarked audio data versus original audio data;
Source: [ZS2008c]

Intensive analysis in the previous analysis showed that the *rMAC* extraction and the watermark embedding inside a particular audio frame were not completely separated due to *spectral leakage* effects (as explained in Appendix A.1.3.3): embedding a watermarking into a certain Fourier coefficient *does* have an influence on neighboring coefficients in the spectrum too. If such coefficient is picked by the *rMAC* feature selection, the *rMAC* which the coefficient contributes to might flip. This is often observed for said "weak" bits for which the continuous measure d' as in Equation (3.1) has an absolute value close to zero. Therefore, the process of *rMAC* feature ex-

[49] In the reviewers' comments of the paper submission to [ZS2009b] the notable false-positive rate was requested for improvement.

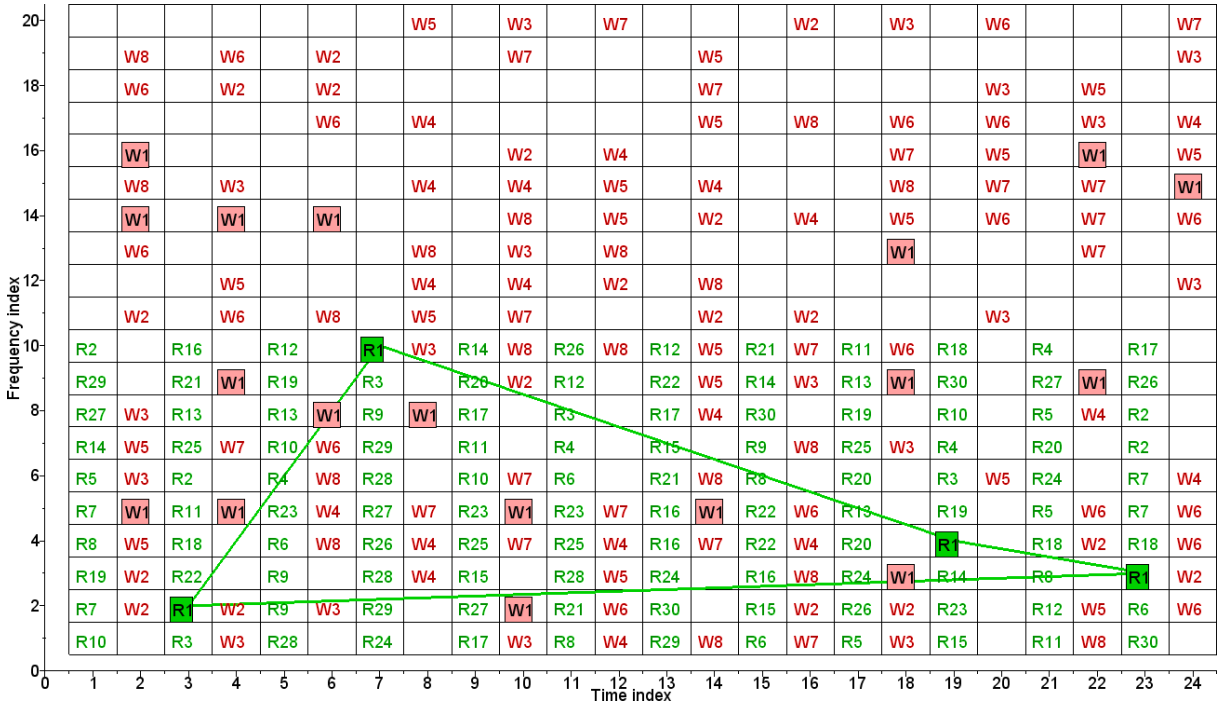


Figure 3.13.: Proposed alignment scheme for time-frequency indices for *rMAC* bits Rm and watermark message bits Wn embedding. For simplification purposes, coefficients are separated in "odd/even order" along the time axis.

traction and the robust watermarking scheme involved are proposed to be appropriately *aligned* as follows:

> **Proposal:** *rMAC* feature selection and watermark embedding are carried out on disjoint sets of audio frames (denoted as "alignment" in the following).

For achieving this, the audio data is separated internally along the time axis by first partitioning the audio frames in every protected audio section into two classes:

- The majority of audio frames is used exclusively for embedding the watermark message.
- The minority of audio frames is used exclusively for the *rMAC* extraction.

The multiplexing along the time axis helps avoiding undesired leakage effects in the course of watermark embedding. The multiplexing is facilitated by the convention that no overlapping of the FFT frames was carried out in the previous rMAC computation.

This is also the major contribution beyond an earlier work by the thesis author: in [ZS2009b] it was only defined that spectral coefficients e_{k_0,t_0} at given time-frequency index must not be assigned for both, hashing and watermarking; nevertheless it was permitted to assign different spectral coefficients e_{k_0,t_0} and e_{k_1,t_0} from the same time time index t_0 to either hashing *or* watermarking. Thus, the proposed alignment is unlike other works for example by *Steinebach* or *Gulbis* [SD2003b, GMS2008a, Gul2013].

Example: Figure 3.13 shows a visualization of the alignment scheme: the Figure virtually combines the pseudo-random assignments as shown in Figures 3.3 and 3.10 above. For simplification purposes, coefficients are multiplexed in *"odd/even order"* along the time axis.

In the implementation of this thesis work, the time indices of the *rMAC* extraction will be pseudo-randomly scattered but dense enough to detect even short sections of tampering. As a consequence, the detection algorithm is formally not only dependent on the secret *rMAC* key $\mathcal{K}_2$ but also on the given secret watermarking key $\mathcal{K}_1$.

The proposed alignment not only has an impact on robustness and embedding rate/payload. It also affects the transparency: for achieving a dense coverage of the *rMAC*, the *rMAC*'ed audio frames must be placed dense enough along the time axis. This is vital to avoid that the algorithm misses tampering with the audio data that has a very short duration of less than a second for example. By the multiplexing the watermark embedder is, in effect, *switched off* and *on* again many times per second. This adds a second source of quality loss because an interrupted (virtually stuttering) watermarking distortion sounds more annoying than a *constant* one. Hence, the contribution of this Section has an impact on all test results in the Section 5 after all. An explicit "sanity check" of the alignment scheme can be found in its Subsection 5.4.1.

3.4.2 Embedded Message Payload

The embedded watermark message is determined by the following elements:

rMAC

Obviously, the *rMAC* is an important part of the embedded watermark message payload. An *rMAC* of $M = 128$ *bit* is used which means the main portion of the total payload.

Time Code

> The work in progress described in this paragraph was published initially in [ZS2009b].

As discussed in Section 2.3.4.3, targeted attacks have to be prevented that allow undetected modification by deleting or exchanging complete audio segments (see [HM2000]). For this, as a minor enhancement, a *time code* is embedded in addition to the *rMAC* of each audio segment. For simplicity, the index of the audio segment is used: the first segment is labeled with the integer time code "0", the next one with "1", and so on. This allows future verification that the order of the audio section is correct and no sections are missing. Attacks by deleting or exchanging complete audio segments in terms of a "*Holliman* attack" can be detected. A time code of 8 to 10 *bit* is prefixed to the watermark message.

This quite obvious, simple but indeed effective enhancement can also be found in the works of *Holiman, Petrovic* or *Park* [HM2000, Pet2005, PTW2007]. It is different from the work by *Chen Fan et al.* [CHW2008] which proposes to implement a *linked chain* of feature extraction and feature embedding: the authors describe that the audio features of a given audio segment are embedded into the following segment, and so on. By contrast, this thesis work embeds the time code in its correspondent segment. This has the advantage that every segment inside an audio file is *self-contained* with regards to all meta data required for its authentication, even if only file fragments are available.

CRC Check

A CRC-16 is appended to the watermark message. It allows verifying if the watermark was retrieved correctly.

Forward Error Correction (FEC)

For improving the robustness of the embedded watermark message, forward error correction techniques (FEC) from communication theory are used in this thesis work. Unlike CRC coding, FEC not only allows for detection but even for *correcting* bit errors in the watermark message with a certain probability. Such transmission errors can be caused by distortions due to admissible post-processing (like lossy encoding, analog recording etc.) or malicious acts of tampering with the watermarked audio data. See Section 2.1.2.8 for an explanation and standard references on FEC.

Here, the software implementation uses an adaptation of the standard *Turbo* coding algorithm which was introduced originally in [BGT1993].

> The adaptations and improvements to Turbo coding for watermarking applications were developed in a thesis work by *Berchtold* [Ber2008] under supervision of the thesis author.

In this thesis work Turbo encoding increases the input message length N (net) to the output length of $3N + 4$ (gross). This is caused by internal stages of *redundant convolutional coding, interleaving*, and *puncturing* in the Turbo encoder. Its details are not elaborated upon in this thesis and the interested reader is referred to [Ber2008].

Synchronization Pattern

Each audio segment carrying the watermark message is prefixed by a watermarking *synchronization* sequence. Formally, this is also a watermark message but it is of fixed value, like a start code. The watermark detector at first tries to find and to lock/engage in the known(!) *"Sync"* sequence. If the detection score exceeds a predefined threshold, the subsequent message retrieval and *rMAC* verification is triggered. In the investigated approach a temporal synchronization accuracy on the detector side as precise as ±1 PCM sample(s) can be regularly achieved. This outperforms the accuracy of the original algorithm by *Steinebach* which synchronizes at an accuracy of ± 10 PCM samples. The details are not elaborated upon in this thesis.

Message Payload

Hence the embedded message payload eventually can be written[50] as

$$\{\ \text{Sync} + \text{Message\_1}\ \}\{\ \text{Sync} + \text{Message\_2}\ \}\{\ \text{Sync} + \text{Message\_3}\ \}\ldots$$

with

$$\text{Message\_i} = \mathsf{TurboCoding}\big(\ \text{rMAC\_i} + \text{TimeCode\_i} + \mathsf{CRC}(\ \text{rMAC\_i} + \text{TimeCode\_i}\ )\ \big) \quad .$$

In the experimental evaluation in this thesis work the approximate duration of the "Sync" sequence typically is 2 seconds while the watermark message requires 10 to 15 seconds of duration each.

[50] The "+" symbol shall denote the *concatenation* of binary strings

3.4.3 Implementation Details

In the following, the software implementation of the algorithm will be outlined briefly. The extraction of the *rMAC* and consecutive embedding is conducted as follows (see Figure 3.14 for a simplified flowchart).

1. Input to the software implementation are the cover data file and a text file containing the secret keys (denoted as "Cover Audio Data" and "Secret Key" in Figure 3.14).
2. In addition, a configuration file controlling technical settings for both *rMAC* extraction and watermarking has to be passed to the input (denoted as "Technical Config Data"). The most important technical settings are
   - the frequency bandwidth used for *rMAC* extraction and watermark embedding in *Hertz* (which symbolically defines the "height" of the virtual "area" in the time-frequency domain as visualized in Figure 3.13),
   - the number of frequency coefficients that is used per message bit (which influences the robustness of the watermark and the total duration required for embedding),
   - the percentage of time-indices from which the extraction of the *rMAC* is allowed (which controls how dense the *rMAC* extraction is carried out),
   - the percentage of frequency coefficients at each time-step in the "area" that are allowed to be utilized for embedding (which influences the transparency and the total duration required for embedding),
   - the intended length of the *rMAC* and the time code in *bit* etc.

   From these settings, it can be calculated by simple arithmetic how many frames are required for embedding the complete watermark message (which symbolically defines the "length" of the "area" in the time-frequency domain).

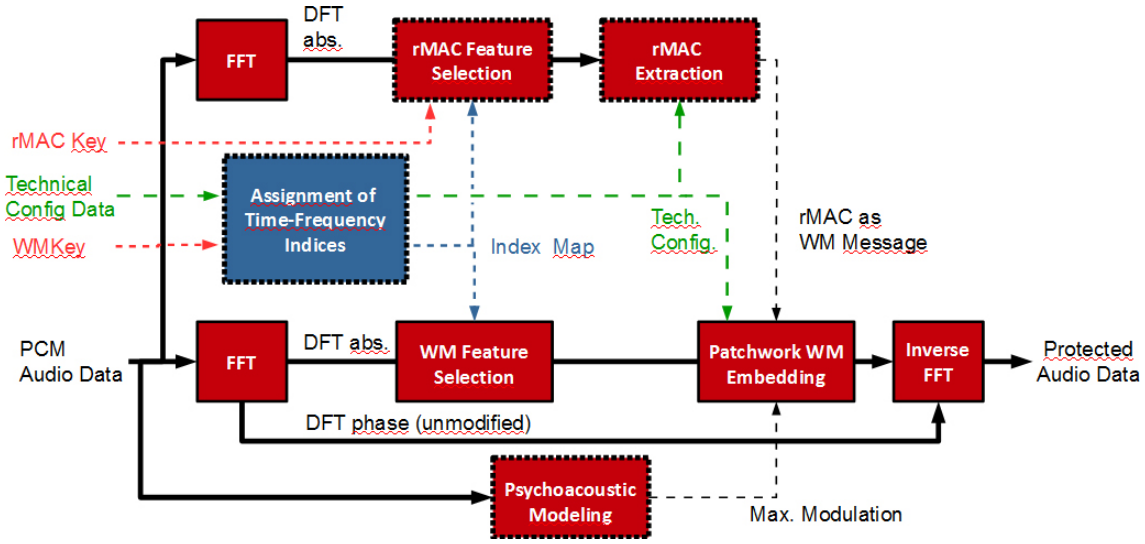


Figure 3.14.: Proposed implementation of *rMAC* extraction during watermark embedding. Thick lines: data flow of audio data; dashed lines: data flow of meta data; dotted boxes: components that were developed or enhanced in this thesis work

3. Dependent on the secret key, the pseudo-random assignment "Assignment of Time-Frequency Indices" is calculated. For this, an array is allocated first which stores the

status of each time-frequency index. In simple words it serves as a ”map” about the status the time-frequency indices (”Index Map”). The ”length” and ”height” of the map is allocated according to the ”Technical Config Data”. Its entries are initialized by setting each entry as *”undefined”*.

4. The pre-defined number/percentage of frames (i.e. their time indices) used for *rMAC* extraction is pseudo-randomly picked first: in the index map they are marked as *”reserved”* for *rMAC* extraction and hence will be skipped during for watermark embedding. All spectral coefficients at these reserved time-steps are marked as potential *rMAC* extraction indices in the index map. In practice, 10 to 40 percent of the total time indices are assigned for *rMAC* computation.
5. For each *rMAC* bit index, the respective set of N indices is pseudo-randomly picked and the *rMAC* bit index is entered into the index map (denoted as ”R*m*” in Figure 3.13). In the *rMAC ”serial mode”*, as proposed in Section 3.2.4), only the frequency indices are chosen pseudo-randomly while correspondent time indices are set in regularly increasing order in sequence according to the *rMAC* bit index.
6. For the majority of leftover time indices, the scattered ”cloud” of watermarking coefficients is pseudo-randomly picked and entered into the index map. For each gross message bit (denoted as ”W*m*” in Figure 3.13) a number of indices is reserved for the Patchwork embedding, according to the Technical Config Data. No time-frequency index is used twice for watermarking and *rMAC* extraction. Eventually, valid states of every entry in the index map represent, in turn, *”*rMAC*’ed”*, *”watermarked”*, or *”unmodified”* (or, in fact, remain *”undefined”* which can indicate program errors at runtime). The index map is then passed as an input to the software libraries for the *rMAC* computation and the watermark embedding.
7. Then the *rMAC* computation is carried on the input audio data in the spectral domain out while respecting the index map. Its output is the input to the watermark embedding algorithm (denoted as ”rMAC as Watermark Message”).
8. The gross message payload is constructed according to the explanation in Section 3.4.2: The *rMAC* and the time code are concatenated with their correspondent CRC code, error correction coding is carried out, and the sync pattern is prefixed. This gross message payload is ”translated” into correspondent index map entries that control which coefficients have to be increased or decreased in the course of Patchwork embedding.
9. The Patchwork embedding is carried out. It has to respect the dynamic psychoacoustic model about the permitted degree of modulation of the DFT absolute coefficients and the static index map and ”Technical Config Data”. The DFT phase coefficients remain unmodified by this algorithm.
10. Finally, the modified DFT spectrum is transformed back to the time domain using the inverted Fourier transform. Final output is the watermarked PCM data which is eventually wrapped into an output *WAVE* file.

The overall software architecture was realized in *C/C++* as follows: An executable implementing the design described above was implemented as a *command line tool*. It serves as a testbed for the different algorithms developed in this thesis work. It imports the developed *rMAC* functionality and the *MPEG* psychoacoustic model as a linked software library. The tool also serves as a parser for the input and output WAVE files, the configuration files and the key files. The tool relies on the following external functionality:

- The commercial watermarking software *SITMark Audio*, developed at the *Fraunhofer Institute SIT*[51]: The proposed embedding algorithm uses the ”Patchwork WM embedding” library of the *SITMark* suite. The same is true for implementing the *rMAC* verification into the *SITMark* detector software (not shown here). The handling of the index map and the ”Assignment of Time-Frequency Indices” were implemented/added into the separate *SITMark* source code for handling repeating utility functions which are used in both watermarking embedding and detection. This includes auxiliary software libraries for WAVE file read/write access and pseudo-random number generation (as source code).
- The 3rd party tool *TooLAME*[52]: This MPEG Audio Layer II encoder internally implements the perceptual model according to [ISO1993b, Appendix D, pp. 5].
- The 3rd party library *FFTW*[53] for performance optimized computation of the Fast Fourier Transform [FJ2005].
- The 3rd party library *Math Kernel Library (MKL)*[54] for performance optimized computation of elementary mathematical operations. In the software this is very useful for the numerous and parallel computations of exponential functions or log functions (during Patchwork embedding or for handling of *Decibel* values).

Using these external optimization toolsets at embedding time, a processing speed of 5x to 10x times faster than real-time playback speed can be achieved. That is, five to ten minutes of input audio data (stereo, 44.1 kHz) require one minute of processing (Standard PC workstation, at present: CPU *Intel Core i5*@3Ghz).

[51] *SITMark Audio* system, version ”AP 2013/06”, source code provided free of charge for academic research by courtesy of *Fraunhofer Institute for Secure Information Technology SIT*, Darmstadt, Germany, *https://www.sit.fraunhofer.de/en/mediasecurity*

[52] *TooLAME* encoder project by *Mike Cheng* under LGPL, *http://sourceforge.net/projects/toolame*

[53] *FFTW* (”...the fastest Fourier transform in the West”), by *Massachusetts Institute of Technology*, Cambridge, USA, under commercial license as part of *SITMark Audio*, *http://www.fftw.org*

[54] *Math Kernel Library (MKL)* by *Intel Corp.*, under commercial license as part of *SITMark Audio*, *https://software.intel.com/en-us/intel-mkl*

Chapter 4

Security of the Proposed Approach against Brute-Force Attacks

This Chapter discusses theoretical security aspects of the proposed audio hashing and audio watermarking algorithms.

An obvious objective for an attacker in integrity and authentication watermarking applications is carrying out data modifications without being detected. Different attack scenarios have to be distinguished

1. With regards to data *integrity* this could be realized by applying the intended integrity breaches in way that they are not detected by the *rMAC* algorithm in the first place. It we assume that the secret *rMAC* key remains undisclosed, this implies:
   - Reconstructing the coverage of the *rMAC*'ed spectral components, especially which time-frequency indices are not covered in the *rMAC* calculation anyway. There, data modifications can not be detected on the basis of the *rMAC* value (but maybe on their impact on the watermark message), or
   - Reconstruction *how* the indeed covered spectral components contribute to the *rMAC* bits, especially which coefficients contribute to a particular *rMAC* bit and how. Attacks on these covered components could be masked by sparing/skipping them during the attack Or slight modifications could be applied to *other* components so that they compensate for the attack in the calculation of an given *rMAC* bit (like the sums in Equation (3.1) .

   The latter would put the attack in a position *as if* he/she had direct access to the *rMAC* key.
2. An alternative integrity attack on given, previously protected content could be realized by attacking the watermark message instead of the *rMAC* calculation. After applying malicious doctoring, the watermark message could be replaced so that it appears to match the *rMAC* of the doctored content again. If we assume that the secret watermark key remains undisclosed this implies, again, reconstructing which and how spectral components are involved in the regular process. This would put the attack in a position *as if* he/she had access to the watermark key directly.

3. With regards to data *authentication*, this knowledge could even be used for creating a forged audio data set, i.e. to make it up completely *from scratch*. The forgery can be prepared so that it contains an embedded watermark message that appears to match the current state of the *rMAC* of the forged data.

Note that the first aspect in the previous list corresponds to the challenges on integrity watermarking described Section 2.3.4.3, especially to the aspects of undetected modification (supported by information leakage and oracle attacks). The second and third aspect correspond to the attacks as described in Section 2.3.4.3, especially removal, de-syncing and oracle/sensitivity attacks.

To summarize, vulnerabilities by targeted attacks on the watermark embedding and *rMAC* extraction, especially the involved pseudo-random feature selection processes should be analyzed. Both will be carried out in the following.

4.1 Brute Force Attack on the Watermark Embedding/Detection

This Section describes the security of the Patchwork embedding approach used in this thesis with respect to brute force attacks. This is done in analogy to the related works by *Cox* or *Bas/Furon* [CDF2006] or [BF2013]. It discusses the actual difficulty for an adversary to get access to the watermarking message. This is expressed in terms of the *effective key length* as outlined in Section 2.3.4.3.

Without knowledge of the secret key, the pseudo-random pattern of modified FFT coefficients in the frequency-time domain is obscure to an adversary in the first place. Instead it has to be guessed directly by an adversary. Formally this means to guess the respective time-frequency *indices* (k,t) from the space of all possible indices[55] $\{(k_n,t_n)\} \in \{1\dots W\}\times\{1\dots T\}$ to get access to the hidden watermark message. The quantity T denotes the duration of the watermark message (measured in frames or discrete time-steps). The value W is the frequency band*width* that is used for embedding. Usually, not the completely available band-width from zero up to the Nyquist frequency is being utilized because very low and high frequencies are not well suited for watermark embedding.

As explained in Section 2.3.5, for each of the M message bits, two disjoint sets

$$A := \{a_{i_n}\} \quad , \quad B := \{b_{j_n}\} \quad , \quad n \in \{1\dots N\} \subset \mathbb{N} \quad .$$

of N FFT coefficients each are pseudo-randomly picked and then modified. The $2N$ coefficients per bit are chosen pseudo-randomly without replacement from $W \cdot T$ possible positions in the frequency-time domain. For the following, not the *value* of the FFT coefficients a_{i_n}, b_{i_n} are relevant, but their respective *indices*

$$i_n := (k_n^{(a)}, t_n^{(a)}) \ , \ j_n := (k_n^{(b)}, t_n^{(b)}) \ \in \{1\dots W\}\times\{1\dots T\} \ \subset \mathbb{N}\times\mathbb{N} \quad .$$

The full set of all indices of coefficients that contribute to a message bit will be denoted as

$$\mathscr{A} := \{i_n\} = \left\{(k_n^{(a)}, t_n^{(a)})\right\} \ , \ n \in \{1\dots N\}$$
$$\mathscr{B} := \{j_n\} = \left\{(k_n^{(b)}, t_n^{(b)})\right\} \ , \ n \in \{1\dots N\} \quad .$$

[55] Note for the avoidance of confusion that the terms "index" or "'indices", resp., correspond to the time-frequency index as a *pair* $(k,t) \in \mathbb{N}\times\mathbb{N}$ consisting of a frequency index k and time index t

4.1.1 Combinatorial Analysis of Full Recovery

At first, it is assumed that a successful attack requires to guess correctly *all* $N+N$ frequency-time indices that are used for each message bit. Then, the attack space Ω for the m-th message bit contains all possible sets of combinations of $2N$ frequency-time indices

$$\Omega = \left\{ (\mathscr{A}_p, \mathscr{B}_q) \,:\, \left|\mathscr{A}_p\right| = \left|\mathscr{B}_q\right| = N \;,\; \mathscr{A}_p \cap \mathscr{B}_q = \emptyset \;,\; \forall p,q \right\} \quad .$$

Each single of the $2N$ indices could have been assigned to set $\mathscr{A}$, or to set $\mathscr{B}$ or to neither of the two. Thus, the above problem can be expressed as the standard combinatorial mental experiment of finding the total number of combinations for

> *"depositing $W \cdot T$ objects in three boxes, with a small number N of objects in 'Box $\mathscr{A}'$, another N objects in 'Box $\mathscr{B}'$ box, and all remaining $W \cdot T - 2N$ objects in the 'third' box".*

For the above standard problem, the total number of elementary events is given by the so-called *multinomial coefficient* from basic combinatorics:

$$|\Omega| = \binom{W \cdot T}{N \;;\; N \;;\; W \cdot T - 2N} := \frac{W \cdot T!}{N!\; N!\; (W \cdot T - 2N)!} \quad .$$

Corollary

The previous expression can be rewritten as

$$\begin{aligned} |\Omega| &= \binom{W \cdot T}{N; N; W \cdot T - 2N} = \frac{(W \cdot T)!}{N!\; N!\; (W \cdot T - 2N)!} = \frac{(W \cdot T)!}{N!\; N!\; (W \cdot T - 2N)!} \cdot \frac{(W \cdot T - N)!}{(W \cdot T - N)!} \\ &= \frac{(W \cdot T)!}{N!\; (W \cdot T - N)!} \cdot \frac{(W \cdot T - N)!}{N!\; (W \cdot T - 2N)!} = \binom{W \cdot T}{N} \cdot \binom{W \cdot T - N}{N} \quad . \end{aligned} \tag{4.1}$$

The last expression corresponds to the mental experiment of the *urn problem* that

> *"N white balls out of total $W \cdot T$ balls have to be drawn without replacement, then another N black balls without replacement from the remaining $W \cdot T - N$",*

which is a standard problem in combinatorics for example for computing the odds in lotteries.

4.1.2 Combinatorial Analysis of Partial Recovery

In the previous Section, the assumption was that only one single combination corresponds to the correct sets $\mathscr{A}$ and $\mathscr{B}$, while all other $|\Omega|$-1 are not. This true, but in practice an attacker does not need to guess and recover the *full* sequence correctly. A small number of incorrectly assigned coefficients will contribute to the detector response to Equation (2.5) as a distortion. Because of the robustness of the Patchwork scheme, the detector can cope with such distortion anyway as with any other distortion caused by lossy compression, analog recording etc. Obviously, that significantly reduces the number of attempts for breaking the scheme. To overcome this security challenge one could increase the total number N of coefficients per message bit. As a drawback this will in return reduce the embedding data rate/capacity of the watermark transmission. These considerations show that robustness, security, and capacity of the watermark are mutually antagonistic.

In the following, it is estimated, how many combinations exist for an *event* $\mathscr{E}$ that represents a successful attack on the message bit i.e. a *partial* recovery of H out of N indices in $\mathscr{A}$ and $\mathscr{B}$ for which $N-H$ "misses" can be tolerated by the adversary (with $0 \leq H \leq N$) .

Upper Bound

An upper bound for the number of combinations that represent a successful attack can be obtained from the following argument: By definition of the embedding/detection algorithm, $N+N$ coefficients' indices are assigned to the two sets $\mathscr{A}'$ and $\mathscr{B}'$. Given by the assumption, the *event* $\mathscr{E}_1$ has to be defined and analyzed that

- a number of H indices is correctly guessed and assigned to $\mathscr{A}$ without replacement; then
- another H indices are correctly assigned to $\mathscr{B}$ without replacement,

which implies that the remaining (N-H) "slots" in A' and the remaining (N-H) slots in B', resp., are assumed to be filled-up by other indices (j,t). The following cases have to be distinguished:

- Case 1: The filled-up coefficient can be picked from indices that are actually unmarked or that belong to a different message bit $m' \neq m$. This represents the event $\mathscr{E}_1 = \mathscr{E}$ and it will be very likely, compared to the other cases: usually there are much more coefficients unused for any message bit than for the m-th bit.
- Case 2: The filled-up coefficient can belong to the m-th bit but is assigned to the *wrong/opposite* set. This case will decrease the detector response as coefficients from the wrong/opposite set will cancel out the embedded spectral feature to some extent.
- Case 3: The filled-up coefficient can belong to the m-th bit and is – by chance – assigned to the *correct* set, too (i.e. it correctly adds to the other H coefficients already in the set) In contrast to Case 2, this case will even improve the detection response, which would be fortunate from the attackers point of view.
- Case 4: ...any combination thereof.

Closer combinatorial analysis shows that the number of combinations $|\mathscr{E}_1(H)|$ for this event can be expressed as

$$|\mathscr{E}_1(H)| = \binom{N}{H}\binom{N}{H}\binom{W\cdot T-2H}{N-H}\binom{W\cdot T-H-N}{N-H} \tag{4.2}$$

This becomes plausible by considering the following

- The first and second term reflect that a partition of H indices were correctly guessed and assigned in $\mathscr{A}'$ and another H in $\mathscr{B}'$.
- The third term reflects that (N-H) indices have to be filled-up to $\mathscr{A}'$. For this, only $r_1 := W\cdot T-2H$ are left from the previous steps ("first fill-up").
- The fourth term reflects that a number of (N-H) indices have to be filled up to $\mathscr{B}'$ ("second fill-up"). For this, only $r_2 := r_1-(N-H) = W\cdot T-2H-(N-H) = W\cdot T-N-H$ coefficients are available, still.

For plausibility it should be noted that for the case $H=N$ it is

$$H=N \Rightarrow |\mathscr{E}_1(H)| = \binom{N}{N}\binom{N}{N}\binom{W\cdot T-2N}{0}\binom{W\cdot T-2N}{0} = 1\cdot 1\cdot 1\cdot 1 = 1 \quad ,$$

i.e. there is only one combination for which all $H \equiv N$ coefficients are correct, as already said.

Lower Bound

A lower bound for the number of combinations that represent a successful attack is given by the following argument: The previous Section described a case-by-case analysis of the "filled-up' coefficients. For estimating a lower bound, the *event* $\mathscr{E}_2$ is assumed that the filled-up indices are all picked in Case 1, i.e. only at indices that *actually* neither belong to $\mathscr{A}$ nor $\mathscr{B}$ which is very likely.

Example: From the figures in the example in Section 4.1.1, it can be seen that this restriction is a feasible assumption: In the experimental evaluation, the frequency bandwidth is approximately $W = 500$ indices (out of 1024) which represents approximately half of the spectrum. The typical duration of the watermarked sequence is $T = 200$. That is, $W \cdot T \approx 100,000$.

If H=25 correct coefficients out of N=30 total are regarded as sufficient for detection, another $5 + 5$ coefficients have to be filled-up to $\mathscr{A}$ and $\mathscr{B}$. It is much more likely that these will be picked from the (100,000-60) indices that neither belong to $\mathscr{A}$ nor to $\mathscr{B}$ than (by chance) from the 10 missing/remaining free slots that actually have belonged to $\mathscr{A}$ or $\mathscr{B}$.

The number of available coefficients for the first fill-up then reduces to $W \cdot T$-2N, the second fill-up reduces to $W \cdot T$-2N-(N-H) = $W \cdot T$-3N+H. Thus, the following expression provides the lower bound for the brute force attack:

$$|\mathscr{E}_2(H)| = \binom{N}{H}\binom{N}{H}\binom{W \cdot T - 2N}{N-H}\binom{W \cdot T - 3N + H}{N-H} \tag{4.3}$$

Result – Summary for Lower and Upper Bounds

If the pseudo-random patterns in the frequency-time domain are equally distributed, the elementary events in Ω are equally likely. Combining Equations (4.2) and (4.3), the probability p_m for a successful attack on the m-th message bit, if H out of the N coefficients in each group are assumed to be guessed correctly, can be narrowed down to

$$\begin{aligned} \frac{|\mathscr{E}_2(H)|}{|\Omega|} \leq p_m \leq \frac{|\mathscr{E}_1(H)|}{|\Omega|} \\ \frac{\binom{N}{H}\binom{N}{H}\binom{W \cdot T - 2N}{N-H}\binom{W \cdot T - 3N + H}{N-H}}{\binom{W \cdot T}{N;N;W \cdot T - 2N}} \leq p_m \leq \frac{\binom{N}{H}\binom{N}{H}\binom{W \cdot T - 2H}{N-H}\binom{W \cdot T - N - H}{N-H}}{\binom{W \cdot T}{N;N;W \cdot T - 2N}} \end{aligned} \quad . \tag{4.4}$$

Corollary – Asymptotic Behavior

For plausibility note that it one sets $H = N$, it is

$$\begin{aligned} H = N \quad &\Rightarrow \quad \frac{\binom{N}{N}\binom{N}{N}\binom{W \cdot T - 2N}{0}\binom{W \cdot T - 2N}{0}}{\binom{W \cdot T}{N;N;W \cdot T - 2N}} \leq p_m \leq \frac{\binom{N}{N}\binom{N}{N}\binom{W \cdot T - 2N}{0}\binom{W \cdot T - 2N}{0}}{\binom{W \cdot T}{N;N;W \cdot T - 2N}} \\ &\Rightarrow \quad \frac{1}{|\Omega|} \leq p_m \leq \frac{1}{|\Omega|} \\ &\Rightarrow \quad p_m = \frac{1}{|\Omega|} \quad , \end{aligned}$$

as can be expected: if it is demanded that all coefficients ($H = N$) need to be correctly guessed, only one out of $|\Omega|$ combinations is correct.

For completeness note that for $H = 0$ it is found (using Equation (4.1)) that

$$\begin{aligned} H = 0 \quad &\Rightarrow \quad \frac{\binom{N}{0}\binom{N}{0}\binom{W\cdot T-2N}{N}\binom{W\cdot T-3N}{N}}{\binom{W\cdot T}{N;N;W\cdot T-2N}} \leq p_m \leq \frac{\binom{N}{0}\binom{N}{0}\binom{W\cdot T}{N}\binom{W\cdot T-N}{N}}{\binom{W\cdot T}{N;N;W\cdot T-2N}} \\ &\Rightarrow \quad \frac{1\cdot 1\cdot \binom{W\cdot T-2N}{N}\binom{W\cdot T-3N}{N}}{\binom{W\cdot T}{N;N;W\cdot T-2N}} \leq p_m \leq \frac{1\cdot 1\cdot \binom{W\cdot T}{N;N;W\cdot T-2N}}{\binom{W\cdot T}{N;N;W\cdot T-2N}} \\ &\Rightarrow \quad \frac{\binom{W\cdot T-2N}{N}\binom{W\cdot T-3N}{N}}{\binom{W\cdot T}{N;N;W\cdot T-2N}} \leq p_m \leq \quad 1 \quad . \end{aligned} \tag{4.5}$$

That would be the probability that the attacker guesses *all* coefficients wrong, by chance.

4.1.3 Numerical Examples

The success probability is calculated for typical settings of the proposed integrity watermarking[56]. Assumed is that the frequency bandwidth is $W = 500$ which represents approximately half of the spectrum if the sample rate is 44.1 kHz. It is furthermore assumed that the duration of the watermarked sequence is $T = 200$ which represents approximately 9 seconds duration in this case. The coefficient groups A and B are assumed of size $N = 30$ each.

Full Recovery

According to Equation (4.1) up to $\binom{100{,}000}{30;30;100{,}000-60} \approx 1.396 \cdot 10^{235}$ attempts are required to guess *all* N indices in $\mathscr{A}$ and $\mathscr{B}$! This corresponds to the theoretical limit by the effective key length of $-\log_2(1/10^{235}) \cong 781\ [bit]$.

Partial Recovery

If only H=25 out of 30 coefficients are regarded as sufficient for detection, according to Equation (4.4) the success probability p_m can be narrowed down to

$$\frac{\binom{30}{25}\binom{30}{25}\binom{99{,}940}{5}\binom{99{,}935}{5}}{\binom{100{,}000}{30;30;99{,}940}} \leq p_m \leq \frac{\binom{30}{25}\binom{30}{25}\binom{99{,}950}{5}\binom{99{,}945}{5}}{\binom{100{,}000}{30;30;99{,}940}} \quad .$$

The numerical result is

$$1.0035 \cdot 10^{-179} \leq p_m \leq 1.0045 \cdot 10^{-179} \quad .$$

The upper and lower bounds in this example are within margins of less than ± 1%. This shows that the simplifying assumptions for obtaining the lower bounds for brute-force attacks are of minor relevance only. In addition, using Equation (4.5) it can be shown that the probability of missing all $2N$ slots is at least 0.964.

[56] All numerical results in this Section calculated in *WolframAlpha* or *Mathematica*, by *Wolfram Research* *http://www.wolframalpha.com*

Finally, the maximum number of brute force attacks is $1/p_m \approx 9.955 \cdot 10^{+178} \approx 2^{594}$ combinations. That corresponds to an *effective key length* as defined by *Bas* and *Furon* of

$$-\log_2(p_m) \cong 594\ [bit] \quad .$$

This expresses the very high difficulty to get access to the watermark message if *all* time-frequency indices of watermarked coefficients need to be estimated. Note that the figures can be expected to be *even better* because the assumption about the bandwidth W and the duration T are rather conservative. In the experimental evaluation as described in Section 5.1.2) the actual figures are 5 to 20 percent greater than the above assumptions.

4.1.4 Additional Remarks

In addition, a few comments on the results of this results need to be made:

- Note that the distributions of detection scores for unmarked versus marked content do not fully separate (recall Figure 2.12). Hence, if an attacker attempts to recover the watermark, he will very likely be irritated by seemingly present message bits caused by false alarms. This will even increase the efforts for an attacker.
- The previous calculation is only exactly true for the first of the M message bits. As every coefficient (its time-frequency index, resp.) is used only once for embedding, the number of available indices for consecutive message bit indices is increasingly reduced.
- The latter is especially significant if the coverage of watermarked time-frequency indices is very high. For example, in the empirical evaluation of this thesis work, a common coverage is 99%: Assume that $M-1$ message bits could be recovered for what ever reason, i.e. all message bits except the last one. Then, the coefficients for this very last message bit needs to be estimated only from the remaining 1%. Nevertheless, according to the numerical example, the value of $N_2 := 0.01 \cdot WT$=1,000 is still a large number. In this case, the success probability for N_2'=1000 and H=5 can be narrowed down to $3.1007 \cdot 10^{-17} \leq p \leq 3.4486 \cdot 10^{-17}$ which represents an effective key length of at least 260 *bit*.
- The previous elaborations apply to the attack scenario that the unmarked cover audio is not available. In the opposite case, an attacker can identify the 99% watermarked indices by comparison of the cover with the watermarked content (e.g. subtraction in the Fourier spectrum). Nevertheless, this would ”only” allow sorting out the 1% unmarked indices and excluding them from the attack calculations. The attacker still would need to face $0.99 \cdot WT$ possible indices for estimating the location of the first message bit.

4.2 Brute Force Attack on the *rMAC* Extraction

According to Equation (3.1) in the proposed robust hashing approach, each hash bit depends on N quantized and standardized Fourier coefficients $e''_{i_1},...,e''_{i_N}$ where N' is an even number, $N' = 2 \cdot k$, $k \in \mathbb{N}$. Their respective indices will be denoted again as

$$i_n := (k_n, t_n) \in \{1...W\} \times \{1...T\} \ , \ n = 1...N' \quad .$$

4.2.1 Combinatorial Analysis

As was pointed out in Section 3.2.1, the construction of the feature selection in the *rMAC* algorithm and the Patchwork embedding are formally quite similar. This is can be seen from the expression in Equation (3.2) which can be generalized (while renaming the indices[57]) as

$$d'_{k,t} = \underbrace{\sum_{i=1}^{N'/2} e_{k_i,t_i}}_{\sim \mathscr{A}_p} - \underbrace{\sum_{j=N'/2+1}^{N'} e_{k_j,t_j}}_{\sim \mathscr{B}_q} \tag{4.6}$$

As a result, the estimation of the effort of brute force attack efforts on the *rMAC* will share similar properties to those on the watermark message. Especially, the estimates as in Equation (4.4) apply for the calculation of brute force attacks on *rMAC* as well.

Nevertheless, *in practice* there is a big difference in the technical settings of the watermarking embedding versus *rMAC* extraction: Preliminary results show that for embedding, each of the groups has to hold a few dozen coefficients each (typically $N \geq 50$ per group). By contrast of the watermark embedding, the analyzed *rMAC* quantity depends on a small number of coefficients. Preliminary results show that rather $N' \leq 8$ coefficients per group provide sufficient sensitivity to integrity breaches.

For analyzing the consequences, the above considerations have to be refined: if it is assumed that a number of H_1 indices is picked correctly in the first set of coefficients and H_2 in the other. Then it can easily be shown that the estimates of upper and lower bounds from Equations (4.2) and (4.3) can be be generalized to

$$\left|\mathscr{E}'_1(H_1,H_2)\right| = \binom{N'/2}{H_1}\binom{N'/2}{H_2}\binom{W\cdot T - H_1 - H_2}{N'/2 - H_1}\binom{W\cdot T - N'/2 - H_1}{N'/2 - H_2} \tag{4.7}$$

$$\left|\mathscr{E}'_2(H_1,H_2)\right| = \binom{N'/2}{H_1}\binom{N'/2}{H_2}\binom{W\cdot T - 2N'/2}{N'/2 - H_1}\binom{W\cdot T - 3N'/2 + H_1}{N'/2 - H_2} \tag{4.8}$$

[57] Note: For the avoidance of confusion about the notation, *rMAC*-related counters are denoted as N' while watermarking-related counters have been denoted with N. For example, in the original algorithm [HOK2001b] N'=4 coefficients are processed. By contrast, in the original watermarking algorithm [Ste2003] or its extension as proposed in Section 2.3.5.1 the quantity "N" in Equation (2.5) denotes the total number of coefficients *per group* (hence $2N$ coefficients total are processed in watermarking).

As a result, this can be summarized to

$$\frac{\left|\mathscr{E}_2'(H_1,H_2)\right|}{|\Omega|} \leq p_m \leq \frac{\left|\mathscr{E}_1'(H_1,H_2)\right|}{|\Omega|}$$

$$\frac{\binom{N'/2}{H_1}\binom{N'/2}{H_2}\binom{W\cdot T-N'}{N'/2-H_1}\binom{W\cdot T-3N'/2+H_1}{N'/2-H_2}}{\binom{W\cdot T}{N'/2\ ;\ N'/2\ ;\ W\cdot T-N'}} \leq p_m \leq \frac{\binom{N'/2}{H_1}\binom{N'/2}{H_2}\binom{W\cdot T-H_1-H_2}{N'/2-H_1}\binom{W\cdot T-N'/2-H_1}{N'/2-H_2}}{\binom{W\cdot T}{N'/2\ ;\ N'/2\ ;\ W\cdot T-N'}} \quad , \tag{4.9}$$

which is an extension to Equation (4.4), while making an adaption of the notation about N and N', resp.

4.2.2 Numerical Examples

The success probability of a brute force attack on the *rMAC* is calculated for typical settings of the proposed integrity watermarking[58]. It is again assumed that the duration of the watermarked sequence is $T = 200$ frames.

The following **Examples 1-3** assume that the rMAC calculation is carried out in the rMAC "scattered" operation mode. The reader is reminded that this means that time-indices can be any random value between 1 and T. In light of the technical settings in the evaluation in Chapter 5 it is assumed that roughly $W = 300$ out of 1024 coefficients are used and that each *rMAC* bits is derived from $N' = 8$ coefficients.

Example 1 – Full Recovery

If we require that all N' indices of the m-th *rMAC* bit have to be guessed correctly, then only one out of

$$|\Omega| = \binom{60,000}{4;4;60,000-8} = 2.9146 \cdot 10^{35}$$

permutations is the correct one. Hence, the expected success probability of such brute-force attack on the m-th *rMAC* bit is

$$p_m = 1/|\Omega| = 3.431 \cdot 10^{-36}$$

which corresponds to an *effective key length* of $-\log_2(p_m) = 117$ [bit].

Example 2 – Partial Recovery

Then we demand that only seven out of eight *rMAC* coefficients have to be guessed correctly ($H_1 = 4$, $H_2 = 3$). This allows studying the sensitivity of a particular *rMAC* bit dependent on finding the last/eighth coefficient in a systematic manner. We find the following estimation:

$$\frac{\binom{4}{4}\binom{4}{3}\binom{60,000-8}{4-4}\binom{60,000-12+4}{4-3}}{\binom{60,000}{4;4;60,000-8}} \leq p_m \leq \frac{\binom{4}{4}\binom{4}{3}\binom{60,000-4-3}{4-4}\binom{60,000-4-4}{4-3}}{\binom{60,000}{4;4;60,000-8}}$$

$$8.233 \cdot 10^{-31} \leq p_m \leq 8.233 \cdot 10^{-31}$$

[58] All numerical results in this Section calculated in *WolframAlpha* or *Mathematica*, by *Wolfram Research* *http://www.wolframalpha.com*

$$\Rightarrow p_m = 8.233 \cdot 10^{-31}$$

This corresponds to an effective key length of

$$-\log_2(p_m) \cong 99\ [bit] \quad .$$

Example 3 – Fewer indices per rMAC bit

This estimation demonstrates also that it is feasible to use more than four coefficients, in contrast to the original work in [HOK2001b]. If we assume that only $N' = 4$ coefficients are used and that three out of the four correspondent indices need to be guessed ($H_1 = 2, H_2 = 1$) it is found:

$$\frac{\binom{2}{2}\binom{2}{1}\binom{60{,}000-4}{2-2}\binom{60{,}000-6+3}{2-1}}{\binom{60{,}000}{2;2;60{,}000-4}} \leq p_m \leq \frac{\binom{2}{2}\binom{2}{1}\binom{60{,}000-2-1}{2-2}\binom{60{,}000-2-2}{2-1}}{\binom{60{,}000}{2;2;60{,}000-4}}$$

$$3.704 \cdot 10^{-14} \leq p_m \leq 3.704 \cdot 10^{-14}$$

This corresponds to a much small effective key length of only $-\log_2(p_m) \cong 44\ [bit]$.

The increased coverage by using a few more Fourier coefficients increases both the sensitivity (true positives) and the brute-force efforts of an adversary. How this influences the specificity (true negatives and false positives, resp.) at that same time is investigated empirically in the following Chapter.

Example 4 – "Serial Mode"

In example 4, the "serial" mode is used: here, the time-indices of all N' FFT coefficients that contribute to a given rMAC bit are concentrated within a very small range along the time axis. In the worst case, that time index is deterministic and *fixed*. This significantly reduces the number of possible time-frequency indices.

In light of the software implementation of the thesis work and the settings used in the evaluation in Chapter 5 it is assumed that for a given rMAC bit index m, the time indices are expected to be from a know small range of T'=5. Hence, the number of possible indices reduces from $W \cdot T = 300 \cdot 200 = 60{,}000$ to $W \cdot T' = 300 \cdot 5 = 1{,}500$.

According to Equation (4.4), the success probability for the attacker per brute-force attempt increases to $p_m = 1.3666 \cdot 10^{-14}$. This corresponds to an effective key length of $-\log_2(p_m) \cong 62\ bit$.

Chapter 5

Experimental Evaluation

The earlier publications by the thesis author provided test results on particular aspects of

- the proposed *rMAC* approach [ZS2007, ZS2008a, ZS2008b, ZS2008c],
- its integration into authentication watermarking [ZS2009b, ZMS2012], or
- improving the general watermarking performance about robustness versus transparency [SZ2008b].

The objective of this Chapter is to provide *consolidated* test results using the same (and larger) base of test data across all different aspects investigated in this thesis.

The Chapter concludes with a summary of test results and identified guidelines for the practical application of the findings in Section 5.5.

5.1 Test Setup

This Section describes the audio test data used, technical settings of the watermarking and *rMAC* algorithms and the different simulated attacks.

5.1.1 Audio Test Data

The detection success was tested on a set of PCM audio files of different genre and sound quality like pop music, classical music, audio books, interview recordings, camera recording in the field etc. (44.1 *kHz*, 16 *bit*, stereo, 90 test files of 3:00 minutes each, total duration 270 minutes). This covers a wide range of different audio content domains like music or voice recordings at both hi-fidelity and low quality recording and production conditions and environments. The intention is to use a wide test basis set of *representative* sounds for the evaluation:

- Obviously, using voice content (interviews, video sound tracks etc.) is motivated by the discussion in Section 1.1 on protecting digital evidence and our cultural heritage against deliberate doctoring as an intentional, malicious act.
- The motivation for using music content is twofold: breaches of integrity on music data rather caused by unintended acts like data loss, mixing-up data or other careless post-processing operation by mistake. In addition, different music genres allow evaluating the

performance of the thesis result on a wide range of sound characteristics (much wider than voice content provides).

An exhaustive listing of the audio content used can be found in the Appendix B.

For simplicity of the evaluation, in most test runs the audio files were divided into short snippets of 30.0 seconds (i.e. six per test file) so that each can hold *one* complete watermark message (i.e. ten to twenty seconds, depending on technical settings). The total number of the snippets is 540.

5.1.2 Technical Settings

The following technical configuration settings were used:

- The FFT frame size for *rMAC* calculation and watermark embedding is 2048 samples which provides a sufficient frequency resolution of $44,100 \text{ Hz}/2048 = 21.5 \text{ Hz}$.
- The watermark was embedded in the frequency range of 500 to 12,000 Hz. 99% of the FFT coefficients available in this range were watermarked.
- Each watermark message bit is embedded "into" $N = 2 \cdot 75$ Fourier coefficients. From preliminary tests this value was found to be a reasonable trade-off between robustness and embedding rate.
- In contrast, the "sync bits" of the prefixed synchronization watermark sequence are embedded with much higher redundancy settings: for each of the four sync bits, approximately $2 \cdot 2500$ coefficients are watermarked in 10 consecutive frames. In total 40 frames are used which represents approximately 2 seconds (at sample rate 44.1 *kHz*). This tuning for high robustness shall allow at least the detection of sync bits if strong distortions or attacks are applied on the audio material.
- An *rMAC* of 128 *bit* is extracted from each snippet and embedded as a robust watermark. The frequency range selected for *rMAC* extraction is set to 100 to 6000 Hz (which by fact turns into 86 to 5,986 Hz given by technical limitations of the discrete FFT frequency resolution). The chosen frequency range is the most relevant with regards to perceptual integrity.
- If not specified otherwise, the *rMAC* is extracted in the *"scattered mode"* as defined above. Each *rMAC* coefficient depends on $N' = 8$ pseudo-randomly picked Fourier coefficients. This value was chosen in contrast to the original work by *Haitsma, Kalker*, and *Oostveen* [HOK2001a, HOK2001b]. The intention is to aggravate security attacks as it was discussed in Section 4.2.
- A time code of 10 *bit* is included in the watermark message. This allows to distinguish 1024 different protected segments per audio file.
- The *rMAC* shall be extracted from a percentage of 33% of the available audio frames at pseudo-random time indices. That means that two third are used for embedding the final watermark message.

5.1.3 Embedding payload

The embedded watermarking data is determined by the following aspects:

Watermark Message

The major part of the message consists of the *rMAC*. It is appended by the block index as a time code. An appended CRC-16 allows verifying if the original *rMAC* value H_0 was detected *correctly* in the first place. Hence, an embedding payload of 154 *bit* is required for the experiments (see Table 5.1):

	Data length (in *bit*)
rMAC	128
Time code	10
CRC	16
Total	$M_1 = 154$

Table 5.1.: Test settings: net embedding payload

Error Correction

For improving the robustness of the embedded message forward error correction techniques (FEC) as outlined in Section 2.1 were applied.

> Here, the software implementation uses an advanced *Turbo* coder as developed in [Ber2008] (Master thesis by *Berchtold* under supervision of the thesis author).

Given by its technical settings, the actual *gross* length M_2 of the Turbo-encoded message becomes

$$M_2 = 3M_1 + 4 = 466\ bit \quad .$$

This demonstrates that the embedding capacity requirements are much higher in integrity watermarking than it is the case for copyright watermarking applications.

5.1.4 Simulated Attacks

The modifications to the audio content ("attacks") were simulated using 3rd party software for common audio processing operations. For example, the software tools *SoX*[59], *Lame*[60] or *Nero AAC codec*[61] were utilized.

Admissible Data Modifications

A number of attacks on the audio were tested which are relevant because they are legitimate and common in audio processing:

[59] *SoX – Sound eXchange* toolbox, version 14.4.1, under LGPL, *http://sox.sourceforge.net*

[60] *Lame* MP3 codec, version 3.97, under LGPL, *http://lame.sourceforge.net*

[61] *Nero AAC Codec*, version 1.5.4.0 (encoder) 1.5.1.0 (decoder), under free license for *"personal non-commercial...purposes"*: *http://www.nero.com/deu/company/about-nero/nero-aac-codec.php* (retrieved January 2016, now offline)

- **MP3/AAC compression**: obviously, lossy compression is widely applied in multimedia production, processing, distribution/broadcast and archiving. In the experimental evaluation, *MP3* and *AAC* compression at different bit rates were applied as legitimate actions. Before watermark detection the data was decoded back to uncompressed PCM/WAV format.
- **Requantization**: increasing the bit resolution is a common task in professional sound engineering studios or in media archives. In the experiments, the 16 *bit* input data was converted to 24 and 32 *bit* resolution and eventually back to 16 *bit* before watermark detection.
- **Resampling**: changing the temporal resolution is a common task in audio post-processing too. Also in the course of lossy compression such resampling might be carried out by an *MP3/AAC* encoder internally and automatically. Moderate resampling down to 32 *kHz* or up to 48 *kHz* was carried out, followed by resampling back to the original sample rate of 44.1 *kHz*. Note that 32 *kHz* resampling includes a low-pass filtering near 16 *kHz* for anti-aliasing i.e. it is a somewhat lossy operation.
- **Soft noise**: For simulating thermal and quantization noise in the course of DA/AD conversion, very soft white noise at -80 *dB* and at -70 *dB* average power level was added. For example, adding -80 *dB* noise is in accordance with the recommendation of the *IHC Committee* for simulating DA/AD conversion [IHC2016], see Section 2.3.4.

Simulated Tampering

For simulating manipulations of the audio content that tamper with the semantic content that one can imagine from an adversary, the following attacks were applied on the test data:

- **Deletion**: A short segment of each test snippet was deleted. The deletion was carried out at fixed positions, for example 4.0, 8.0 or 12.0 seconds from the file beginning. Different durations were tested (from 1/8 s up to 8.0 s). This simulates erasing of audio content from a recording.
- **Replacement**: A short segment of the test snippet was copied first and then used for replacing audio content of equal duration (up to 8.0 s). It was inserted *in-place* (i.e. total snippet duration is maintained) in the same file but at different position, for example 4.0, 8.0 or 12.0 seconds off. Different durations were tested. This simulates a typical so-called *copy-and-move attack*. For example, in voice content this represents assembling a forged spoken statement re-using voice content of the same speaker.
- **Audio Mixing**: As an alternative to the previous attack, audio content was re-used by *mixing* it into the same file but at different position, for example 4.0 seconds off (up to 8.0 s). For example, this simulates adding a speaker into a conversation or adding voice-over while maintaining a consistent background noise or room environment.
- **Strong Noise**: A short segment of the snippet was mixed with significantly audible noise (at -20 *dB*) of different duration (up to 8.0 s). This simulates that parts of the content are ”deleted” by rendering them intangible. This can be used for masking the deletion by pretending a technical problem during recording or transmission.

Note that these attacks are quite *synthetic* as they do not reflect syntactic or semantic context. For example, word boundaries in a voice recording or the bar structure in a music song are not respected. Nevertheless, this can be tolerated for the experimental evaluation: the simulated attacks should be equally detectable in a particular frame no matter if the attacked frames fits into its context in a seamless manner or not.

5.2 Quality Metrics

Apart from standard error metrics like false positive and false negative rates, the following criteria are used to assess the performance of the proposed approach.

Watermark Message CRC Checks

About the integrity and availability of the detected watermark message, the following cases are distinguished

1. The message is retrieved completely and its CRC code can be verified correctly, denoted as *"CRC correct"* in the results below.
2. The message is retrieved completely and its CRC check fails due to bit errors in the detected message that even the Turbo coding cannot cope with (denoted as *"CRC failed"*).
3. It can occur that *no (complete) watermark* can be detected. A self-evident reason for this case could be that the attacks on the audio are so intense or that the file was truncated that only the sync watermark is detectable but not the complete watermark message (denoted as *"Watermark lost"*).

For further study of this evaluation, also the bit error rates of the retrieved messages could be evaluated (because in the simulations the actual values of the watermark messages are known).

rMAC Bit Error Rate

In the general content-fragile watermarking model, a potential forgery will be indicated by comparing the *rMAC* value of the *original* and assumed *true* value H_0 of the audio (as provided by the watermark message) with the current value of the *rMAC* H_1 at the detector side. As error metric the *bit error rate BER* β is used, i.e. the Hamming distance between H_0 and H_1 divided by the bit length M.

Note that the occurrence of the observation that the BER is *non-zero* can be caused by different effects:

1. In case of *"CRC correct"*, the attacks on the audio content could introduce distortions which cause bit errors to the *rMAC* so that $H_1 \neq H_0$.
2. In case of *"CRC failed"* the observable watermark message is not H_0 but a slightly different value H_0'. Then, it is very likely that it can be observed that $H_1 \neq H_0'$ and the BER will become non-zero too. It has to be admitted that false negatives can occur if the altered *rMAC* matches the defective message again *by accident*.
3. In case of *"lost watermarks"* there is no pair of rMAC's available for comparison in terms of BER. Hence, the following convention is proposed:

   > **Proposal:** A missing *rMAC* due to a lost watermark will be treated by setting the BER *manually* to the value "0.50".

   Albeit *arbitrary*, this convention is nevertheless reasonable: Any intense integrity breach shall cause an BER of approximately 0.5 anyway. Hence a BER value of $\beta = 0.5$ shall reasonably be treated as an "alarm" anyway. In addition, this convention enables a *unified* representation of all three cases in the test results below.

rMAC Flag Ratio

About localizing the temporal position of an attack, the *rMAC* flag ratio α as defined in 3.2.3 is used.

Objective Difference Grade (ODG)

Automated sound quality assessment in terms of ODG vaules as described in Section 2.4.2 was carried out used in the experiments. The results were obtained using the commercial *OPERA*[62] system. According to its users manual, the *OPERA* system internally implements a cognitive psycho-acoustic model in terms of masking or *cochlear domain* representations for assessing sound quality degradations. At first, the model parameters are estimated from the input *reference file* and the input *test file*. Then it rates the files by their parameters' difference using a 3rd party artificial neural network toolbox[63]. The neural network was trained with results obtained from elaborate listening tests with human test subjects. The sound quality loss is finally expressed in terms of ODG values ranging from -4.0 (*"very annoying"* sound difference) to 0.0 (*"no audible difference"*), see Table 2.5.

Note that the PEAQ metric reflects human perception much more accurate than the widely used "signal-to-noise" (PSNR). The latter is a purely numerical comparison of

- the energy of the difference signal caused by the watermark(ing) (i.e. original PCM signal minus post-processed signal, sample-by-sample), versus
- the energy of the original input signal.

It takes none of the psychophysical properties of human perception into account. Hence, a "good" PSNR value is not a sufficient condition for the watermark being imperceptible.

Also the well accepted ITU standard of *"perceptual evaluation of speech quality (PESQ)"* cannot be used as it is only suitable in the limited scope of band-limited speech data (sample rate: 16 *kHz* at most) as in telephony[64]. Here, *intelligibility* (i.e. tangibility of the spoken words) is considered instead of *fidelity* (i.e. acoustic similarity of the "sound").

[62] *OPERA* audio quality analysis system by *Opticom GmbH*, Erlangen, Germany, *http://www.opticom.de*

[63] *OPERA* uses artificial intelligence toolbox *"Skynet"* by *Cyberdyne Systems Corp.*, Sunnyvale, CA, USA. See the work by *Cameron et al.* for more detail [CH1983].

[64] *ITU* Recommendation P.862, *http://www.itu.int/rec/T-REC-P.862*

5.3 Evaluation of Watermarking Embedding/Detection

At first, extensions to the embedding/detection scheme as presented in Section 3.3 were evaluated.

5.3.1 Embedding Capacity

From the technical configuration settings given above it can be derived by simple arithmetic that 255 frames in total are needed for embedding the net watermark message and its synchronization prefix. This represents a duration of approximately $T = 11.8$ s (at sample rate 44.1 *kHz*). Hence, the net embedding data rate is $M_1/T \approx$ 12.9 *bit*/s. The correspondent gross embedding rate is $M_2/T \approx 39.0$ bit/s. This includes that 1/3 of the audio segment has to remain unmarked because of the alignment/multiplexing scheme.

For comparison: the core algorithm as in [Ste2003], which was used as technical starting point, provides a net embedding rate of nearly 2 bit/s if tuned for comparably robust copyright or transactional/forensic watermarking.

The increased embedding data rate mainly benefits from the proposed mapping of FFT indices to watermark message bits in the time-frequency domain allows making use of the increased frequency bandwidth in a flexible manner. It also allows to cover the available sub band *more densely* (up to 99% of available indices) than in the original algorithm without causing transparency issues.

5.3.2 Watermarking Robustness

Robustness to admissible attacks

With the given settings, the watermark is very robust to the *admissible data* modifications explained before: approximately 98.5% of the watermarks can be detected correctly for the *"re-quantization, re-sampling"*, and *"soft noise"* attack. Only one message out of 540 (i.e. 0.18%) cannot be retrieved *at all* and is regarded as "lost" for the verification. For only very few watermarks (2%) the respective CRC verification failed unfortunately (see Figure 5.1, upper plot).

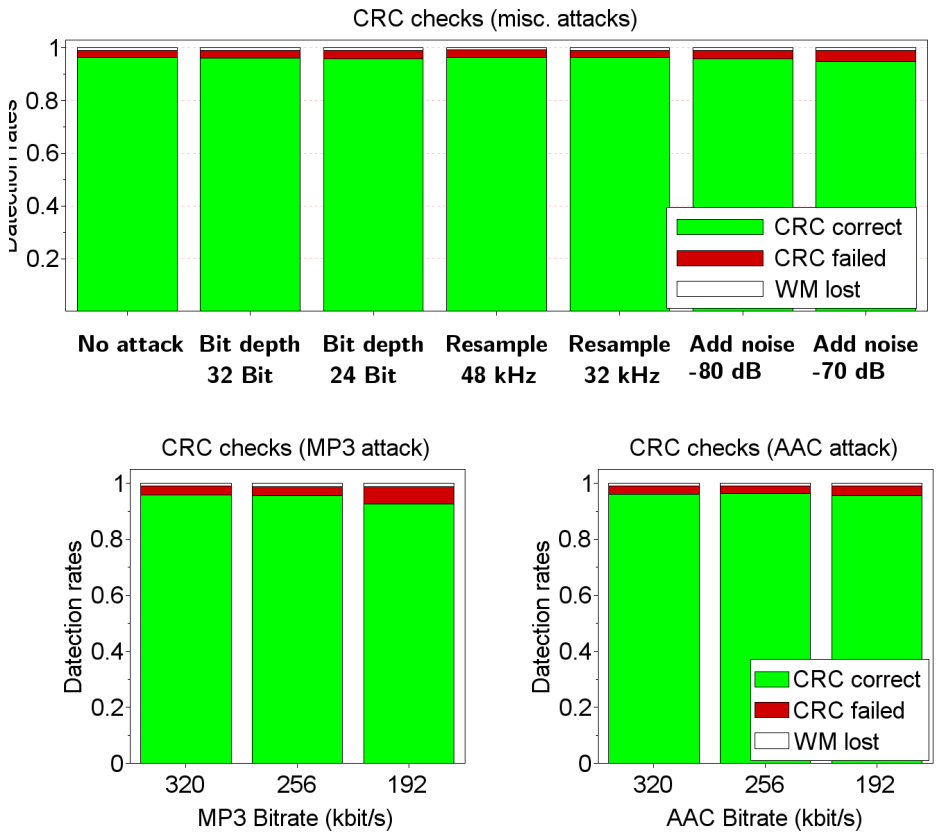


Figure 5.1.: Test result: watermark CRC results – robustness to different kinds of admissible post-processing operations.

For the *"MP3"* and *"AAC attack"* at hi-quality encoding settings (minimum 192 *kbit/s*, stereo) most of the watermark message can still be detected and retrieved correctly again (see Figure 5.1, lower plots). In the Section 5.3.5 below it will be shown that the watermark is robust to *MP3* encoding even at much lower bitrate and sound quality, resp., However, the rate of failed CRC verifications increases to 3 to 7% dependent on the *MP3/AAC* bit rate (most CRC failures obtained for *MP3* attacks at 192 *kbit/s*)

Robustness to malicious attacks

The watermark is also capable to some extent for withstanding many of the *malicious attacks* defined above. The attacks *"audio mixing"*, *"noise mixing"*, and *"replacement"* up to one or two seconds do not drastically reduce the detection success (see Figure 5.2, upper plot and lower right plot). Only for longer duration an increasing portion of watermarked FFT coefficients, and hence message bits is affected (which even the Turbo error coding cannot cope with). Finally, no reliable detection can be expected for an attack duration of 8.0 seconds. This can be expected because the majority of the watermarked audio frames of the 12-second watermark is affected.

It has to be admitted that *no* successful detection can be observed for the *"deletion"* attack. This can be explained as follows: for this attack, not only the audio data of the deleted section is affected. In fact the whole remaining part of the audio file from the spot of deletion until the end of the file is *shifted* to earlier time indices. For example when one second of audio data is removed at position 0:04 *min*. from a file of 30.0 *s* total duration, the remaining 25 seconds will show an offset of -1.0 s relative to their previous position. The detection will mostly fail then

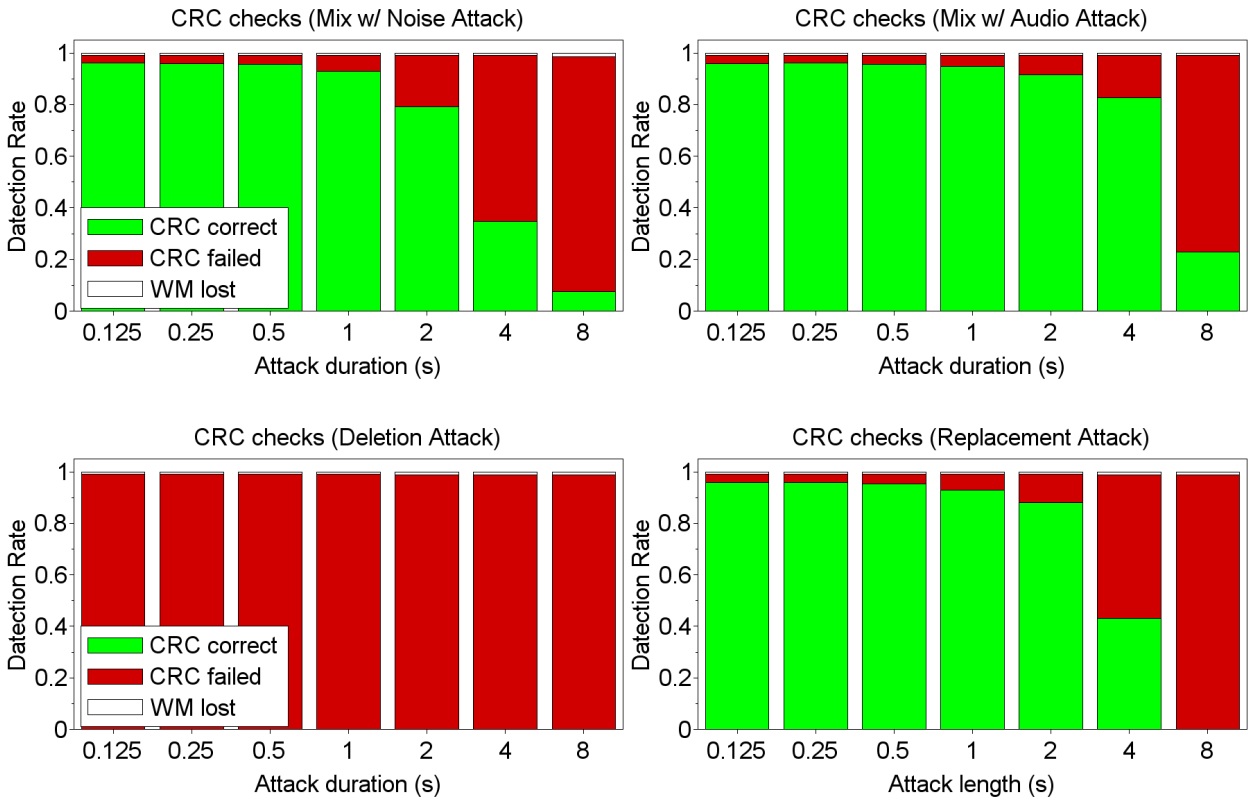


Figure 5.2.: Test result: watermark CRC results – robustness to malicious attacks applied at *pos*=4.0 *s*; from upper left to lower right: *"mix noise"*, *"mix audio"*, *"deletion"*, *"replacement"* attack

(see Figure 5.2, lower left plot) because too many FFT coefficients will be out of their expected temporal position. The same would be observed if audio data was inserted (which is omitted in this simulations).

In another test run the malicious attack were not carried out 4.0s from the file start but right at the file start instead (*pos*=0.0 s). This is relevant in view of the fact that the beginning of the watermarked audio snippet carries the "sync" watermark sequence. It consists of four sync message bits and its duration is approximately 2.5 seconds. Here, the number of lost watermarks increases as can be expected (see Figure 5.3). For example, the *"deletion"* attack prevents almost any detection of the watermark. For the other attacks, the percentage of lost watermarks increases significantly from an attack duration of 0.50 s which is roughly the duration of each sync bit.

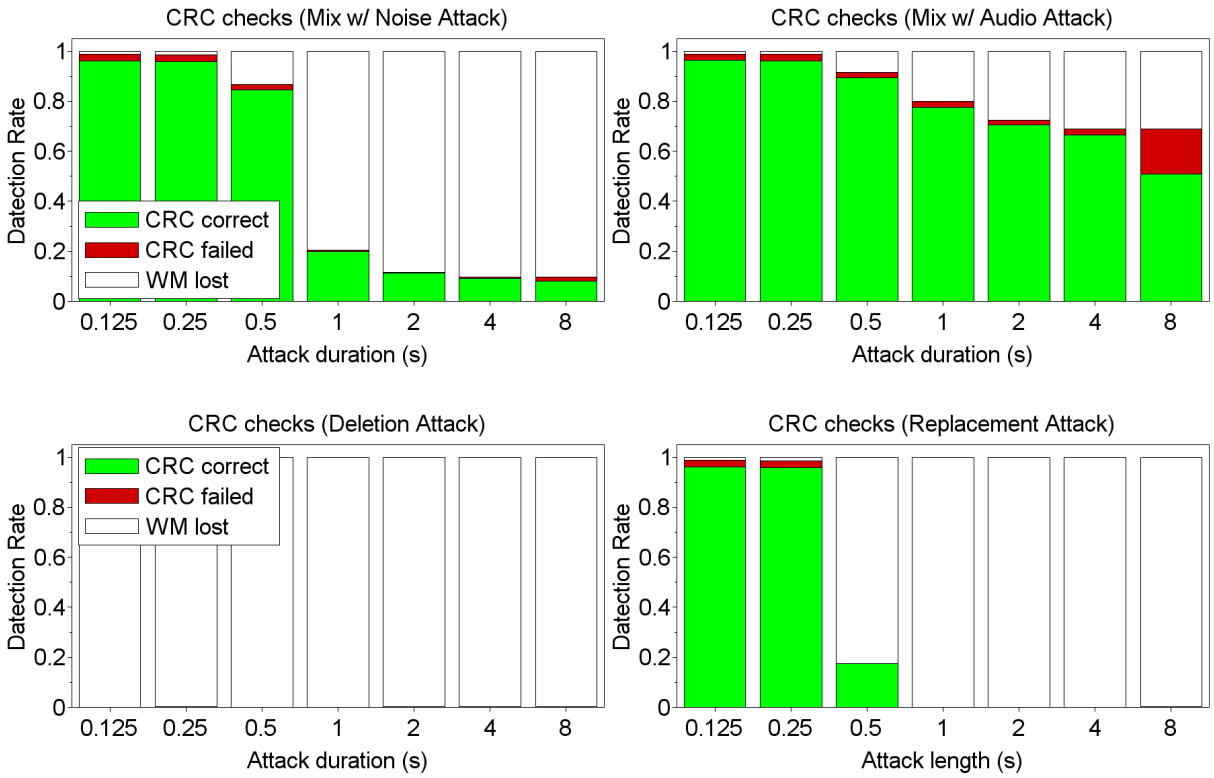


Figure 5.3.: Test result: watermark robustness to malicious attacks applied at *pos*=0.0 *s*; from upper left to lower right: *"mix noise", "mix audio", "deletion", "replacement"* attack

As expected the audio watermark is much more vulnerable in the sync sequence than for attacks applied at other file positions. This behavior is not optimal because missed detections do not allow a precise assessment of the severity of the integrity breach in terms of BER (which will be set manually to 0.5 anyway). Nevertheless, the result allows indicating the *presence* of an integrity issue that should be investigated after all.

5.3.3 Synchronization Precision

What could also be observed is the very high precision of the temporal synchronization position. In un-attacked files the starting sample position of the protected audio segment (namely *pos*=0.0 s) was identified correctly except for three outliers. These outliers were observed from the set of the critical snippets mentioned above in which watermarks were lost or CRC checks failed. The same result could be observed for the admissible and malicious attacks. Synchronizing precisely is vital for avoiding false discoveries due to a temporal misalignment of the authentic *rMAC* with its current value.

Finally, recall that the results in this Section only describe the properties of the watermark embedding. In principle, results could be applied for using the embedding algorithm beyond integrity watermarking. The *rMAC*/watermark combination is investigated in Section 5.4.

5.3.4 Transparency of Watermark Embedding

The results presented in this Section present an evaluation on the transparency of the proposed audio watermarking approach.

The technical settings for embedding (and detection) were the same as for the robustness test above. Then the sound quality degradation caused by the watermarking process was evaluated using the *OPERA* sound quality assessment system. The test was carried out on a *subset* in the test set that contained music files only. A number of 180 snippets (taken from thirty different music files) were evaluated. Total duration represents 60:00 *min*. of duration.

The test had to be limited to music files for the following reason: assessment in terms of the PEAQ sound quality model (see Section 2.4.2) only applies in a meaningful manner if the reference audio content is available in sufficiently *hi-fi* quality in the first place. In other words: an assessment of sound quality on highly distorted, band limited or noisy content is not covered in the PEAQ model. The model neither applies for assessment of the intelligibility of voice data. Closer analysis showed that the *OPERA* system actually created very misleading and inplausible results for *low quality* content like noisy microphone recordings.

The main results obtained using the *OPERA* sound quality assessment system show that a very good sound quality could be achieved for most of the files: Apart from a few outliers, the sound quality of the watermarked audio snippets was assessed as being ”better” than an ODG value of −0.5 (see boxplot in Figure 5.4, left). This is in the transition of difference grades −1.0 (*”audible difference, not annoying”*) and 0.0 (*”no audible difference”*) which is close to the bound of being perceivable at all. The respective median is 0.15 which is an acceptable result.

However, outliers in the results can be observed in which the *ODG* is ”worse” than −1.0 (see Figure 5.4, right).

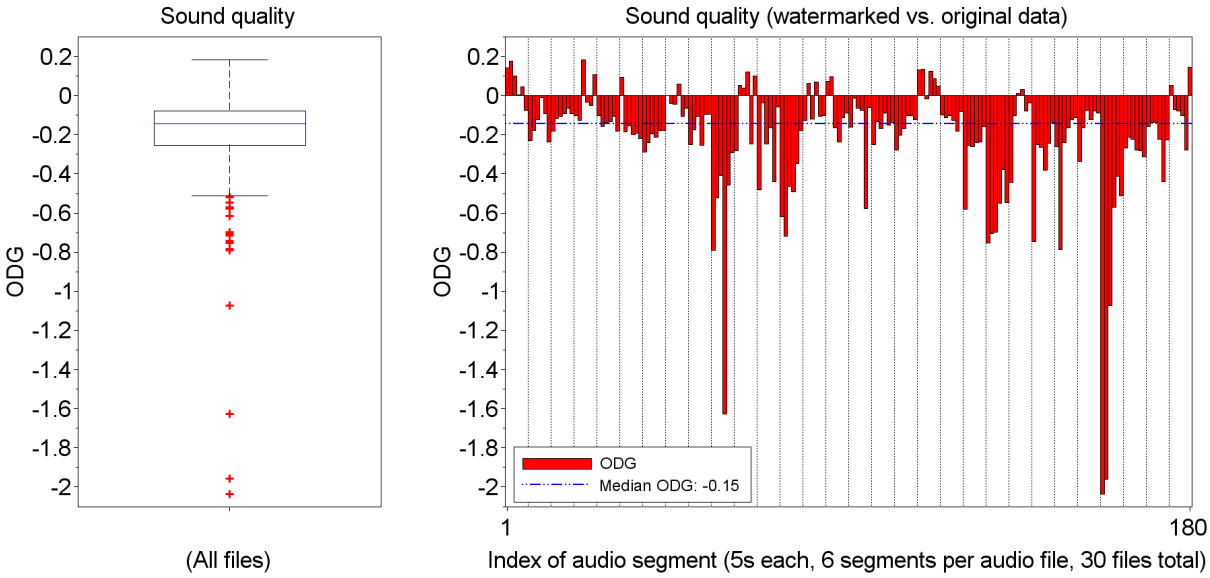


Figure 5.4.: Test results: watermark transparency; left: *ODG* boxplot[65] across all (music!) files; right: *ODG* versus audio snippet index; dashed lines: each group of six snippets correspond to the same test audio file

Closer analysis shows that these outliers are observed in audio snippets from two particular music files:

- One example is a music snippet (labeled as *"M10"* in the list of test data, see Appendix B) that has contained a lot of distortions already in the cover data. The distortions consist of drop-outs, periodically repeating noise, and occasional click artifacts caused by earlier faulty CD ripping. This seems to irritate the *OPERA* system.
- The other example is a music track (*"M27"*) which contains only very low frequency organ sounds: no significant input energy can be observed in the mid and high frequencies.

Especially the second example causing outliers indicates that not all audio content is suitable for watermark embedding with regards to transparency. The absence of sound energy in mid and high frequencies does not allow for hiding the watermark by means of auditory masking with reasonable availability, robustness, and transparency.

Surprisingly, a few watermarked audio snippets are assessed with an *ODG* value *greater* than zero. According to the technical manual of the *OPERA* system such error of measurement has to be expected and accepted. When its artificial neural network was trained by human test subjects, inaccurate and, in effect, *wrong* assignments occurred. Put in simple words: positive *ODG* values reflect that also human listeners would rate the watermarked audio a little "better" than the cover data *by mistake*.

Finally, for completeness note that the *ODG* results obtained were carried out in an early stage of the experiments in which the watermarked frequency band had been set to 0.5-15 *kHz*. The robustness results presented before were obtained with recommended settings of 0.5-12 *kHz*. Hence, transparency results can be expected to even a little better as presented in this Section because with the recommended smaller upper bound of 12 *kHz*, a greater portion of the full spectrum remains untouched.

5.3.5 Psychoacoustic-enhanced Detection

The results presented in this Section refer to the proposal as presented in Section 3.3.2.

As a minor research activity, also the watermark detection was improved by utilizing psychoacoustic modeling not only in the embedder but in the detector algorithm too. In addition to the *MP3* encoding attacks at high quality so far, also encoding at stronger compression settings was carried out at 80, 64 and 48 *kbit/s*, stereo. Especially the "dynamic" approach as explained in Section 3.3.2 was evaluated (denoted as *"Psy Detect: Mode 3"*).

The comparison with results obtained without psychoacoustic functionality (denoted as *"Psy Detect: off"*) shows that the detection significantly benefits from the proposed approach. This is mainly true for medium-quality *MP3* settings at 64 and 48 *kbit/s*. For these bitrates, the detection success of correct CRC increases from 30% to 80% (for 80 *kbit/s*) and from 10% to 60% (for 64 *kbit/s*), resp. The enhanced approach even allows for a non-zero detection success for *MP3*

[65] Boxplots (introduced by *Tukey* [Tuk1977], a.k.a. *"box-and-whisker plots"*): the lower and the upper 25 % quantile of the data samples (denoted as *quartiles* $q1$ and $q3$, resp.) are visualized as the lower and the upper edge of a (thin) rectangular "box" in the samples' domain. Their *median* value is indicated by a (bold) horizontal center line. Outliers are marked with *"+"*-markers. Such outliers are defined (as a convention) as any samples greater than $q3 + 1.5 \cdot \Delta q$ or smaller than $q1 - 1.5 \cdot \Delta q$, resp., in which $\Delta q := (q3 - q1)$ is the *inter-quartile width*. The range of all presumed regular data samples (i.e. except for outliers) are visualized as *whisker* markers.

All boxplots in this thesis created with the *nan*-Toolbox for the *Scilab* programming language by *T. Pettersen/H. Nahrstaedt*, TU Berlin, Germany, *https://atoms.scilab.org/toolboxes/nan*

attacks as strong as 48 *kbit/s* (which is not the case with disabled psycho-acoustic detector). See Figure 5.5 for a visualization of these results.

The increased detection success of complete messages (no matter if *"CRC correct"* or *"CRC failed"*) is a useful feature in authentication watermarking. It allows a better analysis of the impact of audio distortions on the overall detection result: a completely retrieved watermark message allows for the analysis steps described in this work. By contrast, a *"lost watermark"* only allows for treating it according to the "fallback" convention by setting the BER to 0.50 as defined above. True BER assessment or temporal localization of tampering as defined above can not carried out in this case.

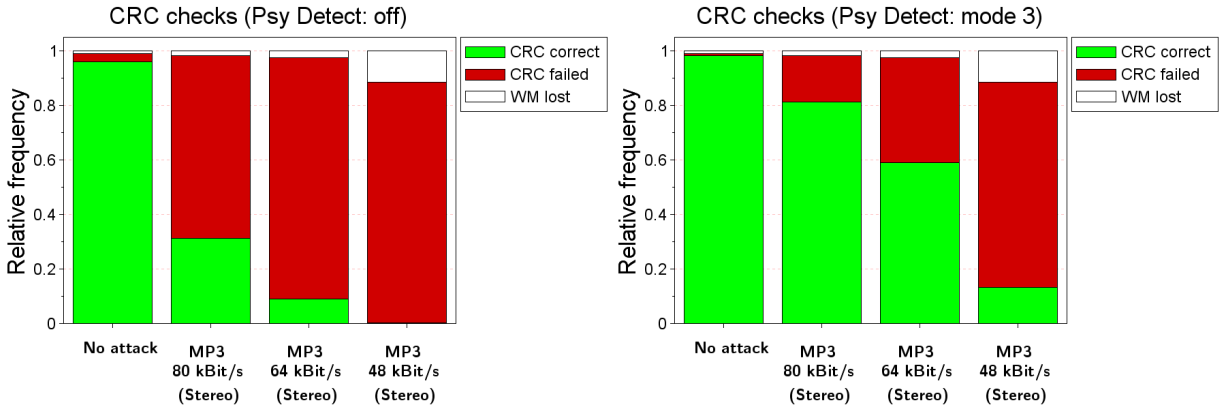


Figure 5.5.: Test results: watermark robustness: detection success after strong *MP3* compression at lower bitrates; left plot: psychoacoustics-based detection enabled; right: ...disabled

In addition, other watermarking applications like copyright or transaction watermarking can benefit even more from this robustness improvement: Complete and correct retrieval of the message is *vital* in order to trace back copyright infringements or information leakage even when the content undergoes very strong attacks. A fallback convention to a failed CRC or a lost watermark does not exist and hence the trace to the dishonest user is eventually "lost" too.

5.4 Evaluation of *rMAC*-based Audio Authentication Watermark

This Section investigates the robustness and distinction performance of the *rMAC* extraction and its integration into Patchwork audio watermarking. This represents the main contribution of this thesis work.

5.4.1 Background level

The results presented in this Section refer to the proposal as presented in Section 3.4.1 in which the alignment of *rMAC* extraction and watermark embedding is described.

At first it was investigated which detection results can be observed if the protected audio is not a subject to *any* kind of attack. The authentication watermark is verified immediately after the embedding stage in order to conduct a sort of ”sanity check”.

As a result, for most of the test files the *rMAC* shows no bit errors (see Figure 5.6). Nevertheless, for a share of approx. 3.5% of the snippets (19 out of 540) the respective BER is non-zero, namely 0.50 instead. The occasional occurrence of such false discoveries as ”background noise” can be explained again by the fact that some audio snippets are rather unsuitable for embedding a watermark in a reliable manner. From these snippets, the watermark message cannot be retrieved correctly in terms of their CRC (or no watermarking message can be retrieved *at all*). All critical files show (at least partially) a very low total sound volume or are extremely band-limited in the spectrum.

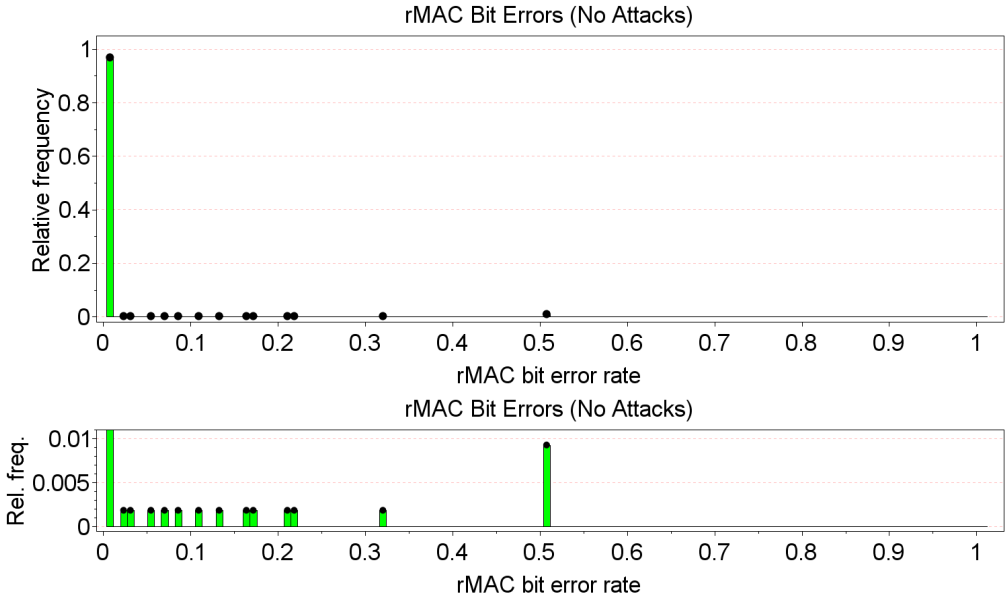


Figure 5.6.: *rMAC* bit Error Rate (BER) – No attack upper plot: histogram of observed bit error rates. lower plot: enlarged view (zoomed to small histogram entries)

One example is the same critical test file *”M27”* that had to be discussed in Section 5.3.4 already: this file shows CRC failures on the retrieved watermarks in all of its six snippets and hence bit errors in the *rMAC* verification. Its lack of signal energy in mid and high frequencies does not allow for reliable watermark detection.

[65] Note: an equivalent visualization result is given as boxplot in the most left column in Figure 5.7

Nevertheless, the test results demonstrate that earlier test results as published in [ZS2009b] could be improved significantly. False discoveries caused by an interference of the watermarking process and the *rMAC* extraction are avoided. For supporting this result a test run could be carried out (for further study) that compares the false discoveries with the alignment switched *on* versus *off* using the thesis test set.

Anyway, the remaining false alarms that are still observable are not caused by such interference in the first place but rather by natural limitations of digital watermarking by itself in the case of critical audio material.

5.4.2 Robustness to Admissible Modifications in Audio

BER for Different Technical Attack Settings

In the next stage of experiments the robustness of the approach to admissible attacks was investigated. Here, the results with regards to BER of the *rMAC* for the requantization or resampling attacks are found as being similar to the background level *without* attacks. Apart from the (same) outliers, the BER range remains below 0.02 (see Figure 5.7). The BER range for the noise attack are significantly higher: the BER range reaches up to 0.08. Except for the *MP3* attack at 192 *kbit/s* for lossy *MP3/AAC* attacks the BER remains smaller than 0.08, too (see Figure 5.7). Although the large majority of results turn out to be as expected, a number of outliers/false discoveries propagated from the background BER level can be observed though.

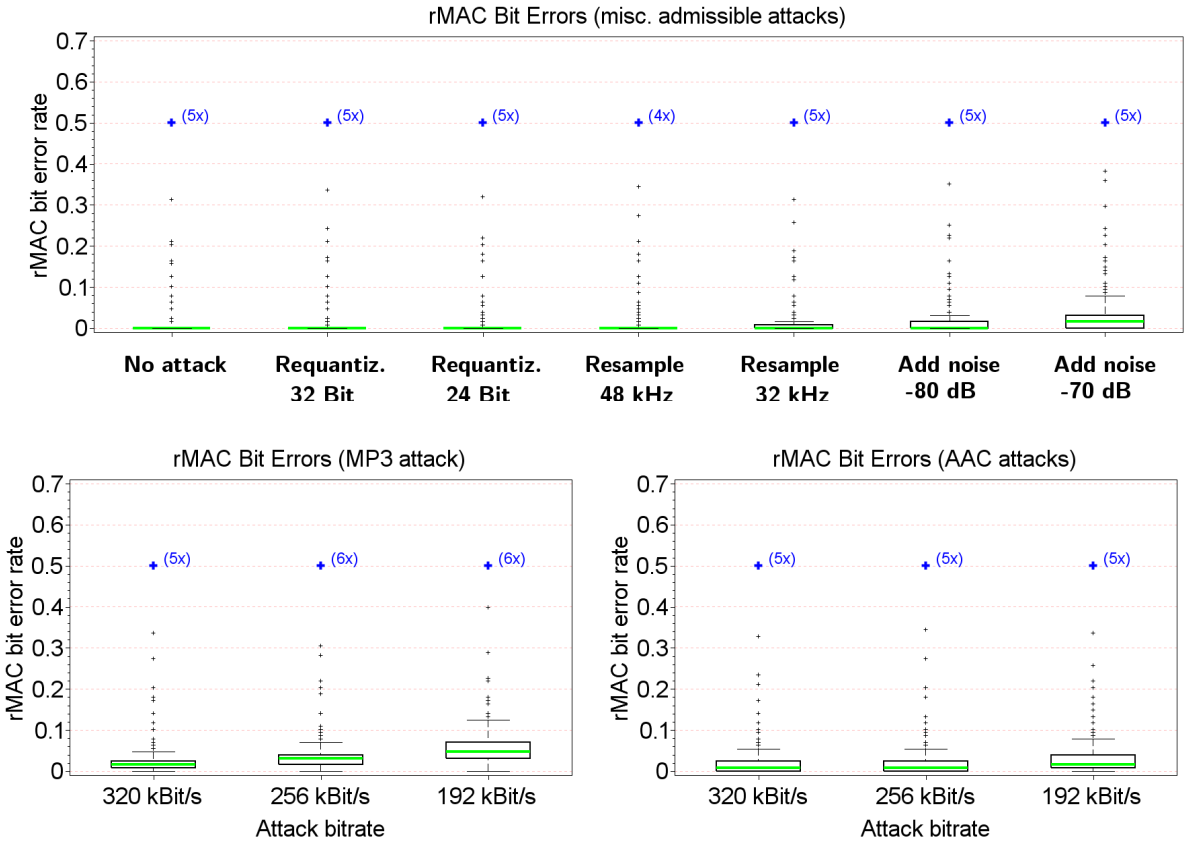


Figure 5.7.: *rMAC* BER – admissible attacks: boxplots of bit error rate; numbers in braces[66]: total count of BER results being exactly 0.5000

BER versus Quality Loss

For a better interpretation of these results, also the *perceived* quality loss caused by the attacks was considered. For this, the BER is compared against the attacks' quality loss in terms of *ODG* values. This allows better comparison across the different kinds of attacks and across the different audio examples.

For example, it is not obvious to the reader which of the lossy attacks *"resampling to 32* kHz" versus "MP3 *encoding at 192* kbit/s" has a stronger impact on the sound quality. But this test setup is also reasonable because the same attack can have different influence on the sound quality dependent on the given fidelity of the input audio or its total volume. For example, adding a very soft noise at -70 *dB* average SPL to very *silent* audio content shall have a stronger influence on the *rMAC* than for *loud* audio content.

The results for the different attacks are found to be as follows[67]:

- Firstly, for *all* "admissible attacks" as defined above, the quality loss in terms of the *ODG* is "better" than -1.0 (except for one single outlier, see below). This means that the range of attacks was designed in a reasonable manner, see upper plots in Figure 5.8.
- The "soft noise" attack remains nearly inaudible for most of the test files. Only for approximately 5% (27 out of 540) of test snippets, the quality loss is worse than *ODG*=-0.2. Here, the *rMAC*-watermarking approach behaves as desired: the more the quality loss of the attack (i.e. the smaller its *ODG* value), the greater is the BER. The correspondent BER remains smaller than 0.10, see Figure 5.8, upper left plot.
- The requantization to 24/32 *bit* (and then back to 16 *bit*) is in fact a lossless operation except for numerical inaccuracies. This is correctly indicated by the *OPERA* system as it does not reduce the perceived sound quality, see Figure 5.8, upper left plot: the *ODG* values remain near 0.0 (or are even positive by mistake). However, the BER can reach up to 0.10 for nearly 20 outliers. Most likely this is caused by numerical inaccuracies in the audio processing tool which was used for requantization. This can influence *rMAC* bits more likely that are "weak" (as defined in Section 2.2.3) occasionally.
- The same is true for the lossy compression to *MP3* and *AAC* as can be verified in Figure 5.8, upper right. To elaborate further on this, also stronger *MP3* compression attacks were conducted. It can be seen clearly that the BER steadily increases with increasing quality loss, see Figure 5.8, lower plot. For example, for attacks at *ODG*=-1.5 or worse (which is at the transition from being *"not annoying"* to *"slightly annoying"*), all BER results are 0.10 or greater, as desired.
- For the resampling attack most attacks have an *ODG* at approximately 0.0, see Figure 5.8, left. However, a few outliers suffer from a notable quality loss showing *ODG* values that can be as "poor" as -2.0. Closer analysis shows that these outliers were caused for resampling at 32 *kHz* (mostly in test files labeled as *"M05"* and *"M18"*, see Appendix B). Note that 32 *kHz* resampling in principle is a slightly *lossy* operation due to the low-pass filtering near 16 *kHz* for anti-aliasing (as explained in Section A.1.2.1. However, the *rMAC* watermarking is invariant to these effects resampling as the BER remains at a value of 0.0 which is a little inconsistent.

[67] For technical reasons slightly different technical configuration settings were used in this experiment. This "BER versus *ODG*" results are not fully compatible with the previous Section but nevertheless demonstrate the sensitivity characteristics of the combined rMAC-watermarking approach.

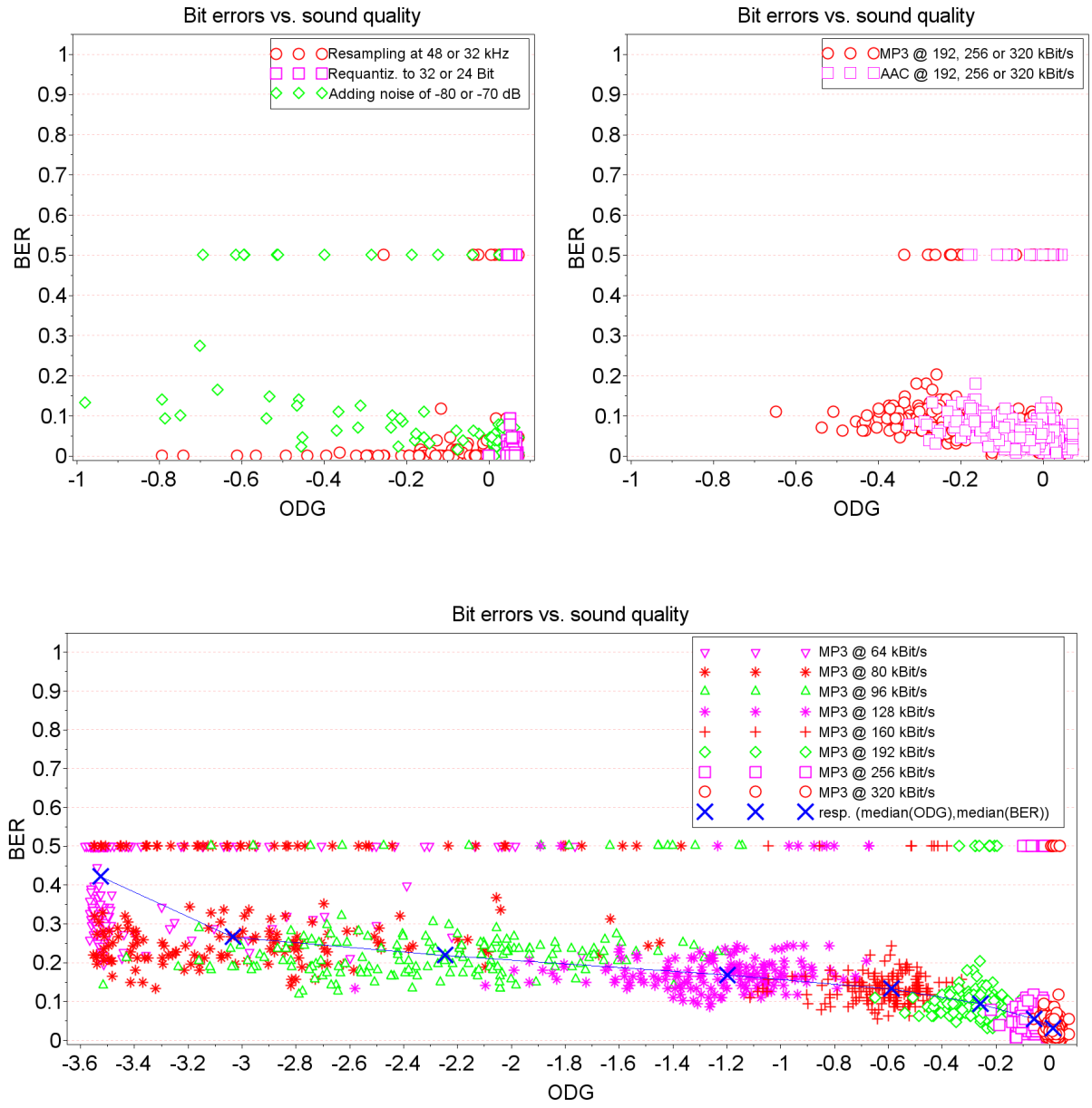


Figure 5.8.: *rMAC* bit error rate versus objective difference grade (*ODG*):
upper Figures: graduated sensitivity for different admissible attacks; note: one single outlier for 32 *kHz* resampling attack at (-2.0;0.0) not displayed;
lower Figure: graduated sensitivity for *MP3* encoding

To summarize, data modifications that cause minor or even no change of the perceived audio quality are causing only few bit errors in the RMAC up to BER=0.10 – if at all. A formal inconsistency for 32 *kHz* resampling is caused by the fact that the upper frequency bound of the *rMAC* extraction and the watermark embedding was intentionally set at small as 6 *kHz* and 12 *kHz*, resp.

5.4.3 Sensitivity to Malicious Data Modification

In the next stage of experiments, the test data was subject to the *malicious* attacks as defined above. The attacks were applied 4.0 seconds from the beginning of the audio snippets in increasing duration (1/8 s to 8.0 s).

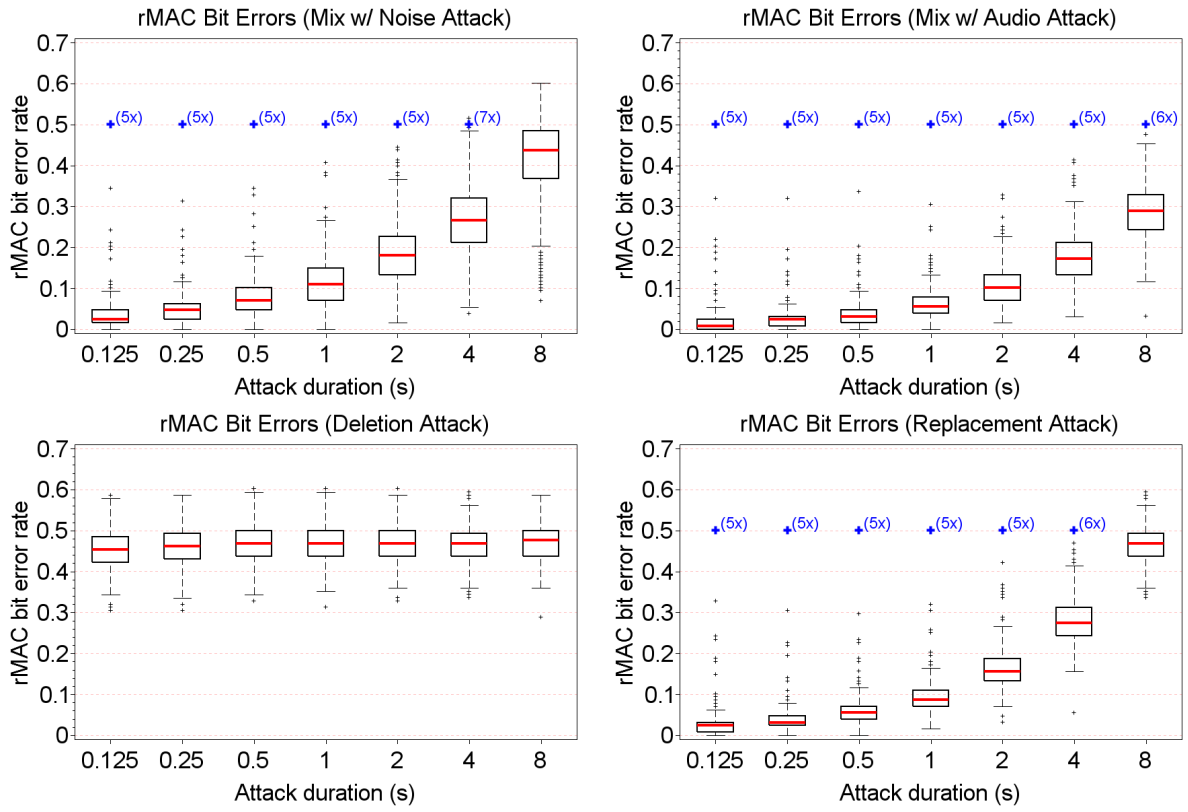


Figure 5.9.: *rMAC* BER – malicious attacks

The test results are found as follows:

- For the *"mix noise"* attacks the *rMAC* shows a graduated sensitivity to the attacks (see Figures 5.9, left): as desired the BER increases if the attack duration increases from 1/8 to 8.0 seconds. Apart from outliers (as defined in *Tukey* boxplots) the BER range increases from [0.00;0.09] to [0.21;0.60]. Up to an attack duration of 1 second the lower bound remains "attached" to zero. Nevertheless, from an attack duration of 4 seconds, the BER distribution clearly separates from the results obtained from admissible attacks.
- The BER results for the *"mix audio"* attack (see Figures 5.9, right) are approximately 1/3 smaller than for the previous *"mix noise"* attack. This is understandable because the volume of the audio sections that are mixed into the audio can vary and can even be zero. By contrast, the all mixed noise sections have constant equal (loud) volume at average -20 Decibel.
- The results of the replacement are fairly similar as for the "mix noise attack" (see Figure 5.9, right). The lower bound of the distribution "detaches" from zero for an attack duration of only 1 second.
- In all cases, a number of outliers can be observed for which the bit error rate (mostly) is significantly higher than the ranges given before. Closer analysis showed that this is

mainly caused by inaccurate retrieval of the watermark message (*"CRC failed", "Watermark lost"*)

- For the deletion attack the sensitivity to behave very differently. For *all* values of the duration, the BER range is approximately 0.3 to 0.6 which strongly indicates the tampering. The seemingly high sensitivity in this case can easily be explained in analogy to the detection results in Section 5.3.1: deletion of even only one second of audio causes a temporal offset in the remaining audio content following the spot of deletion. This affects *all* audio frames after the spot of tampering and hence the extracted *rMAC* bits. In addition. The same result could be expected for a different attack by *inserting* audio content (which was omitted in the simulations for simplicity).
- Closer analysis shows that results are more or less similar, no matter if the correspondent *rMAC* are extracted in series or if they are randomly scattered in time

To summarize, for three out of the four attacks the presented approach does not show a sufficient sensitivity to for an attack duration shorter than two seconds. This can be explained by the fact that such short attacks affect too few Fourier coefficients and hence too few *rMAC* bits.

5.4.4 Optimization of BER Decision Thresholds

Results from the previous Section showed that reasonable decision thresholds are difficult to define. This can be seen by looking at the correspondent histograms. For this, all BER results from the different admissible attacks are concatenated first. The same is carried out for all results malicious attacks. It can be seen that both distributions do not separate very well at small BER values, see Figure 5.10 (upper plot).

As consequence, the objective of identifying malicious attacks of very short duration has to be *relaxed if only the BER is used as decision criterion*:

> **Convention:** For the remainder of this Section the attacks are regarded as malicious only if they are applied for a duration of at least 2.0 seconds.

Under this concession, the correspondent BER distributions separate fairly well again, see Figures 5.10 (lower plot). In either case, a notable peak at BER=0.50 can be seen in the plots of the malicious attacks. They are caused mostly by the particularly different outcome of the *deletion attack* and some contributions by the hard-coded results from "lost watermarks" as well.

An even more meaningful analysis can be obtained from the *receiver operating characteristic (ROC)* curve[68] Under the relaxed objective of identifying malicious of at least 2 seconds, the ROC curve across all possible BER decision thresholds is well concentrated in the left and upper area. The area under the curve[69] is AUC=0.984 which is a good result (see Figure 5.11, right) for a classifier. The best discrimination in terms of minimized sum of false positives and

[68] *"Receiver operating characteristic curve (ROC curve)"*: The ROC curve is a common plotting style for visualizing the performance of a binary classifier. Defined as xy-plot in the unit square of the *true positive rate (TPR)* versus *false positive rate (FPR)* in dependency on the samples' decision threshold. In the terminology of descriptive statistics: equivalent to plotting the *sensitivity* versus (*1-specificity*). For good classifiers, the ROC curve should be concentrated in the far left and far upper section, resp.

[69] *"Area under curve (AUC)"*: The AUC is a relevant parameter of the ROC curve because it is equivalent to the probability that the *rMAC* classifier will rank a randomly chosen *malicious* attack as an "alarm" instead of a randomly chosen measurement from an *admissible* attack or the background level, see [Faw2006]. For good classifiers, the (AUC) should be close to 1.0.

negatives is achieved if the BER threshold is set to $thr_{\max} = 0.078$ (see Figure 5.11, left). This means a Hamming distance of $H_{\max} := 128 \cdot 0.078 = 10$ bit or more. In this case error rates of approximately 4% false positive and negative at the same time have to be expected (values obtained from Figure 5.11, right). If a true positive rate of 100% (i.e. false negative rate of 0%) is desired, a Hamming distance of only 4 *bit* can be tolerated. Closer analysis shows that the majority of false positives are caused by missed/failed detections with an incorrect CRC (BER is set arbitrarily to 0.5 in this case).

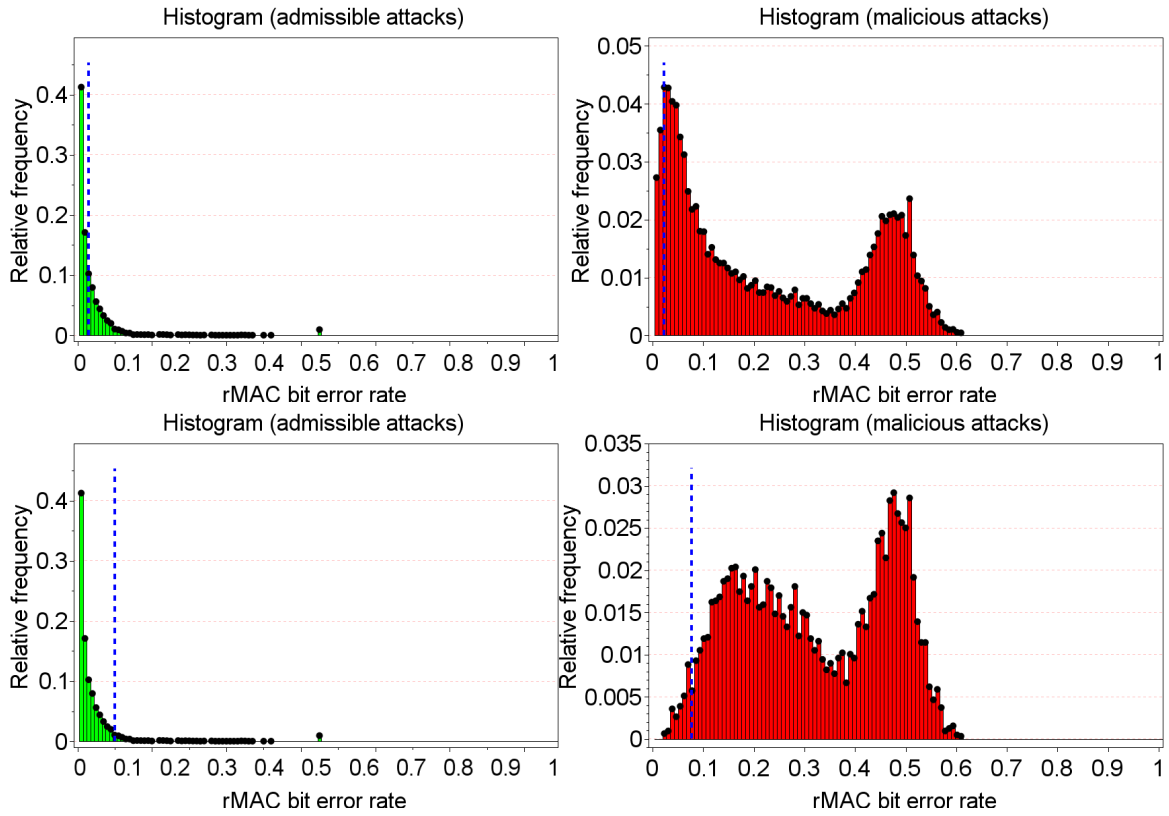


Figure 5.10.: *rMAC* histograms: admissible vs. malicious attacks; upper right: all malicious attacks (duration: 0.125-8 s), lower right: only malicious attacks of duration 2-8 s; black dots: non-zero entries, dashed line: threshold for minimum error rates

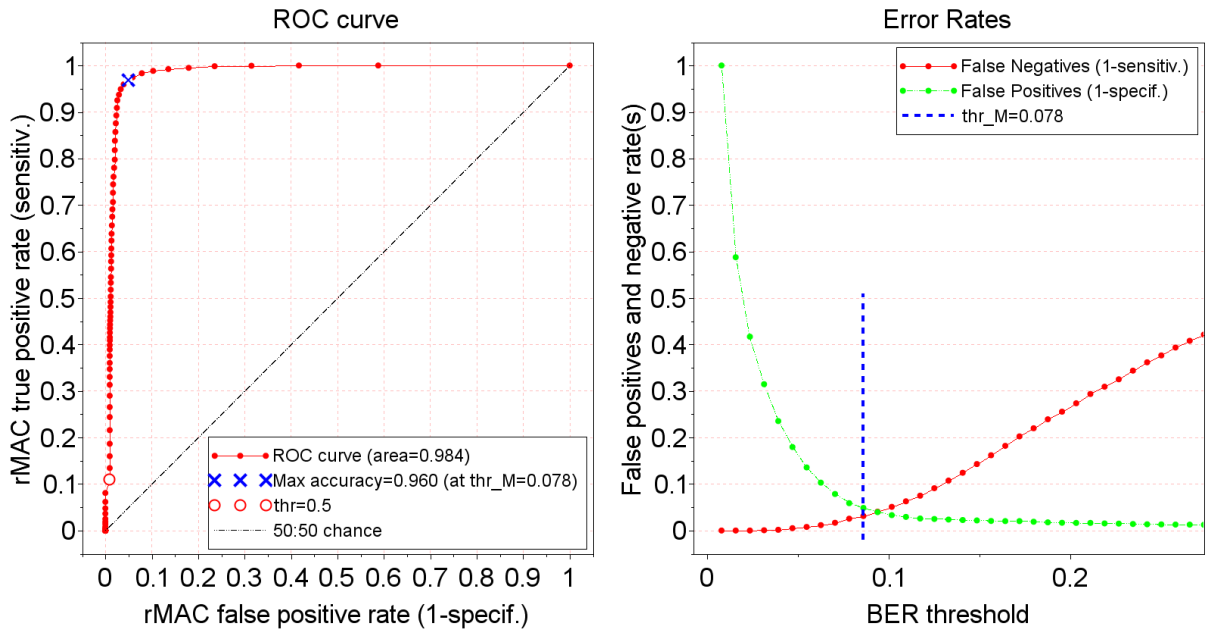


Figure 5.11.: left: ROC curve[70], ×-marker: result for minimum sum error rates (maximum accuracy), legend: note that the given accuracy values assume equal prevalence of admissible and malicious attacks; right: error rates of admissible vs. malicious attacks, dotted line: threshold for minimum error rates

5.4.5 *rMAC* Feature Standardization

The results presented here are referring to the proposal for feature standardization in Section 3.2.2.

For the analysis of the randomness property of the audio hash it was analyzed how the individual *rMAC* bits are distributed. It is found that the relative frequency that a particular *rMAC* bit has value ”1” is approximately 0.5 for all bits, as desired. The correspondent mean is 0.492, see Figure 5.12.

For comparison, a second test run was carried out in which the standardization was *disabled* intentionally. As expected, the *rMAC* bits are much more unequally distributed, see Figure 5.12. This means that without the feature standardization many of the *rMAC* bits would be vulnerable to targeted attacks. This can be seen from the following example: in this second test run the first *rMAC* bit seems to have the value ”0” with a relative frequency of 0.90 (see most left bar in 5.12). Recall that this value was obtained across all different audio snippets. The relative frequency can serve as a reasonable estimator of the probability for any *future rMAC* computation from arbitrary audio data – even audio that was tampered with. Hence, even after a malicious data modification this *rMAC* bit will be ”0” with (same) high probability of 0.9. The attack will finally be missed in most cases. This shows that the standardization is *vital* for obtaining randomly distributed *rMACs* and for maximizing its sensitivity.

[70] All ROC curve plots created with the *nan*-Toolbox for the *Scilab* programming language by *T. Pettersen/H. Nahrstaedt*, TU Berlin, Germany, *https://atoms.scilab.org/toolboxes/nan*

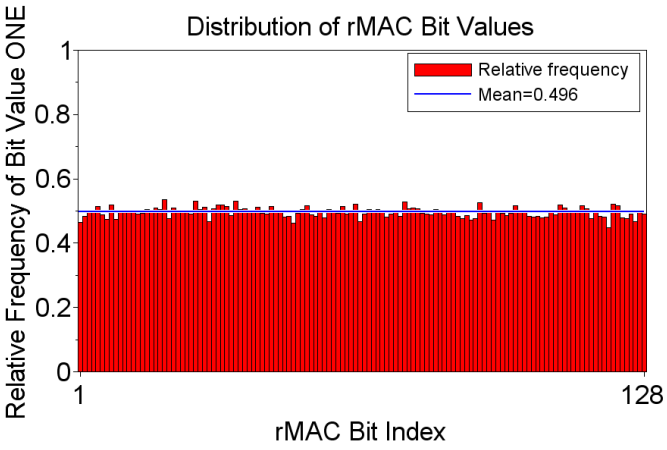


Figure 5.12.: Distribution of *rMAC* bit values (feature standardization *enabled*)

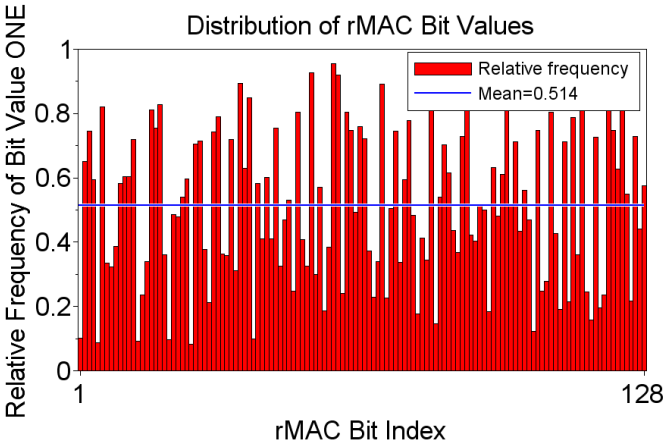


Figure 5.13.: Distribution of *rMAC* bit values (feature standardization *disabled)*

5.4.6 Distinction Performance

In another analysis it was compared how the system can distinguish audio data which are *completely distinct* from one another instead of partially tampered. In this case it is required that the hash values are different and are completely independent. For this, all 540 *rMAC* values calculated at embedding time were mutually compared with one another in terms of their Hamming distance, or BER, resp. Total number of comparisons is $\binom{540}{2} = 145,530$.

As expected, the BER distribution is concentrated around the center: The average BER is 0.500 (see Figure 5.14). This result shows that *rMAC* bits in this case are identical only *by accident* which is a necessary requirement for mutual independence.

Note that this very positive result was enabled not before the feature standardization step (as evaluated in Section 5.4.5) was developed. Without it the BER distribution of actually distinct content would not be centered around 0.50 but at a smaller value. An explicit evaluation of this was not conducted for simplicity. The interested reader is referred to the explanation in Section 3.2.2: Especially Figure 3.5 visualizes earlier preliminary results in which the average BER drops to approximately 0.30 if standardization was *disabled*.

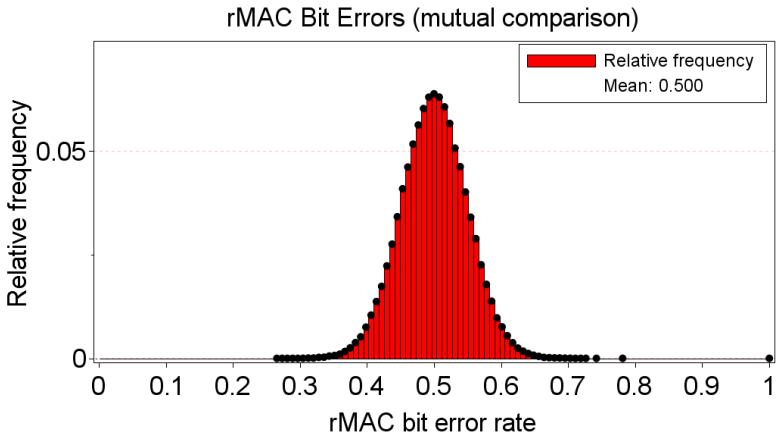


Figure 5.14.: Test result: *rMAC* BER comparison for mutually distinct audio content

5.4.7 Temporal Localization of Tampering

This Section provides results about the capabilities of the approach with regards to localizing a potential integrity breach.

Note that the temporal localization of audio data modification is implemented in two ways:

1. The pseudo-random assignment of *rMAC* bits to time indices can be evaluated so that the position of the modification can be identified as described in Sections 3.2.3 and 3.2.4 The theoretical lower bound of temporal resolution is *by time index*, i.e. *by frame*, which means a split second.
2. The embedded time code as proposed in Section 3.4.2 allows indicating is a complete watermark message was lost. Hence, the temporal resolution is given by the duration of each watermark message, which means approximately 12 seconds in these experiments.

The experimental evaluation of both cases is given in the following.

5.4.7.1 Localization Approach in "Scattered Mode"

This evaluation corresponds to the proposal in Section 3.2.3.

As a first result, the *rMAC* flag ratio for the admissible attacks is mostly zero, except for a few outliers in every time step, see middle plot in Figure 5.15. This is in line with the results in Section 5.4.2.

Then, an analysis is carried out for the *"mix audio"* attack. This attack is considered for demonstrating the performance because test results so far showed that the algorithm is least sensitive to this attack. Hence results in this Section can be seen as the "worst case" in terms of sensitivity.

For this attack, at the shortest duration (1/8 s) the difference between *rMAC* flag ratios in the actually attacked versus un-attacked audio segments can be seen well, see middle plot in Figure 5.15. In the correspondent ROC curve, the AUC is 0.780 which is still a fair result. If the decision threshold is set to 0.01 (i.e. just above 0.00) a true positive rate of 0.70 and, at the same time, a true negative rate of 1-0.17=0.83 can be achieved.

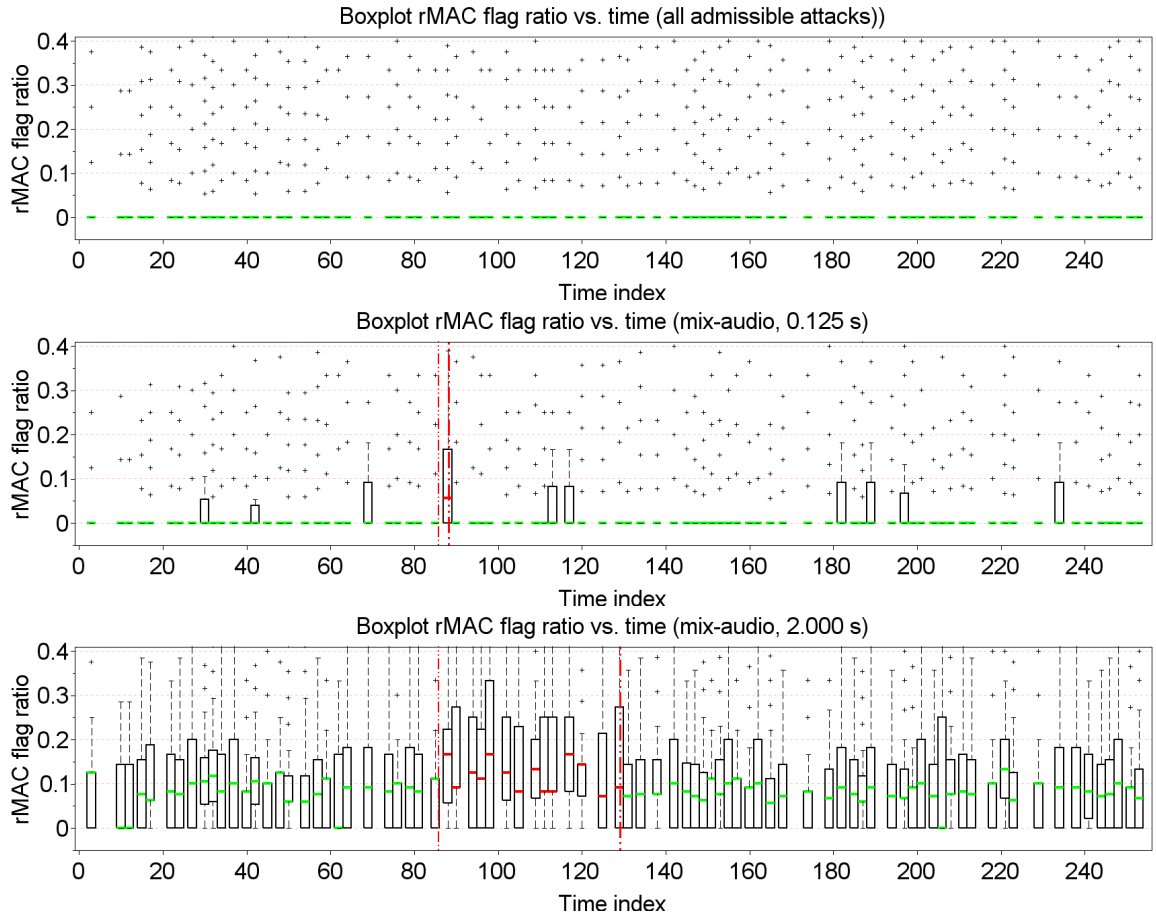


Figure 5.15.: Test result: *rMAC* flag ratio versus time: one boxplot per protected time index; bold center line: respective median value; dashed vertical lines: separator between attacked and un-attacked portions (start: time index 86 = 4.0 seconds)

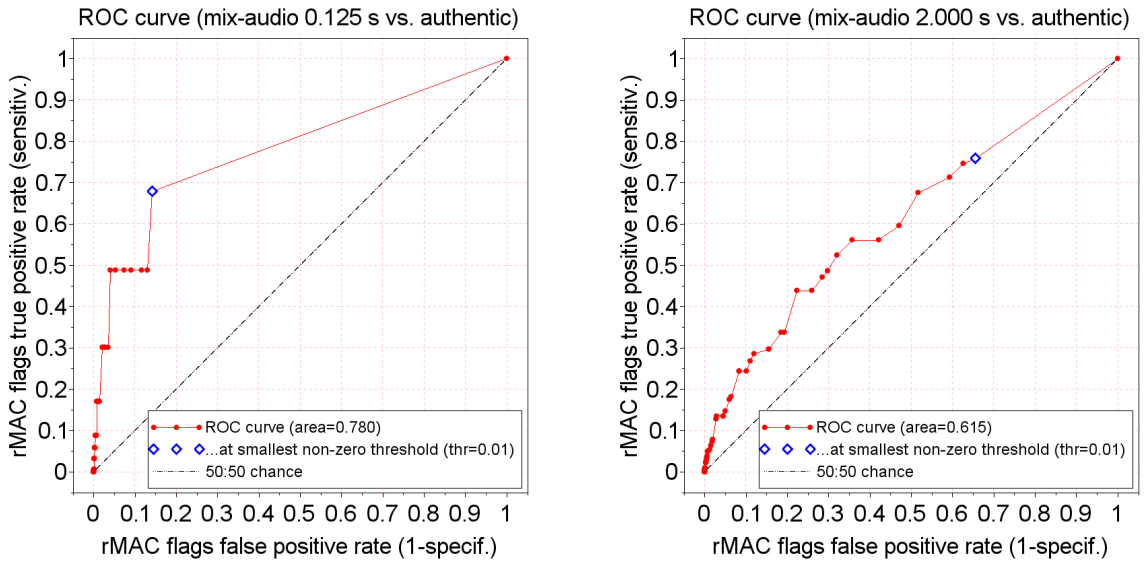


Figure 5.16.: Test result: *rMAC* flag ratio ROC curves; attack duration 0.125 s and 2.0 s ("scattered mode")

In contrast, for a longer attack duration of 2 seconds, the results are quite poor indeed. A lot of false positives are flagged *outside* the attacked section attacks by mistake, see Figure 5.15, lower plot. The correspondent AUC is only 0.615 and the ROC curve rather approaches the diagonal of the unit square. The reason is that the rather long attack duration affects a notable percentage of *rMAC* bits. Hence many time indices are flagged as potential candidates. Because of the pseudo-random scattering, this can accumulate at time indices other than the actually attacked ones *by accident*.

Attack duration	*"mix noise"*	*"deletion"*	*"replacement"*	*"mix audio"*
1/8 s	0.89	0.49	0.88	0.78
1/4 s	0.85	0.49	0.84	0.74
1/2 s	0.77	0.50	0.77	0.68
1 s	0.71	0.50	0.74	0.66
2 s	0.61	0.50	0.66	0.61
4 s	0.55	0.51	0.59	0.57
8 s	0.50	0.52	0.52	0.54

Table 5.2.: Test result: AUC measure; all attacks, all durations ("scattered mode")

Closer analysis shows that the AUC measure for the *"mix noise"* and *"replacement"* attack behaves more or less equally to the *"mix audio"* attack, see Table 5.2. Again, results from the *"deletion"* attack are different. The very low AUC results for *any* attack duration near 0.5 mean that the classification is carried out correctly only *by coincidence*.

The exhaustive collection of time plots and ROC curves across *all* attacks and all attacking durations (up to 8 s) can be found in Appendix C.

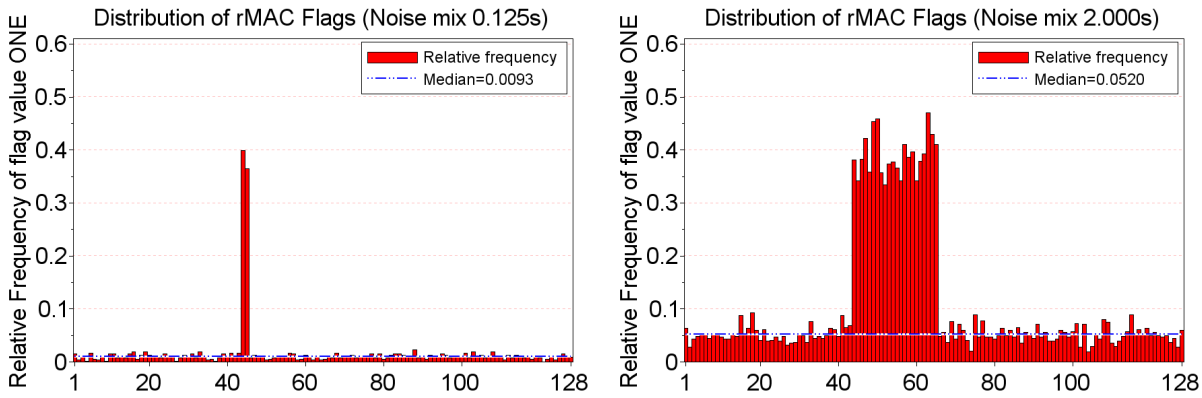


Figure 5.17.: Test result: *rMAC* flags for noise mixing attack at 1/8 and 2.0 seconds (”serial mode”)

5.4.7.2 Localization Approach in ”Serial Mode”

The results presented here refer to the proposal of modified *rMAC* extraction in Section 3.2.4.

For the plausibility of the *”serial mode”* scheme it is at first analyzed *which rMAC* bits are flagging the data modification. The *”noise adding”* attack is carried out at different length. Output of this experiment every detection is an array (of length 128) of ”*rMAC* flags” that indicates which *rMAC* bit index is detected as being attacked. From the analysis of these flag array across all 540 test files, it can be seen that the flagging is according to the definition of the *”serial mode”*. For example, for the short audio attack of 1/8 s duration (only) two *consecutive rMAC* flags are affected namely at bit index 44 and 45 (see Figure 5.17, left). These indices correspond well to the correct temporal location of the attack at $44/128 \cdot 11.8\ \text{s} \approx 4.13\ \text{s}$. With increasing attack duration, the number of consecutive *rMAC* flags increases accordingly.

Again for a few percent of files also *outside* the attacked section attacks are flagged by mistake. Also in the *”serial mode”* the individual standardized spectral coefficients depend on all other spectral coefficients to a certain extent. With increasing duration of the attack, this effect can build up. Closer analysis shows that for the attack duration of 8 seconds (i.e. two third of the total duration), distinction performance is no better than coin flipping. Presumably, the effect on the averages and standard deviations m_{k_0} and s_{k_0} is so intense that *all rMAC* bits are affected with a notable probability. It should be noted that the duration of 8.0 s represents two third of the total duration of the watermark.

Note that the *ideal* value for the relative frequency for *rMAC* flags being ”one” to indicate an integrity breach is 0.5, not 1.0: even under an actual attack, the attacked *rMAC* bit can be identical to its original/authentic value with probability 0.5 *by chance*. However, in the tests the maximum relative frequency for the attacked *rMAC* bits was observed to be rather 0.4 than 0.5.

Now, as in the previous Section the *rMAC* flag ratios are analyzed ”translating” the *rMAC* bit indices to time indices. As an example, the ”mix audio” is considered again. It can clearly be seen that the attacked and un-attacked areas clearly separate: In the attacked areas, the range of the *rMAC* flag ratio is varying between zero and 0.5, see Figure 5.19. In the remaining sections the ratio is mostly zero, except for some outliers as ”background level”.

However, it must be noted that in a notable number of cases no successful detection is triggered. As already said even under ideal conditions the *rMAC* bits indicate a malicious attack only with 50% probability. Hence, also the *rMAC* flag ratios as a derived quantity suffer from a notable

false negative rate. This becomes more clear from looking at the ROC curves again. For the attack durations of 1/2 and 2 seconds, a true positive rate of approximately 0.5 can be achieved while the false positives remains smaller than 0.05, see Figure 5.19. It is very apparent that the characteristic for both attack durations are very similar. Also the AUC measure is relatively constant (approximately 0.75) across different attack lengths. Actually, closer analysis shows that this behavior can also be observed across *all* attack durations up to 4 seconds and also for the ”mix noise” and ”replacement” attack. Hence, the modified ”serial mode” assignment provides an almost *ideal* classification performance over a large range of attack durations.

Only exception is again the ”deletion” attack, as it was observed for the ”scattered mode” scheme already. Note that this is not a critical issue: although a temporal localization can not be carried out with the presented algorithm, the large overall bit error rate is a significant indicator for this attack, at least.

An exhaustive analysis of the AUC measure across all attacks is given in Table 5.3. The interested reader can study the correspondent boxplots and ROC curves in Appendix C.

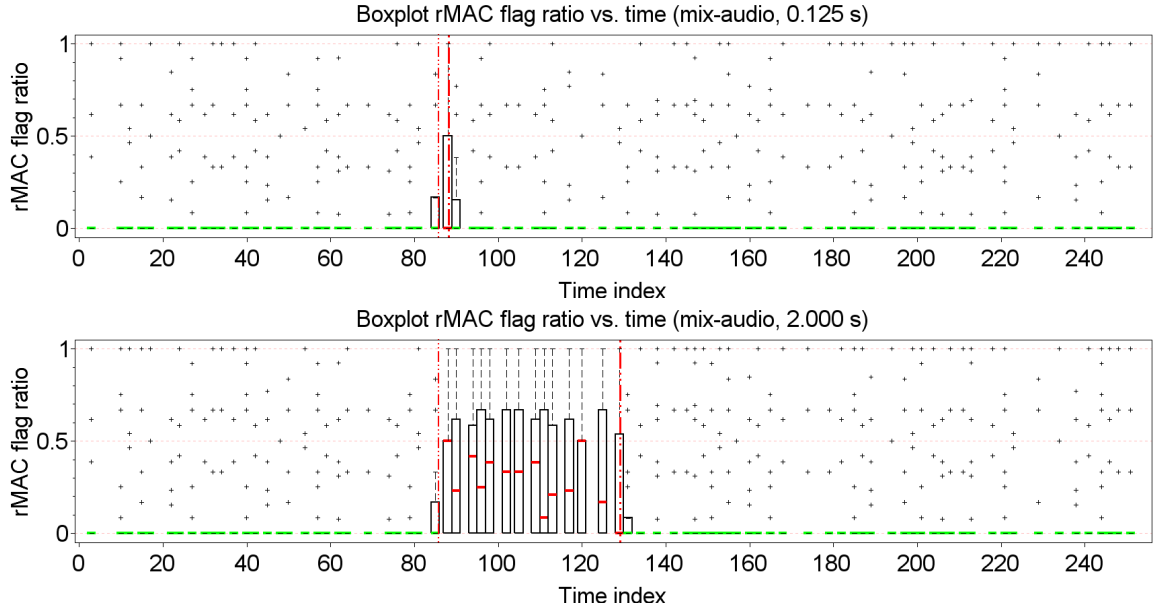


Figure 5.18.: Test result: *rMAC* flag ratio versus time: one boxplot per protected time index; bold center line: respective median value; dashed vertical lines: separator between attacked and un-attacked portions (start: time index 86 = 4.0 seconds)

Attack duration	*”mix noise”*	*”deletion”*	*”replacement”*	*”mix audio”*
1/8 s	0.79	0.51	0.83	0.72
1/4 s	0.80	0.52	0.86	0.77
1/2 s	0.81	0.51	0.87	0.78
1 s	0.80	0.52	0.86	0.77
2 s	0.78	0.52	0.85	0.76
4 s	0.75	0.52	0.82	0.75
8 s	0.59	0.52	0.56	0.70

Table 5.3.: Test result: AUC measure/ROC curve; all attacks, all durations (”serial mode”)

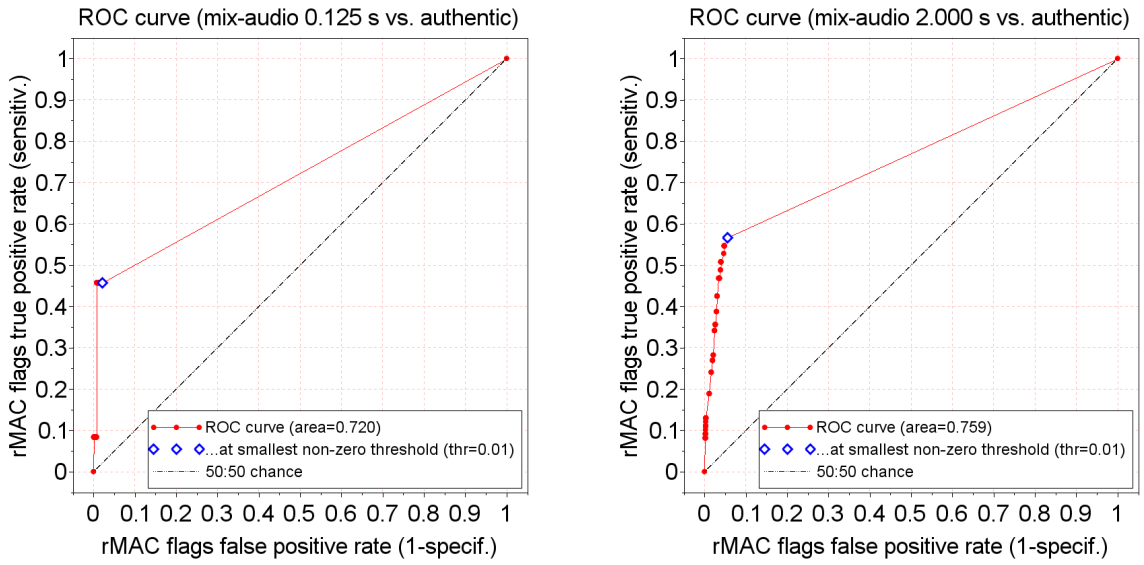


Figure 5.19.: Test result: *rMAC* flag ratio ROC curves; attack duration 0.125 s and 2.0 s

5.4.7.3 Time-code Verification

This Section refers to the proposal in Section 3.4.2.

This rather trivial aspect of the protection scheme was not evaluated for this Chapter again. Test results and execution examples were published the results as published in [ZS2009b]. In that work the test files were prepared so that they contain a number of protected audio sections in series. Each had a duration of 3.0 seconds including the respective block index as watermark message. Then, a section of 9 second duration was deleted from the second half of the test file(s). Finally, the series of differences between the detected value of the block index and the respective expected value (denoted as ”block offset”) are evaluated. It was shown that the deletion is indicated correctly by a block offset of -3 from the spot of deletion, as expected (see Figure 5.20). This means that the embedded block index is suited for indicating data deletions that have a duration longer than the duration of the individual watermarks.

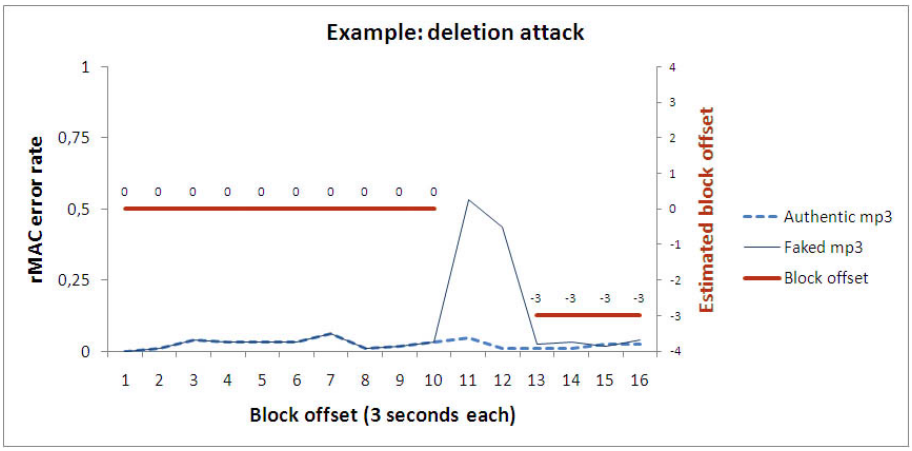


Figure 5.20.: Example: detecting a replacement attack. Figure taken from [ZS2009b]

Note that the (known) information about the duration of the watermark (ergo the protected audio segments) *implicitly* provides a very important information related to the integrity across protected segments: for authentic content, the temporal offset between consecutive protected segments must be *constant*. In the experiments described in this section, a watermarked segment indicated by its sync sequence) can be expected every 12 seconds. Breaches of integrity for example by insertion of content in some cases can *only* be detected on the basis of this timing information. This is especially relevant if a malicious insertion attack is carried out exactly *between* two protected segments. For simplicity, this aspect is not further investigated in this work either.

5.4.8 *rMAC* Key-based Verification of Authenticity

In the last Section of this Chapter the key-dependency of the feature selection as described in Section 3.2.1 is investigated. It is demonstrated how the correct verification result depends on the knowledge of the correct *rMAC* key.

In the preparation for this experiment, all audio snippets were protected with the *rMAC*-watermarking approach like in all experiments described above. For simplicity neither admissible nor malicious attacks were applied afterwards. On the detector side the detection of the watermark message was carried out using the correct watermark key as usual. The subsequent verification of the *rMAC* is carried out using both the correct *rMAC* key and eight randomly chosen *wrong rMAC* keys.

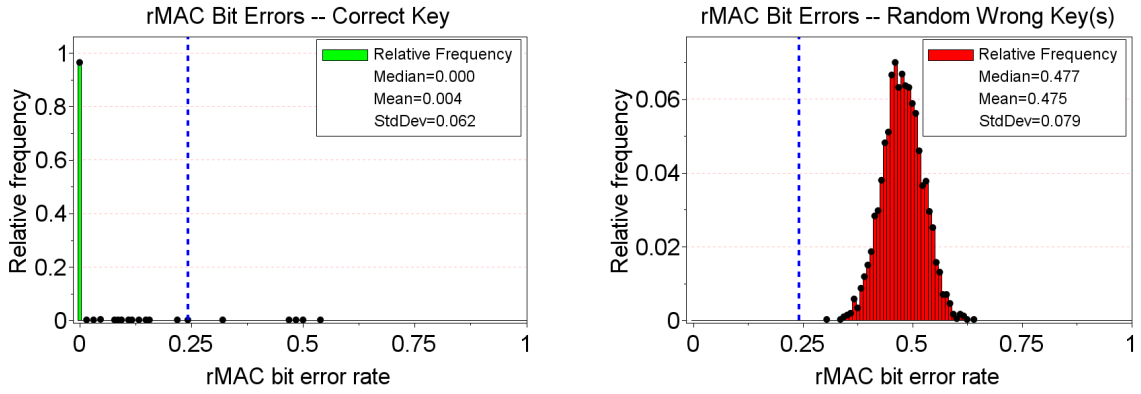


Figure 5.21.: Test result: BER for correct and wrong authentication key, dashed lines: BER threshold correspondent to maximum decision accuracy

As a first result, almost all 540 watermark messages were retrieved correctly. A small number of CRC failures (20 in total) and one single lost watermark had again to be accepted as "background level" though. The subsequent *rMAC* verification process behaves as expected:

- For the verification using the correct authentication key the BER distribution is accumulating at the value 0.00 (Figure 5.21, left).
- For the verification using the wrong key(s) the BER distribution is accumulating at approximately 0.50 (Figure 5.21, right). Only very few false alarm or false acceptance can be observed occasionally.

As a consequence the binary classification shows a very low error rates in terms of false negatives and positives. The optimum can be achieved for the decision threshold of $thr_{\max} = 0.250$. The

correspondent ROC curve is almost perfectly concentrated in the far left and upper area. The AUC area is 0.993 (see Figure 5.22, right).

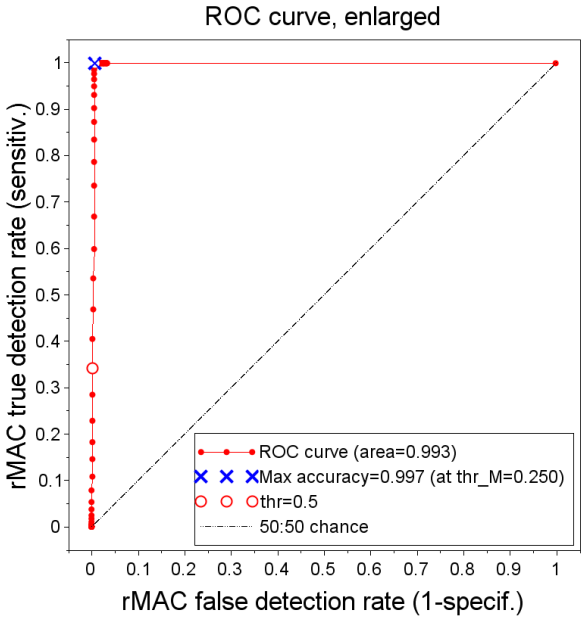


Figure 5.22.: Test result: ROC curve for correct versus wrong authentication key

The almost *ideal* result can be explained by the characteristic of the feature extraction: the *rMAC* keys are being used as a seed for a random number generator. This means for even the slightest difference in the keys it can be expected that the pseudo-random patterns for the *rMAC* extraction are independent. Thus *rMAC* bits are pairwise identical only by accident i.e. with 50% probability. As a consequence the correspondent BER distribution is centered around BER=0.50, as desired. As a consequence only authorized users are able to create the *rMAC* code and verify it at a later date.

5.5 Discussion of Test Results

At first, the *rMAC* approach shall be compared to the original audio fingerprinting work and the original audio watermarking core algorithm by *Haitsma, Kalker* [HOK2001a, HOK2001b] and *Steinebach* [Ste2003, SZ2006a] which were used as a technical starting point. Then in the remainder of this Section the test results in light of the Related Work in authentication watermarking are discussed.

5.5.1 Summary of Test Results

About the watermark embedding, significant improvements could be achieved: using the flexible, ”scattered” embedding pattern in the time-frequency domain, a sufficient capacity improvement could be achieved. At the same time, robustness and transparency could be even improved further by the proposed optimization or integration of enhanced psychoacoustic models in the embedding and detection stage.

About the combined *rMAC*-watermarking approach, it can be said that it shows an acceptable robustness to admissible transformations. It is significantly more sensitive to *MP3* compression than the original work by *Haitsma et al.* At the compression rate of 128 *kbit/s* the correspondent BER of the proposed *rMAC*-watermarking system is nearly 0.20 in comparison to values nearly 0.10 as for the original works. The former value clearly separates from rather inaudible attacks at BERs smaller than 0.10. This is a desired feature: albeit ”not annoying” in terms of the PEAQ model, the compression level of 128 *kbit/s* is already assessed as being ”perceivable” to an average listener. Hence the result is in line with the objective of separating perceivable from imperceptible data modifications.

For the defined malicious attacks, the BER can be used as a criterion for detection if the attack duration can be expected to be two seconds or more. Reasonable decision thresholds can be identified for this. For example, if voice content is protected, this means a sensitivity to doctoring on a *sentence* level.

If an even shorter attack duration can be expected, the *rMAC* flag ratio is found to be a feasible criterion. This allows to indicate doctoring that has a duration as short as 1/8 s. For the example of voice content this allows detecting modifications at a *syllable* level.

For such short attack durations, both ”serial mode” and ”scattered mode” in the *rMAC* extraction allow true positive rates of 0.50 to 0.70 if a false positive rate of less than 0.05 is desired at the same time. This is about the theoretical bound of the ideal classifier under the constraints of the *rMAC* algorithm: each single *rMAC* bit indicates an actual attack only with 50% probability after all. Because of the mostly uncorrelated extraction of the *rMAC* bits, no greater true positive rate can be achieved. The internal ”avalanche effect” introduced by the standardization step seems to be neglectable.

In the transition of attack durations from 1/4 to 2 s, the proposed ”serial mode” of operation provides more a significant classification because it intrinsically avoids many false positives. On the downside, it features a reduced security to targeted attacks because the time-indices of evaluated Fourier coefficients is not fully random.

Like the original work, the proposed *rMAC* clearly separates *distinct* content by an BER of approximately 0.50, as desired. These results mean that the key-dependent (hence pseudo-random) assignment in the time-frequency domain or the discarded overlapping of FFT frames does not reduce the distinction performance in principle. In addition, the experimental results provide strong indication that that the authentication can only be carried out in a consistent

manner, if the correct *rMAC* key is known. Without knowledge of this key, no successful verification could be forged in the simulations.

Nevertheless, it has to be admitted that a number of false positives have to accepted. This is mainly caused by shortcomings of the enhanced embedding algorithm in the presence of ”critical” audio cover data. In 3 to 4 percent of the test snippets, no watermark message could be embedded in reliable manner and the consecutive rMAC verification had to fail. This demonstrates the limits of applicability of watermarking technology in general. Preliminary results on the the sole rMAC verification [ZS2008a, ZS2008c] showed that the rMAC algorithm by itself is not affected by such conditions.

5.5.2 Comparison to the Related Work in Authentication Watermarking

About the work by *Radhakrishnan* and *Memon* [RM2002] the comparison is difficult for different reasons: First of all, only very little empirical results were given by the authors (two ”audio clips” with unspecified total duration). Furthermore, the assessment of sound quality loss caused by attacks the authors use the so-called *perceptual entropy* (see [Joh1988, PS2000] for its definition) instead of the objective difference grades (*ODG*). This makes the comparison of sensitivity and specificity difficult. Also some important technical settings for the attacks were not given, for example the range of the cut-off frequencies of the low-pass filtering attack or if the *MP3* attacks at 64 *kbit/s* were carried out in stereo or mono. But qualitatively, the results are at least resembling those in this thesis: the bit error rate gradually increases with increasing intensity of attacks. Malicious attacks seem to cause higher BER than admissible attacks. Nevertheless, notable false positives and negatives have to be accepted. Finally, it is important to notice that the *rMAC* from this thesis work is much less vulnerable to birthday attacks *in theory* because it is of length 128 *bit* instead of 32 *bit* as in the work by *Radhakrishnan et al.*

The same is true for the authentication watermarking approach by Steinebach [SD2003b]. It defines so-called *feature checksums* as authentications codes which have a length of only 4 *bit* in the experimental evaluation by the original author. However, as these feature checksums are derived from perception-based properties like *loudness* and *brightness* features or directly from a highly subsampled audio spectrum it can well tolerate a number of admissible attacks on audio like lossy compression. Interestingly the work by *Steinebach* suffers from the same issue like this thesis by the fact that watermarking detection cannot be carried out fully reliable: a number of watermarking bit errors and hence incorrect or even missed detections have to be accepted (see *”CRC failed”* and *”Watermark lost”* results above). But in comparison, the false positive rate of the presented *rMAC* watermark (average BER: 0.004) is much smaller than in the work by Steinebach (average BER of the ”Nothing” attack on the RMS feature: 0.07 at best, see [SD2003b, Figure 9]).

The work by *Hong-Xia Wang* [WF2010] also presents an insecure modeling of the final authentication code: In the experimental chapter of that publication, a watermark message of only 4 *bit* is constructed from the spectral centroid feature(s) and eventually embedded as authentication code. Albeit a very innovative alternative approach, the same is true for the work by *Gomez/Cano* [GCdG$^+$2002]: Arbitrary audio data is mapped to an ”alphabet” of 16 different audio descriptor units (ADU) which corresponds again to a key length of 4 *bit*.

Also the contribution by *Gulbis* [GMS2008a, Gul2013] shows a rather short length of the embedded authentication code: the so-called feature vector has a length of 17 *bit* for every audio section of 0.34 s. It is doubtful that this is truly secure with respect to deliberate collision or even 2nd pre-image attacks.

The approaches presented by *Petrovic* or *Park* [Pet2005, PTW2007] are only partly comparable to this thesis work: The watermark message consists of a time-dependent time code instead of a content-dependent authentication code. Hence if an attack is carried out *in-place* only very intense modifications can be detected based on the detection score or failed/missed detections. This is even more true because the robustness in the approach by *Petrovic* seems to be even higher than in this thesis work.

For example, authors claim that the watermark survives a ”perceptual codec” at bitrates as low as 4 *kbit/s*. At the same time, embedding rates are a little smaller than this thesis work as the authors describe that the 40 *bit* message is embedded every 8 seconds. About the temporal localization of deletions or insertions, the works by *Petrovic*, *Park*, and in this thesis seem to be comparable. As said before, for this thesis work a successful synchronization precision up to one PCM sample can be expected. Hence even short insertions of milliseconds of audio into a file can be indicated (indirectly) by an offset between actual and expected temporal position.

A comparison with the work by *Ching-Te Wang* [WLC2007] cannot be given here because the original work lacks a precise description of the algorithm and also test results about the *sensitivity* to malicious attacks. Its security to targeted attacks on the availability is also doubtful as the embedding is done in the infrasound frequencies at 1-20 Hz which can be attacked easily without sound quality degradation. In comparison, this thesis work embeds the authentication watermark in a wider range from 500-12,000 Hz which makes targeted removal attacks much more difficult.

The approaches by *Zhao* [ZS2009a, ZS2010] are capable of identifying deletions or removal but only if the protected audio undergoes no further acts of strong lossy compression attacks. Being *semi-fragile* approaches the degree of data modification is not measured in terms of BER but by the deviation of the detector’s correlation coefficient from its ideal value of 1.0. The approaches can well indicate malicious acts of deletion or insertion of *”short”* sections of audio (of unspecified duration, unfortunately). About admissible attacks the results provided by the authors are somewhat inconsistent with respect to the notion of *perceptual identity*: for example, the requantization attacks at 8 *bit* resolution actually cause annoying quantization noise but are not indicated by the *Zhou* approach (correlation coefficient reported as *”1.0000”*). On the other hand, *MP3* compression attacks at 56 *kbit/s* (presumably stereo) are clearly audible too but only causes little deviation of the correlation coefficient (”0.9918”) from 1.0. Unfortunately, rather imperceptible attacks (like *MP3* compression at higher bitrates) were not tested by the authors so that it cannot be assessed if such moderate attacks can disguise malicious acts. In contrast, the test results from this thesis suggest that such strong *MP3* compression or the adding of audible quantization noise would be indicated successfully by the proposed approach.

Finally, the comparison with the works by *Fan, Wu* or *Yuan/Huss* is omitted as these works are beyond the scope of this thesis work due to their high fragility [CS2004, CHW2008, CZL2010] or because they are bitstream solutions for voice-coded data in GSM [YH2004] or G.723 [WK2001] format.

The same is true for the comparison with the recent fragile approaches by *Huang, Echizen et al.* [HEN2010, HEN2011, HOEN2014] for audio content or by *Rigoni et al.* [RFF2016] for audio-visual multimedia content.

Chapter 6

Summary and Conclusion

This final Chapter summarizes the major contributions of this work with regards to the thesis challenges and objectives. Correspondent publications by the thesis author are recalled and commented. This includes student theses that were conducted under supervision of the author. Then the achievements shortcomings of the proposed approach are discussed and the overall feasibility of the developed authentication mechanisms is concluded. Finally, potential impact is outlined and possible future research work is suggested.

Please note that a list of publications by the author can be found on pp. 157.

6.1 Main Contributions and Correspondent Publications

Modern audio technologies provide numerous opportunities for for malicious doctoring, forging/disguising the origin, or inadvertent data modification of digital audio recordings. Common crypto primitives for verifying the integrity and authenticity of digital media like crypto-hashes and signatures offer secure means of protection in this case. However they feature a number of shortcomings in the context of everyday handling of multimedia data: because of their extremely high sensitivity to *any* kind of data modification also common imperceptible, hence *benign*, post-processing operations like format conversion of lossy encoding are flagged as a ”false alarm” with the standard primitives. This thesis work investigated alternatives from the field of multimedia security for overcoming the shortcomings in crypto primitives.

With regards to data integrity this thesis work proposed a modified definition of the notion of ”integrity”: instead of an *exact* verification of data identity (i.e. bit-by-bit) the requirements were relaxed in this work so that only *perceptible* data modifications have been in focus (*”perceptual identity”*).

Based on that, it proposed and investigated an approach for authentication watermarking. For this, research was conducted in audio hashing, audio watermarking, perceptual modeling, and in combing these technologies by means of ”content-fragile watermarking” as summarized in the following.

6.1.1 Robust Message Authentication Code (*rMAC*) for Audio Data

This contribution corresponds to the publications [ZS2007, ZS2008a, ZS2008b, ZS2008c, ZMS2012] by the thesis author.

The approach of this thesis work was to investigate if and how audio hashing technology can serve as a robust message authentication code (denoted as *rMAC*) to discriminate perceivable data modifications from inaudible ones. The proposed *rMAC* algorithm is influenced by one of the most frequently cited audio hashing algorithms by *Haitsma, Kalker*, and *Oostveen* [HOK2001a, HOK2001b]. The concept and its experimental evaluation were published by the author in first or co-authorship [ZS2007, ZS2008a, ZS2008b, ZS2008c, ZMS2012]. The experimental evaluation in these papers showed that the algorithm allows well to separate inaudible from audible data modifications. The same is also true for distinguishing content which is completely distinct from one another. Hence the proposed scheme allows integrity verification on a perceptional level.

In addition, the work [TNSZ2009] motivated the proposed randomization of the feature selection. Protection against protocol attacks by creating *false positives* is vital to avoid that the perceptual hash appear useless. The secret key dependent feature selection provides that data verification can be carried out by the authorized user only. That is, the scheme can be used for data source authentication to a certain extent.

An intermediate standardization processing step of the input spectral data provides better statistical properties of the *rMAC* and hence a higher security to targeted attacks. By design the *rMAC* algorithm also implicitly allows the temporal localization of tampering and assessing the degree of data modification (if applicable).

6.1.2 Enhancement in Patchwork Audio Watermarking

The following contribution corresponds to the publications [SZ2008b, ZS2009b].

As another objective, the *rMAC* authentication codes should be kept ”inside” the protected audio media without causing overhead by means of digital watermarking in a transparent manner. An existent Patchwork watermarking [Ste2003] was used as a starting point for significant enhancements to general Patchwork watermarking schemes. The focus was on improving the overall capacity, transparency and robustness performance as recalled in the following. Finally note that audio watermarking-based applications beyond this thesis work (like tracing copyright infringements or *2nd screen* services) could benefit from these results.

Capacity Enhancement in Patchwork watermarking: At first the embedding capacity of the Patchwork scheme in the Fourier domain had to be increased. For this a flexible assignment of FFT coefficients to the watermark message bits across low, mid, and high frequencies was developed. The capacity enhancement was published in [ZS2009b] and successfully evaluated more elaborately in this thesis. In a few percent of the tested audio material incorrect or even missed watermark detections had to be accepted though. The overall idea was also presented in co-authorship in [BZSS2011] where it was applied to collusion-secure anti-piracy watermarking beyond the context of this thesis work.

Robustness Enhancement by Psychoacoustic Modeling in the Detector: A minor research activity in the context of this thesis was a further improvement of watermarking robustness. The detection success in the presence of strong attacks could be improved by utilizing psychoacoustic modeling also in the watermark detection step. This was elaborated upon

in a supervised student thesis [Mer2007], published in co-authorship [SZ2008b], and evaluated again in this thesis on a larger test set.

Robustness Enhancement by Adaptations to Turbo Coding: In addition, forward error correction techniques known from communication theory could be adapted and ”tuned” for watermarking application. Research on enhanced *Turbo* coding algorithms was conducted in a student thesis [Ber2008] under supervision by the thesis author. The detail on improving Turbo coding were used but not elaborated upon in this thesis.

6.1.3 *rMAC*-enabled Authentication Audio Watermarking

This contribution corresponds to the publications [ZS2009b, ZMS2012]

The above works on audio hashing and watermarking were eventually integrated as core components into a ”content-fragile watermarking” system. An alignment mechanism was proposed so that both components are compliant with each other. It was initially published in [ZS2009b], and successfully evaluated in this thesis work. Its main capabilities are the following:

Classifying Tampered vs. Authentic Content – Assessment of the Degree of Data Modification: The approach allows classifying doctored versus authentic content. Using the bit error rate (BER) as criterion, a false positive and negative rate of 4% each at the same time can be achieved. If a false negative rate of 0% is desired, the false positive rate at the same time increases to 20%. Main cause for the notable error rates is that audio watermarking algorithm sometimes fails when the protected audio content is unsuitable for carrying an embedded watermark. It has to be admitted that these figures apply if the duration of deliberate tampering is at least 2.0 seconds. Hence, this means that the goal of identifying even very short sections of data modification (less than a second) cannot be achieved if only the total BER is used as criterion.

Temporal Localization of Tampering: To overcome this shortcoming two approaches were proposed and evaluated successfully. One approach (”scattered mode”) is based on a deliberate evaluation of the single *rMAC* bit results on the detector side. It provides a very good sensitivity to data modifications of very short duration up to 1/2 second. For longer durations of integrity breach, the second approach (”serial mode”) is more suitable: it features a deliberate modification of the fully pseudo-random assignment to facilitate the first approach. The latter was investigated in a supervised student thesis [Mun2011] and presented in [ZMS2012].

Compared to the related work on audio authentication watermarking this work presents an alternative approach which provides a key-based authentication code. It was evaluated on a wider test base than many other works. Using a content-dependent authentication code means a significant security improvement of authentication watermarking with regards to copy attacks (as opposed to many works in the state of the art).

6.1.4 Security Analysis of Patchwork Watermarking and *rMAC* Approach

It was also investigated how secure the algorithms are in terms of ”effective key length”. For this the combinatorial properties of the modified Patchwork and *rMAC* algorithm were analyzed. The result is that computational needs for brute force attacks on the system can still be regarded as very high: With realistic assumptions on the time-frequency patterns of picked coefficients,

the effective key length is approximately 100 *bit* for the rMAC in ”scattered mode” and 62 bit in the (more suitable) ”serial mode”.

The effective key length of the proposed Patchwork audio watermarking is at least 260 bit for reasonable technical settings and assumptions.

The analysis and the results have not been published explicitly but are elaborately documented in this thesis in Chapter 4. These results can also be applied to the core algorithm as published earlier by *Steinebach et al.* in application areas like copyright protection.

6.1.5 Benchmarking

This contribution corresponds to the publication [DSLZ2004]

In an early publication also benchmarking of authentication watermarking was investigated and published in co-authorship [DSLZ2004]. On a conceptual level, it was proposed by the thesis author to extend the (formerly) existent *Stirmark Benchmark Audio* system (*SMBA*) by *Lang, Dittmann* and *Steinebach* [SDS$^+$2001]. The objective is performance evaluation of *integrity* watermarking. Like the earlier *SMBM* standards for robustness and security attacks, also the sets of simulated tampering attacks and the set of audio test data was proposed to be standardized.

However the concept was never implemented as the overall *SMBA* project was discontinued by the primary authors. Nevertheless, the inspiration for feasible simulations of attacks were eventually utilized in the experimental evaluation in this thesis.

To summarize, the previous overview on publication shows that this thesis work covers different fields in audio processing. Apart from the major research efforts in authentication watermarking, also techniques of audio hashing, watermarking and perceptual modeling by themselves were investigated and improved as the need arose.

6.2 Conclusion on Limitations and Achievements

The previous section showed that most of the thesis objectives could be achieved – but not all of them. The most important findings in this regard are listed in the following.

Availability of the *rMAC* Authentication Codes

First of all, it could be observed that not all audio test data is suitable for watermark embedding. Here, issues were observed occasionally for audio content that contains a very narrow audio spectrum or which at very low overall volume. However, this unfortunate characteristic could have been expected in the first place: it is known from the different application areas of watermarking that not *every* audio content is well suited for watermark embedding *a priori*. For example in anti-piracy transaction watermarking this is much less relevant: Usually content like music songs, audio books or main feature films offer at least a few file positions that are suitable for embedding the (only) copyright notice or user ID a few times. But in integrity watermarking, the whole file must be covered for a complete protection because the watermark messages vary dynamically over the file's duration.

Hence these rare critical data cannot be protected in a reliable way because false alarms have to be expected. In fact, closer analysis in the experimental evaluation showed that most of the observed false positive errors could be explained by deficiencies in the watermarking embedding/detection scheme, not in the rMAC algorithm.

Life-cycle of Existent Content

Another limitation becomes obvious in light of the natural life-cycle of a media: Unlike audio forensics (any) authentication watermarking protection, but also any cryptography-based mechanism too, becomes effective not before the watermark is embedded. This means that it is desired that the embedding is carried out at an early stage in the life-cycle of the audio media. A natural realization to achieve this is to include the embedding step in a recording device like a voice recorder or when audio content is initially imported to an audio archive. This would allow to address the challenges in protecting digital evidence or for preserving the cultural heritage.

About the latter, the nature of digital watermarking to modify the protected content actively can be of relevance. As a consequence, watermarking could require to conduct a transparency approval by responsible parties, e.g. listening tests by sound engineers. Here, usually the unmodified original versions are preserved and sometimes authenticated using crypto hashes and signatures, resp. However, integrity watermarking is feasible for being used for working copies that are distributed internally or to external users.

Error Rates

Another aspect to discuss are the error rates. It is notable that the binary classification of the proposed approach in general provides a very good classification of admissible versus malicious attacks. A decision threshold can be identified so that false positive and negatives can be reduced to 4% for both at the same time. This is mainly caused by shortcomings of the watermark embedding/detection scheme: the rMAC authentication code cannot retrieved from every protected audio sample which – in doubt – has to be treated as an alarm. These figures mean that the automated verification step might require some manual inspection afterwards.

To summarize, these figures also mean that the proposed solution cannot be seen as a *universal* protection mechanism for *any* audio data under *any* circumstances. Compared to other protec-

tion mechanisms like crypto hashes, crypto MACs, digital signatures or checksums this means a notable shortcoming. Nevertheless the proposed scheme appears to be suitable to a wide range of different kinds of common audio content though.

One way of overcoming the presence of false positives, false negatives, and missed verifications by technical means could be implementing a *verifying stage* already during embedding. The user should at least be notified that the content cannot be protected so that alternative actions can be taken *in advance*.

Symmetric Security Primitive

The dependency of the rMAC extraction from a key provides a means for verification of both, the integrity and the authenticity. This is facilitated by the key dependency of the consecutive watermarking step. But the reader is reminded that that the investigated approach is a *symmetric* security primitive: both the watermark key and the *rMAC* key are symmetric shared secrets. This means that it has to be assumed that the verification step has to be carried out in a *trusted* environment only. A truly *public* verification cannot be provided based on the proposed system. As a consequence, a trusted access and distribution for the secret/private *rMAC* extraction and watermarking key has to be provided.

It has to be noted that this is no particular property of the content-fragile watermarking approach as investigated in this thesis by itself. Instead it is a common property of real-world watermarking systems in general. This includeds academic research as well as commercial products and services. There, the watermarking key is expected to remain private to the entities that have access to the watermark detector software.

The same is true for key-dependent crypto MAC functions. Note that the symmetric nature of nowadays watermarking schemes were not in focus of this thesis. Nevertheless, the proposed *rMAC* approach and the principles of *rMAC*/watermarking alignment might be extended to be compatible to *asymmetric* watermarking schemes of the future.

Continuous Notion of ”Perceptual Integrity/Identity”

Not a true limitation but rather a new notion is the following: Regarding the security of the investigated approach a more general conclusion arises: The overall concept of using bit error rates (BER) in the output of a (whatsoever) hash function as a continuous measure for the integrity of a media must appear *odd* if seen from a *cryptographer*’s perspective. In crypto-based standard approaches (like crypto hashes/MACs or digital signatures) the verification essentially provides an exact *”yes/no”* criterion instead of a continuous measure.

An example: A BER bound of, for example, 0.10 in the presence of ”admissible audio file format conversion” will never be observed using a crypto hash function, given by its algorithmic design. Instead the expected BER in the output of a crypto hash function becomes 0.50 in the presence of even the slightest, even lossless, modification to the input data.

Furthermore, it has to be admitted that also the concept of ”weak’ and ”strong” audio fingerprint/hash bits as discussed in this thesis could appear curios compared to cryptographic standard algorithms. For the latter all bits in a standard crypto hash contribute equally to the overall significance of the hash value. Already the notion of ”perceptual identity” as defined in the Introduction includes a certain degree of inexactness: The subjective sound quality loss due to attacks are different from person to person.

Conclusion

To overcome these contradictory perspectives, the reader is reminded that the presented approach can solve a *classification* challenge than a strict cryptographic *security* challenge. As long as estimates for false negatives and positives are given reasonable decision thresholds can be set. Additionally, the proposed approach shares common properties similar to those in cryptography: examples are the observation that the *rMAC* correctly distinguishes totally different content by an average BER of 0.5 again; or by the observation that successful verification requires the correct and exact *rMAC* key; or by the analysis and estimation in terms of ”effective key length”.

To conclude, the investigated approach shall not be seen as a mean to *replace* crypto-based integrity authentication mechanisms in terms of the provided security level or, for example, as in light of national digital signature acts regulating data integrity. It can rather serve as an *additional* means of verification. Like the technique of *audio forensics*, watermarking serves as an *additional* protection mechanism that provides *indications* of malicious activities instead of *court-proof* evidence. Here, watermarking allows to keep the authentication codes virtually inside the protected media which is an alternative to common storage and archiving standards.

Finally, the outstanding property of this thesis work is to verify the integrity if lossless post-production operations or lossy compression can be expected. The *rMAC* authentication code and the watermark protection are present even when the protected content is transmitted and circulating even when transcoded to other file formats. In contrast, integrity and authenticity protection using crypto hashes and digital signatures, resp., would fail: the transcoding would cause false alarms even if the transcoded copy was perpetually identical to the original version. This means that the proposed approach can even provide a protection in application scenarios in which crypto-based techniques cannot be used in principle.

Such potential impacts are discussed in the following Section.

6.3 Potential Impact on Applications

Since the invention of audio recording devices, a vast number of analog and digital audio media has been created. The issue of the credibility and provenience of these media becomes apparent for the examples that were given in the Motivation of this thesis (see Section 1.1), namely content preserving the *cultural heritage* and *digital evidence* contained in audio. The thesis result can have an impact as outlined in the following.

Preserving the Cultural Heritage

Many different kinds of audio media contribute to contemporary history. This includes famous historical speeches, oral history interviews, mass media like TV and the radio or the variety in the music. These can be seen as ”exhibits” which are being preserved as historical archival records in public archives. The embedded authentication watermark can ”survive” common post-processing operations in the archive (see [Ros2009]) like converting between standard archiving formats like *BWF* and *RF64* [Cha1997, EBU2009]. Unlike standard crypto-based hashes and signatures, the *rMACs* value do not need to be refreshed after the transcoding.

In this case, for the existing historical archival records the proposed method cannot always provide a *complete* chain of proof: authentication watermarking technique was not used or had even been unavailable for existing media. But newly created content can be protected when it is moved into a public archive for the first time. It can also be used when historical exhibits on analog media (e.g. magnetic tape and records from the vinyl or even shellac era) are digitized. This allows, at least, verifying that the digital copy has not been modified inadvertently or maliciously *inside* the archive.

An alternative is protecting working copies when they are handed out to legitimate users, the latest. In the latter case, 3rd party users could use the embedded watermark to verify that the content was not modified by the user in the meanwhile.

Digital Evidence

Digital media can also be important evidence in different kinds of law enforcement actions. This covers recordings of police interview, video surveillance content (if audio is captured too), or eavesdropping in the course of lawful and strategic interception of any kind of voice communication. The audio contained in these sources can prove or at least indicate suspicious, accusing or exonerating facts.

For content that has already existed in these archives, the proposed method cannot become effective not before the watermark is embedded. Hence it is desired that the watermark embedding is carried out as early as possible in the life-cycle of the audio media. This is done at best *inside* the recording device like a voice recording units in interview rooms, mobile dictaphone devices (or apps) or in monitoring centers in lawful interception. The proposed approach can provide the authentication codes even in the presence of (common) lossy compression and across consecutive steps of transcoding / format conversion.

6.4 Future Research Directions

The above summary also points at interesting aspects that are worthwhile for further investigation. Some ideas for Future Work are outlined in the following.

6.4.1 Fine-tuning of decision thresholds for particular audio content

The optimization of decision thresholds for bit error rates of *rMAC* flag ratios was carried out under the assumption that a rather *universal* audio protection across hi-fi and low quality music and voice content is desired. If it is known *a priori* that, for example, only voice content needs to be protected, other technical settings could be used that provide even greater distinction performance and robustness.

For example, from own research beyond this thesis work it is known that pure voice content like in audio books or voice recordings can ’bear” greater watermarking distortions than hi-fi music content. Watermarking modifications according to the Patchwork scheme can be applied by a few Decibel stronger before they become a transparency issue. This gain in being ”markable” could be used in voice content to overcoming the issues of occasional ”lost watermarks” and for increasing the overall detection performance.

It would be worthwhile to carry out the experimental evaluation again on larger sets from a certain audio data domain for further study. That will allow providing more representative and significant estimates of error rates. It could also help identifying other examples of audio material that is unexpectedly critical to process.

6.4.2 Confidence measure in *rMAC* verification

About the *rMAC* extraction itself, the assessment of bit errors on the detector side can be refined further. The reader is reminded that the *rMAC* is a *binary* quantity. It is verified on the detector side in terms of the Hamming distance or BER, resp. According to its definition in Equation (3.1), each *rMAC* bit $H' \in \{0,1\}$ depends on the *continuous* zero-mean quantity $d' \in \mathbb{R}$. As discussed in Section 2.2.3 or in the work by *Coover* [CH2014] some of these *rMAC* bit correspond to a d' value with smaller absolute value $|d'|$ (so-called ”weak bits”) than for others. In simple words: an *rMAC* on the detector side that consists of many weak bits, for whatever reason, has a lower confidence level than other *rMAC*s. Hence a feasible extension the *rMAC* verification could be to evaluate $|d'|$ so that a confidence measure of the verification result is obtained based on the ”weaknesses” of the *rMAC* bits.

Also psychoacoustic modeling could be incorporated in this context. The reader is reminded that perceptual audio models were successfully incorporated in the overall watermarking detection. The motivation for this was to allow an assessment and weighting of every spectral coefficient according to its significance for the final detection and retrieval result. Applying this general principle on the *rMAC* extraction, an observed attack on an *rMAC* bit (indicated by its bit flip) should can be regarded as less significant with regards to ”perceptual integrity” if the corresponding N contributing spectral coefficients are by themselves less or even inaudible anyway. Note that this idea shall not be confused with the investigated but later discontinued research approach as outlined at the end of Section 3.2: calculating the *rMAC* bits from spectral coefficients that were subject to psychoacoustic model-based quantization *priory* would not be intended. Instead the weighting should better be carried out *after* the *rMAC* bit modeling is processed.

Both, the former ”weakness” measure and the latter ”inaudibility” measures could be combined in an appropriate manner to obtain eventually a meaningful measure of confidence in the case that integrity breaches are indicated.

6.4.3 Model-based Authentication of Speech Content

In some application scenarios pure voice data has to be protected and acoustic events in the background can be omitted. Here, the message authentication could benefit from different technologies in speech processing as outlined in the following.

Speech Hashing

Speech hashing could be carried out instead of using universal audio hashing (audio fingerprinting, resp.) as it was carried out in this thesis work. This technique incorporates extracts an identifier from voice data using mathematical models for speech production in the human vocal tract. This should allow for an greater robustness to admissible data modifications. For example works from the state of the art in speech hashing use *linear spectrum frequencies* [JLN2008, JJN2009], or they carry out a principle component analysis on *linear prediction coefficients (LPC)* [CW2009] for extracting robust hashes. Especially the latter proposal appears plausible because LPC models are a well-understood ”mathematical standard tool” in many disciplines of digital speech processing.

Phonetic and Phonemic Transcription

As alternative to *rMAC* or voice hashing, elements from automated speech recognition (ASR) techniques could be included for the extraction of a voice data authentication code. One can make use of the fact that across human languages a so-called *phonetic inventory* of approximately 150 different *phones* is known, which can be represented by 8 bit per phone symbol. Here, a phone is defined as the shortest meaningful element of perceivable and distinguishable speech sounds *”without regard to its place in the sound system of a language”* [MWe2009]. A phonetic transcription or pseudo-randomly selected features from intermediate processing steps in ASR could be used to derive a phonetic-based message authentication code.

An even more robust approach could be developed if also the *language* of the spoken words can be assumed to be known *a priori*. Each human language consist of *phonemes* as their shortest *semantically* relevant and distinguishable elements. In a given language the *phonemic inventory* describes existent and non-existent phones and sequences thereof. For instance the English and the German language consists of a phonemic inventory of only 50 to 60 characteristic phonemes which could be represented with as little as 6 *bit* each [Mad2013].

As an example, the oral statement by a suspect during a police interview, saying

”...I am guilty...”

can be transformed to its phonemic representation

”... /aɪ æm gɪltɪ/ ...”

using the *International Phonetic Alphabet (IPA)* notation[71]. To achieve this technical, ASR technology often implements hidden *Markov* modeling (HMM) [Rab1990] and suitable *acoustic* and *language models* to reflect the ”dictionary” of the particular language. An IPA-like phonemic transcription or pseudo-randomly selected internal HMM-quantities could be investigated to compute a phonemic message authentication code.

It is very promising to utilize elements from ASR for protecting voice content as this technology is very mature[72]: reliable solutions are nowadays included commercially in operating systems (like in *Microsoft Windows, Apple MacOS/iOS* or in *Google Android*), used for taking dictations in word processor software conveniently, or for controlling military equipment[73]. Current ”speech-to-text” transcription software in monitoring centers for strategic surveillance of telephone communication is capable for analyzing and recognizing even heavily distorted/noisy speech content at virtual ”*NSA*-grade” level[74].

Biometric Speaker Verification

From the perspective of both perceptual and semantic integrity of voice content it is of course vital to know *who* gave a certain oral statement. Carrying out ’replay attacks” by exchanging one speaker by another would be very difficult to carry out because an adversary would need to be very precise in repeating the pronunciation and the timing of the original data. But such attacks are possible in principle though. Phonetic and phonemic authentication are invariant to such malicious attack. Hence identification of the individual speaker as the very initial source of the voice data in question would also have to be considered. About this, speaker recognition technology from biometry could verify that the authenticity of speech data is maintained.

The technologies outlined above often use the same mathematical tools like MFCCs, LPCs or HMM as a source of inspiration. To summarize, it would be worthwhile investigating speech-based message authentication can be carried out in a reliable and robust manner which is – at the same time – difficult to forge.

6.4.4 Joint Audio-Video Authentication

Another feasible extension of this thesis work could be realized in the context of integrity protection for video data in different ways. At first, the reader is reminded that sensitive audio information can also be contained in the soundtrack of multimedia-based ”video” content. Here, video watermarking could be used in addition to audio watermarking for increasing the embedding capacity. For example, this could allow to further reduce the duration of protected snippets which would in return increase the sensitivity to malicious acts of tampering.

As an alternative, content-fragile watermarking for audio and video could be combined: audio hashes (like the investigated *rMAC*) approach could be embedded by means of video watermarking, and *vice versa*. This allows joint approaches using both, the video and audio track for *mutual* authentication purposes, as proposed already in 2000 by *Dittmann* [DMS2000] or very

[71] *International Phonetic Alphabet (IPA)* developed by the *International Phonetic Association*. *https://www.internationalphoneticassociation.org*

[72] The interested reader is referred to an exhaustive list of ASR tools and APIs which is kept up-to-date at *Wikipedia*: *http://en.wikipedia.org/wiki/List_of_speech_recognition_software*

[73] Example: *DARPA* project *Robust Automatic Transcription of Speech (RAST)*: *http://www.darpa.mil/program/robust-atuomatic-transcription-of-speech*

[74] Andrew Fishman, *”The Computers are Listening”*, in: *The Intercept* (2015) *https://theintercept.com/2015/05/11/speech-recognition-nsa-best-kept-secret* (by *First Look Media*)

recently by *Rigoni* [RFF2016]. It also saves oneself from elaborate mechanisms for avoiding influence on the hashing process caused by the embedding process.

6.4.5 Miscellaneous Watermarking Applications

The proposed mechanism for audio hash extraction and enhanced audio watermarking can also have an impact on watermarking applications beyond this thesis context: for example, elements of the proposed *rMAC* can be used by itself in audio hashing applications as explained in Section 2.2.2. For example, it was argued by *Fridrich* [FG2000] that key-dependent feature extraction is vital for (more) secure audio hashing.

> This was justified by the work published in co-authorship with *Thiemert* [TNSZ2009].

The latter work showed that audio hash based filtering mechanisms could be circumvented by deliberate attacks on the hash. For this, the internal (pseudo) randomization will make such attacks much more difficult.

Another impact can be seen in improvements in general Patchwork audio watermarking: The commercial application of the audio watermarking core algorithm can benefit by a higher robustness and security in anti-piracy activities for example music, movies or video games as it was discussed in co-authorship in [BSL+2013].

6.4.6 Authentication of Time and Location of Recording

Another reasonable extension to the proposed protection mechanism could be covering the *context* of audio data with regards to *the time and the place* of its initial creation/recording. For example, when protecting *video* data, the additional embedding capacity provided by the video track could be utilized for embedding such meta data for geographic and temporal referencing. This can not only prevent deliberate acts of forgery. It can also help resolving inadvertent mix-up of media content and facilitate organizing media repositories in the future.

The source of such geographic or temporal information can be

- global navigation satellite systems (such as *GPS*, *GALLILEO*, or *GLONASS*),
- timing signals by radio (like the *DCF77* timing signal[75]),
- time servers (according to the *NTP* protocol as in RFC 5905 [MMBK2010]) or
- trusted digital time-stamping services (as in X.509/RFC 3161 [ACPZ2001]).

This extension is already known on a conceptual level for example as described and patented in [Whi2004, Rho2011] or developed in military video intelligence systems as for example announced[76] and demonstrated[77] by the commercial watermarking supplier *Digimarc Inc.*

[75] DCF77 timing signal: transmitted by the operator *Media Broadcast GmbH*, Köln, Germany, on behalf of the German federal *National Metrology Institute (Physikalisch-technische Bundesantalt, PTB), http://www.ptb.de*

[76] *Digimarc* press release: *http://defensenews-updates.blogspot.de/2009/09/digimarc-awarded-contract-to-enhance-us.html* (2009)

[77] *Digimarc*'s watermarking solution for video surveillance in military unmanned aerial vehicles, demonstrated at the conference *IS&T SPIE Electronic Imaging/SSWMC*, invited demo session, San Francisco, USA, Jan. 2012. Video integrity watermarking was described to be embedded in unmanned aerial vehicles "*on-the-fly*" (*sic!*)

List of Publications by the Thesis Author

The following sections provide the list of all publications by the thesis author as main creator or in co-authorship.

Publications linked to this PhD Thesis

The following overview itemizes the publications in connection with this PhD work (29 total, self-citations omitted) and by which other works they were being cited. They are included in the overall Bibliography, too.

[ZMS2012] Sascha Zmudzinski, Badar Munir, and Martin Steinebach. Digital audio authentication by robust feature embedding. In: *Electronic Imaging 2012 - Media Watermarking, Security, and Forensics XIV*. IS&T SPIE, Jan 2012

Subject: Perception-based audio authentication watermarking - improved temporal localization of tampering

Cited by: [YbQyZt2013, HOEN2014, DDZ2013, ZHHQ2016]

[ZS2009b] Sascha Zmudzinski and Martin Steinebach. Perception-based audio authentication watermarking in the time-frequency domain. In: Stefan Katzenbeisser, editor, *Information Hiding Conference 2009*, LNCS. Springer, 2009

Subject: Perception-based audio authentication watermarking incorporating "*rMAC*"

Cited by: [HEN2010, HEN2011, QKBD2010, LFL2010, XLY2011, LZW2012, LZW2013, ZLY+2013, Kno2013, DDZ2013]

[TNSZ2009] Stefan Thiemert, Stefan Nürnberger, Martin Steinebach, and Sascha Zmudzinski. Security of robust audio hashes. In: *IEEE International Workshop on Information Forensics and Security, WIFS 2009, London, UK*, Dez 2009

Subject: Proof of concept for security attacks on existent perceptual hashing

Cited by: [Kno2013]

[SZ2008b] Martin Steinebach and Sascha Zmudzinski. Optimiertes Auslesen digitaler Audiowasserzeichen. In: *D-A-CH Security 2008: Bestandsaufnahme, Konzepte, Anwendungen, Perspektiven*. Syssec, Basel, 2008

Subject: Including perceptual modeling in the detection stage

Cited by: -

[ZS2008a] Sascha Zmudzinski and Martin Steinebach. Content-based message authentication coding for audio data. In: *Ammar Alkassar, Jörg Siekmann (Editors): Proceedings of "Sicherheit 2008 - Schutz und Zuverlässigkeit", 2.-4. April 2008, Saarbrücken, GI-Edition - Lecture Notes in Informatics (LNI), P-128. Köllen Verlag, Bonn*, 2008

Subject: Key-based and robust perceptual hashing ("*rMAC*"); work-in-progress

Cited by: -

[ZS2008b] Sascha Zmudzinski and Martin Steinebach. Psycho-acoustic model-based message authentication coding for audio data. In: *10th ACM Workshop on Multimedia and Security (ACM MMSEC'08), Oxford, UK, September 22-23, 2008*, 2008

Subject: Key-based and robust perceptual hashing ("*rMAC*"); work-in-progress

Cited by: [JJN2009, HPG2009, LVRY2010, QKBD2010, WCG2012]

[ZS2008c] Sascha Zmudzinski and Martin Steinebach. Robust audio hashing for audio authentication watermarking. In: Edward J. Delp III, editor, *Proceedings of SPIE: Security, Steganography, and Watermarking of Multimedia Contents X*, January 2008

Subject: Key-based and robust perceptual hashing ("*rMAC*"); work-in-progress

Cited by: [GMS2008b, GM2010]

[ZS2007] Sascha Zmudzinski and Martin Steinebach. Robust message authentication code algorithm for digital audio recordings. In: Ping Wah Wong Edward J. Delp III, editor, *Proceedings of SPIE Volume: 6505, Security, Steganography, and Watermarking of Multimedia Contents IX*. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham/Wash, 2007

Subject: Key-based and robust perceptual hashing ("*rMAC*"); work-in-progress

Cited by: -

[ZSN2005] Sascha Zmudzinski, Martin Steinebach, and Sergey Neichtadt. Vertrauenswürdigkeit von Audiodaten – Digitale Wasserzeichen und Verifikation der semantischen Integrität. In: *Sicherheit 2005: Sicherheit – Schutz und Zuverlässigkeit, Beiträge der 2. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)*. Köllen Verlag, 2005

Subject: Analysis of content-based audio retrieval for key-based perceptual hashing

Cited by: -

[DSLZ2004] Jana Dittmann, Martin Steinebach, Andreas Lang, and Sascha Zmudzinski. Advanced audio watermarking benchmarking. In: *SPIE Int. Symposium on Electronic Imaging, Security and Watermarking of Multimedia Contents, San Jose, USA*, 2004

Subject: Proposed extension of *Stirmark* benchmarking suite for authentication watermarking

Cited by: [MHJM2004b, MHJM2004a, MHJM2005a, WFM2005, MHJM2005b, MHJSM2006, LA2008, Nut2012]

Publications on Other Subjects

The following overview itemizes publications by the author *beyond* the scope of this thesis.

Watermarking Algorithms

[ZSB2012] Sascha Zmudzinski, Martin Steinebach, and Moazzam Butt. Watermark embedding using audio fingerprinting. In: *Transactions on Data Hiding and Multimedia Security VIII*, volume 7882 of *Lecture Notes in Computer Science*. Springer, Aug 2012

Subject: Enabling fast watermark embedding in combination with audio fingerprinting

[SZP2012] Martin Steinebach, Sascha Zmudzinski, and Dirk Petrautzki. Forensic audio watermark detection. In: *Electronic Imaging 2012 - Media Watermarking, Security, and Forensics XIV*. IS&T SPIE, Jan 2012

Subject: Detecting the presence of a watermark from short sections of audio data

[SZB2010] Martin Steinebach, Sascha Zmudzinski, and Moazzam Butt. Robust hash controlled watermark embedding. In: *32nd Annual Symposium of the German Association for Pattern Recognition (DAGM 2010) - Special Workshop on Pattern Recognition for IT Security, September 22-24, 2010, Darmstadt, Germany*. DAGM, Sep 2010

Subject: Enabling fast watermark embedding in combination with audio fingerprinting

[SZ2008a] Martin Steinebach and Sascha Zmudzinski. Evaluation of robustness and transparency of multiple audio watermark embedding. In: III Delp, Edward J., Ping Wah Wong, Jana Dittmann, and Nasir D. Memon, editors, *SPIE Int., Symposium on Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, San Jose, USA*, volume 6819, 2008

Subject: Robustness of Patchwork watermarking to multiple embedding using different watermarking keys

[ZSN2006] Sascha Zmudzinski, Martin Steinebach, and Sergey Neichtadt. Robust audio-hash synchronized audio watermarking. In: Eduardo Fernández-Medina and Mariemma I. Yagäe, editors, *4th International Workshop on Security in Information Systems (WOSIS 2006), Paphos, Cyprus*, 2006

Subject: Combining watermarking with fingerprinting/robust hashing for improvement of robustness to time-stretching and pitch-shifting

[SZ2006b] Martin Steinebach and Sascha Zmudzinski. Robustheit digitaler Audiowasserzeichen gegen Pitch-Shifting und Time-Stretching. In: *Sicherheit 2006: Sicherheit - Schutz und Zuverlässigkeit. Beiträge der 3. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)*. Köllen Verlag, 2006

Subject: Improving robustness to time-stretching and pitch-shifting

[SZSL2003] Martin Steinebach, Sascha Zmudzinski, Stefan Schäfer, and Andreas Lang. Robustheitsevaluierung digitaler Audio-Wasserzeichen im Rundfunkszenario. In: *2. Thüringer Medienseminar der FKTG – Rechte digitaler Medien, Erfurt, Germany*. Film- und Kinotechnische Gesellschaft FKTG, 2003

Subject: Evaluation of robustness against FM radio transmission

Efficiency of Watermarking

[ZSB2012] Sascha Zmudzinski, Martin Steinebach, and Moazzam Butt. Watermark embedding using audio fingerprinting. In: *Transactions on Data Hiding and Multimedia Security VIII*,

volume 7882 of *Lecture Notes in Computer Science*. Springer, Aug 2012

Subject: Enabling fast watermark embedding in combination with audio fingerprinting

[SZB2010] Martin Steinebach, Sascha Zmudzinski, and Moazzam Butt. Robust hash controlled watermark embedding. In: *32nd Annual Symposium of the German Association for Pattern Recognition (DAGM 2010) - Special Workshop on Pattern Recognition for IT Security, September 22-24, 2010, Darmstadt, Germany*. DAGM, Sep 2010

Subject: Enabling fast watermark embedding in combination with audio fingerprinting

[SZF2004] Martin Steinebach, Sascha Zmudzinski, and Chen Fan. The digital watermarking container: Secure and efficient embedding. In: *Proceedings of the ACM Multimedia and Security Workshop, 20.-21. September 2004, Magdeburg, Germany*, 2004

Subject: Increasing processing speed of transaction watermarking for online shops ("Watermarking Container")

[SZ2004a] Martin Steinebach and Sascha Zmudzinski. Complexity optimization of digital watermarking for music-on-demand services. In: *Technology, economy, social and legal aspects of virtual goods – 2nd International Workshop, Ilmenau, Germany May 27-29 2004 (VIRTUAL GOODS 2004)*. Technische Universität Ilmenau, 2004

Subject: Increasing processing speed of transaction watermarking for online shops ("Watermarking Container")

Watermarking Applications

[BSL+2013] Waldemar Berchtold, Marcel Schäfer, Huajian Liu, Fabio Touceira Takahashi, Andre Schmitz, Sascha Zmudzinski, Martin Steinebach, and Jonas Wieneke. Video game watermarking. In: *Electronic Imaging 2013 – Media Watermarking, Security, and Forensics 2013*. IS&T SPIE, Jan 2013

Subject: Watermarking for the protection of video game content

[WSZ2008] Patrick Wolf, Martin Steinebach, and Sascha Zmudzinski. Adaptive security for virtual goods – building an access layer for digital watermarking. In: Rüdiger Grimm, editor, *Virtual Goods. International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorp. the 4th International ODRL Workshop (VIRTUAL GOODS 2008) , October 16-18, 2008 Poznan, Poland*. Publishing House of Poznan University of Technology, 2008

Subject: Java Framework for integration of watermarking executables in different environments ("Watermarking Algorithm Manager")

Watermarking-supported Protocols

[ZSKR2010] Sascha Zmudzinski, Martin Steinebach, Stefan Katzenbeisser, and Ulrich Rührmair. Audio watermarking forensics: detecting malicious re-embedding. In: *IS&T SPIE Electronic Imaging 2010 Conference - Media Forensics and Security XII, San Jose, January 2010*, 2010

Subject: Integration of watermarking in key exchange protocols; vulnerability of watermarking against multiple embedding with the same watermark key

[RKSZ2010] Ulrich Rührmair, Stefan Katzenbeisser, Martin Steinebach, and Sascha Zmudzinski. Watermark-based authentication and key exchange in teleconferencing systems. In: *11th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security (CMS 2010), 31 May - 02 June, 2010, Linz, Austria*, number 6109 in LNCS, Security

and Cryptology. Springer, Mai 2010

Subject: Detecting man-in-the-middle attacks in key exchange by digital watermarking

Collusion Resistant Fingerprint Watermarking

[BZSS2011] Waldemar Berchtold, Sascha Zmudzinski, Marcel Schäfer, and Martin Steinebach. Collusion-secure patchwork embedding for transaction watermarking. In: *Electronic Imaging 2011 - Media Watermarking, Security, and Forensics XIII*. IS&T SPIE, Jan 2011

Subject: Modifying Patchwork watermarking for intrinsic collusion security on transaction watermarking

[SBZS2010] Marcel Schäfer, Waldemar Berchtold, Sascha Zmudzinski, and Martin Steinebach. Zero false positive 2-secure fingerprinting watermarking. In: *The 12th ACM Workshop on Multimedia and Security 2010 (ACM MMSEC2010), 08.-09. Sep 2010, Rome, Italy*, 2010

Subject: Zero false-positive collusion-secure fingerprint coding for transaction watermarking

[SBH+2010] Marcel Schäfer, Waldemar Berchtold, Margareta Heilmann, Sascha Zmudzinski, Martin Steinebach, and Stefan Katzenbeisser. Collusion secure fingerprint watermarking for real world applications. In: *SICHERHEIT 2010, Berlin*. Gesellschaft für Informatik, 2010

Subject: Optimizing code length of collusion-secure fingerprint coding for transaction watermarking

[SZ2006a] Martin Steinebach and Sascha Zmudzinski. Countermeasure for collusion attacks against digital watermarking. In: *Electronic Imaging 2006 – Security, Steganography, and Watermarking of Multimedia Contents VIII*. IS&T SPIE, Jan 2006

Subject: Collusion-security by pre-warping in the Fourier phase domain

Digital Audio Forensics

[SYZW2015] Martin Steinebach, York Yannikos, Sascha Zmudzinski, and Christian Winter. Advanced Multimedia File Carving. In: Anthony T.S. Ho and Shujun Li, editors, *Handbook of Digital Forensics of Multimedia Data and Devices*, chapter 6, pages 219–269. Wiley IEEE Press, September 2015

Subject: Co-authorship in an overview on current forensic file carving techniques for digital images, audio and video in a handbook on multimedia forensics

[ZTS2012] Sascha Zmudzinski, Ankit Taneja, and Martin Steinebach. Carving and reorganizing fragmented mp3 files using syntactic and spectral information. In: *AES 46th Conference on Audio Forensics 2012, 14-16 June 2012, Denver, CO, USA*. Audio Engineering Society, June 2012

Subject: File carving for reconstruction of deleted and fragmented MP3 files during forensic inspection

Partial Encryption

[SZB2005] Martin Steinebach, Sascha Zmudzinski, and Thorsten Boelke. Audio watermarking and partial encryption. In: Ping W. Wong Edward J. Delp III, editor, *Proceedings of SPIE - Volume 5681, Security, Steganography, and Watermarking of Multimedia Contents VII*. SPIE, Bellingham, USA, 2005

Subject: Embedding and detection of watermarking in encrypted content

[SZ2004b] Martin Steinebach and Sascha Zmudzinski. Partielle Verschlüsselung von MPEG Audio. In: Patrick Horster, editor, *IT Security & IT Management, Proceeding D-A-CH Security 2004*, 2004

Subject: Embedding and detection of watermarking in encrypted content

Patents

[SBZS2012] Marcel Schäfer, Waldemar Berchtold, Sascha Zmudzinski, and Martin Steinebach. Verfahren zur Erzeugung von Transaktionswasserzeichen und Auswerteverfahren zur Kundenrückverfolgung. DPMA German Patent DE102010044228 A1, March 2012. (application: Sep. 2010, granted: Okt. 2012)

Subject: 2-collusion-secure fingerprint coding for transaction watermarking

[SZ2009b] Martin Steinebach and Sascha Zmudzinski. Method for embedding a multi-bit digital watermark in media data. EPO European Patent EP2012269 A1, January 2009. (application: July 2007, granted: May 2011)

Subject: Efficient creation of watermarked content

[SZL2007] Martin Steinebach, Sascha Zmudzinski, and Huajian Liu. URL watermark as filter for on-line directories. EPO European Patent EP1739952 A1, January 2007. (application: July 2005, granted: April 2008)

Subject: Watermark-enabled prevention of copyright infringements of image content on websites during crawling and indexing

[SZ2007] Martin Steinebach and Sascha Zmudzinski. Countermeasure for collusion attacks in digital watermarking. EPO European Patent EP1739617 A1, January 2007. (application: July 2005, granted: Sep. 2014)

Subject: Prevention of collusion attacks of transaction watermarking

Mainstream IT Press

Finally, the following publication in the public mainstream press is mentioned for completeness. Although not being reviewed in a *scientific* context, the work was thoroughly edited by the publishing company and it reached a *large* audience in the general public (with a print run of more than 300,000 copies).

[SZ2009a] Martin Steinebach and Sascha Zmudzinski. Individuell gestempelt – Die Technik hinter digitalen Audio-Wasserzeichen. Number 9/2009 in c't-Magazin für Computertechnik, pages 142–146. Heise Verlag, Hannover, April 2009

Subject: Overview article on State of the Art in audio watermarking

Award

"Best Paper Award" granted by *The Digital Watermarking Alliance* and the societies *IS&T* and *SPIE* for the following publication on improving audio watermarking detection by deliberate non-blind preprocessing:

[SZN2011] Martin Steinebach, Sascha Zmudzinski, and Stefan Nürnberger. Re-synchronizing audio watermarking after nonlinear time stretching. In: *Electronic Imaging 2011 - Media Watermarking, Security, and Forensics XIII*. IS&T SPIE, Jan 2011. (**Best Paper Award)**

Subject: Enabling watermark detection after time-stretching

Bibliography

[ABHB2009] Emad E. Abdallah, A. Ben Hamza, and Prabir Bhattacharya. Watermarking 3D Models Using Spectral Mesh Compression. *Signal, Image and Video Processing*, 3(4):375–389, 2009.

[ABV2009] Michael Arnold, Peter G. Baum, and Walter Voeßing. A phase modulation audio watermarking technique. In: Stefan Katzenbeisser and Ahmad-Reza Sadeghi, editors, *Information Hiding*, volume 5806 of *Lecture Notes in Computer Science*, pages 102–116. Springer Berlin Heidelberg, 2009.

[ACB+2014] M. Arnold, Xiao-Ming Chen, P. Baum, U. Gries, and G. Doerr. A phase-based audio watermarking system robust to acoustic path propagation. *Information Forensics and Security, IEEE Transactions on*, 9(3):411–425, March 2014.

[ACPZ2001] C Adams, P. Cain, B. Pinkas, and R. Zuccherato. *RFC 3161: Internet X.509 Public Key Infrastructure: Time-Stamp Protocol (TSP)*. Number 3161 in Request for Comments (RFC). Internet Engineering Task Force (IETF), April 2001. (updated by RFC 5816).

[AH2000] E. Allamance and J. Herre. Secure delivery of compressed audio by compatible bitstream scrambling. In: *108th AES convention: 2000 February 19-22, Paris (Preprint - AES 5074-5173)*, New York, NY, USA, 2000. Audio Engineering Society (AES).

[AHH+2001a] E. Allamanche, J. Herre, O. Hellmuth, B. Froba, and M. Cremer. AudioID: Towards Content-Based Identification of Audio Material. In: *110th Audio Engineering Society Convention*. AES, May 2001.

[AHH+2001b] E. Allamanche, J. Herre, O. Helmuth, B. Froba, T. Kasten, and M. Cremer. Content-Based Identification of Audio Material Using MPEG-7 Low Level Description. In: ISMIR [ISM2001].

[AKS2003] André Adelsbach, Stefan Katzenbeisser, and Ahmad-Reza Sadeghi. On the insecurity of non-invertible watermarking schemes for dispute resolving. In: T. Kalker, I. Cox, and Y.M. Ro, editors, *International Workshop on Digital Watermarking (IWDW 2003)*, volume 2939 of *Lecture Notes in Computer Science*, pages 355–369. Springer, Berlin, Heidelberg, 2003.

[And1996] R. Anderson, editor. *First Workshop on Information Hiding, Cambridge, U.K., May 30–June 1, 1996, ISBN 3540619968*, volume 1174 of *Lecture Notes in Computer Science*. Springer, August 1996.

[ANR1974] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, Jan 1974.

[Arn2000] Michael Arnold. Audio watermarking: features, applications and algorithms. In: *IEEE International Conference on Multimedia and Expo 2000 (ICME 2000), New York, USA, 30. Juli–2. August*, pages 1013–1016. IEEE Press, 2000. IEEE Catalog Number 00HT8532.

[ASW2003] Michael Arnold, Martin Schmucker, and Stephen D. Wolthusen. *Techniques and Applications of Digital Watermarking and Content Protection*. Artech House computer security series. Artech House, 2003.

[BBC+2004] Mauro Barni, Franco Bartolini, Vito Cappellini, Massimiliano Corsini, and Andrea Garzelli. Digital watermarking of 3D meshes. In: *SPIE Int. Symposium on Electronic Imaging, Security and Watermarking of Multimedia Contents, San Jose, USA*, volume 5208, pages 68–79, 2004.

[BBF+2000] M. Barni, F. Bartolini, J. Fridrich, M. Goljan, and A. Piva. Digital watermarking for the authentication of AVS data. In: *10th European Signal Processing Conference (EUSIPCO 2000)*, pages 1–4, Sept 2000.

[BC2007] S. Baluja and M. Covell. Audio fingerprinting: Combining computer vision data stream processing. In: *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, volume 2, pages 213–216, April 2007.

[BDRF+2013] Tiziano Bianchi, Alessia De Rosa, Marco Fontani, Giovanni Rocciolo, and Alessandro Piva. Detection and classification of double compressed mp3 audio tracks. In: *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '13, pages 159–164, New York, NY, USA, 2013. ACM.

[Ber2008] Waldemar Berchtold. Optimierung der Robustheit und Klangqualität digitaler Audio-Wasserzeichen-Verfahren im Kontext von angepassten Vorwärtsfehlerkorrektur-Algorithmen. Diplomarbeit, Hochschule Darmstadt, Germany, Dep. of Mathematics, December 2008.

[BF2012] Patrick Bas and Teddy Furon. Are 128 bits long keys possible in watermarking? In: *Proceedings of the 13th IFIP TC 6/TC 11 International Conference on Communications and Multimedia Security*, CMS'12, pages 191–191, Berlin, Heidelberg, 2012. Springer-Verlag.

[BF2013] Patrick Bas and Teddy Furon. A New Measure of Watermarking Security: The Effective Key Length. *IEEE Transactions on Information Forensics and Security*, 8(8):1306 – 1317, July 2013.

[BGML1996] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal, MIT Media Lab*, 35(3,4):313–336, 1996.

[BGT1993] C. Berrou, A Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In: *IEEE International Conference on Communications, 1993 (ICC '93), Geneva, Switzerland, 23-26 May 1993.*, volume 2, pages 1064–1070 vol.2, May 1993. (see French patent no. 9105279, European patent no. 92460011.7, US patent no. 07/870483).

[BHMS2007a] F. Balado, N.J. Hurley, E.P. McCarthy, and G. C M Silvestre. Performance analysis of robust audio hashing. *IEEE Transactions on Information Forensics and Security*, 2(2):254–266, June 2007.

[BHMS2007b] F. Balado, N.J. Hurley, E.P. McCarthy, and G. C M Silvestre. Performance of philips audio fingerprinting under additive noise. In: *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, volume 2, pages 209–212, April 2007.

[BHT1963] B. Bogert, M. Healy, and J. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In: *Proc. Symp. on Time Series Analysis*, pages 209–243, 1963.

[BP1998] P. Bassia and I. Pitas. Robust audio watermarking in the time domain. In: *9th European Signal Processing Conference (*EUSIPCO'98*)*, pages 25–28, Island of Rhodes, Greece, 8–11 1998.

[BRC1960] R. C. Bose and D. K. Ray-Chaudhuri. On a class of error correcting binary group codes. *Information and Control*, 3(1):68–79, March 1960.

[Bru2004] Richard Brunner. *Urheber- und leistungsschutzrechtliche Probleme der Musikdistribution im Internet*. PhD thesis, Universitaet Augsburg, 2004.

[BSL+2013] Waldemar Berchtold, Marcel Schäfer, Huajian Liu, Fabio Touceira Takahashi, Andre Schmitz, Sascha Zmudzinski, Martin Steinebach, and Jonas Wieneke. Video game watermarking. In: *Electronic Imaging 2013 – Media Watermarking, Security, and Forensics 2013*. IS&T SPIE, Jan 2013.

[BSN2002] C. Busch, M. Schmucker, and M. Nesi, P.and Spinu. Evolution for music score watermarking algorithm. In: *W.P: Wong; E.J. Delp. Security and Watermarking of Multimedia Contents IV (SPIE Proceedings Series 4675)*, pages 181–193. SPIE, 2002.

[BSS2013] Waldemar Berchtold, Marcel Schaefer, and Martin Steinebach. Leakage detection and tracing for databases. In: ACM, editor, *Proceedings of The 1st ACM Workshop on Information Hiding and Multimedia Security (IH & MMSEC 2013), June 17-19, 2013 Montpellier, France*. ACM, June 2013.

[BT1959] Ralph Beebe Blackman and John Wilder Tukey. *The Measurement of Power Spectra, from the Point of View of Communications Engineering*, chapter Particular Pairs of Windows, pages 95–101. Dover Publications, 1959.

[BTH1996] Laurence Boney, Ahmed H. Tewfik, and Khaled N. Hamdy. Digital watermarks for audio signals. In: *International Conference on Multimedia Computing and Systems*, pages 473–480, 1996.

[BW2004] Rainer Böhme and Andreas Westfeld. Statistical characterisation of MP3 encoders for steganalysis. In: Jana Dittmann and Jessica J. Fridrich, editors, *Proceedings of the 6th workshop on Multimedia*

& Security, MM&Sec 2004, Magdeburg, Germany, September 20-21, 2004, pages 25–34. ACM, 2004.

[BZLB2014] Frank Breitinger, Georg Ziroff, Steffen Lange, and Harald Baier. Similarity Hashing Based on Levenshtein Distances. In: Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics X*, volume 433 of *IFIP Advances in Information and Communication Technology*, pages 133–147. Springer Berlin Heidelberg, 2014.

[BZSS2011] Waldemar Berchtold, Sascha Zmudzinski, Marcel Schäfer, and Martin Steinebach. Collusion-secure patchwork embedding for transaction watermarking. In: *Electronic Imaging 2011 - Media Watermarking, Security, and Forensics XIII*. IS&T SPIE, Jan 2011.

[CABD2012] Xiao-Ming Chen, Michael Arnold, Peter G. Baum, and Gwenaël J. Doërr. AC-3 Bit Stream Watermarking. In: *2012 IEEE International Workshop on Information Forensics and Security, WIFS 2012, Costa Adeje, Tenerife, Spain, December 2-5, 2012*, pages 181–186, 2012.

[Cas1979] Kenneth R. Castleman. *Digital Image Processing*. Prentice Hall Professional Technical Reference, 1st edition, 1979.

[CB2007] M. Covell and S. Baluja. Known-audio detection using waveprint: Spectrogram fingerprinting by wavelet hashing. In: *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, volume 1, pages 237–240, April 2007.

[CC2010] C.-Y. Chang and S. Clark. Practical linguistic steganography using contextual synonym substitution and vertex colour coding. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, Stroudsburg, PA, USA*, 2010.

[CCW2003] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu. Semantic context detection based on hierarchical audio models. In: *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 109–115, New York, NY, USA, 2003. ACM Press.

[CDF2006] Ingemar J. Cox, Gwenaël J. Doërr, and Teddy Furon. Watermarking is not cryptography. In: Y.Q. Shi and B. Jeon, editors, *International Workshop on Digital Watermarking (IWDW 2006)*, volume 4283 of *Lecture Notes in Computer Science*, pages 1–15. Springer, Berlin, Heidelberg, 2006.

[CDXZ2015] Ning Chen, J. Stephen Downie, Hai-dong Xiao, and Yu Zhu. Cochlear pitch class profile for cover song identification. *Applied Acoustics*, 99:92 – 96, 2015.

[CFF2005] F. Cayre, C. Fontaine, and T. Furon. Watermarking security: theory and practice. *Signal Processing, IEEE Transactions on*, 53(10):3976–3987, 2005.

[CGB+2002] P. Cano, E. Gómez, E. Batlle, L. de Gomes, and M Bonnet. Audio fingerprinting: Concepts and applications. In: *2002 International Conference on Fuzzy Systems Knowledge Discovery (FSKD'02), Singapore, November 2002*, 2002.

[CGW2012] Wei-Hong Chuang, Ravi Garg, and Min Wu. How secure are power network signature based time stamps? In: Ting Yu, George Danezis, and Virgil D. Gligor, editors, *ACM Conference on Computer and Communications Security (CCS)*, pages 428–438. ACM, 2012.

[CH1983] James Cameron and Gale Anne Hurd. *Terminator*. Pacific Western Productions, Beverly Hills, CA, USA, movie script, 4th draft edition, April 1983.

[CH2014] B. Coover and Jinyu Han. A power mask based audio fingerprint. In: *IEEE International Conference on Acoustics, Speech and Signal Processing 2014 (ICASSP) 2014*, pages 1394–1398, May 2014.

[Cha1997] R. Chalmers. The Broadcast Wave Format – an Introduction. Technical Report (EBU Technical Review), European Broadcasting Unit (EBU), July 1997.

[CHW2008] Fan Chen, Hong-Jie He, and Hong-Xia Wang. A fragile watermarking scheme for audio detection and recovery. In: *Image and Signal Processing, 2008. CISP '08. Congress on*, volume 5, pages 135 –138, may 2008.

[CK2001] S. Craver and S. Katzenbeisser. Security analysis of public-key watermarking schemes. In: *Proceedings of the SPIE, Mathematics of Data/Image Coding, Compression, and Encryption IV, volume 4475, pages 172–182, July 2001.*, 2001.

[CKLS1996] Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. Secure spread spectrum watermarking for images, audio and video. *IEEE Signal Processing Society – 1996 International Conference on Image Processing (ICIP '96)*, 1996.

[CKLS1997] Ingemar J. Cox, Joe Kilian, Frank Thomson Leighton, and Talal G. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6:1673–1687, 1997.

[CL1997] Ingemar J. Cox and Jean Paul M. G. Linnartz. Public watermarks and resistance to tampering. In: *In International Conference on Image Processing (ICIP'97*, pages 26–29. IEEE, 1997.

[CM2002] Ingemar J. Cox and Matt L. Miller. The first 50 years of electronic watermarking. *Journal of Applied Signal Processing*, pages 126–132, April 2002. NEC Research Institut.

[CMB+2007] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

[CMBM2001] I. Cox, M. Miller, J. Bloom, and M. Miller. *Digital Watermarking*. The Morgan Kaufmann Series in Multimedia Information and Systems. Elsevier Science, 2001.

[CMYY1997] Scott Craver, Nasir D. Memon, Boon-Lock Yeo, and Minerva M. Yeung. Can invisible watermarks resolve rightful ownerships? In: *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 310–321, 1997.

[CMYY1998] Scott Craver, Nasir Memon, Boon-Lock Yeo, and Minerva M. Yeung. Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *IEEE Journal on Selected Areas in Communications*, 16(4):573–586, 1998.

[CS2004] Nedeljko Cvejic and Tapio Seppänen. A novel scheme for merging digital audio with watermarking and authentication. In: *IEEE 6th Workshop on Multimedia Processing*, 2004.

[CS2007a] M. Casey and M. Slaney. Fast recognition of remixed music audio. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1425–IV–1428, April 2007.

[CS2007b] Nedeljko Cvejic and Tapio Seppänen, editors. *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. Information Science Reference, Hershey, NY, USA, July 2007.

[CT1965] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[CT1999a] Edoardo Charbon and Ilhami Torunoglu. Watermarking layout topologies. In: *IEEE Asia South-Pacific Design Automation Conference*, pages 213–216, 1999.

[CT1999b] Christian Collberg and Clark Thomborson. Software watermarking: Models and dynamic embeddings. In: *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '99, pages 311–324, New York, NY, USA, 1999. ACM.

[CW2001] B. Chen and G.W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. on Information Theory*, 47(4):1423–1443, May 2001.

[CW2009] Ning Chen and Wang-Gen Wan. Speech hashing algorithm based on short-time stability. In: Cesare Alippi, Marios Polycarpou, Christos Panayiotou, and Georgios Ellinas, editors, *Artificial Neural Networks - ICANN 2009*, volume 5769 of *Lecture Notes in Computer Science*, pages 426–434. Springer Berlin Heidelberg, 2009.

[CWL+2001] Scott A. Craver, Min Wu, Bede Liu, Adam Stubblefield, Ben Swartzlander, Dan S. Wallach, Drew Dean, and Edward W. Felten. Reading between the lines: lessons from the sdmi challenge. In: *Proceedings of the 10th conference on USENIX Security Symposium - Volume 10*, SSYM'01, pages 10–10, Berkeley, CA, USA, 2001. USENIX Association.

[CX2013] Ning Chen and Hai-dong Xiao. Perceptual audio hashing algorithm based on Zernike moment and maximum-likelihood watermark detection. *Digital Signal Processing*, 23(4):1216 – 1227, 2013.

[CZ2008] Ning Chen and Jie Zhu. A multipurpose audio watermarking scheme for copyright protection and content authentication. In: *Multimedia and Expo, 2008 IEEE International Conference on*, pages 221–224, 2008.

[CZL2010] Ning Chen, Meng-Yao Zhu, and Sen Liu. A new fragile audio watermarking scheme. In: *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, pages 367–372, 2010.

[Dam1988] I.B. Damgård. Collision free hash functions and public key signature schemes. In: D. Chaum and W.L. Price, editors, *Advances in Cryptology, EurocryptŠ87, LNCS 304*, pages 203–216. Springer, 1988.

[Dau1992] Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[DBI2004] Xiaoxiao Dong, M.F. Bocko, and Z. Ignjatovic. Data hiding via phase manipulation of audio signals. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing 2004 Proceedings (ICASSP 2004)*, volume 5, page 377, 2004.

[DBP1996] H. Dobbertin, A. Bosselaers, and B. Preneel. RIPEMD-160, a strengthened version of RIPEMD. *Fast Software Encryption - LNCS*, 1039:71–82, 1996.

[DDPP2013] S. V. Dhavale, R. S. Deodhar, D. Pradhan, and L. M. Patnaik. Robust multiple stereo audio watermarking for copyright protection and integrity checking. In: *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013)*, pages 9–16, Oct 2013.

[DDZ2013] Rui-Hong Dong, Yan-Jun Di, and Qiu-Yu Zhang. A robust content authentication algorithm of speech based on perceptual hashing. *Sensors & Transducers*, 161(12):375, 2013.

[dGCG+2003] L. de Gomes, P. Cano, E. Gómez, M. Bonnet, and E. Batlle. Audio watermarking and fingerprinting: For which applications? *Journal of New Music Research*, 32(1), 2003.

[Dit2000a] Jana Dittmann. *Digitale Wasserzeichen*. Springer Verlag Berlin Heidelberg, April 2000.

[Dit2000b] M. Dittmann, J.; Steinebach. A framework for a secure MIDI eCommerce. In: *Proceedings of the Information Property, Intellectual Property and New Technology Conference (KnowRight 2000), Vienna, 25th - 29th September 2000*, pages 51–57. österreichische Computer Gesellschaft, 2000.

[Dit2001] Jana Dittmann. Content-fragile watermarking for image authentication. In: Wong and Delp III [WDI2001].

[DKSV2004] Jana Dittmann, Stefan Katzenbeisser, Christian Schallhart, and Helmut Veith. Provably secure authentication of digital media through invertible watermarks. *IACR Cryptology ePrint Archive*, 2004:293, 2004.

[DL2004] Peter Jan O. Doets and R.L. Lagendijk. Theoretical modeling of a robust audio fingerprinting system. In: *Fourth IEEE Benelux Signal Processing Symposium*, pages 101–104, 2004.

[DMS2000] J. Dittmann, A. Mukherjee, and M. Steinebach. Media-independent watermarking classification and the need for combining digital video and audio watermarking for media authentication. In: *Information Technology: Coding and Computing, 2000. Proceedings. International Conference on*, pages 62–67, 2000.

[DS2013] P. K. Dhar and T. Shimamura. An SVD-based audio watermarking using variable embedding strength and exponential-log operations. In: *Informatics, Electronics Vision (ICIEV), 2013 International Conference on*, pages 1–6, 2013.

[DSLZ2004] Jana Dittmann, Martin Steinebach, Andreas Lang, and Sascha Zmudzinski. Advanced audio watermarking benchmarking. In: *SPIE Int. Symposium on Electronic Imaging, Security and Watermarking of Multimedia Contents, San Jose, USA*, 2004.

[Dwo2005] Morris J. Dworkin. SP 800-38B. Recommendation for Block Cipher Modes of Operation: The CMAC Mode for Authentication. Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, MD, United States, 2005.

[EBU2004] EBU. EBU Tests of Commercial Watermarking Systems. BBC White Paper WHP 101, European Broadcasting Unit (EBU), January 2004. (also published at 116th AES Convention, Audio Engineering Society, May 2004).

[EBU2009] EBU. MBWF/RF64: An Extended File Format for Audio – A BWF-compatible Multichannel File Format Enabling File Sizes to Exceed 4 Gbyte. Technical Specification EBU Tech 3306-2009, European Broadcasting Unit (EBU), July 2009.

[EJ2001] D. Eastlake, 3rd and P. Jones. *RFC 3174: US Secure Hash Algorithm 1 (SHA1)*. Number 3174 in Request for Comments (RFC). Internet Engineering Task Force (IETF), September 2001. (updated by RFC 4634, RFC 6234).

[ES2009] Yousof Erfani and Shadi Siahpoush. Robust Audio Watermarking Using Improved TS Echo Hiding. *Digital Signal Processing*, 19(5):809–814, September 2009.

[Far2009] Hany Farid. Seeing is not believing. *IEEE Spectrum*, 46(8):44–51, August 2009.

[Faw2006] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[Fec1860] Gustav Theodor Fechner. *Elemente der Psychophysik (Erster Theil)*. Breitkopf und Härtel, Leipzig, 1860.

[FG1999] J. Fridrich and M. Goljan. Images with self-correcting capabilities. In: *International Conference on Image Processing (ICIP 99)*, volume 3, pages 792–796 vol.3, 1999.

[FG2000] J. Fridrich and M. Goljan. Robust hash functions for digital watermarking. In: *Proc. of International Conference on Information Technology: Coding and Computing, 2000 (ITCC 2000), Las Vegas, Nevada, USA, March 27-29, 2000, pp. 173-178*, pages 178–183, March 2000.

[FGD2001] J. Fridrich, M. Goljan, and Rui Du. Invertible authentication watermark for jpeg images. In: ITTC 2001 [ITT2001], pages 223–227.

[FJ2005] M. Frigo and S.G. Johnson. The Design and Implementation of FFTW3. *Proceedings of the IEEE (Special Issue on Program Generation, Optimization, and Platform Adaptation)*, 93(2):216–231, Feb 2005.

[FKK2004] Chuhong Fei, Deepa Kundur, and Raymond H. Kwong. Analysis and design of authentication watermarking. In: Edward J. Delp and Ping Wah Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents*, volume 5306 of *Proceedings of SPIE*, pages 760–771. SPIE, 2004.

[FLWL2013] Ming-Quan Fan, Pei-Pei Liu, Hong-Xia Wang, and Heng-Jian Li. A semi-fragile watermarking scheme for authenticating audio signal based on dual-tree complex wavelet transform and discrete cosine transform. *Int. J. Comput. Math.*, 90(12):2588–2602, December 2013.

[FM1933] Harvey Fletcher and W.A. Munson. Loudness, its definition, measurement, and calculation. *Journal of the Acoustical Society of America*, 5:82–108, 1933.

[FM2012] M. Fallahpour and D. Megias. High capacity logarithmic audio watermarking based on the human auditory system. In: *IEEE International Symposium on Multimedia (ISM)*, pages 28–31, 2012.

[FM2015] M. Fallahpour and D. Megías. Audio watermarking based on fibonacci numbers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8):1273–1282, Aug 2015.

[Fou1822] Jean Baptiste Joseph Fourier. *Théorie analytique de la chaleur*. Chez Firmin Didot, père et fils, 1822.

[Fri2002] Jessica Fridrich. Security of fragile authentication watermarks with localization. *Proc. SPIE*, 4675:691–700, 2002.

[Gao2005] Wen Gao. AVS standard - Audio Video Coding Standard Workgroup of China. In: *14th Annual International Conference on Wireless and Optical Communications, 2005. WOCC 2005*, pages 54–, April 2005.

[Gau1866] Carl Friedrich Gauss. Nachlass: Theoria interpolationis methodo nova tractata. In: *Werke*, volume 3, pages 265–330. Königliche Gesellschaft der Wissenschaften, Göttingen, 1866. (posthumously).

[GBYK2013] A. Ghobadi, A. Boroujerdizadeh, A.H. Yaribakht, and R. Karimi. Blind audio watermarking for tamper detection based on lsb. In: *Advanced Communication Technology (ICACT), 2013 15th International Conference on*, pages 1077–1082, 2013.

[GCdG+2002] E. Gomez, P. Cano, L. de Gomes, E. Batlle, and M. Bonnet. Mixed watermarking-fingerprinting approach for integrity verification of audio recordings. In: *International Telecommunications Symposium ITS2002, Natal, Brazil*, 2002.

[Ger1981] Jan J. Gerbrands. On the relationships between SVD, KLT and PCA. *Pattern Recognition*, 14(1-6):375–381, 1981.

[GK1942] A. George and M.H. Kiesler. Secret communication system, 1942. US Patent 2,292,387.

[Gla1982] P. G. W. Glare, editor. *Oxford Latin Dictionary*. Oxford University Press, 1982. (via online access at "Latdict – latin-dictionary.net" project (retrieved May 2015).

[GLB1996] Daniel Gruhl, Anthony Lu, and Walter Bender. Echo hiding. In: Anderson [And1996], pages 295–315.

[GM2010] Michael Gulbis and Erika Müller. Content-based audio authentication using a hierarchical patchwork watermark embedding. In: *SPIE Photonics Europe*, pages 77230N–77230N. International Society for Optics and Photonics, 2010.

[GMS2008a] M. Gulbis, E. Muller, and M. Steinebach. Content-based authentication watermarking with improved audio content feature extraction. In: *International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008 (IIHMSP '08)*, pages 620 –623, aug. 2008.

[GMS2008b] Michael Gulbis, Erika Müller, and Martin Steinebach. Content-based authentication watermarking with improved audio content feature extraction. In: *Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08 International Conference on*, pages 620–623. IEEE, 2008.

[GS2012] Catalin Grigoras and Jeff M. Smith. Advances in ENF Analysis for Digital Media Authentication. In: *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*, Jun 2012.

[GSM+2012] Gary Grutzek, Julian Strobl, Bernhard Mainka, Frank Kurth, Christoph Poerschmann, and Heiko Knospe. Perceptual hashing for the identification of telephone speech. In: *10. ITG Symposium Proceedings of Speech Communication*, pages 1–4, September 2012.

[Gul2013] Michael Gulbis. *Authentifizierung von Audiodaten mittels inhalts-fragiler Wasserzeichen*. PhD thesis, Universitaet Rostock, Germany, Fakultaet fuer Informatik und Elektrotechnik, 2013. ISBN 3832225072.

[GVL1996] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[Haa1910] Alfred Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[HALT2013] Nawal Hamdouni, Abdellah Adib, Sonia Djaziri Larbi, and Monia Turki. A blind digital audio watermarking scheme based on emd and uisa techniques. *Multimedia Tools Appl.*, 64(3):809–829, June 2013.

[Ham1950] Richard Wesley Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 26(2):147–160, 1950.

[HBMS2007] Neil J. Hurley, Felix Balado, Elizabeth P. Mccarthy, and Guenole C. M. Silvestre. Performance of philips audio fingerprinting under desynchronisation. In: *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 133–134, Vienna, Austria, September 23-27 2007.

[HC2012] Hwai-Tsu Hu and Wei-Hsi Chen. A dual cepstrum-based watermarking scheme with self-synchronization. *Signal Processing*, 92(4):1109 – 1116, 2012.

[He2008] Xing He. *Watermarking in Audio: Key Techniques and Technologies*. Cambria Press, London, UK / Amherst, NY, USA, 2008.

[Hel1863] H. v. Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg, Braunschweig, 1st edition, 1863.

[Hem1961] Emil Frank Hembrooke. Identification of sound and like signals, 1961. United States Patent, 3004104.

[HEN2010] Xuping Huang, Isao Echizen, and Akira Nishimura. A new approach of reversible acoustic steganography for tampering detection. In: *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, pages 538–542. IEEE, 2010.

[HEN2011] Xuping Huang, Isao Echizen, and Akira Nishimura. A reversible acoustic steganography for integrity verification. In: HJ. Kim, Y.Q. Shi, and M. Barni, editors, *International Workshop on Digital Watermarking (IWDW 2010)*, volume 6526 of *Lecture Notes in Computer Science*. IEEE, Springer, Berlin, Heidelberg, 2011.

[HG1997] F. Hartung and B. Girod. Fast public-key watermarking of compressed video. In: *Proc. IEEE Int. Conf. on Image Processing*, pages 528–531 vol.1. IEEE, October 1997.

[HGT2015] Guang Hua, Jonathan Goh, and Vrizlynn L. L. Thing. Time-spread echo-based audio watermarking with optimized imperceptibility and robustness. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(2):227–239, February 2015.

[HH2015] Hwai-Tsu Hu and Ling-Yuan Hsu. Robust, transparent and high-capacity audio watermarking in {DCT} domain. *Signal Processing*, 109:226 – 235, 2015.

[HHC2014] Hwai-Tsu Hu, Ling-Yuan Hsu, and Hsien-Hsin Chou. Variable-dimensional vector modulation for perceptual-based {DWT} blind audio watermarking with adjustable payload capacity. *Digital Signal Processing*, 31:115 – 123, 2014.

[HHS+2016] Guang Hua, Jiwu Huang, Yun Q. Shi, Jonathan Goh, and Vrizlynn L.L. Thing. Twenty years of digital audio watermarking — a comprehensive review. *Signal Processing*, 128:222 – 242, 2016.

[HJB1985] M. T. Heideman, D. H. Johnson, and C. S. Burrus. Gauss and the history of the fast Fourier transform. *Archive for History of Exact Sciences Magazine (ASSP), IEEE*, 34(3):265–277, October 1985.

[HK1999] Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings of the IEEE, special issue on protection of multimedia content*, 87(7):1079–1107, July 1999. invited paper.

[HM2000] M. Holliman and N. Memon. Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. *Trans. Img. Proc.*, 9(3):432–441, March 2000.

[HOEN2014] Xuping Huang, Nobutaka Ono, Isao Echizen, and Akira Nishimura. Reversible audio information hiding based on integer coefficients with adaptive hiding locations. In: Y. Shi, HJ. Kim, and F. Pérez-González, editors, *International Workshop on Digital Watermarking (IWDW 2013), Digital-Forensics and Watermarking, revised papers*, volume 8389 of *Lecture Notes in Computer Science*, pages 376–389. Springer, Berlin, Heidelberg, 2014.

[HOK2001a] J.A. Haitsma, J.C. Oostveen, and A.A.C. Kalker. A highly robust audio fingerprinting system. In: ISMIR [ISM2001].

[HOK2001b] J.A. Haitsma, J.C. Oostveen, and A.A.C. Kalker. Robust audio hashing for content identification. In: *Content based multimedia Indexing (CBMI) 2001, Brescia Italy*, 2001.

[HPG2009] Helge Hundacker, Daniel Pähler, and Rüdiger Grimm. URM – usage rights management. In: Rüdiger Grimm, editor, *7th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorp. the 5th International ODRL Workshop (VIRTUAL GOODS 2009), September 22, 2009, Nancy, France*, 2009.

[HQ2002] Gael Hachez and Jean-Jaques Quisquater. Which directions for asymmetric watermarking? In: *XI European Signal Processing Conference*, 2002.

[HS2005] P. Hoffman and B. Schneier. *RFC 4270: Attacks on Cryptographic Hashes in Internet Protocols*. Number 4270 in Request for Comments (RFC). Internet Engineering Task Force (IETF), November 2005.

[HS2006] Enrico Hauer and Martin Steinebach. Temporal synchronization of marked mpeg video frames based on image hash system. In: *Security, Steganography, and Watermarking of Multimedia Contents VIII. Edited by Delp, Edward J., III; Wong, Ping Wah. Proceedings of the SPIE, Volume 6072, pp. 362-372 (2006).*, volume 6072, pages 607219–607219–9, 2006.

[HSWZ2013] Oren Halvani, Martin Steinebach, Patrick Wolf, and Ralf Zimmermann. Natural language watermarking for german texts. In: ACM, editor, *Proceedings of The 1st ACM Workshop on Information Hiding and Multimedia Security (IH & MMSEC 2013), June 17-19, 2013 Montpellier, France*, June 2013.

[HvdVKB2000] Jaap Haitsma, Michiel van der Veen, Ton Kalker, and Fons Bruekers. Audio watermarking for monitoring and copy protection. In: *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA '00, pages 119–122, New York, NY, USA, 2000. ACM.

[IHC2016] IHC. Evaluation Procedure for Audio Watermark Competition (ver. 5). In: *IHC Evaluation Criteria and Competition*. IHC Committee, chaired by Keiichi Iwamura, Tokyo University of Science, project website: *http://www.ieice.org/iss/emm/ihc/en*, May 2016.

[ISM2001] *2nd International Symposium of Music Information Retrieval (ISMIR 2001), Indiana University, Bloomington, Indiana, USA October 15-17, 2001*, 2001.

[ISO1993a] ISO. Appendix D: ”Psychoacoustic Models”, In: MPEG-1 – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 MBit/s, Part 3: Audio. ISO norm 11172-3 (Appendix D), International Organization for Standardization (ISO), Geneva, Switzerland, May 1993.

[ISO1993b] ISO. MPEG-1 – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 MBit/s – Part 3: Audio. ISO norm 11172-3, International Organization for Standardization (ISO), Geneva, Switzerland, May 1993.

[ISO2000] ISO. MPEG-2 – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC). ISO norm 13818-7, International Organization for Standardization (ISO), Geneva, Switzerland, October 2000.

[ISO2002] ISO. Information technology, Multimedia Content Description Interface, Part 4: Audio. ISO norm 15938-4 (a.k.a. ”MPEG-7”), International Organization for Standardization (ISO), Geneva, Switzerland, 2002.

[ISO2003] ISO. Acoustics – Normal Equal-loudness-level Contours. ISO norm 226, International Organization for Standardization (ISO), Geneva, Switzerland, 2003.

[ITT2001] *Proceedings of the International Conference on Information Technology: Coding and computing (ITTC 2001),2–4 April 2001, Las Vegas, USA*. IEEE, April 2001.

[ITU1988] ITU. Pulse code modulation (pcm) of voice frequencies. ITU-R Recommendation G.711, International Telecommunications Union (ITU), 1988.

[ITU1998] ITU. Method for objective measurements of perceived audio quality. ITU-R Recommendation BS.1387, International Telecommunications Union (ITU), 1998.

[JJN2009] Yuhua Jiao, Liping Ji, and XiaMu Niu. Robust speech hashing for content authentication. *Signal Processing Letters, IEEE*, 16(9):818–821, Sept 2009.

[JLLN2008] Yuhua Jiao, Mingyu Li, Qiong Li, and XiaMu Niu. Key-dependent compressed domain audio hashing. In: *Intelligent Systems Design and Applications, 2008. ISDA ’08. Eighth International Conference on*, volume 3, pages 29–32, Nov 2008.

[JLN2008] Yuhua Jiao, Qiong Li, and Xiamu Niu. Compressed Domain Perceptual Hashing for MELP Coded Speech. In: *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, 2008.

[Joh1988] J.D. Johnston. Estimation of perceptual entropy using noise masking criteria. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2524–2527. IEEE, 1988.

[JYL+2009] Dalwon Jang, C.D. Yoo, Sunil Lee, Sungwoong Kim, and T. Kalker. Pairwise boosted audio fingerprint. *IEEE Transactions on Information Forensics and Security*, 4(4):995–1004, Dec 2009.

[JYLN2007] Yuhua Jiao, Bian Yang, Mingyu Li, and XiaMu Niu. MDCT-Based Perceptual Hashing for Compressed Audio Content Identification. In: *IEEE 9th Workshop on Multimedia Signal Processing, 2007 (MMSP 2007)*, pages 381–384, Oct 2007.

[KA2015] Mohammed Khalil and Abdellah Adib. Informed audio watermarking based on adaptive carrier modulation. *Multimedia Tools and Applications*, 74(15):5973–5993, April 2015.

[Kal2001] T. Kalker. Considerations on watermarking security. In: *IEEE Fourth Workshop on Multimedia Signal Processing*, pages 201–206, 2001.

[KB2013] K. Khaldi and A. O. Boudraa. Audio watermarking via emd. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):675–680, March 2013.

[KBC1997] H. Krawczyk, M. Bellare, and R. Canetti. *RFC 2104: HMAC – Keyed-Hashing for Message Authentication*. Number 2104 in Request for Comments (RFC). Internet Engineering Task Force (IETF), February 1997.

[Ker1883] Auguste Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, IX:5–38, 161–191, 1883.

[KH1990] A. Khotanzad and H. Hong, Y.˙ Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, May 1990.

[KM2003] Darko Kirovski and Henrique S. Malvar. Spread-spectrum watermarking of audio signals. *IEEE Transcations on Signal Processing*, 51(4):1020–1032, April 2003.

[Kno2013] Heiko Knospe. Privacy-enhanced perceptual hashing of audio data. In: *The 11th International Conference on Security and Cryptography (SECRYPT 2014), Vienna, Austria, 28-30 August 2014*, pages 549–554, 2013.

[Kor2006a] Jesse Kornblum. Fuzzy hashing. In: *Digital Forensics Research Conference (DFRWS)*, 2006. Presentation slides for [Kor2006b];
available online: *http://www.dfrws.org/2006/proceedings/12-Kornblum-pres.pdf*.

[Kor2006b] Jesse Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3:91–97, September 2006.

[KP2000] S. Katzenbeisser and F.A.P. Petitcolas. *Information Hiding Techniques for Steganography and Digital Watermarking: Stefan Katzenbeisser, Fabien A.P. Petitcolas, Editors*. Computer Security Series. Artech House, 2000.

[KY2001] Khalid A. Kaabneh and Abdou Youssef. Muteness-based audio watermarking technique. In: *Proceedings of the 21st International Conference on Distributed Computing Systems Workshop (*ICDCSW *'01)*, 2001. 0 -7695 -1080-9/01 IEEE.

[LA2008] Yiqing Lin and Waleed H Abdulla. Multiple scrambling and adaptive synchronization for audio watermarking. In: Y.Q. Shi, HJ. Kim, and S. Katzenbeisser, editors, *International Workshop on Digital Watermarking (IWDW 2007)*, volume 5041 of *Lecture Notes in Computer Science*, pages 440–453. Springer, Berlin, Heidelberg, 2008.

[Lak1976] J. Radford Lakey. Temporal masking level differences: The effect of mask duration. *The Journal of the Acoustical Society of America*, 59(6):1434–1442, 1976.

[Lam1997] D. Laming. *The Measurement of Sensation*. Oxford Psychology Series. Oxford University Press, Oxford, UK, 1997.

[Lan1995] Chris A. Lanciani. Auditory perception and the mpeg audio standard. Technical report, Georgia Institute of Technology, School of Electrical and Computer Engineering, 1995.

[LAOVdK2003] A.N. Lemma, J. Aprea, W. Oomen, and L. Van de Kerkhof. A temporal domain audio watermarking technique. *Signal Processing, IEEE Transactions on*, 51(4):1088–1097, 2003.

[LB2012] Thomas Lenarz and Hans-Georg Boenninghaus. *Hals-Nasen-Ohren-Heilkunde*. Springer-Lehrbuch. Springer, 14th edition, 2012.

[LDS2003] Andreas Lang, Jana Dittmann, and Martin Steinebach. Psycho-akustische modelle für stirmark benchmark - modelle zur transparenzevaluierung. In: *GI Jahrestagung (Schwerpunkt "Sicherheit - Schutz und Zuverlässigkeit")*, pages 399–410, 2003.

[Lee2014] Suk-Hwan Lee. DNA sequence watermarking based on random circular angle. *Digit. Signal Process.*, 25:173–189, February 2014.

[LFL2010] Bai-Ying Lei, Jian Feng, and Kwok-Tung Lo. *Digital Watermarking Techniques for AVS Audio*. INTECH Open Access Publisher, 2010.

[Li2006] Yingjiu Li. Publicly verifiable ownership protection for relational databases. In: *Proceedings of the 2006 ACM Symposium on information, Computer and Communications Security*, pages 78–89, 2006.

[Liu2008] Huajian Liu. *Digital Watermarking for Image Content Authentication*. PhD thesis, TU Darmstadt, Germany, September 2008.

[LLC2000] Chun-Shien Lu, Hong-Yuan Mark Liao, and Liang-Hua Chen. Multipurpose audio watermarking. In: *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 282–285 vol.3, 2000.

[LMP2004] R. Lancini, F. Mapelli, and R. Pezzano. Audio content identification by using perceptual hashing. In: *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 1, pages 739–742 Vol.1, 2004.

[LMW1998] C.U. Lee, K. Moallemi, and R.L. Warren. Method and apparatus for transporting auxiliary data in audio signals, October 13 1998. US Patent 5,822,360.

[Log2000] B. Logan. Mel frequency cepstral coefficients for music modeling. In: *1st International Symposium of Music Information Retrieval (ISMIR 2000), Plymouth, Massachusetts October 23-25, 2000*, Online proceeding only: `http://ciir.cs.umass.edu/music2000/` (URL retrieved October 12th, 2013), 2000.

[Loh2008] Janina Lohrmann. *Urheberrechtsverletzungen im Web 2.0*. Grin Verlag, Muenchen, 2008.

[LS1879] Charlton T. Lewis and Charles Short. *Latin Dictionary: Founded on Andrews' edition of Freund's Latin dictionary*. Oxford University Press / Clarendon Press, Oxford, UK, 1879. (via online access at "Perseus Digital Library" at Tufts University, Medford, MA, USA (retrieved Aug. 2014)).

[LS2006a] Huajian Liu and M. Steinebach. Semi-fragile watermarking for image authentication with high tampering localization capability. In: *Automated Production of Cross Media Content for Multi-Channel Distribution, 2006. AXMEDIS '06. Second International Conference on*, pages 143–152, Dec 2006.

[LS2006b] Huajian Liu and Martin Steinebach. Digital watermarking for image authentication with localization. In: *Proceedings of the International Conference on Image Processing, ICIP 2006, October 8-11, Atlanta, Georgia, USA*, pages 1973–1976, 2006.

[LST2013] B. Lei, I. Y. Soon, and E. L. Tan. Robust svd-based audio watermarking scheme with differential evolution optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2368–2378, Nov 2013.

[LSZ+2012] Baiying Lei, Ing Yann Soon, Feng Zhou, Zhen Li, and Haijun Lei. A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition. *Signal Processing*, 92(9):1985 – 2001, 2012.

[LVRY2010] Claudio Lucchese, Michail Vlachos, Deepak Rajan, and Philip S Yu. Rights protection of trajectory datasets with nearest-neighbor preservation. *The International Journal on Very Large Data Bases*, 19(4):531–556, 2010.

[LY2000] Xin Li and Heather Yu. Transparent and robust audio data hiding in subband domain. In: *International Conference on Information Technology: Coding and Computing (ITCC 2000)*, pages 74–79, September 2000.

[LYH2014] Da Luo, Rui Yang, and Jiwu Huang. Detecting double compressed amr audio using deep learning. In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2669–2673, May 2014.

[LYK2009] Yu Liu, Hwan Sik Yun, and Nam Soo Kim. Audio Fingerprinting Based on Multiple Hashing in DCT Domain. *Signal Processing Letters, IEEE*, 16(6):525–528, June 2009.

[LZ2016] Haiyan Liu and Xiong Zenggang. Overview of Multifunctional Audio Watermarking based on Quantization. *International Journal of Simulation – Systems, Science and Technology*, 17(21):14.1–14.6, 2016.

[LZW2012] Wei Li, Bilei Zhu, and Zhurong Wang. On the music content authentication. In: *Proceedings of the 20th ACM international conference on Multimedia*, pages 1101–1104. ACM, 2012.

[LZW2013] Wei Li, Xiu Zhang, and Zhurong Wang. Music content authentication based on beat segmentation and fuzzy classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–13, 2013.

[Mad2013] Ian Maddieson. Vowel quality inventories/consonant inventories. In: Matthew S. Dryer and Martin Haspelmath, editors, *World Atlas of Language Structures (WALS Online)*. Max Planck Institute for Evolutionary Anthropology, 2013. URL: http://wals.info/.

[Mal1999] S.G. Mallat. *A Wavelet Tour of Signal Processing*. Referex Engineering. Academic Press, 1999.

[Men1995] J. Mendel. *Lessons in estimation theory for signal processing, communications, and control*. Prentice Hall, Englewood Cliffs, NJ, USA, 1995.

[Mer2007] Till Merlé. Optimiertes Auslesen digitaler Audiowasserzeichen. Diplomarbeit (student thesis), Technische Universität Darmstadt, Department of Computer Science, Supervisor: Claudia Eckert, August 2007.

[MF2010] H. Malik and H. Farid. Audio forensics from acoustic reverberation. In: *IEEE International Conference on Acoustics Speech and Signal Processing 2010 (ICASSP 2010)*, pages 1710–1713, 2010.

[MHJM2004a] David Megias, Jordi Herrera-Joancomarti, and Julia Minguillon. A robust frequency domain audio watermarking scheme for monophonic and stereophonic PCM formats. In: *Euromicro Conference*, pages 449–452. IEEE, 2004.

[MHJM2004b] David Megías, Jordi Herrera-Joancomartí, and Julià Minguillón. An audio watermarking scheme robust against stereo attacks. In: *Proceedings of the 2004 workshop on Multimedia and security*, pages 206–213. ACM, 2004.

[MHJM2005a] David Megías, Jordi Herrera-Joancomartí, and Julià Minguillón. Robust frequency domain audio watermarking: a tuning analysis. In: I.J. Cox, T. Kalker, and Lee HK., editors, *International Workshop on Digital Watermarking (IWDW 2004)*, volume 3304 of *Lecture Notes in Computer Science*, pages 244–258. Springer, Berlin, Heidelberg, 2005.

[MHJM2005b] David Megías, Jordi Herrera-Joancomartí, and Julià Minguillón. Total disclosure of the embedding and detection algorithms for a secure digital watermarking scheme for audio. In: *International Conference on Information and Communications Security*, pages 427–440. Springer, 2005.

[MHJSM2006] David Megias, Jordi Herrera-Joancomart, Jordi Serra, and Julia Minguillon. A benchmark assessment of the wauc watermarking audio algorithm. In: *Electronic Imaging 2006. Security, Steganography, and Watermarking of Multimedia Contents VIII*. International Society for Optics and Photonics, 2006.

[MLL2008] Xiaohong Ma, Xin Li, and Wenlong Liu. A new audio watermarking method for copyright protection and tampering localization. In: *Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on*, page 358, June 2008.

[MMBK2010] D Mills, J. Martin, J. Burbank, and W. Kasch. *RFC 5905: Network Time Protocol Version 4: Protocol and Algorithms Specification*. Number 5905 in Request for Comments (RFC). Internet Engineering Task Force (IETF), 2010. (obsolets RFC 1305, RFC 4330).

[MNS2001] M. Monsignori, P. Nesi, and M.B. Spinu. Watermarking music sheets. In: Heung-Yeung Shum, Mark Liao, and Shih-Fu Chang, editors, *Advances in Multimedia Information Processing (PCM 2001)*, volume 2195 of *Lecture Notes in Computer Science*, pages 646–653. Springer Berlin Heidelberg, 2001.

[Moo1995] Brian C.J. Moore, editor. *Hearing*. Handbook of Perception and Cognition (2nd ed.). Academic Press Inc., San Diego, CA, USA, 1995.

[Moo2005] Todd K. Moon. *Error Correction Coding: Mathematical Methods and Algorithms*. Wiley-Interscience, 2005.

[Mun2011] Badar Munir. Perception-based verification of audio data. Master thesis, Hochschule Darmstadt, Fachbereich Elektrotechnik und Informationstechnik, Supervisor: Manfred Götze, 2011.

[MV2001] M. Kıvanç Mıçak and Ramarathnam Venkatesan. A perceptual audio hashing algorithm: A tool for robust audio identification and information hiding. In: I.S. Moskowitz, editor, *Lecture Notes in Computer Science, 4th International Workshop Information Hiding, IH 2001, Pittsburgh, PA, USA, April 25-27, 2001, ISBN 3540427333*, volume 2137, August 2001.

[MVO1996] Alfred J. Menezes, Scott A. Vanstone, and Paul C. Van Oorschot. *Handbook of Applied Cryptography*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1996.

[MW2012] M. Malekesmaeili and R.K. Ward. A novel local audio fingerprinting algorithm. In: *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 136–140, Sept 2012.

[MW2014] Mani Malekesmaeili and Rabab K. Ward. A local fingerprinting approach for audio copy detection. *Signal Processing*, 98:308–321, May 2014.

[MWe2009] Merriam-Webster Online Dictionary, 2009. http://www.merriam-webster.com.

[NAK+2016] Sabry S. Nassar, Nabil M. Ayad, Hamdy M. Kelash, Hala S. El-sayed, Mohsen A. M. El-Bendary, Fathi E. Abd El-Samie, and Osama S. Faragallah. Efficient audio integrity verification algorithm using discrete cosine transform. *International Journal of Speech Technology*, 19(1):1–8, 2016.

[NK2011] Hieu Cuong Nguyen and Stefan Katzenbeisser. Security of copy-move forgery detection techniques. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing 2011 (ICASSP 2011)*, pages 1864–1867, May 2011.

[NK2016] V. Neethu and R. Kalaivani. Efficient and robust audio watermarking for content authentication and copyright protection. In: *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–6, March 2016.

[NQ1998] Klara Nahrstedt and Lintian Qiao. Non-invertible watermarking methods for mpeg-encoded audio. Technical report, Department of Computer Science, University of Illinois, USA, June 1998.

[NTNB2006] Yuta Nakashima, Ryuki Tachibana, Masafumi Nishimura, and Noboru Babaguchi. Estimation of recording location using audio watermarking. In: *8th workshop on Multimedia and security (MM&Sec '06:)*, pages 108–113, New York, NY, USA, 2006. ACM Press.

[NU2015] N. M. Ngo and M. Unoki. Robust and reliable audio watermarking based on phase coding. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 345–349, April 2015.

[Nut2012] M Nutzinger. Real-time attacks on audio steganography. *Journal of Information Hiding and Multimedia Signal Processing*, 3(1):47–65, 2012.

[NXE+2013] I. Natgunanathan, Yong Xiang, S.S.M. Elbadry, Wanlei Zhou, and Yang Xiang. Analysis of a patchwork-based audio watermarking scheme. In: *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on*, pages 900–905, 2013.

[NZAF2012] M. Nouri, Z. Zeinolabedini, B. Abdolmaleki, and N. Farhangian. Analysis of a novel audio hash function based upon stationary wavelet transform. In: *Application of Information and Communication Technologies (AICT), 2012 6th International Conference on*, pages 1–6, Oct 2012.

[ODG2014] C. Ouali, P. Dumouchel, and V. Gupta. A robust audio fingerprinting method for content-based copy detection. In: *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2014.

[OKH2001] Job Oostveen, Ton Kalker, and Jaap Haitsma. Visual hashing of video: application and techniques. In: Wong and Delp III [WDI2001].

[OS2004] A. V. Oppenheim and R. W. Schafer. DSP history - From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95–106, September 2004.

[OSMA2005] Hamza Özer, Bülent Sankur, Nasir Memon, and Emin Anarim. Perceptual audio hashing functions. *EURASIP Journal of Applied Signal Processing*, 2005:1780–1793, January 2005.

[PAK1998] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In: David Aucsmith, editor, *Lecture Notes in Computer Science Second Workshop on Information Hiding, Portland, Oregon, USA, April 14–17, 1998, ISBN 3540653864*, volume 1525, pages 218–238, September 1998.

[PB1961] W.W. Peterson and D.T. Brown. Cyclic codes for error detection. *Proceedings of the IRE*, 49(1):228–235, January 1961.

[Pet2000] Fabien A. P. Petitcolas. Watermarking schemes evaluation. *IEEE Signal Processing*, 17(5):58–64, 2000.

[Pet2005] R. Petrovic. Digital watermarks for audio integrity verification. In: *Telecommunications in Modern Satellite, Cable and Broadcasting Services, 2005. 7th International Conference on*, volume 1, pages 215–220, 2005.

[Pre1993] Bart Preneel. *Analysis and Design of Cryptographic Hash Functions*. PhD thesis, KU Leuven, Belgium, 1993.

[PS2000] Ted Painter and Andreas Spanias. Perceptual coding of digital audio. In: *Proceedings of the IEEE*, volume 88, no. 4, pages 451–513. IEEE, 2000.

[PTW2007] Chang-Mok Park, Devinder Thapaa, and Gi-Nam Wang. Speech authentication system using digital watermarking and pattern recovery. *Pattern Recognition Letters*, 28:931–938, June 2007.

[QKBD2010] Kun Qian, Christian Kraetzer, Michael Biermann, and Jana Dittmann. Audio annotation watermarking with robustness against DA/AD conversion. In: *IS&T/SPIE Electronic Imaging, Media Forensics and Security II, SPIE Proc. 7541*. International Society for Optics and Photonics, 2010.

[QZ2006] Xiaomei Quan and Hongbin Zhang. Data hiding in mpeg compressed audio using wet paper codes. In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 727–730, 2006.

[Rab1990] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In: *Readings in speech recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., 1990. ISBN 1558601244.

[RD1956] D.W. Robinson and R.S. Dadson. a redetermination of the equal-loudness relations for pure tones. *British Journal for Applied Physics*, 7:166–181, 1956.

[RFF2016] Ronaldo Rigoni, Pedro Garcia Freitas, and Mylène C.Q. Farias. Detecting tampering in audio-visual content using QIM watermarking. *Information Sciences*, 328:127 – 143, 2016.

[Rho2011] G.B. Rhoads. Digital watermarking video captured from airborne platforms. US Patent US8085976, December 27 2011.

[Riv1992] Ronald L. Rivest. *RFC 1321: The MD5 Message-Digest Algorithm*. Number 1321 in Request for Comments (RFC). Internet Engineering Task Force (IETF), April 1992.

[RKSZ2010] Ulrich Rührmair, Stefan Katzenbeisser, Martin Steinebach, and Sascha Zmudzinski. Watermark-based authentication and key exchange in teleconferencing systems. In: *11th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security (CMS 2010), 31 May - 02 June, 2010, Linz, Austria*, number 6109 in LNCS, Security and Cryptology. Springer, Mai 2010.

[RM2002] R. Radhakrishnan and N. D. Memon. Audio content authentication based on psychoacoustic model. In: *Proc. SPIE Vol. 4675, p. 110-117, Security and Watermarking of Multimedia Contents IV, Edward J. Delp; Ping W. Wong; Eds.*, pages 110–117, April 2002.

[Ros2009] John Ross. Preservation of Archival Sound Recordings. Technical Recommendations, Association for Recorded Sound Collections (ARSC), Technical Committee, April 2009.

[RS1960] I. S. Reed and G. Solomon. Polynomial Codes Over Certain Finite Fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

[RXM2003] R. Radhakrishnan, Z. Xiong, and N. Memom. On the security of the visual hash function. In: *Proc. SPIE, Security and Watermarking of Multimedia Contents V, San Jose, CA, 2003, vol. 5020*, 2003.

[S+1997] Stephen W. Smith et al. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Pub., San Diego, USA, 1997. Full access online at http://www.dspguide.com (retrieved August 2014).

[SAN2013] T. Shibuya, M. Abe, and M. Nishiguchi. Audio fingerprinting robust against reverberation and noise based on quantification of sinusoidality. In: *IEEE International Conference on Multimedia and Expo 2013 (ICME 2013)*, pages 1–6, July 2013.

[SBH+2010] Marcel Schäfer, Waldemar Berchtold, Margareta Heilmann, Sascha Zmudzinski, Martin Steinebach, and Stefan Katzenbeisser. Collusion secure fingerprint watermarking for real world applications. In: *SICHERHEIT 2010, Berlin*. Gesellschaft für Informatik, 2010.

[SBK+2017] Marc Stevens, Elie Bursztein, Pierre Karpman, Ange Albertini, Yarik Markov, Alex Petit Bianco, and Clement Baisse. Announcing the first SHA1 collision. *Google Security Blog, https://security.googleblog.com/2017/02/announcing-first-sha1-collision.html*, February 2017.

[SBZS2010] Marcel Schäfer, Waldemar Berchtold, Sascha Zmudzinski, and Martin Steinebach. Zero false positive 2-secure fingerprinting watermarking. In: *The 12th ACM Workshop on Multimedia and Security 2010 (ACM MMSEC2010), 08.-09. Sep 2010, Rome, Italy*, 2010.

[SBZS2012] Marcel Schäfer, Waldemar Berchtold, Sascha Zmudzinski, and Martin Steinebach. Verfahren zur Erzeugung von Transaktionswasserzeichen und Auswerteverfahren zur Kundenrückverfolgung. DPMA German Patent DE102010044228 A1, March 2012. (application: Sep. 2010, granted: Okt. 2012).

[SD2003a] Martin Steinebach and Jana Dittmann. Capacity-optimized mp2 audio watermarking. In: *Proceeding of SPIE*, volume 5020, pages 44–54, 2003.

[SD2003b] Martin Steinebach and Jana Dittmann. Watermarking-based digital audio data authentication. *EURASIP Journal on Applied Signal Processing*, 10:1001–1015, 2003.

[SDS+2001] Martin Steinebach, Jana Dittmann, Christian Seibel, Lucilla Croce Ferri, Fabien A.P. Petitcolas, Nazim Fates, Caroline Fontaine, and Frederic Raynal. Stirmark benchmark: Audio watermarking attacks. In: *2001 International Symposium on Information Technology (ITCC 2001), 02-04 April 2001, Las Vegas, NV, USA*, volume 00, pages 173–178, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

[Seo2014] J.S. Seo. An asymmetric matching method for a robust binary audio fingerprinting. *IEEE Signal Processing Letters*, 21(7):844–847, July 2014.

[SH2001] Christian Siebenhaar, Frankand Neubauer and Jürgen Herre. Combined compression / watermarking for audio signals. In: *110th AES Convention, May 12-15, 2001, Amsterdam, The Netherlands*. Audio engineering Society (AES), 2001.

[Sha1948] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[Sha1949] C. E. Shannon. Communication in the Presence of Noise. *Proceedings of the Institute of Radio Engineers (IRE)*, 37(1):10–21, January 1949.

[SJL+2005] J.S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, and C.D. Yoo. Audio fingerprinting based on normalized spectral subband centroids. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 (ICASSP 2005)*, volume 3, pages 213–216, March 2005.

[SMW2005] A. Swaminathan, Y. Mao, and M. Wu. Security of feature extraction in image hashing. In: *IEEE Conference on Acoustic, Speech and Signal Processing 2005 (ICASSP 2005)*, 2005.

[SSA+2008] Alexander Sotirov, Marc Stevens, Jacob Appelbaum, Arjen Lenstra, David Molnar, Dag A. Osvik, and Benne de Weger. MD5 Considered Harmful Today: Creating a rogue CA certificate. In: *25th Chaos Communication Congress (25C3), 27-30 Dec., 2008*, 2008.

[SSSS2008] M. Shirali-Shahreza and S. Shirali-Shahreza. Software watermarking by equation reordering. *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, pages 1–4, April 2008.

[ST2004] Yôiti Suzuki and Hisashi Takeshima. Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, 116(2):918–933, August 2004.

[Ste1961] S. S. Stevens. To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system. *Science (New York, NY, USA)*, 133(3446):80–86, January 1961.

[Ste2003] Martin Steinebach. *Digitale Wasserzeichen fuer Audiodaten*. PhD thesis, TU Darmstadt, Germany, 2003. ISBN 3832225072.

[Stö1995] Horst Stöcker, editor. *Taschenbuch mathematischer Formeln und moderner Verfahren*. Verlag Harri Deutsch, Thun/Frankfurt, Germany, 1995.

[SYZW2015] Martin Steinebach, York Yannikos, Sascha Zmudzinski, and Christian Winter. Advanced Multimedia File Carving. In: Anthony T.S. Ho and Shujun Li, editors, *Handbook of Digital Forensics of Multimedia Data and Devices*, chapter 6, pages 219–269. Wiley IEEE Press, September 2015.

[SZ2004a] Martin Steinebach and Sascha Zmudzinski. Complexity optimization of digital watermarking for music-on-demand services. In: *Technology, economy, social and legal aspects of virtual goods – 2nd International Workshop, Ilmenau, Germany May 27-29 2004 (VIRTUAL GOODS 2004)*. Technische Universität Ilmenau, 2004.

[SZ2004b] Martin Steinebach and Sascha Zmudzinski. Partielle Verschlüsselung von MPEG Audio. In: Patrick Horster, editor, *IT Security & IT Management, Proceeding D-A-CH Security 2004*, 2004.

[SZ2006a] Martin Steinebach and Sascha Zmudzinski. Countermeasure for collusion attacks against digital watermarking. In: *Electronic Imaging 2006 – Security, Steganography, and Watermarking of Multimedia Contents VIII*. IS&T SPIE, Jan 2006.

[SZ2006b] Martin Steinebach and Sascha Zmudzinski. Robustheit digitaler Audiowasserzeichen gegen Pitch-Shifting und Time-Stretching. In: *Sicherheit 2006: Sicherheit - Schutz und Zuverlässigkeit. Beiträge der 3. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)*. Köllen Verlag, 2006.

[SZ2007] Martin Steinebach and Sascha Zmudzinski. Countermeasure for collusion attacks in digital watermarking. EPO European Patent EP1739617 A1, January 2007. (application: July 2005, granted: Sep. 2014).

[SZ2008a] Martin Steinebach and Sascha Zmudzinski. Evaluation of robustness and transparency of multiple audio watermark embedding. In: III Delp, Edward J., Ping Wah Wong, Jana Dittmann, and Nasir D. Memon, editors, *SPIE Int., Symposium on Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, San Jose, USA*, volume 6819, 2008.

[SZ2008b] Martin Steinebach and Sascha Zmudzinski. Optimiertes Auslesen digitaler Audiowasserzeichen. In: *D-A-CH Security 2008: Bestandsaufnahme, Konzepte, Anwendungen, Perspektiven*. Syssec, Basel, 2008.

[SZ2009a] Martin Steinebach and Sascha Zmudzinski. Individuell gestempelt – Die Technik hinter digitalen Audio-Wasserzeichen. Number 9/2009 in c't-Magazin für Computertechnik, pages 142–146. Heise Verlag, Hannover, April 2009.

[SZ2009b] Martin Steinebach and Sascha Zmudzinski. Method for embedding a multi-bit digital watermark in media data. EPO European Patent EP2012269 A1, January 2009. (application: July 2007, granted: May 2011).

[SZB2005] Martin Steinebach, Sascha Zmudzinski, and Thorsten Boelke. Audio watermarking and partial encryption. In: Ping W. Wong Edward J. Delp III, editor, *Proceedings of SPIE - Volume 5681, Security, Steganography, and Watermarking of Multimedia Contents VII*. SPIE, Bellingham, USA, 2005.

[SZB2010] Martin Steinebach, Sascha Zmudzinski, and Moazzam Butt. Robust hash controlled watermark embedding. In: *32nd Annual Symposium of the German Association for Pattern Recognition (DAGM 2010) - Special Workshop on Pattern Recognition for IT Security, September 22-24, 2010, Darmstadt, Germany*. DAGM, Sep 2010.

[SZF2004] Martin Steinebach, Sascha Zmudzinski, and Chen Fan. The digital watermarking container: Secure and efficient embedding. In: *Proceedings of the ACM Multimedia and Security Workshop, 20.-21. September 2004, Magdeburg, Germany*, 2004.

[SZL2007] Martin Steinebach, Sascha Zmudzinski, and Huajian Liu. URL watermark as filter for on-line directories. EPO European Patent EP1739952 A1, January 2007. (application: July 2005, granted: April 2008).

[SZN2011] Martin Steinebach, Sascha Zmudzinski, and Stefan Nürnberger. Re-synchronizing audio watermarking after nonlinear time stretching. In: *Electronic Imaging 2011 - Media Watermarking, Security, and Forensics XIII*. IS&T SPIE, Jan 2011. (**Best Paper Award)**.

[SZP2012] Martin Steinebach, Sascha Zmudzinski, and Dirk Petrautzki. Forensic audio watermark detection. In: *Electronic Imaging 2012 - Media Watermarking, Security, and Forensics XIV*. IS&T SPIE, Jan 2012.

[SZSL2003] Martin Steinebach, Sascha Zmudzinski, Stefan Schäfer, and Andreas Lang. Robustheitsevaluierung digitaler Audio-Wasserzeichen im Rundfunkszenario. In: *2. Thüringer Medienseminar der FKTG – Rechte digitaler Medien, Erfurt, Germany*. Film- und Kinotechnische Gesellschaft FKTG, 2003.

[SZTB1998] Mitchell D. Swanson, Bin Zhu, Ahmed H. Tewfik, and Laurence Boney. Robust audio watermarking using perceptual masking. *Signal Processing*, 66(3):337–355, 1998.

[Tac2003] Ryuki Tachibana. Audio watermarking for live performance. In: *Proceeding of SPIE*, volume 5020, pages 32–43, 2003.

[TBSS2013] Daniel Trick, Waldemar Berchtold, Marcel Schäfer, and Martin Steinebach. 3D Watermarking in the Context of Video Games. In: *Proceeding IEEE 15th International Workshop on Multimedia Signal Processing 2013 (MMSP 2013), Pula, Italy, September 30 - October 2, 2013*. IEEE, October 2013.

[Tea1980] Michael Reed Teague. Image analysis via the general theory of moments∗. *J. Opt. Soc. Am.*, 70(8):920–930, Aug 1980.

[Tes1903] Nikolai Tesla. System of signaling., 1903. US Patent 725,605.

[TLMA2016] Omar Tayan, Lamri Laouamer, Tarek Moulahi, and Yasser M Alginahi. Authenticating sensitive speech-recitation in distance-learning applications using real-time audio watermarking. *International Journal of Advanced Computer Science and Applications*, 7(6):398–407, 2016.

[TNSZ2009] Stefan Thiemert, Stefan Nürnberger, Martin Steinebach, and Sascha Zmudzinski. Security of robust audio hashes. In: *IEEE International Workshop on Information Forensics and Security, WIFS 2009, London, UK*, Dez 2009.

[TTA2006] U. Topkara, M. Topkara, and M. J. Atallah. Hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions. In: *Voloshynovskiy, Dittmann, Fridrich (editors): Proceedings of the 8th workshop on Multimedia and Security, MMSec 2006, Geneva, Switzerland, September 26-27, 2006*, 2006.

[TTB+1998] Thilo Thiede, William C. Treurniet, Roland Bitto, Thomas Sporer, Karlheinz Brandenburg, Christian Schmidmer, Michael Keyhl, John G. Beerends, Catherine Colomes, Gerhard Stoll, and Bernhard Feiten. Peaq-der künftige itu-standard zur objektiven messung der wahrgenommenen audioqualität. *Bericht der 20. Tonmeistertagung*, 20, 1998.

[Tuk1977] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[UM2011] M. Unoki and R. Miyauchi. Reversible watermarking for digital audio based on cochlear delay characteristics. In: *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2011 Seventh International Conference on*, pages 314–317, 2011.

[UM2012] M. Unoki and R. Miyauchi. Detection of tampering in speech signals with inaudible watermarking technique. In: *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on*, pages 118–121, 2012.

[Ven2004] Ilaria Venturini. Counteracting oracle attacks. In: *MM&Sec '04: Proceedings of the 2004 workshop on Multimedia and security*, pages 187–192, New York, NY, USA, 2004. ACM Press.

[VPP+2001] S. Voloshynovskiy, S. Pereira, T. Pun, J. J. Eggers, and J. K. Su. Attacks on digital watermarks: Classification, estimation-based attacks, and benchmarks. *IEEE Communications Magazine*, 39:118–126, 2001.

[Wan2003] Avery Li-Chun Wang. An industrial strength audio search algorithm. In: *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 27-30, 2003*, 2003.

[Wan2006] Avery Li-Chun Wang. The Shazam music recognition service. *Communications of the ACM - Music Information Retrieval*, 49(8):44–48, August 2006.

[WC2008] Wen-Chih Wu and O.T.-C. Chen. Robust echo hiding scheme against pitch-scaling attacks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)*, pages 1737–1740, 2008.

[WCG2012] Jeff Warren, Michael Clear, and Ciaran Mc Goldrick. Metadata Independent Hashing for Media Identification & P2P Transfer Optimisation. In: *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012 International Conference on*, pages 58–65. IEEE, 2012.

[WDI2001] Ping Wah Wong and Edward P. Delp III, editors. *IS&T/SPIE 13th Int. Symposium on Electronic Imaging San Jose, Security and Watermarking of Multimedia Contents, CA, USA, Jan. 2001*, volume 4314. SPIE–The International Society for Optical Engeneering, 2001.

[Web1834] Ernst Heinrich Weber. *De Pulsu, Resorptione, Auditu et Tactu: Annotationes Anatomicae et physiologicae*. C. F. Koehler, Leipzig, 1834.

[WF2010] Hong-Xia Wang and Ming-Quan Fan. Centroid-based semi-fragile audio watermarking in hybrid domain. *Science China Information Sciences*, 53(3):619–633, 2010.

[WFM2005] Foo Say Wei, Xue Feng, and Li Mengyuan. A blind audio watermarking scheme using peak point extraction. In: *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 4409–4412. IEEE, 2005.

[Whe1859] W. Whewell. *History of the Inductive Sciences: VIII. Acoustics*, volume 2 of *History of the Inductive Sciences: From the Earliest to the Present Time*. D. Appleton and Company, New York, NY, USA, 3rd edition, 1859.

[Whi2004] C. White. Automatic location identification and categorization of digital photographs. US Patent US20040201702, October 14 2004.

[WHT2011] Jian Wang, Ron Healy, and Joe Timoney. A robust audio watermarking scheme based on reduced singular value decomposition and distortion removal. *Signal Processing*, 91(8):1693–1708, 2011.

[Wid1961] B. Widrow. Statistical analysis of amplitude-quantized sampled-data systems. *AIEE Transactions on Applications and Industry*, pages 1–14, January 1961.

[WK2001] Chung-Ping Wu and C.-C. Jay Kuo. Speech Content Integrity Verification Integrated with ITU G.723.1 Speech Coding. *itcc*, 00:0680, 2001.

[WLC2007] Ching-Te Wang, Chiu-Hsiung Liao, and Tung-shou Chen. Audio-signal authenticating system based on asymmetric signature schemes. *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, 00:656–661, 2007.

[WM2000] Ping Wah Wong and Nasir Memon. *Secret and public key authentication watermarking schemes that resist vector quantization attack*, volume 3971, pages 417–427. SPIE, 2000.

[WM2013] Nana Wang and Chaoguang Men. Reversible fragile watermarking for locating tampered blocks in 2d vector maps. *Multimedia Tools and Applications*, 67(3):709–739, 2013.

[WNSM2008] Huiqin Wang, Ryouichi Nishimura, Yoiti Suzuki, and Li Mao. Fuzzy self-adaptive digital audio watermarking based on time-spread echo hiding. *Applied Acoustics*, 69(10):868 – 874, 2008.

[WNY2009] Xiang-Yang Wang, Pan-Pan Niu, and Hong-Ying Yang. A robust digital audio watermarking based on statistics characteristics. *Pattern Recognition*, 42(11):3057–3064, 2009. cited By (since 1996)15.

[WSK2000] Chung-Ping Wu, Po-Chyi Su, and C.C. Jay Kuo. Robust and efficient digital audio watermarking using audio content analysis. In: *IS&T/SPIE 12th Int. Symposium on Electronic Imaging San Jose, Security and Watermarking of Multimedia Contents, CA, USA, Jan. 2000*, volume 3971, pages 382–392. SPIE–The International Society for Optical Engeneering, 2000.

[WSY2013] Christian Winter, Markus Schneider, and York Yannikos. F2S2: Fast forensic similarity search through indexing piecewise hash signatures. *Digital Investigation*, 10(4):361–371, 2013.

[WSZ2008] Patrick Wolf, Martin Steinebach, and Sascha Zmudzinski. Adaptive security for virtual goods – building an access layer for digital watermarking. In: Rüdiger Grimm, editor, *Virtual Goods. International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorp. the 4th International ODRL Workshop (VIRTUAL GOODS 2008) , October 16-18, 2008 Poznan, Poland*. Publishing House of Poznan University of Technology, 2008.

[WWZ+2013] Xinkai Wang, Pengjun Wang, Peng Zhang, Shuzheng Xu, and Huazhong Yang. A norm-space, adaptive, and blind audio watermarking algorithm by discrete wavelet transform. *Signal Processing*, 93(4):913 – 922, 2013.

[WYY2005] Xiaoyun Wang, Yiqun Lisa Yin, and Hongbo Yu. Finding Collisions in the Full SHA-1. In: Victor Shoup, editor, *Advances in Cryptology - CRYPTO 2005: 25th Annual International Cryptology Conference, Santa Barbara, California, USA, August 14-18, 2005, Proceedings*, volume 3621 of *Lecture Notes in Computer Science*, pages 17–36. Springer, 2005. (see remark in footnote below[78]).

[XLY2011] Xiangyang Xue, Wei Li, and Yue Yin. Towards content-based audio fragment authentication. In: *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1249–1252, New York, NY, USA, 2011. ACM.

[YbQyZt2013] Huang Yi-bo, Zhang Qiu-yu, and Yuan Zhan-ting. Algorithm for evaluating speech perceptual hash similarity after slight tampering occurs. *Information Technology Journal*, 12(16):3591, 2013.

[YH2004] Song Yuan and Sorin A. Huss. Audio watermarking algorithm for real-time speech integrity and authentication. In: *MM&Sec '04: Proceedings of the 2004 multimedia and security workshop on Multimedia and security*, pages 220–226, New York, NY, USA, 2004. ACM Press.

[YK2001] In-Kwon Yeo and Hyoung Joong Kim. Modified patchwork algorithm: A novel audio watermarking scheme. In: ITTC 2001 [ITT2001].

[YL2005] Chia-Mu Yu and Chun-Shien Lu. Robust non-interactive zero-knowledge watermarking scheme against cheating prover. In: *MM&Sec '05: Proceedings of the 7th workshop on Multimedia and security*, pages 103–110, New York, NY, USA, 2005. ACM.

[YWN2015] Shansan Yao, Yunsheng Wang, and Baoning Niu. An efficient cascaded filtering retrieval method for big audio data. *IEEE Transactions on Multimedia*, 17(9):1450–1459, Sept 2015.

[YWZ2009] Xiaoyuan Yang, Xiang Wu, and Minqing Zhang. Audio digital signature algorithm with tamper detection. In: *Information Assurance and Security, 2009. IAS '09. Fifth International Conference on*, volume 1, pages 15–18, 2009.

[YZLX2013] Litao Yu, Liehuang Zhu, Dan Liu, and Yuzhou Xie. A Novel Audio Information Hiding Scheme based on rMAC. In: *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, volume 212 of *Lecture Notes in Electrical Engineering*, pages 895–902. Springer, January 2013.

[78] The paper only contains an approach for decreasing the complexity from 2^{80} (brute force) to 2^{69}. The final reduction to 2^{63} was presented during the "Rump Session" by *A. Shamir* on behalf of the authors.

[Zer1934] Frederik 'Frits' Zernike. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. *Physica*, 1(7):689 – 704, 1934.

[ZF1990] E. Zwicker and H. Fastl. *Psychoacustics – Facts and Models*. Springer, Berlin, Heidelberg, New York, 2nd updated edition, 1990.

[Zha2015] Jinquan Zhang. Audio dual watermarking scheme for copyright protection and content authentication. *International Journal of Speech Technology*, 18(3):443–448, 2015.

[ZHHQ2016] Qiu-Yu Zhang, Wen-Jin Hu, Yi-Bo Huang, and Si-Bin Qiao. Research on universal model of speech perceptual hashing authentication system in mobile environment. In: *International Conference on Intelligent Computing*, pages 99–111. Springer, 2016.

[ZLY+2013] Liehuang Zhu, Dan Liu, Litao Yu, Yuzhou Xie, and Mingzhong Wang. Content integrity and non-repudiation preserving audio-hiding scheme based on robust digital signature. *Security and Communication Networks*, 6(11):1331–1343, 2013.

[ZMS2012] Sascha Zmudzinski, Badar Munir, and Martin Steinebach. Digital audio authentication by robust feature embedding. In: *Electronic Imaging 2012 - Media Watermarking, Security, and Forensics XIV*. IS&T SPIE, Jan 2012.

[ZS2003] Bin B. Zhu and Mitchell D. Swanson. Multimedia authentication and watermarking. In: David Dagan Feng, Wan-Chi Siu, and Hong-Jiang Zhang, editors, *Multimedia Information Retrieval and Management*, Signals and Communication Technology, pages 148–177. Springer Berlin Heidelberg, 2003.

[ZS2007] Sascha Zmudzinski and Martin Steinebach. Robust message authentication code algorithm for digital audio recordings. In: Ping Wah Wong Edward J. Delp III, editor, *Proceedings of SPIE Volume: 6505, Security, Steganography, and Watermarking of Multimedia Contents IX*. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham/Wash, 2007.

[ZS2008a] Sascha Zmudzinski and Martin Steinebach. Content-based message authentication coding for audio data. In: *Ammar Alkassar, Jörg Siekmann (Editors): Proceedings of "Sicherheit 2008 - Schutz und Zuverlässigkeit", 2.-4. April 2008, Saarbrücken, GI-Edition - Lecture Notes in Informatics (LNI), P-128. Köllen Verlag, Bonn*, 2008.

[ZS2008b] Sascha Zmudzinski and Martin Steinebach. Psycho-acoustic model-based message authentication coding for audio data. In: *10th ACM Workshop on Multimedia and Security (ACM MMSEC'08), Oxford, UK, September 22-23, 2008*, 2008.

[ZS2008c] Sascha Zmudzinski and Martin Steinebach. Robust audio hashing for audio authentication watermarking. In: Edward J. Delp III, editor, *Proceedings of SPIE: Security, Steganography, and Watermarking of Multimedia Contents X*, January 2008.

[ZS2009a] Hong Zhao and Dong-Sheng Shen. A new semi-fragile watermarking for audio authentication. In: *International Conference on Artificial Intelligence and Computational Intelligence, 2009. AICI '09*, volume 3, pages 299–302, Nov 2009.

[ZS2009b] Sascha Zmudzinski and Martin Steinebach. Perception-based audio authentication watermarking in the time-frequency domain. In: Stefan Katzenbeisser, editor, *Information Hiding Conference 2009*, LNCS. Springer, 2009.

[ZS2010] Hong Zhao and Dongsheng Shen. An audio watermarking algorithm for audio authentication. In: *Information Theory and Information Security (ICITIS), 2010 IEEE International Conference on*, pages 807–809, 2010.

[ZSB2012] Sascha Zmudzinski, Martin Steinebach, and Moazzam Butt. Watermark embedding using audio fingerprinting. In: *Transactions on Data Hiding and Multimedia Security VIII*, volume 7882 of *Lecture Notes in Computer Science*. Springer, Aug 2012.

[ZSKR2010] Sascha Zmudzinski, Martin Steinebach, Stefan Katzenbeisser, and Ulrich Rührmair. Audio watermarking forensics: detecting malicious re-embedding. In: *IS&T SPIE Electronic Imaging 2010 Conference - Media Forensics and Security XII, San Jose, January 2010*, 2010.

[ZSN2005] Sascha Zmudzinski, Martin Steinebach, and Sergey Neichtadt. Vertrauenswürdigkeit von Audiodaten – Digitale Wasserzeichen und Verifikation der semantischen Integrität. In: *Sicherheit 2005: Sicherheit – Schutz und Zuverlässigkeit, Beiträge der 2. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)*. Köllen Verlag, 2005.

[ZSN2006] Sascha Zmudzinski, Martin Steinebach, and Sergey Neichtadt. Robust audio-hash synchronized audio watermarking. In: Eduardo Fernández-Medina and Mariemma I. Yagäe, editors, *4th International Workshop on Security in Information Systems (WOSIS 2006), Paphos, Cyprus*, 2006.

[ZST2004] Bin B. Zhu, Mitchell D. Swanson, and Ahmed H. Tewfik. When Seeing Isn't Believing (Multimedia Authentication Technologies). *IEEE Signal Processing Magazine*, 21:40–49, March 2004.

[ZTS2012] Sascha Zmudzinski, Ankit Taneja, and Martin Steinebach. Carving and reorganizing fragmented mp3 files using syntactic and spectral information. In: *AES 46th Conference on Audio Forensics 2012, 14-16 June 2012, Denver, CO, USA*. Audio Engineering Society, June 2012.

[Zwi1961] E. Zwicker. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961.

[ZXH2015] Y. Zhang, Z. Xu, and B. Huang. Channel capacity analysis of the generalized spread spectrum watermarking in audio signals. *IEEE Signal Processing Letters*, 22(5):519–523, May 2015.

Curriculum Vitae

Sascha Zmudzinski

e-mail: sascha.zmudzinski@posteo.de

Education

2007 - 2017	PhD studies in Information Security, *Technische Universität Darmstadt*, Department of Computer Science Final degree: *"Doktor-Ingenieur (Dr.-Ing.)"*.
1994 - 2001	Undergrade and postgrade studies in Physics, *Goethe-Universität Frankfurt* Final degree: *"Diplom-Physiker"* (comparable with *Master's* degree)

Professional Experience in Academia

2016 - 2017	*Fraunhofer Secure Information Technology Institute* SIT, Darmstadt Project Manager on Data Leakage Prevention
2007 - 2014	*Fraunhofer Secure Information Technology Institute* SIT, Darmstadt – Deputy Head of Department "Media Security and IT Forensics" – Project Coordinator of "Information Rights Management" for *Center for Advanced Security Research Darmstadt (CASED)* – Project Manager on Audio Data Security
2002 - 2006	*Fraunhofer Integr. Publication and Information Systems Institute IPSI*, Darmstadt Project Manager in Audio Data Security
2002	*Goethe-Universität Frankfurt*, Department of Applied Physics Research Associate in Computer Vision and Robotics

Award

"Best Paper Award" granted by *The Digital Watermarking Alliance* and the societies *IS&T* and *SPIE* for a publication on audio pre-processing for improving audio watermarking detection; joint work together with *Martin Steinebach* and *Stefan Nürnberger* [SZN2011].

Appendix A

Theoretical Background

This Chapter introduces the technical background of the proposed approach for perception based audio data authentication. For this, basic principles of digital audio data representations and human auditory perception are explained.

A.1 Representations and Transformations of Digital Audio

This Section describes digital representations of audio data for further processing that are relevant for this PhD thesis.

A.1.1 The Physical Nature of Sound

The nature of *acoustic sound* and its perception by a human listener has been under research since ancient times: Already the Greek philosopher and scientist *Aristotle* (381-322 B.C.) described sound as a physical phenomenon of compression and decompression of the air that can be observed when it is

> *"...contracted and expanded...and again struck by the impulses of the breath and the strings",*

see *Whewell et al.* [Whe1859, p. 24]. Physically, acoustic sound is the phenomenon of the variation of the *pressure* $p = p(\vec{x}, t)$ in the air or in any other suitable gaseous, liquid or solid propagation media caused by a *sound source* and propagating as a longitudinal wave through the media. Such sources can be, for example, musical instruments (*"strings"*) or human voice (*"breath"*).

At a remote and *fixed* location $\vec{x}_0$ in the propagation medium, the amplitude of the sound (pressure) wave is a purely time-dependent quantity $p(t) = p(\vec{x}, t)|_{\vec{x}=\vec{x}_0}$. The magnitude of the pressure variations is usually many orders of magnitude smaller than the constant hydrostatic/barometric pressure. This oscillation can be measured using suitable technical sensor devices (e.g. a microphone) and can be processed further in a computer, as in this thesis work. Alternatively, it can be perceived by the ear of a human listener and the consecutive cognitive processing in his auditory system. The following two Subsections describe how the recorded sound is digitized and then represented in the temporal domain as PCM samples or in spectral domain as Fourier coefficients for further processing as in this thesis work.

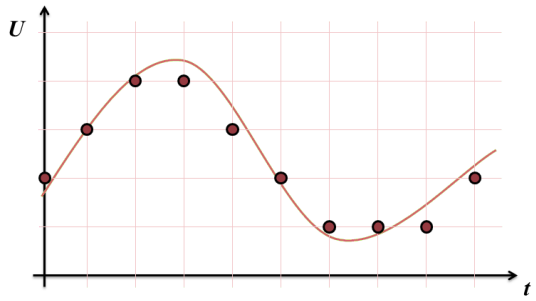


Figure A.1.: Sampling and quantization – continuous line: analog input signal U; dots: sampling points/nodes; grid: given resolution of time-axis and amplitude-axis for input signal

A.1.2 Temporal Domain Representation

The data domain of this work is digital audio data. Thus, in order to process or analyze audio by a computer, it needs to be available and accessed in digitized form. Unless the audio data is created synthetically and digitally by a computer in the first place, it needs to be captured with suitable devices. Then it must be converted from *analog to digital* and finally saved in a suitable standard format.

An important standard for audio signals is the *Pulse-Code-Modulation (PCM)* as explained in the following [ITU1988]. Other common file formats based on (lossy) spectral representations, (e.g. *MP3, AAC, AC-3*) or linear prediction modeling (as in *GSM* or *G.7xx* voice codecs) are not considered in this Section.

The analog input signal can be the continuous, time-dependent electrical voltage amplitude $U(t)$ for example captured with a microphone or electrically picked-up from a musical instrument. Then, this signal is expressed as a time-series which is discrete in time and amplitude by means of *sampling* and consecutive *quantization*.

A.1.2.1 Signal Sampling and (Anti-) Aliasing

The continuous signal is analyzed at regular points in time or instants (see vertical grid lines in Figure A.1). Those instantaneous values are considered specific, representative measurements, or *samples*, of the original signal. All other continuous states of the input signal *between* the sampling instants are neglected. However, the *Sampling Theorem* by *Nyquist/Shannon* [Sha1948, Sha1949] states that this approximation is sufficiently precise under the condition that the *sampling frequency* f_S is at least twice as high as the highest frequency component present in the signal. Oscillations at frequencies higher than the so-called *Nyquist frequency* $f_N = 1/2 \cdot f_S$ will be "overlooked" in the course of sampling and cannot be observed (correctly).

Even more: if a signal component at a frequency $f_N + f_2$ greater than the Nyquist frequency was present, it can be shown that it would be interpreted by the sampler as the presence of a low frequency signal at frequency (f_2 mod f_N) instead. This is often denoted as *aliasing*. This terminology has its origin in computer science, in which the phenomenon of one a particular piece of data being accessible or "appearing" several times is commonly called aliasing. For avoiding such "phantom" sounds as aliasing artifacts, the input signal usually is subject to an appropriate *anti-aliasing* low-pass filter that removes any frequency components above f_N first.

Typical values of the sampling frequency (often denoted as *sampling rate*) are

- 44.1 and 48 kHz for consumer media like MP3, Audio CDs, DVD */BluRay* sound tracks,
- 8 kHz or 16 kHz for telephony,
- 48 kHz and multiples thereof up (to 192 kHz) in professional music production or in Digital Cinema.

in practical applications.

A.1.2.2 Signal Quantization

As explained, the sampling step finally provides a series of real-valued, i.e. numerical samples (x_t). Usually, an *integer* representation with a certain bit length B, i.e.

$$x_t \in [-2^B; +2^B - 1] \subset \mathbb{Z} \quad , \quad t = 1, 2, 3, ...$$

or a normalized *floating point* representation, i.e.

$$x_t \in [-1.0; +1.0] \subset \mathbb{R} \quad , \quad t = 1, 2, 3, ...$$

is used for representing the original signal in the time domain. Most common is a *linear* quantization of the analog signal.

As for every numerical value processed by a computer, a feasible data representation needs to be chosen. Typical examples in PCM audio processing are

- 16 bit *integer* representation on Audio CDs for consumers,
- 24-32 bit *integer* or 32 bit *float* normalized to the range $[-1.0; 1.0]$ in Digital Cinema, professional music production systems or in multimedia processing SDKs like in *AudioUnits* under *Apple iOS*,
- 8 or 14 bit integer representation for voice data in telephony.

For integer representation of music-like sound data, the quantization steps are typically *linear*. For example, on Audio CDs, 16 bit signed integer linear-quantization is used, i.e. *quantized* sample values are in the range $-32,767$ to $+32,768$. This provides a range and resolution of the volume dynamics that offers sufficient sound quality to most of the users. For speech transmission, *logarithmic* quantization steps (e.g. A-law, μ-law characteristics) are more suitable, especially at low bit resolutions in telephony [ITU1988].

The representation as integer or float values is only an approximation of the original real-valued input sample. Due to rounding errors, the quantized digital value differ from its original/actual analog value by up to the half of the quantization step size (see Figure A.1). Given by the *Quantization Theorem* by *Widrow* [Wid1961], rounding errors can be modeled as a uniformly distributed additive noise, the so-called *quantization noise*. If the quantization steps are sufficiently small, reconstructing the original signal by interpolation from those quantized nodes is possible.

To achieve a reasonable sound quality, appropriate sample settings of rate and quantization must be chosen dependent on the application and the audio content characteristics (voice, music etc.). In this thesis work, typical settings for sample rate/resolution are 44.1 kHz/16 bit like on Audio CDs which offers sufficient sound quality to most users.

A.1.3 Spectral Audio Representation in the Fourier Domain

In the course of this thesis work also *spectral* representations of audio data are utilized to a large extent.

The readers of this thesis probably are a little familiar with the concept of ”spectrum” and ”frequencies” of audio already. For example, it is common knowledge that most musical instruments provide a certain range of *tones* they can generate. The colloquial wording for describing such tones as ”low” or ”high” reflect that most humans have a natural understanding that they can be *linearly ordered* according to the their physical *frequency*. Quantitatively, the audio spectrum can be measured by computationally analyzing the digitally sampled audio signal. For this, suitable mathematical transforms have been well known in theory and in practice for long and are utilized in this thesis work as well.

With regards to perceptual hashing applications, the majority of commercial audio hashing/fingerprinting systems perform the analysis on the amplitude spectrum of suitable spectral transforms (not necessarily DFT) or on quantities derived from it. The same is true for many audio watermarking algorithms in commercial services.

A very common spectral transform is the *Fourier transform*, named in honor of *Jean Baptiste Joseph Fourier* (1768-1830). According to *Heidemann* [HJB1985] its theoretical foundations were developed independently by *Fourier* and *Johann Carl Friedrich Gauß* (1777-1855) already in the first years of the 19th century but only published much later [Fou1822, Gau1866]. Nowadays, the Fourier transform is a vastly used ”mathematical tool” in numerous theoretical subjects in natural sciences as well as an analysis technique in signal processing.

A.1.3.1 Definition of the Discrete Fourier transform

In this thesis' context especially the *discrete Fourier transform (DFT)* for time-discrete digitally sampled audio data will be utilized. The following simplified explanation of the DFT is based on the compact description in the mathematical handbook by *Stöcker* [Stö1995] or in the *textbook* by Smith [S$^+$1997].

At first, the PCM audio signals x_t are expected to be partitioned and analyzed in short sections of length L, i.e. $t = 1...L$. Such audio section is often denoted as a *frame* in audio processing. The DFT itself is described as a decomposition of any real-valued frame data $x_t \in \mathbb{R}$ into a sum or superposition of sinusoids. Using the exponential notation of trigonometric sine/cosine functions the decomposition is defined as

$$x_t = \sum_{k=0}^{L-1} X_k \cdot e^{i2\pi k/L \cdot t} \quad . \tag{A.1}$$

In most cases the *Fourier coefficient* X_k is a *complex* quantity. It is calculated as follows:

$$X_k = 1/L \sum_{t=0}^{L-1} x_t \cdot e^{-i2\pi k/L \cdot t} \quad \in \mathbb{C} \quad . \tag{A.2}$$

Note about the wording:

- The transition $\mathscr{F} : (x_t) \longrightarrow (X_k)$ as in Equation (A.2) is commonly denoted as the *(forward) Fourier transform* or *Fourier analysis*.

- The transition $\mathcal{F}^{-1} : (X_k) \longrightarrow (x_t)$ in Equation (A.1) is commonly denoted as the *inverse Fourier transform* or *Fourier synthesis*.

Polar representation – Magnitude and Phase

In practice, the *polar* representation of the complex-valued Fourier coefficients

$$X_k = \mathsf{Re}[X_k] + i \cdot \mathsf{Im}[X_k] = |X_k| \cdot e^{-i\phi_k} \in \mathbb{C} \quad .$$

is more descriptive than the *Cartesian* representation: Here, the Fourier transform is described by *magnitude* $r_k := |X_k|$ and *phase* ϕ_k which can easily be obtained from basic properties of complex numbers:

$$r_k := |X_k| = \sqrt{X_k^2} = \sqrt{\mathsf{Re}[X_k]^2 + \mathsf{Im}[X_k]^2} \in \mathbb{R} \quad ,$$
$$\phi_k = \tan^{-1}\left(\frac{\mathsf{Re}[X_k]}{\mathsf{Im}[X_k]}\right) \in [-\pi, +\pi) \subset \mathbb{R} \quad .$$

The reader should note:

- The magnitude r_k at index k describes how much oscillatory behavior the signal shows at frequency $f_k = k/L$. This can also be seen from Equation (A.2) in which X_k formally is the cross-correlation coefficient between the (x_t) and the complex oscillations $e^{-i2\pi k/L \cdot t}$. This quantity by itself is dimensionless.
- The phase ϕ_k describes the initial angular offset of the oscillation at frequency f_k; loosely spoken the phase describes if the particular f_k oscillation behaves more ”sine-like” or ”cosine-like”.
- The Equation (A.1) also implies that the transform is not meaningfully defined for an *instant* of sound: no spectrum can be calculated from a a single PCM sample, i.e. when $L = 1$. For a graphic explanation the reader should recall that measuring frequency components implies measuring *oscillations*. The nature of any oscillation is characterized by a periodic behavior of a quantity *over time* – not during an instant. At least the duration of one period length T of the oscillation must be observable somehow, i.e. it must be $L \geq T > 1$ which is also in line with the Sampling Theorem.
- Closer analysis of the fundamental properties of the DFT shows [Stö1995]: because the input frame data is real-valued it follows that the magnitude spectrum shows the following *symmetry*:

  $$x_t \in \mathbb{R} \quad \Rightarrow \quad r_k = r_{L-k+1} \quad k = 1...L \quad .$$

  That is, only the first $L/2$ coefficients with $k = 1, ..., L/2$ are significant: The upper half of coefficients for $k = (L/2+1), ..., L$ is redundant and is discarded in this thesis work.
- A different wording for the set of r_k is the *amplitude spectrum* or loosely and imprecisely: *”the spectrum”*. The latter reflects also that the magnitudes are often much more relevant so that that the phase spectrum is negligible in many applications.

A.1.3.2 Frequency versus Temporal Resolution Trade-off

It should be recalled that the definition in Equation (A.1) is given for audio signals that were partitioned in frames previously. The calculated amplitude and phase spectrum are valid and

defined only for the duration L of the particular frame. But for regular music, voice etc. the sound, and so its spectrum, will obviously vary over time and the audio will have to be partitioned in many consecutive and independent frames. For obtaining a meaningful transform result for each of the frames, its temporal duration must be short enough so that the signal is sufficiently *stationary*: its sound characteristics shall not change (too much) during the elapsed playing time $\Delta t = L/f_S$ $[s]$.

On the other hand, the precision of the spectral representation, the *frequency resolution* Δf, must be sufficient as well for avoiding to loose spectral information. By comparing with the Nyquist theorem it can easily be shown that every DFT coefficients describes a frequency interval of

$$\Delta f = f_N/(L/2) = 2f_N/L = f_S/L = 1/(\Delta t) \quad [\text{in } Hz]$$

so that the basic *uncertainty relation* of the Fourier transform is

$$\Delta f \cdot \Delta t = 1 \quad .$$

In practice a suitable trade-off between frequency resolution and temporal resolution has to be identified.

Example: In this thesis work, a typical frame length with regards to sufficient stationarity is L=2048 samples which corresponds to Δt=46.4 msec (at sample rate of 44.1 kHz). Hence, the frequency resolution is Δf=21.5 Hz.

Example: In this thesis work, the framelength $N = 2048$ PCM samples is commonly used and the audio content typically has $f_S = 44.1$ kHz sample rate. That means,

- the temporal resolution is $\Delta t = N/f_S = 2048/44100$ $[s]$ which represents 46 milliseconds,
- the frequency resolution is $\Delta f = f_S/N = 22.050/1024$ $[Hz]$ which is approximately 21.5 Hz.

That means that for example the first magnitude coefficient r_1 represents not only the DC component at 0.0 Hz but the averaged presence of *all* frequency components from 0.0 to 21.5 *Hz*. Note that for distinguishing sounds at low frequencies, this is a rather low resolution. For example, the frequency range of the lowest octave on a standard 88 keys piano keyboard (from tone $A0$ to $A1$) is from 27.5 to 55 Hz. The correspondent Fourier coefficients represent in that range cannot distinguish the presence of single half tones.

A.1.3.3 Spectral leakage

What has been concealed in the *simplified* description so far was that the Fourier analysis as in Equation (A.2) in fact not only approximates the framed signal (x_t) in the domain $t \in \{1, 2, 3, ..., L\}$. Closer theoretical analysis shows that the *infinite periodic extension* by the sample values as if it was

$$\tilde{x}_t = x_{(t \bmod L)}$$

in the domain $t \in \{-\infty, ..., +\infty\}$ is actually approximated. This means: the synthesis relation as in Equation (A.1) describes the signal also beyond the frame boundaries as if it was *periodically repeated* infinite times. This can cause inaccuracies in the DFT results which is demonstrated by the following empirical...

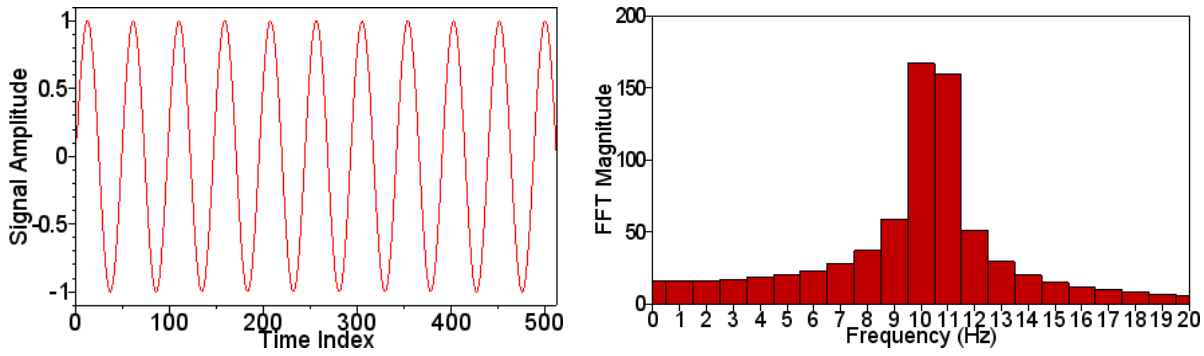


Figure A.2.: Left Figure: Example audio frame with sine tone at 10.5 Hz;
Right Figure: correspondent (magnitude) spectrum – a peak is visible at 10 and 11 Hz. Significant magnitudes are visible at other frequencies due to *spectral leakage*

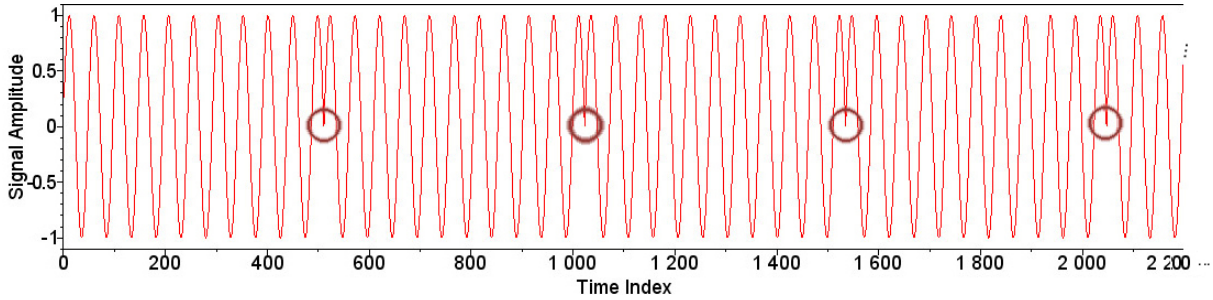


Figure A.3.: Example: Periodic extension of 10.5 Hz signal. Discontinuities can be observed as sharp dips (see circle markers) every 512 samples.

...Example: In Figure A.2 a pure sinusoid signal at frequency 10.5 Hz is subject to a 512-point DFT. Instead of a peak/spike the output spectrum indicates seeming frequency components in the whole domain of 0-20 Hz (and beyond, actually) i.e. at frequencies other than 10 to 11 Hz. A portion of the signal energy figuratively ”leaks” to side-lobes at frequencies adjacent to the actual main peak. The notable behavior in the example can be explained by the fact that the periods of the 10.5 Hz oscillation do not ”fit” in the 512-sample DFT frames. Thus, the virtual periodic extension (see Figure A.3), which the Fourier transform *actually* approximates, shows significant notches every 512 samples. The audio playback of such discontinuous curve shape obviously sounds different from the pure sine tone: impulses noise or clicks become audible. Those distortions introduce additional frequency components in the spectrum.

This effect is called *spectral leakage*. Leakage occurs for every audio signal that contains frequency components that do not ”fit” in the analysis frame window. For real-world audio content (music, voice etc.) such leakage can be expected as well. Hence, it will become relevant in the watermarking and perceptual hashing investigated in this thesis work too.

An abstract and formal explanation of the leakage is the following: In the previous example, the notion of a *”pure sinusoid signal at frequency 10.5 Hz ”* strictly speaking is very idealized: it

would only correct if the signal duration was infinite. In fact, the framing of the signal can be expressed as sample-wise multiplication of the input signal with the function

$$w_r^{(\text{rect})} = \begin{cases} 1.0\ , \text{ if } t \in \{1...L\} \\ 0.0\ , \text{ else} \end{cases}$$

which is has a rectangular shape in the time domain, the so-called *rectangular window* (function). From the fundamental convolution theorem of the DFT it is known that the spectrum of such sample-wise product can be calculated as the *convolution* of the DFT of the signal with the DFT of the window function. Hence, the spectrum of the sample-wise product obviously will differ from the spectrum of the "pure" sinusoid.

A.1.3.4 Windowing

The common approach in signal processing for reducing the leakage effect is applying a multiplication of the input signal with a suitable *window function* other than the rectangular window. The shape of the window function is defined so that it approaches *zero* towards the frame boundaries without discontinuities. Because the windowed frame data vanishes at the frame boundaries, its periodic extension becomes seamless again and there appear no discontinuities at the frame boundaries.

As an example, the cosine-based *von Hann* window (commonly denoted as *Hanning* window) was introduced by *Blackmann* and *Tukey* [BT1959] in honor of the meteorologist *Julius von Hann* (1839-1921) with

$$w_t^{(\text{Hanning})} = \begin{cases} 1/2\left(1+\cos(2\pi/L \cdot t)\right)\ , \text{ if } t = 1...L \\ 0.0\ , \text{ else} \end{cases} \quad ,$$

Other common window types are the *Blackman* window, the *Hamming* window (not to be confused with the *Hanning/von Hann* window), which is named in honor of the mathematician and computer scientist *Richard Hamming* (1915-1998):

$$w_t = 0.54 + 0.46\cos(2\pi t/L) \quad ,$$

or or the *Bartlett* window, see [S$^+$1997, Chap. 9 and 16].

Example: In the example in Figure A.4 a *Hanning window* was applied on the pure sine-like audio signal at 10.5 Hz from the earlier example. Closer analysis and comparison of the plots Figures A.2 and A.4 (right Figures, resp.) shows that the leakage effect in is reduced by two orders of magnitude in the side-lobes due to windowing.

Although the shape of the windowed signal is indeed different from the original, closer analysis of the theory of the DFT shows that the position of the peaks in the magnitude spectrum is not influenced by the windowing: leakage effects and "phantom" frequency components are reduced significantly. Closer analysis shows that leakage effects are not completely removed, though. Leakage is an immediate consequence of partitioning the audio input into frames – no matter if/which a windowing is applied.

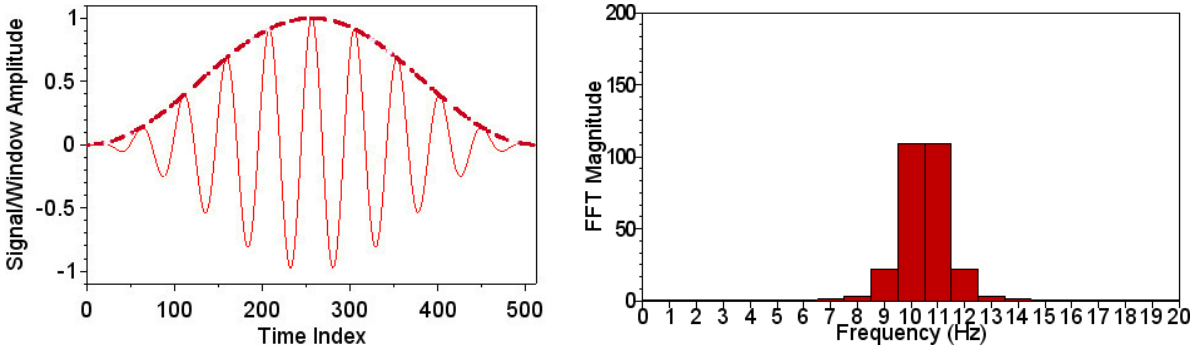


Figure A.4.: Left Figure: Example audio frame data (solid line: 10.5 Hz sine tone) that was windowed (dashed line: *Hanning* window);
Right Figure: correspondent Fourier magnitude spectrum; observation: spectral leakage is reduced

A.1.3.5 Fast Fourier Transform (FFT)

In practice, a numerical optimization of the DFT algorithms is used very often, namely the *Fast Fourier Transform (FFT)*. According to *Heidemann et al.* [HJB1985] not only the roots of the general DFT but also of the FFT optimization go back to the work by *Gauß* (not *Fourier*) from the 19th century [Gau1866]. It was rediscovered and adapted to modern computing by *Cooley* and *Tukey* in the 1960s [CT1965].

The FFT mainly utilizes an algorithmic *divide-and-conquer* design for the DFT for reducing computational efforts. For this, in most FFT implementations the frame length L has to be a power of two, i.e. $L = 2^k, k \in \mathbb{N}$ or the product of small prime numbers at least. The algorithmic detail go beyond the scope of this overview. The interested reader can find a comprehensible explanation of the overall FFT algorithm in the mathematical handbook by *Stöcker* [Stö1995].

In this thesis work, a number of numerically optimized FFT implementations are used for example

- the *"Fastest Fourier Transform in the West (FFTW)"* as introduced by *Frigo* and *Johnson* [FJ2005] from *Massachusetts Institute of Technology, USA*[79],
- the *"Math Kernel Library (MKL)"* by the *Intel Corp.* which is a math library optimized for *Intel* compatible CPUs[80].

In practice the FFT allows a significant reduction in algorithmic complexity: it can be shown that calculating the DFT according to Equation (A.2) has a complexity $O(L) \propto L^2$. In contrast, the FFT has a lower algorithmic complexity of only

$$O_{\mathrm{FFT}}(L) \propto L \log_2(L) \quad .$$

Example: In this thesis work a frame size of $L = 2048$ will commonly be used. Using the FFT-optimization instead of the plain DFT "textbook definition" reduces computational needs to $2048 \cdot \log_2(2048)/2048^2 \approx 1/186$ i.e. by more than 99%!

A.1.4 Other Spectral Audio Representations

[79] *FFTW Project: http://www.fftw.org*
[80] Intel *Math Kernel Library (MKL): https://software.intel.com/en-us/intel-mkl*

A.1.4.1 The Cepstrum

The *cepstrum* (as defined by *Bogert, Healey* and *Tukey* [BHT1963]) is a common "mathematical tool" in signal processing. Using the notation as given in Section A.1.3, the cepstrum C_k related to a Fourier audio magnitude spectrum $X_k = \mathscr{F}\{(x_t)\}$ is defined as the inverse Fourier transform of the logarithm of the Fourier magnitude spectrum:

$$(C_l) = \mathscr{F}^{-1}\Big\{\log(|(X_k)|^2)\Big\} \quad .$$

For this representation, the original authors creatively defined the somewhat unique term of the *"cepstrum"* as a paraphrase of the word "'spectrum'. Furthermore, the independent index l of the cesptral coefficients C_l is denoted as the *quefrency* (instead of *frequency*). This shall reflect that the relevant data domain is neither the frequency domain nor truly the temporal domain.

It can be shown that by its analytical properties the cepstrum is well suited for many signal processing tasks like identifying pitch/fundamental frequencies and detecting/compensating for echoes in signals. The latter property is relevant for watermarking approaches based on *echo hiding* as explained in Section 2.3.4 or MFCC-based audio hashing from the state of the art as recalled in Section 2.2.4.

A detailed description of the cepstrum can be found in the comprehensible work on the *"History of the Cepstrum"* by *Oppenheim* and *Schafer* [OS2004].

A.1.4.2 The SVD Domain

The SVD is a decomposition/factorization of an arbitrary matrix M (the *measurement matrix*) according to

$$M = USV^T \quad .$$

Here, U and V are orthogonal matrices containing the so-called *singular vectors* as columns. The matrix $S = \mathsf{diag}(s_1^2, s_2^2, ...)$ is a diagonal matrix containing the so-called *singular values* s_i which are the sorted non-zero eigenvalues of the matrix $M^T M$. The SVD allows for identifying/removal of dependencies and noise in the measurement input. It is mathematically similar to the techniques of *principal component analysis (PCA)* or the *Karhunen-Loeve* transform. Independent signal components can be identified by the presence of singular values that are significantly greater than those of noise components. Such components can be suitable for carrying an audio watermark embedding or serving as robust features in audio hashing.

In numerous fields of natural science and engineering the SVD formalism is used as a "mathematical tool" for for example solving systems of linear equations, least-squares estimation or in approximating a matrix by another matrix of lower rank etc. It must be admitted that the SVD factorization and the spectrum of singular values result are a highly *abstracted* representation of M. Nevertheless, in watermarking the spectrum of the singular values easily allows identifying independent – and hence *robust* – signal components which are suitable for embedding.

The detailed explanation of the SVD and its properties would go beyond the scope of this brief overview. It can be studied in the comprehensible textbooks by *Mendel* or by *Golub/van Loan* [Men1995, GVL1996]. The mathematical similarity between SVD, PCA and KL transform is elaborated on by *Gerbrands* [Ger1981].

A.1.4.3 The DCT, Wavelet and MFCC domain

The *Mel Frequency Cepstrum Transform* is another common transform in signal processing. Here, the spectrum is expressed in terms of the Mel frequency cepstral coefficients (MFCCs) derived from the cepstrum representation (as explained before in Section A.1.4.1). In extension, if features a sub-sampling of the linearly spaced quefrency scale to the *Mel* scale (as explained in Section A.2.6). The MFCC representation reflects human perception in terms of critical band characteristics. It is a useful ”mathematical tool” in music analysis, speech/speaker recognition and in watermarking and audio hashing as well.

The mathematical detail of MFCCs go beyond this overview and the explanation on the cepstrum in this thesis. The interested reader can find a very comprehensible explanation of MFCCs and their utilization by *Logan* [Log2000].

Other frequently utilized transforms, which are not explained in detail here, are the *Discrete Cosine Transform* (DCT) (introduced in the 1970s by *Ahmed* [ANR1974]) and the *Wavelet* transform (introduced as early as 1909 by Haar [Haa1910]) which are distantly related to the DFT. The interested reader can find a very comprehensible explanations of these transforms and their utilization as ”mathematical tools” in multimedia processing in the works by *Smith* [S$^+$1997] on DCT or in [Cas1979, Dau1992, Mal1999] on Wavelets.

A.2 Psychophysics – Modeling of Human Sound Perception

This Section explains the foundations of *psychophysics*, i.e. the modeling of the relationship between the objective physical properties of sound as a stimulus and its correspondent subjective auditory perception. Most important findings and quantities that are relevant for understanding this thesis work are explained in the following.

According to *Helmholtz* [Hel1863, p. 27] research on acoustical sound has been conducted already in ancient times for example by *Aristotle, Pythagoras* and in early modern times for example by *Galilei, Newton, Euler* or *Bernoulli*. At those times, research was rather focused on the creation and propagation of sound, the theory of vibrating strings in musical instruments and other fundamental *objective* physical quantities like *frequency, sonic velocity, pitch* etc.

From the 19th century, the *subjective* aspect of human perception of sound became increasingly under investigation. Since then the overall concept of *psychophysics* has been increasingly investigated in physiology and psychology. For example, *Helmholtz* explained that sounds can be grouped into either *"musical"* or *"noises"* (see [Hel1863, p. 14, 19]). It is interesting to see that this view is *still* persistent to some extent in modern *MP3/AAC* audio encoding techniques in which so-called *"tonal"* and *"non-tonal components"* are identified during encoding.

A.2.1 Anatomy of the Human Ear

Many characteristics of human auditory perception can be understood from the anatomy of the inner ear (see *Lennarz* [LB2012]): Sound waves pass the outer *ear canal* and the *ear drum*, then propagate through the middle ear bones (*auditory ossicles*) to the spiral *cochlea* in the inner ear. Inside the cochlea the sound is perceived by the sound receptors, the so-called *hair cells* on the *organ of Corti*. The *outer* hair cells can react on incoming sound by mechanical contraction and hence *actively* amplify the sound. The actual *conscious* perception of sound by the brain is eventually stimulated triggered by the sound receptors of the *inner* hair cells. It was found that hair cells are present in *bundles* of many dozens of hair cells. Their tips are mechanically tethered with their neighboring bundles by so-called *tip links*.

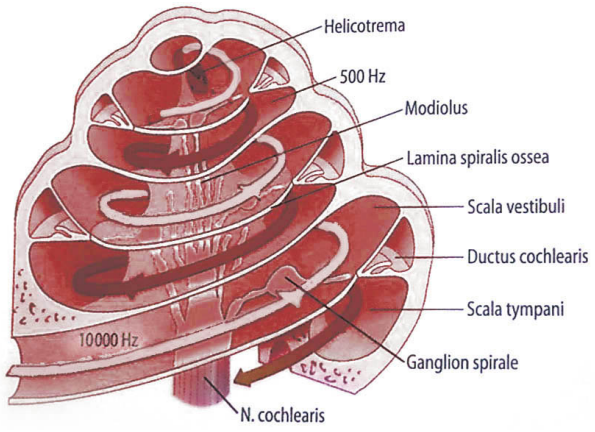


Figure A.5.: Anatomy of the human *cochlea*: sound waves entering from the lower left. Regions being sensitive for frequencies at "500 *Hz*" and "10,000 *Hz*" are tagged accordingly; Source: [LB2012]

As can be seen from Figure A.5, the sensation of certain frequencies correspond to certain spatial areas in the organ of *Corti*: high frequencies are perceived at the beginning of the outer

section of the organ. In return, low frequencies are perceived in the inner section. The selective frequency sensitivity of the hair cells and their coupling via tethered tip links causes some remarkable properties of the human sound sensation as explained in the following.

A.2.2 The *Weber-Fechner* Law

In the 19th century *Weber* discovered that the *just noticeable difference* of two stimuli like sound is proportional to the magnitude of the stimuli themselves [Web1834]. From this result *Fechner* derived [Fec1860, pp. 64, pp. 134] that the magnitude of subjective perception s of a stimulus is proportional to the *logarithm* of its objective intensity I so that

$$s \propto \log_{10}(I) \quad ,$$

known as the term *Weber-Fechner Law*. According to the authors, it applies to acoustical perception like loudness or pitch but also to sensation of brightness of light, temperature, heaviness of weights etc.

It should be mentioned that the logarithmic model was later questioned and generalized to other sensations by *Stevens* by showing that *"a* power *law, not a log function, describes the operating characteristic of a sensory system"* [Ste1961] so that

$$s' \propto I^n \quad ,$$

with $n \approx 0.54$ to 0.60 for audio loudness. However, both *Fechner*'s log law and *Stevens'* power law describe the loudness characteristics qualitatively alike, namely as a strictly monotonic curve that is downward increasing. The following explanations stick to the *Weber-Fechner* definition for simplicity.

The interested reader can find an *exhaustive* analysis of *Fechner's* versus *Stevens'* results in the text book on *"The Measurement of Sensation"* by *Laming* [Lam1997] from 1997.

More modern research on psychophysics from the 20th century made important discoveries with respect to perception of loudness and frequencies. The most important eventually led into the development of efficient lossy audio coding and are being utilized for this thesis work.

A.2.3 The *dB* and *Phon* scale for sound pressure levels

An objective measure for the perceived loudness in quantitative figures instead of qualitative terms as "loud", "soft", "fortissimo" or the like was defined by *Heinrich Barkhausen* (1881-1956). For this, at first the objective intensity of sounds with average sound pressure $\bar{p}$ is expressed (in line with the *Weber-Fechner* law) as *sound pressure levels (SPL)* on the logarithmic *Decibel (dB)* scale:

$$L := 10 \cdot \log_{10}\left(\frac{\bar{p}^2}{p_0^2}\right) \quad [\text{in } dB] \quad .$$

At the time of its first definition the value of the so-called *reference pressure* p_0 was regarded as the minimum sound pressure required so that a sound is just noticeable to an average listener (at 1 kHz). Note that these *Decibel* values are actually *dimensionless* as they are the logarithm of the dimensionless ratio $\bar{p}^2/p_0^2$. The usage of the (pseudo) unit "*dB*" shall remind of the logarithmic dependency.

Then, for describing and comparing the subjective *loudness* of different tones or noises with different characteristics, timbre etc., the *Phon* scale was defined by *Barkhausen*:

> Phon *scale: a* loudness level *of "n* Phon*" of an arbitrary sound or noise means that it appears as loud as the reference tone increased by n* dB *to an average listener.*

As a rule of thumb, loudness level differences of approximately ±1 *Phon* are just noticeable by an average listeners.

Helmholtz also describes that sound can be characterized by its *"pitch"*. The pitch concept implies that tones can be characterized by their (base) *frequency*. It was said that the range of frequencies that can be *"perceived at all"*, are in the range of *"16 to 38,000" Hz*. oscillations". Note that the author used the terminology *"oscillations"* because the physical unit *Hertz (Hz)* was officially introduced not until 1935 in honor of the physicist *Heinrich Hertz* (1857-1894). Actually, more modern research from the 20th century as collected by *Zwicker and Fastl* [ZF1990] found the upper frequency limit rather to be approximately at 20,000 *Hz*, at best.

A.2.4 Absolute Hearing Threshold in Silence and Equal Loudness Perception

In the 1930s, *Fletcher* and *Munson* provided empirical results about how sensitive the human ear is to different frequencies [FM1933]. For example, the human ear is most sensitive at approximately 4 kHz and the total range of sensitivity is from 20 Hz to 20 kHz, approximately. In Figure A.6 each so-called *masking curve* corresponds to equal loudness for a given loudness level. The lowest curve in the Figure was regarded as the *absolute hearing threshold* in silence (corresponding to the reference level p_0).

The correct shape of the equal-loudness-curve, especially for low frequencies smaller than 1 *kHz* has been under research, debate and refinement for decades [RD1956, ST2004]. The results finally became an *ISO* standard for *"Equal-loudness-level Contours"* in 2003 [ISO2003].

The research proved that the human ear has fascinating capabilities as an organ of perception: The area of sensitivity covers many orders of magnitude across frequencies and sound pressure levels (which for example outperforms the human eye in perception of light).

A.2.5 Auditory Masking

Frequency Masking

In extension to the work by *Fletcher* and *Munson* another important psychoacoustic effect is *frequency masking* [ZF1990, Moo1995]. Masking means that the presence of a so-called *masking tone* (commonly denoted as: *masker*) significantly reduces the sensitivity to sounds at other frequencies. The results from lab testing as shown in Figures A.7 and A.8 demonstrate how the shape of the contour curves are lifted or flattened, resp. Here, tone-like maskers and noise markers cause different shapes of the updated contour curve modification and, hence, have to be treated slightly differently in mathematical modeling of masking. As a surprising example, even a white noise masker that would be imperceptible by itself (i.e. being *below* the absolute hearing threshold) *does* reduce the ear's sensitivity to an additional test tone (see dotted line in Figure A.7).

Frequency masking is relevant for complex sounds like music or voice data. Here, the given "mixture" of spectral components interacts with itself. Some components will be reduced in loudness or can even be extinguished completely – although they are physically present.

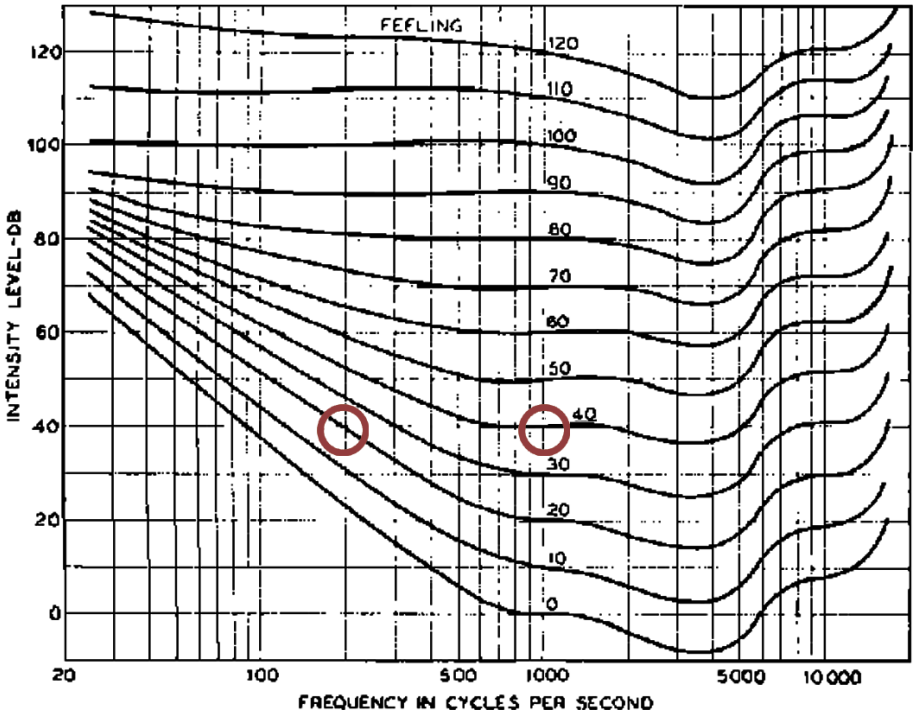


Figure A.6.: Equal loudness level contours (in Phon); threshold of pain denoted as *"feeling"*; Horizontal axis: frequency (in Hz); Vertical axis: input sound pressure level (in *dB*); Example (see circle markers): A test tone with sound pressure level 40 *dB* at 1 *kHz* is perceived at 40 *Phon*. Another test tone at the same intensity but at 200 *Hz* is perceived 20 *Phon* softer;Source: *Fletcher, Munson* [FM1933]

In the course of this PhD thesis, the frequency masking effect is utilized for transparent watermark embedding, more reliable watermarking detection and it is discussed for identifying perceptually relevant components in audio hashing.

Temporal Masking

For completeness note that also *temporal* masking can be observed and modeled: a masking tone can attenuate or even annihilate the perception of consecutive sound events in the respective near future (*"forward masking"*). Especially for very short sound events (like pulses etc.) It can last up to 150 msec if there is a large volume difference between the masker and the consecutive test tone. Surprisingly, maskers can even influence the perception of other sound *ex post*. This *"backward masking"* has a range up to 50 msec. seemingly into the past. Both are thought to be caused by time-dependent process during the neural reception [Lak1976].

Temporal masking can and will be ignored in the course of this thesis work for simplicity: the temporal resolution of the FFT framing as carried out in this thesis work ($\approx$ 46 *msec*) is quite in the same range as backward masking. Also forward masking can be ignored in sufficient approximation for most of the sound content.

A.2.6 Critical Band Characteristic

It was a discovery by *Fletcher* in the 1940s that several masking tones present inside a certain frequency range effectively contribute as *one* single masker, see *Zwicker* [ZF1990, pp. 149].

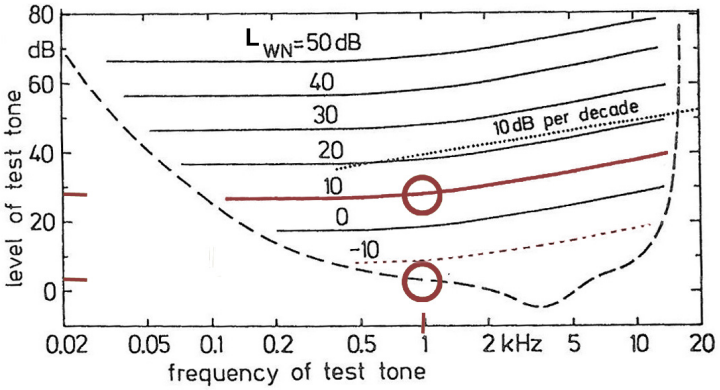


Figure A.7.: Equal loudness level contours in the presence of noise maskers; dashed lines: hearing threshold in silence for comparison
Solid lines: equal loudness level contours in the presence of broad band white noise at levels L_{WN} from -10 to +50 *dB*; Circle markers: example of a white noise masker at L_{WN}=10 *dB* lifts the absolute hearing threshold at 1 *kHz* by 25 *dB*
Source: *Zwicker et al.* [ZF1990, p. 62]

That frequency range is called a *critical band* (CB). Experimental results show that for low frequencies, the critical bands have equal width of 100 *Hz*. From 500 *Hz*, the critical band width increases exponentially, see Table A.1 for detail. In total, approximately 26 critical bands fit in the frequency range of 20 to 20,000 Hz in which the ear is sensitive to sound.

CB	1	2	3	4	5	6	7	8	...	22	23	24
f_1 [in *Hz*]	0	100	200	300	400	510	630	770	...	7700	9500	12000
f_2 [in *Hz*]	100	200	300	400	510	630	770	920	...	9500	12000	15500
Bark	0	1	2	3	4	5	6	7	...	21	22	23

Table A.1.: Critical bands: f_1 and f_2: lower and upper CB bounds in *Hz*; last row: correspondent *Bark* range
Source: rounded values taken from [ZF1990, Zwi1961]

To count the index of the critical band, the non-linear *Bark* scale (named in honor of *Heinrich Barkhausen*) was proposed by *Zwicker* in the 1960s [Zwi1961, ZF1990]. For example, the value of 5 *Bark* corresponds to the cut-off frequency of the *5th* critical band at 510 *Hz*. The mapping from Hz to *Bark*, and vice versa, can be modeled by

$$[Bark] = 12\tan^{-1}\big(0.00076\cdot[Hz]\big) + 3\tan^{-1}\left(([Hz]/7500)^2\right)$$
$$[Hz] = 52548/([Bark]^2 - 52.66\cdot[Bark] + 690.39) \quad .$$

Note that the CB characteristic does not imply that a masker does not have an influence beyond the range of ±0.5 *Bark* around its respective center frequency. Instead, the CB characteristic reflects how a set of maskers within a CB range accumulates and eventually contributes to the overall masking threshold. This is exploited for the psycho acoustic modeling as conducted in this thesis work, and by other authors in the Related Work on integrity protection, as for example by *Gulbis* [GMS2008a, Gul2013].

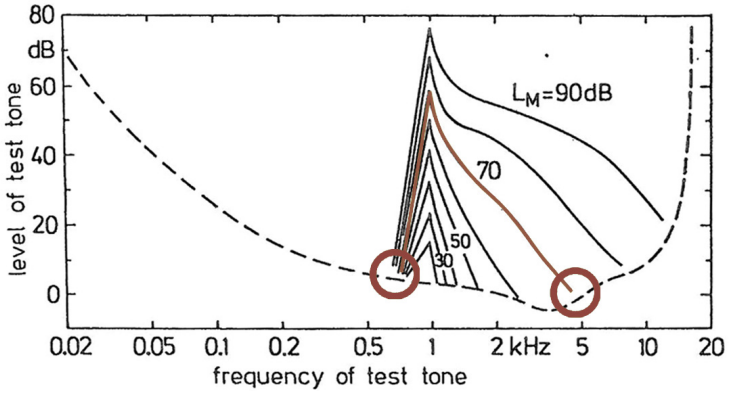


Figure A.8.: Equal loudness level contours in the presence of tone maskers; dashed lines: hearing threshold in silence for comparison
Solid lines: equal loudness level contours in the presence of a tonal masker at 1 *kHz* at different levels L_M from +30 to +90 *dB*; Circle markers: example of a 1 *kHz* masker at 70 *dB* affects the hearing threshold in a broad range from 0.6 to 5 *kHz*; Source: *Zwicker et al.* [ZF1990, p. 68]

The related *Mel* scale is also common in psycho acoustics. It is defined as

$$100\,Mel = 1\,Bark \quad .$$

It is the foundation for the so-called *Mel-frequency cepstral transform (MFCC)* widely used in speech processing and audio fingerprinting algorithms (see A.1.4.3).

Note that both frequency masking and CB characteristics can be explained from the anatomy of the hair cells in the human ear: The nonlinear coupling/tethering by their tip-links can influence excited hair cell bundles in a certain spatial neighborhood in terms of masking. The overall coupling is however spatially limited due to damping which explains ranges of both masking and CBs.

Appendix B

List of Audio Test Files

The following audio content was used in the experimental evaluation. The total duration of the files was trimmed to 3:00 *min*. In most test runs in the experimental evaluation each file was further divided in snippet of 0:30 *min*. duration.

Total duration of the audio test set is 4.5 hours (2.66 *GibiByte*).

B.1 Audio books

The following audio book examples were captured from Audio CDs.

File No.	Title/Content
A 01-07	Hermann Kurze, ”Thomas Mann – Leben und Werk”
A 08-10	Rainer M. Schroeder, ”Das Geheimnis der weissen Moenche”
A 11-19	Marcel Reich-Ranicki, ”Mein Leben”
A 20-23	Barbara Auer liest Tobe Jansson, ”Mumins lange Reise”
A 24-26	Tschechow, ”Erzaehlungen”
A 27-30	Heinz Ludwig Arnold, ”Die Gruppe 47”

Table B.2.: Audio test set (audio books)

B.2 Music songs

The following music examples were captured from Audio CDs.

File No.	Title/Content
M 01	Adrian Bond – Die Calmly Like A German
M 02	Adrian Bond – The Gods of Old
M 03	Conjunto Explosao Do Samba – Barra Pesada
M 04	Exit 100 – Beams A Gonna Plaint
M 05	Black Pete – Saviour
M 06	Black Pete – Vicious
M 07	Cassandra Wilson – Blue in Green
M 08	Neneh Cherry – Buffalo Stance
M 09	Chora Viola Chora – Tempestade Do Amor
M 10	Phillip Glass – Cloudscape

File No.	Title/Content
M 11	David C. Arthur – An Evening in D.C.
M 12	Thomas Newman – Dead Already
M 13	Juergen Knieper – Der Himmel Ueber Berlin
M 14	D-Flame – Es tut mir leid
M 15	Dirk Darmstadter – Best Day
M 16	DJ Fraternity – Alec Empire
M 17	Joao Nogueira – Espere Oh! Nega
M 18	Fleshhouse – The Flamingo People
M 19	Commodores – Gonna Blow Your Mind
M 20	Suicidal Tendencies – Gotta kill Cpt. Stupid
M 21	Grant Lee Phlillips – Under the Milky Way
M 22	Orchestra Claude Debussy - Gymnopedie No .1
M 23	Jungle Brothers - I'll House You
M 24	Indigo Girls – Language to the Kiss
M 25	Jeff Buckley – Lover, You Should've Come Over
M 26	Benito di Paula – Jeito de Felicidade
M 27	Phillip Glass – Koyaanisqatsi
M 28	K's Choice – Dad
M 29	Lloyd Cole – Late Night
M 30	Exit 100 – Louie Frantic

Table B.4.: Audio test set (music songs)

B.3 Voice recordings

Most of the following examples of voice recordings were captured from *Youtube* as secondary source.

File No.	Title/Content
V 01	"Michael D. Bailey police interrogation"
V 02	"Dalia Dippolito questioned by Police"
V 03	"Russel Williams confession"
V 04	"Police officer slaps U.S. soldier"
V 05	"Harvey Wince child abuser"
V 06	"Man in court accused of murder"
V 07	"Collateral Murder" (Wikileaks)
V 08	"President Reagan's address at Brandenburg Gate"
V 09	"How to survive a traffic stop"
V 10	"Klaus Kinksi unmoegliches Interview"
V 11	"Threatened with psychological evaluation"
V 12	"Michael Jackson/Oprah Winfrey, interview
V 13	"Entlastungsmaterial Lothar Koenig"
V 14	"Am Strassenrand, Polizei stoppt Motorradfahrer"
V 15	"Bushido, Stress ohne Grund Interview"
V 16	"Tagesschau 18.10.89, Honecker tritt zurueck"
V 17	"Helmut Schmidt, Rede an die Nation"
V 18	"Gerhard Schroeder, Elefantenrunde"

File No.	Title/Content
V 19	”Helmut Kohl gegen Helmut Schmidt”
V 20	”Helmut Kohl vs. Herbert Wehner”
V 21	”Juergen Trittin Kurzintervention”
V 22	”Martin Luther King Jr., I have a dream!”
V 23	”Sunday Times Interview with President al-Assad
V 24	”Cpt. Schettino sulla telefonata con De Falco”
V 25	”Challenger disaster live footage”
V 26	”Der Nuernberger Prozess (Urteil, 1/9)”
V 27	”Nixon, raw Watergate tape, smoking gun”
V 28	”Marcel Reich Ranicki lehnt Deutschen Fernsehpreis ab”
V 29	”Tagesschau zu Tschernobyl”
V 30	”WM 1990 Achtelfinale Deutschland Holland 2:1”

Table B.6.: Audio test set (misc. voice recordings)

Appendix C

Exhaustive Test Results

This Appendix provides more exhaustive test results about detail about the experimental evaluation.

C.1 *rMAC* Flag Ratio: Boxplots versus Time – Admissible Attacks

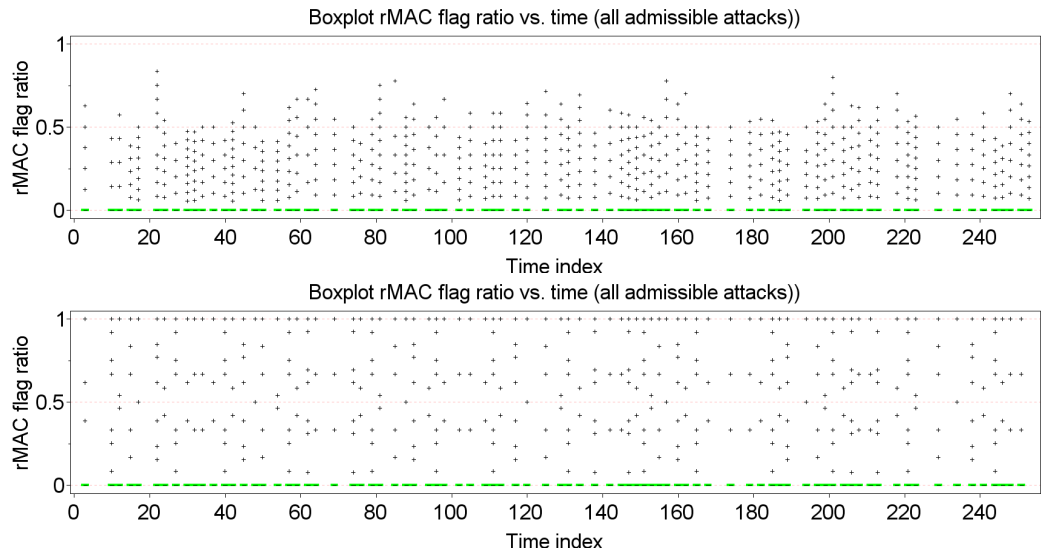


Figure C.1.: *rMAC* flag ratio versus time as boxplots; only admissible attacks applied (obviously all over the file); upper: ”scattered mode”, lower: ”serial mode”

C.2 *rMAC* Flag Ratio: Boxplots versus Time – Malicious Attack – "Scattered Mode"

"Scattered Mode" – Mix noise

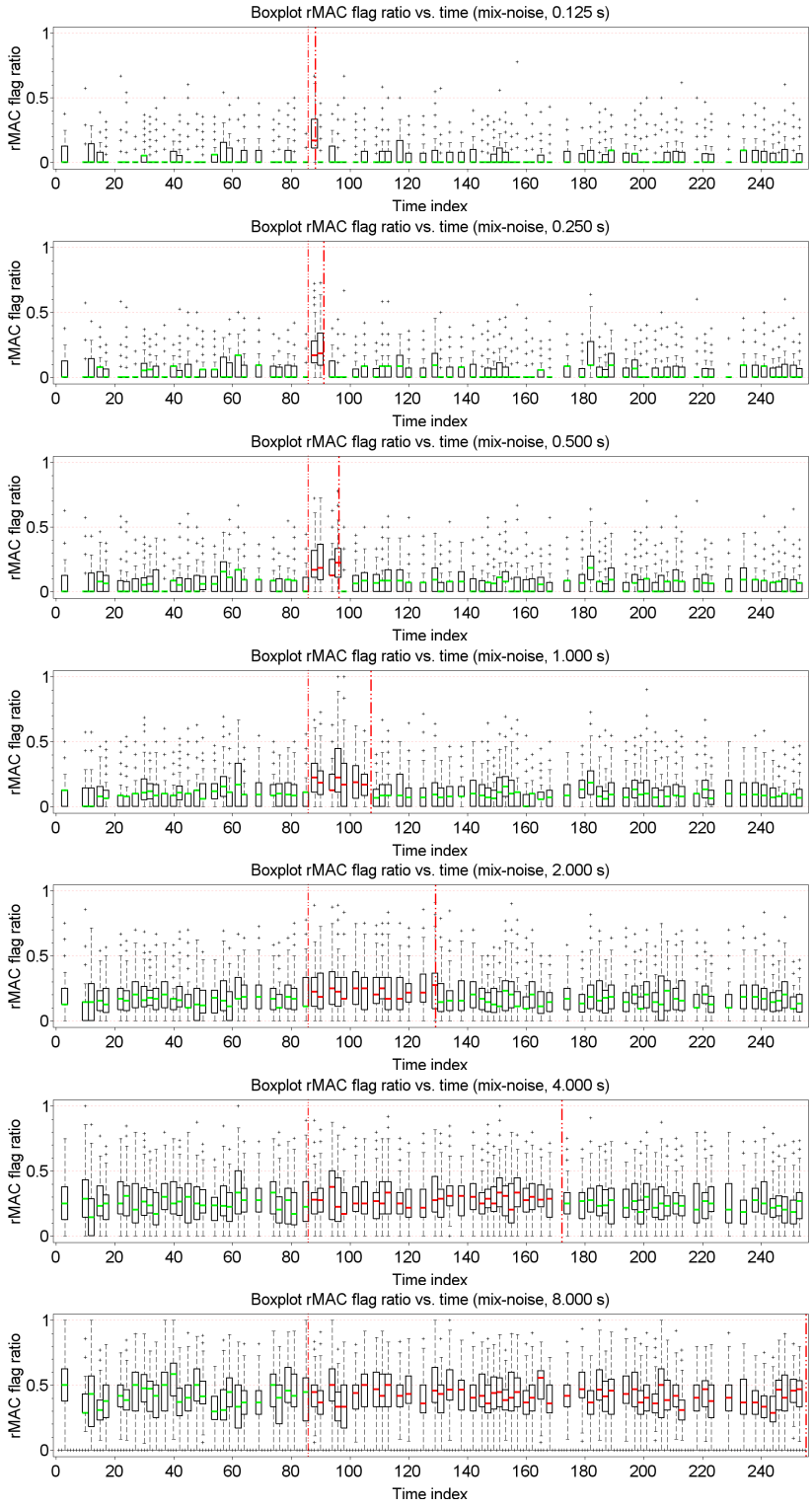


Figure C.2.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Scattered Mode” – Deletion

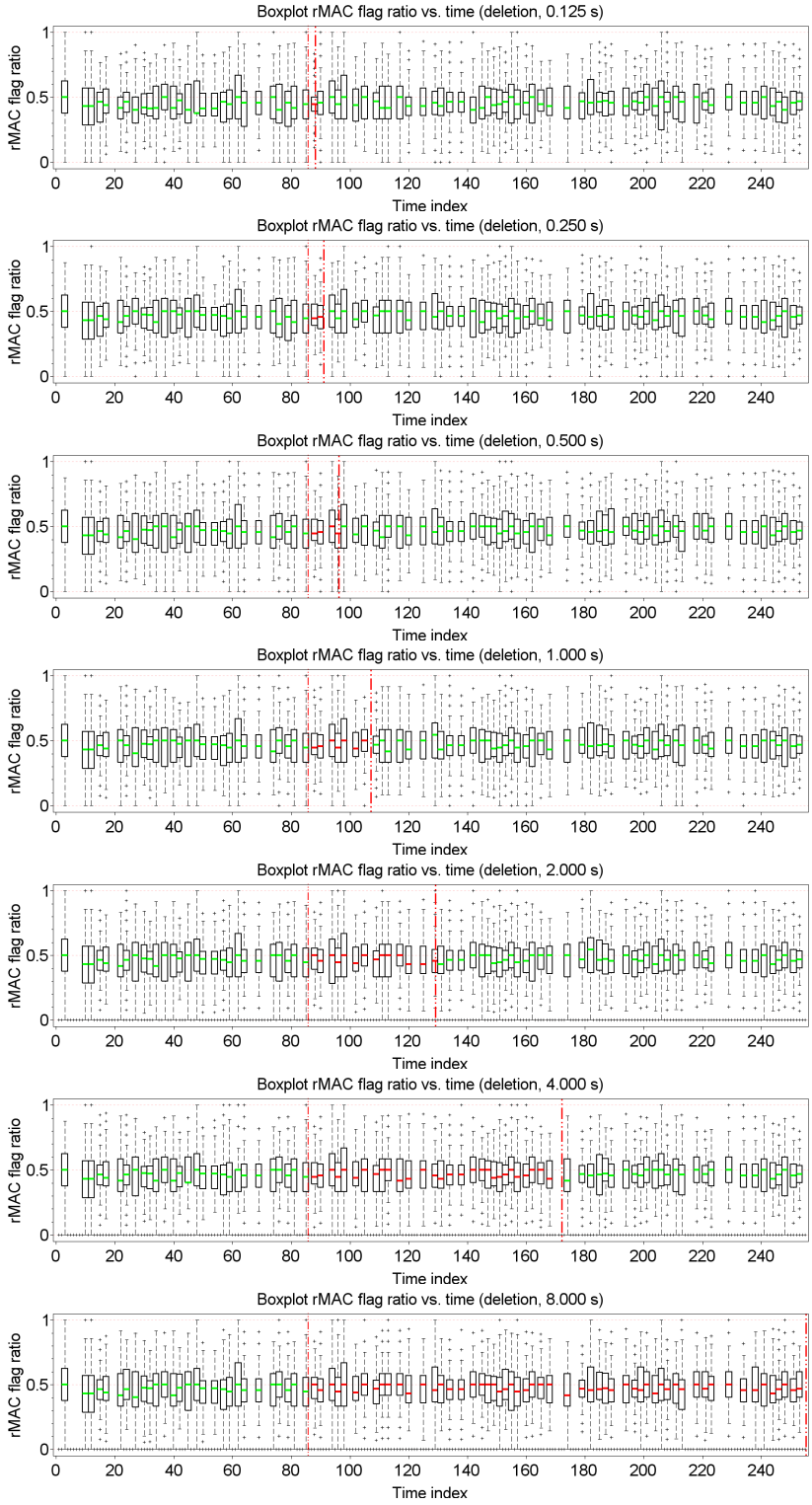


Figure C.3.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Scattered Mode” – Replacement

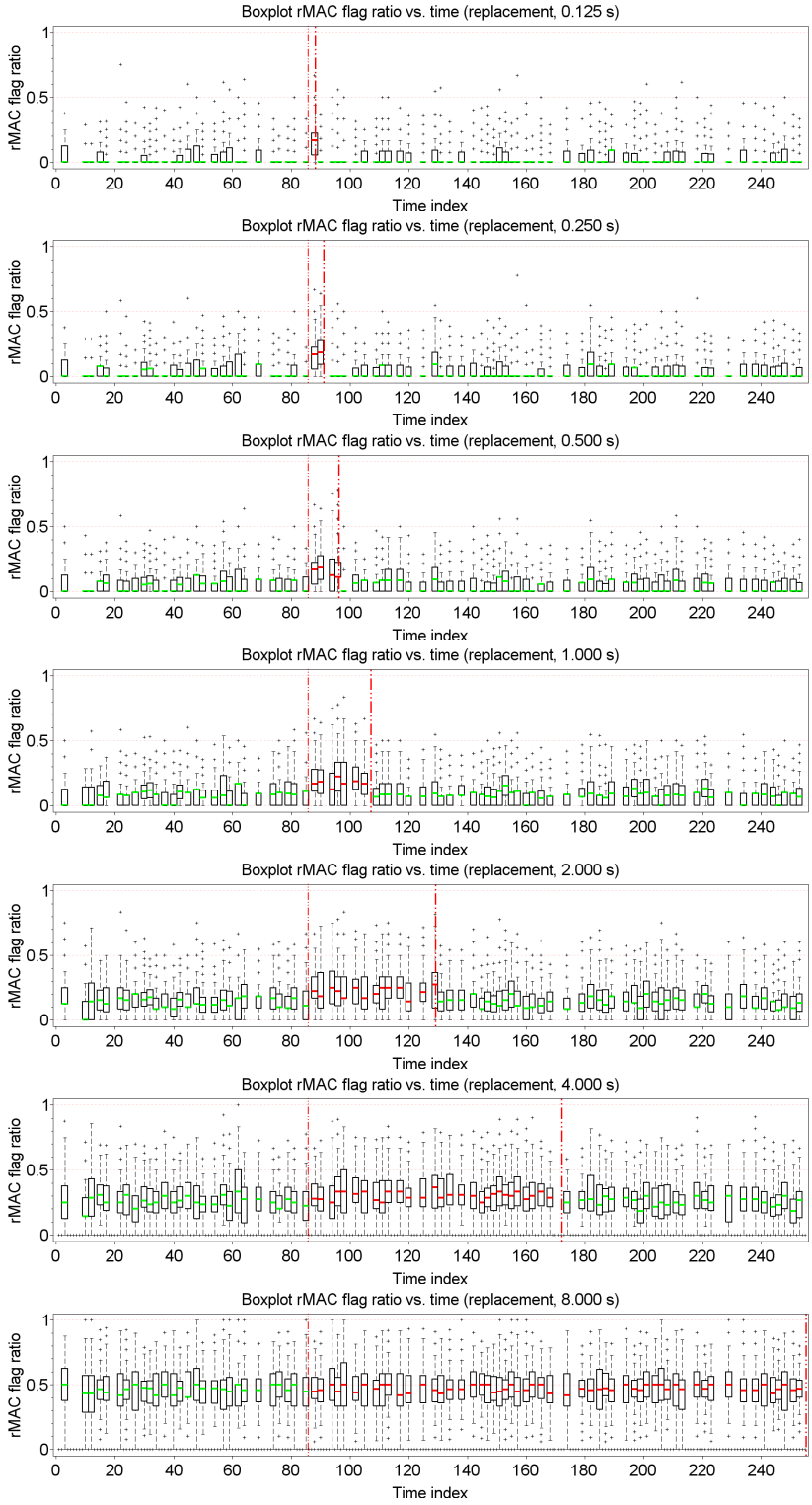


Figure C.4.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Scattered Mode” – Mix audio

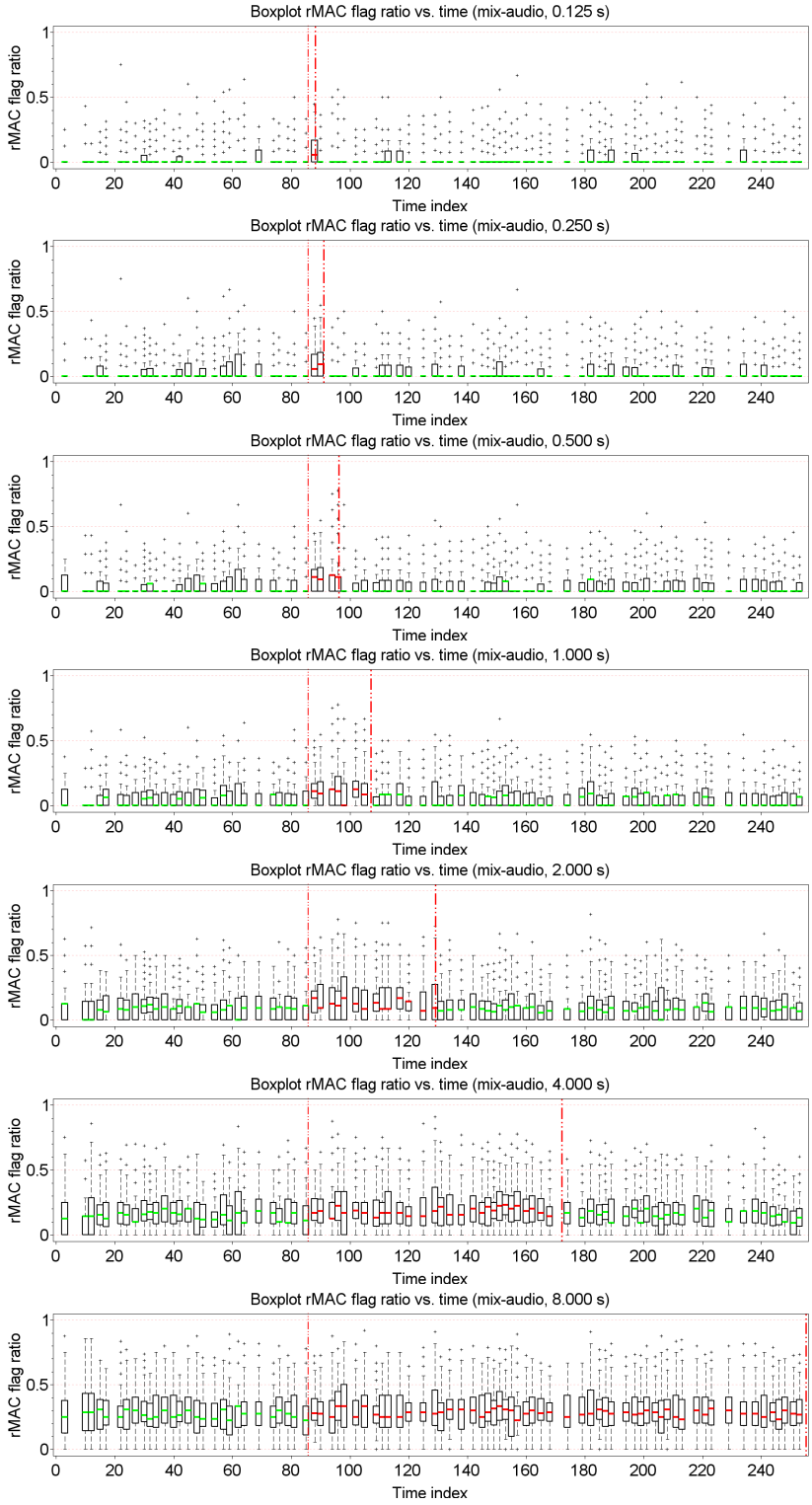


Figure C.5.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

C.3 *rMAC* Flag Ratio: Boxplots versus Time – Malicious Attack – "Serial Mode"

"Serial Mode" – Mix noise

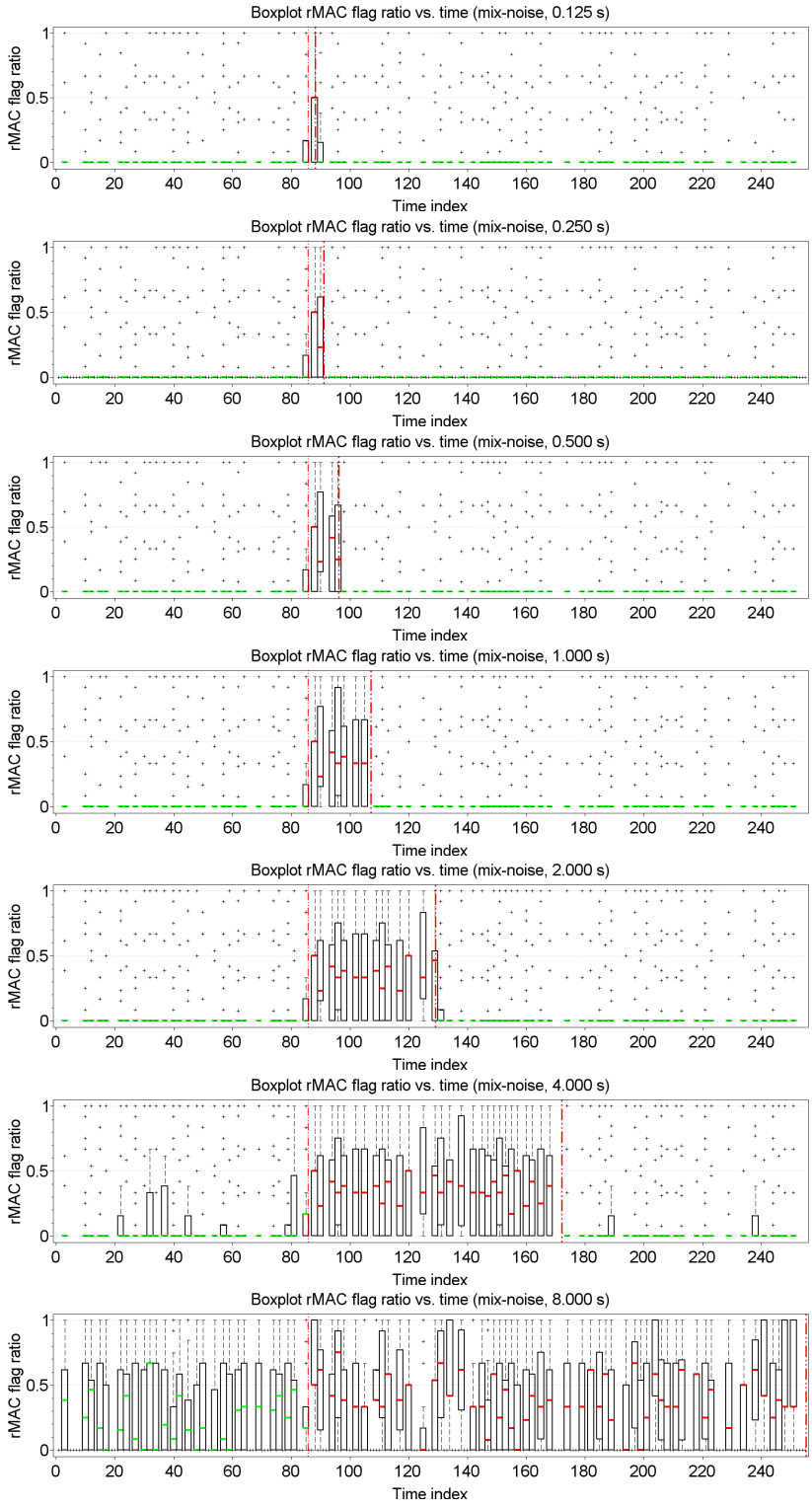


Figure C.6.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Serial Mode” – Deletion

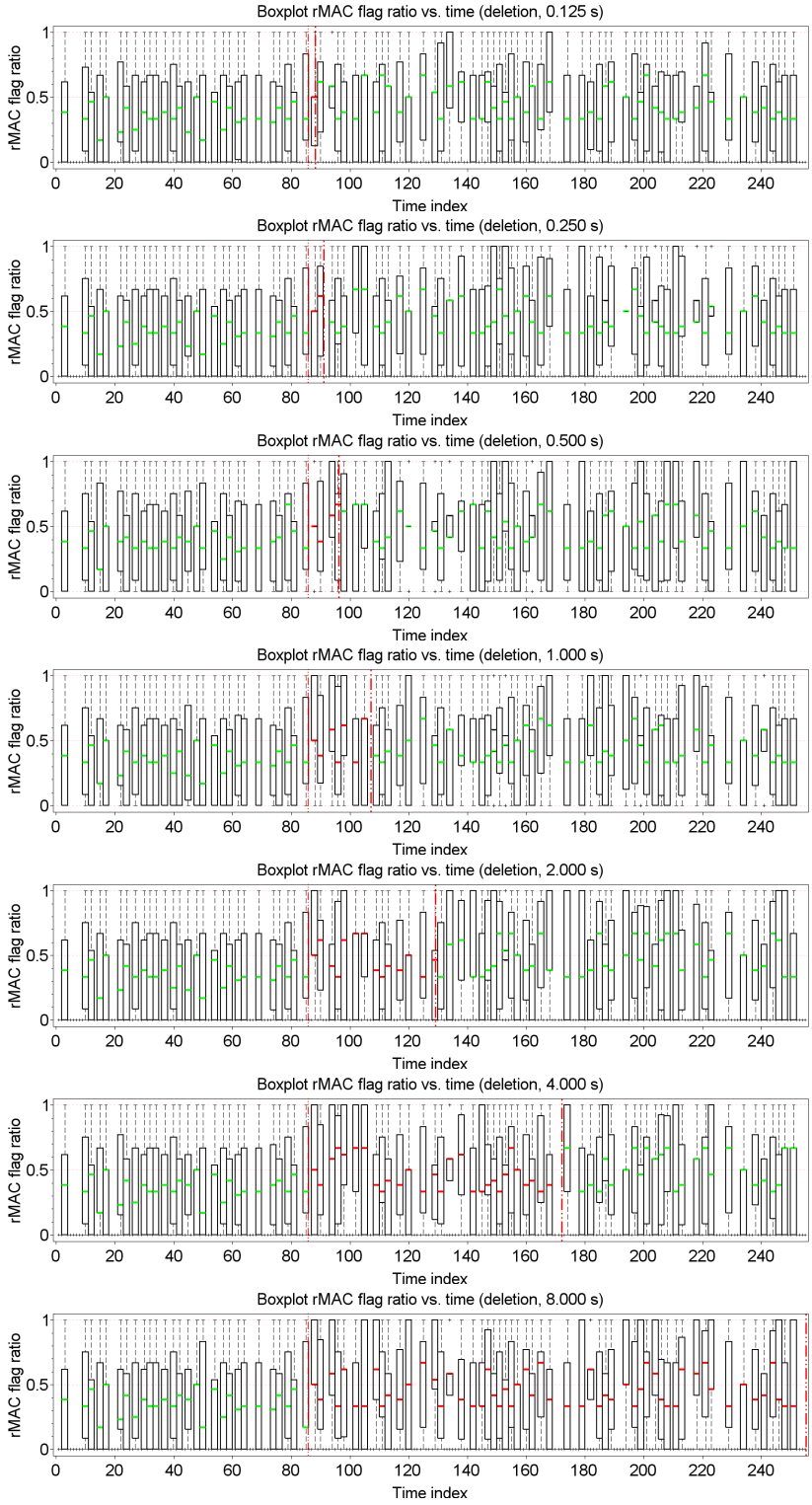


Figure C.7.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Serial Mode” – Replacement

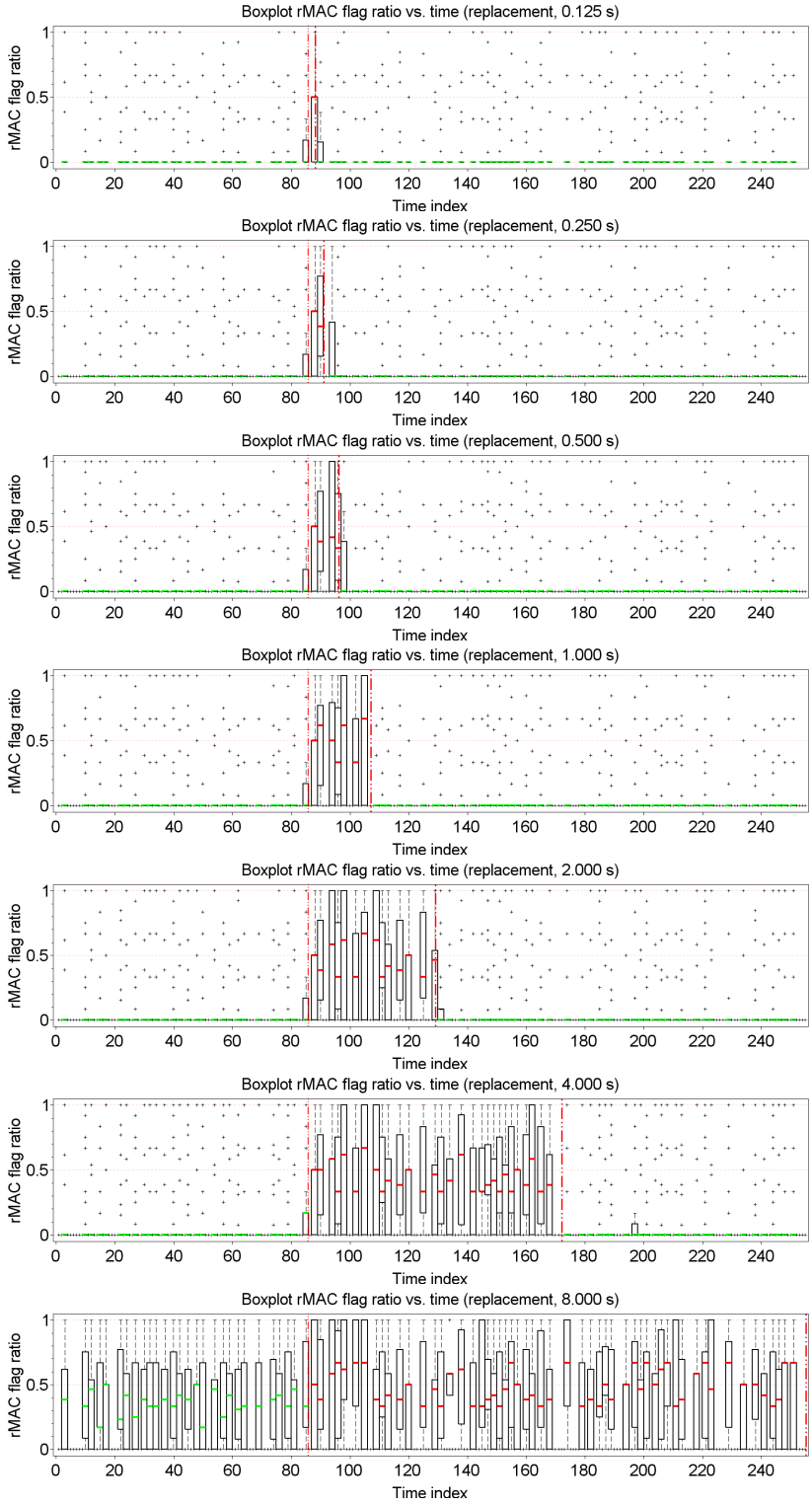


Figure C.8.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

”Serial Mode” – Mix audio

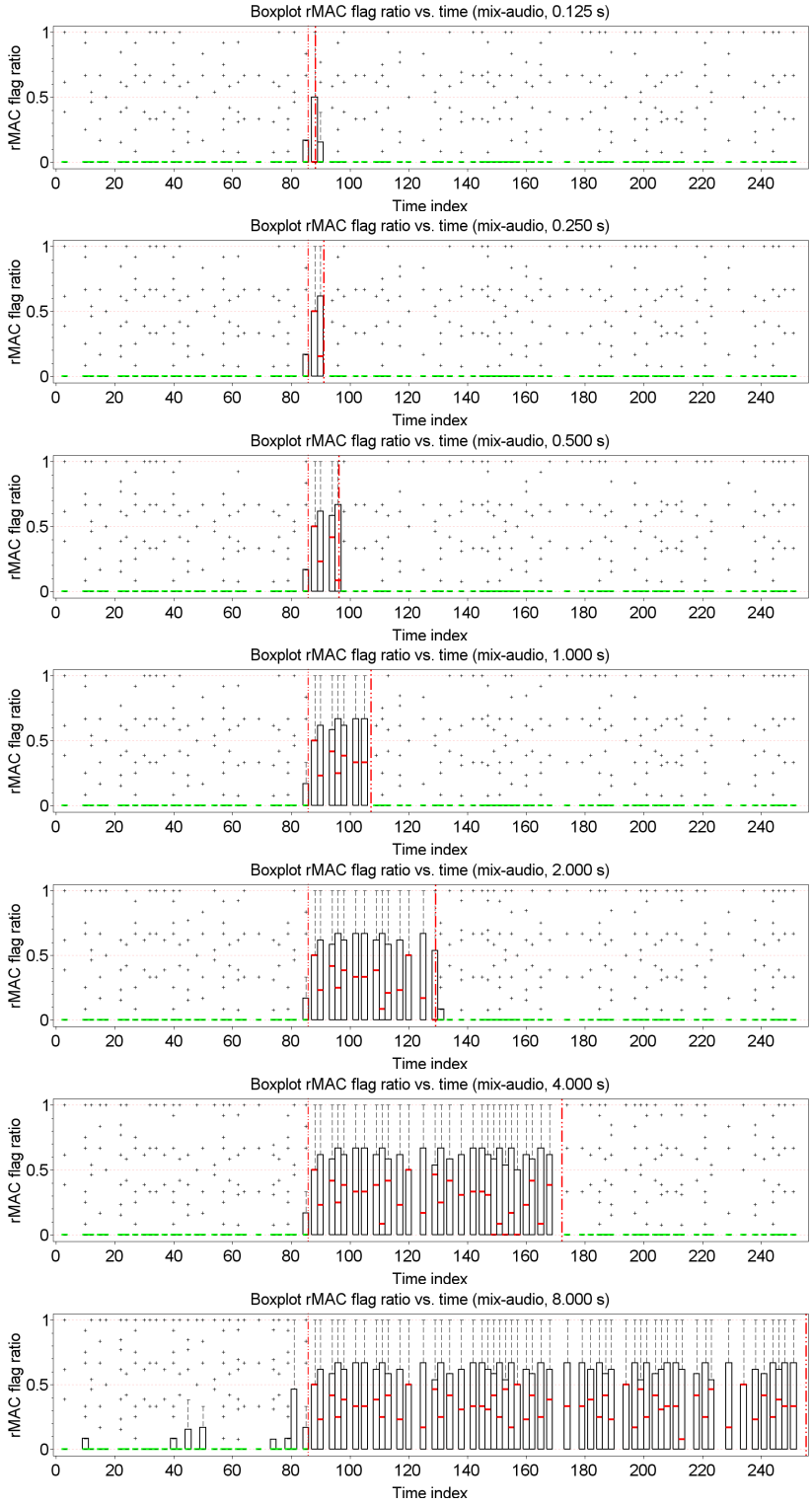
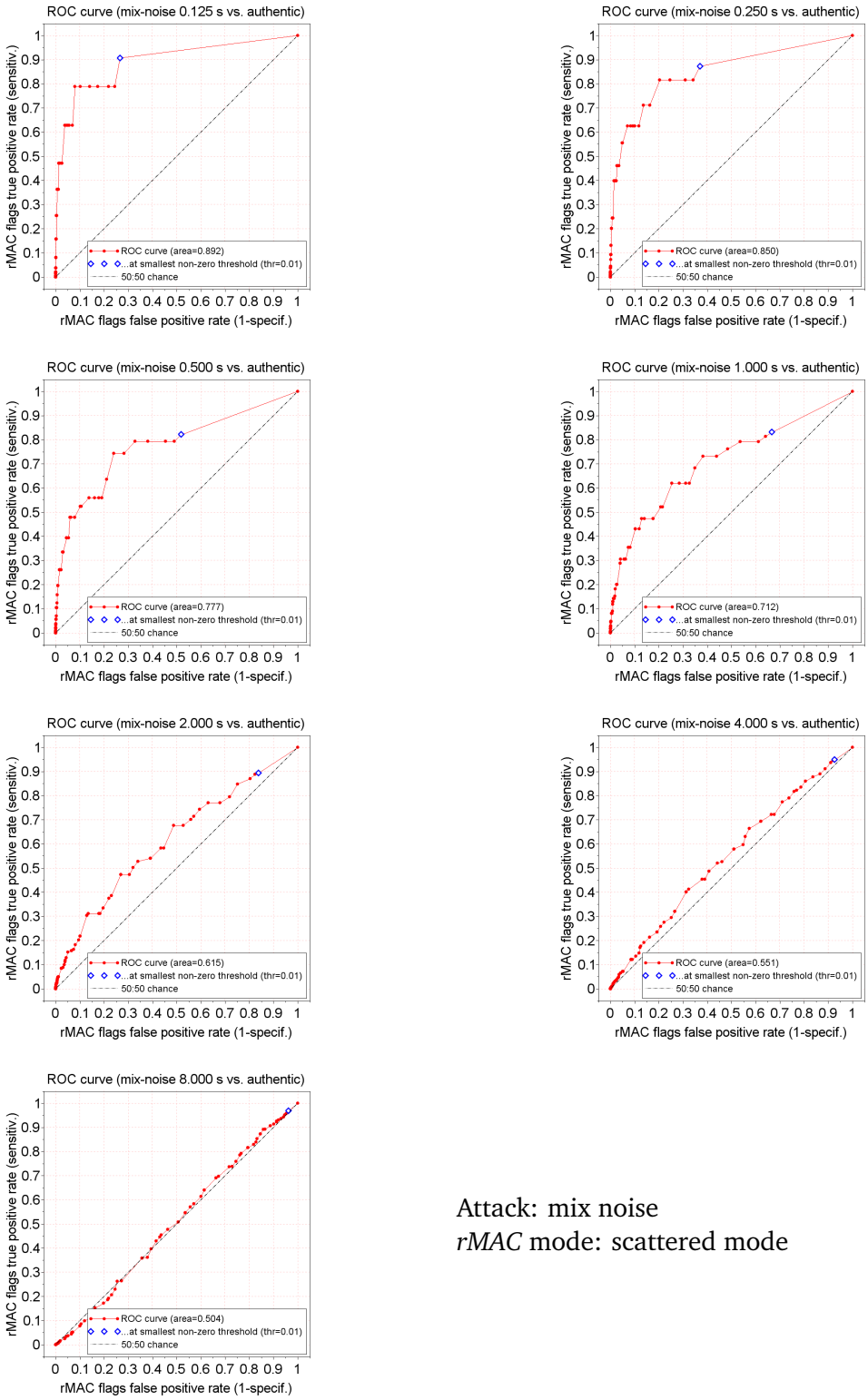


Figure C.9.: *rMAC* flag ratio versus time as boxplots; dashed lines enclose attacked portion in audio snippets (starting at time index 86 = 4.0s from the beginning)

C.4 *rMAC* Flag Ratio: ROC curves – ”Scattered Mode”

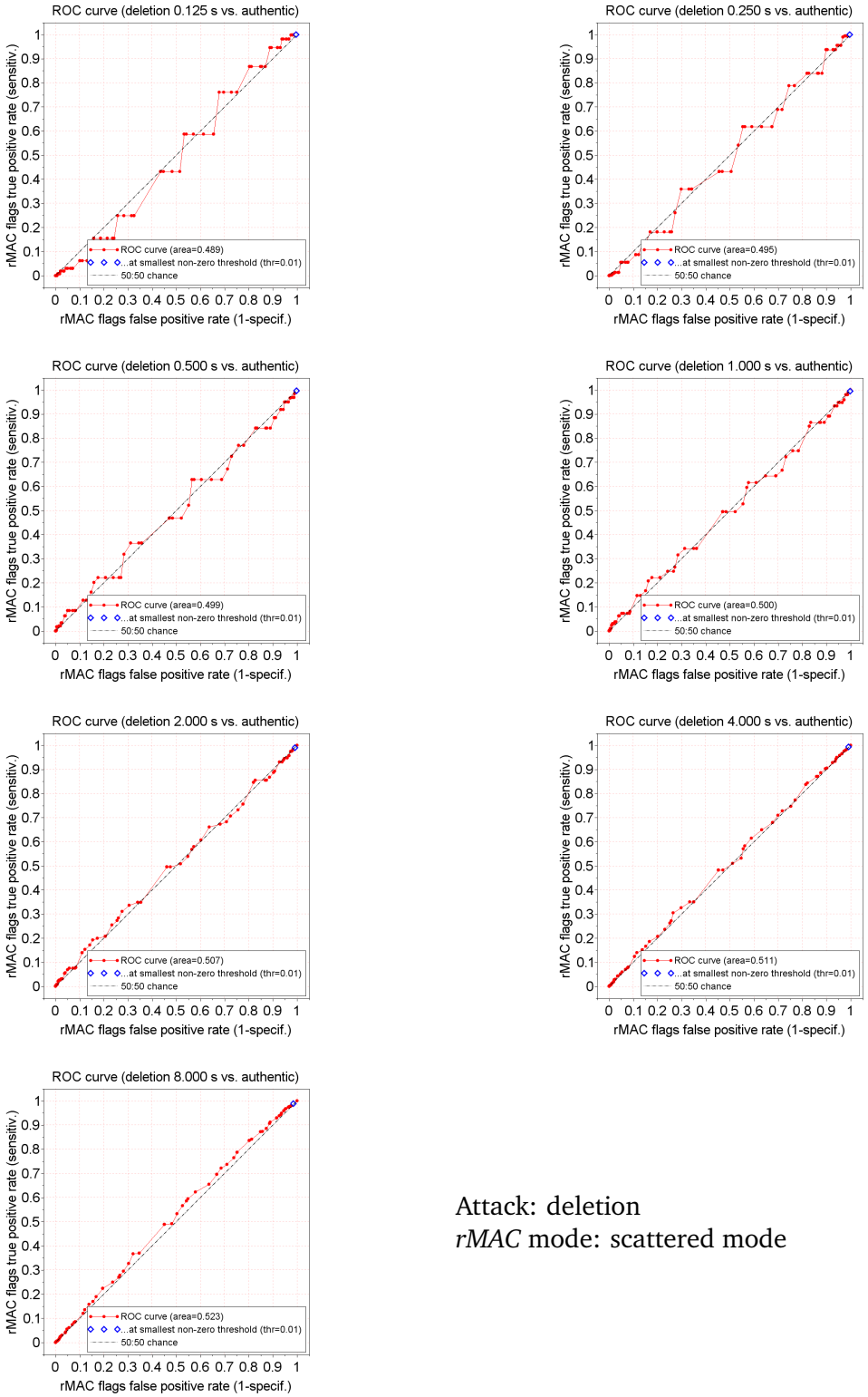
”Scattered Mode” – Mix Noise



Attack: mix noise
rMAC mode: scattered mode

Figure C.10.: *rMAC* flag ratio as ROC plots

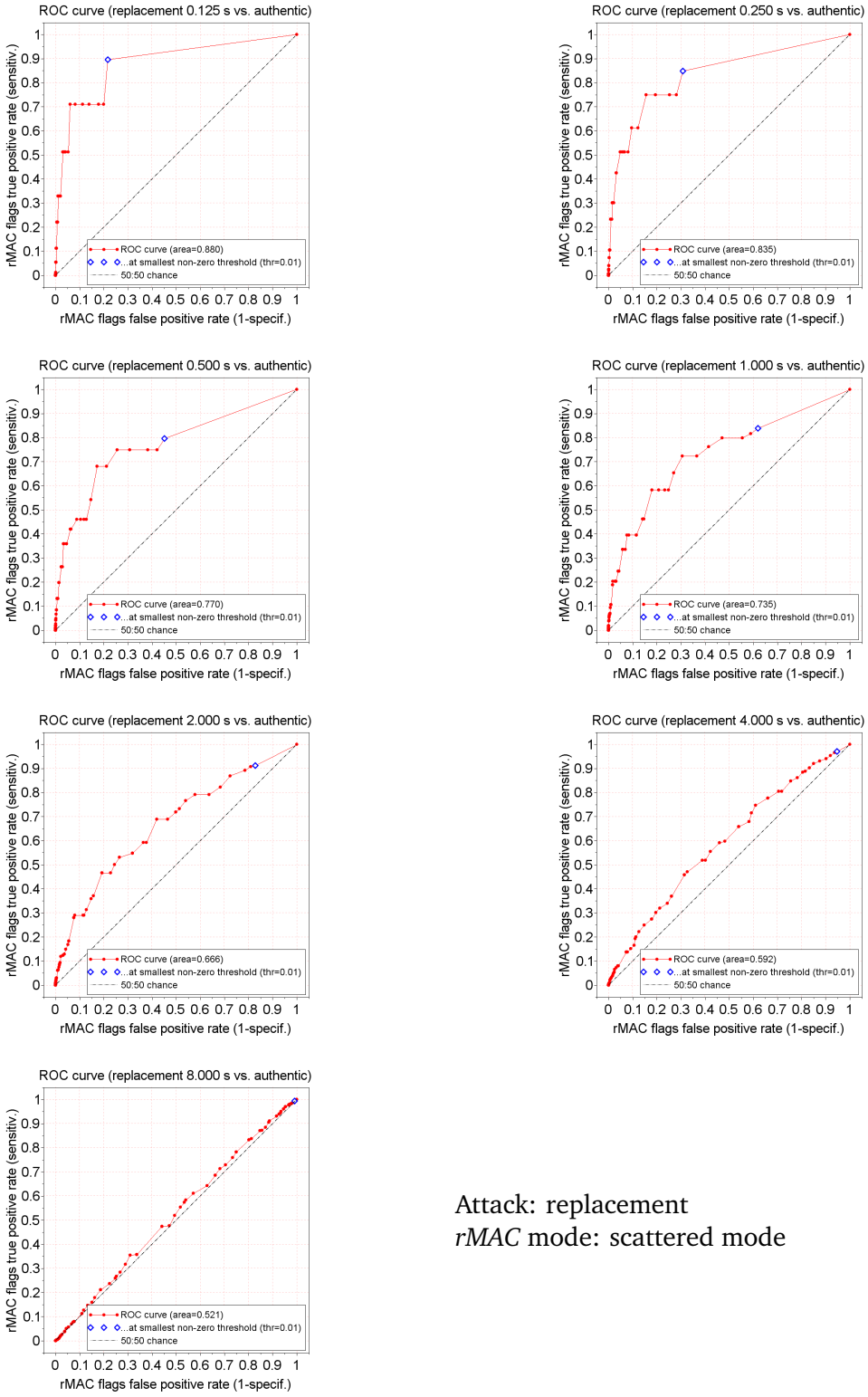
”Scattered Mode” – Deletion



Attack: deletion
rMAC mode: scattered mode

Figure C.11.: *rMAC* flag ratio as ROC plots

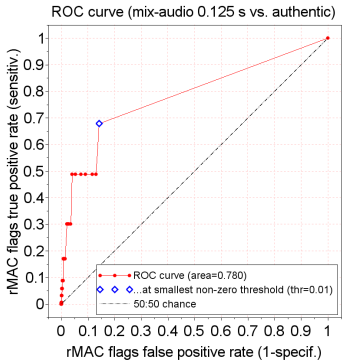
”Scattered Mode” – Replacement

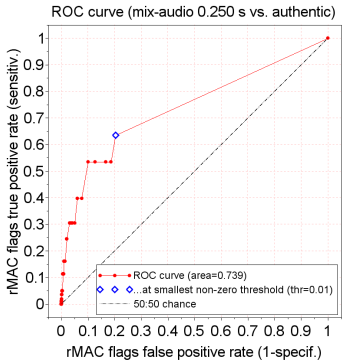


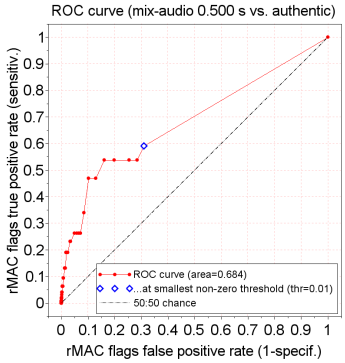
Attack: replacement
rMAC mode: scattered mode

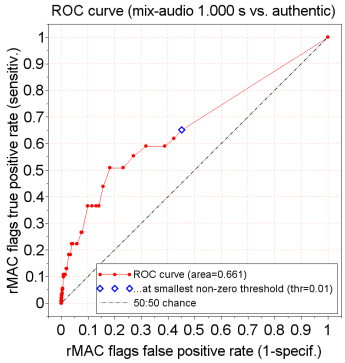
Figure C.12.: *rMAC* flag ratio as ROC plots

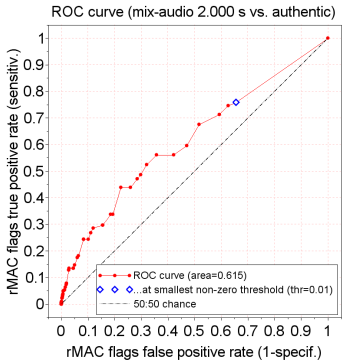
”Scattered Mode” – Mix Audio

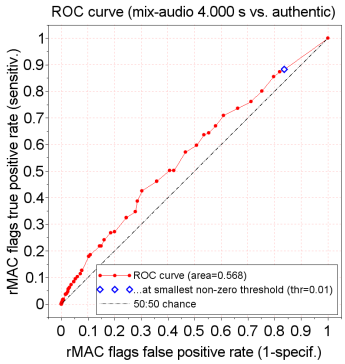


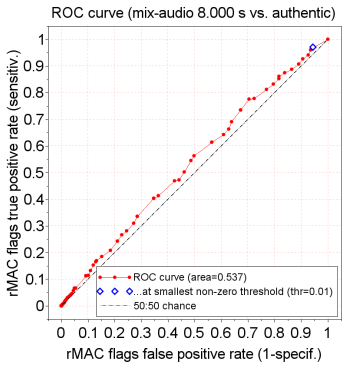








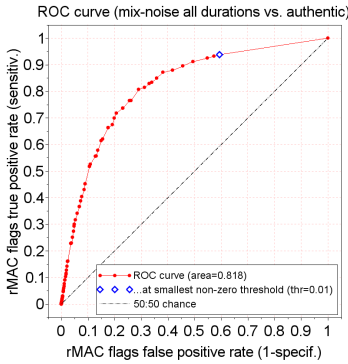




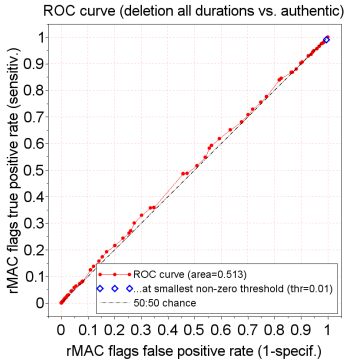
Attack: mix audio
rMAC mode: scattered mode

Figure C.13.: *rMAC* flag ratio as ROC plots

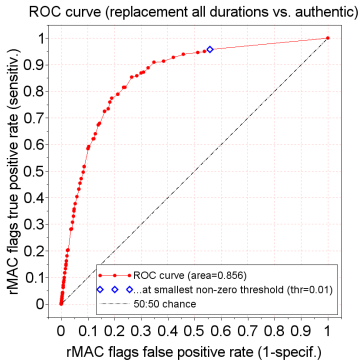
”Scattered Mode” – All Attack Durations Combined



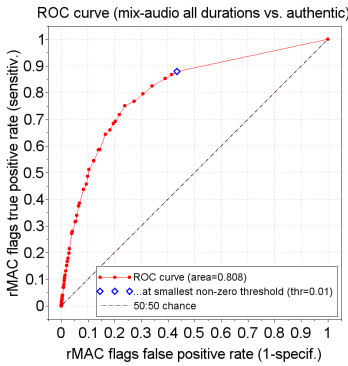
Attack: mix noise
rMAC mode: scattered mode



Attack: deletion
rMAC mode: scattered mode



Attack: replacement
rMAC mode: scattered mode

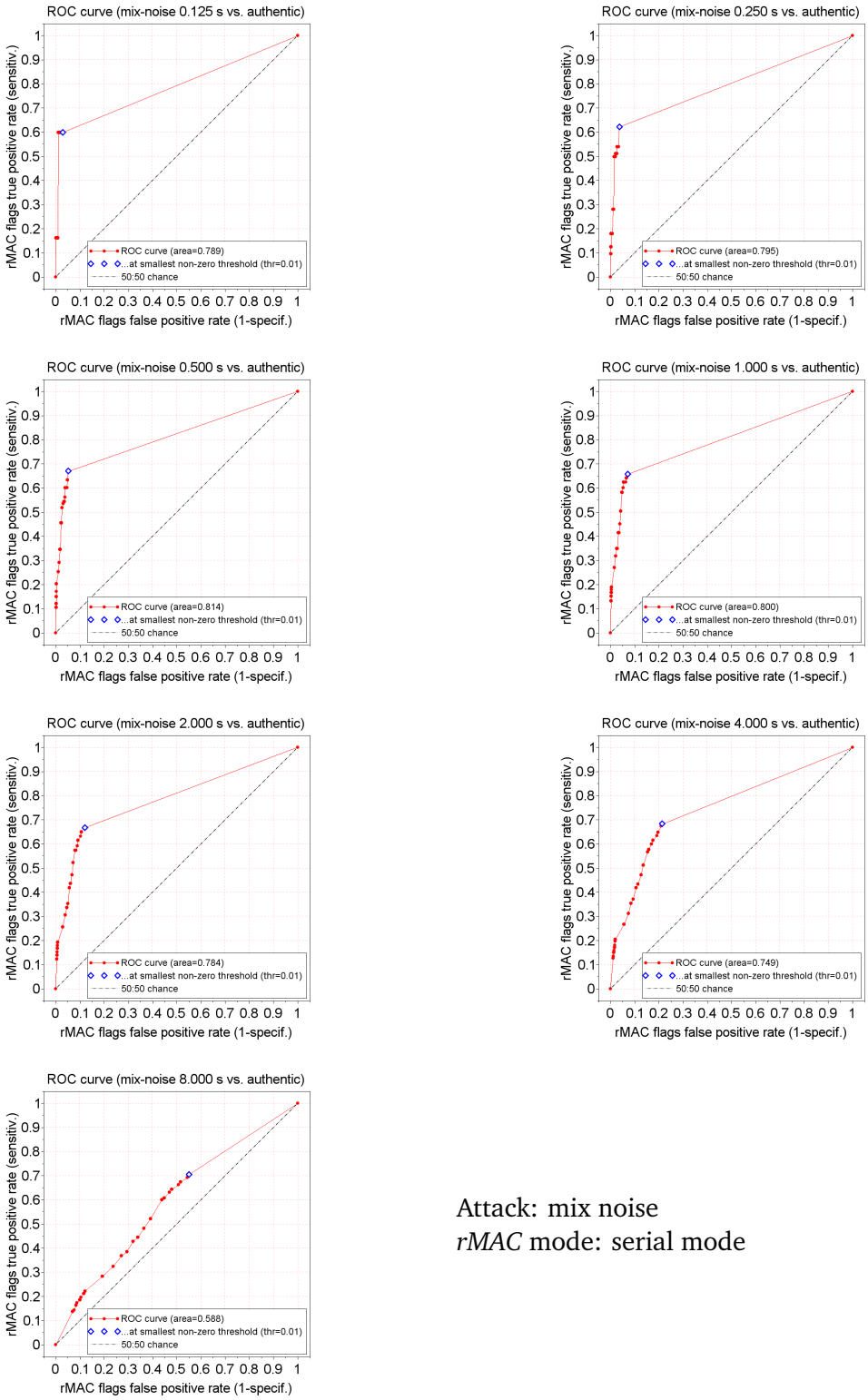


Attack: mix audio
rMAC mode: scattered mode

Figure C.14.: *rMAC* flag ratio as ROC plots including all attack durations from 1/8 s to 8 s

C.5 *rMAC* Flag Ratio: ROC curves – ”Serial Mode”

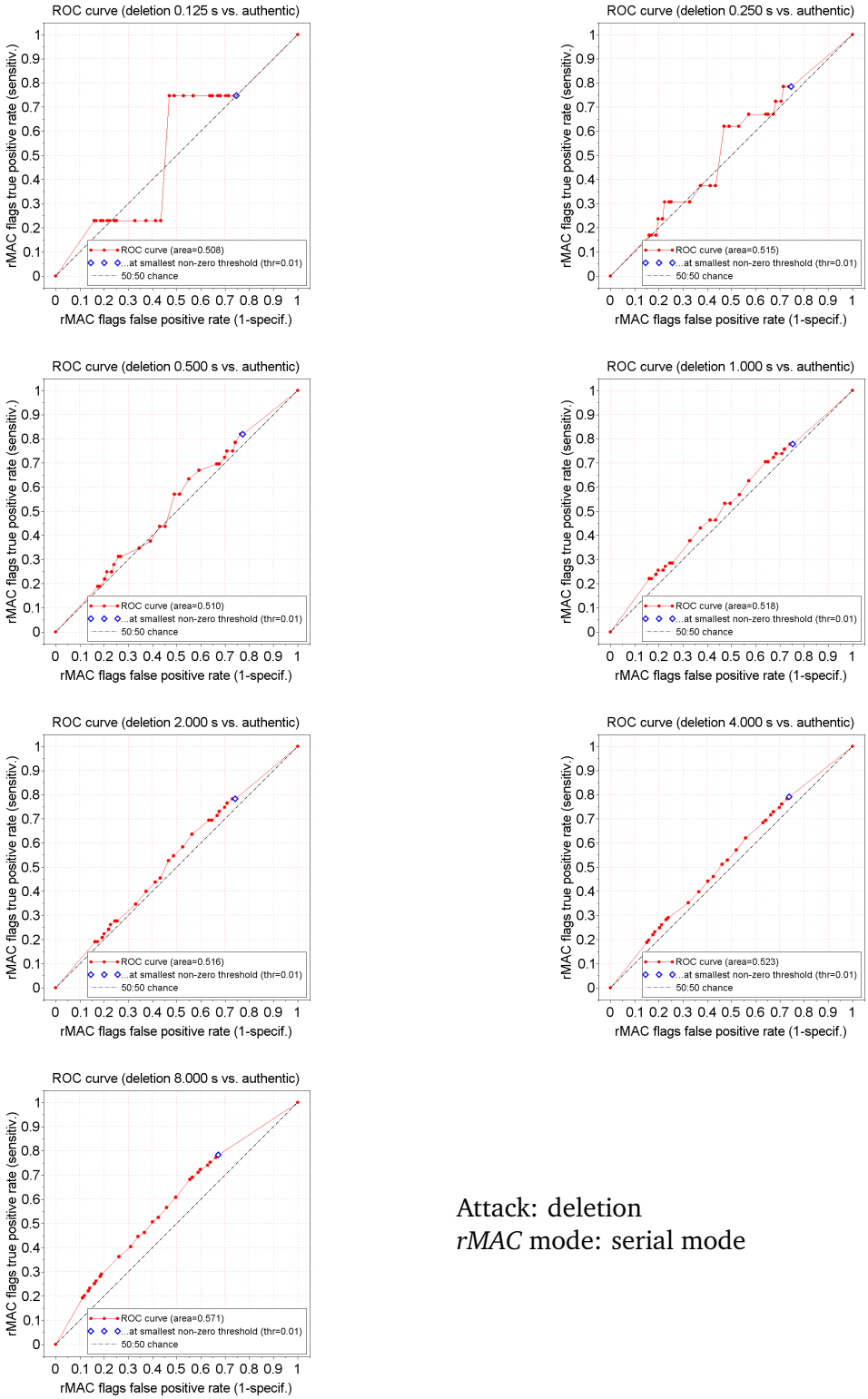
”Serial Mode” – Mix noise



Attack: mix noise
rMAC mode: serial mode

Figure C.15.: *rMAC* flag ratio as ROC plots

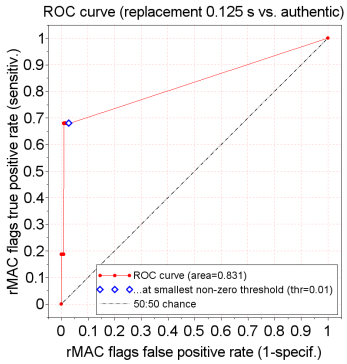
”Serial Mode” – Deletion

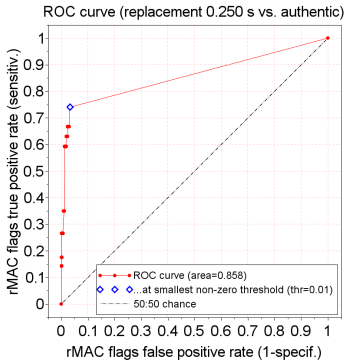


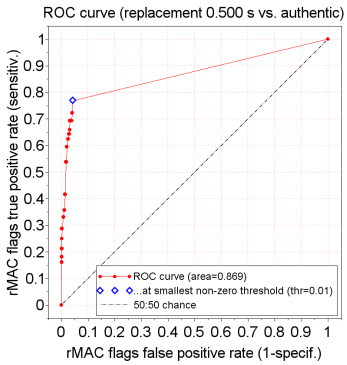
Attack: deletion
rMAC mode: serial mode

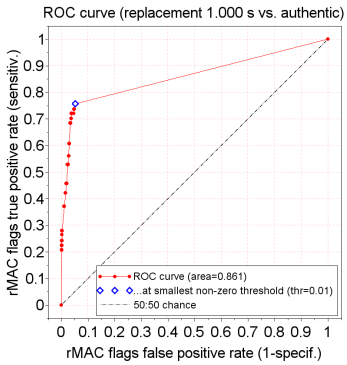
Figure C.16.: *rMAC* flag ratio as ROC plots

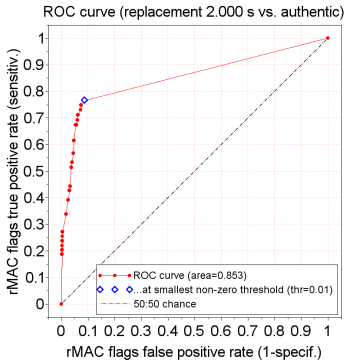
”Serial Mode” – Replacement

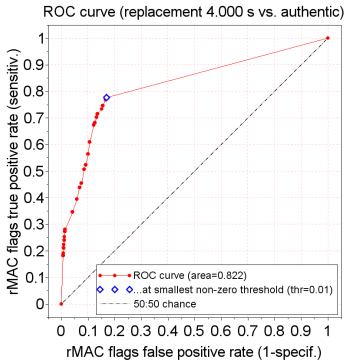


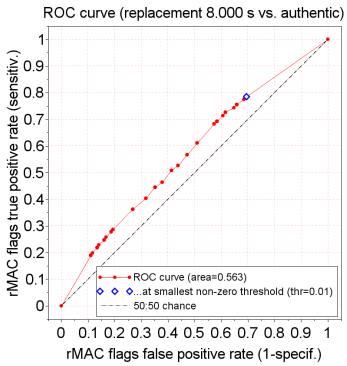








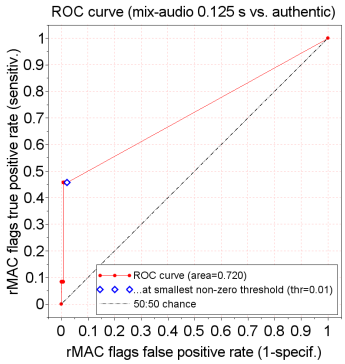


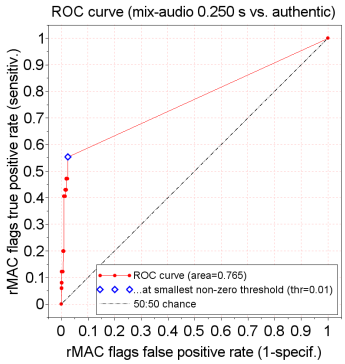


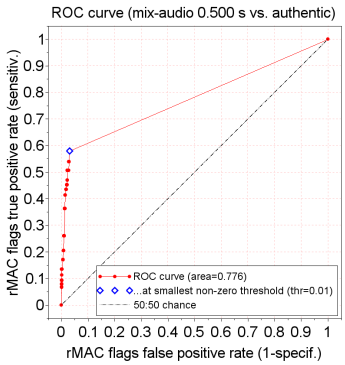
Attack: replecement
rMAC mode: serial mode

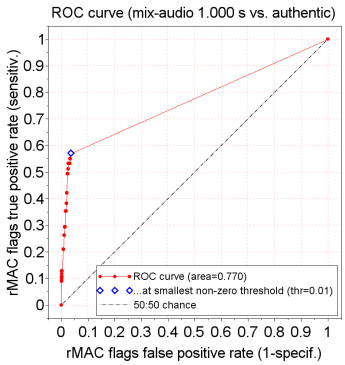
Figure C.17.: *rMAC* flag ratio as ROC plots

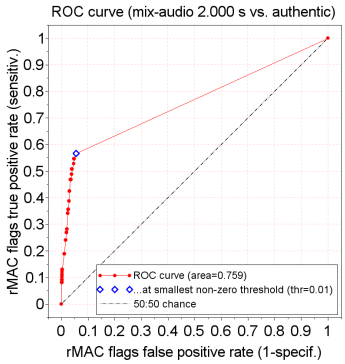
”Serial Mode” – Mix audio

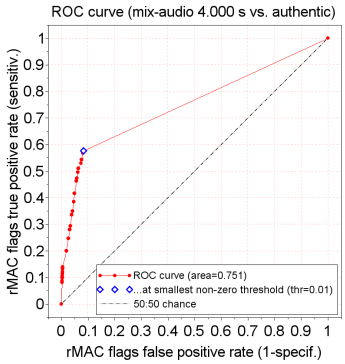


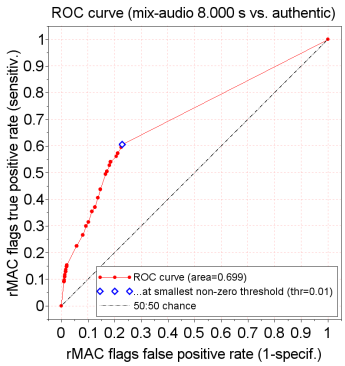








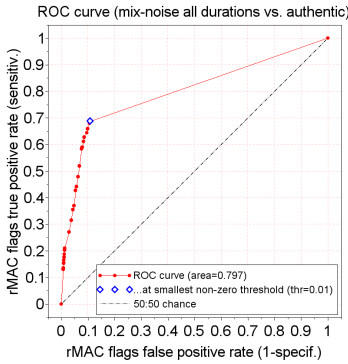




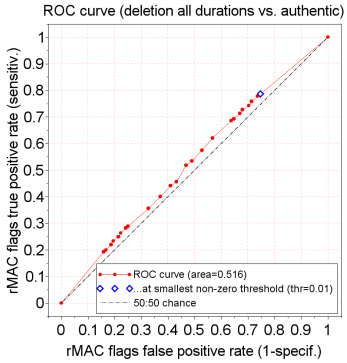
Attack: mix audio
rMAC mode: serial mode

Figure C.18.: *rMAC* flag ratio as ROC plots

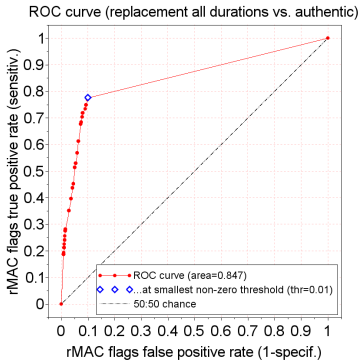
”Serial Mode” – All Attack Durations Combined



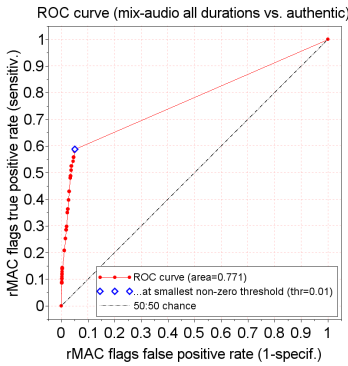
Attack: mix noise
rMAC mode: serial mode



Attack: deletion
rMAC mode: serial mode



Attack: replacement
rMAC mode: serial mode



Attack: mix audio
rMAC mode: serial mode

Figure C.19.: *rMAC* flag ratio as ROC plots including all attack durations from 1/8 s to 8 s