OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

FACULTY OF
COMPUTER SCIENCE

Statistical Pattern Recognition for Audio Forensics – Empirical Investigations on the Application Scenarios Audio Steganalysis and Microphone Forensics

Dissertation zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von: Dipl.-Inform. Christian Krätzer
geboren am 12. Dezember 1978 in Halberstadt

Gutachter:

Prof. Dr. Jana Dittmann, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

Prof. Dr. Stefan Katzenbeisser, Technische Universität Darmstadt, Darmstadt, Germany

Prof. Dr. Scott A. Craver, Binghamton University, Binghamton, NY, United States of America

Magdeburg, Germany
23. Mai 2013

Abstract

Media forensics is a quickly growing research field struggling to gain the same acceptance as traditional forensic investigation methods. Media forensics tasks such as proving image manipulations on digital images or audio manipulations on digital sound files are as relevant today as they were for their analogue counterparts some decades ago. The difference is that tools like Photoshop now allow a large number of people to manipulate digital media objects with a processing speed far beyond anything imaginable in times when image, audio or video manipulation of analogue media was a hardware- and labour-intensive task requiring skill and experience.

Currently, there exists a large number of research prototypes but only a small number of accepted tools that are capable of answering specific questions regularly arising in court cases, such as the origin (source authenticity) of an image or recording, or the integrity of a media file. The reason for this discrepancy between the number of research prototypes and the number of solutions accepted in court has to be sought in the very nature of most judicial systems: Judges have to distinguish between valid bases for evidence and expert testimonies on the one hand, and 'junk science' on the other hand. In most cases, this distinction reflects a long and cumbersome process that relies on support from the respective scientific communities.

Statistical pattern recognition (SPR) is a powerful methodology that has received a lot of attention in academia and industry since the 1930s. In many application fields, it makes possible efficient solutions for classification-based problems, such as the decision whether an e-mail is spam or not. Many current media forensics methods already implement SPR to solve very specific classification or decision problems.

This thesis introduces an SPR-based solution concept for audio signal analyses that leave the narrow tracks of such specific forensic methods. Instead, a general-purpose audio forensics concept is discussed, providing the basis for the verification of different authenticity and integrity aspects of audio signals. From the wide range of potential application scenarios that can be used to illustrate the significance of such an approach, audio steganalysis (the detection of hidden communication channels in audio data) and microphone forensics are chosen here to represent one research field focussed on integrity and one focussed on source authenticity.

The introduction of such a currently unavailable generalised approach does not solve the issue of making all current SPR-based audio forensics analyses acceptable in court. This goal would be far too ambitious for a PhD project. But the author considers it to be an important step in helping the discourse between judges on the one hand and researchers working on media forensics approaches on the other hand. A positive outcome of this discourse will improve the chances of forensic science passing hurdles such as the Daubert criteria used by judges to keep 'junk science' out of their courtrooms.

The three research challenges for this thesis are:

- *Is there a generalised SPR approach (i.e. an adaptable solution concept) for media forensics that can be used for addressing multiple audio forensics investigation goals?*

- *How can the usefulness of application scenario specific instantiations of a generalised audio forensics approach be measured?*

- *Can instantiations of the generalised audio forensics approach be used to adequately implement the application scenarios of audio steganalysis and microphone forensics?*

In order to address these challenges, detailed methodology, concept and design considerations are made, leading to an approach that is highly flexible (i.e. considers components like classification algorithms to be exchangeable off-the-shelf products), deterministic, verifiable (integrating feature selection as a means of plausibility verification for the classification process) and easily adaptable to different application scenarios. Furthermore, the possibility of assessing the usefulness of implementations of application scenario specific instantiations of the introduced concept is discussed against the background of the Daubert criteria.

The two instantiations of the introduced generalised approach selected as examples illustrate the instantiation process and the corresponding performance assessment. In the substantial empirical investigations performed in audio steganalysis, detection performances similar to those presented in the state-of-the-art in specialised audio steganalysis solutions are achieved. For the microphone forensics scenario, the approach introduced here is assumed to outperform the current state-of-the-art, as it shows similar detection performances while overcoming context limitations from which most of the specific approaches suffer.

Despite the progress made in both application scenarios, which illustrates the significance of the introduced approach, it has to be admitted that the results discussed here still fall short of the ultimate (yet for a PhD thesis unrealistic) goal of providing audio forensic solutions acceptable for court proceedings.

Deutschsprachige Version des Abstract

Medienforensik ist ein derzeit schnell wachsendes Forschungsfeld, welches anstrebt, eine ähnliche Akzeptanz wie etablierte forensische Disziplinen zu erlangen. Aufgaben der Medienforensik, wie beispielsweise der Nachweis von Manipulationen an digitalen Bild- oder Tondaten, sind heute im selben Umfang relevant, wie es ähnliche Techniken für Fotomaterial oder analoge Tonträger zuvor waren. Der Unterschied zu diesen liegt in der Verfügbarkeit von Werkzeugen wie zum Beispiel Photoshop, die den Durchsatz bei Manipulationen erhöhen, während gleichzeitig das benötigte Know-how sinkt.
Derzeit gibt es eine Vielzahl von prototypischen Lösungen aus der Forschung, welche in der Lage wären, spezielle Fragestellungen der Medienforensik, wie zum Beispiel den Nachweis der Herkunft (Quellenauthentizität) oder der Integrität, zu führen. Allerdings gibt es nur wenige solche Lösungen, die auch vor Gericht Bestand haben. Der Grund für diese Diskrepanz liegt in der Natur der meisten Rechtssysteme: Richtern obliegt es, zwischen geeigneten Grundlagen für die Beweisführung und pseudowissenschaftlichen Ramschwissenschaften zu unterscheiden. Zumeist geht der Zulassung einer wissenschaftlichen Methode im Gericht ein langwieriger Zulassungsprozess voran, bei dem die Richter auf die Zuarbeit der entsprechenden Forschungsbereiche angewiesen sind.

Statistische Mustererkennung (SPR, engl.: *statistical pattern recognition*) ist eine sehr mächtige Methodik, die seit den 1930er Jahren in Forschung und Industrie eine starke Akzeptanz erfährt. In vielen Bereichen ermöglicht sie effiziente Lösungen für klassifikationsbasierte Probleme, wie zum Beispiel die Frage, ob eine E-Mail Spam ist oder nicht. Viele aktuelle Ansätze der Medienforensik nutzen diese Methodik für die Lösung spezieller Klassifikations- oder Entscheidungsprobleme. In dieser Dissertation wird ein SPR-basiertes Lösungskonzept vorgestellt, welches die ausgetretenen Pfade hochspezialisierter forensischer Methoden verlässt und ein Universalwerkzeug für unterschiedliche forensische Analysen der Authentizität und Integrität von Audiodaten darstellt. Aus dem umfangreichen Pool an möglichen Einsatzszenarien für ein solches Universalwerkzeug werden hier zwei Beispiele zur Illustration der Praxisrelevanz des vorgestellten Ansatzes vorgestellt. Dabei handelt es sich um die Audiosteganalyse (die Erkennung von verdeckten Kommunikationskanälen in Audiodaten) und die Mikrofonerkennung. Durch diese Auswahl wird jeweils eine Lösung für die Verifikation der Integrität von Audiodaten und eine Lösung für die Verifikation der Quellenauthentizität betrachtet.
Die Einführung eines solchen, derzeitig fehlenden, generalisierten Ansatzes wird nicht automatisch alle derzeitig auf statistischer Mustererkennung basierenden Verfahren der Medienforensik für Audiodaten vor Gericht verwertbar machen. Dies wäre ein viel zu ambitioniertes Unterfangen für eine einzelne Dissertation. Allerdings betrachtet der Autor die hier durchgeführten Arbeiten als einen wichtigen Schritt im Diskurs zwischen Richtern einerseits und Forschern im Bereich der Medienforensik andererseits. Ein positiver Ausgang dieses Diskurses sollte die Chancen dafür erhöhen, dass Forschungsergebnisse erfolgreich die Hürden nehmen, welche von Richtern genutzt werden, um Pseudowissenschaften aus den Gerichtssälen fernzuhalten (zum Beispiel die Daubert-Kriterien der US-Rechtsprechung).

Die drei wissenschaftlichen Herausforderungen, die in dieser Arbeit aufgegriffen werden, sind:

- *Gibt es einen generalisierbaren SPR-Ansatz (d. h. ein anpassbares Lösungskonzept) im Bereich der Medienforensik, welcher verwendet werden kann, um mehrere unterschiedliche Untersuchungsziele der forensischen Audiodatenanalyse zu erfüllen?*
- *Wie kann die Nützlichkeit der Instantiierungen eines solchen generalisierten Ansatzes für praktische Untersuchungen in unterschiedlichen Anwendungsszenarien bewertet werden?*
- *Können solche Instantiierungen genutzt werden um adäquate Lösungen für die Anwendungsszenarien der Audiosteganalyse und der Mikrofonforensik zu erzeugen?*

Um Antworten auf diese Fragen zu finden, werden Methodik, Konzepte und mögliche Designs detailliert betrachtet. Diese führen zu einem Ansatz, welcher hochgradig flexibel (d. h., Komponenten wie zum Beispiel Klassifikationsalgorithmen werden als leicht austauschbare Elemente angesehen), deterministisch, nachvollziehbar (durch Nutzung von Merkmalsselektion für die Verifikation der Plausibilität im Klassifikationsprozess) und für unterschiedliche Einsatzzwecke leicht adaptierbar ist.

Des Weiteren wird vor dem Hintergrund der Daubert-Kriterien diskutiert, wie die Nützlichkeit von Instantiierungen eines solchen generalisierten Ansatzes bewertet werden kann.
Für die beiden beispielhaft ausgewählten Anwendungsszenarien werden sowohl der Instantiierungsprozess durchgeführt als auch die Leistungsfähigkeit entsprechend bewertet. Für das Anwendungsszenario der Audiosteganalyse werden dabei in den umfangreichen empirischen Analysen Ergebnisse erzielt, welche äquivalent zu denen spezialisierter Ansätze aus dem aktuellen Stand in Forschung und Technik sind. Für die Mikrofonforensik kann für den hier vorgestellten Ansatz angenommen werden, dass er leistungsfähiger ist als die derzeit existierenden spezialisierten Alternativen, da er bei vergleichbaren Erkennungsraten geringere Einschränkungen bezüglich aufgenommener Inhalte aufweist.

Trotz der wissenschaftlichen Erfolge, die für beide Anwendungsszenarien erzielt werden, sei an dieser Stelle erwähnt, dass die hier erzielten Ergebnisse immer noch weit davon entfernt sind, dem schlussendlichen Ziel jeder forensischen Methode – der Gerichtsverwertbarkeit – nahezukommen. Dies zu erreichen wäre ein Ziel, welches für eine Dissertation viel zu hoch gesteckt ist.

Acknowledgements

This work was prepared during my time as a research assistant at the Multimedia and Security research group (AMSL) at the Faculty of Computer Science of Otto-von-Guericke University Magdeburg, Germany.

First of all, I would like to express my gratitude to Prof. Dr.-Ing. Jana Dittmann for all our joint work during the last nine years, for supporting my work at the AMSL and for supervising this thesis. Her advice and interest have been essential to this work. In addition to the constant support she gave to my own research activities, I learned a lot in terms of project-based research work (research and development, management, and acquisition) while working at AMSL. The experience from research projects such as ECRYPT, SHAMAN, ECRYPT II, POWER, Digi-Dak, HEU, DigiDak+ and KOMMmodel contributed much to my professional self-perception. I also am grateful to her for allowing me to spend quite a lot of time teaching classes and supervising students, two activities I really like.

Furthermore, I would like to thank all my current and former colleagues at AMSL, who have always provided a friendly and motivating working atmosphere. In particular, I would like to thank Stefan Kiltz, best man at my wedding and office mate for the last seven years, as well as Silke Reifgerste, who often made the impossible possible.

During my time at AMSL a large number of bachelor and master theses were (co-)supervised by me. Some of those theses contributed to this work, especially to the creation of the microphone recording setups, the implementation of some of the information hiding (IH) algorithms, and the development, improvement as well as evaluation of the audio feature extractor AAFE used here. Therefore, I want to thank our former students Stefan Sokoll (work on AAFE), Sebastian Heutling (AAFE and one IH algorithm), Christian Zeitz (AAFE), Reyk Hillert (AAFE), Marcel Dohnal (creation of microphone recording set *RS3*), Thomas Naumann (AAFE), Jan Leif Hoffmann (implementation of an IH algorithm), Jörg Wissen (creation of recording set *RS1*), the team consisting of Nataliya Kulyk, Xiangyu Wang and Shen Liu (creation of recording set *RS2*), Carmen Pohl (for helping me to create recording sets *RS15*, *RS16*, *RS17* and *RS18*) and Christian Spillker (AAFE).

I would like to give special thanks to my (former) colleagues and co-authors on the conference and workshop papers accompanying this PhD thesis: Thomas Vogel (joint work on audio steganalysis), Robert Buchholz (joint work on spectrum-based microphone forensics), Andrea Oermann (theoretical foundations of microphone forensics), Maik Schott (potential integration of microphone forensics into trustworthy archiving solutions), Andreas Lang (for our joint work on watermarking benchmarking), Kun Qian (context modelling for the recording process) and Claus Vielhauer (for his brilliant idea of transferring the Mel-cepstral based signal analysis from biometric speaker verification to the domain of steganalysis).

My work on detectability analyses for information hiding schemes has been supported in part by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT[1]. The work on recording source authentication has been supported in part by the European Commission through the FP7 ICT Programme under Contract FP7-ICT-216736 SHAMAN[1].

During my time at AMSL, additional research activities, which are not related to the content of this thesis, were funded in part by research projects of the European Commission (project ECRYPT II; ICT-2007-216676), and the EOARD (project POWER; EOARD grant number FA8655-09-M3061) as well as German federal funding (projects HEU, KOMMmodel, Digi-Dak and DigiDak+ (German Federal Ministry of Education and Science (BMBF) through the Research Programmes "Digi-Dak (Digitale Fingerspuren)" & "DigiDak+ Sicherheits-Forschungskolleg Digitale Formspuren" under Contract No. FKZ: 13N10818)).

[1] Disclaimer for EU FP6 and FP7 projects: The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

As a researcher who has published papers in IT-security related workshop and conference proceedings as well as a journal, I would like to express my gratitude to all reviewers involved in the publication process. They mostly remained anonymous but nevertheless with their comments helped to shape the research presented within this thesis.

Two reviewers who helped to improve the final version of this thesis and do not remain unknown to me are Prof. Dr. Stefan Katzenbeisser (TU Darmstadt, Germany) and Prof. Dr. Scott A. Craver (Binghamton University, USA), the two external reviewers and also members of the doctoral committee for this thesis. I want to use this chance to thank them and the other members of my doctoral committee (Prof. Dr. Holger Theisel, Prof. Dr. Rudolf Kruse and Prof. Dr. Myra Spiliopoulou, all Otto-von-Guericke-University Magdeburg, Germany).

Chet D. Hosmer, Chief Scientist & Sr. Vice President of Allen Corp., and Walter Bruehs, FBI Forensic Audio, Video, and Imagery Analysis Unit, both had a strong influence on this PhD project. Their invited talks at the ACM Multimedia and Security Workshop 2011 in Buffalo, NY, as well as the following discussions with the author convinced me to spin the evaluation criteria towards discussing the Daubert compliance of audio forensics methods. For an academic researcher like me, it was impressive to hear from those two experts in the fields of (practically applied) steganalysis (C. D. Hosmer) and forensic imagery analysis (W. Bruehs, one of the FBIs Photographic Evidence Examiners) about their first-hand experience in forensic media object analyses in U.S. legal procedures. In this regard, I would also like to thank Jessica Fridrich and her research group for the breakthrough they seem to have achieved for the whole research field with their PRNU-based camera forensics approach in the law case *United States of America v. Nathan Allen Railey*[2], where this image forensics approach successfully passed the Daubert hearing.

I do not only consider Jessica Fridrich to be an outstanding researcher but I also know her to be a patient and friendly listener. I have to thank her, Tomáš Filler and Jan Kodovský for their moral support at one of the toughest points of this PhD project (August and September 2011). Together they managed to point out (without stating the obvious) that the only thing that slowed me down into a complete stalemate at this point was my own uncertainty. After this realisation, the solution of the problem was simple.

Anja Schüler has been a true friend of the family during the last few years. Appart from all the help she provided to my wife Judith and me, she read early versions of this thesis, giving valuable feedback especially on the accessibility of chapter 1 and the English and German versions of the abstract.

Last but not least I wish to thank my family: Ich danke meiner gesamten Familie, welche dieses Unterfangen über Jahre hinweg tatkräftig unterstützt hat. Besonderer Dank gilt dabei natürlich meinen Eltern, ohne die weder mein Studium noch die Promotion möglich gewesen wären, und deren Einstellung zur Bildung mir immer ein leuchtendes Vorbild war.

Mehr als jeder Andere mit mir gelitten hat meine geliebte Frau Judith. Für ihre Geduld mit mir und diesem Promotionsvorhaben bin ich ihr unendlich dankbar!

Mein letzter Gruss geht an unseren Nachwuchs Oskar und Emma Krätzer, die seit November 2012 dieses Unterfangen begleiten: Jetzt, da diese Dissertation endlich abgeschlossen ist, gehören Papas Feierabende, Wochenenden und Urlaubstage endlich euch! ☺

Christian Krätzer
Magdeburg, May 23rd, 2013

Legal notices:

The author has absolutely no legal training. All legal considerations made in this thesis (especially the considerations on the forensic compliance of methods and the Daubert standard) are layman interpretations of freely available material, which are made to the best of the author's knowledge. If the content of this thesis is used in any legal proceedings, the reader must consult appropriate legal counsel for the corresponding jurisdiction.

[2] United States District Court for the Southern District of Alabama, August 2nd, 2011.

Contents

Acronyms

AAFE	AMSL Audio Feature Extractor
AAST	AMSL Audio Steganalysis Toolset
AC	alternating current
ACM	Association for Computing Machinery
A/D	analogue to digital, as in analogue to digital converter
AES	Advanced Encryption Standard
AMSL	Advanced Multimedia and Security Lab, Otto-von-Guericke University Magdeburg, Germany
API	application programming interface
ASCII	American Standard Code for Information Interchange
ASCLD	American Society of Crime Laboratory Directors
AUR	area under ROC
BOSS	Break Our Steganographic Scheme
BOWS	Break our Watermarking Scheme
BPP	Bit per pixel
CCITT	Comité Consultatif International Téléphonique et Télégraphique (now ITU)
CD	compact disc
CPU	central processing unit
DC	direct current
DCT	discrete cosine transform
DES	Data Encryption Standard
DFD	Digital Forensic Database
DFT	discrete Fourier transform
DVD	Digital Versatile Disc
DWT	discrete wavelet transform
ECC	error correcting code
EICAR	European Institute for Computer Antivirus Research
ENF	electric network frequency
ENFSI	European Network of Forensic Science Institutes
FAVIAU	Forensic Audio, Video, and Image Analysis Unit of the FBI
FBI	Federal Bureau of Investigation
FFT	fast Fourier transform
FLAC	Free Lossless Audio Codec

FMFCC	Filtered Mel-Frequency Cepstral Coefficients
FN	false negative
FP	false positive
FPR	false positive rate
FRE	Federal Rules of Evidence
FSAAWG	Forensic Speech and Audio Analysis Working Group
FT	Fourier transform
GB	gigabyte
GMM	Gaussian Mixture Models
GSM	Global System for Mobile Communications
HUGO	steganographic system used in the BOWS contest
Hz	Hertz
IAI	International Association for Identification
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IH	information hiding
INTERPOL	International Criminal Police Organization
IP	internet protocol
ISBN	International Standard Book Number
ISO	International Organization for Standardization
IT	information technology
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
JPEG	Image encoding standard ISO/IEC 10918-1 introduced by the Joint Photographic Experts Group
LSB	least significant bit
MAC	media access control
MB	megabyte
MD5	Message-Digest Algorithm 5 – a cryptographic hash algorithm
MFCC	Mel-Frequency Cepstral Coefficients
MIDI	Musical Instrument Digital Interface
MMA	MIDI Manufacturers Association
MP3	MPEG-1 or MPEG-2 audio layer III
MPEG	Moving Picture Experts Group
NIST	National Institute of Standards and Technology

PCA	principal component analysis
PCM	Pulse-code modulation
PCMU	Pulse Code Modulation μ-law as defined in CCITT/ITU-T recommendation G.711
PDF	Probability density function
PR	pattern recognition
PRNG	Pseudo-Random Number Generator
PRNU	photo response non uniformity
RAM	random-access memory
REWIND	EU FP7 project REVerse engineering of audio-Visual content Data
RMS	root mean square
ROC	Receiver Operation Characteristic
SARC	Steganography Analysis and Research Center – a company based in Fairmont, WV, USA
SMBA	StirMark Benchmark for Audio
SPIE	Society of Photo-Optical Instrumentation Engineers
SPR	statistical pattern recognition
SPSS	Statistical Product and Service Solutions
SQAM	Sound Quality Assessment Material
SVM	support vector machine
SWGFAST	Scientific Working Group on Friction Ridge Analysis, Study, and Technology
TN	true negative
TP	true positive
USB	Universal Serial Bus
USENET	Unix User Network
VoIP	Voice over IP
WEKA	An open source data mining suite developed and maintained by the Machine Learning Group at University of Waikato, New Zealand
ZIP	File format used for data compression and archiving

Nomenclature

$\bar{H}_{histpos}$	Histogram of the time-domain sample values in a window of the audio signal, divided by the window size
$\bar{S}_i^k$	Stream of audio frames
ϵ	Comparison descriptor for theoretical considerations on steganographic security
κ	Kappa statistics used within this thesis a detection performance metric – see section 4.1.4
$\mathrm{FT}()$	Fourier transform
$\mathrm{FT}^{-1}()$	Inverse Fourier transform
T	Sample-time; $\mathrm{T} = 1/f_{sample}$
$\mathrm{win}()$	Windowing function
time()	Unix time command
arraySort$_{pointer}$()	Function sorting an array and returning the value at position $pointer$ after the sorting process
BandFilter()	Helper function removing in the frequency-domain the audio content between 200 and 6819.59 Hz
fr(h)	Frequency belonging to coefficient-in-the-frame index h
LSB()	Function returning the LSB-value of a sample
LSB$_0(s_{i,j}^k)$	Boolean check whether the LSB-value of a sample $s_{i,j}^k$ is ‘0’
LSB$_1(s_{i,j}^k)$	Boolean check whether the LSB-value of a sample $s_{i,j}^k$ is ‘1’
MelScaleTransformation()	Helper function performing a Mel-scale transform of a spectrum
$\tilde{S}(t)$	Analogue audio signal in time-domain
$\tilde{Y}(f)$	Frequency-domain representation of an analogue audio signal
$\tilde{Y}_C(f)$	Frequency-domain representation of an analogue-to-digital converted audio signal
$\tilde{Y}_E(f)$	Frequency-domain representation of an analogue audio signal after environmental shaping
$\tilde{Y}_P(f)$	Frequency-domain representation of an analogue audio signal directly after its generation / projection
$\tilde{Y}_R(f)$	Frequency-domain representation of an analogue audio signal after recording using a microphone
$\tilde{Y}_T(f)$	Frequency-domain representation of an analogue audio signal after transmission from a microphone to a processing unit
A_{S*}	one of the five steganography algorithms used in this thesis
A_{W*}	one of the five watermarking algorithms used in this thesis
$accuracy$	Classification accuracy
Age	Microphone aging influence to a microphone recording
Alg	data hiding algorithm

B_i^k	Bark scale spectrogram
$C_k(n\mathrm{T})$	Cover object as digital audio signal with k channels
$coef$	Number of frequency bands returned by the Fourier transform of an audio signal
cv	test mode 10-fold stratified cross-validation
d	Dimensionality of the feature space considered in a pattern recognition (PR) problem
$D(f)$	Discolouration function resulting from short-term reflections of audio signals
e	Enhancement factor imposed by medium-term reflections of audio signals
$Embed_{Alg}()$	Embedding function for a data hiding algorithm Alg
f	Frequency
$F_{driver}(f)$	Loudspeaker amplification function
$F_{membrane}()$	Transfer function of the microphone membrane
$F_{mic}(f)$	Frequency response function of the microphone
$F_{reverb}(f)$	Echoes and / or reverberation influences to an audio signal
f_{sample}	Sampling frequency (or Nyquist frequency)
$F_{samp}(f)$	Sampling function
$F_{tran}(f)$	Non-linear distortion during the transmission of the signal from the microphone to recording device
$framecount$	Number of frames in an channel of the audio data stream
gf_*	Segmental feature – for the resolution of the wildcard $*$ see table 4.1 in section 4.1.1
GF_{all}	set of global features containing all 17 such features computed by AAFE v.2.0.5
H	Histogram of the time-domain sample values in a window of the audio signal
h	Index for the frequency coefficients in the spectrum of a window of the audio signal
i	Frame index in a stream of audio frames
j	Sample-in-the-frame index in a stream of audio frames
k	Number of channels in an audio signal
key	Embedding key for a data hiding algorithm
l	Lower bound of the frequency range of a driver in a loudspeaker
$length$	Number of complete audio frames in a framed audio data stream
low	Lower frequency bound for a formant
M_*	A microphone used in the microphone forensics evaluations
$MembCharacteristics$	Membrane characteristics of a microphone
$message$	The message embedded by a data hiding algorithm
$Mount$	Mounting influence to a microphone recording
n	Sample index in a sampled audio signal, with $n \in \mathbb{N}$; $1 \leq n \leq length$

n_{driver}	Number of drivers in a loudspeaker
$N_{ENF}(f)$	Electric network frequency (ENF) influence to the recording
$N_{envi}(f)$	Environmental noise influences to a recorded audio signal
$N_{ls}(f)$	(Thermal) noise that a loudspeaker generates in the playback signal
$N_{mic}(f)$	Thermal noise generated by the microphone
$N_{quan}(f)$	Quantisation noise in A/D conversion of an audio signal
$N_{thermal}(f)$	Thermal noise of the A/D converter
$N_{tran}(f)$	Thermal noise imposed to the signal by the transmission environment
o	Overlap of neighbouring audio frames
Or	Orientation influence to a microphone recording
P_a	Percentage agreement between multiple classifiers or a classifier and ground truth – see section 4.1.4
P_c	Distribution of the cover objects in a steganographic scheme
P_s	Distribution of the stego objects
P_{chance}	Probability of chance agreement between multiple classifiers or a classifier and ground truth – see section 4.1.4
$pointer$	Index in a sorted array returned by the function arraySort$_{pointer}$()
q	naive classifier throughput performance metric
q_{new}	classifier throughput performance metric
R	Quantisation range
$runtime$	Classifier runtime (training and testing)
$S(t)$	Time-domain representation of an analogue-to-digital converted audio signal
$S^k(n\mathrm{T})$	Sampled audio data stream
S_i^k	Sampled, quantised and windowed digital audio signal
sf_*	Segmental feature – for the resolution of the wildcard $*$ see table 4.1 in section 4.1.1
SF_{all}	set of segmental features containing all 590 such features computed by AAFE v.2.0.5
$SFQ_{Classifier}()$	transfer function for a detector (i.e. classifier) into the benchmark
$ST^k(n\mathrm{T})$	Steganogram as output of the application of an data hiding algorithm to an audio signal
$streamlength$	Overall number of audio samples in a sampled audio signal
$strength$	The embedding strength parametrised for a data hiding algorithm
t	Time
$thresh$	Silence threshold
$time$	normalised runtime description
$trte$	test mode independent training and testing with two different audio sets

u	Upper bound of the frequency range of a driver in a loudspeaker
up	Upper frequency bound for a formant
w	Frame size of an audio frame
Y_i^k	Frequency-domain representation (spectrum) of a window of the audio signal
aats389	Set of 389 files used as multi-genre cover source for steganalysis
ahss1	Set of 10 files used as single-genre cover source for steganalysis
longfile	Long recording used as single-genre cover source for steganalysis
*R**	Recording location in the microphone forensics evaluations
ref10	Set of 10 reference files used in the microphone forensics evaluations
ref2	Set of two reference files used in the microphone forensics evaluations; subset of *ref10*
*RS**	Recording sets used in the microphone forensics evaluations
testset24	Set of 24 files used as multi-genre cover source for steganalysis
*weka.classifiers.**	The set of 74 supervised classification algorithms provided by WEKA (version 3.6.1) – see section 4.1.3
*weka.clusterers.**	The set of eight clustering algorithms provided by WEKA (version 3.6.1) – see section 4.1.3

List of Figures

1

Introduction and Motivation

Trust in digital-born or digitised media strongly depends on their source as well as their content. If we receive the media object from a source we trust, we also tend to trust the information that is contained within the object. The reason is that we assume that the source has authenticated the information and that it vouches for the integrity of the data object – otherwise we would have no trust in this source.
If the content is too implausible we might override our initial trust assumption for the source. This is very well illustrated by the example of the Ztohoven nuclear bomb prank: On June 17th, 2007 viewers of a Czech television channel, watching a web cam program monitoring weather in various Czech mountain resorts, could see something that appeared to be a nuclear explosion taking place in the Krkonose or Giant Mountains in the northern part of the Czech Republic. This prank (the explosion was obviously not real, even though it looked realistic) was implemented by the Czech prankster group Ztohoven by manipulating the web cam data stream. It won its authors the 2007 'NG 333' prize for young artists by Pragues National Gallery together with a cash prize of 333,000 Czech Koruna (at that time about US$ 18,350). In the criminal investigation following the event, a court cleared the members of Ztohoven of charges, but several members had to pay a fine of 50,000 Czech Koruna each (US$ 2,400) for tampering with a television broadcast.
This example shows that, in an information society working on digital information objects which are much easier to manipulate than their analogue counterparts, we need (security) mechanisms to establish trust in information regardless of its origin. This trust is especially based on the authenticity and integrity of digital objects. The assurance of authenticity and integrity in media is directly linked to trust and therefore to the value of the information.

That the question of authenticity and integrity is not only limited to what we can see in images or on TV, but is also relevant for the domain of audio signals considered in this thesis is very well illustrated by the Watergate Tapes (a.k.a. the Nixon White House tapes or the Nixon's Watergate tape) example. In [Maher10], the official 87-page forensic report [Bolt74] on this prominent historical example, also known as the '18 $^1/_2$ minute gap', is summarised as follows:

> *"The watershed event for audio forensics was arguably the 1974 investigation of a White House conversation between U.S. President Richard M. Nixon and Chief of Staff H.R. Haldeman recorded in the Executive Office Building in 1972. Investigators discovered that the audio recording contained an unexplained section lasting 18 $^1/_2$ minutes during which buzz sounds could be heard but no discernable speech sounds were present. Due to the highly specialized nature of the technical evidence,* [the] *Chief Judge* [...] *appointed a special Advisory Panel on White House Tapes to give expert advice to the court. The advisory panel consisted of six technical experts,* [...], *with the court's direction* '[...] to study relevant aspects of the tape and the sounds recorded on it' [Bolt74]. [...] *The advisory panel performed a series of objective analyses of the tape itself, the magnetic signals on it, the electrical and acoustical signals generated by playback of the tape, and the properties of the recording equipment used to produce the magnetic signals on the tape. Analysis included observation of the audio signals as well as magnetic development of the domain patterns and head signatures on the tape. Ultimately the panel determined that the 18 $^1/_2$-minute gap was due to several overlapping erasures performed with a specific model*

> *of tape recorder that differed from the device that produced the original recording. The panel's conclusion was based primarily on the characteristic start/stop magnetic signatures present on the subject tape."*

In this example the integrity violation proved by the advisory panel via means of source authentication, led to the resignation of U.S. president Richard M. Nixon, a historic event without precedent. But modern day audio recording and audio manipulation do no longer impose head signatures to the recordings. In [Davis12], Julie Hirschfeld Davis and Greg Stohr report upon a case of digitally tampered audio recordings of a statement by U.S. president Barack H. Obama's top Supreme Court lawyer, Solicitor General Donald Verrilli, in the U.S. Supreme Court. The tampered material, which was used in a Republican Party Internet advertisement to attack president Obama's health-care law, shows how easily digital audio manipulations can nowadays be applied to influence the public opinion. This case of audio material tampering was evident even without a sophisticated forensic analysis. In contrast to the Watergate Tapes example, it was a rather sophisticated manipulation, but the original material is available to the public, and therefore proving the manipulation was easy. The interesting question is: How many audio signal manipulations altering the outcome of court cases or the public opinion have gone unnoticed?

Besides the questions on source authenticity and integrity illustrated above, another hot topic to be addressed in the context of this thesis is the issue of growing importance of digital steganography. Steganography (from ancient Greek: *steganos graphein* – in English: covered writing; see [Wölfel11] or [Fridrich09]) is the art and science of hidden communication. Its counter-science is steganalysis, attempting to detect the existence of hidden communications in observed communication channels. History is full of descriptions of steganographic methods from all epochs: the Ancient Greek, the Middle Ages, European Renaissance, the World Wars, the Cold War era, etc. In his most influential book 'The Codebreakers' ([Kahn1996]), David Kahn illustrates how steganography and cryptography (as well as their counter-sciences steganalysis and cryptanalysis) have been used as alternatives for confidential communication throughout the times. With the emergence of the digital age, the 1990s also saw the emergence of new digital steganographic methods, followed by the need for corresponding counter-methods (i.e. for steganalysis methods).
Despite the facts that no statistics exist on the actual usage of digital steganography, its relevance is illustrated by such facts as Hollywood studios taking up the idea for their spy movies, it making its way into contemporary popular literature (e.g. Tom Clancy's 'Dead or Alive', [Clancy12]) and its usage being strongly advertised in Jihadist newsletters (see [Givner-Forbes07]).

Regarding the issue of trust connected to such authenticity and integrity questions, this thesis elaborates two perspectives on forensic sciences, the information technology (IT) perspective and the society perspective:

The place of forensics in IT-security:
The notion of **trust** in information technology (IT) encompasses three different components: a **security** component, a **safety** component and a **non-technical** component, which combines all human influence factors for a trust assumption (i.e. our own experiences, knowledge, fears, etc.). From these three components, only the IT-security is directly in the focus of this thesis.
Following the notation Matt Bishop established for IT-security in [Bishop03], security can be seen as a process that is governed by **security requirements** (expressed in the form of **security aspects**), **security policies** (as control mechanisms) and **security measures** or **security mechanisms** (which actually implement security and which therefore can be classified based on the addressed security aspects). Regarding the definition of security aspects, different conflicting models exist (cf. for example [Bishop03] and [Eckert11]). Authenticity and integrity, the small selection of the established security aspects considered in this thesis, nevertheless are conceptually present in most of all these models. Regarding **authenticity**, two distinct concepts exist, entity authenticity and data authenticity [Kiltz08]. The first one, which is outside the scope of this thesis, establishes the genuineness of human entities (e.g. users in a system) and non-human entities in the sense that an entity is really who it claims to be. Mechanisms here are for human entities either single factor authentication using knowledge (e.g. passwords), posses-

sion (e.g. authentication via smart cards) or biometric characteristics (see e.g. Ross et al. [Ross06] and Vielhauer [Vielhauer05]), or multi-factor authentication (combinations of the single factor authentication possibilities). For non-human entities (e.g. processes or devices), authentication measures include for example the usage of (unique) device identifiers (e.g. media access control (MAC) addresses of network interface cards), vendor identifiers and serial numbers for CPUs, etc.). Data authenticity, as the second and for this thesis relevant facet of authenticity, is sometimes also called source authenticity. It is the assessment of the origin, genuineness, originality, truth and realness of (digital) data objects.

Data integrity in computer science and telecommunications refers to the integrity of resources. Integrity requirements describe how the integrity of the system can be ensured (prevention), or it reports if the resource, for example information, has been altered or manipulated (detection) or they enable the system to be recovered into a consistent state (recovery). Integrity is therefore the quality or condition of being whole, complete and unaltered. It also refers to the consistency, accuracy, and correctness of data [Kiltz08].

Regarding the measures and **mechanisms** that implement IT-security, and here especially authenticity and integrity, we have to distinguish between **active mechanisms** (e.g. digital watermarking (see e.g. Cox et al. [Cox08] and Dittmann [Dittmann00]) or steganography (see e.g. Fridrich [Fridrich09])) and **passive mechanisms**. While active mechanisms introduce a priori changes to objects for protection (e.g. to ensure authenticity and/or integrity or to implement a channel for hidden communication), their passive counterparts work without modification of the protected data. The latter include mechanisms that have to be applied prior to an expected incident (like cryptographic hashing, perceptual hashing, encryption, etc.) as well as **forensics**, a concept for security mechanisms that can be applied a posteriori. The forensic sciences can be divided into many distinct sub-categories, such as crime scene forensics and IT-forensics. Only IT-forensics is relevant for this thesis. In this field, a distinction is usually made between the focus on systems (e.g. hard drive or network forensics) and the focus on media objects (**media forensics**).

In this thesis, considerations focus on passive, forensic security mechanisms aiming to establish trust in audio material. The motivation for this media forensics approach is nicely summarised in a statement from Oermann et al. [Oermann05] (which itself is based on a quotation from the Random House Dictionary of the English Language [Flexner87]):

> *"The motivation of our work is determined through its forensic background. The word* 'forensic' *is defined as* 'pertaining to, connected with, or used in courts of law or public discussion and debate.' (The Random House Dictionary of the English Language, Second Edition – Unabridged. Random House, Inc., New York, 1987; [Flexner87]). *Hence, the term forensic audio may be defined as the application of audio knowledge, technologies and methodologies to questions of civil and criminal law or public discussion and debate. Forensics imply the assurance of integrity and authenticity of information."*

The role of forensics in the human society

One of the oldest and most respected forensic sciences is fingerprint analysis. Since being established in the 19th century, it has been used worldwide to solve criminal cases by matching latent fingerprint traces found e.g. at crime scenes to samples acquired from suspects. But even an established practice like fingerprint analysis has to face new challenges once in a while. One extremely influential challenge has been the amendment of the U.S. Federal Rules of Evidence (FRE) in the year 2000 as a result of the two ground-breaking cases **Daubert**[3] and **Kumho Tire**[4]. This amendment to the FRE led during a short time after its instalment to several court cases, where the presiding judge refused to allow the admission of fingerprint analysis results in court (see [SWGFAST11]).

The Scientific Working Group on Friction Ridge Analysis, Study, and Technology (SWGFAST), a highly respected, non-profit expert group for fingerprint analysis acknowledged and supported by the American

[3] *Daubert v. Merrell Dow Pharmaceuticals, Inc.* United States Court 509 U.S. 579, 1993.

[4] *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 119 S. Ct. 1167, 1999.

Society of Crime Laboratory Directors (ASCLD), the International Association for Identification (IAI) and the Federal Bureau of Investigation (FBI), draws the following conclusions for forensic scientists from the Daubert and Kumho Tire cases in [SWGFAST11]:

> "[...] *the definition of science, the scientific method, and scientific evidence can no longer be used as loosely as experts have been doing. It is no longer sufficient to call yourself a forensic scientist in order to be considered a scientist. It is no longer sufficient to say that something is a subject of forensic science in order for a court to agree that it is dealing with science. Simply saying it does not make it so. The courts may, and many will, require the experts to show that they know what the scientific method consists of and provide the scientific basis for their conclusions. By the same token, each discipline will be judged by its own standards and upon its own experience.* [...] *It also means that forensic scientists can no longer expect to rely on the fact that courts have long accepted and admitted evidence of their expert conclusions. The court can relitigate the admissibility of a certain type of expert evidence if a litigant can make a credible argument that there has been no previous scientific inquiry of the validity of the assumptions on which a forensic field has long rested. Decades of judicial precedent no longer preclude reviewing whether existing precedent satisfies Daubert and Kumho Tire."*

This summarising statement from a forensic specialist group trained (and providing training) for giving expert testimony in U.S. courts - notably the busiest judicial system in the world – **defines the requirement for the investigation of the fitness of all forensic methods: potential success in the struggle for admissibility in court**.

Up until today, the consideration of forensic compliance, e.g. to the so called Daubert criteria (see section 2.2), has been grossly neglected in the field of media forensics. The main reason for this fact has to be sought in the current practice of implementing myriads of highly specialised approaches for very narrow application scenarios. This leads to a wide landscape of forensic tools and approaches, most of them promoted by only one person or one research group. Naturally, very few of these approaches will ever see the necessary field penetration and acceptance that would be necessary for passing the hurdle set by the Daubert case arguments (or Daubert criteria).

Some application scenarios (like e.g. image steganalysis) are beginning to see approaches that achieve at least some degree of generalisability, but approaches that span different application scenarios are still amiss in the field of media forensics. Regarding the focus of this thesis, the author considers it necessary to strive for application scenario spanning, generalised forensic approaches. Those could, on the one hand, be easily adapted to fit the needs of specific application scenario. On the other hand, they might more easily reach the widespread acceptance required to be admissible in court. Within this thesis, one such adaptable, general-purpose approach is introduced. Even though the work described here cannot alone achieve the introduction and implementation of a truly universal, Daubert-compliant audio forensic approach, it is intended to lay the foundation for a process aiming at this distant goal. Therefore, it is intended to facilitate the understanding between the IT and the society perspectives of forensics summarised above. The intended benefit of this is that it allows these two to better understand each other: On one hand, it shows IT-security researchers working on the development of forensic methods a precise picture of the compliance requirements installed by the society; on the other hand, it is intended to help non-technicians to understand the challenges that researchers in IT-security face.

Structure of the following sections, composing the rest of chapter 1
The following section 1.1 presents a more detailed view on the topics discussed in this thesis as well as on the associated terminology. With audio steganalysis and microphone forensics it introduces the two specific media forensics application scenarios considered here in order to be solved by the application of statistical pattern recognition (SPR) techniques. Based on this, the addressed research challenges are identified in section 1.2. In section 1.3 the corresponding objectives are derived from the research challenges. Section 1.4 gives a brief summary of major outcomes of the thesis, followed in section 1.5 by the description of the outline of this thesis.

1.1 Brief reflections on the state-of-the-art for audio forensics approaches relevant for this thesis

This section is dedicated to a brief description of the scope of this thesis and the corresponding state-of-the-art in current scientific work, to facilitate the understanding of the research gap addressed as well as the objectives and contributions described in the following sections. More detailed analyses of the state-of-the-art relevant for this thesis are presented in chapter 2.
According to a well-established definition (cf. [Goodman07]) given in [Palmer01], **forensics** is:

> "*The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.*"

This thesis focuses on a special part of the huge field of forensics, the area of **IT-forensics** or, more precisely, on selected topics of the sub-discipline of **media forensics** for audio signals. The **investigation goals** considered in this thesis are special problems of either **integrity verification** by **manipulation detection** or **device/source authentication**.

In this thesis only two application scenarios from the vast field of media forensics - **audio steganalysis** and **microphone forensics** - are considered for integrity (both application scenarios) and authenticity (microphone forensics) investigations.

Audio steganalysis, which has the goal of detecting audio file manipulations by steganography (i.e. steganographic message embedding), can be considered in this field to be a rather well researched and published domain, having gathered a large number of scientific publications since the early 1990s. In general, the field distinguishes between **universal, blind steganalysis** (which assumes that it is possible to construct one detector that is able to successfully perform steganalysis for all steganography approaches possible in one application domain) and **application specific steganalysis** (a.k.a. targeted steganalysis; specialised detection approaches that focus on one specific steganographic technique or tool). Since the boundaries between steganography and **digital watermarking** are blurred (see Cox et al. [Cox08]), steganalysis can also be used to perform statistical analyses on the detectability or security of digital watermarking schemes.
In comparison to audio steganalysis, microphone forensics, aiming primarily at the authentication of the recording source (i.e. microphone) for digital audio signals, is a rather young and immature research field. It is one goal of this thesis to strongly improve the degree of maturity of this field.

Regarding the basic **methodology** applied in the state-of-the-art work in these fields to implement audio forensics performing security mechanisms, it can be stated that most of them are based on **pattern recognition (PR)**. Pattern recognition focuses on enabling machines to distinguish between classes of objects while ignoring any background (or noise) influences. If the different sub-types of PR approaches (template matching, statistical pattern recognition, neuronal networks, etc.; cf. [Duda01]) are considered, for steganalysis the **statistical pattern recognition (SPR)** is currently by far the dominating PR sub-type. The audio steganalysis approach introduced in this thesis also follows this methodology. In contrast, in the start-of-the-art in microphone forensics, all currently existing approaches perform **template matching** (which, in comparison to SPR, uses a simpler method to represent the decision basis for classification – instead of complex statistical modelling, it uses templates, usually a set of candidate elements, to represent the classes to be distinguished). The result of this choice is that all state-of-the-art approaches in this field show severe content limitations in their applicability. Therefore, a new SPR-based microphone forensics approach is established within this thesis to overcome these limitations.

The detailed analysis of the state-of-the-art in both application scenarios (see sections 2.5 and 2.6) shows that most of the research work done in this field concentrates on selected aspects of the pattern recognition pipeline rather than acknowledging the whole process. Summarising this research work, it has to be stated that currently no generalised approaches exist for implementing a wide range of audio forensic applications scenarios. Instead, selected tasks are addressed by specific solutions, which hardly seems to be efficient.

To advance the current state-of-the-art in research, this thesis introduces a general-purpose approach for audio forensics that could easily be adapted to fit the needs of specific application scenarios. It is signal processing based (instead of being based on information theory), using with SPR an established technique for solving decision problems. The intended goal is to offer an alternative to using a myriad of highly specialised tools that are hardly capable of keeping up with the evolution speed of today's information society. Furthermore, such a general-purpose approach might more easily gain the necessary field penetration and acceptance required to pass the hurdle set by the Daubert criteria.

Figure 1.1 shows the introduced general-purpose approach as an abstract descriptor of a forensic audio authenticity or integrity verification mechanism.

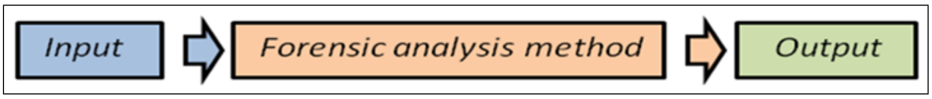


Figure 1.1: Abstract descriptor of a forensic audio authenticity or integrity verification mechanism

The core part of this general-purpose approach – the forensic analysis method – is implemented in this thesis by a **statistical pattern recognition (SPR) pipeline**. It consists of signal processing and pattern recognition components which operate in two different operational modes: training (or model generation) and application (or testing). If the SPR pipeline is split into functional entities, the following main components emerge: **pre-processing** (increasing the distance between the pattern and the background), **feature extraction** (information reduction by projection of the signal onto a pre-defined feature space) and **training** or **classification** (the generation of classifier models or the application of these models to assign a class to the pattern within the signal). In contrast to other work (e.g. [Vielhauer05]), the **signal acquisition** is not considered here to be an integral part of the actual pattern recognition pipeline. Instead, it is one component in a more complex **signal preparation** block (implementing the blue box labelled 'input' in figure 1.1) that feeds into the pattern recognition process. This is necessary due to the performed plausibility considerations (see below), which consider influences to the signal generation process as well as the persistence of the patterns under selected post-processing operations. The **evaluation** (performed on the output of SPR-driven forensic analyses) encompasses in this thesis considerations on:

- The **detection performance** of the implemented SPR processes,
- The **plausibility** (incl. the sensitivity of the implemented processes against malicious or non-malicious attacks)
- The **forensic conformity** (within this thesis considered to be equivalent to the Daubert standard, the requirements for forensic compliance in U.S. federal level judicial matters)

Classification accuracy is the most commonly used metric in data mining and pattern recognition to characterise the detection performance in PR problems. It describes the performance of recognition systems such as classifiers by computing in supervised applications the ratio between true decisions and the number of overall decisions in closed-set evaluations.

Regarding the issue of **plausibility testing**, the results generated by the application of machine learning strategies lack the intuitive verification that usually accompanies human decision-making processes. Therefore, additional effort has to be invested after the application of strategies like SPR to establish whether the results of the automated process are reasonable. In practice, this means to establish that the patterns trained and detected are really the patterns that the user wants to distinguish between and

that side-effects as well as external influence factors are known for the pattern recognition process[5]. Therefore, the plausibility of the application of PR mechanisms in a forensic application scenario is directly linked to the trust we place into the decisions (and training basis) of corresponding PR-based security mechanisms.
Furthermore, if such a forensic security mechanism sees field application, it will have to deal with countermeasures – so called attacks or **anti-forensics**. Therefore, the author considers it important to include selected countermeasures or attacks into the plausibility considerations within this thesis.

Apart from these considerations on efficiency (i.e. detection performance and plausibility), all forensic methods should aim at fulfilling some form of forensic conformity. The requirements considered within this thesis for forensic methods are imposed by the Daubert standard, a set of rules defined in the 1990s by the United States of America Supreme Court based on [USC93]/[USCA95]. This standard is regarding the trust in, or the admissibility of, expert witnesses' testimony as evidence in legal proceedings (for detailed explanations see e.g. the explanations on the Federal Rules of Evidence (FRE) rule 702 in [LLI11]). Although the Daubert standard is only mandatory for United States of Americas federal legal proceedings, it is widely regarded as a good (if not the best established) set of recommendations for judges on how to evaluate the usefulness of scientific (as well as non-scientific) expert testimony (see e.g. [HC-STC05]). A more detailed analysis of the Daubert standard and its relevance for this thesis is given in section 2.2.

1.2 The research challenges addressed in this thesis

The **research challenges** addressed in this thesis can be summarised in the following three questions:

(a) Is there a generalised statistical pattern recognition (SPR) approach (i.e. an adaptable solution concept) for media forensics that can be used for addressing multiple audio forensics investigation goals?

(b) How can the usefulness of application scenario specific instantiations of a generalised audio forensics approach be measured?

(c) Can instantiations of the generalised audio forensics approach be used to adequately implement the application scenarios of audio steganalysis and microphone forensics?

Although there exists a huge variety of implementations of classification algorithms (as the core part of any SPR implementation), no universal approach can currently be found in the state-of-the-art for forensic analyses of audio signals. All existing approaches are closely modelled on very specific investigation goals. This means that in this field no investigations have yet been performed, whether it is possible to achieve with generalised approaches a detection performance equivalent to application scenario specific approaches.

This gap in the current state-of-the-art is addressed in this thesis, because the introduction of generalised approaches – universal tools that can be easily adapted to solve specific problems (like audio steganalysis or microphone forensics) – would be an important step towards efficiency (in terms of forensic compliance) in audio media forensics.
For building a generalised statistical pattern recognition (SPR) based mechanism for audio forensics, the

[5] A classifier tends to learn the simplest pattern described by the features extracted from the set of signals under observation. A rather renowned story in the data mining community to illustrate this fact tells of scientists in a military project trying to train a neural network to classify images as containing either tanks or trees. The story is summarised in [Bersano-Begey97] as follows: scientists present pictures of trees and pictures of tanks to the neural network to train it. After sophisticated pre-processing of the images, these are fed in a neural network and, after considerable training, the network is able to classify each image correctly. However, when it is tested on other images, the network seems to classify every image as trees, even when it contains a tank. After careful study the scientists finally resolve the mystery: in all the images used in the training, those containing trees were always taken in broad daylight, while those containing tanks were always taken in a darker setting! Thus, the network had learned to distinguish the (trivial matter of) differences in overall light intensity rather than recognising the presence of tanks.
Sometimes, this possibly apocryphal story is told claiming to aim at the distinction between American and Russian tanks.

following main open research issues are addressed in this thesis: definition of a useful general-purpose (i.e. not specifically designed for one application but suitable to cover a number of different usage scenarios) audio feature set, investigation of methods for classifier selection, evaluation set designs for plausibility testing and forensic conformity considerations.

1.3 The research objectives for this thesis

For the considerations in thesis, it is necessary to extend the detail of the considerations on the abstract descriptor of an audio authenticity or integrity verification mechanism, as it is described above in figure 1.1. Its three main methodical and conceptual components (or 'building blocks') for the required investigations reflect the research challenges specified in the previous section, as shown in figure 1.2. Ordered according to the corresponding research challenge, these 'building blocks' are: The **forensic analysis method**, represented by the **statistical pattern recognition (SPR) pipeline** (research challenge (a)). The **output**, or more precisely, the corresponding **evaluation methodology and concepts** for generating interpretable output, is the concern of research challenge (b). The **input**, i.e. the **signal preparation**, is in this thesis a synonym for the considered application scenario for forensic audio signal authenticity and integrity verification. Together with the corresponding evaluation considerations, it addresses research challenge (c).

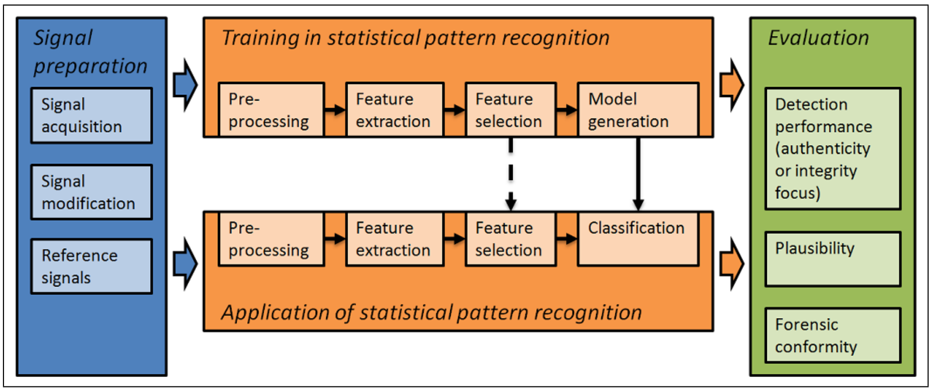


Figure 1.2: The 'building blocks' or major methodical and conceptual components for the investigations performed in this thesis

The goals of this thesis are to perform investigations that allow answering the three research challenges specified in the previous section. These challenges are, in reflectance of the 'building blocks' shown in figure 1.2, translated into the following, more specific research objectives: Analyses on whether a generalised statistical pattern recognition (SPR) approach for forensic audio signal analysis is possible (research objective 1), which performance indicators are useful (research objective 2) and how the generalised SPR approach can be used to implement selected specific application scenarios (here, audio steganalysis and microphone forensics; resulting in research objectives 3 and 4). These **research objectives** can be more closely specified as:

- **Research objective 1:** In order to introduce a generalised approach, the objective is to analyse the state-of-the-art in both selected application scenarios (audio steganalysis and microphone forensics) and to discuss the existing alternative solution principles. The important questions are: Which existing methods and concepts can be used for the intended approach? Which methodological and conceptual deviations have to be made in this thesis from the paradigms currently used in the state-of-the-art?
  It is assumed here that the necessary deviations from the established paradigms are less severe in the case of audio steganalysis (where SPR is already applied as the predominant principle) than for the much younger research field of microphone forensics (which is currently mostly addressed by template matching approaches). This assumption has to be verified.
  The research effort to answer the questions imposed by this research objective has to include

considerations on a new high-dimensional, simple to compute, general purpose audio feature set as well as considerations on concepts for feature and classifier selection. Considerations on conceptual components of the SPR process which only aim at the performance enhancement in already established processes (e.g. the detection performance enhancing effect of pre-processing or the design of application-specific classification algorithms) are considered to be outside the direct scope of this thesis and are therefore reserved for future research. Also, a complete (mathematical) formalisation of the process is reserved for further work.
The investigations on this objective contribute to the answer for research challenge (a) 'Existence of a generalised SPR approach for audio forensics'.

- **Research objective 2:** The alternatives presented as main performance indicators in the state-of-the-art (ranging from classification accuracy to correlation-based schemes for template matching approaches) for both application scenarios are analysed here. The main question is: Whether the main performance indicator currently used for describing the detection performance in both application scenarios (the classification accuracy) is suitable for comparable performance measurements? In case the answer to this question is negative, alternatives that allow for a direct comparison between different classification-based solutions have to be discussed.
  A further question regarding the performance of SPR-driven audio forensics mechanisms is: Whether further performance indicators apart from the detection performance are required for the evaluation of such schemes?
  The investigations on this objective contribute to the answer for research challenge (b) 'Applicable performance measures'.

- **Research objective 3:** The following two audio forensic application scenarios are considered within this thesis:

  - The well established application scenario of audio steganalysis – in order to validate the practical performance of the resulting application scenario specific instantiation of the introduced general-purpose approach
  - The sparsely researched application scenario of microphone forensics – in order to validate the practical performance of the resulting application scenario specific instantiation and to see whether this solution can overcome the restrictions of the current template matching based state-of-the-art in this field

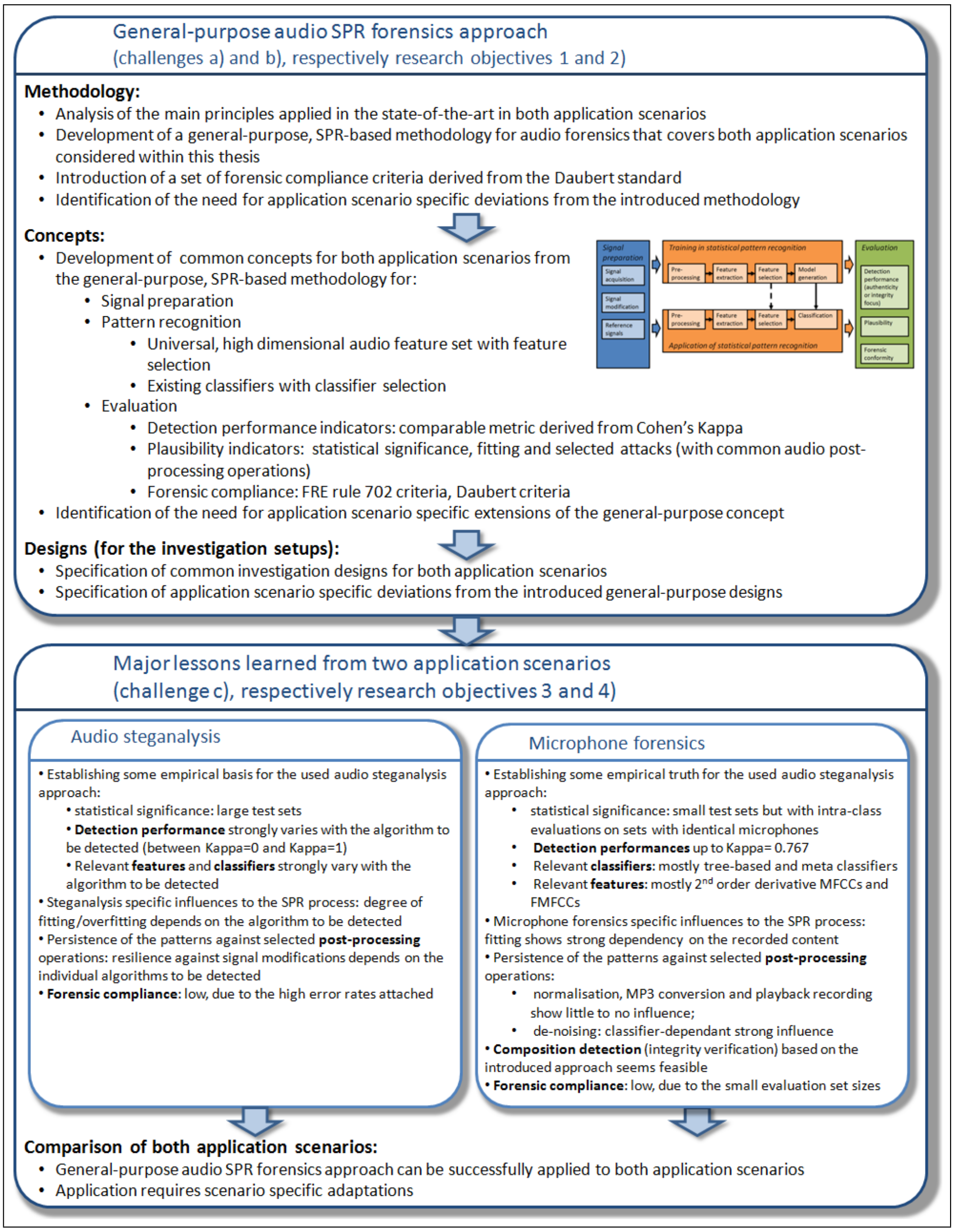
  When applying the general-purpose audio forensics SPR approach introduced in this thesis to these application scenarios, the following question has to be answered: How do suitable investigation setups have to be designed?
  The practically achieved results of the application scenario specific adaptations have to be compared against the performance of the corresponding state-of-the-art approaches.
  For audio steganalysis, the investigations include (see section 4.3) the following application scenario specific intrinsic influences: Number of feature vectors in training, balancing of error rates in a two-class setup, suitable classifiers (from a preexisting portfolio provided by WEKA, suitable features, content selection as well as content dependent and independent training and testing, and two-class vs. multi-class setups. Investigations on influences outside the SPR process for audio steganalysis are: Embedding domain and algorithm identification, influence of the key scenario in steganography and selected common audio post-processing operations (MP3 conversion and de-noising).
  For microphone forensics, the investigations include (see section 4.3) the following application scenario specific intrinsic influences: Number of feature vectors in training, suitable classifiers (from a preexisting portfolio provided by WEKA), suitable audio features and the influences of using content selection as well as content dependent and independent training and testing influences the detection performance in microphone forensics. Investigations on influences outside the SPR process for microphone forensics are: The recording environment, microphone orientation, mounting of the microphone, selected common audio post-processing operations (normalisation, MP3 conversion and de-noising), as well as playback recording and composition attacks.

All investigation results on this objective contribute to the answer for research challenge (c) ‘Adequate implementation of mechanisms for the chosen application scenarios’.

- **Research objective 4:** The goal behind the comparison of the two application scenario specific instantiations is to show the prospects and current limitations of the introduced general-purpose audio SPR forensics approach. The question behind this objective is: How large are steps required to adapt the general-purpose approach to a specific application scenario, like the two exemplary chosen for this thesis? The investigation results produced here are a further contribution to the answer for research challenge (c) ‘Adequate implementation of mechanisms for the chosen application scenarios’.

1.4 Summary of the main contributions of the thesis

The **main thesis contributions** to the addressed research challenges are summarised in figure 1.3.



General-purpose audio SPR forensics approach
(challenges a) and b), respectively research objectives 1 and 2)

Methodology:
- Analysis of the main principles applied in the state-of-the-art in both application scenarios
- Development of a general-purpose, SPR-based methodology for audio forensics that covers both application scenarios considered within this thesis
- Introduction of a set of forensic compliance criteria derived from the Daubert standard
- Identification of the need for application scenario specific deviations from the introduced methodology

Concepts:
- Development of common concepts for both application scenarios from the general-purpose, SPR-based methodology for:
  - Signal preparation
  - Pattern recognition
    - Universal, high dimensional audio feature set with feature selection
    - Existing classifiers with classifier selection
  - Evaluation
    - Detection performance indicators: comparable metric derived from Cohen’s Kappa
    - Plausibility indicators: statistical significance, fitting and selected attacks (with common audio post-processing operations)
    - Forensic compliance: FRE rule 702 criteria, Daubert criteria
- Identification of the need for application scenario specific extensions of the general-purpose concept

Designs (for the investigation setups):
- Specification of common investigation designs for both application scenarios
- Specification of application scenario specific deviations from the introduced general-purpose designs

Major lessons learned from two application scenarios
(challenge c), respectively research objectives 3 and 4)

Audio steganalysis

- Establishing some empirical basis for the used audio steganalysis approach:
  - statistical significance: large test sets
  - **Detection performance** strongly varies with the algorithm to be detected (between Kappa=0 and Kappa=1)
  - Relevant **features** and **classifiers** strongly vary with the algorithm to be detected
- Steganalysis specific influences to the SPR process: degree of fitting/overfitting depends on the algorithm to be detected
- Persistence of the patterns against selected **post-processing** operations: resilience against signal modifications depends on the individual algorithms to be detected
- **Forensic compliance**: low, due to the high error rates attached

Microphone forensics

- Establishing some empirical truth for the used audio steganalysis approach:
  - statistical significance: small test sets but with intra-class evaluations on sets with identical microphones
  - **Detection performance** up to Kappa= 0.767
  - Relevant **classifiers:** mostly tree-based and meta classifiers
  - Relevant **features:** mostly 2nd order derivative MFCCs and FMFCCs
- Microphone forensics specific influences to the SPR process: fitting shows strong dependency on the recorded content
- Persistence of the patterns against selected **post-processing** operations:
  - normalisation, MP3 conversion and playback recording show little to no influence;
  - de-noising: classifier-dependant strong influence
- **Composition detection** (integrity verification) based on the introduced approach seems feasible
- **Forensic compliance:** low, due to the small evaluation set sizes

Comparison of both application scenarios:
- General-purpose audio SPR forensics approach can be successfully applied to both application scenarios
- Application requires scenario specific adaptations

Figure 1.3: Overview of the main contributions of this thesis

As a first, necessary contribution a common methodology and solution concepts for SPR-based audio forensics are described. The methodological work on this step includes a thorough analysis of the state-of-the-art in the two selected application scenarios of audio steganalysis and microphone forensics. Based on this analysis, a general-purpose SPR-based methodology is introduced.
From the general-purpose methodology, solution concepts are derived. The concepts fit the three major 'building blocks' of signal preparation, the SPR pipeline and the evaluation. The main parts of the research work on these major 'building blocks' are focused on the second and third component, the SPR pipeline and the evaluation.

Regarding the **introduction of a general purpose audio statistical pattern recognition (SPR) forensics approach** for multiple audio forensics application scenarios, this thesis focuses on two main concepts: a universal, high-dimensional audio feature set (here with 590 segmental features computed for frames of the audio signal) and a classifier selection approach for the variety of existing classification methods.
The high-dimensional feature set proposed and applied here contains time-, frequency- and cepstral-domain features capable of handling the patterns used in different audio forensics analyses (as shown for audio steganalysis and microphone forensics within this thesis). Nevertheless, the thesis investigations show feature selection concepts are required for the adaption to the application scenario.
The existence of large numbers of different classifiers (like in the WEKA data mining environment [Hall09] used extensively within this thesis) as well as the thesis results which show that the choice of a wrong classifier leads to low detection performances, result in the proposal of concepts for application scenario specific classifier selection.

For the **evaluation**, the classification accuracy, as the major **performance indicator** used in the state-of-the-art in this research field, is replaced by other, more suitable performance indicators. The reason for this step is that the accuracy does not allow for a direct comparison between different classification problems or solutions. Within this thesis it is substituted by simplified Kappa statistics[6]. This metric, which is used in the WEKA [Hall09] implementation, is derived from Cohen's Kappa (cf. [Cohen60], [Carletta96]). So far, the usage of this detection performance indicator is uncommon in the research field of media forensics. Its introduction to the considered research field is deemed necessary in this thesis because it facilitates direct comparison between different classifiers and classification problems.
In addition to the metric for detection performance, **plausibility indicators** are proposed here. These indicators include statements on: statistical significance, fitting (consideration of different influences in the signal generation process) as well as deceivability (persistence of the patterns under selected attacks like audio signal post-processing operations or anti-forensics).
Furthermore, considerations on the **forensic conformity** of audio forensic methods under are presented. In this thesis, the requirements for forensic conformity or compliance are considered to be imposed to forensic methods by the assumedly most active legal system currently existing – the U.S. legal system. This assumption on the necessity of compliance is a rather novel approach for the two considered application scenarios (from the current state-of-the-art in research on this field, only the electric network frequency (ENF) approach [Grigoras03] for microphone forensics developed by Catalin Grigoras makes in [Grigoras09] some steps in this direction) as well as the whole field of academic research on media forensics. The thesis contributions in this field include the proposal of a scheme for the discussion of the forensic compliance of audio forensic methods. Unfortunately, this scheme cannot be turned into a metric for performance measurement or estimation, since the foundation of these considerations (the Daubert standard) is currently not codified, i.e. it is a set of verbally described requirements which leave room for the judicial interpretation required in any specific court case.

[6] In short, the used Kappa statistic κ measures the agreement of prediction with the true class (i.e. the agreement normalised for chance agreement) in the range $[-1, 1]$. A value of $\kappa = 1$ indicates perfect agreement and $\kappa = 0$ indicates chance agreement for the overall classification. Negative values for κ imply the choice of a classifier model trained for a different classification problem. For a complete description of the used Kappa statistics see section 4.1.4

Since both application scenarios considered are strongly different, the considerations are accompanied by an identification of the need for application scenario specific deviations from the introduced general-purpose methodology and concepts.

For the **instantiation of the general purpose audio SPR approach for audio steganalysis**, the thesis shows:

- The instantiation of the generalised SPR process for steganalysis follows within this thesis the general trend in this research field. The practical results for steganalytical detection (based on the introduced detection performance indicator) strongly vary between the different data hiding algorithms used in the steganalysis evaluations. Results for the Kappa statistics between $\kappa = 0$ and $\kappa = 1$ show that there are algorithms in the set under evaluation which result in a very high detection performance for the used audio features, while other algorithms are not (or only barely) detectable with the same setup. This means that the capability of detecting the usage of data hiding algorithms by application of pattern recognition techniques is strongly dependant on the strength of the pattern imposed by the embedding function of the algorithm under investigation.
- The plausibility results show, that the SPR-based audio steganalysis seems to be negatively influenced by other audio signal processing operations. Therefore, if its application as a specialised integrity verification mechanism is considered, the implementation of the mechanism should undergo extensive plausibility evaluations against other audio signal modifications (encoding, re-sampling, etc.) that are likely in the considered application field. Furthermore, if specific counter-forensics methods have to be expected, these should also be integrated into the evaluation process.
- The research results for detection performance and plausibility imply in their combination that, for application specific steganalysis, a reliable detector for selected algorithms might be implementable (if the embedding process generates a statistically significant footprint as a pattern in the computed audio features), but that this result cannot be achieved in general. There will always be some information hiding (IH) algorithms where no reliable detection is possible under any constraints or where counter-forensics can be successfully applied to obliterate the embedding pattern.
- The implications of the Daubert-compliance driven evaluation lead to the realisation that the current practice in steganalysis is merely aiming at detection of steganographic traffic. It is therefore only the first stage of a two-phase process that would be required under Daubert considerations to aim for admissibility in court. The second phase would be the verified detection necessary to bind the steganalysis result to a law case (by clarifying who sent which message to whom), which is in this thesis described as the 'ideal steganalysis process'. Addressing this second phase is outside the scope of this thesis, because it would require a reliable detection, which cannot be guaranteed with the rates of detection errors presented here.

In respect to the **instantiation of the general purpose audio SPR approach for microphone forensics**, the investigations show:

- In microphone forensics, the need for the newly introduced approach is explained by the strong context dependency of the current, mainly template matching driven approaches in this rather young and so far sparsely researched application scenario. Here, the instantiation of the generalised SPR approach for microphone forensics proposed within this thesis is considered a method to overcome this context dependency. The potential implication of this contribution for a forensic investigation is a larger number of recorded audio signals that might successfully undergo source authentication and integrity verification, i.e. it will also work even if no ENF traces, clear reverberation-patterns or other requirements of the alternative approaches in this field are met.
- The results achieved show a very good authentication performance in small scale evaluations (results with values for the introduced detection performance indicator Kappa statistics close to $\kappa = 1$ for ideal setups with different microphones and ideal content; significant (using a Kappa to

statistical confidence mapping based on the work of Landis and Koch [Landis77]) results with κ up to 0.767 for plausible setups with sets of identical microphones and a wide range of recorded content).

- The considerations on the plausibility indicators for the investigations performed show a high degree of resilience of the introduced approach regarding common audio signal post-processing operations (like MP3 conversion or normalisation). Traces of the recording patterns imposed by the original microphone are shown to survive to some extend a playback recording procedure.

- For the integrity verification results, the introduced approach also shows its potential for the detection of compositions of audio material recorded with different microphones. Based on the seminal results on this matter presented here, future work will have to establish how reliable means for audio signal integrity verification can be achieved using the SPR-based approach proposed in this thesis in combination with the state-of-the-art in this field (especially the ENF-based work from [Rodríguez10]).

- The implications of the Daubert-compliance driven evaluation show the current work is still a long way from being fit for court. It has to be admitted that the size of the experiments performed might still lack generalisability, but the detection performances achieved in evaluations on sets of identical microphones are very promising.

The thesis contributions regarding the **comparison of the two application scenario specific instantiations** show that these two application scenarios, despite being completely different in their application goals, share strong similarities when it comes to the possible solution approaches. Both can be projected onto a detection or pattern recognition problem and can therefore be solved in similar ways.
The investigations performed in both application scenarios follow a sound evaluation methodology. Regardless of the theoretical merits of the introduced evaluation methodology, the performed investigations have to face severe **limitations** in practice. For audio steganalysis, where the number of freely available embedding algorithms is rather small, the evaluations within this thesis are improved in their significance by including digital audio watermarking algorithms in the evaluation set, boosting the number of algorithms under evaluation from 5 to 9. For microphone forensics, the evaluations are limited in terms of set sizes and compositions by the quantity and quality of the lab equipment available. For both application scenarios, the results achieved within this thesis allow for some generalisation, but future research will have to find suitable approaches to establish a wider test base using the evaluation methodology proposed here. That such an extension of the test base is already possible in selected forensic application scenarios is demonstrated in [Goljan09]. The benefit of this extension would be the possibility to present answers for Daubert hearings that carry the required statistical significance, as shown impressively for Jessica Fridrichs PRNU-based digital camera forensics approach in the law case *United States of America v. Nathan Allen Railey*[7], where this image forensics approach successfully passed the Daubert hearing.

It has to be mentioned here that the applicability of the proposed general purpose audio statistical pattern recognition (SPR) approach is not limited to the two chosen application scenarios. In fact, it could be transferred to further forensic research topics which share the same essential characteristics, like e.g. voice recognition, speech recognition, speaker recognition, audio coder verification, gunshot characterisation, audio signal quality verification, etc. Future work will have to investigate the effort required for the adaptation to these additional application scenarios.
Apart from these contributions to the specific research objectives formulated, general and specific open questions and challenges are summarised in chapter 8, outlining future work in this field. For the specific open questions regarding classifier benchmarking and information fusion[8], first solution ideas that have been developed in the context of this thesis are already presented in section 8.2.

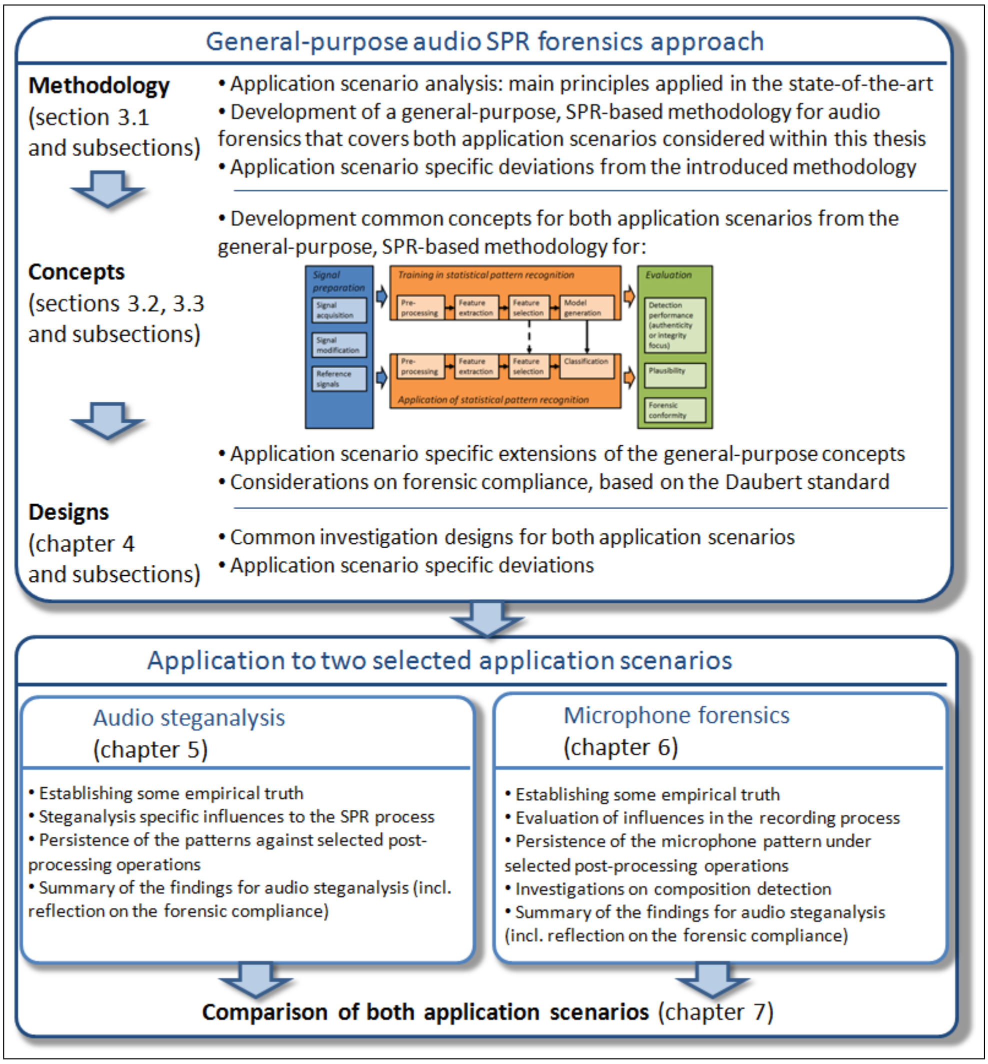
[7] United States District Court for the Southern District of Alabama, August 2nd, 2011 – For a short summary of the relevant part of the proceedings see: http://blog.al.com/live/2011/07/expert_witnesses_link_camera_t.html

[8] Within the context of this thesis, information fusion is considered to address the question of how to combine different decision or expert systems. For more details on fusion see section 8.2.2

1.5 Thesis outline

The complete thesis consists of 8 chapters and 3 appendices. It is structured as follows:

Chapter 2 presents the required fundamentals in and for audio steganography, audio steganalysis and microphone forensics. If the 'building blocks' metaphor and the application scenario specific forensic investigation introduced in section 1.3 are re-used here, then the basics for all these building-blocks and application scenarios are addressed within this chapter. To facilitate understanding, chapter 2 begins in sections 2.1 and 2.2 with descriptions on the application goals of forensics and the evaluation of forensic methods regarding their fitness in (U.S.) legal proceedings (i.e. the compliance to the Daubert standard). Section 2.3 presents the audio signal and signal processing basics that are required for the understanding of the application scenarios, while section 2.4 gives brief summaries on pattern recognition in general and statistical pattern recognition as its sub-category to be applied within this thesis. Sections 2.5 and 2.6 give comprehensive summaries of the state-of-the-art in both considered application scenarios (audio steganalysis and microphone forensics).



General-purpose audio SPR forensics approach

Methodology (section 3.1 and subsections)

- Application scenario analysis: main principles applied in the state-of-the-art
- Development of a general-purpose, SPR-based methodology for audio forensics that covers both application scenarios considered within this thesis
- Application scenario specific deviations from the introduced methodology

Concepts (sections 3.2, 3.3 and subsections)

- Development common concepts for both application scenarios from the general-purpose, SPR-based methodology for:

- Application scenario specific extensions of the general-purpose concepts
- Considerations on forensic compliance, based on the Daubert standard

Designs (chapter 4 and subsections)

- Common investigation designs for both application scenarios
- Application scenario specific deviations

Application to two selected application scenarios

Audio steganalysis (chapter 5)

- Establishing some empirical truth
- Steganalysis specific influences to the SPR process
- Persistence of the patterns against selected post-processing operations
- Summary of the findings for audio steganalysis (incl. reflection on the forensic compliance)

Microphone forensics (chapter 6)

- Establishing some empirical truth
- Evaluation of influences in the recording process
- Persistence of the microphone pattern under selected post-processing operations
- Investigations on composition detection
- Summary of the findings for audio steganalysis (incl. reflection on the forensic compliance)

Comparison of both application scenarios (chapter 7)

Figure 1.4: Outline of the thesis

Figure 1.4 displays the outline of the chapters 3 to 7. As shown in this figure, chapters 3 and 4 contain the introduction of the general-purpose audio statistical pattern recognition (SPR) approach and the designs for its implementation within this thesis. Its application to two selected application scenarios and their comparison are addressed in chapters 5, 6 and 7.

Chapter 3 develops the methodology and concepts for the investigations performed within this thesis. In sections 3.1.1 and 3.1.2, the methodology considerations for both selected application scenarios are introduced and compared with the corresponding principles presented in the state-of-the-art (sections 2.5 and 2.6 respectively). On this basis, section 3.1.3 performs a comparison between those two application scenario specific solutions and derives the methodology for the general-purpose approach introduced here to address research challenge (a) 'Existence of a generalised (SPR) approach for audio forensics' as well as research objective 1.
Section 3.2 projects the methodology considerations into investigation concepts. Here, first a set of common concepts is formulated in section 3.2.1, followed by descriptions on application scenario specific concept extensions in sections 3.2.2 and 3.2.3.
Section 3.3 introduces the concept applied in this thesis to use the Daubert standard as general means of results discussion for forensic methods. This is done for two reasons: one hand it leads to the definition of **investigation tasks for the practical investigations on both application scenarios**, on the other hand it contributes to research challenge (b) 'Applicable performance measures' and the corresponding research objective 2, by specifying requirements for such performance measures.
The section 3.4 performs a restriction of the considerations within this thesis to narrow down the focus of the practical investigations.

In **chapter 4** the experiments performed within this thesis are outlined. This chapter is divided into common design decisions for both application scenarios (section 4.1) and application scenario specific adaptations to the common designs (sections 4.2 and 4.3). Within the descriptions of the common design criteria, the two core components of the prototype used for the practical investigations, the feature extractor and the classification mechanisms are discussed.

The **chapters 5 and 6** contain summaries of the results achieved while applying the introduced audio forensics approach to audio steganalysis (chapter 5) and microphone forensics (chapter 6). Those research results address research objective 3, as defined in section 1.3.

Chapter 7 performs a comparison of the performances achieved for both application scenarios and compares these with the performance achieved by state-of-the-art approaches. This is done to show that two exemplary chosen instantiations of the introduced general-purpose approach are performing adequately. Thereby, research challenge (c) 'Adequateness of the introduced approach' as well as research objective 4 are addressed.

The **chapter 8** performs a brief summary of the work presented and draws the necessary conclusions in regard to the defined research challenges and objectives. Furthermore, it contains descriptions of ongoing and future research that is outside the focus of this thesis but builds on the foundations laid here.

The **appendices** contain details about the audio features used in within this thesis (appendix A, starting on page 189), the experimental setups for the audio steganalysis application scenario (appendix B, starting on page 197) and the experimental setups for the microphone forensics application scenario (appendix C, starting on page 201).

2

Thesis Fundamentals and Summary of Related Work

In this chapter the fundamentals required for the development of the methodology, concepts and designs in chapters 3 and 4 are discussed.

The chapter begins in section 2.1 with an introduction on forensics and media forensics. This is followed in section 2.2 by an analysis of the most prominent requirement for forensic methods – the admissibility in court – which is in this thesis considered to be synonymous with the compliance to the Daubert standard as established within the U.S. Federal Rules of Evidence (FRE) rule 702. These judicial considerations are used in the following chapters as one component of the set of performance criteria developed for the investigations.

Section 2.3 summarises required basics on audio signals, while section 2.4 recapitulates the required basics on the statistical pattern recognition (SPR) process pipeline.

The analyses on the current state-of-the-art in the two chosen application scenarios audio steganalysis and microphone forensics are performed in sections 2.5 and 2.6 respectively.

2.1 Forensics in the context of this thesis

This section provides an extension to the rather brief introduction of forensics in general and media forensics as presented in section 1.1 above.

In [Palmer01] **IT-forensics** or digital forensic science is defined as being:

> *"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations."*

In the scope of this thesis this definition is extended to encompass also analogue sources. Usually, the focus in forensic investigations is set to legal proceedings (or "*relating to courts of law*" – the Oxford Dictionary, Oxford University Press, 2012). Nevertheless, recent works, like [Rekhis07] or [Kiltz09], also focus on the usage of forensics in the general investigation of security related incidents. Kiltz et al. [Kiltz09] extend the scope of forensics in their considerations by including besides maliciously caused incidents also hardware failure or software errors. The original considerations in the forensic model by Kiltz et al. only include computer and network forensics, but they can also be applied for **media forensics**, as done within this thesis. Their forensic model (as described in [Kiltz09]) consists of three main aspects: a phase driven description of the forensic process, the classification of forensic methods into classes and a classification scheme for forensically relevant data types. Using their terminology, the statistical pattern recognition (SPR) based forensic mechanisms considered within this thesis would fall into the category of explicit means of (intrusion) detection and would have to be deployed in the phase of strategic preparation of the forensic process to facilitate **online investigations** by an passive security mechanism, as they are considered here, or in the data investigation phase for post-mortem or **a posterior analyses**. The considered data types would be raw (audio) data as input and hardware data (for the application example of microphone forensics) or details about data as output data of the mechanisms.

The **investigation goals in media forensics** are usually either **device/source authentication** or **integrity verification** by **manipulation detection**.

Before digital audio signals mostly superseded **analogue audio signals** as the normal output of recording pipelines in the late 1990s, forensic audiotape analysis (see e.g. [Bolt74]) was used to establish whether a recording on a tape is an original or a copy and to find out whether a recording was made on a given recorder (i.e. to authenticate the recording source). When electronic equipment is used to record an audio signal on tape, it captures not only the intended signal but also the idiosyncrasies and characteristics of the recorder itself [Grigoras03]. Tape recorders left start, stop and pause signatures coming from record and erase heads. Koenig [Koenig90], Pellicano [Pellicano90], Dean [Dean91], Molero [Molero01] and others have shown that by combining waveform, spectral analysis and magnetic patterns analysis, the forensic audio examiner was in most cases able establish the originality of a analogue recording and/or authenticate the recording.

But in the digital world, with digitally recorded audio signals, no recorder idiosyncrasies and characteristics are left behind, but instead other phenomena become usable for the authentication of recorded signals. For the forensic research on digital audio (which if the selected media for this thesis) **authentication** found in literature pursuit different investigations, such as:

- Speech recognition (human speech as well as specific content) as well as forensic speaker recognition [Neustein11]
- MP3 encoder detection (Böhme and Westfeld [Böhme04]): aims for the identification MP3 encoders by artefacts left in the encoded data stream
- Quality estimation for MP3 files (e.g. [Yang09], [D'Alessandro09])
- Double compression of audio signals (e.g. [Yang10]): were a detection of double compressed speech signals is performed based on compression artefacts
- Gunshot analysis (e.g. [Maher06], [Maher07])
- Microphone forensics (or rather recording setup authentication)

Within this thesis, from this list only the application scenario of microphone forensics is considered for the performed concept, design and empirical investigation work on source authentication. The trust placed in sensor data is a necessary condition for most sensor-based systems. This trust is in many cases based on a trust assumption on the source. With microphone forensics (as a specialised form of sensor forensics) a passive security mechanism is considered here that evaluates a posterior the source authenticity of an analogue source for audio data.

Regarding the potential for audio **integrity verification** by manipulation detection, examples for different audio forensics investigations that can be identified today are:

- Deletion detection for segments in MP3 files ([Yang08], [Yang12]): In their papers the authors introduce a format conversion dependent method for locating forgeries (insertions and deletions) in MP3 files by time-domain based analyses of encoder frame offsets.
- Steganalysis (also known as stegoanalysis) is the counter-science to steganography, like cryptanalysis is to cryptography. Steganography, as the art or science of hidden communication, is a historical concept transferred from an ancient Greek origin (the transcription from ancient Greek: *steganos graphein* does in English mean: covered writing; see [Wölfel11] or [Fridrich09]) into today's digital world. While it was originally ([Wilkins41]) considered as a preferable alternative to communication encryption, modern business and warfare rely on cryptographic means instead of steganography to ensure confidentiality. Communication hidden by steganographic means is nowadays considered relevant mostly in application scenarios that originate from espionage or terrorism. Considering this, together with the large number of existing steganography tools for

various digital data formats (see [Fridrich09]), a strong need can be identified to implement steganalysis tools as security mechanisms that detect the usage of steganography in observed environments.

- Audio composition detection: Using the authentication approach of microphone forensics to determine whether an audio signal shows the characteristics of only one or multiple sources – in the latter case a composition attack has to be assumed [Kraetzer11].

Within this thesis, only the last two application scenarios (audio steganalysis and microphone forensics as a method for audio composition detection) are considered for the concept, design and empirical investigation work on audio integrity investigations.
It has to mentioned that the consideration of steganalysis as an integrity verification or manipulation detection problem is rather non-intuitive since the counter-science to a confidentiality addressing security mechanism is considered, but here the focus is narrowed down to the problem of detecting steganography (or data hiding in general) that has been performed by cover modification instead of cover synthesis or cover selection. Therefore, the problem to be solved can be reformulated as: Does an audio data object originate from a 'natural' source or was it maliciously modified by steganographic means?

2.2 The Daubert standard and its relevance for this thesis

This section introduces one of the major evaluation criteria for forensic methods: the **Daubert standard** – or, to be more precise: the requirements of the Federal Rules of Evidence (FRE) rule 702 and the (so called) Daubert criteria. The U.S. law case *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786 (1993) [USC93] / [USCA95] or short *'Daubert'*, has encouraged in U.S. federal jurisdiction a trend toward greater judicial scrutiny of scientific evidence. This is of importance for this thesis because Daubert has also to be considered for forensic evidence. The Daubert standard consists of a set of rules for the admission of expert testimony (i.e. the interpretation of evidence by experts) defined in 1995 by the United States of America Supreme Court. Compliance to this standard is considered within this thesis as the target for every forensic approach, including the two exemplary methods for media forensics considered here.

In sections 2.5 and 2.6 the so called Daubert criteria, introduced in the following subsections, are used to assess the state-of-the-art in the two chosen application scenarios. In chapter 3 these criteria are integrated into the evaluation methodology for this thesis and in chapter 7 they are applied in performance assessment of the developed solutions for the two application scenarios and for comparison to the performance of the alternative approaches summarised in the state-of-the-art.

2.2.1 Rules governing the interpretation of forensic results in court

The usage of forensic SPR-driven security mechanisms, as proposed within this thesis, implies the intention to use the output of these methods in legal proceedings (i.e. in court). The author has no background in law to completely evaluate the whole set of legal challenges to the admissibility of traces or even evidence that is generated in this way, but some basics on this matter have to be discussed here, because this admissibility would be the ultimate benchmark for every forensic method.
All legal considerations presented here are based on freely available material concerning the U.S. legal system at the federal level. The U.S. legal system one of the most active in the world with large numbers of trials involving all kinds of forensic investigations being held every day. As a result, within this legal system strict rules for the integration of the results of forensic investigations have been established. These rules, the Federal Rules of Evidence (FRE)[9], define the framework within which evidence can be admitted into court. Even if these rules are in their original form only applicable on U.S. federal level,

[9] These rules constitute a ground-breaking law reform in U.S. federal law, specifying strict general rules instead of only relying on constitutional rights and precedents. They were approved by the U.S. Supreme Court and the Congress passed the FRE in 1975. The FRE became effective for all U.S. federal courts on July 1st, 1975.

their concepts for handling forensic data have, to the best of the author's knowledge[10], influenced many other judicial systems worldwide.

In general, **forensic results**, like the ones considered within this thesis, have to be interpreted by experts to the court. The reason for this lies in the assumption that any judge (or jury) will lack the expert knowledge to completely interpret the findings of a forensic investigation on his/her own and that therefore expert testimony is strictly required in court proceedings[11]. If the expert's opinion helps the fact finder in understanding the significance of factual data, then the expert witness is essential for the case and its opinion evidence is admissible.

Using the terminology of U.S. jurisdiction, the trial judge acts as a form of 'gatekeeper', assuring that scientific expert testimony truly proceeds from reliable (or scientific) knowledge. Considerations on relevance and reliability require the trial judge to ensure that the expert's testimony is 'relevant to the task at hand' and that it rests 'on a reliable foundation'. According to [SWGFAST11], the primary rules that are relevant for the presentation of forensic evidence in court (i.e. that apply to expert witnesses) in the Federal Rules of Evidence (FRE) are FRE rule 702 ("*Testimony by Experts*") and FRE rule 703 ("*Bases of Opinion Testimony by Experts*").

In the year 2000 **FRE rule 702** ("*Testimony by Experts*") stated (see e.g. [LLI10a]; FRE as amended April 17th, 2000, effective December 1st, 2000): "*If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.*" In 2011 FRE rule 702 was slightly amended[12].
When analysing this rule, it can be seen that regarding the admissibility of an expert, the judge has to establish whether the following four points are met:

- **Qualification of a witness as expert:** First, a witness has to qualify as an expert. For the description of the involved process for U.S. legal system see e.g. [SWGFAST11] or [Jackson08]. The conclusion of this process is that the presiding judge decides whether the witness may offer opinion testimony[13] as an expert.

- **Type of knowledge considered:** The first seven words of FRE rule 702 specify different types of knowledge that an expert can offer. The question is which kind of knowledge is generated by the media forensics methods considered in this thesis? According to [SWGFAST11] this question is addressed by the U.S. supreme court in the original precedent *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 119 S. Ct. 1167 (1999) [USC99]. In this case the court states that the same evaluation criteria used in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786 (1993) [USC93] to determine whether testimony offered as scientific knowledge is reliable should also govern the admissibility of testimony for the "*technical*" and "*other specialized knowledge*" [LLI10a] types of knowledge addressed in FRE rule 702. Therefore, distinguishing between science, applied science, technology, or experience-based expertise is not required.

[10] **Important notice:** The author has absolutely no legal training. All legal considerations made within this thesis are therefore layman's interpretation of freely available material, which are made to the best of the author's knowledge. If the content of this thesis is intended to be used in any legal proceedings, the reader must consult appropriate legal counsel for the corresponding jurisdiction.

[11] A good summary of the role science and scientists in the resolution of legal disputes is given in [Jackson08].

[12] FRE rule 702 as amended by the United States Supreme Court Apr. 26, 2011, (eff. Dec. 1, 2011) now reads: "*A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.*" This amendment was made to improve the readability of the rule and has, for this thesis, no influence on its interpretation.

[13] The judge can also define/limit the extent to which the expert is permitted to testify.

- **Who is addressed by the expert:** Basically, there are two entities the expert has to convince. First, the judge, to get admitted in pre-trial hearings, and second the 'fact finder' (the "*trier of fact*" in FRE rule 702 [LLI10a], either a jury in normal cases or a judge in non-jury trials) at the trial itself. In the context of this thesis only the first entity is relevant.

- **Qualification:** Any expert has to testify upon the five criteria listed in FRE rule 702 "*knowledge, skill, experience, training, or education*" [LLI10a]. This information helps the judge to decide whether an expert can be admitted to trial in a specific case and helps the 'fact finder' (i.e. usually the jury) to assign corresponding weights to each expert's testimony in the decision process.

If these four points are established, the judge determines for the case whether an expert is qualified to testify under FRE rule 702. The April 2000 (effective December 2000) amendment of FRE rule 702 includes three further requirements which must also be met. The goal of these additional requirements is to make it easier to present effective scientific and technical expert testimony whenever such evidence is warranted and provide a basis for the exclusion of opinion testimony that is not based on reliable or mature methodology. These additional requirements are ([LLI10a]): "[...] *if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.*" In April 2011 another requirement was added to this list ([U.S. Congress11]) "[...] *the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue* [...]".

These four requirements from FRE rule 702 are translated into the following evaluation criteria for forensic investigations within this thesis ([LLI10a]):

- **FREC0:** "*the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue*"

- **FREC1:** The investigation (which leads to the corresponding expert testimony) is "*based upon sufficient facts or data*"

- **FREC2:** The investigation is based upon "*reliable principles and methods*", preferably scientific methodology and knowledge[14]

- **FREC3:** The methods are applied "*reliably to the facts of the case*"

In the notes on FRE rule 702 published by the Legal Information Institute at Cornell Law School in December 2010 ([LLI10b]) the current regulations regarding the interpretation of this rule for U.S. federal courts are summarised as follows: "*Rule 702 has been amended in response to* Daubert v. Merrell Dow Pharmaceuticals, Inc., *509 U.S. 579 (1993), and to the many cases applying Daubert, including* Kumho Tire Co. v. Carmichael, *119 S.Ct. 1167 (1999). In Daubert the Court charged trial judges with the responsibility of acting as gatekeepers to exclude unreliable expert testimony,* [...]". The main result of this amendment, the so called Daubert hearings, is discussed in section 2.2.2.

The second important rule regarding expert testimony is **FRE rule 703** ("*Bases of an Expert's Opinion Testimonys*"; see e.g. [U.S. Congress11]). Since it is not directly relevant for this thesis, it is discussed in less detail here. Basically, this rule specifies the two different types of testimony experts are allowed

[14] For the possible legal differentiation between accepted reliable principles and methods on one hand and scientific methodology on the other hand see the discussion on the admission or rejection of expert testimony given by dactyloscopic experts of the FBI in the case *United States of America v. Plaza, Acosta, and Rodriguez* (United States District Court, E.D. Pennsylvania, March 13, 2002). In this case the presiding judge in the Daubert hearings initially prohibited the prosecution to call dactyloscopic experts in front of the jury, arguing that these are, despite being certified practitioners in the century old practice of latent fingerprint comparison, no scientific experts and have therefore to be excluded under the Daubert criteria. The judge also identified other 'flaws' in the dactyloscopic methods, which seemed to him in to be in conflict with the Daubert criteria. In further hearings in this case the FBI (and the scientific and legal experts it was able to present) could establish that dactyloscopy is a reliable method and that certified practitioners (despite being no scientists) can be admitted as experts in court.

to offer in court. The first is the testimony about first-hand knowledge. Here, the expert acts as a fact witness, testifying on observations made in examining evidence – which is the important component of this rule within this thesis. The second permitted type is the testimony based on reports or examinations made by others.

The **remaining Federal Rules of Evidence (FRE) regarding opinions and expert testimony** (rule 701 "*Opinion Testimony by Lay Witnesses*", rule 704 "*Opinion on an Ultimate Issue*", rule 705 "*Disclosing the Facts or Data Underlying an Expert's Opinion*" and rule 706 "*Court-Appointed Expert Witnesses*"; see [U.S. Congress11]) are of little relevance for this thesis and are listed here only for the sake of completeness. For a detailed analysis of the relevance[15] of these additional rules in the presentation of forensic evidence via expert testimony see [SWGFAST11].

Regarding other federal rules that might be generally relevant for the presentation of forensic results by an expert [SWGFAST11] lists the Federal Rules of Evidence (FRE) rules 401 ("*Test for Relevant Evidence*"), 402 ("*General Admissibility of Relevant Evidence*"), 403 ("*Excluding Relevant Evidence for Prejudice, Confusion, Waste of Time, or Other Reasons*") and Article X[16] ("*Contents of Writings, Recordings, and Photographs*") as well as the Federal Rules of Criminal Procedure rule 16 ("*Discovery and Inspection*").

FRE rule 401 demands that the evidence be relevant to the case at hand. It defines relevant evidence as "[...] *has any tendency to make a fact more or less probable than it would be without the evidence;*" and "[...] *the fact is of consequence in determining the action.*" (see [U.S. Congress11]). FRE rule 403 allows a judge to exclude certain relevant evidence as a matter of judicial discretion. The rule states (see [U.S. Congress11]): "*The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.*" FRE rule 402 completes the matter on admissibility by simply stating that "*Irrelevant evidence is not admissible.*" (see [U.S. Congress11]). FRE rules 1001 through 1008 address the contents of writings, recordings, and photographs. These rules set forth the definitions and requirements regarding what constitutes originals or duplicates and the admissibility of each, even if the original is lost or destroyed (see [U.S. Congress11]). The Federal Rules of Criminal Procedure rule 16 section G ("*Expert Witnesses*") specifies the need to provide accurate information on the witness' qualifications and a written summary of the testimony that is intended to be used under FRE rules 702, 703 or 705 (see [U.S. Congress10]).

2.2.2 Daubert challenges to forensic methods

Regarding the second and third point of the list given above in section 2.2.1 in the analysis of FRE rule 702 ('Type of knowledge considered' and 'Who is addressed by the expert') it has to be summarised that if something is declared to be 'science' in regard to FRE rule 702 then the criteria for the evaluation of scientific methods introduced in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) [USC93] have to be applied by the judge to make the expert prove this declaration. These criteria and their relevance for this thesis are discussed in this section.

In 1923 the court in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) made a first suggestion how to proceed with the admission of expert testimony based on novel forensic techniques. The court in **Frye** suggested [U.S. Congress23]: "*Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone*

[15] Mostly the mechanisms in FRE rule 701 intended to prevent the introduction of an expert as a lay witness and the possibility for experts to state upon 'ultimate issue' in FRE rule 704(a).

[16] This article focuses on the documents that are intended to be used in court. It is obvious in the context of this thesis that any documentation of the forensic processes must comply with the highest imaginable standards for the integrity and authenticity of the evidence. Such matters as a complete and documented chain of custody, evidence security from the time it is initially received to the time it leaves the laboratory and a documentation of the analysis steps that allows for a complete reproduction of the results are crucial for ensuring that evidence will be admitted in court.

the evidential force of the principle must be recognized, and while the courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs." In Frye (or the Frye standard as it is also referred to) the court concluded that the polygraph test that was intended to be used in this case could not be admitted because it lacked the required general acceptance in the corresponding research fields. Prior to this seminal ruling in Frye, according to [SWGFAST11], the competence of an expert was equivalent to his success in real life. In [SWGFAST11] it is summarised as: "*If a person earned a living selling his or her knowledge in the marketplace, then that person would be considered an expert who could testify at trial. Although not very sophisticated, this early principle of 'marketplace acceptance' (a concept we might in the post-Daubert parlance equate to some early form of peer review) served the law in a more or less acceptable manner for a great number of years.*"

The Frye standard was in 1975 partially replaced by the newly introduced **Federal Rules of Evidence (FRE)**. As might be noticed from the discussion of the FRE in section 2.2.1 above, they contained in the original version no special rule that, when dealing with 'scientific' evidence, novel or otherwise, ensured that science-based testimony is reliable and, therefore, admissible. Therefore all evidence was considered admissible if relevant, provided its use in court was not outweighed by "*unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence*", as stated in FRE rule 402 [U.S. Congress11].

The next relevant step in legal developments on expert testimony (and therefore the means of introducing forensic sciences into court) occurred in 1993, when the U.S. Supreme Court made another ground-breaking decision on expert testimony in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) [USC93]. **Daubert** was in 1999 followed by another important court case, *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999). Both Daubert and Kumho Tire arose out of civil lawsuits. An extensive and intelligible summary of the proceedings in the Daubert cases (original and the affirmation in the U.S. Court of Appeals) is presented in [SWGFAST11]. The main point of interest for this thesis is that the court unanimously held that Frye did not survive the enactment of the FRE. In interpreting FRE rule 702, the court in Daubert stated that if the admissibility of scientific evidence is challenged, it is the function of the trial court to act as 'gatekeeper' to determine whether proffered opinion evidence is relevant and reliable. The U.S. Supreme Court specified several flexible and non-exclusive criteria (the so called Daubert criteria or **Daubert standard**) to guide other courts when they have to consider in deciding whether a scientific field is sufficiently reliable to warrant admission of opinion evidence. As a further important milestone, in 1999 in *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999) the U.S. Supreme Court applied the Daubert criteria of proof of reliability to all forms of expert opinion testimony (i.e. scientific, applied science, technological, skill and experience). Also, the court in Kumho Tire made it clear that the list of Daubert criteria was meant to be helpful and is not a definitive checklist but rather a flexible, non-exclusive recommendation. As a result no attempt has been made in US law to 'codify' these specific criteria. Other U.S. law cases have established that not all of the specific Daubert criteria can apply to every type of expert testimony.

The outcome of Daubert and Kumho Tire led in April 2000 to the amendment of FRE rule 702 described in section 2.2.1. The Scientific Working Group on Friction Ridge Analysis, Study, and Technology (SWGFAST) draws in [SWGFAST11] the following conclusions for forensic scientists from Daubert and Kumho Tire: "*It means that the definition of science, the scientific method, and scientific evidence can no longer be used as loosely as experts have been doing. It is no longer sufficient to call yourself a forensic scientist in order to be considered a scientist. It is no longer sufficient to say that something is a subject of forensic science in order for a court to agree that it is dealing with science. Simply saying it does not make it so. The courts may, and many will, require the experts to show that they know what the scientific method consists of and provide the scientific basis for their conclusions. By the same token, each discipline will be judged by its own standards and upon its own experience.* [...] *It also means that forensic scientists can no longer expect to rely on the fact that courts have long accepted and admitted evidence of their expert conclusions. The court can relitigate the admissibility* *of a certain type of expert evidence if a litigant can make a credible argument that there has been no previous scientific inquiry of the validity of the assumptions on which a forensic field has long rested. Decades of judicial precedent no longer preclude reviewing whether existing precedent satisfies Daubert and Kumho Tire. Long-recognized forensic disciplines have been and are being challenged, with more to come.*"

This summarising statement, from a forensic expert group trained (and providing training) for court appearances as expert witnesses, defines the requirement for the investigation of the fitness of the exemplary selected statistical pattern recognition (SPR) based forensic methods considered in this thesis. The actual evaluation criteria applied for the investigations to be performed here are the criteria FREC0, FREC1, FREC2 and FREC3 specified in section 2.2.1 above as well as the criteria introduced by Daubert for the so-called **Daubert hearings**.

The specific criteria, explicated by the Daubert court and used within this thesis as evaluation criteria for forensic investigations, are [LLI10b]:

- **DC1:** "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"
- **DC2:** "*whether the technique or theory has been subject to peer review and publication*"
- **DC3:** "*the known or potential rate of error of the technique or theory when applied*"
- **DC4:** "*the existence and maintenance of standards and controls*"
- **DC5:** "*whether the technique or theory has been generally accepted in the scientific community*"

While the criteria DC2 to DC5 are self-explanatory (including the fact that publication in DC2 means 'open publication'), DC1 is summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*".

The Daubert criteria are widely accepted in the classical field of the medical forensics (see e.g. [Lally03], [Fulero09] and [Pinkl09]). It also can be, and is, applied in the much younger field of IT-forensics (see e.g. [Meyers04], [Nelson10]). It has to be admitted that the field of media forensics, which is the focus of this thesis, is still lacking maturity in this regard. Here, only very specific methods applied in this field already fulfil the Daubert criteria sufficiently. Overviews over the more mature techniques in this field are given in [Bijhold07] and [Daeid10].
Regarding digital camera forensics (as one of the most mature research fields in media forensics) a major breakthrough can be seen in the law case *United States of America v. Nathan Allen Railey* (United States District Court for the Southern District of Alabama[17], August 2nd, 2011). In the Daubert hearings of this case, the method of digital camera authentication based on intrinsic characteristics of its image acquisition sensory developed by Jessica Fridrich and her group (see e.g. [Goljan09]) got accepted for the first time as forensic evidence. The FBIs Forensic Audio, Video, and Image Analysis Unit (FAVIAU) established in the Daubert hearings that this approach meets all necessary criteria (DC1 to DC5) and the presiding judge furthermore decided that this evidence (or more precisely the FBI expert testimony based on this media forensic analysis) also meets the Federal Rules of Evidence (FRE) rule 702 criteria (FREC0 to FREC3). This is an important success for the whole research field of media forensics.

Considering the methods for audio material collected in [Bijhold07] by forensic experts from the Netherlands Forensic Institute, the German Bundeskriminalamt and the United States of America Federal Bureau of Investigation (FBI) Digital Evidence Laboratory, the two relevant research fields are, on one hand, forensic audio analysis and, on the other hand, speaker identification. While the latter is outside the scope of this thesis, the former, which looks into authentication, speech enhancement, transcription

[17] For a short summary of the relevant part of the preceedings see:
http://blog.al.com/live/2011/07/expert_witnesses_link_camera_t.html

of linguistic content / disputed utterance examination, and the analysis of non-speech events, lists with the electric network frequency (ENF) approach ([Grigoras09]) exactly one method for the authentication of digital audio material. This, in conjunction with the fact that the ENF analysis method can only be applied under rather specific circumstances, boldly underlines the current immaturity of this research field and the necessity of the investigations performed in this thesis.

In [SWGFAST11] an extensive and intelligible review of Daubert hearings regarding challenges to the admissibility of friction ridge individualisations ('fingerprints') for forensic identification of human beings is presented. This review is addressing all Daubert criteria for this forensic field as well as further concerns raised by judges in Daubert hearings. Despite the fact that the two exemplary selected statistical pattern recognition (SPR) based audio forensic methods discussed within this thesis are far away from having the same degree of maturity as the decades old practice of fingerprint verification, this methodology of analysing a forensic method is transferred to this thesis. The combination of the criteria directly derived from the Federal Rules of Evidence (FRE) rule 702 (FREC0 to FREC3; see section 2.2.1) and the criteria explicated by the Daubert Court (DC1 to DC5) is referred to within this thesis as the **Daubert criteria**. They are used in the investigations on the two selected application scenarios as performance indicators for an estimation of their currently achieved forensic compliance.

2.3 Selected fundamentals on audio signals

An audio signal, as it is perceived by a human listener, is a representation of pressure variations in air. Usually only pressure variations in the frequency range of 20 to 20,000 Hz are considered as being 'audible' since this is the (idealised) range of human hearing. An audio signal can be visualised in different forms, reflecting different transform domains. The most common forms are time-domain (which is the natural domain for audio signals) and frequency-domain representations of the signal. The latter is most often generated in audio by using Fourier transforms.

Digital audio signals are either 'born digitally', e.g. as MIDI[18] files or, more commonly, generated as analogue signals and recorded by microphones as shown in figure 2.1 below. The case of digitally born audio signals is neglected for the remainder of this thesis for two reasons: first, because it is less common than the case of recording and second, because it violates the basic assumptions for the two considered application scenarios, being that intrinsic traces of the used microphone are present for microphone forensics and that plausible (i.e. complex) covers are chosen for steganography in the audio steganalysis application scenario.

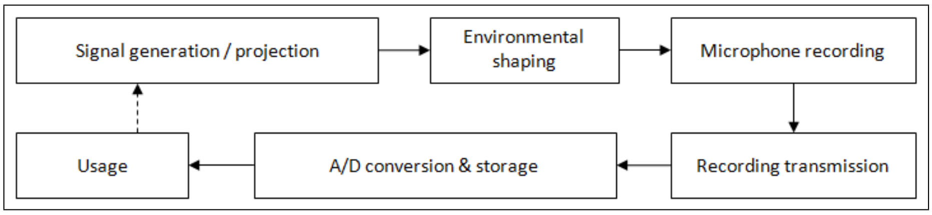


Figure 2.1: Natural audio signal lifecycle

Possible signal generation and projection processes can include human speech, instrumental music, environmental noise, loudspeaker projections, etc. After generation / projection the signal is subject to environmental shaping (e.g. the reverberation influence of the room) and superposition with other

[18] MIDI: Musical Instrument Digital Interface – an industry-standard protocol maintained by the MIDI Manufacturers Association (MMA, http://www.midi.org). It aims at enabling electronic musical instruments (synthesizers, drum machines, etc.), computers and other electronic equipment (MIDI controllers, sound cards, samplers) to communicate and synchronise with each other, either directly via cable connections or via MIDI-files. MIDI initially made no provision for specifying timbre. I.e. every MIDI synthesiser had its own methods for producing the sound from MIDI instructions. This situation was solved by the introduction of the General MIDI reference set in 1991, which created a standard set of 128 default sound types (piano, organ, guitar, strings, etc.) for sound generation.

audio signals. In the third step shown in figure 2.1 the signal is recorded by a microphone, introducing the intrinsic traces to the signal that is used for microphone identification within this thesis. In the recording transmission the still analogue signal is transferred as voltage variations from the microphone to the analogue-to-digital-conversion device (A/D converter). In this transit the signal is prone to electromagnetic distortions. Finally, the signal is low-pass filtered, sampled and quantised by the A/D converter. Also, if stereo microphones or multi-microphone arrays are used, the two- or multi-channel signal is composed. In most cases the signal generated by the A/D converter is in PCM[19] format. For storage the signal is either kept in this high-resolution format or compressed into other formats like MP3[20].

For the remainder of this thesis the performed pattern recognition observations are mostly restricted to uncompressed (never compressed) PCM signals with a sampling rate of 44.1 kHz with a resolution of 16 Bit in quantisation (CD-quality). This is done because it is the most common uncompressed end user format and sees wide-spread end user usage through its employment in the CD Audio standard as well as in DVDs and on Blu-Ray discs.

2.3.1 Audio signal representation

The nature of audio signals (or sound) is described by Pressnitzer et al. [Pressnitzer00] as follows: "*A vibrating body creates in the surrounding air the propagation of a pressure wave, in the same manner that to agitate an object on a surface of water provokes the propagation of wavelets. This is the physical reality of sound, the variation of acoustic pressure over time, and this reality is unique. It can, nevertheless, be represented in different ways according to the information that one wishes to emphasize.*"
Various representations for audio signals (or sound) can be found in daily use. They differ in many aspects and no representation is capable of displaying all characteristics of a sound perceived. Some of the most commonly used representations come in the form of music scores or various analogue or digital audio signal representations. Especially for the latter, a large set of different presentation forms can be identified in literature, e.g. time-domain, frequency-domain, time-frequency or psychoacoustical representations. Within this thesis the focus for the considerations of audio signal representations is limited to the two most common ones: the time-domain and frequency-domain representations. In the following two sub-sections the nomenclature used within the thesis for these two basic representations is introduced.

Digital audio signals in time-domain

Time-domain representations of audio signals are trying to directly capture the physical reality of sound. Therefore, this presentation is most likely to be considered the 'natural' domain for audio signals.

As shown in figure 2.2 a sampled and quantised digital audio signal is fully represented by k arrays of audio samples $S^k(n\mathrm{T})$ where T is the sample-time ($\mathrm{T} = 1/f_{sample}$ with f_{sample} being the sampling frequency; see Shannon-Nyquist Theorem [Shannon49]), n is the index of a sample in the corresponding array ($n \in \mathbb{N}$; $1 \leq n \leq streamlength$; $streamlength$ being the overall number of audio samples in

[19] PCM: Pulse-Code Modulation – is one method invented by Alec Reeves in 1937 to time- and amplitude discretely representation of sampled analogue signals. It is the standard format for digital audio in computers and various Blu-ray, compact disc (CD) and Digital Versatile Disc (DVD) formats. A PCM stream is a digital representation of an analogue signal, in which the magnitude of the analogue signal is sampled regularly at uniform intervals, with each sample being quantised to the nearest value within a range of digital steps. For a more detailed explanation of the Pulse-Code modulation see [U.S. Congress05].

[20] MP3: The MPEG standard MPEG-1 (and MPEG-2) Audio Layer 3 is more commonly referred to as MP3. It is a patented digital audio encoding format using lossy data compression. It is a common audio format for consumer audio storage, as well as a de facto standard of digital audio compression for the transfer and playback of music on digital audio players. Compressed representations, like MP3, usually employ redundancies in the perceived signal to reduce the storage space required. This can be done either lossy (e.g. in MP3) or lossless (e.g. Free Lossless Audio Codec (FLAC) - the Free Lossless Audio Codec http://flac.sourceforge.net/) and normally operates in a trade-off between storage space requirement and audio quality.

the signal) and k is the number of audio channels that exist in parallel within the signal. Therefore, k is usually either 1 (mono audio signals) or 2 (stereo).

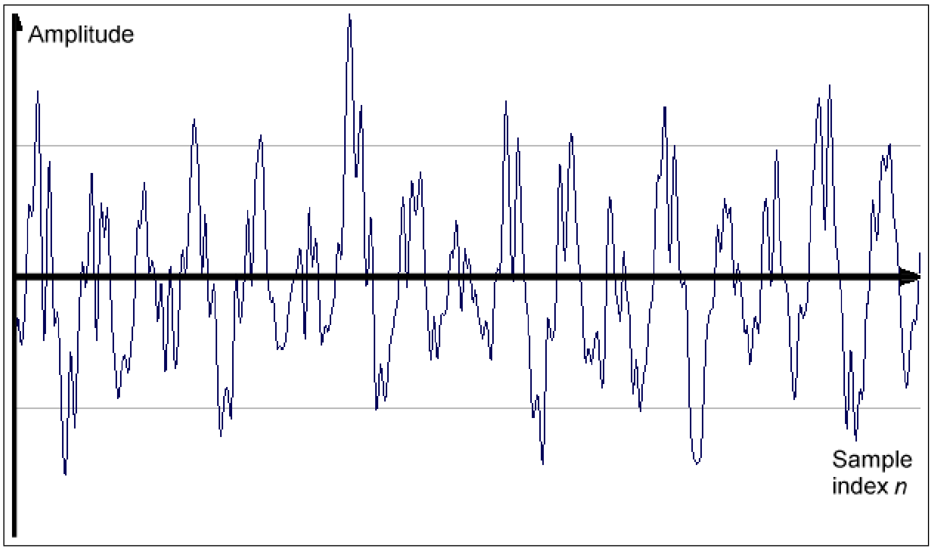


Figure 2.2: Time-domain representation of the audio signal $S^k(n\mathrm{T})$ for the case $k = 1$ (mono) – x-axis: sample index, y-axis: sample value as amplitude

For the purpose of audio signal analysis, each of the k arrays, representing the sample stream for one audio channel, in $S^k(n\mathrm{T})$ is first framed and then windowed. As shown in figure 2.3 in the framing, by the definition of a frame size w and an overlap o ($w, o \in \mathbb{N}$; $o < w$), the sample stream $S^k(n\mathrm{T})$ for each audio channel is split into $framecount$ frames $\bar{S}^k_i$ where i is the frame-index ($i \in \mathbb{N}$; $0 < i \leq framecount$). For given values for the frame size w and offset o the number of resulting complete frames can be computed for a sample stream of size $length$ as $framecount = \lfloor \frac{length}{w-o} \rfloor$ with a rest of $length \mod (w - o)$ samples.

Considering the frame size and the overlap between the frames, the start sample for each frame is at position $(iw - o(i-1) - (w-1))\,\mathrm{T}$ and the last sample in the frame is at $(iw - o(i-1))\,\mathrm{T}$.

A further processing step applies a windowing function (a.k.a. analysis window) $\mathrm{win}()$ to the frames $\bar{S}^k_i$, producing the sampled, quantised and windowed digital audio signal $S^k_i = \bar{S}^k_i \mathrm{win}(w)$. The choice of windowing function is strongly dependant on the intended post-processing. For signal processing operations sophisticated windowing functions like Hamming or Hann [Blackman59] windows are used. In signal analysis, like in this thesis, often a rectangular window (also known as Dirichlet window [Kumar10]) is used, which actually leaves the signal unmodified[21].

This time-domain notation for audio signals allows for considerations on feature extraction as required for the statistical pattern recognition (SPR) based application examples within this thesis. As explained in detail in section 2.4.2 these features are either computed globally (over the complete signal, or for each audio channel independently), locally (for individual data samples) or segment-wise (here for windows of the audio signal). All three different types of features are supported by the introduced notation for audio signals. If it is necessary for computation operations like e.g. feature extraction to access individual samples in a window S^k_i, these samples are denoted in the following as $s^k_{i,j}$, where i is the frame-index as above, k is the channel and j is the sample-in-the-frame index ($j \in \mathbb{N}$; $1 \leq j \leq w$).

[21] In this case S^k_i and $\bar{S}^k_i$ are equivalent.

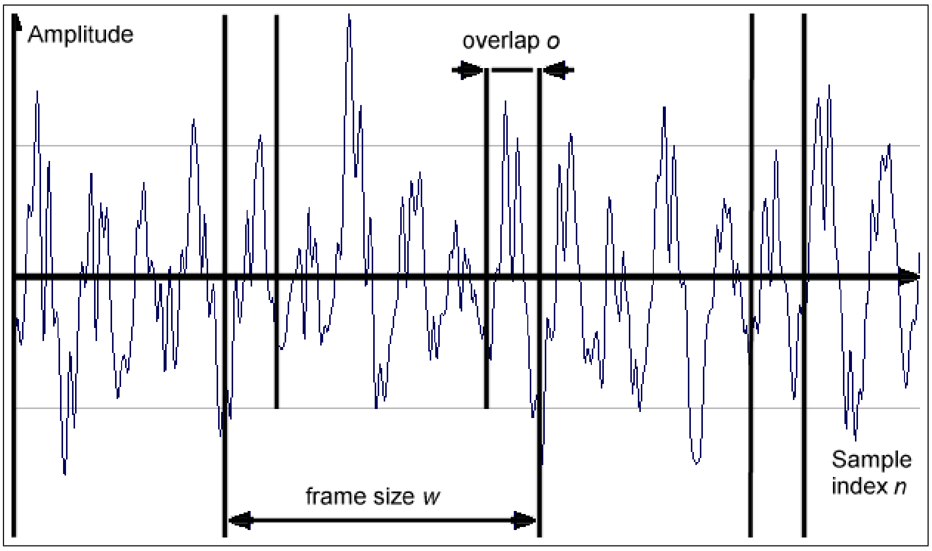


Figure 2.3: Framed digital audio signal S_i^k for the case $k = 1$ (mono) – the overlap between consecutive frames in this example is set to 20%, every second frame is marked with yellow background colour

Digital audio signals in frequency-domain

The frequency-domain representation of an audio signal is usually generated by the application of a Fourier transform to the time-domain audio signal. The $coef$-band spectrum ($coef = \lfloor \frac{w}{2} \rfloor$), as the real output of the Fourier transform, of a window S_i^k of the audio signal is denoted here by $Y_i^k = \mathrm{FT}(S_i^k)$. An important note for the computations performed here is, that the the spectrum used is an absolute spectrum (i.e. every coefficient has a value of 0 or larger). If it is necessary for computation operations like e.g. feature extraction to access individual coefficients in the spectrum Y_i^k of a window S_i^k, these coefficients are denoted in the following as $y_{i,h}^k$, where i is the frame-index as above, k is the channel and h is the coefficient-in-the-frame index ($h \in \mathbb{N}$; $1 \leq h \leq coef$).The phase component of the audio signal, as the second output of the transform, is irrelevant for this thesis.

The frequency-domain representation is required within this thesis, amongst other reasons, to perform a sophisticated context modelling for the microphone forensics application scenario. This context modelling of the influence factors in the recording process is performed in section 2.3.2.

2.3.2 Audio signal generation

For the work within this thesis two different types of audio signals have to be considered especially. These two types are: first, audio signals generated in microphone recordings, and second, the audio files generated by modification based audio information hiding. For the first type, previous work as [Oermann05] and [Kraetzer07c] postulate and demonstrate that the recording source (here the combination of the used microphone and attached recording equipment as a sensory unit) leaves a statistical imprint in the audio material, which can be used for source authentication. For the second type, the modification based audio information hiding, previous work (e.g. [Fridrich09]) illustrates that the modifications performed change the statistics of the audio material, which can be used for implementing steganalysis for this kind of material. Nevertheless, the implementation of a security mechanism requires sophisticated context modelling for the determination of possible influence factors for the problem at hand. Such context modelling for the two application scenarios considered in this thesis is presented below.

Context modelling for microphone recordings

A frequency-domain based context model describing the audio recording process is the one initially presented in [Kraetzer11] and generalised and extended in [Kraetzer12b]. In frequency-domain the different influence factors in the recording process can be modelled much easier than in time-domain.

The resulting model helps to understand the exact influences to the audio signal during the steps of the recording process and is therefore of importance especially for the considerations on microphone forensics within this thesis. In [Kraetzer11] it is assumed that the signal projection happens via loudspeaker. The more generalised model presented in [Kraetzer12b] shows how other forms of generation of the recording input could be covered by a slightly adapted context model.
As shown in [Kraetzer12b], an audio recording process within this thesis is described using a pipeline which consists of five segments.

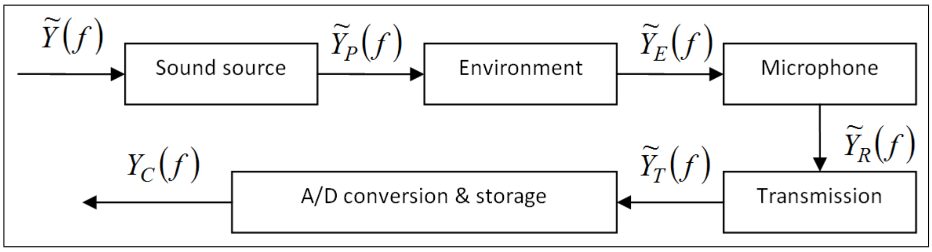


Figure 2.4: Recording process pipeline – context model (based on [Kraetzer12b])

Audio signals can be considered as either continuous or discrete signals in time- or frequency-domain. For the context modelling for the microphone recordings a frequency-domain representation of the signals is considered here, because it is more appropriate for the modelling of the analogue influence factors than a time-domain representation. Let a function $\tilde{S}(t)$ denote the analogue audio signal in time-domain, thus $\tilde{Y}(f)$, its representation in frequency-domain could be easily achieved by a Fourier transformation as shown in equation 2.1:

$$\tilde{Y}(f) = \mathrm{FT}\left(\tilde{S}(t)\right) \tag{2.1}$$

In Figure 2.4 $\tilde{Y}_P(f)$, $\tilde{Y}_E(f)$, $\tilde{Y}_R(f)$ and $\tilde{Y}_T(f)$ denote the analogue audio signals after each processing segment, while the output of the analogue-to-digital conversion process $Y_C(f)$ and its time-domain counterpart $S(t)$ computed via inverse Fourier transform as $S(t) = \mathrm{FT}^{-1}(Y_C(f))$ denote the final audio signal as the result. Due to the fact that for microphone recordings the number of recorded channels per microphone k is usually[22] equal to 1, the signal $S(t)$ is thereby equivalent to $S^k(n\mathrm{T})$ introduced above.
If a loudspeaker is used as the sound source, the processing operations within the recording pipeline can be modelled as follows:

$$\tilde{Y}_P(f) = \sum_{n_{driver}} \int_u^l F_{driver}(f)\tilde{Y}(f)df + N_{ls}(f) \tag{2.2}$$

A typical loudspeaker consists of multiple drivers ([Davis97]), as individual electrodynamic drivers provide quality performance over at most about three octaves. Equation 2.2 simulates the process of a loudspeaker with n_{driver} different 'drivers' playing the audio signal. Depending on different driver types (full-range, subwoofer, woofer, mid-range or tweeter), the upper and lower frequency values (u and l) range, and the amplifying function $F_{driver}(f)$ could be simplified into a constant amplifying factor in ideal circumstances. Furthermore, $N_{ls}(f)$ denotes the (thermal) noise that the loudspeaker generates in the playback signal.

After the audio signal is generated by the sound source, it is usually distorted by various environmental factors before it reaches to the microphone. There are mainly three aspects of such distortions: reflections, reverberation and addition of environmental noise. As introduced by [Pawera03], there are three types of reflections. The short-term reflections, which arrive at the microphone only fractionally later (0.8 to 20 ms) than the direct sound, produce discolouration. The medium-term reflections, with delay times of usually more than 40 ms, enhance the volume of the direct sound. The long-term reflections,

[22] Stereo or multi-channel recordings are generated using more than one microphone.

with delay times longer than 80 ms, create echoes. If there exist multiple long-term reflections in the sound field and the reflections reach an energy intensity equal to the one of the direct sound, the reverberation is triggered. Equation 2.3 describes all three distortions as follows:

$$\tilde{Y}_E(f) = e \int_f D(f)\tilde{Y}_P(f)df * F_{reverb}(f) + N_{envi}(f) \tag{2.3}$$

In equation 2.3, $D(f)$ denotes the discolouration function resulting from short-term reflections, e denotes the enhancement factor imposed by medium-term reflections, and the convolution with $F_{reverb}(f)$ simulates the possible distortion from the echoes and / or reverberation [Takala92]. The consistency of this convolution is the characteristic verified for an audio recording in the forensics approach presented in [Malik10].
When the recording process is accomplished in an anechoic environment, then $F_{reverb}(f)$ can be considered as an constant value of 1. The possible distortion caused by environmental noise is denoted by $N_{envi}(f)$ in the equation.

$$\tilde{Y}_R(f) = \int_h F_{mic}(f)\tilde{Y}_E(f)df + N_{mic}(f) + N_{ENF}(f) \tag{2.4}$$

Equation 2.4 simulates the process of a microphone collecting the signal. In this equation $F_{mic}(f)$ denotes the frequency response function of the microphone, $N_{mic}(f)$ denotes the thermal noise that the microphone generates, and $N_{ENF}(f)$ denotes the electric network frequency (ENF) influence (which is the characteristic used for the ENF approaches to recording setup forensics, see e.g. [Grigoras07]).
We assume for our approach that the specificity of a microphone is decided by the characteristics ($F_{membrane}(MembCharacteristics)$) of the membrane in the microphone with its unique vibration behaviour and interaction with the other parts of the microphone. Other influences to be considered here are the orientation of the microphone to sound sources, the microphone mounting and possible aging phenomena of the microphone. These influences are modelled within our context model as multiplicative influences Or (orientation), $Mount$ (mounting and Age (aging). So far no sophisticated model exists for the estimation of these influences; therefore we assume them to be Gaussian distributed with a mean of 1 and a small variance – which would, for these multiplicative influences, imply that they have only a very small influence. Thus $F_{mic}(f)$ can be considered as a function as follows:

$$F_{mic} = F_{inf}(Or, Mount, Age)F_{membrane}(MembCharacteristics) \tag{2.5}$$

Usually $N_{mic}(f)$ can be considered as a constant as it contributes a rather minor influence on the recorded signal compared to $F_{mic}(f)$.
This modelling of the microphone response is independent from the actual microphone type (condenser, electrets, pietzo, etc.; see [Pawera03]). The type only determines the strength of the influences.

$$\tilde{Y}_T(f) = \int_h F_{tran}(f)\tilde{Y}_R(f)df + N_{tran}(f) \tag{2.6}$$

Equation 2.6 describes the distortion caused by the signal transmission. Here, $F_{tran}(f)$ denotes the possible non-linear distortion during the transmission of the signal from the microphone to recording device. The component $N_{tran}(f)$ denotes the thermal noise coming from the transmission environment.

$$Y_C(f) = \int_0^{f_{sample}} F_{samp}(f)\tilde{Y}_T(f)df + N_{quan}(f) + N_{thermal}(f) \tag{2.7}$$

The equation 2.7 summarises the process of analogue-to-digital (A/D) conversion and storing the audio as an audio file. In the equation f_{sample} denotes the sampling frequency (a.k.a. Nyquist frequency), $F_{samp}(f)$ the sampling function, $N_{quan}(f)$ denotes the quantisation noise, and $N_{thermal}(f)$ the thermal noise of the A/D device. As a last step in the modelling, the signal is projected back into time-domain $S(t) = \mathrm{FT}^{-1}(Y_C(f))$.
The Sampling Theorem [Shannon49] states that continuous-time signals can be fully represented with discrete-time samples of the signal, if we sample the signal often enough, i.e. using a sampling frequency

(f_{sample}) at least twice as high as the highest frequency in the sampled signal.
Therefore, for a sufficiently high sampling frequency (e.g. 44100 Hz for CD-quality audio material) the signal $S(t)$ can be fully represented by a stream $S(n\mathrm{T})$ of $n\mathrm{T}$ samples, where T is the sample time ($\mathrm{T} = \frac{1}{f_{sample}}$) and n the sample index $n \in \mathbb{N}$ ([Bosi03]). Following the sampling operation, a quantisation step is performed by the mapping of continuous amplitude values of the signal into codes that can be represented with a finite number of bits R (also called quantisation range). This thesis only focuses on scalar and uniform quantisation (the mapping of an amplitude value is not influenced by previous or following amplitude values and equally sized ranges of input amplitude are mapped onto each code) for the considered audio signals, further quantisation types are discussed in detail in [Bosi03]. The quantisation transforms the time-discrete but continuous-amplitude samples to a stream of time- and amplitude-discrete samples $S(n\mathrm{T})$. For further processing this stream is usually split into the individual audio channels k, framed and windowed as described above in section 2.3.1. A good reference on further details on A/D conversion is [Robin00].

Context modelling for modification based audio information hiding / audio steganalysis

The context modelling for modification based audio information hiding is much simpler than the context modelling for microphone forensics. The influence of the embedding of a data hiding scheme is here modelled as:

$$ST^k(n\mathrm{T}) = Embed_{Alg}(C^k(n\mathrm{T}), key, message, strength) \tag{2.8}$$

In equation 2.8 the embedding function $Embed_{Alg}()$ for a data hiding algorithm Alg is supplied with the following inputs: $C_k(n\mathrm{T})$ denoting the cover object (a digital audio signal), the embedding key key, the message embedded ($message$) and the embedding strength $strength$. The output of the embedding is the steganogram $ST^k(n\mathrm{T})$ as a digital audio signal.
The task of the steganalysis is then to decide whether an audio signal $S^k(n\mathrm{T})$ is a (unmodified) cover object $C_k(n\mathrm{T})$ (e.g. a microphone recording of some speech) or a steganogram $ST^k(n\mathrm{T})$.
More extensive context modelling on modification based steganography, which strongly exceeds the requirements for this thesis, is presented e.g. in [Winkler11].

2.4 Selected fundamentals on (statistical) pattern recognition

A **pattern**, from the French *patron*, is a type of theme of recurring events, objects or characteristics. Therefore, the pattern is the basis which allows the classification of objects or events into distinct classes or sets.
If using a definition given by Bebis [Bebis06], then **pattern recognition (PR)** is in general the study of how machines can observe their environment, learn to distinguish patterns of interest from their background signals and make sound and reasonable decisions about categories of the patterns. Therefore the key objectives in pattern recognition are to process the sensed data to eliminate noise, perform a suitable information reduction (by feature extraction), hypothesising the models that describe each class population and, given a sensed pattern, choosing the best-fitting model for the assignment to the class associated with the model. In short: Pattern recognition is the act of taking in raw data, processing it into features and taking an action based on the ‘category’ (or class) of the pattern [Duda01].

In literature (cf. [Duda01], [Bebis06]) the following five distinct classes of pattern recognition approaches are considered:

- **Syntactic pattern recognition** utilises the structure of the patterns. Instead of carrying an analysis based strictly on quantitative characteristics of the pattern, here the interrelationships between the primitives (the components which compose the pattern) are emphasised. Typical patterns which are subject to syntactic pattern recognition research are therefore characters, fingerprints, chromosomes, etc. The analogy between the structure of some patterns and the syntax of a language which has a solid theoretical basis is a very attractive one. In [Friedman99] it is stated that: “*By introducing concepts like a formal grammar and a language the design* *syntax classifiers is enabled, so that these can classify a given pattern presented as a string of symbols. In general, given a specific class, a grammar whose language consists of patterns in this class is designed. For an unknown new pattern a syntax classifier analyses the pattern (a string) in a process called parsing and determines whether or not that string belongs to the language (class).*"

- **Template matching** can be subdivided into two approaches: feature-based template matching and global template-based matching. The feature-based approach uses features, such as edges or corners in image analysis, and distance based solutions as the primary match-measuring metrics to find the best matching template for a candidate input. Since this approach does not consider the entirety of the training objects and the candidate input, but only extracted features, it is generally more computationally efficient when working with larger digital objects than the global template-based approach.
  The global template-based approach uses the entire template with generally a sum-comparing metric (e.g. using cross-correlation [Cole04]).

- Training patterns of various classes overlap often, for example when they originate from similar statistical distributions. In this case a **statistical pattern recognition (SPR)** approach is appropriate, particularly when the various distribution functions of the classes are known. If these distributions are not known they must be approximated using the training patterns. Sometimes the functional form of these distributions is known and one must only estimate its parameters. However, in some applications even the distribution's form is unknown and must (approximately) be found. The model for a pattern may be a single specific set of features, though the actual pattern sensed has been corrupted by some form of noise.
  A statistical classifier must also evaluate the risk associated with every classification which measures the probability of misclassification. Statistical pattern recognition (SPR) is focussing on the statistical properties of the patterns (generally expressed in probability densities) this approach receives the most attention in [Duda01] and is the one most widely considered in the chosen application fields for this thesis.

- In contrast to the statistical approach the **structural pattern recognition approach** tries to describe the structure of objects that intuitively reflects the human perception. The features become primitives (sub-patterns), fundamental structural elements, like strokes, corners or other morphological elements.
  Next, the primitives are encoded as syntactic units from which objects are constructed. As a result, objects are represented by a set of primitives with specified syntactic operations. For instance, if the operation of concatenation is used, objects are described by strings of (concatenated) primitives. [Pekalska05] states: "*The strength of the statistical approach relies on well-developed concepts and learning techniques, while in the structural approach, it is much easier to encode existing knowledge on the objects*".

- The **neural network pattern classification** (also known as neural pattern recognition or neural net approach) is considered a close descendant of SPR despite its somewhat different intellectual pedigree. It assumes as other approaches before that a set of training patterns and their correct classifications is given. The architecture of the network which includes input layer, output layer and hidden layers can be very complex. It is characterised by a set of weights and activation functions which determine how any information (input signal) is being transmitted to the output layer. The neural network is trained with training patterns and adjusts the weights until the correct classifications are obtained. It is then used to classify arbitrary unknown patterns.

The considerations within this thesis are restricted to statistical pattern recognition (SPR) and template matching, because these two classes of pattern recognition approaches appear to be the most significant ones in existing media forensics approaches, the two application scenarios considered in this thesis. For more detailed information about the mentioned classes of pattern recognition approaches the author refers to literature focussing on this topic like e.g. [Duda01], [Friedman99].

Since the processes for template matching and SPR are very similar[23] one general (but simplified) **pattern recognition pipeline** can be presented for both these classes. Figure 2.5 shows this general pattern recognition pipeline. The process is generally divided into two phases: First, the phase of reference data acquisition, processing and storage and second the assignment of a class to candidate signals based on the knowledge generated in the first phase.

The first phase is most often called **training phase** but it is also known in specific application scenarios as training, registration phase, enrolment, etc.; the second phase is commonly called **testing** but also for this phase different names like classification, field application or authentication can be found in specific application scenarios.

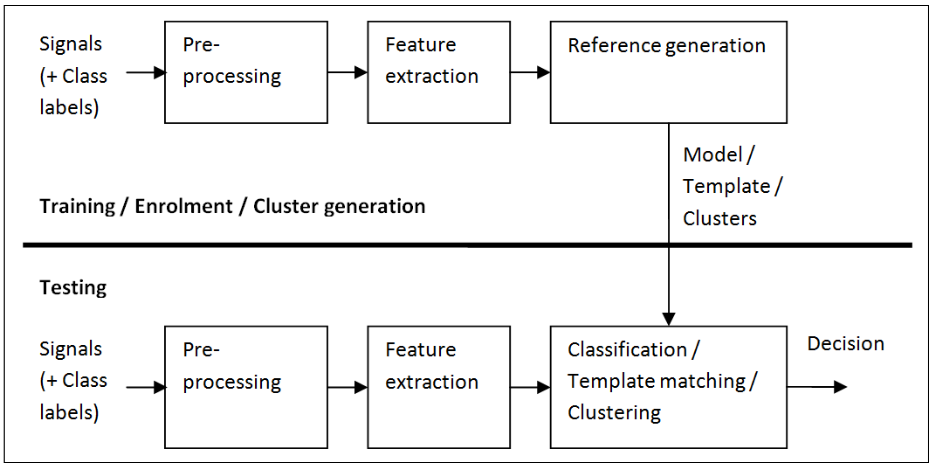


Figure 2.5: General pattern recognition pipeline (simplified)

The three basic operations performed in the training phase are pre-processing, feature extraction and the reference data set generation. Depending on the kind of PR used (here supervised statistical pattern recognition (SPR), clustering or template matching) the output of the reference generation operation is either called a model (for supervised SPR), a template (in template matching) or a set of clusters in clustering.

In the testing the same pre-processing and feature extraction operations are performed like in training. Based on the training output (the statistical model, the set of templates or the set of clusters) the assignment of a class to the candidate signal is performed. As stated already above for figure 2.5 this general pipeline is a simplified projection of the actual processed employed in practise. In the following sections this simplified pipeline is extended by additional operations required for this thesis like feature selection.

If the correct class labels of the candidate signals in testing are known they can be used to evaluate the accuracy achieved in the assignment of the classes for the candidate signals. If they are not known (as in most industrial applications of pattern recognition based solutions) an appropriate number of tests should be performed prior to field deployment of such a solution to investigate its performance in normal conditions expected for the application scenario and under extreme conditions to establish trust into this solution.

It has to be noted at this point, that certain publications assign further additional blocks to this pipeline. The most common additional blocks are: the signal acquisition (see e.g. [Vielhauer05]), which is in this thesis assigned to the conceptual entity of signal preparation outside of the pattern recognition problem, and the segmentation (see e.g. [Duda01]), which, if it would be required within is thesis, would be considered here to be part of the pre-processing.

[23] In fact template matching is so close to SPR that some authors consider them to be one class of pattern recognition techniques, with template matching covering the cases with the simplest expression of the class statistics in the model – the explicit description of a PDF by explicitly naming all associated instances – a that allows for an equally simple classification by distance-based classifiers.

2.4.1 Pre-processing

The task of **pre-processing** is to address two distinct topics. First, it could be used to maximise the distance between pattern and noise. In [Duda01] noise is defined in very general terms as: "[...] *any property of the sensed pattern due not to the true underlying model but instead to randomness in the world or the sensors. All non-trivial decision and pattern recognition problems involve noise in some form.*" Therefore it has to be clearly stated that noise is neither the intra-class nor the inter-class variance in the objects to be classified. So, the pre-processing does not enable the distinction between the different classes but only makes it easier. The pre-processing can either be content-based (analysing the syntax and or semantics of the content) or content-insensitive.

The second function of the pre-processing is to prepare the signal by transforming it into the input format expected by feature extraction operation, e.g. by windowing (see the considerations on digital audio signals in time-domain in section 2.3.1).

2.4.2 Feature extraction

Statistical pattern recognition (SPR) is generally **feature** based. The features originate from a feature extractor, whose purpose is to reduce the data (in this thesis audio data) into a d-dimensional feature space by measuring certain pre-selected properties. These features (or, more precisely, the values of these features) are then passed to a classifier that evaluates, with a model based decision boundary (or sometimes a simple threshold), the evidence presented and makes a decision as to the class of the object under evaluation.

There exist two completely distinct approaches for **feature design**: features are either especially designed for an application scenario, which, despite the fact that it is sometimes also called intuition-based feature design, usually requires strong domain knowledge, or the features to be used are transferred from other, similar signal processing domains.

The two general types that are most commonly considered in this context in literature (e.g. [Shyu98]) are **local and global features**. Local as well as global features are either determined content based or without higher-level content analysis. A good example for content based local features is the determination of minutiae in fingerprint images; an example for local features computed without higher-level content analysis could be the colour-value distance between one pixel and the next in a row in an image. For content based global features an example could be the existence of a specific object (e.g. a dolphin) in an image; an example for global features computed without higher-level content analysis could be the entropy of a complete signal. It is obvious that the global features perform the strongest information reduction, while especially the local features computed without higher-level content analysis provide very little information reduction.

As an in-between for local and global features a third class, the segment-wise computed features (also known as **segmental features** or intra-window features) can be determined. They could be considered as being a global feature (e.g. entropy) applied only to a segment of the whole signal or as the evaluation of local features for a whole segment (e.g. the number of colour-value changes in and image block). Also this segmental approach to feature computation is often employed when features are extracted in a transform domain representation of the original signal (e.g. in frequency-domain representations of audio or image signals) since many established domain transforms are working segment-wise (a.k.a. window-wise). In terms of information reduction are the medium between local and global features.

2.4.3 Feature selection

Because the addition of irrelevant or even confusing features often confuses pattern recognition systems they should be removed prior to classification to optimise the classification accuracy achieved. At the same time a reduction of the dimensionality of the feature space results in a decreased computational complexity[24], this is equivalent to an increased classifier throughput (or a lower cost in a cost-based

[24] It should be mentioned at this point that feature selection itself might be a computational complex task which easily outweighs simple classification problems.

benchmarking). An additional result of this process, especially in scientific applications, is that it results in less complex, more easily interpretable representation of the target concept and therefore the basis of the problem.

This process of selecting the significant features for a problem is called **feature selection** (or attribute selection). Its integration in the statistical pattern recognition (SPR) pipeline is shown in figure 2.6.

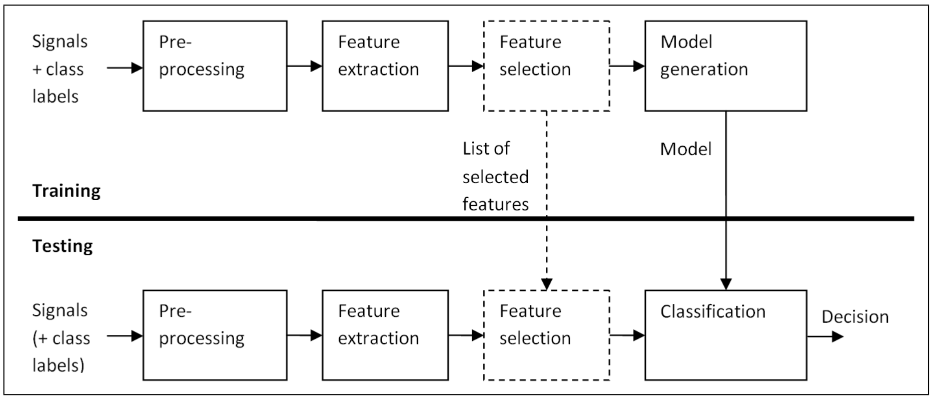


Figure 2.6: Statistical pattern recognition pipeline – précised and extended from figure 2.5

Following the methodology proposed for feature selection by [Witten05] the best way to select features would be a manual selection based on a deep understanding of the learning problem and the actual meaning of the features. Since the required knowledge is not always available different approaches for automatic feature selection are available.

For this exist two fundamentally different approaches for scheme-independent **feature evaluation**: The first one is using **filters** on the feature set to find the most promising features and the second one is using a classifier in a **wrapper** to identify the significant features. The problem with both approaches is according to [Witten05] the lack of a universally accepted measure of 'relevance', which is in this thesis represented by using the influence on the achieved classification accuracy.

A simple approach for the filter approach would be the evaluation of subsets of features until one is found that distinguishes all instances uniquely. This can easily be done using exhaustive search beginning with an empty set, although at tremendous computational costs. Additionally this approach faces problems from bias and a strong tendency for overfitting (see [Witten05]).

For the wrapper method a simple approach would be to employ quite stable classifiers like decision trees to identify the significant features and discard insignificant or distracting ones (which would, if they were used at all, be used very far down in the tree). The set of significant features could then be used in classifiers which react notoriously bad to irrelevant features, like e.g. nearest-neighbour methods.

In every feature selection approach the feature evaluator or feature set evaluator is accompanied by a **search method** which navigates the evaluator though the feature space, since exhaustive search is impractical on all but the simplest problems. Details on different search methods like forward selection or backward elimination can be found in [Witten05]).

More **advanced methods for feature evaluation** like e.g. symmetric uncertainty [Witten05], principal component analysis (PCA; [Lu07]) or random projection [Blum06] based methods allow not only for the removal of irrelevant or even confusing features but also eliminate redundancies in the set. The first uses entropy and joint entropy between features and performs then a correlation-based feature selection. The PCA transforms the data linearly into a lower-dimensional space with the drawback of being very computational expensive (the time taken to find the transformation is cubic in the number of attributes). The random projection is projecting the data into a subspace with a predetermined number of dimensions. Random projections are computationally much cheaper than PCA-based methods but normally show worse accuracy than those employing PCA. For more details on basic and advanced methods for feature selection see e.g. [Witten05] chapters 7.1 and 10.8.

Another approach, that has to be mentioned here as a **possible extension to the introduced scheme**, is the usage of methods form **analytical statistics** (e.g. variance analysis or factor analysis) to determine the suitability of features.

Regarding the **implementation of feature selection methods**, there already exist multiple applicable software solutions that can be used for performing feature selection. One such tool that has to be mentioned in this context, because it is used to implement feature selection within this thesis is the renowned open source data mining suite WEKA [Hall09]. For details on the various filter- and wrapper-based feature selection mechanisms implemented in WEKA see [Hall09].

2.4.4 Classification

Within practical investigations performed in this thesis the considerations on the assignment of the class to the candidate signals are limited on supervised statistical pattern recognition (also known as **classification**) and unsupervised statistical pattern recognition (**clustering**).

Regarding the nature of the classification process and its relation to the preceding feature extraction [Duda01] states the following: "*The conceptual boundary between feature extraction and classification* [...] *is somewhat arbitrary: an ideal feature extractor would yield a representation that makes the job of the classifier trivial; conversely, an omnipotent classifier would not need the help of a sophisticated feature extractor. The distinction is forced upon us for practical, rather than theoretical reasons. Generally speaking, the task of feature extraction is much more problem and domain dependent than is classification, and thus requires knowledge of the domain.*"

Hence, **classification** is, in most general terms, the task of recovering the model[25] that generated the patterns for each class. In this process the application of different classification techniques might be useful depending on the type of candidate models themselves.

The first question to be answered in a pattern recognition (PR)-based solution approach is always the question for the type of problem at hand (i.e. **classification problem identification**). The types of problems to distinguish between are: single-class, two-class and multi-class. While the first is mostly suitable for anomaly or outlier detection, the two-class classification is employed where either only two classes have to be distinguished or where a complex classification is split into a larger number of two-class classifications. The multi-class classifiers are the ones which propose most often the naïve solution approach for pattern recognition problems found in practical applications. The answer to the question to which class a problem belongs determines which classification algorithms can be applied to the solution of the problem.

A closely related question is the question whether the actual classes are explicitly known for all samples in the classification process (classification) or not (clustering). The answer to those questions is directly influencing the set of algorithms available as choices for implementing a solution to a practical problem.

To the knowledge of the author no agreed upon strategy for classifier selection exists in the research field of data mining. It seems to be common to establish the suitability of classifiers empirically. The approach pursued in this thesis is based on this assumption and performs a brute force testing through a large set of available classifiers. Since the number of implemented classifiers (and different possible parametrisations for those) found as standalone implementations and in data mining suites is extremely high, the author restricts the set of choice for this thesis to the set of classifiers implemented in the renown data mining software suite WEKA[26] (version 3.6.1; [Hall09]) and their default parameters.

Based on the general idea that different classification techniques might be useful depending on the type of candidate models themselves, methodologies for the evaluation of those different classification tech-

[25] Statistical pattern recognition (SPR) uses models, which are typically mathematical in form, to describe the classes.

[26] WEKA is open source and freely available for download, accepted in scientific communities with pattern recognition tasks, platform independent and easy to automate for a large number of test. Additionally, due to the form of its implementation, it can be combined with external (reliable) time measurement mechanisms to determine the cost of an operation. It is comparable in functionality and usability with commercial suites like IBMs SPSS Modeler (http://www-142.ibm.com/software/products/de/de/spss-modeler) or other open source solutions like Orange (http://www.ailab.si/orange/).

niques become a requirement. These evaluation or benchmarking methodologies are already a necessity when using individual classifiers but they become even more important in the prospect of information fusion on the output of different classification techniques.

Two general concepts from decision theory (the science which includes as the most prominent sub-field the pattern classification) to be mentioned in this context are the **cost** and **risk** of such a decision. The cost of the application of pattern recognition can be expressed as a function of its computational complexity. In [Duda01] it is stated on this subject: "[...] *we may ask how an algorithm scales as a function of the number of feature dimensions, or the number of patterns or the number of categories. What is the trade-off between computational ease and performance? In some problems we know we can design an excellent recognizer, but not within the engineering constraints. How can we optimise within such constraints?*"

Closely associated to the cost of the application of a pattern recognition approach is the notion of its risk. In [Duda01] it is stated on this subject: "*We should realize that a classifier rarely exists in a vacuum. Instead, it is generally to be used to recommend actions (put this fish in this bucket, put that fish in that bucket), each action having an associated cost or risk. Conceptually, the simplest such risk is the classification error: what percentage of new patterns are called the wrong category. However the notion of risk is far more general, as we shall see. We often design our classifier to recommend actions that minimize some total expected cost or risk. Thus, in some sense, the notion of category itself derives from the cost or task. How do we incorporate knowledge about such risks and how will they affect our classification decision? Finally, can we estimate the total risk and thus tell whether our classifier is acceptable even before we field it? Can we estimate the lowest possible risk of any classifier, to see how close ours meets this ideal, or whether the problem is simply too hard overall?*"

Generalisation is the handling of objects in testing, which are outside the model defined by the training samples, i.e. novel patterns. In most cases a complex model (i.e. represented by a complex decision boundary) will not provide a good generalisation since it tends to overfit – to closely match the particular training samples, rather than some underlying characteristics (e.g. the probability distributions of the categories) or true model.

In most PR problems, however, the amount of data that can be obtained easily for training is often quite limited. For other pattern recognition problems it might be easy to obtain training examples which as a consequence might lead to overly excessive model sizes which result in to complex classification operations. Therefore, one of the most important questions in pattern classification is how to optimise the trade-off between the complexity of the decision boundary on one hand and generalisation and **overfitting** on the other hand.

This problem can be translated in practise with the question: How should one design training sets for the derivation of classifier models suitable to solve the pattern recognition model at hand?

A problem, often contributed to classification, but actually inherent with the general problem of pattern recognition is the so called **curse of dimensionality**. The problem is the exponential growth of the volume of the d-dimensional space or hyper-volume in which the classifier model is created as a function of dimensionality [Bellman61]. By this exponential growth, problems tend to become intractable as the number of the dimensions increases. To rephrase this general mathematical problem for pattern recognition (PR) problems, it can be said that: the training of models from a finite number of data samples in a high-dimensional feature space requires an enormous amount of training data to ensure that there are several samples with each possible combination of values.

To consider this problem from a practical side would mean to address the Hughes effect[27] (or Hughes phenomenon; named after Gordon F. Hughes) established in [Hughes06]: With a fixed number of training samples, the predictive power would be reduced by increasing the dimensionality.

Regarding the implementation of classification methods, there exist multiple applicable software solutions that can be used for performing feature selection. Two of these solutions, which are used within

[27] Not to be confused with the similarly named, but completely unrelated, Hughes effect in electromagnetism named after Declan C. Hughes.

this thesis, are libSVM[28] [Chang11] and WEKA [Hall09]. While libSVM implements different support vector machines as classifiers, WEKA is a large open-source data mining suite with 74 supervised classification methods and 8 clustering algorithms implemented (figures for version 3.6.1). For details on the various classifiers implemented in WEKA see [Hall09].

2.5 State-of-the-art in audio steganography and audio steganalysis

The goals and requirements for audio steganalysis (or steganalysis in general) cannot be described without considering, at least to some extent, also the goals and methods of steganography. As introduced in section 1.1 steganography is the art and science of hidden communication. Steganalysis as its counter-science is supposed to detect these hidden communication attempts. Regarding the security aspects addressed, steganography focuses on the confidentiality of communications, while steganalysis is considered within this thesis as a mechanism aiming at reliable integrity verification of audio data against steganographic modifications. Steganalysis can be implemented either as a security mechanism protecting a communication channel against un-allowed modification (the online or warden setup) or a security mechanism that performs a-posterior forensic analyses on communication channel traffic records.

The following overview includes in section 2.5.1 a brief summary on general steganographic characteristics and considerations on the security of steganographic schemes. These descriptions include: a summary on the three existing basic approaches to perform steganography, the steganographic channel model, two existing theoretical methods to model steganographic security and a notion on the practical security of steganographic schemes.

Based on these facts an identification of the state-of-the-art in steganalysis is performed in section 2.5.2. The considerations in this section include some required considerations on the extend of recent steganographic research and the extreme diversity in audio steganography approaches, which is then put in contrast to the number of available audio steganography tools and the (nearly non-existent) diversity in (audio) steganalysis approaches. These considerations are accompanied by a brief discussion of the different existing sets of goals for steganalysis and the possible setups (online as well as forensic) that can be imagined. Existing field studies for this field are discussed as well as the results achieved under laboratory conditions, followed by a brief indication on the numbers of available (commercial and non-commercial) steganalysis toolsets/steganalysis detectors. At the end of this section, for the sake of completeness a standard-like end-user guide on steganalysis (or rather steganalytical benchmarking of steganographic tools) is summarised.

The section 2.5.3 summarises the principal methodologies and basic concepts used in steganalysis. It is oriented on the design of a statistical pattern recognition (SPR) pipeline and contains considerations on: the patterns observed, the input signals, pre-processing, feature design and selection, template or model sizes. The descriptions are made considering the most dominant pattern recognition (PR) approach in this field (statistical pattern recognition (SPR)) and different classification approaches. The section is concluded by observations on empirical evaluations and performance indicators.

2.5.1 General steganographic characteristics and the security of steganographic schemes

According to [Fridrich09], one of the most widely accepted text-book references on steganography, there exist **three different basic approaches to perform steganography**: In steganography by **cover selection**, the sender in the communication scenario has access to a set of different classes of media objects that can be used to establish the steganographic communication (covers) and both, the sender and the receiver, share a communication channel as well as a codebook that assigns a meaning to each class. In **steganography by cover synthesis** the sender creates the cover so that it conveys the desired message. In the third, and until today most studied steganography paradigm, the **steganography by**

[28] An open source library for support vector machines, available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

cover modification, the sender modifies cover objects to embed the message. The main difference between the first two and the last basic approach is that only the last one introduces (by the required modifications) potentially changes to the original source characteristics for the cover objects. Only this modifying approach to steganography, which is used by the majority of existing steganographic approaches [Fridrich09] is considered in this thesis.
Another basic differentiation between steganographic channels is discussed in [Gianvecchio07]. There, the authors distinguish **covert storage and covert timing channels**. Within this thesis only the first are considered. The latter type, focussing on the timing-behaviour of systems (e.g. the response times of a web server), is completely outside of the scope of this thesis.

Figure 2.7 shows the composition of the typical steganographic channel for steganography by cover modification.

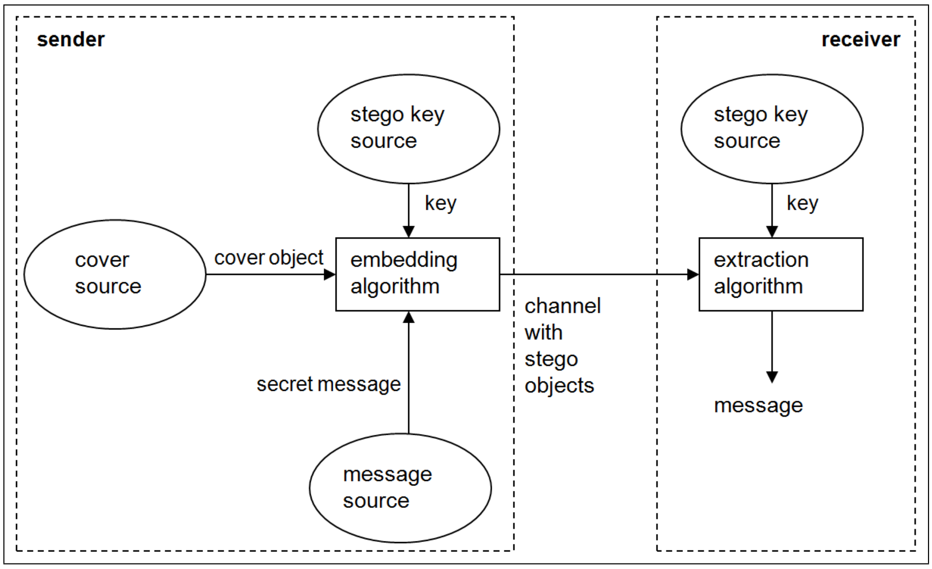


Figure 2.7: Elements of the steganographic channel for steganography by cover modification (based on [Fridrich09])

The two main functions in the steganographic channel model are the embedding and extraction functions. The first is usually parametrised with a cover (or set of covers), a stego key and the message(s) to be transmitted. It has to be made explicit here that in many practical steganography schemes/tools the messages are also encrypted prior to their embedding. The output of the embedding is then referred to as the stego object. The extraction function requires the appropriate key for retrieval and the stego object to extract the message. What is not shown in figure 2.7 is the fact that usually an attacker is modelled on the channel between the embedding and detection processes. This attacker, who is supposed to perform steganalysis, is assumed to work under compliance to Kerckhoffs' principle[29] and is either modelled as passive (simply reading and analysing the channel) or active (modifying the objects transmitted on the channel). This thesis is focussed on the passive attack scenario and here especially on cover-stego-attacks (for an analysis of the different attack scenarios, i.e. models for starting knowledge that can be applied, in steganalysis see e.g. [Franz00]).
A simple audio steganography channel modelling is introduced as part of the context modelling in specific audio signal generation application scenarios in section 2.3.2. More detailed considerations on steganographic channel modelling can be found in [Winkler11], where e.g. possible pre-processing operations for the embedding are integrated into the modelling.

According to [Fridrich98], there exist three major basic properties to describe information hiding algorithms based on cover modification: their detectability, the capacity and the robustness. The detectability is sometimes termed transparency or also security of a steganographic scheme. The capacity

[29] I.e. the attacker has complete knowledge of all algorithms/methods used for communication, the security of the scheme therefore relies only on the used key; see [Katzenbeisser00], [Kerckhoffs83].

describes the message size that can be embedded. In most cases this is done in relation to the cover size (e.g. in Bit per pixel). The robustness is describing the difficulty to remove hidden information from the stego object. These three properties are generally considered to be mutually competitive; it is not possible to optimise a scheme for all three properties at the same time. This competitiveness requires them to be optimised for each implementation of an application scenario.

For this thesis the detectability of a steganographic scheme is its most important property, since it is the one directly addressed by steganalysis. Nevertheless, due to the relationship between the three characteristics, the other two show also influence the process of steganalysis: A longer message (a higher capacity required) implies that more changes have to be made in the embedding, which in most cases is equivalent to a higher probability of detection. Also, an increase of the robustness of a scheme (e.g. by performing error-coding on the message) most often influences the statistical detectability, since it often increases the message length.

In a heuristic way the detectability can be described as the capability of resisting a steganalytical investigation, which is in many publications considered to be equivalent to the security of the steganographic scheme[30].

Theoretical considerations on steganographic security

There are two different approaches in literature how to **theoretically describe steganographic security**, the information-theoretic and the complexity-theoretic approach. Despite the fact that this thesis is focussed on the practical security of steganographic schemes, the two theoretical approaches are briefly summarised below and compared, based on the descriptions given in [Fridrich09]. In this comparison it is emphasised that the information-theoretic approach is by far the most widely accepted approach in the theoretical investigation on the security of steganographic schemes.

Information-theoretic approach to steganographic security: Based on the definition of information-theoretic security, it should be impossible for an attacker to design a steganalysis method for a perfectly secure steganography scheme that can reliably distinguish between cover and stego objects. Following the argumentation by Christian Cachin ([Cachin04]), the goal of steganography can be reformulated as constructing a steganographic scheme that assures that the distribution of the stego objects P_s is as close as possible to the distribution of the covers P_c as possible. Thus, it would be hard for an attacker to reliably decide whether an observed candidate object is drawn from the covers or from the stego objects, even if he has access to the steganographic scheme as postulated by Kerckhoffs' law. Therefore the distance between the P_c and P_s can be used as a measure for the security of a steganographic scheme. In [Cachin04] the Kullback-Leibler divergence [Kullback59] is used as a similarity measure between P_c and P_s.

If the Kullback-Leibler divergence is equal to zero the steganographic scheme is perfectly secure[31]. If the Kullback-Leibler divergence for a steganographic scheme is smaller than a value ϵ then the scheme is called ϵ-secure. If two steganographic schemes are compared in this approach, then the one with the smaller ϵ is to be considered more secure.

In [Cachin04] steganalysis is constructed as a hypothesis test[32] with the null-hypothesis being that a candidate object is an unmodified cover and the alternative hypothesis is that the candidate object is a stego object. This binary classification leads to four possible outcomes: true positive (TP) statements when a cover object is correctly identified as a cover object, true negative (TN) statements when a stego object is correctly identified as stego object, false positive (FP, a.k.a. statistical type I error or false alarm) when a cover object is wrongfully identified as a stego object, and false negative (FN, statistical type II error or missed detection) when a stego object is wrongfully identified as a cover.

[30] In [Fridrich09] one example is presented where the un-detectability is not equivalent to steganographic security. In this example the sender makes a digital photography of its secret message (steganography by cover synthesis) and attaches the image to an e-mail sends it to the recipient. Since the image is a plausible cover object for the source of senders images (its digital camera) an automatic steganalysis mechanism would not detect the presence of the message, while it would be obvious for a human warden.

[31] Perfectly secure steganographic systems are therefore most likely based on the basic steganographic approaches of cover synthesis or cover selection (where the stego objects follow P_c).

[32] Generally in literature on information theoretic approaches to steganalysis (like [Cachin04] or [Ker07b]) it is assumed that steganalysis is a binary problem, i.e. there exists only one embedding method and it is known.

Other publications, like [Ker07b] extend this information-theoretic approach from Cachin, e.g. by including considerations on the effect of long-term repeated steganalysis to the security of a steganographic scheme as well as considerations on the relationship between cover size and safe embedding capacity. Nevertheless, the feasibility of this information-theoretic approach to describing steganographic security strongly depends on the assumption that there is a reliable probabilistic model for the cover objects (i.e. for P_c). However, as shown in [Böhme08] it is for most media types (including audio) it still not clear how to describe and estimate P_c and P_s in practice. The second problem of the information-theoretic approach is the fact that it by definition ignores complexity issues. In this sense it is concerned only with the possibility of constructing an attack rather than its practical realisation.

Complexity-theoretic approach to steganographic security: To address the lack of considerations on the complexity of attacks to stego systems, i.e. the feasibility of such attacks, Hopper et al. [Hopper02] as well as Katzenbeisser and Petitcolas [Katzenbeisser02] introduced in 2002 independently from each other complexity-theoretic definitions of steganographic security. To summarise these approaches it can be said that they are based on two common principles:

First, the requirement to know P_c, established for the information-theoretic approach described above, is replaced by a much weaker assumption. For the complexity-theoretic approaches instead the availability of two oracles is assumed. The first of these oracles samples from the set of covers according to their distribution over the cover channel, the second oracle is returning stego objects generated by using an unknown key.
Second, the security of the steganographic system is then established by means of a probabilistic game between the warden in the prisoners' problem scenario and an external referee called 'judge'. The warden is enabled to train for some time using the cover oracle and is then asked by the judge to distinguish between outputs of the two oracles. The steganographic system is considered secure in the complexity-theoretic sense, if the detection accuracy of the warden minus the probability of guessing correctly in this scenario is close to zero and therefore negligible.

Practical considerations on steganographic security

Regarding the **practical security of steganographic schemes**, which is the part of the considerations on steganographic security that is relevant for this thesis, it can be assumed that it follows generally the same path like for symmetric cryptography[33]. There, schemes like the Data Encryption Standard (DES) or the Advanced Encryption Standard (AES) cannot be proven to be secure, but are considered to be so as long as nobody is able to produce an attack to these schemes that is faster than a brute-force search for the key and as long as they provide an adequate key space size. The design of these cryptographic methods is the result of a long-term cyclical interplay between cryptographers and cryptanalysts. Digital steganography cannot be considered as being a field of research that is as mature as digital cryptography, but nevertheless it takes the same path of development as its more mature counterpart[34]. Theoretical models, like the described information-theoretic and complexity-theoretic approaches to steganographic security, and existing practical attacks give the designers of steganographic systems ideas necessary to design new generations of algorithms or schemes (e.g. [Orsdemir08]).
Therefore, in practice the security of a steganographic scheme is often understood neither in the information-theoretic sense nor in the complexity-theoretic sense but rather as the inability to practically construct a reliable steganographic detector using existing attacks or modifications thereof [Fridrich09]. The approach in this thesis, to constructing a media security mechanism which is capable of performing steganalysis for audio material, is following exactly these practical considerations.
Regarding the evaluation of steganalytic approaches, similar to cryptanalysis, different attack scenarios are defined, based on the assumed starting knowledge for the attack. Here, one the most widely used attack scenarios are the so called 'cover-stego-attacks', which assume the analysis performing entity to

[33] The majority of the steganographic approaches existing today are of a symmetric nature. Asymmetric approaches like [Craver98], [Guillon02] or the Publimark tool (developed and maintained by G. Guelvouit, see http://www.gleguelv.org/soft/publimark/index.html) are rather uncommon.

[34] For example see the Break Our Steganographic Scheme (BOSS) open contest organised by Tomáš Pevný, Tomáš Filler and Patrick Bas between September 9th and December 15th, 2010 (see http://boss.gipsa-lab.grenoble-inp.fr).

have access to marked and unmarked versions of the same file for the training of the detector. This scenario is compliant with Kerckhoffs' principle which would allow the analyser access to the embedding algorithm so that he can use any key he wants to generate his own marked versions of any training set he would like to use. Other attack scenarios would be more restrictive in modelling the capabilities of the analyser. They are omitted here, because they violate the sound reasoning imposed by Kerckhoffs' principle.

Besides the attack scenarios, also different goals for the steganalysis can be defined. The most prominent of these goals is the detection of the hidden communication. Besides this goal sometimes very specific secondary goals are defined like the determination of the embedding strength or payload size (see e.g. [Ker04]). Also, application specific steganalysis goals might arise from specific steganography approaches, like e.g. information pooling in the in the case of Andrew Kers batch steganography and pooled steganalysis [Ker07a].

In [Fridrich09] two **main classes of steganalysis approaches** are identified. These two classes are the **statistical steganalysis** and the **system attacks**. The first, which is the class which is in the focus of this thesis, performs statistical analyses to detect the presence of a message embedded into a cover object. The second class, which is excluded from the considerations within this thesis, is looking for telltale information about the usage of steganography, which do not originate directly from the embedding process. An example for the latter class would be the usage of SARCs Steganography Analyzer Artifact Scanner (see http://www.sarc-wv.com/products/stegalyzeras/), which claims to be able to detect the download and installation of over 1000 steganography tools on Microsoft Windows machines.

Nissar et al. perform in [Nissar10] a slightly different distinction. They consider a classification of approaches into **statistical steganalysis** and **signature steganalysis**. The few existing examples that can be found in literature for the latter class are mostly limited to the detection of specific (in most cases older and less sophisticated) steganographic tools under rather specific embedding conditions. Examples for such signature steganalysis are approaches like the visual steganalysis by Westfeld and Pfitzmann [Westfeld99] or the small number of approaches summarised in [Nissar10]. Due to their specific focus and the associated low generisability, these signature steganalysis considerations are excluded from this thesis.

If the statistical analyses approach to steganography is considered, there can be two sub-classes identified: targeted (a.k.a. application specific) and blind (a.k.a. universal) steganalysis. The first is building on Kerckhoffs' principle, i.e. it assumes knowledge about the steganography algorithm (or at least the embedding strategy) that was potentially used. The blind steganalysis in contrast is building on the assumption that absolutely no knowledge about the potential usage of steganography is given. The communicating parties might or might not use steganography. Even if they were using steganography, the techniques used would be unknown to the steganalysis performing observer.

If statistical steganalysis is to be performed 'in the field', [Fridrich09] describes the corresponding setup as the process of "*forensic steganalysis*". This practical **forensic steganalysis process** is considered in [Fridrich09] to encompass the following six steps:

1. "*Identification of web sites, Internet nodes, or computers that should be analyzed for steganography.*"
2. "*Development of algorithms that can distinguish stego images from cover images.*"
3. "*Identification of the embedding mechanism,* [...]"
4. "*Determining the steganographic software.*"
5. "*Searching for the stego key and extracting the embedded data.*"
6. "*Deciphering the extracted data and obtaining the secret message (cryptanalysis).*"

In this process blind (step 2) as well as targeted steganalysis (steps 3 to 5) are integrated into the larger scheme. Besides the fact that this process model for practical steganalysis illustrates very well the actual complexity of the problem, it can be used to define **steganalytical success levels** that

could be achieved. Similar to cryptanalytic performance levels (like unconditionally secure vs. computationally secure cryptography) such success levels achievable with existing steganalysis approaches could be used in practice to model the security of a steganography algorithm. In early publications like [Katzenbeisser00], an attempt to perform steganalysis on a steganographic scheme was considered successful if the existence of hidden information can be detected. Unfortunately, this first step of the forensic steganalysis (or lowest steganalytical success level) is of very limited use under the consideration of the Daubert standard.
As highlighted in section 2.2, the 2011 amendment of rule 702 ("*Testimony by Experts*") of the U.S. Federal Rules of Evidence (FRE) states [U.S. Congress11]: "*A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the experts scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.*" These requirements, which are within this thesis translated into the evaluation criteria FREC0, FREC1, FREC2 and FREC3, imply that the mere detection of the presence of hidden information by steganalysis is of no relevance as long as it cannot directly be tied to the case. To do so, the entities involved into the steganographic communication with their roles (sender, receiver), the message content communicated as well as communication related metadata (e.g. time of the message embedding, time of the transmission, etc.) should be established. Therefore, acceptable success levels for steganalysis under Daubert (and FRE) considerations would have to include every step in the forensic steganalysis process described above. This would include, amongst others, the successful message extraction and decryption (if encrypted, which has to be assumed due to the fact that many modern steganographic tools include strong encryption).

This forensic steganography setup can also act as a basis for steganography or **steganalysis benchmarking**, by using the steganalytical success levels for the comparison of the performance of different steganography or steganalysis schemes. While for simple steganographic algorithms and appropriate steganalysis techniques it might be possible to successfully achieve all six steps of the forensic steganalysis process, sophisticated steganographic approaches will show extreme resilience against success. Besides the fact that the sixth step in the process might actually mean strong ciphers like AES or even unconditionally secure one-time pads, also other parts or this process would prove to be extremely burdensome. A good example would be the determination of the used **payload size** (depending on the implementation of the steganalysis process, this has to be performed either in step 2, 3, 5 or 6). Andrew Ker summarises the general consensus[35] in this field in [Ker07b] as: "*All a steganalysis method can ever hope to do is to detect changes in the cover, rather than the payload itself, and the existence of adaptive source-coding techniques (e.g. [Fridrich06]) mean that the number of changes is not necessarily proportional to the payload size.*"

Besides the forensic (or offline) setup, which is in the focus of this thesis, also an alternative approach to steganalysis has to be mentioned: in the so called **online setup** (or warden scenario), the steganalyser would be used to monitor a communication channel for the usage of steganography and in case a hidden channel is detected it would initiate an action, e.g. terminate the channel. Obviously, in this second approach different goals and requirements would have to be formulated. The goal would be limited to the detection of the presence of a steganographic pattern in blind or targeted steganalysis. The steps 3 to 6 from the forensic steganalysis process described above would not be necessary. Regarding the requirements, the approach would most likely have to be real-time capable, which make this approach also difficult to implement in practice.

Summarising the considerations made here on steganographic schemes and their security, it has to be highlighted that the basic approach of steganography by cover modification dominates this field in research and corresponding literature. The practical considerations on security of steganographic

[35] Despite this generalisation, there exist a number of scientific publications (e.g. [Ker04]) which perform something like quantitative steganalysis on simpler (i.e. non-adaptive) techniques to estimate the embedded payload size. It has to be stated here that this form of analysis is outside the scope of this thesis.

schemes, to be used within this thesis, are here a synonym for statistical detection and a successful performance in forensic steganalysis. A successful application of steganalysis as a forensic method under consideration of the Daubert criteria would require forensic process to be successfully completed, tying the outcome of the steganalysis directly to the case at hand.

2.5.2 Identification of the state-of-the-art in (audio) steganalysis

To understand the research work currently going on in steganalysis, again some additional light has to be shed on **steganographic research**. In the last few years there have been hundreds of scientific publications on digital steganalysis published every year. An recent statistic published in [Fridrich09] states that in the IEEE publications alone in 2008 more than 200 papers have been published that contain in their keywords either 'steganography' or 'steganalysis' – considering the facts that the market share of the IEEE on scientific publications in this field is not so overwhelming and that the search criteria applied do not include 'data hiding' the estimated number of unreported publications might be far higher. Another statistics presented in [Johnson08] and [Johnson98] reports for 2007 alone more than 450 newly released steganographic software applications or new versions of such software. Even though most of the steganography publications and tools focus on digital images (an estimated 56%, see [Fridrich09]) also digital audio material is considered by a fairly large share of approaches (an estimated 14.8%). When analysing this fast-growing research field, nearly all tools implement steganography by cover modification.

The number of scientific publications on steganography has grown so fast and numerous that by 2005 a new class of scientific publications in this field emerged in a large quantity: survey papers on steganography and steganalysis. While a majority of these publications like [Cheddad10] and [Li11] focus on image steganography (and steganalysis), others, like [Bandyopadhyay08], consider a wider range of media including audio. Some, like [Jayaram11], [Nosrati12] and [Santhi12], focus entirely on audio steganography. In [Meghanathan10] the authors claim to perform "*the first comprehensive survey*" on audio and video steganalysis algorithms.

Existing audio steganography approaches and tools

From the existing literature, the different **audio steganography approaches**[36] existing can be roughly grouped in three classes:

- Time-domain (or natural domain; see the considerations on time-domain signals in section 2.3.1) approaches
- Transform domain[37] approaches
- Application format specific approaches

In time-domain embedding, the audio signal is directly modified. The most common approach is least significant bit (LSB) modification, where only the last (least informative) bit is changed in each byte (see e.g. [Bender96], [Katzenbeisser00]). Another early example is the echo hiding approach, which adds an echo to the original signal to represent a bit from the secret message [Gruhl96]. In an original echo hiding system, an attenuated and delayed copy of the original signal is added to itself, whereby the actual delay determines the secret bit. For small delays, the human ear cannot distinguish between the original signal and an echo. [Erfani07] propose a method for a more transparent echo hiding, where the embedding transparency is investigated at the sender. This way, the sender can decide whether an echo can be introduced and with which parametrisation. Hence, the signal distortion is minimised.

[36] Approaches that simply append their data to an audio file like in the case of the tool Data Stash (produced by Skyjuice Software, Singapore; http://www.skyjuicesoftware.com/software/ds_info.html) are in this thesis not considered to be audio steganography approaches.

[37] Common transforms for audio material are: the discrete Fourier transform (DFT) which converts the signal in frequency coefficients and corresponding phases, the discrete cosine transform (DCT), which converts the signal by using a representation relying on two sets of frequency coefficients, and further, the discrete wavelet transform (DWT), which transforms the audio signal into multiple frequency bands, supplying detail and approximation coefficients.

Beside the echo, another important musical effect in time-domain is the reverberation. The scheme from Nian et al. [Nian06] uses reverberation characteristics to represent a bit from the secret message. In this approach two different artificial reverberation impulse responses, which are indistinguishable for a human listener, are used to represent the '0' and '1' bits of the transmitted secret message.

An example for transform domain embedding could be modifying the phase information after a DFT with a phase coding algorithm (e.g. Bender et al. [Bender96]; Kruus et al. [Kruus03]). After the modifications, all data has to be translated back to time-domain for transmission. Another approach for transform domain embedding is the spread spectrum approach (e.g. [Matsuoka06]). Here the main idea is to spread a representation of a steganographic signal with an originally low bandwidth over one or multiple larger frequency bands. In general, algorithms working in a transform domain are more complex than those working directly with the time-domain signal.

In the application format specific approaches, specific audio file or data stream format characteristics are exploited to perform the steganographic embedding. For example Tian et al. [Tian09] construct three different embedding schemes employing the characteristics of a G.729a codec, while Aoki in [Aoki08] implements a steganographic system which exploits the actual representation of audio samples in the G.711 μ-law signed linear Pulse-code modulation (PCM) (PCMU) technique.

Most curiously, the huge number of different audio steganography approaches presented in scientific papers does not manifest itself into tools. There exist only a very limited number of available audio steganography tools. An analysis of the two large web-repositories on steganography tools gives the following figures:

- http://www.stegoarchive.com (last updated in January 2006):
  - For Windows: 12 audio out of 50 (from these 12 are 4 commercial only)
  - For Java implementations: 1 out of 11,
  - For Mac 1 out of 3
  - For Linux 3 out of 23
- http://stegtools.jjtc.com/ (publicly updateable list of steganographic tools as of Nov. 14th, 2010):
  - 14 out of 105 tools are audio steganography schemes (6 out of these 14 are commercial products)

Existing audio steganalysis approaches and tools

Considering the diversity in audio steganographic schemes found in literature, one would assume that the **steganalysis approaches** as the counter-science would show the same diversity. But this assumption is wrong. Except from some very few candidates (e.g. Westfeld and Pfitzmanns steganalysis using visual attacks [Westfeld99]) nearly all perform statistical pattern recognition (SPR) – see the summary on audio steganalysis techniques presented in [Meghanathan10].
Furthermore, most publications on steganography and steganalysis (e.g. [Li11]) trivialise the latter into a simple two-class decision problem: either a data object is an unmodified cover or a stego object. The normal way in literature to tackle this decision problem is to use supervised classification, first, to train classifiers and, second, to compute the classification accuracies on known good (cover) and known bad (stego) samples in artificially constructed evaluation sets with known classes for all objects. It is true that such SPR-based approaches might be efficient for solving the steganalysis problem, but in practical application it is less trivial (due to the facts that more one steganography tool exists and that most of these tools allow for different parametrisations, influencing the pattern generated by the message embedding) and to achieve reliable and plausible results is much harder. Most recent publications on steganalysis (e.g. [Fridrich09]) acknowledge the need for formulating the steganalysis problem as a multi-class problem.

A further interesting point to be mentioned in the context of this thesis is the mismatch between research and development/application of (commercial) **tools** in the field of steganalysis. For other communication based threat scenarios in IT-security, like viruses/malware or email spam, a large range of commercial detectors is available. But in steganalysis, it contrast to the hundreds or even thousands of research publications focussing on this topic, only few open source or research demonstrator steganalysers are found together with an even smaller number of commercial **steganalysis detectors**. Most of the research demonstrators, like for example the Steg_IDS[38] written by Angela D. Orebaugh at George Mason University in Fairfax (Virginia, USA) are still lacking maturity.
Also, for those few commercial tools available the focus of application is in many cases not statistical steganalysis but instead on system attacks[39]. In the context on this thesis it has to be explicitly mentioned that the number of available tools for audio steganalysis is much smaller than for image steganalysis. In fact it is limited to a small number of research prototypes.

Existence and maintenance of standards and controls concerning the operation of methods

In general, **forensic steganalysis setups** would need to fulfil the criteria of the Daubert standard (see section 2.2). Here, a strong discrepancy can be seen in the state-of-the-art in this field: The standard requires that the existence and maintenance of standards and controls concerning the operation of methods. But yet such standards, or even a significant number of **field studies** for this field, are missing.

One of the extremely rare examples, where steganalysis is applied in large scale field evaluations and is reported upon in scientific publications, is the work of Niels Provos and Peter Honeyman in [Provos02]. In their paper, the authors criticise current state-of-the-art in steganalytical approaches at the point of time of their publication (like [Farid01] and [Fridrich00]) as being practically infeasible, due to faulty basic assumptions (two-class problem description and statistical overfitting to the training sets). In contrast to these publications Provos and Honeyman construct a multi-class SPR-based image steganalysis detector called Stegdetect. Each candidate image is considered to be member of one of four classes, either it is an unmodified cover image or it is the result of the application of one out of three different steganographic tools (JSteg, JPHide and OutGuess 0.13b) which have been amongst the state-of-the-art at this point of time. Stegdetect is then applied blindly (without knowledge about the true class) to two million images downloaded from eBay auctions and one million images obtained from USENET archives. As a result, Stegdetect implies that over 1% of all images seem to have been steganographically altered (mostly by JPHide) and therefore contain hidden messages. Based on these findings, Provos and Honeyman describe in [Provos02] also a second tool called Stegbreak for plausibility considerations, i.e. for verifying the existence of messages hidden by JPHide in the images identified by Stegdetect. Their verification approach is based on the assumption that at least some of the passwords used as embedding key for the steganographic embedding are weak passwords[40]. Based on this assumption, they implement for Stegbreak a dictionary attack using JPHide's retrieval function and large (about 1,800,000 words) multi-language dictionaries. This attack is applied to all images that have been flagged as stego objects by the statistical analyses in Stegdetect.

To verify the correctness of their tools, Provos and Honeyman insert tracer images into every Stegbreak job. As expected the dictionary attack finds the correct passwords for these tracer images. However, they do not find any single genuine hidden message. In their paper, they offer four possible interpretations of this result, either: a) there is no significant use of steganography on the internet, b) they have been analysing images from sources that are not used to carry steganographic content, c) nobody uses steganographic systems that could be found with their detector, or d) all users of steganographic systems carefully choose passwords that are not susceptible to dictionary attacks. Even though the result of this large scale investigation is negative, the methodology and concepts behind the work in [Provos02] are remarkable. Even more so, since they also perform throughput considerations (throughput for Stegdetect is given in Kilobit of images per seconds; the throughput for Stegbreak is given in words per second

[38] See: http://www.securityknox.com/Steg_project.pdf

[39] The SARC steganalyser (see http://www.sarc-wv.com/products/stegalyzeras/) claims to detect the download and installation of over 1126 steganography applications on Microsoft Windows computers.

[40] This reasoning is built on statistics presented for weak and strong passwords in [Klein90].

for the dictionary attack) for their analysis tool-chain, something that is also strongly amiss in most steganalysis publications.

Not exactly a field study, but lab studies (or closed-set experiments) to be mentioned in this context are [Kharrazi05] and [Kharrazi06]. These publications do not only present classification accuracies computed in image steganalysis benchmarking but also look into plausibility and complexity/throughput issues. With these additional considerations they are much closer to fulfilling the Daubert criteria as well as the necessities for practically applied steganalysis than most other publications in this field. In [Kharrazi06] Kharrazi et al. evaluate seven different steganographic embedding techniques against three established universal image steganalysis techniques on a cover data set of 100,000 randomly collected JPEG grey-scale images of medium image quality. In their investigations, which are relevant for the plausibility required in forensic application scenarios for steganalysis, they consider the effect of different image properties (size, texture and source) as well as post-processing operations (compression and re-compression) on the performance of steganalysis techniques. Regarding the throughput analysis, which would be important for an online setup for steganalysis, Kharrazi et al. present figures (in hours) for the embedding time required by the different steganographic algorithms as well as the classifier times (training plus testing) in cross-validation on a given test set size and a given machine.

As a third research effort to be mentioned in this context, the BOSS (Break Our Steganography System; see http://exile.felk.cvut.cz/boss/BOSSFinal/) was the first large scale, scientific and public challenge on image steganalysis. This challenge, which has to be considered to be somewhere in between a field study and a large lab experiment, was organised between September 2010 and January 2011 by Tomáš Pevný (University of Binghamton, USA), Tomáš Filler (Technical University of Prague, Czech Republic) and Patrick Bas (Centre National de la Recherche Scientifique in Lille, France). The provided material in the BOSS included: training databases containing cover- and stego images, the embedding algorithm as tool and as algorithm description and a test set of 1,000 images of which it was unknown whether they were cover or stego images. So far the setup is compliant to Kerckhoffs' principle and might be considered a plausible field study for steganalysis performance in application specific steganalysis. What reduces the BOSS into the category of an experiment under laboratory conditions are the facts that: the ratio between cover and stego objects was known to the attackers for the test set (the type of each image (cover or stego) was chosen according to a Bernoulli process with equal probabilities for cover and stego objects), the embedding rate was fixed for all stego objects (0.4 BPP) and that an oracle existed telling upon submission the accuracy[41] achieved. These conditions made the setup less realistic than the setup used by Provos and Honeyman in [Provos02]. In summary of this largely popular challenge, for which the log files have shown participation from 96 different researchers or research groups in the world, it has to be stated that the best detection accuracy achieved in this evenly distributed two-class classification setup has been 80.3% (see the official ranking on http://exile.felk.cvut.cz/boss/BOSSFinal/; for a detailed description of the best performing steganalysis approach see [Fridrich11]) – a performance which would without doubt lead to the exclusion by any judge of the used steganalysis result in forensic testimony.

In comparison to the three studies summarised above, the majority of the steganalysis research is much further away from the requirements on the existence and maintenance of standards and controls concerning the operation of methods. This is especially true for the field of audio steganalysis. This fact is strongly supported in [Meghanathan10] where the authors attribute the current immaturity of audio steganalysis to "*the existence of advanced audio steganography schemes and the very nature of audio signals to be high-capacity data streams*".

Figures on achieved detection performances

As mentioned above, most scientific publications trivialise steganalysis into a simple two-class decision problem and focus on reporting classification accuracies for supervised classification obtained under targeted steganalysis evaluation setups. The reported **detection performance** (i.e. in nearly all cases

[41] Including mechanisms to prevent the solving of the challenge by simple iterative guessing and checking.

the classification accuracy), achieved under laboratory conditions, is often close to 100%. On the other hand, there exist information theoretic proofs that perfectly secure (i.e. statistically undetectable) steganography systems can exist (see e.g. [Herrera-Joancomarti07], [Katzenbeisser00]). The necessary proofs are not repeated here, but it has to be stated that such a perfectly secure steganographic system is hardly practical since a steganalyser would become sceptical when intercepting the random messages required for establishing the security. These two contradicting facts – good performance in targeted steganalysis under laboratory conditions for practical steganography schemes on one hand and the proof for the potential existence of perfectly undetectable steganography on the other hand – show how far away steganalysis in general is from ever becoming a Daubert-conform media forensics approach.

Regarding the performance of steganalysis on terms of **throughput**, which would be sine qua non for any online application scenario for steganalysis but is also of importance for any forensic investigation[42], it has to be stated that except for a selected few (e.g. [Provos02], [Kharrazi06]) hardly any scientific publication looks into this factor. Nevertheless, some authors include considerations that aim directly or indirectly at the reduction the complexity of the classification task and thereby the improvement of the throughput. A good example for this class of publications is [Miche06] where the authors perform a feature selection in a statistical pattern recognition (SPR) based steganalysis approach.

Besides the tool-driven perspective on achievable detection performances, a completely different view on the plausibility of stenography and steganalysis is presented in [Givner-Forbes07]. Here, not the forensic considerations of the Daubert standard are in the focus of the evaluation of steganalytical method, but instead **instructions are given for potential end-users** on how to evaluate the actual security of existing steganographic tools. Following the instructions, it is simple to identify all tools that are not compliant with Kerckhoffs' principle. Furthermore, basic techniques are explained that allow estimating the perceptual and statistical impact of steganography by modification for steganographic tools. Also, the influence of strong encryption prior to embedding and other considerations to achieve a higher degree of communication security (i.e. making successful steganalysis much harder) are discussed.

Summarising the state-of-the-art in steganalysis as presented in this section, it has to be said that most of the work found in literature so far is limited to investigations on the performance against individual steganographic algorithms (i.e. steganalytical benchmarking of steganography approaches), not on considerations as a global forensic security mechanism that could be implemented and applied online as a tool or in forensic investigations. The majority of approaches are statistical pattern recognition (SPR) based and the complete research field of audio steganalysis is far away from showing any compliance to the Daubert standards. Additionally, it has to be mentioned again in this summary that, while there are virtually hundreds of scientific papers published on audio steganography and steganalysis, the number of actually available audio steganography and steganalysis tools is extremely small.

2.5.3 Principal methods and concepts employed in the state-of-the-art in audio steganalysis

For the different sets of goals for steganalysis (application specific steganalysis, universal audio steganalysis and steganography or steganalysis benchmarking) and the possible setups (online- as well as forensic) the following methodology considerations are currently made by the stat-of-the-art: The **patterns observed** in nearly all existing steganalysis approaches are the modifications made in the embedding process of steganography by cover modification schemes. As highlighted in section 2.5.2, the majority of the approaches are driven by statistical pattern recognition (SPR) based detectability analysis. Therefore, the **input** in most approaches is under lab conditions generated training and test sets of cover and stego objects. The only noticeable exception to be mentioned here is the work by Provos and Honeyman in [Provos02], where the test set is 'taken from the wild' instead of being an artificial construct (see section 2.5.2).

[42] In theory a forensic analysis might be unbound in time and invested effort, but in practical forensic investigations, e.g. by police investigators, the time and effort that can be allocated for one case is usually strongly limited.

For the statistical pattern recognition (SPR) based approaches, the evaluated **pre-processing** operations include in most cases windowing, due to the fact that most feature extractors include also transform domain features.
Some authors also use more sophisticated pre-processing to generate the supposedly 'unmarked' references required to build the classifier models. Examples to be mentioned here are the work of Xue-Min Ru et al. [Ru05] where the references are reconstructed via linear predictive coding, benefiting from the very nature of the continuous wave-based audio signals or from Özer et al. [Özer03] as well as Avcibas [Avcibas06] by using a de-noising functions. Besides the pre-processing required for these self-generated reference signal based approaches, some other approaches found in literature employ even more complex pre-processing operations. An example for such an approach is the work by Micah K. Johnson et al. [Johnson05], where the authors use the pre-processing to perform a dimensionality reduction of the signal by application of a principal component analysis in training to generate a low dimensional linear basis. In the pre-processing in testing the analysed signal is projected onto the linear basis.

Regarding **feature design** approaches that can be applied to audio steganalysis, two main approaches can be identified here: Intuition-based feature design, which to be successful requires expert knowledge on the domain covered, or the transfer of features from other (similar) problem domains.
For the intuition-based feature design, a good example would be the usage of LSB features to detect LSB steganography, like in [Dittmann05] where the impact of LSB modification-based features (LSB-ratio and LSB-change ratio) are used to implement a steganalysis system. Another interesting approach in this class is proposed in [Johnson05] where the root mean square errors between frequency-domain representations of the signal and their projections onto a linear basis formed by a principal component analysis (PCA) form an error distribution from which the first four statistical moments are used as feature vector.
To design features for one application field by transfer of concepts from other problem domains is also a rather common methodology. An example in audio steganalysis that should be mentioned here is the usage of Mel-Frequency Cepstral Coefficients (MFCC) based features in [Kraetzer07a]. These features originate in biometric speaker recognition and are transferred successfully from their original application domain to audio steganalysis. In case of the MFCCs, these features have been further developed and extended due to their good performance in audio steganalysis. Examples for feature sets derived from the original MFCCs are the band-pass filtered version (FMFCCs) introduced in [Kraetzer07a] and the 2nd order derivative MFCCs from [Liu09].
In general, the resulting features from these design approaches could be either local features, global features or segmental features (see section 2.4.2) with or without using higher-level content analysis. It has to be noted that for audio signals with their high data rate the usage of local features is rather uncommon. Instead segmental features are used in most cases.

Regardless of which feature design approach is applied and which kind of features is generated, subsequent **feature selection** (see section 2.4.3) should be used to validate the significance of all elements in the feature vector and, if necessary, eliminate insignificant features. Results of applied feature selection in steganalysis, like e.g. [Miche06], show that it is often possible to reduce the complexity of the classification problem at hand by feature reduction while at the same time maintaining the same or at least similar classification accuracies.

Regarding the resulting **template or model sizes**, the impact of the steganographic embedding is usually represented either as feature vectors extracted from the audio signal or as completely trained statistical model. The first case is more likely since it allows a dynamic adaptation of the detector (e.g. by adding/enrolling new steganographic algorithms). The latter case, the storage of trained models, is less flexible but would be faster in field application since the necessary and time consuming process of generating a classifier model from the feature vectors has already been performed. Especially in an online setup, the latter alternative would be much preferred due to the performance (throughput) requirements of such a setup. Here, a bank of detectors (one for each algorithm that might be used) trained as two-class detectors would be the most likely scenario.

If the feature vectors are used to represent an algorithm, then the number of feature vectors and their dimensionality define the size of the model size. If trained models are used for the representation of algorithms then the size is determined by the input used in the model generation process and the way the classifier model is represented/stored.
In general it has to be mentioned that the model sizes tend to be rather large (due to the high dimensionalities of the feature spaces used and the large number of windows that have to be considered for representing the wide range of possible audio signals sufficiently) despite no figures on the exact model sizes are presented in literature. An interesting exception from the large model sizes rule is presented in [Johnson05] where the size is reduced by the performed dimensionality reduction of the signal considered and the resulting extremely small feature vector dimensionality.

As mentioned in section 2.5.2, the majority of the steganalysis approaches found in literature follows the **pattern recognition (PR) approach** of statistical pattern recognition (SPR). Regarding the closely related question of the **classification approaches** to be used, there are two different schools of thought found in literature: modelling the steganalysis as a two-class classification problem, which is done by the majority of authors, or describing practically applied steganalysis as a multi-class classification problem. For the first approach for practical applications a set of specifically trained (i.e. application specific) steganalysis detectors would be operated in a parallel setup. The classifiers of choice for most publications that follow this trend are classical two-class classifiers like e.g. support vector machines (SVMs). For the latter (universal) approach a multi-class classifier is used e.g. in [Provos02] for assigning one input candidate to exactly one class out of a set of predefined classes (here steganographic tools/schemes). The multi-class approach is assumed to show problems if it comes to the scaling of this approach. Especially for a large number of classes (steganographic embedding schemes) to be distinguished, the high dimensionality of the feature space required to represent all these embedding impacts together with the large number of training samples for statistically significant models will make the model extremely complex. As a result the classification based on this model would be rather slow.
In contrast to such a complex model, a network of two-class classifiers might be easier to construct and maintain. It would feature the trends in modern computing machines, i.e. multi-processor / multi-core as well as parallel-, grid- and cloud-computing. Also the integration of new steganalysis tools to the repository of the steganalysis toolset would be much easier – in case of the multi-class approach such a scaling would result in the necessity to update the complex 'global' model. It is imaginable that future work on steganalysis might combine networks of two-class classifiers with multi-class classifiers, e.g. to use the first ones to identify the embedding method/domain (e.g. time-domain LSB replacement) and the latter ones to identify the actual tool that was used to embed the data.

When steganalysis is considered to be a forensic security mechanism solving a data integrity verification problem, it would have to undergo extensive empirical evaluations to meet the **Daubert requirements** (see section 2.2) connected to forensic investigations. Problems arise for steganalysis in this context especially from three major points of the Daubert factors: the number of existing empirical investigations, the error rates associated to the methods have to be known (precisely and plausibly) and the current lack of standards and controls concerning the (implementation and) operation of steganalytical methods.
Regarding the first and second of these points, so far only one large-scale 'in field' empirical investigation is known to the author for the entire field of steganalysis. Besides the investigation of Provos and Honeyman in [Provos02], which had to be concluded with a rather unsatisfying end (no steganographic message was successfully extracted from three million images under suspicion), hundreds of publications exist on detectability benchmarking for steganography algorithms, where detection accuracies close to 100% are presented under lab conditions. The **plausibility** of these steganographic detectability benchmarking approaches is so far not completely addressed. Only few publications so far (e.g. Kharrazi et al. in [Kharrazi06], see section 2.5.2) consider the impact of typical, non-malicious signal post-processing operations. An even smaller number of publications considers active countermeasures against steganalysis (as discussed e.g. in [Orsdemir08]) in their work.

Regarding the **performance indicators** required in the necessary empirical evaluations, so far the (classification or matching) accuracy dominates in the existing literature. It is used extensively in the supervised classifications performed in steganographic setups under lab conditions. Few authors use in their publications other metrics such as the precision and the ROC (the Receiver Operation Characteristic curve of a transmission system; see [Li11]), the AUR (area under ROC; see [Kharrazi06]), specific points on the ROC curve (see [Fridrich09]) or a combination of accuracy and false-positive rate [Johnson05].
The only existing scientifically published upon real life steganalysis [Provos02] has no suitable detection performance indicators for the performed 'in the wild' study, but analyses the throughput of the system. Similar throughput analyses are performed by Kharrazi et al. in [Kharrazi06], where the authors present figures (in hours) for the embedding time required by the different steganographic algorithms as well as the classifier times (combined and training and testing) in cross-validation on a given test set size and a given machine.

Summarising the principal methods and concepts employed in the state-of-the-art in audio steganalysis, is has to be mentioned that statistical pattern recognition (SPR) based two-class classification setups in closed-set experiments dominate the research in this field. Promising extensions to such setups (like work on information fusion) are as rare as the usage of advanced, comparable performance metrics or the considerations required for forensic setups (e.g. plausibility analyses), which are considered a necessarily within this thesis.

2.6 State-of-the-art in microphone forensics

This section summarises the state-of-the-art in the application scenario of microphone forensics. The existing alternative approaches are identified and their basic principles, prospects and constraints are described in detail. This is done to build the basis for the development of the methodology for this thesis in chapter 3 and to allow in section 6.5.2 for comparisons between the existing state-of-the-art and the newly developed approach for microphone forensics.
Parts of this section build upon basics introduced in section 2.2 and 2.4.

The **goals for microphone forensics**, as it is considered within this thesis, can be described best by the specification of the **addressed security aspects** and the **considered media data**: The addressed security aspects are: reliable, a-posteriori source authenticity and integrity verification[43] of (never compressed) PCM encoded audio recordings in CD quality, without side-information on the signal. This translates to a security mechanism solving the source authentication problem, which is the main task in microphone forensics.

The following overview over the state-of-the-art approaches for microphone forensics is generated based on four different types of activities:

1. An analysis of corresponding publications registered in the well established Digital Forensic Database (DFD)[44]

2. An analysis of the deliverables of the EU FP7 research project REWIND[45] (REVerse engineering of audio-Visual content Data)

3. A review of the few currently existing survey papers compiled by criminal investigators and researchers on this field, starting with early works like [Owen88], [Bijhold07] and [Maher10] (which still put a strong emphasis on analogue audio signals). The later surveys considered in the preparation of this overview of the state-of-the-art (e.g. [Brixen07], [Rumsey08], [Koenig09], [Tibbitts09],

[43] Here: Verification of consistency of global (i.e. recording source intrinsic and content independent) phenomena against audio stream composition from multiple sources.

[44] http://www.cs.dartmouth.edu/ farid/dfd/index.php/topics – maintained by Hany Farid at the Computer Science Department of Dartmouth College (Hanover, NH, USA)

[45] http://www.rewindproject.eu/

[Maher10] and [Gupta12]) mostly, or entirely, focus on forensic audio signal analysis on digital signals.

4. An intensive search on additional microphone forensics related publications not covered by the first three types

Currently[46], according to the information gathered from these sources, the existing approaches in this field can be classified into three classes: side-information based authenticity and integrity analyses using the electric network frequency (ENF) of a recording setup, content-based consistency analyses on local phenomena found in the time-domain representation of a recording and a microphone response based pattern recognition. In the following section 2.6.1 the current state-of-the-art approaches for these three classes are introduced in detail. The principal methodologies and basic concepts are compared in section 2.6.2. The structure of this section is based on the design of a pattern recognition (PR) pipeline for statistical pattern recognition (SPR) or template matching and contains considerations on: the patterns observed, the required input signals, pre-processing, feature design and selection, template or model sizes. The descriptions are made under the consideration of the most dominant pattern recognition approaches applied in this field and the different classification approaches. The descriptions are concluded by observations on the assessment under the Daubert criteria and performance indicators.

2.6.1 Identification of the state-of-the-art in microphone forensics

The oldest and most widely published-upon approach for microphone forensics is the **electric network frequency (ENF) based approach** [Grigoras03]. It is based on the realisation that, when digital equipment with an AC power supply (i.e. is not battery powered) is used to record an audio signal, the 50/60 Hz[47] ENF as well as its harmonics become part of the recorded signal. The reason for this is that digital equipment normally lacks ideal voltage regulators and perfect shielding.

The investigations in [Grigoras03] on the ENF show that is displays variations in the form of fluctuations of up to ± 0.6 Hz. Its spectrogram shows that the ENF with these fluctuations is a continuous function. The microphone (or rather digital recording setup) forensics approach in [Grigoras03] proposes to constantly measure the ENF within a power grid and store its development over time in a reference database. This registered ENF is then used as side information on the influence imposed to recording equipment by this electronic phenomenon and therefore for assessing the authenticity and integrity of digital audio/video evidence (see e.g. [Grigoras07]). Selected state-of-the-art publications on ENF-based microphone forensics are [Cooper08], [Grigoras05], [Grigoras07] and [Nicolalde09]. Newer publications like e.g. [Rodríguez10] shift the focus from authentication of audio signals towards precise integrity verification on the basis of ENF discontinuity analysis.

The complete electro-physical requirements for this approach are summarised in a corresponding European Network of Forensic Science Institutes (ENFSI) standard ([Grigoras09]). The core functionalities of this approach can be summarised as a pre-registration (template generation) for the ENF at the recording location, extraction of ENF-related features from the recorded signal and a correlation based template matching of the feature vector against the pre-registered ENF side-information.

However, some existing drawbacks of the approach are limiting its field application:

1. This approach does not work for the devices which use direct current (DC) power supply, such as typical handheld recording devices (either audio only or audio and video, like in video cameras) or mobile phones, which are powered by batteries.

2. In [Brixen08a] it is implied that certain types of microphones are immune to electromagnetic fields (and their changes). In combination with point 1 above, recordings by such microphones would not show any ENF traces (neither from the recording equipment nor picked-up from the surrounding environment).

[46] As of October 2012

[47] In most parts of the world either 50 or 60 Hz electric network frequencies are used, depending on the national regulations.

3. As described in [Grigoras09], the standard approach requires a precise database documenting ENF features at all possible recording locations for the authentication. Constructing such a database requires tremendous work, and as the ENF features might change due to even the slightest adjustment of the power supplying devices, the database would need constant update.

4. The database-based authentication works for devices powered by normal public power supply, yet it does not work if the recording location uses its own power source or uninterruptible power supply, which is not monitored by sensors connected to the forensic database.

5. It is shown in [Grigoras09] that the approach does not work well with the recorded evidence after lossy encoding, such as GSM or MP3 encoding.

6. The approach would work fine with common speech recordings as the ENF is usually outside of the bandwidth occupied in the signal spectrum by human speech, however if the signal has strong frequency components in the frequency band covering the ENF (e.g. background music), it is highly probable that the approach would fail.

7. It does not allow for the authentication of legacy content for which no ENF template exists.

For a more complete analysis of the ENF approach for source authentication the author refers to [Grigoras09], [Brixen08a] and [REW11].

The second class of approaches to be mentioned here is the **time-domain and local phenomena based evaluations**. In 2010 Malik and Farid [Malik10] described a technique to model and estimate the amount of reverberation in an audio recording by correlation based template matching. Because reverberation depends on the shape and composition of a room, differences in the estimated reverberation can be used in a forensic setting for location-based authentication (and/or integrity verification against composition of audio material from multiple source recordings). The computed consistency of the reverberation behaviour can be considered as a special kind of global feature, even though it is based on local phenomena. However, it might not be extractable in each recording, since it is strongly content dependant. Additionally, there exist a wide range of environments which do not display the required constant reverberation behaviour (e.g. any outdoor recording location as well as crowded places). Thus, the application of this approach is seriously limited, or as summarised by [Gupta12]: "*Currently, this measure has been successfully applied to synthesized audio with assumptions that cannot be fulfilled by most real-world signals. Thus, it needs to be generalized for a wider range of applications.*" Its application is further hindered by signal post-processing operations (e.g. blind de-reverberation) performed in many application scenarios, like audio / video conferencing, hands-free telephone, etc – see [REW11].

The third alternative approach is the **microphone response based pattern recognition (PR) approach**. In 2005, Oermann et al. indicated in their theoretical work [Oermann05] that an identification of a microphone as source of a recording might be possible, based on the observation that two different microphone cause noticeable differences in the recorded spectra of the same sound. It has to be mentioned here that the authors of [Oermann05] performed no practical implementation of their idea. A practical realisation and evaluation of this approach can be found in the work by Daniel Garcia-Romero and Carol Espy-Wilson in [Garcia-Romero10], where the authors develop a GMM based template matching to implement automatic acquisition device identification on speech recordings. The main motivation behind their source authentication work is the realisation that the determination of the microphone would improve the performance of speaker recognition approaches. In their evaluations with two sets each comprising of eight microphones[48] they achieve classification accuracies higher than 90 percent. If these results would be generalisable, this would allow for a reliable selection of microphone-specific speaker recognition models and thereby would assumedly result in an increase of the performance of the subsequent speaker recognition.
The most interesting part of the approach by Garcia-Romero et al. is the generation of the template for each microphone, which allows for an extremely compact and recording length independent representation of the microphone response based recording influences. The main drawback of this approach

[48] A low quality set of eight telephone handsets and a normal quality set of microphones.

is that, while it assumedly works quite well on speech signals, it is determined to fail on other, more complex audio signals. Publications like [Dufaux01] and [Moncrieff06] show (for foreground and background audio signals respectively) that, for audio complex signals, even complex GMM (with a high number of Gaussian components) do not succeed in adequately solving the high diversity of the signal. Another (but less severe) drawback of this approach is that it does not allow for integrity verification by any means due to the extreme information reduction in the template generation process.
Another practical realisation and evaluation of the microphone response based pattern recognition approach is the work of Malik and Miller in [Malik12]. In this paper the authors perform threshold based template matching using first- and higher-order statistics of estimated Hu moments to implement microphone authentication. Unfortunately, the paper lacks the necessary detail in the description of the setup of the performed practical investigations which does not allow the reader to speculate on the plausibility of the 100% detection accuracy reported for a test set of 8 microphones.

The author of this thesis is currently not aware of any further alternatives in literature addressing this research field. Nevertheless, besides these three existing classes of alternative approaches for microphone forensics, there exists scientific work in fields that can be considered to be closely related to microphone forensics, mainly in benchmarking and quality assurance, e.g. in microphone impulse response and distortion measurement (see for example [Farina00]). Nevertheless, none of this work is currently capable of authenticating individual recording sources / microphones. Therefore these publications are excluded from the scope of this thesis.

2.6.2 Principal methods and concepts employed in the state-of-the-art in microphone forensics

For solving the source authentication problem, the **patterns observed** in the recorded material by the three existing practical approaches introduced in section 2.6.1 are expressed explicitly as templates.

Regarding the pattern analysis it has to be stated that the input for all three classes of approaches is different:

- The electric network frequency (ENF) approach requires for the authentication an audio recording and the previously registered ENF templates. The ENF templates have to be extracted from the power grid by using a special purpose sensor. For integrity verification it is implied (but only recently evaluated to some extend, see e.g. [Rodríguez10]) that only the audio file would be required.

- For Malik and Farid's ([Malik10]) time-domain and local phenomena based evaluations only the audio file is required for integrity verification. For authentication the reverberation behaviour in the recording would have to be matched against a database of previously registered reverberation templates.

- For the microphone response based template matching approach by Garcia-Romero et al. ([Garcia-Romero10]) the forensic authentication requires at least recordings from the evaluated microphone for training (generation of the template) as well as recordings from a statistically significant number of other microphones, because the matching in their publication is not performed on a distance metric and threshold basis, but by using support vector machines (SVMs). Under ideal circumstances for the forensic authentication the investigator has access to the evaluated microphone and can make the reference recordings in a controlled environment. As already mentioned in section 2.6.1 no means of integrity verification are possible for this approach.
  In Malik and Miller [Malik12] a threshold-based template matching is performed. Since no 'world model' is created here, the approach would only need samples from the microphone to be verified and a suitable threshold. It would not necessarily require a significant number of samples from other microphones to perform its task.

The **pre-processing** currently used in microphone forensics is either content-based as in the approach of Malik and Farid (where the content is analysed for portions of the time-domain signal where the reverberation can be extracted) or content-insensitive like in the ENF-based approach or the microphone response based template matching approach by in [Garcia-Romero10]. For the ENF approaches the pre-processing is performed by band pass filtering (see [Brixen08b]). For the microphone response based template matching the pre-processing consists of a simple windowing (in [Garcia-Romero10] with Hamming windows of a length of 20ms with 50% overlap; in [Malik12] with Dirichlet windows of a length of 4s with 50% overlap).

Regarding **feature design approaches** that can be applied to microphone forensics, two main approaches can be identified here: Intuition-based feature design, which to be successful requires expert knowledge on the domain covered, or the transfer of features from other (similar) problem domains.
For the intuition-based feature design, a good example is the correlation-based feature of Malik and Farid [Malik10] use for their approach. It is computing the consistency of local phenomena (here the reverberation behaviour) within one file.
To design features for one application field by transfer of concepts from other problem domains is also a rather common methodology. Examples in microphone forensics can be seen in the ENF-based approach, where the used global feature (the consistency of the ENF artefacts extracted from a recording with the original ENF for the recording position and time supplied to the verification as side information) originates from evaluations on electromagnetic compatibility and sampling problems in electrical engineering. Another good example for the feature transfer from one domain to another is the usage of the GMMs by Garcia-Romero et al. in [Garcia-Romero10], where the concept originates in biometric speaker recognition. In the work of Malik and Miller ([Malik12]) the used scale invariant Hu moments are transferred to this problem from the domain of image analysis.
The resulting features from the ENF approach are local features (ENF value at a given sample point). The features for Malik and Farid's approach from [Malik10] can be considered on two different logical levels: either, as a special form of segment-wise computed complex feature using higher-level content analysis (the reverberation behaviour in portion of the time-domain signal where the reverberation can be extracted), or as a global feature (a typical reverberation behaviour for a complete audio file and its consistency for this file). For the approaches of Garcia-Romero et al. ([Garcia-Romero10]) and Malik and Miller ([Malik12]) the features are first computed segment-wise and then merged into one global feature template.

Regardless of which feature design approach is applied and which kind of features is generated, subsequent **feature selection** (see section 2.4.3) should be used to validate the significance of all elements in the feature vector and, if necessary, eliminate insignificant features. For the ENF approaches such a feature selection is performed in [Brixen08b] where the author has to summarise the usability of the ENF features as: "*In praxis, approximately 40-60% of the digital recordings in question contain traceable ENF*". For the time-domain and local phenomena based evaluations of Malik and Farid in [Malik10], feature selection operations are not published jet, but it has to be assumed that the features show similar usefulness as the estimation of Eddy B. Brixen on the ENF feature. In [Garcia-Romero10] Garcia-Romero et al. perform for the intermediate features a wrapper-based selection on different speech recording sets to show that their approach of using GMMs and the choices made in implementation and parametrisation are justified for this kind of audio material and the evaluated microphones. In [Malik12] the authors show scatter plots of their features and compare these visually, deriving the statement that: "*It can be observed* [...] *that there are significant inter- as well as intraclass variations* [...]"

Regarding the resulting **template or model sizes**, Grigoras reports in [Grigoras05] a template size of 7384.68 MB for 365 days per monitored power grid. The template here scales linearly with the recorded time. For the Malik and Farid approach in [Malik10], no exact figures exist for the template size, but the template is assumed to consist of several kilobytes of audio signal per microphone/recoding location (i.e. reverberation pattern) combination – under the assumption that the combination displays a constant reverberation behaviour (see section 2.6.1). For the microphone response based template matching approach by Garcia-Romero et al. no precise template size is mentioned in [Garcia-Romero10], but it is of a fixed length ("*This procedure results in a fixed-length template to represent variable-length speech recordings.*" [Garcia-Romero10]) and is has to be assumed that it is relatively small, since it only depends on the number of used Gaussian components. For [Malik12] also the template sizes are assumed to be rather small, due to the high abstraction of the audio signal.

Based on the observed pattern and the feature vectors used, a suitable **pattern recognition (PR)** approach has to be chosen. In the field of microphone forensics, all existing practical approaches introduced in section 2.6.1 are implementing template matching. All use characteristics which must be clearly separable from the recorded content for these approaches to work (see the descriptions on the content limitations for the three approaches in section 2.6.1).

Closely following the question on the used pattern recognition (PR) approach is the question about the **classification approaches** to be used. In general, the main problem in microphone forensics is source authentication. Therefore, it is a classical multi-class classification problem assigning one input candidate to exactly one class out of a set of predefined classes. For the template matching based approaches in microphone forensics, this multi-class classification problem is either solved fast and efficiently by correlation-/distance-based classification ([Malik10]), by distance-based multiple hypothesis testing [Malik12], or, in case of the work of Garcia-Romero et al. ([Garcia-Romero10]), by using a matrix of fast linear SVMs with one model for each partition.

Regarding the **performance indicators** required in empirical evaluation, so far the (classification or matching) accuracy dominates in the existing literature. It is used in the supervised classifications/matchings performed in the ENF analyses and in the evaluations by Garcia-Romero et al. ([Garcia-Romero10]) as well as in [Malik12]. Malik and Farid in [Malik10] do not use any performance indicator, since this publication does not perform any empirical evaluations on their concept.

For the work on the ENF approach considerations on its **plausibility** are found in the existing literature – most prominently published by E.B. Brixen (see e.g. [Brixen08b] and [Brixen08a]). For the other two classes of approaches presented above, the usage of plausibility indicators is completely missing in the literature. Common audio signal post-processing operations (like normalisation, blind de-reverberation, etc.) as acknowledged by [REW11] are completely ignored in the evaluations performed. Due to the inherent insensitivity of template matching approaches to counter-forensics or anti-forensics methods, it has to be assumed that all these approaches are furthermore very easily affected by such targeted attack methods.

Regardless of the approach chosen to implement microphone forensics since the target output is a forensic security mechanism solving a source authentication problem, it should undergo an **assessment under the Daubert criteria** as they are summarised in section 2.2. Table 2.1 shows the authors assessment of the existing approaches introduced above for the state-of-the-art in this field.

Table 2.1: Using the Daubert criteria (see section 2.2) for assessment of the existing microphone forensics approaches

	electric network frequency (ENF)	**time-domain local phenomena (reverberations)**	**microphone response based pattern recognition**	
	Initial publication: [Grigoras03]	Initial publication: [Malik10]	Initial publication: [Garcia-Romero10]	Initial publication: [Malik12]
FREC0: the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue	criteria, that cannot be answered in general, because they are related to the specific court case under consideration			
FREC1: the investigation is based upon sufficient facts or data				
FREC2: the investigation is based upon reliable principles and methods, preferably scientific methodology and knowledge	Rather mature, considered in [Bijhold07]	Only a concept not tested	Pattern recognition / template matching	Pattern recognition / template matching
FREC3: the methods are reliably applied to the facts at hand	criterion, that cannot be answered in general, because it is related to the specific court case under consideration			
DC1: "*whether the expert's technique or theory can be or has been tested*"	yes, large scale tests (see e.g. [Grigoras05])	no	yes, limited closed-set experiments (two sets with eight microphones each)	yes, limited closed-set experiments (one set with eight microphones)
DC2: "*whether the technique or theory has been subject to peer review and publication*"	Publication count: >20	Publication count: 1	Publication count: 1	Publication count: 1
DC3: "*the known or potential rate of error of the technique or theory when applied*"	for ideal circumstances the error rate for authentication is known, for integrity verification it is not known under non-ideal circumstances (see e.g. [Brixen08b]) or even under the assumption of counter-forensics the rates are not known	no	only for two small sets and under ideal (speech only) circumstances	only for a small set and under ideal (environmental noise generated with a 12-inch fan only) circumstances
DC4: "*the existence and maintenance of standards and controls*"	European Network of Forensic Science Institutes (ENFSI) Forensic Speech and Audio Analysis Working Group (FSAAWG) guidelines on ENF analysis in forensic authentication of digital evidence [Grigoras09]	no	no	no

Continued on Next Page. . .

Table 2.1 – Continued

	electric network frequency (ENF)	**time-domain local phenomena (reverberations)**	**microphone response based pattern recognition**	
DC5: "*whether the technique or theory has been generally accepted in the scientific community*"	supporting arguments: large number of publications and citations, applied also for video recordings, document [Bijhold07] compiled by forensic experts from different police forces for an INTERPOL Forensic Science Symposium – opposing arguments: context dependency, does not work for DC powered devices, [Brixen08b]: "*In praxis, approximately 40-60% of the digital recordings in question contain traceable ENF*"	supporting arguments: none known – opposing arguments: context dependency, [Gupta12]: "*Currently, this measure has been successfully applied to synthesized audio with assumptions that cannot be fulfilled by most real-world signals*". [REW11]: hindered by common signal post-processing operations (e.g. blind de-reverberation) performed in many application scenarios, like audio / video conferencing, hands-free telephone, etc	supporting arguments: none known – opposing arguments: context dependency (speech only)	supporting arguments: none known – opposing arguments: only tested with one kind of recording content (environmental noise generated with a 12-inch fan)

Regarding three of the four criteria that are derived directly from the Federal Rules of Evidence (FRE) rule 702 (FREC0, FREC1 and FREC3) no statement can be given here, since these are investigation related criteria, that cannot be answered in general. For FREC2 it can be stated that only the ENF method has currently reached a degree of maturity that allows assigning a certain reliability. For this reason it is the only method for the authentication of digital audio material considered in [Bijhold07].

For the first of the criteria that is derived from the Daubert standard (DC1) it has to be said that only the ENF approach so far underwent substantial empirical testing (see table 2.1).
All approaches have been published upon (Daubert criterion DC2) in reviewed conference proceedings or journals, but only for the ENF approach so far a larger number of scientific publications by different authors from different research organisations exist.
Regarding the known or potential rate of error of the technique or theory when applied (DC3) the following has to be stated for the approaches:

- For authentication purposes the accuracy achieved in the experiments performed by Grigoras on the ENF approach is close to 100%, even on large test sets. For the integrity verification no figures are given, yet. What has been neglected in the empirical evaluations on this approach so far is an analysis of its content dependency, especially on the fact whether the ENF can be extracted from every kind of recording. This has to be doubted for music, because the frequency band in which the ENF is present would in this case also contain part of the recorded content (see e.g. [Brixen08b]). Also common audio signal post-processing operations like de-noising or MP3-conversion are assumed to disable any ENF based authentication or integrity verification attempt.

- The approach of Malik and Farid [Malik10] is at the point of the submission of this thesis only a concept lacking any empirical evaluations. Here evaluations on the context dependency as well as the required environmental conditions for a stable/usable reverberation behaviour would be necessary. This approach is assumed to be robust against simple audio signal post-processing operations like de-noising or MP3-conversion (although this is not explicitly mentioned in [Malik10].

- The approach by Garcia-Romero et al. in [Garcia-Romero10] is evaluated on two sets of eight microphones each and shows on these test sets an accuracy better that 90%. Nevertheless, here also investigations on the context dependency seem necessary. The same holds true for the approach presented by Malik and Miller in [Malik12].

The question about the existence and maintenance of standards and controls (DC4) can only be answered positively for the ENF approach. Here, the performed evaluations and the applicability of this approach are substantiated by the European Network of Forensic Science Institutes (ENFSI) Forensic Speech and Audio Analysis Working Group (FSAAWG) guidelines on ENF analysis in forensic authentication of digital evidence [Grigoras09].

For the hard to assess DC5 (the acceptance in the scientific community), only supporting and opposing arguments can be summarised here. For the ENF approach there exist a lot of supporting arguments (the large number of publications form different authors, the large number of citations – amongst others in [Malik10], the fact that it is also applied for video recording forensics, the document [Bijhold07] compiled by forensic experts from different police forces for an INTERPOL Forensic Science Symposium). Although it is hard to find explicit formulations on opposing arguments in existing literature, it has to be said that the ENF approach is constricted by the fact that it does not work for DC powered devices and its (recording) content dependency (both factors summarised in [Brixen08b]: "*In praxis, approximately 40-60% of the digital recordings in question contain traceable ENF*"). In general E. B. Brixen is performing rather critical analyses on the applicability of ENF based authentication in his publications (see e.g. [Brixen08b] or [Brixen08a]). Regarding the general approach of Malik and Farid (as presented in [Malik10]) several publications like [Gupta12] or [REW11] propose opposing arguments (see table 2.1). For the microphone response based pattern recognition approaches, neither supporting nor opposing arguments can be identified so far in literature. Nevertheless, it has to be mentioned here that they are assumedly also strongly affected by their (recording) content dependency (see section 2.6.1).

3 Methodology and Concepts

This core chapter of this thesis presents on one hand the methodology and on the other hand the concepts for the introduced statistical pattern recognition (SPR) based general-purpose approach for audio forensics. The methodology is considered in this thesis to encompass the principles of methods, rules, and postulates employed by a discipline. From these abstract basic principles the concepts – abstract ideas for solutions – are derived. While a common methodology for addressing both application scenarios is presented, the concepts have to discuss the required application scenario specific adaptations for the authentication and integrity verification work in audio steganalysis and microphone forensics. The theoretical considerations made in this chapter are used as basis for the development of evaluation designs – the practical (solution) plans – in chapter 4.

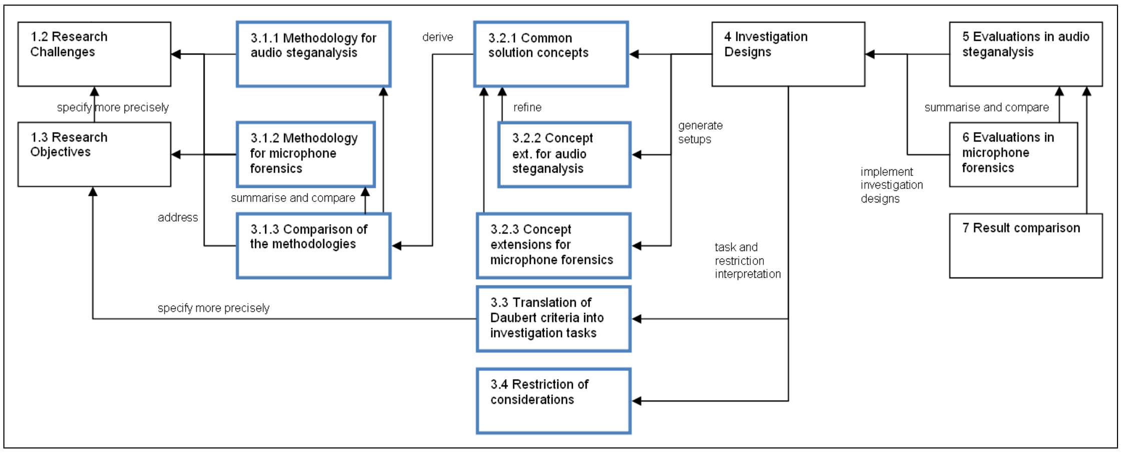

1.2 Research Challenges
3.1.1 Methodology for audio steganalysis
derive
3.2.1 Common solution concepts
4 Investigation Designs
5 Evaluations in audio steganalysis
specify more precisely
refine
generate setups
summarise and compare
1.3 Research Objectives
3.1.2 Methodology for microphone forensics
3.2.2 Concept ext. for audio steganalysis
implement investigation designs
6 Evaluations in microphone forensics
summarise and compare
address
3.1.3 Comparison of the methodologies
3.2.3 Concept extensions for microphone forensics
task and restriction interpretation
7 Result comparison
specify more precisely
3.3 Translation of Daubert criteria into investigation tasks
3.4 Restriction of considerations

Figure 3.1: Integration of the content of chapter 3 into the thesis context

Figure 3.1 shows a rough sketch of the linking between this chapter and the rest of the thesis. Besides the fact that obvious parts are missing (e.g. the fundamentals chapter 2, which is relevant for every other part of the thesis, the summarising chapter 8 and the appendices) also a low level of detail is used in the display and naming of chapters as well as sections. The intention here is to summarise the importance of this chapter within the thesis context:

- First, based on the research objectives specified in section 1.3, the **abstract methodology** (the analysis of the principles of methods, rules, and postulates employed by a discipline) for answering those research objectives is presented in section 3.1. This is done by combining the methodologies employed in the corresponding state-of-the-art for both considered application scenarios with new constructs required for addressing the research objectives defined in section 1.3.

- Second, based on the general methodology, a **concept**, an abstract idea for answering the questions raised by the research objectives (and the research challenges they build upon – see section 1.2), is postulated in section 3.2. Here, a common concept for both application scenarios is introduced in section 3.2.1, before application scenario specific concept modifications are discussed for audio steganalysis (section 3.1.1) and microphone forensics (section 3.1.2).

The analysis of methodology and concepts presented here, is strongly motivated by the Daubert criteria FREC2 ("*the investigation is based upon reliable principles and methods, preferably scientific methodology and knowledge*" adapted from [U.S. Congress11], see section 2.2) and DC1, which is summarised in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*".

For **audio steganalysis**, most of the work in the state-of-the-art is SPR-based, but as stated in section 2.5.3 the existing research still lacks currently the degree of maturity that would be required to pass a Daubert hearing as a forensic method. This motivates the effort invested within this thesis into the investigations on audio steganalysis, hoping that this effort might bring the whole field of research closer to that goal.

To draw a resume on the state-of-the-art in **microphone forensics**, as it was presented in section 2.6, regarding the applicable methods, all different alternative methods that currently exist in scientific literature implement some form of template matching. By their nature, these existing approaches suffer from severe limitations mostly regarding the availability and extractability of the required features (see section 2.6.1 on these limitations). As a consequence, within this thesis a new, SPR-based approach for microphone forensics is introduced, which is more generally applicable, i.e. which is less affected by constraints like the existing approaches.

- As the third part in this chapter, in section 3.3 the research objectives presented in section 1.3 are specified more precisely into **investigation tasks** for practical investigations. This is done by using the criteria of the Daubert standard (as discussed in section 2.2).
- In the fourth and last part of this chapter, in section 3.4 the **scope of the methodological and conceptual considerations is restricted**. This delimitation of the work performed has to be done because, even within the context of a PhD thesis, not every aspect of a complex solution strategy can be developed to cover every possible detail.

The methodology and concepts developed in this chapter are resulting in chapter 4 in tangible design layouts for the investigations including precise descriptions of evaluation setups and used parametrisations.

3.1 Methodology for the general-purpose audio forensics statistical pattern recognition approach

This section is dedicated to the analysis of the principles of methods, rules, and postulates employed by the discipline of statistical pattern recognition to solve the two chosen forensic application scenarios. Therefore, this section focuses on the methodology behind the investigations that are made in this thesis to address the research questions formulated for a general purpose approach addressing both application scenarios.

The state-of-the-art for both application scenarios, as it is presented in sections 2.5 and 2.6, is reconsidered and compared here, to act as the basis for the development of concepts for the investigations in section 3.2. The purpose of this is to address two main scientific points of the Daubert standards: first, the fact that all forensic methods have to be derived from scientific methods and second, to pave the way for the concepts and designs in the following sections, which aim to establish a good picture on the important question of the error rates that have to be associated with the methods.

3.1.1 Introduction of the principles for audio steganalysis applied in this thesis

The work on steganalysis presented in this thesis is focussing on the most prominent basic approach for performing steganography, i.e. steganography by cover modification. Furthermore, audio signal based storage channels are the only covert channel type considered. These specifications are typical for research on audio steganalysis[49].

[49] Audio steganography is usually considering audio files and therefore automatically excludes covert timing channels. Steganography by cover synthesis or cover selection are rather uncommon in this field.

The authors of [Nissar10] point out that there exist only two possible **solution alternatives for steganalysis** under these assumptions: what they call "*signature steganalysis*" and "*statistical steganalysis*". While signature steganalysis might work in case of very poor information hiding algorithms (like Data Stash, see below), it will fail against sophisticated, Kerckhoffs-compliant, context-adaptive IH approaches. Nissar et al. underline this realisation in [Nissar10] by stating: "*From the knowledge of the methods reported in this paper we infer that statistical steganalysis techniques, in any domain, are more robust and give promising results than signature steganalysis.*" Also, the system attacks described in [Fridrich09] are excluded from the considerations within this thesis, being side-channel attacks that aim on avoiding an analysis of the audio signals.
Therefore, statistical steganalysis remains the **only plausible alternative for the steganalysis considerations within this thesis**.
As highlighted in section 2.5, this realisation is consistent with the majority of the current research work in audio steganalysis. The most widely used approach for implementing statistical steganalysis is actually the statistical pattern recognition (SPR) also considered within this thesis. It allows the implicit description of the statistical descriptions of the classes to be distinguished (here, most likely unmodified covers and steganograms) by providing corresponding training samples as representative candidates for each class. The alternative here would be an explicit specification of the class statistics (e.g. via explicit, mathematical specification of the PDFs for each class) but this is uncommon because it would require an extremely precise domain-knowledge for the signals considered for analysis, which hardly exists for any any media format. Other PR approaches besides SPR (like e.g. template matching, see section 2.4) are unlikely to succeed for audio steganalysis, due to the high complexity of the cover signal and the assumedly small embedding impact.
In case the assumption made above that audio files are the signal under analysis would be changed to the case of audio data streams, an possible alternative might arise with anomaly detection from the field of change detection. This approach, which would also be feature-based like SPR, might be used to 'learn' the typical channel characteristics for an audio stream and, given suitable features, detect deviations (anomalies), like onset of an steganographic embedding, in the stream. This process would be similar to an one-class classification approach and would have the advantage that it does not require an explicit or implicit description of the possible anomalies. The downside of this idea is again the highly dynamic nature of audio signals in combination with the assumedly small embedding impact, which assumedly makes this approach also implausible for practical steganalysis.

Regarding the **performance metrics**, it is implied in the analysis of the state-of-the-art performed in section 2.5.3 that so far the classification accuracy dominates in the existing literature. This choice suites the pre-dominant two-class setups for steganalysis benchmarking, but is unsuitable for the multi-class setups made necessary by the forensic steganalysis process. In this thesis, with the Kappa statistics, a more suitable metric is introduced to steganalysis, which allows for a better comparison of classification results achieved on problems of different (class-)dimensionality.

Following the requirements established for **forensic compliance** in the Daubert standard (see section 2.2), in an ideal application of steganalysis as a forensic method all organisational, personnel related and technical issues would be addressed. This means that all evaluation criteria (FREC0, FREC1, FREC2 and FREC3 as well as DC1, DC2, DC3, DC4 and DC5) defined in this thesis on basis of the Daubert criteria and the requirements of the Federal Rules of Evidence (FRE) rule 702 would be addresses. It has to be stated here that the criteria FREC0, FREC1 and FREC3 cannot be addressed within this thesis, since they have to be decided by a judge on a per-case basis. Instead the focus within this thesis is limited to the criteria that aim to establish whether steganalysis is the reliable product of sound 'scientific methodology'.
In contrast to the application scenario of microphone forensics, where the forensic nature of the methods concerned is evident, for steganalysis the forensic application is still a rather neglected perspective. The small number of publications on forensic setups for steganalysis (e.g. the older Provos and Honeyman [Provos02] and the newer [Fridrich09]) is still by far outnumbered by publications performing steganalysis benchmarking (i.e. simply measuring the achieved detection performance under lab conditions).

In this thesis, the forensic application of steganalysis is the main focus of the considerations. To develop a suitable concept for this forensic application, below the ideal forensic audio steganalysis process is idealised as a Daubert standard conform, forensic process model for steganalysis. This generalisation is performed on basis of the forensic steganalysis process from [Fridrich09] (see the summary on this process in section 2.5.1). Since this ideal process turns out to be highly infeasible in practice, is then replaced for the considerations in this thesis by a concept for the construction of practical steganalysis processes under consideration of the Daubert criteria.

The ideal forensic audio steganalysis process

Based on the forensic steganalysis process from [Fridrich09], in figure 3.2 the audio steganalysis process is modelled as an ideal Daubert standard conform, forensic process.

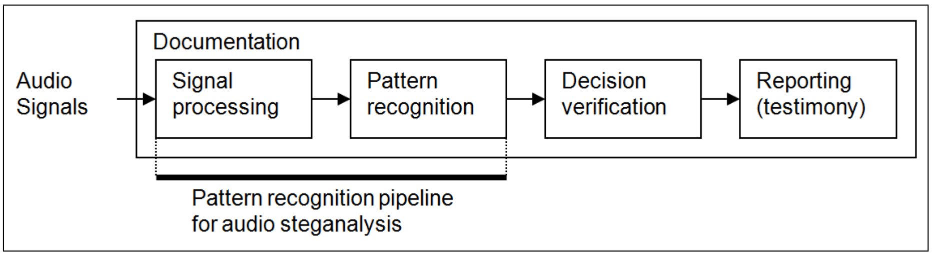


Figure 3.2: Audio steganalysis as an idealised, potentially standard conform, pattern recognition driven, forensic process

In this ideal forensic audio steganalysis process the audio signals would undergo steganalysis in a pattern recognition pipeline, followed by decision verification and the required reporting (e.g. in form of an expert testimony in court). The complete process would be accompanied by a documentation which would have to fulfil all requirements to forensic documentations (completeness, chain of custody, reproducibility, etc.).

A more detailed version of the signal processing, pattern recognition and decision verification phases in this idealised process is presented in figure 3.3. In this figure the pattern recognition pipeline for audio steganalysis shown in figure 3.2 is split into its two components, the training phase and the application (or testing) phase. The naive assumption on the relationship between these two phases would be that they are performed strictly sequentially: first the steganalyser models are trained and then they are applied in field investigations. In a less naive setup the candidate audio signals to be examined in field application would be first analysed for genre, quality and other semantic properties. The result of this analysis would be used to perform a content based selection of the audio material to be used for training. This potential relationship between the candidate audio signals and the training selection is shown in figure 3.3 with a dashed line.

The decision verification phase serves the purpose of tying the result of the steganalysis to the case at hand (see section 2.5.1). If this connection cannot be achieved, the judge would have to dismiss the expert testimony based on the steganalysis findings as not relevant for the case. Therefore success in this phase would be relevant to achieve acceptable success levels for steganalysis under Daubert (and Federal Rules of Evidence rule 702) considerations.

A projection of the six steps of the forensic steganalysis process from [Fridrich09] (see the summary on this process in section 2.5.1) to figure 3.3 would result in the following mapping: Steps 1 to 4 (i.e. identification of the observed channel, blind universal steganalysis, identification of the embedding strategy, and determination of the steganographic software) would be covered by the training and application phase. In practice they would be implemented by a network of consecutively executed pattern recognition pipelines, each focusing on specific tasks, like the identification of the embedding mechanism or the identification of the software used. The steps 5 and 6 ([Fridrich09]: "*Searching for the stego key and extracting the embedded data*" and "*Deciphering the extracted data and obtaining the secret message (cryptanalysis)*") be performed in the verification phase.

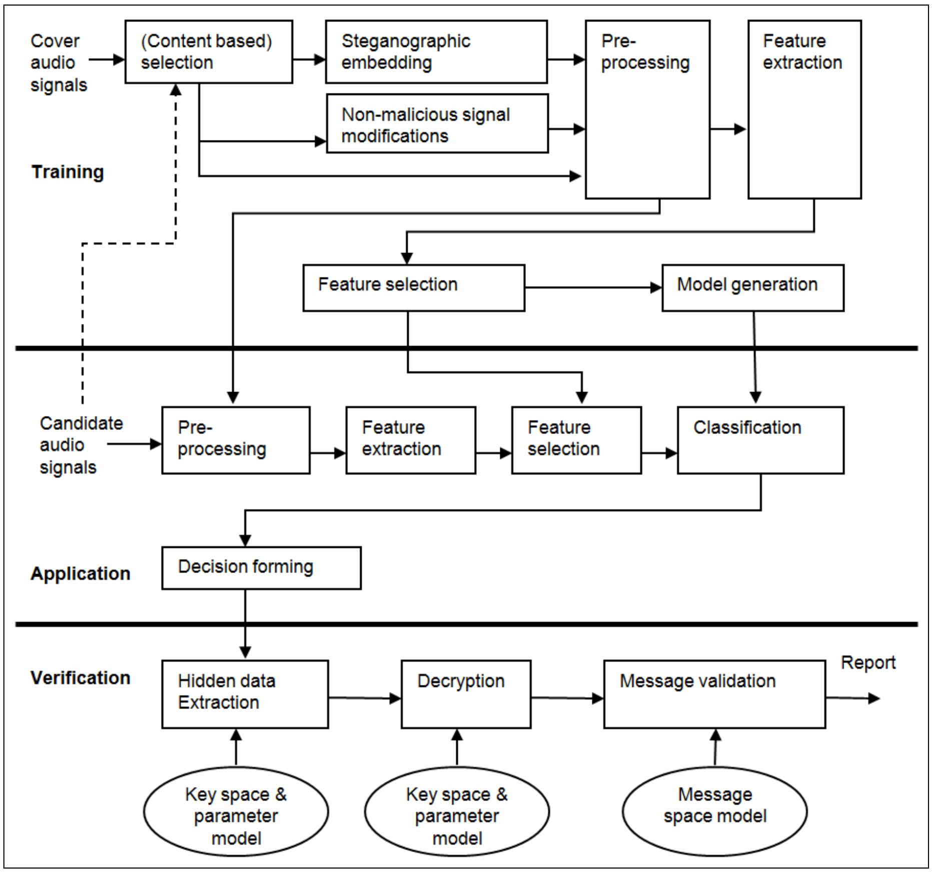


Figure 3.3: Detailed description of audio steganalysis as an idealised, Daubert standard conform, pattern recognition driven, forensic process in case of a single expert system (non-fusion case); the required parametrisation of the individual steps is neglected in this visualisation for the training and application phases

The closest realisation to this detailed, idealised process shown in figure 3.3 is the work of Provos and Honeyman in [Provos02] (see section 2.5.2). In their publication the authors have shown that the verification phase is actually the hard part in this process. In fact it involves three hard problems: the hidden data extraction (which has to be 100% reliable since the next step would automatically fail otherwise), the decryption of message that have been encrypted prior to embedding (see section 2.5.1) and the message validation.
The first might be possible, since the steganalysis system might identify the used steganographic scheme during the classification process and based on the knowledge of the scheme and with an estimation of the used parameters short cut attacks (e.g. dictionary attacks like tested in [Provos02] might be plausible). The second is equivalent to the cryptanalysis problem in a Kerckhoffs-compliant setup. Since hard encryption is nowadays easy to obtain[50] this part is extremely difficult to realise. In case the extracted data was not encrypted or the encryption was successfully broken the message validation renders the message into some format that can be used as evidence in the forensic process. Depending on the type of message communicated, this step might impose another set of problems to the forensic process. In case the message is not in an intuitive form (e.g. an ASCII text) or a well known standard format (e.g. a JPEG image), appropriate mechanisms for its interpretation have to be found (as shown by Provos and Honeyman in [Provos02]).

Summarising the facts presented on the ideal forensic audio steganalysis process, it has to be admitted that it can be applied successfully only in very rare cases. One of the few examples that can be

[50] See e.g. http://www.aescrypt.com/ for an open source AES 256-bit stand alone encryption/ decryption tool or programming libraries like Botan (http://botan.randombit.net/).

mentioned here is Data Stash[51] v1.1b and v1.5. In his system attack analysis[52] on this commercial tool – which unfortunately never made it into a scientific publication but can be easily reproduced by everyone – Guillaume Tena performed successfully all three steps of the verification phase: the hidden data extraction (Data Stash simply appends the data to be 'hidden' after the end of the cover file), the decryption (it uses the cryptographic algorithm Blowfish, but only to encrypt the key for access control purposes) and the message validation (the message is a ZIP compressed data container with a typical ZIP header and without a password defined).

Nevertheless, despite this existing example, which shows that the ideal steganalysis process can be performed successfully under certain, rather severe, circumstances, it has to be assumed that this is only valid for countering a very small number of steganographic tools. In general, this ideal process is infeasible; the main reasons have to be sought in the reliable message extraction, decryption and message validation parts. Therefore, the methodology for this thesis has to focus on a practical steganalysis process instead of the ideal one.

The practical steganalysis process

The **practical steganalysis process** considered in this thesis shifts the goal from the complete forensic steganalysis process (see section 2.5.1 and the considerations on the ideal forensic audio steganalysis process above) to a mere reliable and practically applicable detection in targeted steganalysis in a forensic setup. This is equivalent to restricting the focus of the investigations to steps 2, 3 and 4 (blind universal steganalysis, identification of the embedding strategy, and determination of the steganographic software) of the process model introduced in [Fridrich09].

In terms of achievable Daubert-compliance this automatically means basically two things: first, shifting the focus of the forensic mechanism that leads to the expert testimony from establishing communication contents to the pure fact that a communication took place, and second, weakening the link to a case (Daubert-standard criterion FREC0).

Before the methodology to be used for the investigations in this thesis is designed the terms 'reliability' and 'practicability' have to be précised. For this, first the BOSS contest as described in section 2.5.2 is reconsidered: simplified, it is a Kerckhoffs' compliant two-class setup with an oracle returning the classification accuracy for each steganalysis attempt. The basic assumptions within this contest are rather idealised, an object is either a cover or a stego object. If it is a cover, it would be a natural cover, i.e. authentic with a specific cover source model (in BOSS a digital camera), if it is a stego object, then it is modified by exactly one steganographic algorithm (in BOSS the algorithm called HUGO) using one fixed parametrisation. This is more a steganalytical detection benchmarking of the HUGO algorithm (with a fixed parametrisation) than practical steganalysis. In practice, aiming for forensic analyses or the observation of communication channels 'in the wild', the investigators would face different problems in audio steganalysis:

- There are more classes than unmodified, natural covers. While it might be plausible that images are used as they are generated, for other media, including audio signals, this is rather unlikely. Here, the signals usually undergo rather extensive signal modifications (e.g. by filtering operations, mixing, compression, etc.) prior to any release. Therefore, a class of 'modified' covers exist that does not behave consistent to any source statistic. The consequence, which has to be drawn, is that non-malicious signal modifications have to be considered in the setup of any steganalysis investigation.

- Digital objects like images are either stego objects or covers. For time-discrete media, like audio signals, it might be necessary not to consider the complete media object but instead to analyse parts of the object individually, for only few sections of the data (i.e. some segments of an audio stream) might have been modified. As a consequence, a focus has to be set on the development of segmental features as basis for the statistical pattern recognition based analysis.

[51] http://www.skyjuicesoftware.com/software/ds_info.html

[52] See: http://www.guillermito2.net/stegano/datastash/index.html

- In practical application a multi-class setup is more likely than the two-class modelling for steganalysis – the steganographic channels to be detected would assumedly be created using different steganographic tools with varying parametrisations. This means that based on the idea of universal steganalysis, multi-class approaches or a fusion-based framework or application specific detectors should be considered.
- 'In the wild', there exists no oracle to tell us how well or accurate our detector is performing. Therefore metrics should be designed that allow for performance estimation, as a first step to fulfil the part of the Daubert criteria that focuses on the error rates of a forensic scheme. This performance estimation should focus primarily on the accuracy, but it should also consider the throughput of the detector, which is paramount for any online setup but also crucial for many forensic steganalysis application scenarios.

For considering the **reliability and practicability** for the practical audio steganalysis process, as they are considered within this thesis, the following **basic assumptions for the methodology in practical audio steganalysis for this thesis** are made:

1. The considerations are restricted to the detection of steganography by modification (in a single file scenario – ignoring other, less common, potential setups like e.g.batch steganography [Ker07a]). The focus lies on the mere detection of the hidden communication. Other potential goals, like e.g. payload size estimations are outside the scope of this thesis.
2. The message extraction, decryption and validation from the idealised steganalysis process (see the description of the ideal forensic audio steganalysis process above) are excluded from the considerations here.
3. All considerations are made on the basis that statistical pattern recognition (SPR) based approaches are used to implement the steganalysis (see section 2.5.2).
4. The true distribution of covers and stego objects in real life application scenarios are not known.

Based on these assumptions, the **reliability** tries to access the question: How can the detector performance be estimated if there exists no way to measure it? I.e. bereft of the oracle telling us the detector performance, how can we be sure that we are not simply guessing 'in the wild'? These questions, especially if they aim for statistical generalisability, have to take into account the existence and influence of non-malicious signal modifications.
Regarding the **practicability**, we have to mainly look into the question: Is it feasible to expect a steganalysis result (in the form of a detector response) within a certain amount of time? This throughput-focussed questing is, due to its own complexity, outside the direct focus of this thesis. It is complicated by the potential alternatives that exist for the setup of the steganalysis system (single detector versus fusion-based framework and two-class versus multi-class) and the complexity of the models used in the classification. The model complexity strongly depends on the chosen classifier and amount of data used in training (including the material representing non-natural covers created by non-malicious signal modifications). Initial ideas that arise from the work within this thesis for answering parts of the practicability problem are presented in section 8.2.

Solution methodology for the audio steganalysis performed within this thesis

Here, an instantiation of the practical steganalysis process is performed. Regarding the methodology applied in this thesis for audio steganalysis, the methodology is split into three 'building blocks' (see figure 1.2 in section 1.3). The following basic principles are used within this thesis to implement these three blocks (signal preparation, statistical pattern recognition (SPR) and evaluation):

In the **signal preparation**, the **patterns observed** are the subtle impact caused by a steganographic embedding in an audio file. The **input** required for the statistical pattern recognition (SPR) based steganalysis performed here requires cover audio material as well as stego files for different steganographic algorithms in statistically significant numbers for the training of statistical models. Here, we could distinguish between two application modes: offline and online forensic setups. In the first case an investigator is performing steganalysis after an incident. In an ideal case he would have no restrictions on the computational complexity to spend on the case and he can analyse the signals under investigation regarding their genre, quality and other semantics and select the training material accordingly. In the second case, the online forensic setup, a steganalysis mechanism is operated like a malware detector to observe under real-time constraints a communication channel. In this case, the statistical models for the detection process have to be generated in advance, in the worst case on large, multi-genre set to cover a large number of potential contexts, if the channel content characteristics are not known.

Regarding the **statistical pattern recognition (SPR)** block, the **pre-processing** currently used in audio forensics is either content-based or content-insensitive. The considerations within this thesis are restricted to the latter. Content-based, or better content-analysis based pre-processing would be an extremely powerful, but equally complex methodology extension, which is due to its inherent complexity, reserved for future work.

Regarding **feature design approaches** that can be applied to audio steganalysis, both main approaches identified in section 2.4.2 are applied here: Intuition-based feature design, which to be successful requires expert knowledge on the domain covered, or the transfer of features from other (similar) problem domains.

Feature selection (see section 2.4.3) is used to validate the significance of all elements in the feature vector and, if necessary, eliminate insignificant features. This is included into the investigations performed here for two different reasons: on one hand it has the potential to increase the performance (in terms of throughput) of the classification and on the other hand it allows for the generation of domain knowledge because the identification of significant features also derives by implication also knowledge about the characteristics of the patterns classified.
The development of new feature selection strategies is outside the scope of this thesis. Instead, existing implemented techniques to perform feature selection are chosen from the established open-source data mining suite WEKA.

Regarding the resulting **model sizes**, it has to be assumed that they are very large, due to the large numbers of feature vectors required to sufficiently represent a multi-genre audio context. In case the audio genre used or the algorithm(s) to be detected can be narrowed down, smaller model sizes could be achieved.

Regarding the used **classification approaches**, forensic steganalysis is generally a classical multi-class classification problem assigning one input candidate to exactly one class out of a set of predefined classes. For performance reasons it might be wise to split this multi-class problem into a grid of two-class classification problems.
The development of new classification techniques is outside the scope of this thesis. Instead, existing implemented supervised and unsupervised classification techniques are chosen from the established open-source data mining suite WEKA.

For the **evaluations**, all approaches for audio steganalysis introduced in the state-of-the-art lack **comparable (detection) performance indicators**, a consequent consideration of **plausibility** issues as well as **assessment under the Daubert criteria** in their original literature. Within this thesis, a suitable detection performance indicator is introduced to the field of steganalysis and the introduced practical audio steganalysis approach undergoes plausibility considerations as well as an assessment as a forensic method. This assessment uses the Daubert criteria, as they are summarised in section 2.2, to summarise the potential forensic performance of the scheme. Special focus in the necessary investigations is put on the Daubert criterion DC1, which is summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*".

3.1.2 Introduction of the principles for microphone forensics applied within this thesis

The work on microphone forensics presented in this thesis is focussing primarily on the source authentication aspect of this application scenario. Investigations on the second possible aspect, the integrity verification for recorded material, are considered here to be secondary concerns.

Source authentication for microphone recordings, looking for device specific influences to the recorded material, is a classical pattern recognition (PR) problem. As such it is approached in the currently established state-of-the-art in this field by the application of template matching based approaches (see section 2.6). Within this thesis, an alternative methodology to the currently established template matching based approaches is developed. In this section the methodology for this new approach is introduced, to act as a basis for the conceptual descriptions in sections 3.2.1 and 3.2.3. The new approach uses statistical pattern recognition (SPR) to perform source authentication as well as integrity verification. The main difference between template matching and SPR pattern recognition sub-disciplines lies in the representation of the classification model: while in template matching the feature values that belong to one class are represented explicitly in the model, in SPR these are represented implicitly by statistical models. This fact usually makes SPR-based solutions more complex but also more tolerant to noise influences or strong intra-class variances. The latter reason is considered by the author to be a good motivation to implement a SPR-based microphone forensics approach.

The approach developed within the scope of this thesis can be considered, like the works of Garcia-Romero et al. in [Garcia-Romero10] and of Malik and Miller in [Malik12], as a realisation of the theoretical work of Oermann et al. in [Oermann05] on patterns intrinsic to the microphone response function. Besides the fact that it is more generalisable in its application than the existing approaches (it does not rely on the existence of local phenomena (like e.g. reverberations) or helper data (like electric network frequency templates) and it also works for other recorded content, not only speech signals) it can also be used to actually generate domain knowledge and thereby can answer still open research questions. Such knowledge, which can be answered by explorative statistical pattern recognition (a.k.a. data mining), might be for example the influence of the individual components in the recording process (see the context modelling for microphone recordings done in section 2.3.2).

Besides template matching and SPR **no further alternative solution approaches for forensic recording source authentication seem to be feasible**. Other pattern recognition (PR) sub-disciplines (see section 2.4), are by their very nature unsuitable to solve the problem of detecting the subtle traces / pattern imposed by the microphone in such extremely dynamic content.

For the secondary aspect of integrity verification, the author would consider the application of (feature-based) anomaly detection to be the only possible alternative to PR-based approaches[53].

For considering the **reliability and practicability** for the practical microphone forensics process, as they are considered within this thesis, the following **basic assumptions for the methodology in practical microphone forensics for this thesis** are made:

1. The term ‘microphone’, as used within the practical investigations, describes the whole hardware setup that is connected to the A/D converting device, i.e. the actual microphone, its mounting apparatus, the connecting cables, any pre-amplifier or phantom-power generator as well as the input of the A/D converter. If one of these components is exchanged (except for the mounting apparatus, which is considered in the practical investigations as a variable influence factor) the new setup is, in the context of the thesis, considered as a new ‘microphone’.

2. Acquiring large numbers of correctly labelled audio recordings is an extremely time consuming and burdensome task. Therefore, here the same assumption is made as in [Garcia-Romero10] and [Malik12] that the practicability of the microphone forensics approach can be shown by using test sets of small size but of suitable composition.

3. It is assumed that SPR-based approaches are, due to their higher flexibility, are better suited to cope with changing influences in the recording process than template matching based approaches.

[53] In fact that approach is applied by Grigoras in [Grigoras07] to implement the ENF-based integrity verification by searching for anomalies in the first-order derivative of the ENF-component of an audio signal.

Thereby, the introduced method is assumedly less content dependent that its template matching counterparts in the state-of-the-art.

Based on these assumptions, the **reliability** tries to access the question: How can the forensic performance of the approach be estimated if there exists no way to measure it? This is in microphone forensics closely related to the question: How have suitable test sets to be designed? These questions, especially if they aim for statistical generalisability, have also to take into account the existence and influence of non-malicious signal modifications.

Like with audio steganalysis, in depth investigations on the **practicability** of the approach introduced here are outside the direct focus of this thesis. Initial ideas that arise from the work within this thesis for answering parts of the practicability problem are presented in section 8.2.

Solution methodology for the microphone forensics performed within this thesis

Regarding the methodology applied in this thesis for microphone forensics, the methodology is split into three ‘building blocks’ (see figure 1.2 in section 1.3). The following basic principles are used within this thesis to implement these three blocks (signal preparation, statistical pattern recognition (SPR) and evaluation):

In the **signal preparation**, the **patterns observed** in the recorded material are the subtle traces left by the microphone response and the device characteristic noise introduced by a microphone or more precisely a recording setup in a recording. These patterns (the frequency response function $F_{mic}(f)$ of the microphone and its thermal noise $N_{mic}(f)$ see the context modelling for microphone recordings in section 2.3.2) are here expressed implicitly as statistical models by the microphone response based SPR approach.

The **input** required for the microphone response based SPR approach the forensic authentication requires at least recordings from the evaluated microphone as well as recordings from a statistically significant number of other microphones for training of a statistical model. This is due to the fact that here a kind of ‘world model’ is created for the training of the classifier. This has to be placed in contrast to the template matching based approaches described in the state-of-the-art, were local models are created that not only allow to tell in matching which template is the closest but also allow for an measurement or estimation of the distance to the closest template. If this minimal distance would be larger than a definable threshold, the system could decide that all templates are too far away and that therefore the verification sample does not belong to any of the enrolled devices. With the ‘world model’-centric SPR approach this would not happen. Here, all possible classes in an application scenario would have to be presented in training and the classification always assigns one of these classes to the verification sample. Therefore, for microphone forensics it is necessary to include in the investigation concepts and designs suitable setups for the training input (i.e. using sets of identical as well as different microphones). At this point it has to be explicitly mentioned that it is feasible to assume that for the purpose of a forensic authentication the investigator has access to the device under investigation (here, the evaluated microphone) and can make the reference recordings in a controlled environment.

Regarding the **statistical pattern recognition (SPR) block**, the **pre-processing** currently used in microphone forensics is either content-insensitive or content-analysis based. Due to the complexity of the latter, only content-insensitive pre-processing is used in the investigations on the microphone response based statistical pattern recognition approach developed within this thesis.

Regarding **feature design approaches** that can be applied to microphone forensics, two main approaches identified in section 2.4.2 are applied here: Intuition-based feature design, which to be successful requires expert knowledge on the domain covered, or the transfer of features from other (similar) problem domains.

An example for intuition-based feature design in this thesis is the usage of a (frequency domain) histogram as part of the feature set to be evaluated. This is motivated by known global phenomena,

in this case the intrinsic frequency response (see the context modelling for microphone recordings in section 2.3.2) of the microphone, which is a fact known to any expert in the field of audio recording processing and which is publicly described by microphone manufacturers for their products.
To design features for one application field by transfer of concepts from other problem domains is also a rather common methodology. Here, on the example of the Mel Frequency Cepstral Coefficients (MFCCs), features that originate in biometric speaker recognition are transferred successfully to microphone forensics.
The resulting features from these design approaches can be either local features, global features or segmental features (see section 2.4.2) with or without using higher-level content analysis. Within this thesis the local features are removed from the performed considerations due to the fact that their usage would result in extremely complex and practically infeasible classifier models. Also higher-level content analysis is excluded from the consideration within this thesis and reserved for future work.

Feature selection (see section 2.4.3) is used to validate the significance of all elements in the feature vector and, if necessary, eliminate insignificant features. This is included into the investigations performed here for two different reasons: on one hand it has the potential to increase the performance (in terms of throughput) of the classification and on the other hand it allows for the generation of domain knowledge because the identification of significant features also derives by implication also knowledge about the characteristics of the patterns classified.
The development of new feature selection strategies is outside the scope of this thesis. Instead, existing implemented techniques to perform feature selection are chosen from the established open-source data mining suite WEKA.

Regarding the resulting **model sizes**, it has to be assumed that they are significantly larger than the template sizes for the three existing approaches from the state-of-the-art. In case of the microphone response based statistical pattern recognition (SPR) approach the microphones are represented either as feature vectors extracted from the audio signal or as completely trained statistical model. The first case is more likely since it allows a dynamic adaptation of the authentication system (e.g. by changing the pre-processing routines, adding new features to the feature space, exchanging the classifier, adding/enrolling new microphones, etc.). The latter case, the storage of trained models, is less flexible but would be faster in field application since the necessary and time consuming process of generating a classifier model from the feature vectors has already been performed. If the feature vectors are used to represent a microphone then the number of feature vectors per class and their dimensionality define the size of the template. If trained models are used for the representation of microphones then the template size is determined by the input used in the model generation process and the way the model is generated, represented and stored by the classifier.
The actual size of the training set should be large enough for the chosen SPR-based approach, since it cannot distinguish intuitively between context imposed phenomena and microphone-intrinsic characteristics of the recording. Without sufficient training of a statistical model on suitable training sets of statistically significant size and based on relevant features, it would never be able to tell which kind of influence was responsible for the actual sample values within a window of an audio signal because the same microphone output could be the result of different microphones under different input signals and environmental conditions. Within this thesis a variation of the number of training samples is used to allow for initial estimations on required model sizes.
Regarding the used **classification approaches**, in general, the primary goal in microphone forensics is microphone authentication. Therefore it is a classical multi-class classification problem assigning one input candidate to exactly one class out of a set of predefined classes. The development of new classification techniques is outside the scope of this thesis. Instead, existing implemented supervised and unsupervised classification techniques are chosen from the established open-source data mining suite WEKA.

For the **evaluations**, all approaches for audio steganalysis introduced in the state-of-the-art lack **comparable (detection) performance indicators**, any consideration of **plausibility** issues and, with the exception of the electric network frequency (ENF) based approaches, **assessment of the (potential)**

forensic compliance under the Daubert criteria in their original literature. As one result, the performance of the different approaches is hardly comparable, due to the different dimensionalities of the investigated multi-class problems. Within this thesis, a directly comparable metric is introduced to this field. In section 8.2.1, first ideas on a throughput based performance based metric are discussed. Furthermore, first selected plausibility considerations are made for the introduced SPR-based microphone forensics approach. These plausibility considerations focus on the resilience of the microphone patterns imposed in the recording process against common audio signal post-processing operations. Also, the introduced microphone forensics approach undergoes an assessment as a potential forensic method. This assessment uses the Daubert criteria (as discussed in section 2.2) to summarise the potential forensic performance of the scheme. Special focus in the necessary investigations is put on the Daubert criterion DC1, which is summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*".

3.1.3 Comparison of the principles for the two application scenarios and generalisation into a general-purpose approach

Table 3.1 compares the similarities and differences regarding the main methodology for addressing the two application scenarios considered within this thesis.

Table 3.1: Brief comparison of the similarities and differences in the methodologies considered in this thesis for audio steganalysis and microphone forensics

	Aspect	**Audio steganalysis**	**Microphone forensics**
Signal preparation	patterns observed	impact of the embedding of messages (steganography by modification), influenced by degrees of freedom in cover selection and embedding	microphone response based traces from the recording source (the microphone) – influenced by degrees of freedom in the recording process
	security aspect considered	integrity verification (against steganographic embedding by the implicit verification of consistency of local or global phenomena or by detection of the violation of source intrinsic properties of the cover)	primary source authentication and secondary integrity verification (against composition of audio data from different sources)
	setups	offline forensic (content selection for training data based on semantics of the material under investigation) and online forensic (large, multi-genre set to cover a large number of potential contexts) instantiations of practical steganalysis	(offline) forensic / investigator has access to the microphone(s)
Pattern recognition	pattern recognition approaches	SPR, two-class and multi-class classifications	SPR, multi-class classifications
	pattern recognition and classification pipeline	pre-processing: content-insensitive	
		feature design: Intuition-based feature design and transfer of features from other (similar) problem domains	
		feature selection strategy: combination of existing feature selectors from the established open-source data mining suite WEKA	
		classification: existing techniques are used from the established open-source data mining suite WEKA	
Evaluation	detection performance indicators	detector reliability based on multi-class setups (Kappa statistics)	
	plausibility	selected, common audio signal post-processing operations are used to emulate robustness challenges to the detector or the usage of anti-forensics	
	forensic conformity	discussion on the basis of the Daubert criteria	
	scale of the evaluations	large-scale, representative, multi-genre evaluations	small-scale investigations – limited by the available physical recording setups

As shown in the two separate methodology sections 3.1.1 (audio steganalysis) and 3.1.2 (microphone forensics), the patterns observed in these two SPR application scenarios strongly differ, both having different degrees of freedom in the creation process that influence the pattern as well as the noise component in the signal. In steganalysis, the patterns observed are imposed to the audio material by the embedding of messages in steganography by modification based applications. The audio steganalysis in this case is considered here as being an integrity verification mechanism under consideration of the steganalysis as a malicious integrity violation. Microphone forensics focuses primarily in the authentication of a recording source based on traces resulting from the individual response functions. The consistency of these traces can also be used to verify the integrity of a recording under the expectance of composition attacks.

In the state-of-the-art in audio steganalysis, the most suitable solution method is considered to be statistical pattern recognition (SPR). In this thesis, this approach is transferred to the field of microphone forensics, where the current state-of-the-art so far focussing on the application of template matching (see section 2.6). The reason to do so is the fact that through the more complex (statistical) way of presentation of the basis for classification decisions (i.e. the classification model) the decision becomes more flexible and less content dependant. It is the assumption of the author that thereby one of the major shortcomings of the state-of-the-art approaches – the strong context dependency – might be overcome.

Regarding the dimensionality of the classification problem at hand, the opinions are divided for audio steganalysis. While some authors (e.g. [Provos02]) strongly argue against the modelling as a two-class problem, the majority of publications follow exactly this modelling approach. Both considered major directions (two-class and multi-class classification) are reflected within the considerations in this thesis. Microphone forensics is intuitively a typical multi-class identification setup and handled in this thesis exactly this way. The author is aware of the fact that the state-of-the-art approaches in this field perform with distance-based template matching actually a projection onto a linear search in the template space. A simulation of a similar projection of the introduced SPR-driven multi-class approach onto a sequence of two-class problems might be interesting, because it shifts parts of the overall complexity from the models to the numbers of classifications, which might directly affect the scalability and throughput of the mechanism. Nevertheless, due to the inherent complexity of this alternative it is considered to be outside the scope of this thesis and is reserved for future work.

Regarding the setups considered, steganalysis could either be forensic or online – especially the second option having rather severe implications on the training of the classifiers, while microphone forensics is a classical forensic setup. This has to be adequately reflected in the evaluation designs.

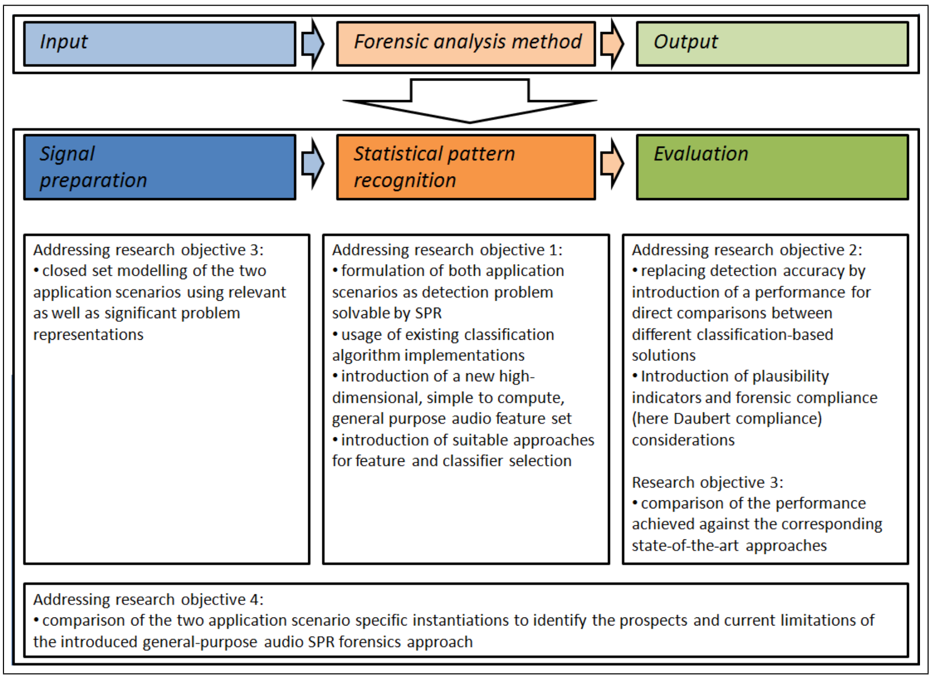
For both application scenarios similar pattern recognition and classification pipeline setups can be applied. Regarding the pipeline setup, it has to be repeated here that the design and implementation of new feature selectors and classifiers is outside of the scope of this thesis, since for these two blocks of the pipeline already suitable solutions exist. Instead, effort is invested in considerations on how to perform feature- and classifier selection.

The performance indicator used within this thesis focuses on comparability of the results. Therefore, the usually used classification accuracy is here replaced by Kappa statistics.

Regarding the plausibility considerations, they are within this thesis considered to be an integral part of the evaluation scheme for forensic methods. This is not entirely uncommon in steganalysis (see e.g. [Orsdemir08]), but for microphone forensics only in the work on the ENF approach considerations on its plausibility are found in the existing literature – most prominently published by E.B. Brixen (see e.g. [Brixen08b] and [Brixen08a]). For the other two classes of microphone forensics approaches presented in section 2.6, the usage of plausibility indicators is completely missing in the literature. Common audio signal post-processing operations (like normalisation, blind de-reverberation, etc.) as acknowledged by [REW11] are completely ignored in the evaluations performed. Due to the inherent insensitivity of template matching approaches to counter-forensics or anti-forensics methods, it has to be assumed that all these approaches are furthermore very easily affected by such targeted attack methods. Here, selected, common audio signal post-processing operations are used to emulate the necessary robustness challenges to the detector or the usage of anti-forensics.

For the discussion of the forensic compliance, the evaluation criteria FREC0 to FREC3 and DC1 to DC5 derived from the FRE rule 702 and the Daubert challenges are used as they are described in section 2.2. Regarding the evaluation and comparison of the introduced approaches the following points have to be highlighted as limitations imposed by the introduced methodology:
For the setup and the associated performance evaluation under the consideration of the Daubert criteria also important would be the verifiability. In this regard both application scenarios strongly differ: for steganalysis the approach by Provos and Honeyman in [Provos02] with its consecutive detection and (dictionary-based) message retrieval, decryption and validation shows that a verification of the forensic scheme is theoretically possible. This procedure (although highly infeasible for a universal steganalysis detector) would be an implementation of the ideal forensic steganalysis process described in section 3.1.1.
For such a posteriori verification of the result in microphone forensics, of the state-of-the-art approaches most do present no general alternative that can be used for independent verification: electric network frequency (ENF) approaches would fail because two microphones recording at the same time and segment of the power grit will show the same ENF pattern, the approach of Malik and Farid [Malik10] would show the same reverberation behaviour for two microphones in the same room and the approach introduced in [Garcia-Romero10] would fail in this regard for non-speech signals. Only the work of [Malik12] might present such a means for independent verification (see section 2.6.1). The combination (fusion) with this approach is an interesting topic reserved for future work.
Another point that influences the concepts is the scale of the empirical evaluations that is possible. In audio steganalysis all evaluations are strictly software-based and therefore for large scale empirical investigations, given the required software (here implementations of the embedding functions) is available. In microphone forensics the whole evaluations are hardware (microphone) based, therefore only a strongly limited number of recording setups can be considered within this thesis.

To draw a resume on the comparisons between both application scenarios, it can be stated that despite the huge differences between the two chosen application scenarios – especially regarding the patterns observed and addressed security aspects – they seem to provide a suitable ground for the introduction of a useful general-purpose statistical pattern recognition (SPR) driven audio forensics approach.



Addressing research objective 3:
- closed set modelling of the two application scenarios using relevant as well as significant problem representations

Addressing research objective 1:
- formulation of both application scenarios as detection problem solvable by SPR
- usage of existing classification algorithm implementations
- introduction of a new high-dimensional, simple to compute, general purpose audio feature set
- introduction of suitable approaches for feature and classifier selection

Addressing research objective 2:
- replacing detection accuracy by introduction of a performance for direct comparisons between different classification-based solutions
- Introduction of plausibility indicators and forensic compliance (here Daubert compliance) considerations

Research objective 3:
- comparison of the performance achieved against the corresponding state-of-the-art approaches

Addressing research objective 4:
- comparison of the two application scenario specific instantiations to identify the prospects and current limitations of the introduced general-purpose audio SPR forensics approach

Figure 3.4: Projection of the abstract descriptor of a forensic audio authenticity or integrity verification mechanism into the general-purpose SPR approach for this thesis (incl. identification of the research challenges relevant for each of the three 'building blocks')

Figure 3.4 shows the projection of the abstract descriptor of a forensic audio authenticity or integrity verification mechanism (as discussed in section 1.1) into the general-purpose statistical pattern recognition (SPR) approach for this thesis. The three 'building blocks' in the abstract descriptor find their individual counterparts in the main components of the introduced general purpose approach. The centre parts of the methodology considerations are, on one hand, the formulation of the two application scenarios as practical detection problems to be solved by statistical pattern recognition and, on the other hand, the decisions for the realisation of the SPR to rely on already existing classification algorithms while introducing a new high-dimensional, simple to compute, general purpose audio feature set as well as suitable approaches for feature and classifier selection.

Further methodology considerations for this thesis are: the closed set modelling of the application scenarios using relevant as well as significant problem representation (i.e. sets of audio material) as well as the extension of the performance evaluation to include, besides the usual detection performance, also plausibility as well as forensic compliance considerations. The work done on research objective 4 identifies the prospects and current limitations of the introduced general-purpose audio SPR forensics approach as an important instrument for future research in both application scenarios as well as for the adaptation of the introduced general-purpose approach to further application scenarios.

Based on the summary given above on the principles for the two chosen application scenarios, the following structure to be used in the following concept and design sections can be derived: First, addressing the degrees of freedom in the creation of the signals (influencing the intrinsic patterns and the noise), second, composition and parametrisation of the SPR pipeline, and third, performance metrics. This structure is applied in sections 3.2.1 (focussing of common concepts), 3.2.2 (specific concepts for audio steganalysis) and 3.2.3 (specific concepts for microphone forensics) respectively. In chapter 4 the sections also use this basic structure, but extend the second point (the parametrisations of the SPR pipeline) into the individual steps in this pipeline.

3.2 Concepts for the introduced general-purpose approach

Within this section the methodology considerations from section 3.1 are projected into investigation concepts considering the research objectives formulated in section 1.3.

Figure 3.5 below shows the placement of the concept or the conception phase in a statistical pattern recognition (SPR) pipeline. It is located on a meta-layer influencing each and every component in the pipeline as well as co-located decisions like evaluation goals and training- and test set designs.

In general, the conception phase addresses the same three distinct parts or 'building blocks' as introduced in section 1.3. As shown in figure 3.5, these parts are: first, the concepts for generating suitable audio training and test sets (reflecting the degrees of freedom in the signal generation process and the influences to patterns and noise), second, the concepts for managing the influence factors in the SPR pipeline and third, the evaluation considerations.

For the descriptions of the conceptual decisions made for these three 'building blocks', this section is in the following sub-divided into the common concepts for both considered application scenarios on one hand (section 3.2.1) and the identification of application scenario specific concepts on the other hand (sections 3.2.2 and 3.2.3 respectively). This inversion of the sequence employed in the previous sections for the descriptions (first applications scenario specific considerations, followed by generalisations) is performed to prevent unnecessary redundancy in the descriptions.

3.2.1 Common concepts for audio steganalysis and microphone forensics

The first set of major conceptual considerations is focusing on the signal preparation, i.e. on the **degrees of freedom in the generation of investigation conditions**. This reflects directly the requirements imposed by the Daubert criteria DC1 (short: 'empirically testable', see section 2.2.2) as well as DC3 (short: 'error rates') and can be roughly translated for SPR-based schemes as the **specification of suitable training and test sets**. Within this thesis practical investigations based on closed-set experiments are performed to evaluate the suitability of the instantiations of the introduced general-purpose

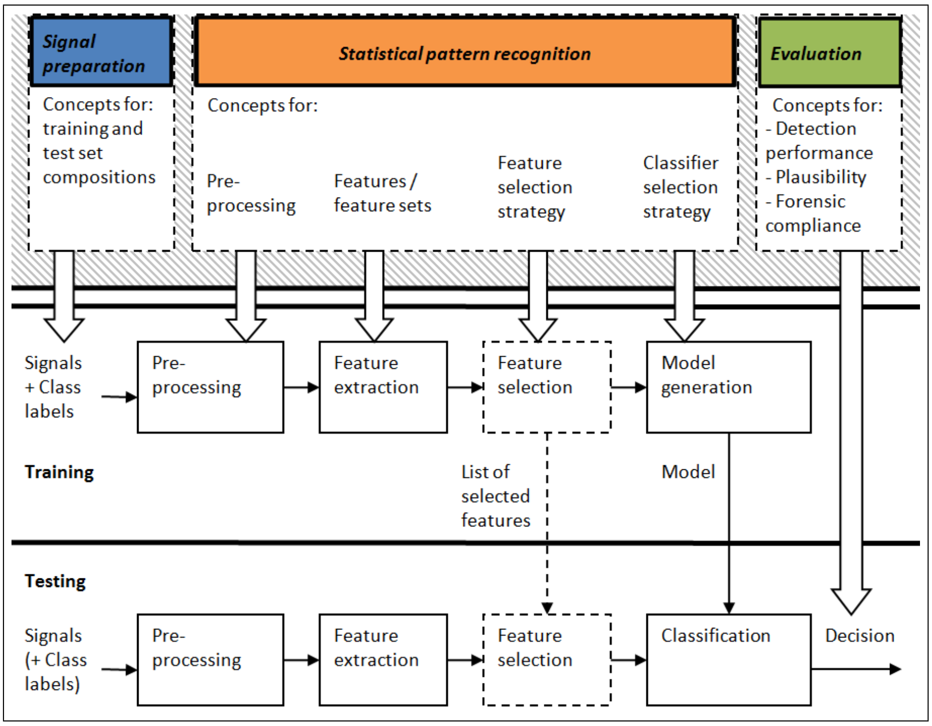


Figure 3.5: General statistical pattern recognition (SPR) pipeline – conception phase (extended from figure 2.5)

approach for the two selected application scenarios. Here, it is required to generate evaluation sets that are representative for the application scenario as well as statistically significant. While the representativeness has to be ensured in the test set design, the significance has to be verified in the investigations. Regarding the **input** considered by the application scenarios, this thesis focuses on (never compressed) PCM-encoded audio recordings sampled and quantised with CD quality (44.1 kHz, 16 Bit – see section 2.3). For the considerations in audio steganalysis, stereo signals taken from audio CDs are used as cover signals. In case of the recordings in microphone forensics the mono signals generated by the microphones are used. In the investigations performed, the correct class labels for all audio signals under consideration are made available.

The second set of important considerations for the common concepts for both application scenarios is regarding the **implementation of the statistical pattern recognition (SPR) signal processing pipeline**, as presented in section 2.4 and shown in figure 3.5 above.
Based on the methodology considerations presented in section 3.1, the decision is here to focus in the development and implementation work accompanying this thesis on feature extraction and use for feature selection as well as classification already existing software solutions (mainly the established open source data mining suite WEKA [Hall09]), integrated into own investigation strategies.
As described in section 2.4.1, **pre-processing** is used to enhance the classification performance, not to enable the successful classification. Therefore the pre-processing applied within this thesis is in the essence restricted to the framing and windowing required for the window-based processing of audio material in feature extraction. The usage of more sophisticated pre-processing operations to enhance the detector/classifier performance is reserved for future work.
To address and implement the **feature extraction**, here our own audio feature extractor labelled AMSL Audio Feature Extractor (AAFE) is developed. In this feature extractor segmental and global features are considered as they are introduced in section 2.4.2. For this thesis it is decided to discard local features as well as higher-level content analysis based features from the considerations. The reason for not considering local features is the small information reduction they provide. For audio signals, which contain in CD-quality 88,200 16 Bit samples per second, the number of computed local features would quickly exceed the practically processable set sizes for statistical pattern recognition based classification approaches. Regarding content analysis it is decided to restrict the consideration to low-level (syntacti-

cal) features and to reserve higher-level (semantical) content analysis based approaches for future work. The features used follow both concepts for feature design introduced in section 2.4.2. Therefore, some of the features are designed intuitively by using the domain knowledge of the author, while others are transferred from other audio signal processing domains.

As pointed out in the introduction of the methodology for this thesis in section 3.1.3, the design and implementation of new **feature selection** methods and **classifiers** is outside of the thesis' scope. Instead, from the set of established solutions for these two blocks of the pipeline, suitable candidates are selected using selection strategies designed within this thesis.

The third major set of conceptual considerations question is focusing on the **evaluations** 'building block'. After the analysis of the **detection performance** metrics used applied in the state-of-the-art in both application scenarios and their shortcomings, it is decided to introduce a new performance metric to these two fields. The benefit provided by this new metric, which is derived from Cohen's Kappa, is to allow for a fair evaluation of the performance in multi-class setups. A variation of the number of training samples and the impact to the achieved classification accuracy is used to allow for initial estimations on required model sizes, which allows some first estimations on the scalability of the introduced solutions. Investigations on **training- and test set sizes**, **context selectivity for training and test sets** and **context dependency for training- and test set generation** are of importance to show, in compliance to the Daubert standard (criterion DC3 short: 'error rates'), under which constraints the performed evaluations show statistically significant results.

The number of evaluations performed for **plausibility investigations** is limited to a practically feasible number – the idea is here to establish the concepts and reserve more detailed analyses for future work. For the investigations on the **forensic compliance**, a first concept derived from the Daubert criteria is introduced. It is highly infeasible to assume that a single PhD-thesis can explore and promote a forensic approach to such an extent that it would be accepted as fully compliant under these requirements. Therefore, here merely the concept is established and first required investigations are performed to outline the roadmap for addressing the question of forensic compliance for the two selected application scenarios as well as similar tasks.

3.2.2 Special concept extensions for audio steganalysis

In this section the necessary deviations from the common concepts, as they are summarized in section 3.2.1, are described for the application scenario of audio steganalysis. The goal for this application scenario is the implementation of the practical steganalysis process as described in section 3.1.1.

Concepts for data hiding

One important conceptual decision made in the signal preparation for this application scenario within this thesis is to rely on the usage of real algorithms for the embedding instead of embedding strategy simulations. The main drawback of this decision is the dependence on the availability of audio steganography tools or algorithms. As already mentioned in the analysis of the corresponding state-of-the-art (see section 2.5.2) there is only a small number of algorithms for audio steganography freely available. Therefore, the decision is made here to use audio steganography as well as audio watermarking algorithms to create an evaluation setup which might allow for some degree of generalisation.

Generally, steganalysis (by cover modification) and digital watermarking are two technically very similar disciplines, which can be summarised under the term data hiding. Nevertheless, the general objectives of steganography and digital watermarking are completely different. In [Cox08] this difference is illustrated as follows: in watermarking the embedded information is always some kind of metadata related to the cover object (owner/buyer information, producer information, integrity verification information, annotations, etc.) while in steganography the message should have no contextual correlation to the cover (instead it is just a message in a hidden communication). Despite these conceptual differences, the changes that audio steganography and watermarking algorithms impose to audio signals are very similar. Therefore, steganalysis (or the statistical analysis of potential cover objects) can be used to perform integrity verification against modification based data hiding operations.

Of a huge importance for the investigations performed are the basic assumptions for the training of statistical models for classification. Within this thesis, the trends in the state-of-the-art are followed in this regard and the possibility of so called 'cover-stego-attacks' is assumed in the forensic setup considered here. This means that the forensic examiner has access to the steganographic algorithms that are potentially used (i.e. a Kerckhoffs compliant setup) and is capable of generating stego objects from cover objects of his choice. This allows an adaptation of the classifier model to be used to the context of the observed channel. In this thesis a multi-genre setup (coverage of a wide set of possible cover contents) is compared with a very specific audio channel (speech only setups, like e.g. in VoIP telephony sessions).

Influence factors in the chosen SPR pipeline for audio steganalysis

For this application scenario the conceptual deviations from the common concepts, as presented in section 3.2.1 are only marginal: For pre-processing and feature selection no deviations exist.
For the design of the extracted features specialised considerations for intuitive designs have to be made. Here, the characteristics of the used algorithms as well as their embedding domains and strategies should be analysed for potential features. An example for such an intuitive feature development is the computation of LSB ratios and change rates to detect classical LSB-replacement based embedding strategies.
Regarding the classification, two-class as well as multi-class setups are considered.

Performance metrics and the assessment under the Daubert criteria for audio steganalysis

The main difference between the performance considerations for audio steganalysis and microphone forensics is the fact, that for the latter all error classes have assumedly the same significance, while for the former the different error classes have in practice strongly different significance. This can be easily illustrated for a two-class setup simply deciding whether an object is a stego object or an unmodified cover; here missed detections are a question of the security level, while falls alarms are a matter of cost. Therefore, appropriate performance metrics for steganalysis should also consider the different error classes (e.g. statistical Type I and Type II errors in the aforementioned two-class setup).
While the detection accuracy is right now the dominant performance metric in this field, it is replaced in this thesis by the Kappa statistics, a metric that fulfils the requirements of comparability and applicability for two-class as well as multi-class setups (implied as being required by Provos et al. in [Provos02]).
Regarding the assessment under the Daubert criteria, for steganalysis multiple cases should be considered: on one hand, either for the verified detection (ideal steganalysis process) or simply the detection (practical steganalysis process), and on the other hand for universal steganalysis (multi-class setup) or algorithm specific steganalysis (two-class setup). The second evaluation dimension in steganalysis would be the question of detection (practical steganalysis process) versus verified detection (ideal steganalysis process). Regarding this question, the verified detection (i.e. the message extraction, decoding and decryption parts) is considered to be outside the scope of this thesis.

3.2.3 Special concept extensions for microphone forensics

Building upon the common concepts for both application scenarios, as presented in section 3.2.1, this section identifies the concept extensions necessary to adapt the general methodology of statistical pattern recognition (SPR) based forensic mechanisms to the application scenario of microphone forensics.
The Daubert standard requires that a method "*must be empirically testable, falsifiable and refutable*" (see [USC93]), has been tested (criterion DC1) and that the attached error rates are known or estimated (criterion DC2). Derived from the methodology for this thesis, the concept to address these requirements is based on the recording process context model (as presented in section 2.3.2) and on targeted modifications on selected components in this process as well as on the statistical pattern recognition pipeline, that allow the estimation of the influence of these modifications.
The following points address: First, the degrees of freedom in the creation of the signals (influencing the intrinsic patterns and the noise) in signal preparation, second, parametrisations for the SPR pipeline, and third, evaluation considerations.

Degrees of freedom in the recording setup

The degrees of freedom in the recording setup in signal preparation are described extensively in the context model of the microphone recording pipeline in section 2.3.2 of this thesis. The investigation concept regarding these degrees of freedom is to perform a number of targeted modifications to allow for an estimation of their impact to the SPR-based microphone forensics approach introduced here.

At this point it has to be explicitly mentioned that it is feasible to assume that for the purpose of a forensic authentication the investigator has access to the device under investigation (here, the evaluated microphone) and can make the reference recordings in a controlled environment.

For reason of practicability the numbers of targeted modifications is limited to a feasible number. Furthermore, that only a limited number of recording hardware setups is available to the author. Nevertheless, the setups have to trying to achieve some form of generalisability even with a strongly limited set of recording setups (see table 3.1 in section 3.1.3). The main concepts to achieve such generalisability are the usage of different types of microphones, sets of identical microphones, different kinds of recorded content and (for specific investigations) the limitation of undesired side-effects by using ideal recoding environments (i.e. a soundproof, anechoic chamber).

Based on the conceptual, design and evaluation work performed within this thesis, future work should extend the analysis of the influence of the degrees of freedom in microphone forensics, either with a systematic evaluation of large numbers of hardware setups or by simulation of different influences.

Influence factors in the chosen statistical pattern recognition pipeline for microphone forensics

Next to the degrees of freedom in the recording setup, the second set of influence factors to the chosen SPR-based microphone forensics approach contains the components in the SPR pipeline and their parametrisations.

Each of the processing operations in this pipeline imposes its own impact factors to the overall result. Like for the degrees of freedom in the recording process, for reasons of practicability in the evaluations within this thesis some of those of influence factors undergo targeted modifications to evaluate their impact, while the remaining are kept constant.

Those influence factors chosen for targeted modification are:

- The used feature set as output of the feature extraction and feature selection
- The choice of classifiers in model generation and classification
- Training- and test set sizes
- Context selectivity for training- and test set generation
- Context dependency between training and test set

Of those influence factors, the impact of different features / feature sets as well as the influence of the choice of classifiers are the most important ones.

Since the beginning of his work on this microphone forensics approach in 2007, the author continually enhanced the used **feature extractor** AMSL Audio Feature Extractor (AAFE) with new segmental and global features without higher-level content analysis. As described in section 3.1.2 for the applicable and applied feature design approaches, some of them are especially designed for microphone forensics, while others are designed for other application areas (e.g. audio steganalysis) of the universal audio feature extractor used. Special focus shall be cast here on the spectrogram features designed in [Dohnal08] and modified for usage in this thesis in [Buchholz09]. These features have been designed with the purpose to reflect the well-known characteristics of the microphones regarding the intrinsic frequency response curves, and, as a consequence, they result in a very good performance for this purpose in [Buchholz09]. Nevertheless, it is shown in this thesis that they are outperformed by features which are motivated by other application areas of the AAFE and transferred into the domain of microphone forensics. For the practical evaluations on microphone forensics performed in this thesis the process of feature selection is strongly integrated into the concept. The **feature selection** serves a dual purpose: First, it identifies suitable and unsuitable features for microphone forensics and thereby generate domain knowledge on

the corresponding problem domain. Second, it improves the throughput of the security mechanism by a reduction of the classification times required, while at the same time keeping the same accuracies.
For the choices on **classification** methods no specific considerations are made within this thesis (see the common concepts as summarised in section 3.2.1).

The influence factors selected above for targeted modification are in the opinion of the author the ones which are strictly required to show that the microphone forensics approach introduced here actually works. All other existing influence factors are choices normally made with the aim of improving an already working approach. Since this thesis (and the accompanying conference and workshop papers) shows that the selected approach is working, those remaining factors are mere options to further improve the performance – their optimisation is reserved for future work. They are conceptually selected here to be kept constant. Examples for this class of influence factors are: the pre-processing alternatives (incl. different window sizes or windowing functions, etc.), classifier parametrisations, etc.

Performance metrics and the assessment under the Daubert criteria for microphone forensics

As summarised in section 3.1.2 for the currently applied methodologies in microphone forensics, the (classification or matching) accuracy is right now applied as the dominant performance metric in the state-of-the-art. The need of more appropriate performance metrics is highlighted for both application scenarios is section 3.2.1. For microphone authentication (which is a classical multi-class) the selection or introduction of new performance metrics is an even stronger necessity than for audio steganalysis (which could be modelled as a series of two-class classifications).
Furthermore, a new metric is required for the integrity investigations for recording composition (or mesh-up) detection. Here, with the Relative Frequency Ratio (the ratio of the number of observations in a statistical category to the total number of observations) a fundamental analysis method is borrowed from analytical statistics to be used as an appropriate metric.

Regarding the plausibility and estimation of the error rates – as requested by the Daubert criteria – specific investigations on (malicious or non-malicious) signal modifications or attacks on the method have to be included. Within this thesis the number of such signal modifications is limited to a set of common audio signal operations (normalisation, de-noising, MP3 conversion) and selected application scenario specific attacks (audio file composition and playback recording).

The general concepts for the assessment under the Daubert criteria are précised here for the special concepts applied within this thesis for microphone forensics. First, it has to be admitted that actually two different assessments would be necessary for this application scenario. The source authentication, as the primary task in for microphone forensics, and the integrity verification should be considered separately as forensic methods under Daubert criteria. This would reflect on one hand the fact that different approaches could be applied and on the other hand their different degree of maturity already achieved. Within this thesis the focus in microphone forensics is mainly on source authentication.

3.3 Concept of using the Daubert standard for general result discussion – definition of investigation tasks

Section 1.3 defines the research objectives for this thesis. The first three objectives can be divided into those that focus on the introduction of a generalised statistical pattern recognition (SPR) approach for forensic audio signal analysis as well as corresponding evaluation strategies (objectives 1 and 2) and the objective focussing on conducting the required empirical investigations on the two selected application scenarios (objective 3). The fourth objective uses the results from objectives 1 to 3 to outline the prospects and current limitations of the introduced general purpose audio SPR forensics approach.

The **introduction of the general purpose audio SPR forensics approach** (research objective 1) is performed in sections 3.1 and 3.2. The considerations on the **design of suitable performance indi-**

cators (research objective 2) and the **instantiation of the general purpose audio SPR approach for audio steganalysis and for microphone forensics** (part of research objective 3) are located in chapter 4. The second part on research objective 3, the **practical investigations on the application scenarios**, is addressed in chapters 5 and 6. The **comparison of the two application scenario specific instantiations** (research objective 4) is found in chapter 7.
Here, additional considerations are presented on the outline and limits of the practical investigations performed within this thesis. The idea is to use the Daubert criteria (see section 2.2 and its subsections) to outline the practical investigations by defining investigation tasks.

Considering the different Daubert criteria, it is obvious that their implications strongly differ in their significance for this thesis. Table 3.2 summarises the criteria and their significance in regards to the specific investigation tasks.

Table 3.2: The Daubert criteria and their significance in regards to the development of specific investigation tasks

Criterion	Description / significance
FREC0	**Description** ([LLI10a]): "*the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue*" **Significance**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.
FREC1	**Description** ([LLI10a]): the investigation (which leads to the corresponding expert testimony) is "*based upon sufficient facts or data*" **Significance**: Case specific, the only significance for this thesis arises from the fact that the term 'sufficient' has to be manifested into training and testing (application / evaluation) set sizes.
FREC2	**Description** ([LLI10a]): the investigation is based upon "*reliable principles and methods*", preferably scientific methodology and knowledge **Significance**: Chapter 2, as well as sections 3.1, 3.2 and chapter 4 of this thesis are dedicated to establish the fact that the two exemplary selected audio forensic methods are implemented as deterministic processes using the decades old methodology of statistical pattern recognition (SPR). Besides the task to show that the application scenarios can actually be solved by SPR, no further tasks are derived from this criterion.
FREC3	**Description** ([LLI10a]): the (forensic) methods are applied "*reliably to the facts of the case*" **Significance**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.
DC1	**Description** ([LLI10b]): "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"; summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*" **Significance**: This criterion imposes the most important task to the practical investigations performed within this thesis: Establish within which limits the proposed media forensic methods can give plausible results.
DC2	**Description** ([LLI10b]): "*whether the technique or theory has been subject to peer review and publication*"; with "*publication*" meaning 'open publication' **Significance**: This criterion is not translated into tasks but instead requires the author to interact with the scientific community relevant for the chosen application scenario. To address this criterion this thesis is submitted for (peer) review, as have been the accompanying conference and journal papers. The review comments received from the latter have helped shaping the described approaches as well as their evaluations.
DC3	**Description** ([LLI10b]): "*the known or potential rate of error of the technique or theory when applied*" **Significance**: This criterion requires the investigations for DC1 to be accompanied by a reliable measurement of error rates achieved.
DC4	**Description** ([LLI10b]): "*the existence and maintenance of standards and controls*" **Significance**: The task that would be derived from this criterion would be the compilation of the work into standards together with or within a standardisation body. This complex process is outside the scope of this thesis.
DC5	**Description** ([LLI10b]): "*whether the technique or theory has been generally accepted in the scientific community*" **Significance**: This criterion is similar to DC2 in its meaning and in the fact that it is not translated into tasks.

Based on table 3.2 the following **investigation tasks** are derived here for practical investigations on the two application scenarios considered within this thesis:

A) Empirical ground truth: show that the application scenarios can actually be solved by statistical pattern recognition (SPR) – derived from FREC2. This task also includes an estimation on what 'sufficient' means in terms of required training and testing (application / evaluation) set sizes – derived from FREC1.

B) Investigations on the impact of application scenario specific intrinsic influences to the SPR process – derived from DC1 and DC3.

C) Investigations on influences outside the SPR process on the performance of the scheme – derived from DC1 and DC3.

Investigation task A documents of refutes whether the introduced general-purpose approach (or more precisely: application scenario specific instantiations thereof) can be used to successfully address multiple audio forensics investigation goals. Therefore, a positive result for this task does answer research objective 1 (which extends research challenge a) – see section 1.3) and builds the ground truth for research objective 3 (which extends research challenge (c)).
Investigation tasks B and C extend the considerations on the two exemplary selected forensic application scenarios and therefore contribute to the answers for research objective 3. The answers given here allow an assessment on how adequately the application scenarios can be implemented with the introduced approach (research challenge c)).

Besides these tasks, which are addressed individually for the two exemplary chosen application scenarios, a summary and comparison is formulated as an additional task:

D) Summary and comparison of the results for both application scenarios.

The answers generated by fulfilling this task addresses the question about performance metrics that allow for a comparison between different application domains – research objective 2 (which extends research challenge (b), see section 1.3) – as well as the questions on limitations of the introduced general-purpose audio SPR forensics approach raised by research objective 4.

Based on these task descriptions (as extensions or précising statements for the research objectives formulated in section 1.3), the corresponding investigation designs for both application scenarios are described in detail in chapter 4. The investigation tasks A) to C) are elaborated in more detail for the audio steganalysis application scenario in section 4.2. The corresponding practical investigations are pr presented chapter 5. For the microphone forensics application scenario, they are elaborated in more detail in section 4.3 and the corresponding practical investigations are presented in chapter 6.
Task D) is addressed in chapter 7.

3.4 Restriction of the considerations within this thesis

In this section the **scope of the methodological and conceptual considerations is restricted**. This delimitation of the work performed has to be done because, even within the context of a PhD thesis, not every aspect can be developed to cover every possible detail.

Outside the scope of the thesis are considerations on:

- Work on the design of new audio pre-processing, feature selection methods and classification algorithms – statistical pattern recognition is used here as a kind of 'off the shelf' product.

- Detailed reflections on information fusion (here, the combination of different forensic mechanisms) – due to the complexity this task imposes of the confidence (considering the distance to a complex, combined decision threshold) estimation for the combined output.

- Work on benchmarking of audio forensic tools – Even though some steps are made into this direction (e.g. in the design of selection schemes for feature and classifier selection) benchmarking work would require a much wider test base. In the evaluations performed within this thesis the authors aim is to achieve some form of (statistical) generalisability. The design and implementation of fair benchmarking strategies would go beyond what is feasible within this thesis.

- Estimation of the achievable security levels and the precise error rates for the introduced audio forensics methods – these considerations would require extensive benchmarking (see the point above).

- Detailed juristic analyses – Since the author possesses absolutely no legal training, all legal considerations made within this thesis (especially on the Daubert standard) are therefore layman's interpretation of freely available material, which are made to the best of the author's knowledge. The intention behind the work on forensic compliance performed here is to derive a performance metric for research in the field of audio forensics. If the content of this thesis is intended to be used in any legal proceedings, the reader must consult appropriate legal counsel for the corresponding jurisdiction.

- A complete (mathematical) formalisation of the media forensic process – Some of the work performed within this thesis (e.g. the microphone recording context model) are backed by a complete mathematical formalisation, these formalisations are restricted to the absolute minimum, because a complete formalisation of the whole process is outside the scope of this thesis.

Based on the investigation tasks defined in section 3.3, the extent of the practical investigations conducted within this thesis is here narrowed down as follows:

Audio steganalysis

Regarding the investigation task A) 'Empirical ground truth', the size of the set of evaluated information hiding algorithms is strongly limited. Selected investigations are performed on a set of nine different algorithms. The majority of the investigations are performed with a subset of three different algorithms that show the worst, the best and an average detection performance.

The investigations on the impact of application scenario specific intrinsic influences to the statistical pattern recognition (SPR) process (investigation task B)) are limited primarily to the relevant features and the performance of existing classifiers (taken from the data mining suite WEKA [Hall09] version 3.6.1) in their default parametrisations. Parameter optimisation operations for the classifiers or other possible performance enhancing operations like pre-processing operations are omitted here.

For the investigation on influences outside the SPR process on the performance of the scheme (investigation task C)) the embedding domain, the key scenario used in the steganographic embedding and content dependant training and testing are addressed here. Other possible influence factors, like e.g. the embedding strength or variations of other embedding parameters are reserved for future work. The number of plausibility considerations is here limited to a small number of audio signal post-processing operations.

Microphone forensics

Because this application scenario is hardware driven and the amount of available hardware is limited only to a small number of practical test setups (with between 4 and 7 different microphones) are considered here. To achieve at least some level of confidence in the results for investigation task A) 'Empirical ground truth', two sets of four identical microphones each are used in the investigations on intra-class detection performance.

The considerations on investigation task B) are for microphone forensics the same as presented above for audio steganalysis.

Regarding the investigation on influences outside the SPR process on the performance of the scheme (investigation task C)) only specific influences are tested here. These include ten different recording environments, different orientations and mountings of microphones, selected source types (human speakers versus exactly one loudspeaker) and content dependency for ten different reference signals. Other potential influence factors, like e.g. used pre-amplifiers, cable, etc are reserved for future work. For the plausibility considerations, the influences of normalisation, MP3 conversion, de-noising and playback recording on the introduced are evaluated. The performed small scale investigations on composition detection focus on four different attack scenarios, which might easily extended in future work.

4

Investigation Designs – Considerations for the Practical Evaluations

In this chapter the abstract methodology and concept considerations from chapter 3 are transferred into precise evaluation setups. The layout of the description follow the same flow already used for the descriptions of the concepts in section 3.2: first the considerations on the aspects (here designs) that are common for both considered application scenarios are discussed and then, in the second part of this chapter the scenario specific design considerations are made.
As an important point in the scenario specific design considerations also a **task and restriction interpretation for each application scenario** is performed on basis of the (abstract) definition of investigation tasks in section 3.3 and the corresponding restrictions postulated in section 3.4.

4.1 Common designs decisions for audio steganalysis and microphone forensics

The design decisions have to precise the conceptual considerations made in section 3.2 and its subsections in eight important points:

- Input format and patterns observed
- Pre-processing applied
- Feature extraction and the construction of the feature space
- Feature selection
- Classification approaches
- Detection performance indicators
- Plausibility indicators
- Forensic compliance considerations

Regarding the first, the second, the seventh and the eights point of this list, only a small refinement is necessary in comparison to the conceptual considerations:

As already stated in section 3.2.1, the **input format** considered in both application scenarios is PCM encoded audio signals sampled and quantised with CD quality (44.1 kHz, 16 Bit – see section 2.3). In the audio steganalysis application scenario, the cover signals are either taken from commercial audio CDs or specifically recorded in this format for the investigations performed here. Therefore, it can be ensured that these signals have never been compressed. For microphone forensics, the recordings are directly made in this format to ensure that no additional processing artefacts (like re-sampling or re-quantisation noise) are impairing the investigations performed.

The **patterns observed** in audio steganalysis are created by the embedding with different information hiding (IH) algorithms (see section 4.2.1 below). For the microphone forensics investigations (except where explicitly stated otherwise in section 4.3.1) the same reference sounds are recorded in parallel by all microphones in a test set. This is done to ensure that the differences in the recordings are really originating from characteristic transfer functions of the different sources (i.e. microphones or microphone pre-amplifier combinations) rather than from changing environmental conditions.

The influence of **pre-processing** operations is restricted within this thesis to the absolute minimum. As already described in the conceptual considerations in section 3.2.1, the universal pre-processing consists of the framing and windowing required for the window-based processing of audio material in feature extraction.

The **plausibility indicators** used exemplary within this thesis are selected signal post-processing operations which are common in audio signal handling. The operations used here are: MP3 encoding (with a bit rate of 128kBit/s using the Lame codec[54]), de-noising (here by re-quantisation for 16 Bit to 8 Bit and back to 16 Bit) and normalisation (up to -3dB of the maximum possible signal level).
Using these post-processing operations, it is possible to show how plausible the achieved detection results are in cases where such a post-processing was applied instead of a steganographic embedding or where the microphones recording characteristic was overlaid by such post-processing influences.

For the forensic compliance considerations the table 3.2 introduced in section 3.3 is re-used in a slightly modified form. Instead of the significance of each of the Federal Rules of Evidence (FRE) and Daubert criteria it contains in the summary of evaluations the progress made within this thesis towards fulfilling that criterion.

For the remaining four points, the following subsections specify the design decisions made to precise the conceptual considerations.

4.1.1 The feature extractor and feature space used

The AMSL Audio Feature Extractor (AAFE) used for feature extraction within this thesis was in 2006 / 2007 intended to be used in a dedicated audio steganalysis software called AMSL Audio Steganalysis Toolset (AAST). The AAST never saw any field application, but the AAFE, as one of its core components, remains in usage and is used in this thesis as the universal audio feature extractor applicable in both application scenarios. The AAFE (or parts of the extractor) is also used by some other researchers in the field of audio steganalysis (e.g. by Qingzhong Liu of the Sam Houston State University of Texas, in the USA, see [Liu09] and [Liu11]).
Since the beginning of the PhD project reflected within this thesis, the AAFE was constantly improved by the author. For the practical investigations presented in chapters 5 and 6, three versions of this audio feature extraction software are of importance:

- AAFE v.1.0.3 (with its 63 dimensional feature vectors) introduced in 2007 in [Kraetzer07a],
- AAFE v.1.0.4 (with a 98 dimensional feature vector) introduced in 2008 in [Kraetzer08a],
- AAFE v.2.0.5 (with 590 segmental and 17 global features) introduced in 2010 in [Kraetzer10]

Table 4.1 compares the feature set compositions extracted by AAFE versions 1.0.3, 1.0.4 and 2.0.5. The detailed feature descriptions for all features are given at the end of the document (page 189) in the appendix “Appendix A: Audio Features used in AAFE”.

[54] http://lame.sourceforge.net/

Table 4.1: Comparison between the feature set compositions extracted by AAFE versions 1.0.3, 1.0.4 and 2.0.5 (considerations for the default window size of 1024 samples)

	AAFE v.1.0.3 (C/C++)	AAFE v.1.0.4 (C/C++)	AAFE v.2.0.5 (MATLAB)
Remarks	63 features; first used in [Kraetzer07a]	98 features; first used in [Kraetzer08a] – extends v.1.0.3	590 segmental and 17 global features; first used in [Kraetzer10] – reimplementation of the AAFE in MATLAB
Segmental features	**time-domain**	**time-domain**	**time-domain**
	sf_{ev} empirical variance sf_{cv} covariance $sf_{entropy}$ entropy sf_{LSBrat} LSB ratio $sf_{LSBflip}$ LSB flipping rate sf_{mean} mean of samples in time domain sf_{median} median of samples in time domain	sf_{ev} empirical variance sf_{cv} covariance $sf_{entropy}$ entropy sf_{LSBrat} LSB ratio $sf_{LSBflip}$ LSB flipping rate sf_{mean} mean of samples in time domain sf_{median} median of samples in time domain	$sf_{entropy}$ entropy sf_{LSBrat} LSB ratio $sf_{LSBflip}$ LSB flipping rate sf_{mean} mean of samples in time domain sf_{median} median of samples in time domain $sf_{zero\_cross\_rate}$ zero-crossing-rate sf_{energy} energy $sf_{RMS\_amplitude}$ RMS amplitude
	frequency-domain	**frequency-domain**	**frequency-domain**
		$sf_{formant\_*}$ 11 formant describing features $sf_{Bark\_1}, \cdots, sf_{Bark\_24}$ 24 band Bark scale spectrogram	$sf_{formant\_*}$ 11 formant describing features $sf_{sp\_centroid}$ spectral centroid sf_{sp_flux} spectral flux $sf_{sp\_rolloff}$ spectral roll-off sf_{sp_bw} spectral bandwidth $sf_{sp\_smoothness}$ spectral smoothness $sf_{sp\_irregularity}$ spectral irregularity $sf_{sp\_entropy}$ spectral entropy $sf_{sp\_base\_freq}$ base frequency $sf_{spec\_1}, \cdots, sf_{spec\_512}$ 512-bin linear spectrogram
	frequency-domain	**frequency-domain**	**frequency-domain**
	$sf_{MFCC\_1}, \cdots, sf_{MFCC\_28}$ 28 Mel-frequency cepstral coefficients $sf_{FMFCC\_1}, \cdots, sf_{FMFCC\_28}$ 28 filtered Mel-frequency cepstral coefficients	$sf_{MFCC\_1}, \cdots, sf_{MFCC\_28}$ 28 Mel-frequency cepstral coefficients $sf_{FMFCC\_1}, \cdots, sf_{FMFCC\_28}$ 28 filtered Mel-frequency cepstral coefficients	$sf_{MFCC\_1}, \cdots, sf_{MFCC\_13}$ 13 Mel-frequency cepstral coefficients $sf_{FMFCC\_1}, \cdots, sf_{FMFCC\_13}$ 13 filtered Mel-frequency cepstral coefficients $sf_{d2FMFCC\_1}, \cdots,$ $sf_{d2FMFCC\_13}$ 13 second-order derivative FMFCCs $sf_{d2FMFCC\_1}, \cdots,$ $sf_{d2FMFCC\_13}$ 13 second-order derivative FMFCCs
Global features			gf_{zcr_total} total zero-crossing-rate $gf_{entropy\_AVE}$ average entropy gf_{LSBrat_AVE} average LSB ratio $gf_{LSBflip\_AVE}$ average LSB flipping rate gf_{mean_AVE} average mean gf_{median_AVE} average median $gf_{zero\_cross\_rate\_AVE}$ average zero-cross-rate gf_{energy_AVE} average energy $gf_{RMS\_amplitude\_AVE}$ average RMS amplitude

Continued on Next Page...

Table 4.1 – Continued

	AAFE v.1.0.3 (C/C++)	AAFE v.1.0.4 (C/C++)	AAFE v.2.0.5 (MATLAB)
			$gf_{sp\_centroid\_AVE}$ average spectral centroid $gf_{sp\_rolloff\_AVE}$ average spectral roll-off $gf_{sp\_bw\_AVE}$ average spectral bandwidth $gf_{sp\_smoothness\_AVE}$ average spectral smoothness $gf_{sp\_irregularity\_AVE}$ average spectral irregularity $gf_{sp\_entropy\_AVE}$ average spectral entropy $gf_{sp\_base\_freq\_AVE}$ average base frequency $gf_{sp\_flux\_AVE}$ average spectral flux

The segmental features sf_* are computed for one window S_i^k of the sampled, quantised and windowed digital audio signal (window index i and channel index k). The global features gf_* are computed over the window-stream of a complete channel k of the input audio signal.

4.1.2 The feature selection performed

For the implementation of the feature selection here five different feature evaluators are used from the WEKA portfolio, together with the search method 'Ranker' (sorting the features by the merit displayed in the single feature evaluations). These used five feature evaluators are *ChiSquaredAttributeEval*, *FilteredAttributeEval*, *InfoGainAttributeEval*, *OneRAttributeEval*, *SymmetricalUncertAttributeEval* from WEKA version 3.6.1. This selection covers filter functions as well as wrapper-based feature selection approaches. For a description of the evaluators see [Witten05].

The output of all five feature evaluators is fused on a per feature basis by computing the arithmetic mean of the ranks returned. The final ranking is obtained by re-sorting the features by this output.

The feature ranking investigations are accompanied by estimations on the true dimensionality of the feature space for each classification problem. To this purpose WEKAs implementation of a principal component analysis (PCA) is used to transform the feature space given into a lower-dimensional feature space containing at least 95% of the original information. The dimensionality of this lower-dimensional space is then assumed to be close to the true dimensionality of the problem, i.e. number of non-correlated components derivable from the AAFE feature set for a given classification problem.

4.1.3 The classification algorithms used

For the classification, existing algorithm implementations provided by WEKA (version 3.6.1) are used. These can be divided into the two major classes of clustering and supervised classification techniques.

Clustering

WEKA (version 3.6.1) provides eight different clustering algorithms (*weka.clusterers.**). These are: *cobweb*, *DBScan*, *EM*, *FarthestFirst*, *FilteredClusterer*, *MakeDensityBasedClusterer*, *OPTICS* and *SimpleKMeans*. If these clustering algorithms are used within this thesis, except for the number of expected clusters no deviation from the default parametrisation is made. All algorithms are used in 'classes-to-clusters' evaluation mode to enable the computation of detection performances. Since the algorithms are applied in this thesis in a black box view, a detailed discussion of their construction principles is omitted here. For such information the author refers to [Witten05].

Alternatively to the complete naming scheme, an abbreviated form is used in this thesis, omitting the prefix '*weka.clusterers.*' (e.g. *SimpleKMeans* instead of *weka.clusterers.SimpleKMeans*).

(Supervised) classification

In WEKA (version 3.6.1) 74 different supervised classification algorithms (*weka.classifiers.**) are implemented. These are grouped into the following eight classes:

- *weka.classifiers.bayes.** (Bayesian classifiers): *BayesNet*, *ComplementNaiveBayes*, *DMNBtext*, *NaiveBayes*, *NaiveBayesMultinomial*, *NaiveBayesMultinomialUpdateable*, *NaiveBayesSimple*, *NaiveBayesUpdateable*
- *weka.classifiers.functions.**: *LibLINEAR*, *LibSVM*, *Logistic*, *MultilayerPerceptron*, *MLRM*, *RBFNetwork*, *SimpleLogistic*, *SMO*
- *weka.classifiers.lazy.**: *IB1*, *IBk*, *KStar*, *LWL*
- *weka.classifiers.meta.** (Meta-classifiers): *AdaBoostM1*, *AttributeSelectedClassifier*, *Bagging*, *ClassificationViaClustering*, *ClassificationViaRegression*, *CostSensitiveClassifier*, *CVParameterSelection*, *Dagging*, *Decorate*, *END*, *EnsembleSelection*, *FilteredClassifier*, *Grading*, *GridSearch*, *LogitBoost*, *MetaCost*, *MultiBoostAB*, *MultiClassClassifier*, *MultiScheme*, *OrdinalClassClassifier*, *RacedIncrementalLogitBoost*, *RandomCommittee*, *RandomSubSpace*, *RotationForest*, *Stacking*, *StackingC*, *Vote*
- *weka.classifiers.mi.**: *CitationKNN*, *MISMO*, *MIWrapper*, *SimpleMI*
- *weka.classifiers.misc.**: *FLR*, *HyperPipes*, *VFI*
- *weka.classifiers.rules.** (rule-based classifiers): *ConjunctiveRule*, *DecisionTable*, *DTNB*, *JRip*, *NNge*, *OneR*, *PART*, *Ridor*, *ZeroR*
- *weka.classifiers.trees.** (tree-based classifiers): *BFTree*, *DecisionStump*, *FT*, *J48*, *J48graft*, *LADTree*, *LMT*, *NBTree*, *RandomForest*, *RandomTree*, *REPTree*, *SimpleCart*

Alternatively to the complete naming scheme, an abbreviated form is used in this thesis, omitting the prefix '*weka.classifiers.*' (e.g. *NaiveBayes* or *bayes.NaiveBayes* instead of *weka.classifiers.bayes.NaiveBayes*). Giving descriptions of these classification schemes, their basic methods of operation and their differences takes up about half of the 500 page textbook [Witten05]. To repeat any of these considerations does not add any value to this thesis, because the algorithms are applied here in a black box view as an 'off-the-shelf' product. Within all evaluations, they are used with their default parametrisation (see [Witten05]).

4.1.4 Detection performance indicators used

As pointed out in the analysis of the state-of-the-art in research on the chosen two application scenarios (see sections 2.5.2 and 2.6.2 respectively), the accuracy (i.e. the ratio between true classifications and all classification attempts in a supervised classification) is currently used as the main performance indicator in both application scenarios discussed.

For the measurement of the classification gain for fair performance evaluation within this thesis it is proposed to use the **Kappa statistics** κ instead of the accuracy. It is basically a single-rater version of Cohen's Kappa (see [Carletta96], [Gwet12] for multi-rater considerations and [Eugenio04] for single-rater considerations derived from Cohen's Kappa) in the range $[-1, 1]$. Therefore the Kappa statistic measures the agreement of prediction with the true class (i.e. the agreement normalised for chance agreement). Equation 4.1 shows the computation of the Kappa statistics κ for an n-class problem:

$$\kappa = \frac{1}{n}\sum_{a=1}^{n}\frac{P_a - P_{chance}}{1 - P_{chance}} \tag{4.1}$$

For each of the n classes P_a is the corresponding percentage agreement (e.g., between the classifier and ground truth) and P_{chance} is the probability of chance agreement. Therefore, $\kappa = 1$ indicates perfect agreement and $\kappa = 0$ indicates chance agreement for the overall classification. Only in rare cases negative κ values are achieved, i.e. the classification performance of a system is worse than simple

guessing at the class. This is most likely the case when the model was trained to distinguish between patterns completely different than the ones actually presented in the evaluations.
For equally distributed classes, P_{chance} for all classes is simply $\frac{1}{n}$. For differently distributed classes [Eugenio04] describes different methods how to calculate of estimate P_{chance}. For the computation of the Kappa statistics within this thesis the WEKA implementation is used, estimating Kappa from the distribution of the classes in the supplied test set.

By using Kappa statistics, it is possible to construct for classification-based investigations a degree of closeness of measurements of a quantity to its actual (true) value that is exempt from the influence of the probability of guessing correctly. Such a metric does allow for direct comparison between the classification performances of classifiers on problems of different classes (e.g. a two-class classification problem like the classical hypothesis testing for an assumable steganographically modified channel and a four-class problem like steganographic algorithm identification on a set of three algorithms (plus unmarked covers) that might have been applied).

Regarding the interpretability of Kappa, [Landis77] presents a mapping between the Kappa value and the agreements of the different raters (see table 4.2). Within this thesis the fact is used that it is actually known to which class a sample belongs in the evaluations performed – a rather unlikely scenario in the normal application fields of Kappa statistics, like studies performed in clinical medicine. Based on this realisation, here the Kappa values are mapped onto statistical confidence using the mapping defined in table 4.2.

Table 4.2: Kappa values, agreements according to [Landis77] and the statistical confidence mapping used in this thesis

Kappa value κ	**Agreement according to [Landis77]**	**Statistical confidence mapping used in this thesis**
$\kappa \leq 0$	No agreement	Poor
$0.01 \leq \kappa \leq 0.20$	Slight agreement	
$0.21 \leq \kappa \leq 0.40$	Fair agreement	
$0.41 \leq \kappa \leq 0.60$	Moderate agreement	Poor to fair
$0.61 \leq \kappa \leq 0.80$	Substantial agreement	Fair to good
$0.81 \leq \kappa \leq 1.00$	Almost perfect agreement	Good

The usage of Kappa in research is not without controversy. Authors like Sim et al. [Sim05] argue that: "[...], *the magnitude of kappa is influenced by factors such as* [...] *the number of categories* [...]". Furthermore, Kappa is generally not easy to interpret in terms of the precision of a single observation, because according to [Reichow11] the standard error of the measurements would be required to interpret its statistical significance. To address this problem Sim et al. propose in [Sim05] multiple evaluations as basis for the construction of a confidence interval around the obtained value of Kappa, to reflect sampling errors.
Both facts (implicit influence of the number of classes as well as the standard error in the measurement) are also considered here. In the statistical confidence mapping introduced for this thesis the first fact should be negligible for the practical investigations, because the number of classes considered (and therefore assumedly also their implicit influence) only varies within a rather small interval. Regarding the second fact, here the actual classes in the investigations are actually known which solves part of this problem. Regarding the precision, it is assumed here (based on the achieved evaluation results in initial tests) that it is high enough to allow for meaningful investigations (i.e. the corresponding confidence interval would be suitably small). Nevertheless, the exact precision will have to be established in future work.
In despite of the drawbacks that might be attached to the usage of Kappa, Sim et al. [Sim05] argue that: "*If used and interpreted appropriately, the kappa coefficient provides valuable information on the reliability of data obtained with diagnostic and other procedures* [...]." – which is exactly the motivation why Kappa is used instead of the mere classification accuracy within this thesis.

4.2 Special task and design adaptations for audio steganalysis

The investigation tasks, that are derived in section 3.3 form the Daubert criteria, the general research challenges and the objectives for this thesis, are here elaborated in detail for the application scenario of audio steganalysis.

For **investigation task A** ("Empirical ground truth") the investigations on large sets of audio material marked by different information hiding (IH) algorithms are used to establish:

- Whether SPR-based audio steganalysis is actually possible with the introduced approach

For **investigation task B** ("Investigations on the impact of application scenario specific intrinsic influences to the SPR process") practical investigations are performed within this thesis on:

- Which influence of the number of feature vectors in training has on the detection performance?
- Whether the error rates in a two-class setup are equally balanced?
- Which classifiers (from a pre-existing portfolio provided by WEKA) are suitable to implement the audio steganalysis?
- Which features from the high-dimensional, simple to compute, general purpose audio feature set of the AAFE introduced here are suitable for audio steganalysis?
- How classification using content selection as well as content dependent and independent training and testing influences the detection performance in audio steganalysis?
- Which results are achieved in two-class vs. multi-class setups?

For **investigation task C** ("Investigations on influences outside the SPR process on the performance of the scheme") practical investigations are performed here on:

- Embedding domain and algorithm identification
- The influence of the key scenario in steganography
- Selected common audio post-processing operations (MP3 conversion and de-noising)

The following sections 4.2.1 and 4.2.2 specify which design adaptations (resp. extensions) have been made to implement the corresponding investigations.

4.2.1 Chosen parametrisations for the generation of evaluation data for audio steganalysis

As shown in figure 2.7, the steganographic embedding usually has three input sources: the cover source, the message source and the key source. Additionally, the IH algorithm itself might have user-specifiable parameters that influence the message shaping or the embedding strategy.
The following decisions are made idea for the investigations performed here:

- Regarding the cover source, multi-genre and specific audio test sets (see below) are used to compare the performance of the introduced scheme on different content types.
- The influence of the message source is outside the focus of this thesis, here only fixed messages (either 'UniversityOfMagdeburg' is embedded repeatedly or Goethes' 'Faust' in an ASCII representation is used as message).
- The key influence is included in the investigations performed here. A fixed-key scenario (using the key 'UniversityOfMagdeburg') and a variable-key (using the MD5-hash value of the filename for each file in a test set as the key for embedding) scenario are compared.

- Regarding the user-specifiable parameters for the IH algorithms, here all schemes (see below) are used with their default parametrisation.

On the following pages, detailed descriptions on used audio test sets, Information Hiding (IH) algorithms and post-processing operations are given.
At the end of the section, a link points to the corresponding experimental setup descriptions in appendix B (starting on page 197).

Audio test sets

As a large scale multi-genre audio test set the *aats389* introduced in [Kraetzer06b] is used. It consists of 389 files classified into four main categories (music, sounds, speech and SQAM). All audio files are PCM encoded WAVE files with 44100 Hz sampling rate, 16 Bit quantisation and 2 channels (stereo) (audio CD format). They have an average duration of about 30 seconds. In the category music are 267 files which are distributed into ten sub-categories (metal (20 files), pop (20), reggae (20), blues (20), jazz (20), techno (20), hip-hop (20), country (20), synthetic (20) and classical). The sub-category classical music (with 87 audio files) is again sub-divided into choir (8 files), string quartet (18), orchestra (21), single instruments (19) and opera (19). The main category sounds is divided into four sub-categories (computer generated (12 files), natural (8), silence (2) and noise (11)). The main category speech has four sub-categories (male (24 files), female (20), computer generated (20) and sports (11)). The main category SQAM [Waters88] contains 16 audio files (9 voice and 7 for instrumental). For more details on this audio test set see [Lang07] where this set was extensively used for audio watermarking benchmarking.
With *testset24* a second audio test set is introduced in [Kraetzer09a] for the investigations performed. It has the same genre structure as *aats389*, but contains only a single file in the audio CD format (i.e. PCM encoded, 16 Bit quantisation, sampled at 44.1 kHz, stereo) per genre.
A third multi-genre testset used is the set *ref10* described in section 4.3.1.
Further test sets used are the two speech-only testsets *longfile* and *ahss1* containing only human speech. The first contains only one long audio file (characteristics: duration 27 minutes 24 seconds, sampling rate 44.1 kHz, stereo, 16 bit quantisation in an uncompressed, PCM encoded WAV-file; [Kraetzer07a]) containing only speech signals of one speaker. The latter contains ten PCM encoded speech files with an overall duration of 65 minutes [Kraetzer08a]. The files are recorded by microphones in 16 Bit quantisation, 44.1 kHz sampling, and mono.

Information Hiding (IH) algorithms

As described in section 3.2.2, the algorithms considered here for the practical investigations include steganography as well as audio watermarking tools. The first are denoted as A_{S*}, the latter as A_{W*}. Table 4.3 provides a brief description of the algorithms, which are all used within this thesis in their default parametrisation.

Table 4.3: Information hiding (IH) algorithms used in this thesis

ID	Name	Ref.	Description
A_{S1}	AMSL LSBStego (version Heutling051208)	[Kraetzer06b]	A steganographic algorithm developed at the AMSL with the intention to use it in VoIP steganography for PCM encoded VoIP channels. The message is embedded by this time-domain algorithm into the LSBs of the audio samples which are not identified as silence. This algorithm is described in detail in [Dittmann06]. Default parametrisation: embedding strength = 100%, silence detection on, error correction (ECC) off
A_{S2}	Publimark (v.0.1.2)	[Dittmann06]	This steganography algorithm is an open source tool, developed by Guelvouit[55]. For the embedding (in time-domain) it uses an asymetric key scenario. The algorithm is described in detail in [Dittmann06]. Default parametrisation: no user parameters besides the keys and message

Continued on Next Page...

[55] http://www.gleguelv.org/soft/publimark/index.html

Table 4.3 – Continued

ID	Name	Ref.	Description
A_{S3}	WaSpStego	[Kraetzer07a]	WaSpStego is a spread spectrum, wavelet-domain algorithm, embedding ECC secured messages into PCM encoded audio files. The embedding is done by the modification of the signum of the lower third of wavelet coefficients of each block. Detection is done by correlating the signums of these coefficients with the output of the PRNG initialised with the same key as in the embedding case. Default parametrisation: block width = 256, embedding strength = 1%
A_{S4}	Steghide (v.0.4.3)	[Kraetzer06b]	This time-domain algorithm embeds a compressed and encrypted (Rijndael with a key length of 128 bits in cipher block chaining mode) message in audio files. For the embedding, a sequence of positions of samples in the cover file is created for embedding, based on a PRNG initialised with the key. A graph-theoretic matching algorithm is used to find pairs of positions such that exchanging their values has the effect of embedding the corresponding part of the secret data. Because most of the embedding is done by exchanging sample values it is implied that the first-order statistics (i.e. the number of times a value occurs in the file) is not changed. Default parametrisation: no user parameters besides the keys and message
A_{S5}	Steghide (v.0.5.1)	[Kraetzer06b]	Modified version of A_{S4}. Default parametrisation: no user parameters besides the keys and message
A_{W1}	AMSL Spread Spectrum Watermarking	[Dittmann06]	This frequency-domain watermarking algorithm works embeds the watermark (and ECC information) as sequences into the frequency coefficients. Default parametrisation: ECC on, lower frequency bound = 2000 Hz, upper frequency bound = 17000 Hz, strength = 50000
A_{W2}	2A2W (AMSL Audio Water Wavelet)	[Dittmann06]	This watermarking algorithm works in wavelet-domain and embeds the watermark on selected zero tree nodes. A detailed description on the algorithm is given in [Dittmann06]. Default parametrisation: encoding = binary, method = ZeroTree
A_{W3}	AMSL Least Significant Bit Watermarking	[Dittmann06]	A simple time-domain LSB watermarking algorithm embedding sequentially into all LSBs of the cover. Default parametrisation: ECC = on
A_{W4}	VAWW (Viper Audio Water Wavelet)	[Lang06]	A wavelet-domain embedding the message into selected wavelet coefficients. An detailed description of the algorithm is provided in [Lang06]. Default parametrisation: threshold = 40, scalar = 0.1

Post-processing operations

The MP3 conversion is one of the most widely used audio signal post-processing operations. Here, it is applied with a common bit rate of 128kBit/s (using the LAME codec[56]) to show the impact of this data reduction to the classification performance achieved in audio steganalysis.
De-noising is implemented here by quantisation to 8 Bit resolution and re-quantisation to 16 Bit.

Experimental setup descriptions

The experimental setups used in chapters 5, 6 and 8 are identified in those chapers by underlined and italic font setting (e.g. *AS-D-SF-scaling*). A summary of the experimental setup descriptions (identifying training and test data, classifiers and features used) for the audio steganalysis application scenario is given in table 10.1 in appendix B (starting on page 197). Originally, this summary was part of this section. It has been move into appendix B to improve the accessibility of the core chapters of this thesis.

[56] http://lame.sourceforge.net/

4.2.2 Chosen operators and parametrisations specialised performance metrics for the pattern recognition pipeline for audio steganalysis

Figure 4.1 summarises the **chosen operators and parametrisations for the pattern recognition pipeline** that are applied for the practical investigations on audio steganalysis in this thesis.

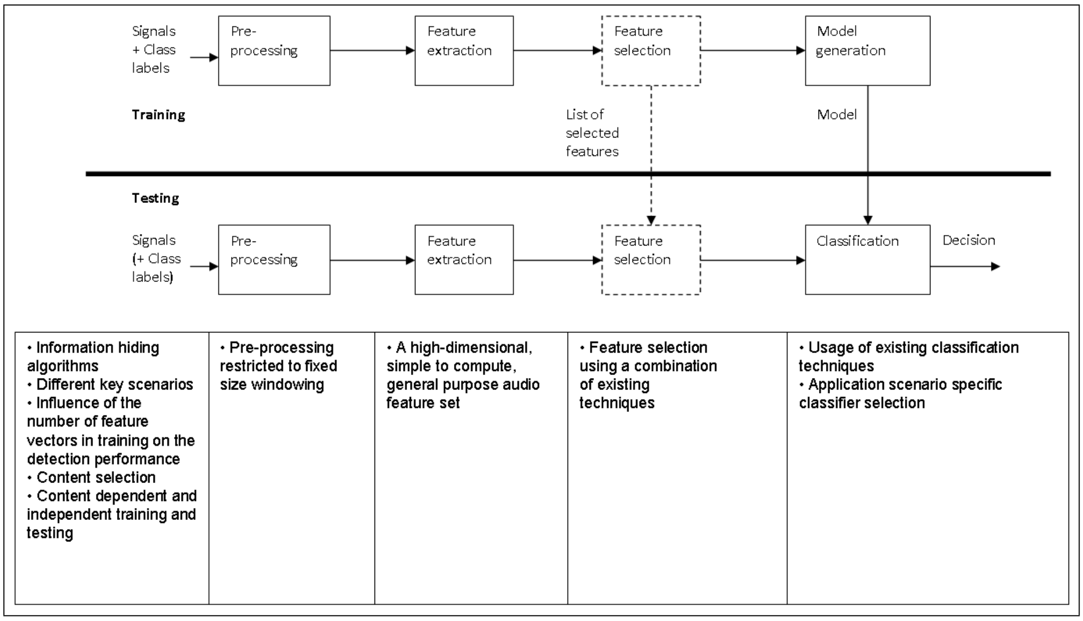

Signals + Class labels
Pre-processing
Feature extraction
Feature selection
Model generation
Training
List of selected features
Model
Testing
Signals (+ Class labels)
Pre-processing
Feature extraction
Feature selection
Classification
Decision
• Information hiding algorithms
• Different key scenarios
• Influence of the number of feature vectors in training on the detection performance
• Content selection
• Content dependent and independent training and testing
• Pre-processing restricted to fixed size windowing
• A high-dimensional, simple to compute, general purpose audio feature set
• Feature selection using a combination of existing techniques
• Usage of existing classification techniques
• Application scenario specific classifier selection

Figure 4.1: Chosen operators and parametrisations for the pattern recognition pipeline for audio steganalysis (based on figure 2.6)

The main idea for the evaluations are: to limit the pre-processing to the absolute minimum (i.e. its possibilities to enhance the performance of the pattern recognition is omitted here and reserved for future work), use the same high-dimensional, simple to compute, general purpose audio feature set as for microphone forensics, and combine existing feature selection and classification algorithms (provided by WEKA) with application scenario specific selection and classification schemes. In detail this means:

- Existing information hiding (IH) algorithms are used to build a representative amount of audio material for the investigations (see section 4.2.1).
- The influence of the number of feature vectors in training on the detection performance is investigated.
- Content selection and content dependent and independent training and testing are investigated as influences to the SPR process
- As evaluation strategies 10-fold stratified cross-validation, percentage split and separate training- and test sets are used
- Pre-processing is restricted to windowing with 1024 samples per non-overlapping, consecutive frame, using Dirichlet window (see section 2.3.1).
- For feature extraction the high-dimensional, simple to compute, general purpose audio feature set of the AAFE (in different version, see section 4.1.1) is used[57].

[57] For one investigation a second feature extractor is used here. It is an audio adaptation of the RS-Analysis (Regular/Singular analysis or dual statistics) approach of Fridrich et al. [Fridrich01], [Ker04] called for the rest of this work *AudioRS*. The implementation used for the corresponding tests is adapted by the author from the ImageRS incorporated by Kathryn Hempstalk into the open source project Digital Invisible Ink Toolkit (http://diit.sourceforge.net/links.html). In contradiction to AAFE, which is primarily an intra-window feature extractor, *AudioRS* is an inter-window (global) feature extractor returning one 19 dimensional feature vector per file instead of one per window. It is only used in a very few evaluations within this thesis Due to its rather low detection performance it was abandoned by the author.

- Feature selection is implemented by a fusion of the feature selectors discussed in section 4.1.2.
- For classification the supervised and unsupervised classification algorithms implemented in WEKA v.3.6.1 (see section 4.1.3) and by *libsvm* [Chang11] are used in their default parametrisations.
- For the classification a $timeout$ boundary of 12 hours (=43,200s) is defined.
- For each atomar WEKA instance 1.6 GByte RAM are allocated.
- Classifier selection is performed on three representative IH algorithms.

The hardware platform for the implementation of the statistical pattern recognition (SPR) solution for this thesis is an array of workstations with a Intel Core 2 Duo E8400 CPU 3GHz with 4 GB RAM, running Microsoft Windows XP, WEKA v.3.6.1 on Java SE 6 (32-bit Windows version) with 1.6 GByte allocated RAM for each WEKA instance (i.e. classifier, clusterer, PCA or feature selector).

For the audio steganalysis application scenario, no **specialised performance metrics** are required for the investigations in this thesis.

4.3 Special task and design adaptations for microphone forensics

The investigation tasks, that are derived in section 3.3 form the Daubert criteria, the general research challenges and the objectives for this thesis, are here elaborated in detail for the application scenario of microphone forensics.

For **investigation task A** ("Empirical ground truth") sets of inhomogeneous (different microphones) and homogeneous (microphones of the same brand and model) recording setups are used to establish:

- Whether SPR-based microphone authentication is actually possible with the introduced approach?

For **investigation task B** ("Investigations on the impact of application scenario specific intrinsic influences to the SPR process") practical investigations are performed within this thesis on:

- Which influence of the number of feature vectors in training has on the detection performance?
- Which classifiers (from a pre-existing portfolio provided by WEKA) are suitable to implement the microphone forensics?
- Which features from the high-dimensional, simple to compute, general purpose audio feature set of the AAFE introduced here are suitable for microphone forensics?
- How classification using content selection as well as content dependent and independent training and testing influences the detection performance in microphone forensics?

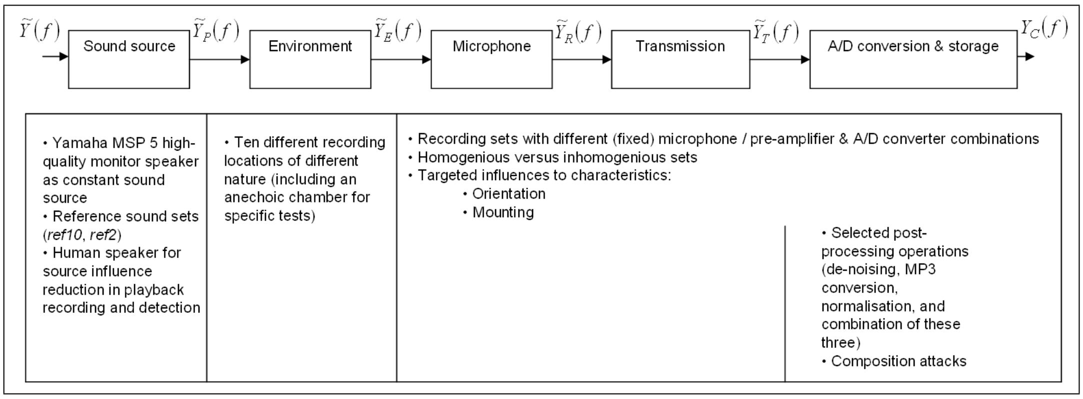
For **investigation task C** ("Investigations on influences outside the SPR process on the performance of the scheme") practical investigations are performed here on the influences of:

- The recording environment
- The microphone orientation
- The mounting of the microphone
- Selected common audio post-processing operations (normalisation, MP3 conversion and denoising)
- Playback recording
- Composition attacks

The following sections 4.3.1 and 4.3.2 specify which design adaptations (resp. extensions) have been made to implement the corresponding investigations.

4.3.1 Chosen parametrisations for the degrees of freedom in the recording process

Figure 4.2 summarises the chosen parametrisations for the degrees of freedom in the recording process.



- Yamaha MSP 5 high-quality monitor speaker as constant sound source
- Reference sound sets (*ref10*, *ref2*)
- Human speaker for source influence reduction in playback recording and detection

- Ten different recording locations of different nature (including an anechoic chamber for specific tests)

- Recording sets with different (fixed) microphone / pre-amplifier & A/D converter combinations
- Homogenious versus inhomogenious sets
- Targeted influences to characteristics:
    - Orientation
    - Mounting

- Selected post-processing operations (de-noising, MP3 conversion, normalisation, and combination of these three)
- Composition attacks

Figure 4.2: Chosen parametrisations for the degrees of freedom in the recording process pipeline (based on figure 2.4)

The main idea for the investigations performed here is to focus on the microphone influence and minimise the remaining degrees of freedom as good as possible. Using the context model for the recording process as introduced in section 2.3.2, this means that:

- To provide control over the input (see equation 2.2) in section 2.3.2, for the majority of the investigations fixed sets of reference signals (*ref10* and *ref2*) are used.
- Live recordings of human speakers are used to reduce the source influence for playback detection investigations.
- For the majority of the investigations exactly one sound source (a very precise Yamaha MSP 5 high-quality monitor loudspeaker) is used to provide a very good transmission function $F_{driver}(f)$ and a minimal noise component $N_{ls}(f)$ for constant playback influences in sequential tests (see equation 2.2).
- A fixed set of 10 different recording locations is used to provide a controlled set of different influences on the environmental shaping (see equation 2.3).
- A (near) perfect recording is created by using an anechoic chamber for minimising e, $N_{envi}(f)$ and the influence of the discolouration function $D(f)$ (see equation 2.3).
- A limited number of microphone and pre-amplifier combinations are used to provide a controlled set of different recording influences $F_{mic}(f)$ and $N_{mic}(f)$ (see equation 2.4; i.e. the intrinsic characteristics used for the microphone forensics approach within this thesis) – here homogeneous as well as inhomogeneous sets of microphones are considered (intra- vs. inter-class variance).
- Ten orientations and eight mounting alternatives are used to investigate on the impact of the orientation and mounting influences (see equation 2.5).
- The influence factor $N_{ENF}(f)$ (equation 2.4) has to be ignored in this thesis, due to lack of corresponding detection and measurement hardware.
- $F_{samp}(f)$ and $N_{quan}(f)$ (equation 2.7), controlled by the analogue to digital conversion, are characteristic for audio signals in CD quality.
- The remaining influence factors ($F_{tran}(f)$, $N_{tran}(f)$, and $N_{thermal}(f)$) (see equations 2.6 and 2.7) are kept constant by a fixed physical setup.
- Selected common signal post-processings (de-noising, MP3 encoding and normalisation) are used for plausibility investigations.

- Four selected composition attack scenarios are used for investigations on the performance of the introduced microphone forensics approach for integrity verification.

On the following pages, detailed descriptions on the reference signals, recording locations, microphone and pre-amplifier combination (recording sets) and post-processing operations are given.
At the end of the section, a link points to the corresponding experimental setup descriptions in appendix C (starting on page 201).

Reference signals

Besides live recordings (only *RS16_ProbM01* and *RS16_ProbM01_playback*), two sets of reference files are used in the microphone forensics experiments conducted within the context of this thesis. These sets are *ref10* and *ref2*.
In *ref10* a number of 10 files (see table 4.4) from the AMSL audio test set described in [Kraetzer06b] were chosen. These reference files represent ten different classes of audio material (music (metal, pop, techno), noise (MLS and white noise), digital silence, a pure harmonic sine at 440 Hz, recorded speech (male and female speaker) and one sample from the SQAM files (Sound Quality Assessment Material; see [Waters88])). All material is provided in 44.1 kHz sampling frequency, 16 Bit quantisation, stereo and PCM coded. The reference files (and the recordings based on these references) have a duration of 30 seconds each.

Table 4.4: The set of reference files *ref10* used in microphone forensics (based on [Kraetzer07c])

Test file	Genre
Metallica-Fuel.wav	music/metal
U2-BeautifulDay.wav	music/pop
Scooter-HowMuchIsTheFish.wav	music/techno
mls.wav	sounds/noise
sine440.wav	sounds/noise
white.wav	sounds/noise
silence.wav	sounds/silence
MariaG-afewboys_nor.wav	speech/female
andreas-D2.wav	speech/male
vioo10_2_nor.wav	sqam/instrumental

The files are played in every room in *R** using a notebook computer and a Yamaha MSP 5 high-quality monitor speaker and the sound is recorded by the microphones in the corresponding recording set (see below). These microphones are in most sets (except *RS7*, *RS8* and *RS9*) mounted in a fixed position together with the notebook, the speaker and the used preamplifiers on a trolley to provide mobility for the fixed set-up.

The set *ref2* is a subset of *ref10* containing only the silence and harmonic sine at 440 Hz parts.

Recording locations

Table 4.5: Microphone forensics recording locations *R01*, *R02*, ..., *R10* (based on [Kraetzer07c])

Recording location *R**	Room number	Description
R01	29R114	large office
R02	29R131	small office
R03	29R140	bathroom
R04	29R146	laboratory
R05	29R307	large lecture hall
R06	audiobox	anechoic chamber
R07	outside1	quiet outside environment
R08	outside2	busy parking lot
R09	corridor	long and narrow corridor
R10	stairs	stone stairwell with strong echo

Since the beginning of the practical investigations on microphone forensics reflected in this thesis, the same set of ten recording locations (rooms, *R01*, *R02*, ..., *R10*) has been used. This set is described in detail in [Kraetzer07c]. It consists of eight rooms and two outside locations of the main building of the Faculty of Computer Science, Otto-von-Guericke University Magdeburg. A plan of these recording locations is shown in figure 4.3.

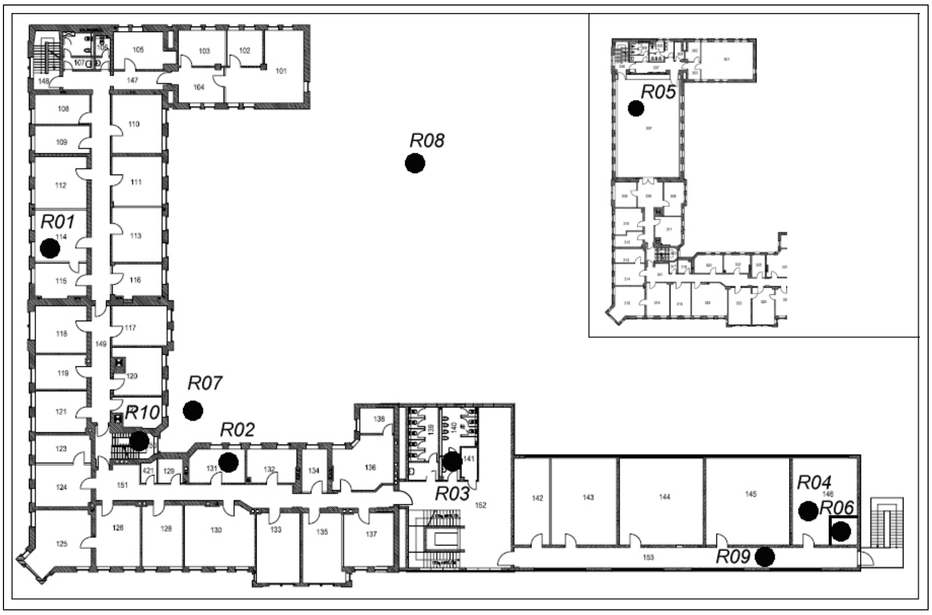


Figure 4.3: Microphone forensics recording locations *R01*, *R02*, ..., *R10* (taken from [Kraetzer07c])

Table 4.5 summarises the description of the recording locations, as introduced in [Kraetzer07c].

Recording sets

For the practical investigations performed in microphone forensics nine different recording sets *RS** have been created: *RS1*, *RS2*, *RS4_Beyer*, *RS4_Rode*, *RS7*, *RS8*, *RS9*, *RS16_ProbM01*, and *RS16_ProbM01_playback*. The apparent gaps in the naming scheme result from the intention of the author to remain compliant with the naming schemes used in the paper publications accompanying this thesis (see the introduction text in chapter 6).
A recording set is described by a microphone M_* and pre-amplifier/soundcard combination, see table 4.6.

Table 4.6: The recording sets *RS** used for microphone forensics

Identifier	Microphone(s)	Pre-amplifier/soundcard
RS1		
M_1	AKG SE 300 B	Millenium Mic 1
M_2	TerraTec HeadsetMaster	Creative Sound Blaster USB
M_3	Shure SM58	Creative Sound Blaster USB
M_4	Tbone T.bone MB45	Millenium Mic 1
RS2		
M_2	TerraTec HeadsetMaster	Creative Sound Blaster USB
M_5	PUX 70TX-M1	UBC 60XLT-2 receiver connected to a Creative Sound Blaster USB
M_3	Shure SM58	Creative Sound Blaster USB
M_6	T.bone MB45	Creative Sound Blaster USB
M_7	AKG SE 300 B (CK93)	Creative Sound Blaster USB
M_8	AKG SE 300 B (CK98)	Creative Sound Blaster USB
M_9	AKG SC600	Creative Sound Blaster USB
RS4_Rode		
M_{16}	Rode NT6	Presonus FireStudio Project 8-port soundcard
M_{17}	Rode NT6	
M_{18}	Rode NT6	

Continued on Next Page. . .

Table 4.6 – Continued

Identifier	Microphone(s)	Pre-amplifier/soundcard
M_{19}	Rode NT6	
RS4_Beyer		
M_{20}	Beyerdynamic Opus 89	Presonus FireStudio Project 8-port soundcard
M_{21}	Beyerdynamic Opus 89	
M_{22}	Beyerdynamic Opus 89	
M_{23}	Beyerdynamic Opus 89	
RS7, *RS8* and *RS9*		
M_{22}	Beyerdynamic Opus 89	Presonus FireStudio Project 8-port soundcard
RS16_ProbM01 and *RS16_ProbM01_playback*		
M_{33}	Integrated microphone of an Audio Advantage Roadie USB soundcard	
M_{34}	Integrated microphone of an Audio Advantage Roadie USB soundcard	
M_{35}	Plantronics (Head-set Master)	Creative Sound Blaster USB
M_{36}	No-name head-set	Creative Sound Blaster USB
M_{37}	Beyerdynamics Opus 89	Presonus Firestudio (line-in)
M_{38}	PUX 70TX-M1 Piezoelectric surveillance microphone	UBC 60XLT-2 receiver connected to the recording notebooks microphone input

The recording set *RS1* is the original recording set from [Kraetzer07c]. It contains four different microphones that are used to record simultaneously the reference signals (see above) played with a Yamaha MSP 5 high-quality monitor speaker.

RS2 is the recording set created for [Buchholz09]. It contains seven different microphones which are used for sequential recording (all on a Sound Blaster USB as pre-amplifier). It contains three dynamic microphones (TerraTec HeadsetMaster, Shure SM58, and T.bone MB45), three condenser microphones (AKG CK93, AKG CK98, and T.bone SC600), and a piezoelectric microphone (the PUX 70TX-M1).

The two sets *RS4_Rode* and *RS4_Beyer* both contain a set of four identical microphones and both recorded in parallel (time synchronous) using a Presonus FireStudio Project 8-port Firewire soundcard. The *RS4_Rode* represents a homogeneous set of four Rode NT6 condenser microphones, while *RS4_Beyer* is a homogeneous set of four Beyerdynamic Opus 89 dynamic microphones. Thereby, the tests performed on *RS4_Rode* and *RS4_Beyer* cover the two most common microphone types in intra-class evaluations. The results can be assumed to be of stronger significance than those achieved on mixed class sets like *RS1* or *RS2*.

The recording set *RS7* is recorded in the anechoic chamber (room *R06*), using the Beyerdynamic Opus 89 microphone M_{22}. By this microphone two different reference sounds (a harmonic sinusoid at 440 Hz and silence) are recorded in eight different microphone orientations, each with 45° offset in the xy-plane from its predecessor, stating with the orientation directly towards the sound generating loudspeaker.

RS8 uses the same microphone as *RS7* but here two microphone orientations with 180° offset in the yz-plane are recorded by M_{22} in *R06*.

The recording set *RS9* also records with M_{22} in the anechoic chamber *R06*. Here, two different reference sounds (a harmonic sinusoid at 440 Hz and silence) are recorded in eight different microphone mounting positions. The distance (50cm) and orientation to the loudspeaker are kept constant in these tests.

The recording set *RS16_ProbM01* is generated by a human speaker reading a text in front of an array of dynamic (M_{33}, M_{34}, M_{35}, and M_{36}), condenser (M_{37}) and piezoelectric (M_{38}) microphones. The recordings with this set are played back using a Yamaha MSP 5 high-quality monitor speaker and the playback is recorded with the same hardware (recording set *RS16_ProbM01_playback*).

For *RS1*, *RS2*, *RS4_Beyer*, and *RS4_Rode* recordings are generated in each of the 10 recording environments *R01*, *R02*, ..., *R10* specified above. For *RS7*, *RS8*, *RS9*, *RS16_ProbM01*, and *RS16_ProbM01_playback*, which are used to investigate specific influence factors to the recording process, recordings are only generated in the anechoic chamber *R06*.

Post-processing operations and attack scenarios for composition attack investigations

Normalisation is a fairly common audio signal post-processing operation. Since it is an amplitude scaling operation in time domain, normalisation is considered here to be a representative for all such operations. It is implemented here with a normalisation factor computed independently for each file in

the material under investigation.
The MP3 conversion is one of the most widely used audio signal post-processing operations. Here, it is applied with a common bit rate of 128kBit/s (using the LAME codec[58]) to show the impact of this data reduction to the classification performance achieved in microphone forensics.
De-noising is implemented here by quantisation to 8 Bit resolution and re-quantisation to 16 Bit.

Four different composition tests are performed in this thesis:

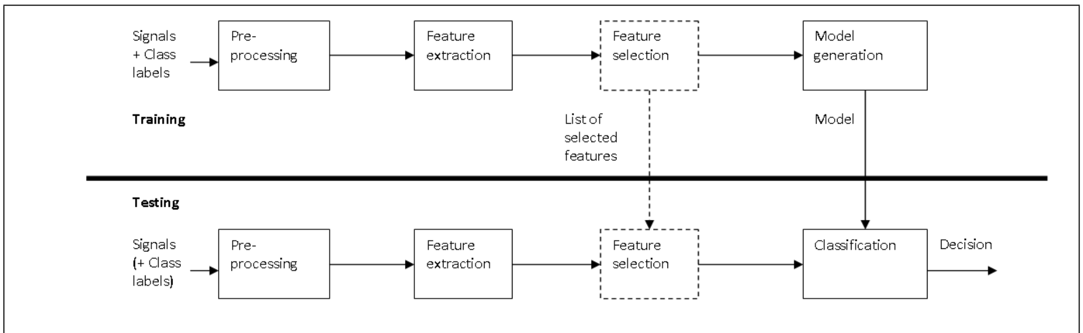
- Microphone recordings of one known microphone made in different locations composed into one stream.
- One known microphone pasted into a stream of completely different known microphone.
- One unknown microphone pasted into a stream of completely different known microphone.
- One unknown microphone pasted into a stream of completely different unknown microphone.

Experimental setup descriptions

The experimental setups used in chapters 5, 6 and 8 are identified in those chapers by underlined and italic font setting (e.g. *Mic-01*). A summary of the experimental setup descriptions (identifying training and test data, classifiers and features used) for the microphone forensics application scenario is given in table 11.1 in appendix C (starting on page 201). Originally, this summary was part of this section. It has been move into appendix C to improve the accessibility of the core chapters of this thesis.

4.3.2 Chosen operators and parametrisations for the pattern recognition pipeline and specialised performance metrics for microphone forensics

Figure 4.4 summarises the **chosen operators and parametrisations for the pattern recognition pipeline** that are applied for the practical investigations on microphone forensics in this thesis.



• Microphones (Recording sets) • Recording locations • Influence of the number of feature vectors in training on the detection performance • Content selection • Content dependent and independent training and testing	• Pre-processing restricted to fixed size windowing	• A high-dimensional, simple to compute, general purpose audio feature set	• Feature selection using a combination of existing techniques	• Usage of existing classification techniques • Application scenario specific classifier selection

Figure 4.4: Chosen operators and parametrisations for the pattern recognition pipeline for microphone forensics (based on figure 2.6)

The main ideas for the evaluations are: to limit the pre-processing to the absolute minimum (i.e. its possibilities to enhance the performance of the pattern recognition is omitted here and reserved for

[58] http://lame.sourceforge.net/

future work), use the same high-dimensional, simple to compute, general purpose audio feature set as for audio steganalysis, and combine existing feature selection and classification algorithms (provided by WEKA) with application scenario specific selection and classification schemes. In detail this means:

- The microphones (recording sets) and recording locations are used to build a representative amount of audio material for the investigations (see section 4.3.1).
- The influence of the number of feature vectors in training on the detection performance is investigated.
- Content selection and content dependent and independent training and testing are investigated as influences to the SPR process.
- As evaluation strategies 10-fold stratified cross-validation, percentage split and separate training- and test sets are used.
- Pre-processing is restricted to windowing with 1024 samples per non-overlapping, consecutive frame, using Dirichlet window (see section 2.3.1).
- For feature extraction the high-dimensional, simple to compute, general purpose audio feature set of the AAFE (in differend version, see section 4.1.1) is used.
- Feature selection is implemented by a fusion of the feature selectors discussed in section 4.1.2.
- For classification the supervised and unsupervised classification algorithms implemented in WEKA v.3.6.1 (see section 4.1.3) are used in their default parametrisations.
- For the classification a $timeout$ boundary of 60 hours (=216,000s) is defined.
- For each atomar WEKA instance 1.6 GByte RAM are allocated.
- Classifier selection is performed on two homogeneous recording sets (*RS4_Rode* and *RS_Beyer* – see section 4.3.1).

The hardware platform for the implementation of the statistical pattern recognition (SPR) solution for this thesis is an array of workstations with a Intel Core 2 Duo E8400 CPU 3GHz with 4 GB RAM, running Microsoft Windows XP, WEKA v.3.6.1 on Java SE 6 (32-bit Windows version) with 1.6 GByte allocated RAM for each WEKA instance (i.e. classifier, clusterer, PCA or feature selector).

Regarding the **specialised performance metrics** for microphone forensics, besides the κ statistics (and *accuracy*; see section 4.1.4) an extension has to be made for the integrity verification tests using composition attacks. For these tests the change rate and the average sequence length in class assignment of an authentication attempt are used to give an indication on the potential integrity violation in an audio recording. The smaller the change rate and the higher the average sequence length in the classifications, the better the classification under these circumstances (i.e. a (suitably) trained classifier which achieves a high detection performance in authentication assigns the frames to a consistent source).

5 Investigations for Application Scenario 1: Audio Steganalysis

This chapter is dedicated to the experimental evaluation of the performance of an instantiation of the introduced general-purpose statistical pattern recognition (SPR) based audio forensics approach for audio steganalysis. It is structured along the investigation tasks A) to C) defined in section 3.3.

The **empirical ground truth** requested in investigation task A) is established for the audio steganalysis approach developed in this thesis in section 5.1. For the performed investigations, this section will show the statistical relevance of the introduced solution approach as well as provide required knowledge for the following evaluations.
In section 5.2 the **impact of steganalysis specific influences to the statistical pattern recognition (SPR) process** is considered (as part of investigation task B)) by investigations on the embedding domain, the key-scenario used, context dependent and independent training and testing as well as the dimensionality of the evaluation setup (i.e. modelling steganalysis as two-class or multi-class problem). For investigation task C) (**Influences to the performance of the scheme, which are outside the SPR process**) in section 5.3 the persistence of the introduced approach against selected post-processing operations is investigated.

At the end of the chapter, the major results of the investigations performed within this chapter are summarised in section 5.4. This summary includes a mapping of the progress made on this application scenario to investigation tasks (as defined in section 3.3) as well as to the Daubert criteria as specified in section 2.2 and its subsections.

As usual for a dissertation project in the field of computer science, in compliance with Daubert criterion DC2 ("*whether the technique or theory has been subject to peer review and publication*" [USC93]) and to give other researchers / reviewers the chance to dispute the theory and its application (Daubert criterion DC5 "*whether the technique or theory has been generally accepted in the scientific community*" [USC93]), parts of the results presented in this chapter have been previously published in workshop, conference proceedings, two technical reports as well as a journal publication. The corresponding papers are (in chronological order):

- **2006**:
  - [Kraetzer06c] presented at the IEEE International Symposium on Circuits and Systems in Kos, Greece, May 21th-24th, 2006.
  - [Kraetzer06a] presented at the BSI-Workshop IT-Frühwarnsysteme, Bonn, Germany, July 12th, 2006.
- **2007**:
  - [Dittmann07] technical report ECRYPT D.WVL.16 Report on Watermarking Benchmarking and Steganalysis, 2007.

  - [Kraetzer07a] presented at the SPIE conference Security, Steganography, and Watermarking of Multimedia Contents IV, IS&T/SPIE Symposium on Electronic Imaging, January 28th-February 1st, in San Jose, CA, USA, 2007.
  - [Kraetzer07b] presented at Information Hiding 2007, June 11th-13th, in St. Malo, France, 2007.

- **2008**:
  - [Kraetzer08a] presented at the SPIE conference Security, Forensics, Steganography, and Watermarking of Multimedia Contents X. Electronic Imaging Conference 6819, IS&T/SPIE 20th Annual Symposium, in San Jose, CA, USA, January 26th-31st, 2008.
  - [Kraetzer08b] presented at the 10th ACM Workshop on Multimedia and Security, September 22nd-23rd, in Oxford, UK, 2008.
- **2009**:
  - [Kraetzer09a] presented at the SPIE conference Media Forensics and Security XI. Electronic Imaging Conference 7254, IS&T/SPIE 21st Annual Symposium, in San Jose, CA, USA, January 18th-22nd, 2009.
- **2010**:
  - [Kraetzer10] presented at the SPIE conference Multimedia on Mobile Devices 2010, Electronic Imaging Conference 7542, IS&T/SPIE 22nd Annual Symposium, in San Jose, CA, USA, January 18th and 19th, 2010.
- **2012**:
  - [Kraetzer12a] published in the journal Transactions on Data Hiding and Multimedia Security VIII, Lecture Notes in Computer Science, Vol. 7228, Springer Berlin / Heidelberg, ISBN: 978-3-642-31970-9, pp. 80-101, 2012.

The major results from these publications are recapitulated in the following sections, where they are further substantiated and accompanied by additional investigations as necessary.

5.1 Establishing some empirical truth for the used audio steganalysis approach

Within this section some basic empirical results are presented, fulfilling investigations task A) as defined in section 3.3. These basic considerations focus in section 5.1.1 on the question whether applying statistical pattern recognition (SPR) for audio steganalysis is feasible in the first place. In the following section 5.1.2 considerations are made on evaluation sizes (in terms of feature vectors used for training) required to achieve reliable answers in the investigations performed. Section 5.1.3 presents investigations on the detector on completely unmarked material, giving an indication on the tendency of the introduced approach to generate false positive errors. In sections 5.1.4 and 5.1.5 application scenario specific classifier and feature selection operations are performed with the aim of identifying suitable candidates for the following investigations.

5.1.1 Detection performance

In [Kraetzer07a], the first publication which uses the audio steganalysis approach considered in this thesis, in classical two-class setups (see experimental setup[59] *AS-Kraetzer2007SPIE-summary*) κ values between 0.142 and 0.950 (detection accuracies between 57.1% and 97.5%) are achieved. The setup for the detection of the nine information hiding (IH) algorithms evaluated there is using AAFE v.1.0.3 (with

[59] The experimental setups used in chapters 5, 6 and 8 are identified by underlined and italic font setting (e.g. *Mic-01*) – they are resolved in appendixes B (audio steganalysis) and C (microphone forensics).

its 63 dimensional feature vectors), a training set size of 64 feature vectors per file of the multi-genre audio set *aats389* and using another 16 feature vectors per file for testing. Table 5.1 identifies the best Kappa values κ achieved for each algorithm in an experiment testing different sub-sets in this feature space. The corresponding feature sets used to obtain the input for the classifications vary, since they are a result of a first and rather coarse feature selection (*AS-Kraetzer2007SPIE-summary* – see table 10.1 in appendix B (starting on page 197)). The feature sets used to achieve the results presented in table 5.1 are for A_{S1} and A_{S5} only the time-domain features of AAFE v.1.0.3. For A_{S2}, A_{S3}, A_{S4}, A_{W1}, A_{W2}, A_{W3}, A_{W4} the combination of time-domain features and FMFCCs shows the best detection performance (see [Kraetzer07a]).

Table 5.1: Kappa values for the best performing detections on the nine IH algorithms in [Kraetzer07a] – setup *AS-Kraetzer2007SPIE-summary*; κ computed from the accuracies reported in [Kraetzer07a]

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
κ	0.142	0.198	0.344	0.214	0.224	0.950	0.432	0.212	0.190

The results presented in table 5.1 show for every evaluated IH algorithm a detection performance that is better than the probability of guessing correctly (which would be equivalent to $\kappa = 0.000$). Nevertheless, the detection performance strongly differs between the different IH algorithms for the used feature set. With $\kappa = 0.142$ the lowest detection performance is achieved for the LSB steganography algorithm A_{S1} (here, the silence detection of the embedding function, which is the major difference between the embedding in A_{S1} and A_{W3}, obviously has strong influence to the detectability), while the best performance is seen with $\kappa = 0.950$ for the spread spectrum watermarking algorithm A_{W1}.

In [Kraetzer08b], which uses the same set of nine IH algorithms, the accuracies presented in table 5.1 are confirmed by similar results using AAFE v.1.0.4 calculating 98 features and 256 windows per file for training and 64 windows per file for testing on the multi-genre audio set *aats389*.

For later publications on this approach (e.g. [Kraetzer09a] and [Kraetzer10]) the mostly information fusion and plausibility focussed investigations are narrowed down from a selection of nine IH algorithms to a representative[60] set of three algorithms (A_{S1}, A_{S3} and A_{W1}). The single classifier accuracies achieved in these later publications (with AAFE v.1.0.4 in [Kraetzer09a] and v.2.0.5 with its 590 segmental and 17 global features in [Kraetzer10]) on multi-genre audio sets are further documenting the potential of the introduced approach of SPR-based audio steganalysis based on the usage of a general-purpose audio feature extractor like the AAFE.

In table 5.2 a selection from the single classifier evaluations performed in [Kraetzer10] is presented. It compares for the global features of AAFE v.2.0.5 two different evaluation strategies (10-fold stratified cross-validation over 40 windows per file as well as training and testing with 40 windows per file from *aats389* for training and 40 windows per file from *testset24* for testing – see experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary* in table 10.1 in appendix B (starting on page 197)). The results show, on one hand, that the chosen approach works quite well, but on the other hand, it also quite well identifies one of the major problems of the SPR approach: the dependency of the detection performance on the availability of representative training material for the building of the classifier model.

Table 5.2 summarises for the three exemplary selected IH algorithms A_{S1}, A_{S3} and A_{W1} as well as the two different evaluation strategies (10-fold stratified cross-validation as well as training and testing with independent sets) the classification performance based on the AAFE v.2.0.5 and global features. A more complete analysis of these evaluations is presented in the classifier selection focussed section 5.1.4. The most significant case is seen in this table for A_{W1} in case of the cross-validation where out of WEKAs (v.3.6.1) 74 classifiers 20 show a κ better than 0.8 (equivalent in this setup with *accuracy* $> 90\%$), 24 further classifiers achieve a κ between 0.6 and 0.8 and additional 7 classifiers perform

[60] Representative in this context means that time-domain, frequency-domain and wavelet-domain embedding strategies are still covered, as well as the set still including steganography as well as audio watermarking algorithms.

with a κ between 0.2 and 0.6. The best classification performance achieved in this test is $\kappa = 0.85$ (*accuracy*$=92.68\%$).

The least significant result is seen for A_{S1} in case of the training and testing with independent sets scenario with the best classifier achieving only $\kappa = 0.15$ (*accuracy* $= 57.29\%$). The differences between the algorithms differ strongly in terms of the detectability of the algorithms. The fact, established in [Kraetzer10], that the watermarking algorithms in the set are generally better detectable than the steganographic algorithms, is supported here.

Table 5.2: Comparison of the detection performance achieved with the AAFE v.2.0.5 global features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*)

	10-fold strat. cross-valid.			**Training and testing**		
	A_{S1}	A_{W1}	A_{S3}	A_{S1}	A_{W1}	A_{S3}
Maximum achieved κ value	0.72	0.85	0.38	0.15	0.50	0.19
Maximum achieved accuracy	86.15%	92.68%	68.94%	57.29%	75.00%	59.38%
Performance histogram:						
$0.2 \leq \kappa < 0.6$ ($60 \leq$ *accuracy* $< 80\%$)	2	7	40	0	40	0
$0.6 \leq \kappa < 0.8$ ($80 \leq$ *accuracy* $< 90\%$)	3	24	0	0	0	0
$0.8 \leq \kappa \leq 1.0$ ($90 \leq$ *accuracy* $< 100\%$)	0	20	0	0	0	0

Regarding the differences between the evaluation strategies a more complex statement is presented by the results in table 5.2. For these investigations, the cross-validation based evaluations in general show better detection performances than the training and testing with independent sets. On a first glance, it seems that the cross-validation is the more appropriate evaluation method because it shows more significant results and is assumed here to be less overfitting. Nevertheless, this evaluation scenario is rather unrealistic since it implicitly assumes access to unmarked versions of the candidate audio signals in an investigation, which at the same point of time would make the complete steganographic process superfluous because the investigator in this case could perform a much simpler and more reliable difference analysis. Therefore the results for the training and testing with independent sets have to be considered in this thesis to be the more relevant – or closer to the practical constraints of a steganography implementing security mechanism.

Equivalent to the global feature results presented above in table 5.2, in table 5.3 the classification performance based on AAFE v.2.0.5 segmental features for detection of the three exemplary selected IH algorithms A_{S1}, A_{S3} and A_{W1} as well as the two different evaluation strategies (10-fold stratified cross-validation as well as training and testing with independent sets) is presented.

Table 5.3: Comparison of the detection performance achieved with the AAFE v.2.0.5 segmental features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *AS-Kraetzer2010SPIE-SF-singleClass-summary*)

	10-fold strat. cross-valid.			**Training and testing**		
	A_{S1}	A_{W1}	A_{S3}	A_{S1}	A_{W1}	A_{S3}
Maximum achieved κ value	0.88	0.93	0.41	0.00	0.79	0.22
Maximum achieved accuracy	94.19%	96.47%	70.36%	50.12%	89.48%	61.22%
Performance histogram:						
$0.2 \leq \kappa < 0.6$ ($60 \leq$ *accuracy* $< 80\%$)	1	4	12	0	5	6
$0.6 \leq \kappa < 0.8$ ($80 \leq$ *accuracy* $< 90\%$)	0	3	0	0	31	0
$0.8 \leq \kappa \leq 1.0$ ($90 \leq$ *accuracy* $< 100\%$)	4	30	0	0	0	0

For the inter-algorithm comparisons the steganographic algorithms A_{S1} and A_{S3} are again much less detectable than the watermarking algorithm A_{W1}, even when the cross-validation based evaluation identifies some classifiers which achieve a $\kappa \geq 0.8$ (an *accuracy* of more than 90%). If the cross-validation and two-set training and testing results are compared the same observation can be made as for the global features: it seems that the cross-validation is the more appropriate evaluation method because it shows more significant results. Yet again, the training and testing with independent sets is considered in this thesis to be closer to the practical constraints of a steganography implementing security mechanism. Unfortunately, using this strategy the investigations on A_{S1} are unable to present any satisfactory results ($\kappa = 0$) for all used classifiers.

Comparing the global and segmental feature results (table 5.2 and table 5.3 respectively), it can be stated that while both classes of features allow for significant classification accuracies, the segmental features seem to outperform the global features in most cases – a more detailed analysis of the performance of the two feature classes in audio steganalysis is presented in sections 5.1.4 and 5.1.5.

Résumé for this section: The investigations performed in this section can be summarised as follows: Notwithstanding the fact that all classifiers are used in default parametrisation – which has to be assumed to be sub-optimal (a fact which would require more detailed considerations on classifier optimisation and -generation, which are outside the scope of this thesis) – the results presented for statistical pattern recognition (SPR) based audio steganalysis, as it is used within this thesis, can be considered to support the assumption that SPR can indeed be used to solve this media forensics application scenario. From the two evaluated test scenarios (10-fold stratified cross-validation and training and testing with independent sets), the cross-validation is assumed here to be less suitable, because it implicitly assumes access to unmarked covers, which seems to hardly feasible for audio steganalysis cases. The strong variance in the reported detection performances implies the need for the following sections to investigate more closely on the internal and external influences to the detection performance.

5.1.2 Influence of the number of feature vectors in training

In [Kraetzer07a] it is implied for A_{S1} using AAFE v.1.0.3 that larger model sizes outperform smaller model sizes in terms of resulting detection performance. The test performed there are run on one hand with 80 feature vectors per file (split in the ratio 64 for training and 16 for testing) and on the other hand with 2600 feature vectors per file (comparing the test cases of 400 feature vectors per file for training and 2200 for testing with the inverted ratio of 2200 feature vectors per file for training and 400 for testing). Unfortunately, different test sets are used for these two tests (*aats389* for the first vs. *longfile* for the second) so that these results cannot be directly compared. Nevertheless, the results indicate an increase of the classification accuracy even when the training set size exceeds 2000 feature vectors per file.

A more substantial and generalisable investigation on the scaling behaviour is performed in [Kraetzer07b] also using the AAFE v.1.0.3 with its 63 dimensional feature vectors (see setup *AS-Kraetzer2007IH-scaling* in table 10.1 in appendix B (starting on page 197)). There the multi-genre test set *aats389* is used for detection performance evaluations for three different training- and test set sizes (16, 64 and 256 feature vectors per file for training and 4, 16 and 64 feature vectors per file for testing). The results of this evaluation are summarised in table 5.4 below.

Table 5.4: Detection performance (κ values) for different model sizes, using from the 63 dimensional feature set of AAFE v.1.0.3 only the time-domain features and the FMFCCs (see experimental setup *AS-Kraetzer2007IH-scaling*, results adapted from [Kraetzer07b]

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
κ value for tr.16 vs. te.4	0.032	0.210	0.273	0.195	0.196	0.869	0.408	0.183	0.115
κ value for tr.64 vs. te.16	0.129	0.199	0.344	0.213	0.217	0.950	0.433	0.211	0.190
κ value for tr.256 vs. te.64	0.244	0.244	0.459	0.244	0.241	0.982	0.497	0.220	0.278

The results presented in table 5.4 show for eight out of the nine IH algorithms a homogeneous increase in the detection performance with an increasing training set size. Only the steganography algorithm A_{S2} presents a slight exception from this rule.

The speech-only test-set *longfile* was used in [Kraetzer07a] to simulate the application scenario of VoIP steganography using the steganography algorithm A_{S1}. This simulation is extended in [Kraetzer07b] to the other IH algorithms evaluated there (see experimental setup *AS-Kraetzer2007IH-scaling_VoIP* in table 10.1 in appendix B (starting on page 197)). Unfortunately the algorithms A_{S4}, A_{S5}, A_{W2}

and A_{W3} were not capable of marking the extremely long audio file composing the test set *longfile*. In the case of A_{W2} the embedding process was terminated with a 'segmentation fault', in the case of A_{W3} and A_{S5} the embedding function terminated with the message 'aborted' without generating the marked output file. For A_{S4} the embedding process was aborted manually after running 40 hours without termination or showing any form of progress. The behaviour of those four algorithms (which is considered to be a result of the extreme file size) is marked in table 5.5 with 'n.a.' (result not available).

Table 5.5: Detection performance (κ values) for different model sizes, using from the 63 dimensional feature set of AAFE v.1.0.3 only the time-domain features and the FMFCCs (see experimental setup *AS-Kraetzer2007IH-scaling_VoIP*, results adapted from [Kraetzer07b])

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
κ value for tr.400 vs. te.2200	0.939	0.922	0.915	n.a.	n.a.	1.000	n.a.	n.a.	0.944
κ value for tr.2200 vs. te.400	1.000	1.000	1.000	n.a.	n.a.	1.000	n.a.	n.a.	1.000

The table 5.5 shows good or even perfect results for the large models used here for classification. The only conclusion, which can be drawn from these test results, is that the ideal model size for this very context limited evaluation seems to be between 400 and 2200 feature vectors per file for the most IH algorithms considered. Only for A_{W1} this number is lower – on the pure speech content in the test set *longfile* it is assumed by the author to be even lower than the size of 256 feature vectors per file evaluated in table 5.4 for the multi-genre test-set *aats389*.
Therefore, these results on the *longfile* test-set with its speech-limited content are of high importance for this thesis. At this model size all five IH algorithms, for which the investigation was possible, show a detection performance of $\kappa > 0.9$ under ideal circumstances[61]. Therefore, this could be used as a first recommendation for a model size for implementing a security mechanism based on a general purpose audio feature extractor like the AAFE v.1.0.3 on known content. If the content is not known or is in its characteristics less ideal than speech signals, a significantly higher number of feature vectors per file in a training set reflecting the application scenario might be required. If the application scenario would be universal steganalysis (suitably represented by the multi-genre audio set *389files*) and we would take the figure of 400 feature vectors per file from above as a rough estimation, then this would result in a vector field of about 155,000 reference vectors for the generation of the classifier models. Considering the feature vector dimensionality involved, this figure would present a highly computational complex problem to any classification algorithm. Even if the model generation could be performed successfully (which is not guaranteed with such large reference data sets) it has to be assumed that the model will be very complex and therefore any classification using this model will be extremely slow.

In [Kraetzer08a] another investigation is performed on the application scenario of (simulated) VoIP steganalysis. There the special purpose audio set *aahs1* (consisting of 10 speech samples from different persons with an average duration of 390s per file; see experimental setup *AS-Kraetzer2008SPIE-VoIP* in table 10.1 in appendix B (starting on page 197)) is used to extend the investigations to all nine IH algorithms evaluated there and to an estimation of the achievable classification performance on a more realistic multi-speaker audio set. Table 5.6 summarises the results achieved in the classification *accuracy* focussed investigations in [Kraetzer08a].

Table 5.6: Detection performance (κ values) for a training-set size of 15000 vs. a test-set size of 1200 per reference file using the complete 98 dimensional feature set of AAFE v.1.0.4 (see experimental setup *AS-Kraetzer2008SPIE-VoIP*, results adapted from [Kraetzer08a])

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
κ value	0.896	0.550	0.996	0.102	0.206	1.000	0.802	0.938	1.000

If the results of table 5.6 are compared to the results presented in table 5.5 it can be noticed that for three out of the five IH algorithms, for which results are present in both tables, the detection performance for the multi-speaker set *aahs1* are lower than for the single-speaker set *longfile*. On a first glance, this seems to contradict the aforementioned assumption that a higher number of feature vectors per file in a training-set should reflect into increased classification accuracy.

[61] Training and test material are drawn from the same source.

On the second glance, this assumed contradiction evaporates if the differences in the investigative setups *AS-Kraetzer2007IH-scaling_VoIP* and *AS-Kraetzer2008SPIE-VoIP* are taken into account. The setups differ in three relevant points[62]: the extracted feature sets, the audio sets used for training and testing and the number of feature vectors used in training.
The feature set extracted in *AS-Kraetzer2007IH-scaling_VoIP* is a direct sub-set of the feature set extracted in *AS-Kraetzer2008SPIE-VoIP*. The feature selection operations performed in [Kraetzer08a] as well as in section 5.1.5 of this thesis show better results for the larger feature space – therefore it is unlikely that this influence factor is responsible for the decrease in the classification performance.
The other two factors, the audio sets used and the number of feature vectors used in training, are most likely both contributing to the decrease in the classification performance. On one hand, the stronger variance in the audio material (ten different speaker characteristics instead of only one) makes the classification problem more complex, while on the other hand the larger number of vectors in training has to be assumed to lead to an overfitting and therefore assumedly also has also a negative effect on the classification accuracies achieved. Unfortunately, the strength of the individual effects cannot be estimated form the investigations performed here.

To give a more fine-granular analysis of the influence of the training set size, the setup from [Kraetzer10] is re-used here to perform a measurement of the achieved detection performance. Figure 5.1 summarises the detection performances achieved for step sizes ranging from 3112 feature vectors for training (equivalent to 4 frames per audio file in the evaluation set or 72 seconds of audio material in total; see experimental setup *AS-D-SF-scaling*) up to 24896 feature vectors (equivalent to 578 seconds of audio material).

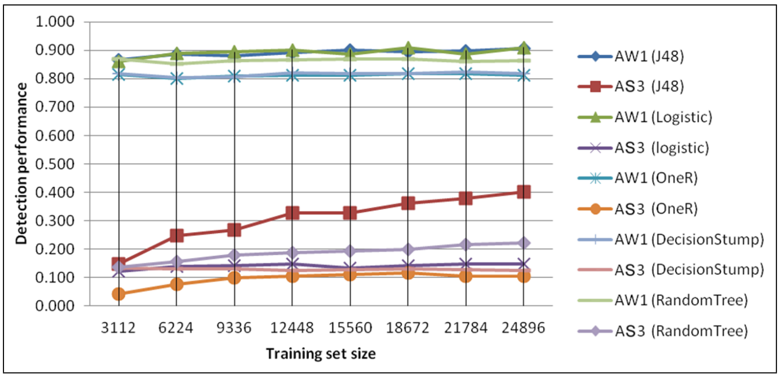


Figure 5.1: Detection performance (κ value) variation with training set size (experimental setup *AS-D-SF-scaling*)

The results for A_{S1} are omitted in this figure because the evaluated classifiers achieved for this algorithm no detection performance other than $\kappa = 0$. As can be seen in figure 5.1, the results for A_{W1} are extremely good for all tested set sizes. A small increase is visible when raising the set size from 3112 to 6224 but further increases do not show any significant effect. For A_{S3} obviously even the largest set size tested is not yet large enough to achieve optimal results.

Résumé for this section The classification performance achieved in SPR is, next to other characteristics, strongly depending on the quality of the model used for classification. The quality of the model itself depends on a number of characteristics. Amongst them is the model size, which is the focus of this section. The model size itself directly reflects the material provided in the training phase. If this training set is chosen too small, then the model is most likely insignificant. If it chosen too large then the curse of dimensionality [Bellman61] will take its toll on the computation time required for training as well as application of the model in classifications, additionally in this case a high probability of an overfitting exists. If a wrong context is trained, due to a wrong choice of training set components, this wrongful content adaptation will also with a high probability render the model useless.

[62] All other influence factors, like the classifier used, the parametrisations and embedding strengths for the IH algorithms, etc are identical.

The investigation results presented in this section indicate that the achieved classification accuracy scales with increasing training set size. But a concrete optimal training set size (and corresponding classifier model size) for all audio steganalysis approaches could not be established in the investigations here, since such a figure is depending on too many variables. For one example specification of the audio steganalysis problem – simulated VoIP steganography and steganalysis using the AAFE (v.1.0.3) with a confidence / trust level of 95% correct classifications – a rough estimate is presented for five chosen IH algorithms at 400 feature vectors per reference file, a figure which would be much too large for the training multi-content or even general-purpose classifier models. Nevertheless, this example specification is very idealistic as shown in the results presented in [Kraetzer08a] and does not allow for any generalisation of this subject. Future investigations in SPR-based steganalysis should extend their considerations on required feature set sizes for context specific training.

Based on the figure given above, the results presented here imply that general-purpose SPR performs suboptimal in steganalysis, where optimised versions of the general SPR tool-set are required (e.g. optimisations by feature selection and training set optimisation) will lead to more practical relevant (confidence and throughput/complexity) steganalysers.

5.1.3 Tests on unmarked sets

Even if it is assumedly more significant if a detector fails to detect a steganographic channel present (i.e. causes false negative errors), it has to be investigated also whether the introduced approach generated an unnecessarily high number of false positive errors. For this reason [Kraetzer07b] presents detection performances from two-class setups achieved in testing on completely unmarked material. The Table 5.7 summarises these results for two different feature sets (the time-domain features computed by AAFE v.1.0.3 vs. a combination of the time-domain features and the FMFCCs computed by AAFE v.1.0.3).

Table 5.7: Detection performance (κ values) for models trained for the corresponding information hiding algorithm (on marked and unmarked versions of *aats389*) and tested on unmarked material; results adapted from [Kraetzer07b]; experimental setup *AS-Kraetzer2007IH-unmarked*

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
only time-domain features	0.101	0.292	0.631	0.791	0.791	0.636	0.802	0.263	0.320
time-domain features and FMFCCs	0.085	0.272	0.506	0.588	0.788	0.979	0.551	0.335	0.231

Based on these detection performances, table 5.8 summarises the corresponding false positive error rates, showing a very inhomogeneous behaviour for the different test cases, which follows the general trend for the results presented within this thesis for the application scenario of audio steganalysis.

Table 5.8: False positive rate (FPR) for the corresponding information hiding algorithm (on marked and unmarked versions of *aats389*) and tested on unmarked material; results adapted from [Kraetzer07b]; experimental setup *AS-Kraetzer2007IH-unmarked*

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
FPR for time-domain features	44.97%	35.40%	18.44%	10.43%	10.43%	18.20%	9.90%	36.86%	34.00%
FPR for time-domain features and FMFCCs	45.77%	36.39%	24.71%	20.62%	10.59%	1.04%	22.46%	33.23%	38.44%

Résumé for this section: The initial results, summarised here for the performance on completely unmarked audio material, show a strong divergence in the false positive error rates achieved. The probability of causing false alarms strongly depends on the individual detector. When steganalysis detectors are introduced with sufficiently high detection performances, future work should incorporate the false positive and false negative error rates into a fair benchmarking scheme for steganalysis detectors, because their proportion is equivalent between the security (false negative rate) and usability (false positive rate) of the approach.

5.1.4 Application scenario specific classifier selection for audio steganalysis

In [Kraetzer08a], as a first classifier comparison for the approach introduced in the context of this thesis, three different **supervised classification algorithms** (*libSVM*, *weka.classifiers.bayes.NaiveBayes* and *weka.classifiers.functions.MLRM*) are tasked with the same classification problem. For each of the nine data hiding algorithms evaluated there a classical two-class classification on marked and unmarked material (ratio 1:1) is performed for the audio test set *aats389* using the 98 dimensional feature set extracted from the audio material by AAFE v.1.0.4 and 256 feature vectors per file for training against 64 feature vectors per file for testing (see experimental setup *AS-Kraetzer2008SPIE-ClassifierComparison* in table 10.1 in appendix B (starting on page 197)). The results of this classifier comparison are presented in table 5.9.

Table 5.9: Detection performance (κ values) for three different classifiers on the same audio steganalysis problem (see experimental setup *AS-Kraetzer2008SPIE-ClassifierComparison*, results adapted from [Kraetzer08a])

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
κ for *libSVM*	0.10	0.31	0.45	0.28	0.27	0.92	0.48	0.28	0.27
κ for *NaiveBayes*	0.05	0.02	0.19	0.04	0.06	0.92	0.04	0.01	0.04
κ for *MLRM*	0.07	0.08	0.24	0.05	0.06	0.96	0.16	0.16	0.08

For *libSVM* the average κ over all nine IH algorithms is 0.37 (0.28 in average for the five steganography algorithms and 0.49 for the watermarking algorithms). The *NaiveBayes* shows an average κ of 0.15 (0.07 on steganography and 0.25 on watermarking). The *MLRM* classifier shows an average κ of 0.21 (0.10 for steganography and 0.34 for watermarking). All three classifiers show significant results ($\kappa > 0$) in the performed two-class classifications of the data hiding algorithms. The best classified algorithm for all three classifiers is A_{W1} with $\kappa > 0.92$ (equivalent to a classification *accuracy* above 95%). Summarising the results for the steganography and watermarking algorithms it can be seen that the watermarking algorithms show for all three classifiers a higher statistical detectability than the steganography algorithms, as expected from the different focus in the algorithm design.

These first classifier performance comparisons presented in [Kraetzer08a] are extended in [Kraetzer10] by an in-depth analysis for three selected information hiding (IH) algorithms. As a point-of-reference, the performance of all 74 supervised classification techniques implemented in WEKA (v.3.6.1) using the global as well as the segmental features in a classical two-class setup is evaluated. Table 5.10 summarises the results for the global features for 10-fold stratified cross-validation on the set *aats389* as well as independent training with *aats389* and testing with *testset24* (see experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*).

Table 5.10: Detailed comparison of the detection performances achieved with the AAFE v.2.0.5 global features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*)

	10-fold strat. cross-validation			**Training and testing**		
	A_{S1}	A_{W1}	A_{S3}	A_{S1}	A_{W1}	A_{S3}
Maximum achieved κ value	0.72	0.85	0.38	0.15	0.50	0.19
Maximum achieved accuracy	86.15%	92.68%	68.94%	57.29%	75.00%	59.38%
Time duration (s)	598.7	603.8	617.3	222	191	180
Performance histogram:						
Errors	13	14	14	13	13	13
$0.00 \leq \kappa < 0.04$ ($50 \leq accuracy < 52\%$)	24	8	11	55	9	16
$0.04 \leq \kappa < 0.20$ ($52 \leq accuracy < 60\%$)	32	1	9	6	2	45
$0.20 \leq \kappa < 0.40$ ($60 \leq accuracy < 70\%$)	2	1	40	0	34	0
$0.40 \leq \kappa < 0.60$ ($70 \leq accuracy < 80\%$)	0	6	0	0	16	0
$0.60 \leq \kappa < 0.80$ ($80 \leq accuracy < 90\%$)	3	24	0	0	0	0
$0.80 \leq \kappa < 1.00$ ($90 \leq accuracy < 100\%$)	0	20	0	0	0	0

The results presented in table 5.10 show the following: the highest achieved maximum κ values in cross-validation and training and testing are seen for A_{W1} ($\kappa = 0.85$ and 0.5 respectively). A_{S3} shows

in average the second best result, while A_{S1} ranks last – confirming our previous findings under similar test setups (e.g. in 5.10). With an average duration a little bit over 600 seconds on our reference test machine, for all tests the cross-validation performs only 3-times slower than the training and test. In all tests the same set of about 13 classifiers (in two cases 14) are not able to successfully generate a decision. The following reasons for this error behaviour are identified: some classifiers terminate with an error stating that they have not enough memory to complete the task, some classifiers are terminated after 12 hours to keep the overall test duration limited (the time for those terminated is not included in the timings presented in table 5.10 and table 5.11), others are cost sensitive classifiers which would not run without a cost file, which could not be modelled for this test.
The lower six rows in table 5.10 basically form histograms of how many classifiers achieved κ values in the corresponding ranges. Again A_{W1} shows the best performance: in cross-validation for 20 classifiers $\kappa > 0.8$ was achieved, with a maximum at $\kappa = 0.85$ (*weka.classifiers.trees.LMT*).
Generally, the results in the training and testing setup are lower than in the cross-validation, this seems to be due to the lower correlation between the test and training data in this case – for a detailed discussion of the different relevancies achieved in this application scenario by cross-validation and two-set training and testing see section 5.1.1. The good results found for A_{S1} in cross-validation (*weka.classifiers.lazy.IB1* $\kappa = 0.70$, *weka.classifiers.lazy.IBk* $\kappa = 0.70$ and *weka.classifiers.lazy.KStar* $\kappa = 0.72$) could not be confirmed in independent training and testing.

Similar to the results for the global features in table 5.10 above, table 5.11 shows the summary of the classifier detection performance for the segmental features for 10-fold stratified cross-validation on *aats389* as well as training with *aats389* and testing with *testset24*.

Table 5.11: Detailed comparison of the detection performances achieved with the AAFE v.2.0.5 segmental features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *Kraetzer2010SPIE-SF-singleClass-summary*)

	10-fold strat. cross-validation			**Training and testing**		
	A_{S1}	A_{W1}	A_{S3}	A_{S1}	A_{W1}	A_{S3}
Maximum achieved κ value	0.88	0.93	0.41	0.00	0.79	0.22
Maximum achieved accuracy	94.19%	96.47%	70.36%	50.12%	89.48%	61.22%
Time duration (s)	104467.5	86056.2	113349.6	30870.0	35852.1	54521.3
Performance histogram:						
Errors	31	29	35	29	27	31
$0.00 \leq \kappa < 0.04$ ($50 \leq accuracy < 52\%$)	35	0	1	45	0	20
$0.04 \leq \kappa < 0.20$ ($52 \leq accuracy < 60\%$)	3	8	26	0	11	17
$0.20 \leq \kappa < 0.40$ ($60 \leq accuracy < 70\%$)	0	3	11	0	1	6
$0.40 \leq \kappa < 0.60$ ($70 \leq accuracy < 80\%$)	1	1	1	0	4	0
$0.60 \leq \kappa < 0.80$ ($80 \leq accuracy < 90\%$)	0	3	0	0	31	0
$0.80 \leq \kappa < 1.00$ ($90 \leq accuracy < 100\%$)	4	30	0	0	0	0

As with the global features above, A_{W1} shows the best detectability with 30 classifiers achieving a $\kappa > 0.8$ in cross-validation. In general, the segmental features seem to outperform the global features in terms of the achieved classification performance, i.e. more higher detection performances are visible in the comparison between the corresponding histograms. Like already shown for the global features, the good results for four classifiers on A_{S1} in cross-validation could not be verified with independent training and testing.

If the results for the global and segmental features are compared directly, besides the slight overall increase in the classification accuracy, two additional facts are noticeable: the number of classification algorithms which cannot fulfil the classification task (row 'error' in the tables) more than doubles and the time duration for the tests increase by at least two magnitudes. The former is due to the increased memory requirement for the segmental features (by factor 20 in comparison to the global features) and a corresponding increase of the number of 'out of memory errors' from WEKA, while the latter is due to the much more complex training and testing tasks at hand and thereby more timeouts at the defined 12 hour timeout.

If **clustering**, as the second general approach to classification, is considered for audio steganalysis, the results can be summarised as: even though clustering is in some cases able to achieve significant detection performance in classes-to-clusters evaluations (here, as highest value $\kappa = 0.45$ is achieved by WEKAs *SimpleK-means* on A_{W1}) the detection performance achieved in clustering is by far outperformed by the accuracies achieved in supervised classifications. Due to this fact a detailed discussion of the clustering results is omitted here.

Résumé for this section: The application scenario specific classifier comparison for audio steganalysis shows that the choice of the classifier has indeed a very strong influence on the outcome of the detection. In [Kraetzer08a] it is shown that, under certain circumstances, WEKA classifiers are able to outperform the SVM classifier *libSVM* (which is in many publications on steganography considered to be the expert classification engine for this two-class classification problem) in terms of practically achieved detection performances. This point highlights why it is so important to perform application scenario specific performance comparisons. The initial work presented within this section is extended in section 8.2.1 by the presentation of details on ongoing work on application scenario based classifier benchmarking for audio forensics applications.

5.1.5 Feature selection for audio steganalysis

In [Kraetzer08b] a first feature selection is performed on the steganalysis approach considered here. This first feature selection is implemented by single feature classification – a naive approach which does not take into account that feature combinations might be more powerful than individual features. The only fact established by these evaluations is that different features are relevant for each of the tested IH algorithms.

Here, the feature selection concept and design introduced in section 4.1.2 is applied to the three IH algorithms in experimental setups *AS-Feature-Selection-GF* and *AS-Feature-Selection-SF* (see table 10.1 in appendix B (starting on page 197)).

Global features

None of the global features computed by AAFE v.2.0.5 shows any significance for A_{S1}, therefore the results for this algorithm are omitted in the feature ranking results presented in table 5.12.

Table 5.12: Ranking of the 10 best global features for A_{W1} and A_{S3}, based on the fused rankings computed by the five selected feature selectors (see experimental setup *AS-Feature-Selection-GF*)

	A_{W1}		A_{S3}	
Final rank	Feature	Average ranking	Feature	Average ranking
1	gf_{zcr_total}	1	$gf_{LSBflip\_AVE}$	1
2	$gf_{zero\_cross\_rate\_AVE}$	2	gf_{LSBrat_AVE}	2
3	$gf_{sp\_bw\_AVE}$	3.8	gf_{zcr_total}	3
4	$gf_{entropy\_AVE}$	4.4	$gf_{zero\_cross\_rate\_AVE}$	4.4
5	$gf_{sp\_centriod\_AVE}$	5.2	$gf_{sp\_bw\_AVE}$	5
6	$gf_{sp\_irregularity\_AVE}$	5.6	$gf_{entropy\_AVE}$	6
7	$gf_{sp\_entropy\_AVE}$	6.6	$gf_{sp\_centriod\_AVE}$	8.6
8	$gf_{sp\_rolloff\_AVE}$	7.4	$gf_{RMS\_amplitude\_AVE}$	9
9	gf_{energy_AVE}	9.6	gf_{energy_AVE}	9.6
10	$gf_{RMS\_amplitude\_AVE}$	10.6	$gf_{sp\_rolloff\_AVE}$	10.2

The best performing features for A_{W1} (the zero-crossing features the average spectral bandwidth and the average entropy) imply strong noise behaviour of the embedding operation of this algorithm, while A_{S3} causes a strong impact to the least significant bits of the audio samples (LSB flipping rate and LSB ratio).

Segmental features

The detection performance on A_{S1} is rather low when using the segmental features. Nevertheless, since κ values larger than 0 are achieved in section 5.1.4, the feature selection results returned for this algorithm are included in table 5.13. If the average ranking is considered for the individual features, it can be seen that it is extremely high in comparison to the other two IH algorithms. This is resulting from the fact that the five individual feature selectors used here in this fused ranking show strongly differing rankings.

Table 5.13: Best 30 segmental features for A_{S1}, A_{S3} and A_{W1}, based on the fused rankings computed by the five selected feature selectors (see experimental setup AS-Feature-Selection-SF)

	A_{S1}		A_{W1}		A_{S3}	
Final rank	Feature	Average ranking	Feature	Average ranking	Feature	Average ranking
1	sf_{spec_146}	69	sf_{spec_369}	2.2	$sf_{entropy}$	14.24
2	sf_{spec_168}	74.8	sf_{spec_363}	3	sf_{LSBrat}	15.56
3	sf_{spec_143}	76.8	sf_{spec_360}	4.6	$sf_{zero\_cross\_rate}$	16.28
4	sf_{spec_151}	77.8	sf_{spec_368}	4.8	sf_{d2MFCC_1}	16.52
5	sf_{spec_149}	78.2	sf_{spec_365}	5.2	sf_{median}	16.68
6	sf_{spec_147}	78.4	sf_{spec_370}	7	$sf_{sp\_entropy}$	17.08
7	sf_{spec_148}	81.8	sf_{spec_350}	10	sf_{MFCC_1}	18.36
8	sf_{spec_145}	84.6	sf_{spec_362}	10	sf_{spec_122}	18.92
9	sf_{spec_163}	86	sf_{spec_371}	12.2	sf_{spec_124}	19.64
10	sf_{spec_144}	86.4	sf_{spec_358}	12.8	sf_{sp_bw}	19.84
11	sf_{spec_160}	86.8	sf_{spec_355}	13.4	sf_{spec_107}	20.04
12	sf_{spec_134}	87.2	sf_{spec_356}	14.2	sf_{spec_95}	20.28
13	sf_{spec_164}	89.4	sf_{spec_361}	15.4	sf_{spec_109}	20.96
14	sf_{spec_150}	92	sf_{spec_364}	15.6	sf_{spec_140}	21.52
15	sf_{spec_167}	92	sf_{spec_373}	16	sf_{spec_123}	21.6
16	sf_{spec_142}	92.6	sf_{spec_357}	16.4	sf_{spec_117}	21.8
17	sf_{spec_155}	92.6	sf_{spec_366}	19	sf_{energy}	22.32
18	sf_{spec_161}	93.4	sf_{spec_344}	19.2	$sf_{RMS\_amplitude}$	22.52
19	sf_{spec_137}	93.6	sf_{spec_341}	19.6	sf_{FMFCC_1}	22.64
20	sf_{spec_135}	94.4	sf_{spec_343}	19.6	sf_{spec_1}	22.8
21	sf_{spec_138}	94.6	sf_{spec_372}	20.4	sf_{spec_82}	23
22	sf_{spec_169}	96.6	sf_{spec_359}	21.2	sf_{spec_98}	23.56
23	sf_{spec_136}	97.8	sf_{spec_348}	21.4	sf_{spec_97}	23.84
24	sf_{spec_165}	99.2	sf_{spec_367}	23	sf_{spec_6}	24.44
25	sf_{spec_158}	99.6	sf_{spec_352}	25	sf_{spec_8}	24.92
26	sf_{spec_162}	100.2	sf_{spec_353}	27.6	$sf_{sp\_rolloff}$	25.56
27	sf_{spec_141}	100.4	sf_{spec_354}	29	sf_{spec_2}	25.88
28	sf_{spec_139}	103	sf_{spec_335}	30	sf_{spec_3}	26.6
29	sf_{spec_224}	103.8	sf_{spec_339}	30	sf_{spec_4}	26.76
30	sf_{spec_140}	104.2	sf_{spec_347}	30	sf_{spec_5}	26.84

For A_{W1} the 30 most significant features are, similar to A_{S1}, all frequency-domain features derived from the energy in certain frequency bands (14.4 kHz to 16.1 kHz). This is plausible, because it is well within the frequency range used by the algorithm for embedding and rather unlikely to be interfering with many of the contents used in the evaluations (like speech signals). The plausibility is fostered by the strong agreement of the five involved feature selectors, which expresses itself in the small values returned for the average ranking.
While A_{S1} and A_{W1} show only frequency-domain features in the top 30 list, for A_{S3} a mix of time-, frequency- and Mel-cepstral-domain features can be observed.

In table 5.14 the 25 worst performing features for each algorithm are identified. For A_{S1} and A_{W1}, most of these are frequency-domain features, while A_{S3} shows here mainly Mel-cepstral-domain features. One further interesting fact is the strong agreement between the five individual feature selectors in case of A_{W1} (see the column 'Average ranking' for A_{W1}, where the vales are close to the range 566 to 590 of the final rank). In the case of A_{S1} the agreement is weaker but still significant, while for A_{S3} a much stronger agreement is shown, implying that the empirical approach to feature selection used within this thesis is not performing optimally and should be accompanied in future work by methods from analytical statistics or inferential statistics (i.e. analysis of variance).

Table 5.14: Worst 25 segmental features for A_{S1}, A_{S3} and A_{W1}, based on the fused rankings computed by the five selected feature selectors (see experimental setup AS-Feature-Selection-SF)

	A_{S1}		A_{W1}		A_{S3}	
Final rank	Feature	Average ranking	Feature	Average ranking	Feature	Average ranking
566	sf_{spec_354}	489.4	sf_{spec_436}	541.6	sf_{d2MFCC_4}	206.2
567	sf_{spec_345}	489.8	$sf_{formant\_I1}$	542.8	sf_{spec_491}	206.52
568	sf_{spec_359}	490.6	$sf_{formant\_U1}$	543	sf_{d2MFCC_5}	206.72
569	sf_{spec_349}	492.4	$sf_{formant\_U2}$	544.4	sf_{d2MFCC_2}	207.36
570	sf_{spec_348}	494.6	$sf_{formant\_A1}$	546.2	sf_{MFCC_4}	208.16
571	sf_{spec_335}	496	$sf_{formant\_O2}$	547.2	$sf_{sp\_base\_freq}$	208.64
572	sf_{spec_318}	496.4	sf_{spec_441}	547.6	sf_{d2MFCC_8}	208.8
573	sf_{spec_334}	497.2	sf_{spec_454}	547.6	sf_{MFCC_6}	208.96
574	sf_{spec_325}	500.8	sf_{spec_482}	547.8	sf_{d2MFCC_6}	209.72
575	sf_{spec_336}	502	sf_{spec_453}	548.6	sf_{d2MFCC_7}	210.12
576	sf_{spec_388}	504.6	sf_{spec_511}	548.6	sf_{MFCC_3}	210.64
577	sf_{spec_390}	505	sf_{spec_408}	550.2	sf_{d2MFCC_12}	211.04
578	sf_{spec_322}	506.2	sf_{spec_455}	550.4	sf_{d2MFCC_9}	211.32
579	sf_{spec_380}	507.6	$sf_{sp\_base\_freq}$	553.6	sf_{MFCC_5}	212.24
580	sf_{spec_327}	508	sf_{spec_23}	555.2	sf_{d2MFCC_11}	212.64
581	sf_{spec_326}	510	$sf_{formant\_A2}$	556.8	sf_{MFCC_13}	213.36
582	sf_{spec_343}	512	sf_{spec_443}	557.2	sf_{d2MFCC_10}	213.84
583	sf_{spec_323}	515.6	sf_{spec_461}	557.4	sf_{MFCC_8}	214.6
584	sf_{spec_324}	517.4	sf_{spec_37}	558.2	$sf_{sp\_irregularity}$	215.12
585	sf_{spec_344}	518.8	sf_{spec_25}	559.2	sf_{d2MFCC_13}	215.6
586	sf_{spec_347}	519.2	sf_{MFCC_12}	559.4	sf_{MFCC_11}	215.72
587	sf_{spec_338}	520.6	$sf_{formant\_E2}$	570.2	sf_{MFCC_10}	218.16
588	sf_{spec_339}	521	sf_{sp_bw}	574.8	sf_{MFCC_7}	219.16
589	sf_{spec_351}	521.6	$sf_{formant\_Singer}$	576.2	sf_{MFCC_9}	219.56
590	sf_{spec_341}	524.6	$sf_{formant\_I2}$	578.8	sf_{MFCC_12}	220.68

PCA-based estimation of the number of uncorrelated features for each classification problem

In experimental setup AS-Feature-Selection-SF/GF-PCA the principal component analysis (PCA) based estimation of the number of uncorrelated features is performed for global and segmental features as described in section 4.1.2 to evaluate the feature (in-)dependency for the performed audio steganalysis. After the PCA, for all three IH algorithms the transformed feature space for the global features is reduced from the original 17 to 11 dimensions. If the same PCA is performed for the segmental features, the results vary slightly for the three different IH algorithms considered A_{S3} shows with 159 dimensions in the transformed feature space the smallest correlation between the original features, while the results for A_{S1} (144) and A_{W1} (148) are close together. In summary it can be said that the feature space of the segmental features shows for all three IH algorithms a strong correlation. The PCA is capable to reduce the dimensionality of the feature space to 25% (which would significantly reduce the time for the classification) while at the same time keeping its expressive power at 95% of the expressive power of the original. Nevertheless, it should be remembered here that the expressive power (i.e. the detection performance) strongly varies between the three evaluated IH algorithms (see section 5.1.4).

Feature selection by feature ranking – the domain knowledge generated

For A_{S1} not much can be derived in terms of domain knowledge from the results presented. As shown in section 5.1.4 the detection performance achieved is extremely low, this leads to the realisation that the currently used features are not suited to allow for the detection of this algorithms on multi-genre audio signals. Here, future research should be invested into analyses on specific genres of audio signals (e.g. speech only content – where more promising detection results have been achieved here). This might lead also to the design of new features that can be used to detect this algorithm.

The results for A_{W1} show a strong effect of the embedding in mid-to-high frequencies, which is consistent with the upper and lower frequency limits set for the embedding algorithm. It is also consistent

with the embedding strategy that the formant features as well as very high frequencies are not affected by the embedding.
For A_{S3} a wide range of features seems to be relevant for the detection. The performed ranking indicates good results for time-domain features, frequency-domain features as well as the fist coefficients in MFCCs, FMFCCs and the second-order derivative MFCCs.

Résumé for this section: For the performed audio steganalysis evaluations, global as well as segmental features show relevance. For all three IH algorithms evaluated different features show relevance, i.e. a set of 'best features' cannot be named for this application scenario.
The varying levels of agreement between the fused feature selectors reduce the confidence in this feature selection strategy applied here, which in turn implies a need for future research to look into analytical statistics and significance analysis for feature selection.
Due to a strong correlation in the feature space, the performed PCA is capable of reducing the dimensionality of the feature space to 25% while maintaining 95% of the original detection performance. Even though the resulting decreased time required for the classifications after the PCA is of limited interest for research, it will be a huge influence factor once audio steganalysis would be implemented into security mechanisms.

5.2 Steganalysis specific influences to the SPR process

An important task for the audio steganalysis approach introduced within this thesis is to determine the influences of the steganalysis setup to the achievable detection performance. Following the design for the evaluations introduced in sections 4.2.1 and 4.2.2, here the influences of the algorithm and embedding domain, the key scenario, the context dependency between training and testing as well as the dimensionality of the classification setup (two-class or multi-class) are considered.

5.2.1 Embedding domain and algorithm identification

In [Kraetzer07b] a model-cross evaluation on model significances is performed to investigate whether similarities between the embedding algorithms can be seen by training a model for the detection of one algorithm and testing it against a different algorithm. The results of these evaluations (experimental setup *AS-Kraetzer2007IH-CrossEval*) are shown in adapted versions in table 5.15.

Table 5.15: Results (κ value) for the cross-algorithm evaluation using the feature set $SF_{std\&FMFCC}$ y-Axis: training material – x-Axis: test material; experimental setup *AS-Kraetzer2007IH-CrossEval* – The embedding domains are: T=time, F=frequency and W=wavelet – see section 4.2.1)

↓ **Training**	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
Embedding domain	**T**	**T**	**W**	**T**	**T**	**F**	**W**	**T**	**W**
A_{S1}	0.244	0.038	0.219	0.137	0.149	0.056	0.235	0.122	0.060
A_{S2}	0.052	0.244	0.240	0.189	0.150	0.335	0.155	0.208	0.033
A_{S3}	0.080	0.117	0.459	0.179	0.182	-0.107	0.293	0.168	0.142
A_{S4}	0.084	0.130	0.263	0.244	0.225	-0.016	0.229	0.222	0.016
A_{S5}	0.072	0.065	0.252	0.211	0.241	0.018	0.265	0.200	0.029
A_{W1}	0.003	0.003	0.004	0.000	0.025	0.982	0.005	0.000	0.007
A_{W2}	0.070	0.063	0.311	0.153	0.168	-0.142	0.497	0.130	0.097
A_{W3}	0.110	0.185	0.262	0.216	0.212	0.013	0.220	0.252	0.028
A_{W4}	0.058	0.060	0.315	0.078	0.092	-0.126	0.251	0.078	0.278

The results in imply that table 5.15 models not only significant for one algorithm but for 'similarity classes' – i.e. cases where Kappa values larger than 0 are achieved off the main diagonal. The arising problem is that some of these implied similarities are not intuitive. An example for an intuitive similarity is the case of A_{S4} and A_{S5} which are different versions of Steghide with only slightly modified embedding strategies. One example for an unexpected similarity is the model for the time-domain algorithm A_{W3}, which achieves also an extremely high detection performance for the wavelet-domain algorithm A_{S3}.
An example, where a model created for one algorithm works quite well for all other algorithms in the

same embedding domain is the model for A_{W4}. Unfortunately, this is not true for most of the cases – an embedding domain detection seems to be infeasible with this evaluation setup.

In general one further insight can be derived from this investigation: Most of the models trained show a very low discriminatory power if it comes to the task of distinguishing between different embedding algorithms. The only exception is the model trained for A_{W1}, for which the second highest result in the corresponding row in table 5.15 has a Kappa value of 0.025. For all other cases there exist test cases off the main diagonal, which return $\kappa > 0.2$. The implication is, that a model based algorithm identification with the current approach seems to be possible only in very specific cases.

When considering single feature-based algorithm identification instead of a large set model-based approach, the results presented in [Kraetzer08b] imply that it might be possible to identify the embedding domain or even the embedding technique used for the creation of a set of stego objects. In that paper, it is stated that frequency domain features are strongly affected by the evaluated frequency-domain embedding strategies and could be used to distinguish between time-domain embedding on one hand and frequency-domain as well as wavelet-domain embedding on the other hand.

Résumé for this section: The similarities presented for the detection of stego objects embedded by one algorithm using models generated for a different algorithms provide a strong argument against the usual two-class modelling of steganalysis. The ideal forensic audio steganalysis process (as discussed in section 3.1.1) requires, amongst others, a reliable identification of the used embedding domain and algorithm. This cannot be achieved with the methods evaluated here. Even though the embedding domain or even the algorithm used might be correctly identified possible in very specific cases, future research work has to be invested into feature-based algorithm identification, to close in on the requirements for the ideal forensic audio steganalysis process.

5.2.2 Key scenario in steganography – influence to steganalysis

In [Kraetzer09a] a set of three data hiding algorithms (A_{S1}, A_{S3} and A_{W1} – for algorithm descriptions see section 4.2.1) is used to generate training and test data for the estimation of the key scenario on the detection performance achieved (see experimental setup *AS-KraetzerSPIE2009-KeyScen*). Each of the three algorithms is working in a different domain: A_{S1} is a time-domain LSB algorithm, A_{S3} a wavelet-domain algorithm and A_{W1} a frequency-domain spread spectrum technique. For the embedding two different key selection strategies are compared. The first ('fixed key') uses exactly one predefined key (*UniversityOfMagdeburg*) for the generation of the stego objects used as training and test files – i.e. in all files the message is embedded using the same key. The second key selection strategy ('variable key') uses the MD5-hash value of the filename for each file in a test set as the key for embedding – therefore it uses for each file in the test set a unique key. All files used in the evaluations are marked by the data hiding algorithms with 100% embedding strength, the message to be embedded is an ASCII version of Goethes' 'Faust'.

The results achieved in this experiment show a rather small impact of the key selection strategies tested on the classification results. Only in 5 out of the 30 direct comparisons between fixed and variable key the difference in the achieved detection *accuracy* is larger than 2% (for details see [Kraetzer09a]). Table 5.16 compares the best detection performances achieved for both key scenarios.

Table 5.16: Best detection performance achieved for two different key scenarios (training set: *aats389_Part1* test set: *testset24*; adapted from [Kraetzer09a] – see experimental setup *AS-KraetzerSPIE2009-KeyScen*)

	fixed key		variable key	
Algorithm	Detection performance (κ)	Feature extractor & classifier	Detection performance (κ)	Feature extractor & classifier
A_{S1}	0.042	AudioRS & *SimpleLogistics*	0.084	AudioRS & *NaiveBayes*
A_{S3}	0.500	AAFE & *ADABoost*	0.542	AAFE & *ADABoost*
A_{W1}	0.792	AAFE & *ADABoost*	0.792	AAFE & *ADABoost*

Résumé for this section: The results presented imply that the choice of the embedding key has only a very limited impact on the detection performance achieved in steganalysis. Nevertheless, if future work achieves the implementation of reliable detectors for audio steganalysis, the impact of the key selection should be re-evaluated together with other embedding parameters (e.g. the embedding strength and embedding strategy).

5.2.3 Classification using content selection as well as content dependent and independent training and testing

In [Kraetzer07a] significantly better results for audio steganalysis are achieved for AAFE v.1.0.3 on the speech only test-set *longfile* in comparison to the multi-genre test-set *aats389* (experimental setups *AS-Kraetzer2007SPIE-longfile* versus *AS-Kraetzer2007SPIE-summary*). The difference in the detection performance on those two sets is most significant for A_{S1}, with $\kappa = 1$ (for a ratio of 2200 to 400 feature vectors per file for training and testing and the time-domain features and the FMFCCs in AAFE v.1.0.3 on the test-set *longfile*) to $\kappa = 0.142$ (ratio of 64 to 16 feature vectors per file and only the time-domain features on test-set *aats389*). Unfortunately, this investigation mixes two different influencing factors: the scaling of the classification accuracy with increasing model sizes (see section 5.1.2) and the content dependency of the steganalysis approach. As a result of this realisation, these results are the motivation to perform more generalisable investigations on the content dependency.

In [Kraetzer08a] we consider two different setups for the statistical pattern recognition (SPR) based audio steganalysis approach: a setup where we know the semantical characteristics of the channel under observation (cover dependent training and testing) and a setup where those characteristics are unknown (cover independent training and testing). The first setup could be created by introducing with content classification an additional, semantical pre-processing to the steganalysis process pipeline. An alternative description of these two setups can be give using the degree of contextual correlation[63] between the training and test sets. For the first setup, the contextual correlation is very high, while in the second case with a very high probability a low correlation would be achieved. The relevance of cover dependent training and testing is illustrated in [Kraetzer08a] on the example of VoIP steganalysis, where a channel with known semantical characteristics (i.e. human speech in VoIP-enabled Internet telephony) is considered.

Channel specific cover dependent training on the example of VoIP steganalysis

Table 5.17 shows the results for a *libSVM*-based classification and cover dependent training and testing (using as feature set all 98 segmental features computed by AAFE v.1.0.3) for a speech-only VoIP-like setup and a multi-genre audio test set. For the speech-only evaluations, the audio test set *ahss1* containing only human speech, which is assumed to be the usual content in VoIP communications, is used for training and large models are generated keeping only 1200 samples per file of this test set for testing. As a result the detection performance for all nine algorithms is in the range $[0.102, 1]$, with $\kappa > 0.8$ for six algorithms.

Table 5.17: Detection performance (κ values) for cover dependent training and testing for *libSVM* (experimental setup *AS-Kraetzer2008SPIE-ContentDependent*) – results adapted from [Kraetzer08a]

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
speech	0.896	0.550	0.996	0.102	0.206	1.000	0.802	0.938	1.000
multi-genre audio	0.104	0.306	0.454	0.276	0.274	0.922	0.482	0.280	0.266

When comparing the results presented in table 5.17 for speech-only and multi-genre material, significant differences in the detection performances achieved can be seen. The models generated on the speech

[63] No formalisation on the degree of correlation between sets of audio material is performed in this thesis. Instead it is assumed that, e.g. two sets of human speech signals show a higher degree of correlation than while speech signals and white noise, or speech and violin music.

set seem to be much more effective in audio steganalysis on speech material than their multi-genre counterparts on multi-genre audio.

Cover independent training and testing – an extreme case

To show which impact a wrong assumption on the channel characteristics may have on the detection performance, table 5.18 compares cover independent (row 1) and cover dependent tests (row 2). As described in experimental setup *AS-Kraetzer2008SPIE-ContentInDependent*, the model for the **cover independent tests** is generated for each algorithm using *ahss1*. Therefore, this model is trained only on marked and unmarked speech signals. The model for the **cover dependent tests** is generated for each algorithm using *ref10* by applying on the marked and unmarked versions of *ref10* a split 80%:20% and using the 80% for training of the model and the remaining 20% as test material against both generated models are tested.

Table 5.18: Detection performance (κ values) for cover independent (row 1) and cover dependent (row 2) *libSVM* classification using all 98 features from AAFE v.1.0.4; models generated on *ahss1* and *ref10* test material generated from *ref10* adapted from [Kraetzer08a]

	A_{S1}	A_{S2}	A_{S3}	A_{S4}	A_{S5}	A_{W1}	A_{W2}	A_{W3}	A_{W4}
Model generated on *ahss1*	0.034	0.032	−0.018	0.030	0.084	0.126	0.018	−0.034	0.018
Model generated on *ref10*	0.334	0.416	0.832	0.440	0.512	0.890	0.636	0.512	0.748

A comparison of the results in table 5.18 shows that the cover dependent training and testing (average detection performance over all nine algorithms: $\kappa = 0.592$) performs for all nine algorithms better than the cover independent training and testing (average $\kappa = 0.032$). The differences in detection performances between both setups are rather large, highlighting the significance of correct training material selection.

Résumé for this section: The results presented in this section highlight two different facts: the impact of the channel specific characteristics to the classification and the need to train models adapted to the application context. Selected audio contents (speech) seem to allow for better detection performances for the introduced statistical pattern recognition (SPR) based audio steganalysis approach.

Furthermore, the results imply that a model generated for one type of audio content performs significantly worse if applied for the classification on different content. As a consequence, the channel specifics characteristics (in the example evaluations presented here: speech in VoIP steganography) should be reflected in the training. Future work in this field should consider the integration of content analysis as a pre-processing operation into SPR-based steganalysis approaches to enable the shift from cover independent to cover type dependent training and testing.

5.2.4 Two-class vs. multi-class setups

In experimental setup *AS-D-SF-multiClass*, instead of a classical two-class setup a multi-class setup as motivated by [Provos02] is implemented. This is ignoring the general trend in the state-of-the-art in this field which mostly models the steganalysis problem as a strictly two-class detection problem. The results of this experiment, which uses five exemplary classifiers selected from WEKAs portfolio on basis of their performance in classifier selection (see section 5.1.4), are summarised in table 5.19.

Table 5.19: Multi-class (unmodified cover, A_{S1}, A_{S3} and A_{W1}) steganalysis results for global end segmental features (AAFE v.2.0.5) – based on experimental setup *AS-D-SF-multiClass*

	Global features	**Segmental features**
Classifier	κ value	κ value
trees.J48	0.313	0.395
functions.Logistic	0.362	0.355
rules.OneR	0.077	0.262
trees.DecisionStump	0.196	0.300
trees.RandomTree	0.197	0.258

The results presented in table 5.19 show that the for all five exemplary selected classifiers detection performances of $\kappa > 0$ can be achieved. The results are not as good as for the two-class setup for

A_{W1}, but with maxima of $\kappa = 0.362$ for the global features and $\kappa = 0.395$ for the segmental features they are still promising.

Table 5.20 and table 5.21 show the confusion matrices for the global and segmental features in experimental setup *AS-D-SF-multiClass* classified by the *trees.J48* implementation of WEKA. As can be seen, the results follow the same general trend as shown in the classifier selection performed in section 5.1.4: A_{W1} contributes the best detection performance to the evaluation, while A_{S1} causes an extremely high number of misclassifications.

Table 5.20: Confusion matrix for experimental setup *AS-D-SF-multiClass*, global features classified with *trees.J48*

↓ **Training**	A_{S1}	A_{W1}	A_{S3}	**original**
A_{S1}	105	11	42	97
A_{W1}	24	216	13	6
A_{S3}	89	12	141	23
original	168	11	30	59

Table 5.21: Confusion matrix for experimental setup *AS-D-SF-multiClass*, segmental features classified with *trees.J48*

↓ **Training**	A_{S1}	A_{W1}	A_{S3}	**original**
A_{S1}	881	79	412	3071
A_{W1}	139	4900	132	113
A_{S3}	768	132	2710	1420
original	2111	78	400	2026

Résumé for this section: As already discussed by [Provos02], steganalysis in practice would be a multi-class detection problem (one class for each possible embedding method) trying to answer a two class decision problem (steganographic communication present or not). Even though the evaluations on multi-class realisations of audio steganalysis performed within this thesis are of very limited nature, they show that such a realisation might be possible in practice. Future work should be invested into the question of multi-class steganalysis (i.e. algorithm identification), extending the dimensionality of the decision problem and comparing it to networks of more typical two-class algorithm specific steganography detectors trying to solve the same problem. The corresponding evaluations would have to compare both approaches with regard to their detection performance as well as their scaling behaviour.

5.3 Persistence of the patterns against selected post-processing operations

The plausibility of audio steganalysis has to verify the influence of (malicious or non-malicious) audio signal processing operations on the classification behaviour. The motivation for this consideration in found in the fact that especially pieces of music undergo rather dramatic modifications between their recording and the roll-out on a CD. One example for such modification is the custom to 'improve' singers voices with artificial reverberation. Table 5.22 summarises the results of an experiment from [Kraetzer10], where we train classifiers for three different data hiding algorithms (A_{S1}, A_{S3} and A_{W1}) and then apply for each these algorithms the '5 best' classifiers in the set onto the segmental and global features extracted from a set of completely unmarked audio material that underwent signal modifications (MP3 conversion and de-noising). An identification of those '5 best' classifiers is given in section 8.2.1. For a more detailed description of the performed evaluation we refer to [Kraetzer10].

A value of $\kappa = 1$ in table 5.22 indicates that the complete test material was rightfully classified as unmarked by the corresponding feature extractor and classifier combination. A value of $\kappa = 0$ (equivalent in this two-class setup to an *accuracy* of 50%) implies that the classifier achieves a similar detection performance as a simple guessing at the correct class would return. A value of $\kappa = -1$ means that the classifier produced false alarms on every input sample. Summarising the evaluation results, it can be stated the de-noising operation output is in nine out of 15 test cases with the global features found 100% ($\kappa = 1$) correct to be 'not marked', in four other cases κ value is above 0.8, while for nine cases negative κ values indicate false alarms rates higher than in the case of simply guessing whether the file is an unmodified cover or not.

Table 5.22: Detection performance (κ values) for the global- and segmental features and the best 5 classifiers from the classifier comparison in [Kraetzer10] for each algorithm (see experimental setup *AS-Kraetzer2010SPIE-SF/GF-singleClass*)

Modification	**Classifier**	A_{S1}		A_{W1}		A_{S3}	
		global features	**segmental features**	**global features**	**segmental features**	**global features**	**segmental features**
MP3 encoding	best	0.136	0.073	0.545	0.023	0.273	-0.664
	2nd	0.136	-0.818	0.409	0.539	0.909	0.423
	3rd	1.000	-0.070	0.273	0.130	-1.000	-1.000
	4th	0.909	-0.861	0.545	0.523	0.136	-1.000
	5th	1.000	-1.000	-0.591	0.511	0.455	-1.000
de-noising	best	-0.409	0.079	1.000	0.998	0.591	0.923
	2nd	-0.409	-0.790	0.909	0.675	1.000	-0.133
	3rd	1.000	0.000	0.682	0.3014	1.000	-1.000
	4th	1.000	-0.965	0.818	0.421	1.000	-1.000
	5th	1.000	-1.000	1.000	0.551	1.000	-1.000

For the MP3 encoding the picture is worse, with only two classifiers achieving $\kappa = 1$ and two further classifiers performing at $\kappa > 0.8$. For this modification, 10 classifiers return negative κ values.
It has to be stated that the segmental features seem to perform significantly worse in these tests if it comes to plausibility against common signal modification operations. None of the 30 segmental test cases summarised in table 5.22 reaches $\kappa = 1$, while eight cases show a false alarm rate of 100% ($\kappa = -1$).

Résumé for this section: As implied by the test results presented in this section, SPR-based audio steganalysis seems to be negatively influenced by other audio signal processing operations. Therefore, if its application as a specialised integrity verification mechanism is considered, the implementation of the mechanism should undergo extensive plausibility evaluations against other audio signal modifications (encoding, re-sampling, etc.) that are likely in the considered application field.

In this investigation, two different types of features are compared: global and segmental audio features. In section 5.1.4 the segmental features show a higher detection performance (paid for by higher computational complexities in feature extraction and classification). Here, the global features seem to be less severely influenced by the signal modifications. Nevertheless, while global features can only give a class assignment (i.e. an indication whether an audio signal is a stego object or not) for a complete file, segmental features might be used to identify which part of the file was modified and which was kept unchanged. These facts imply that a combination (i.e. by decision-level fusion) of both might be beneficial to the overall audio steganalysis problem.

5.4 Summary of the findings for audio steganalysis

In section 3.3 the tasks for the practical investigations performed within this thesis are defined. In this summarising section, the results for the audio steganalysis application scenario are first projected onto these investigation tasks. In the second step performed here, the results achieved are reflected under consideration of the evaluation criteria for forensic investigations derived within this thesis from the Daubert standard (see section 2.2 and its subsections).

5.4.1 Projection of the results onto the defined investigation tasks

The first step required in the investigations is to establish some empirical ground truth (**investigation task A**, as an precising statement for research objective 1 – see section 3.3) to show that the application scenario of audio steganalysis (as it is considered within this thesis) can actually be solved by statistical pattern recognition (SPR).

The fact that audio steganalysis can be classified with this SPR-based approach and with a detection performance much better than the probability of guessing correctly was first demonstrated by us in [Kraetzer07a] for a set of nine IH algorithms. This result is verified within this thesis in section 5.1.1. Results achieved vary for the different IH algorithms under investigation between a detection performance of $\kappa = 0$ (for the algorithm A_{S1}) and $\kappa > 0.9$ (A_{W1}), strongly depending on the evaluated algorithm, the used features, the evaluation strategy and the used classifier. Using the mapping between Kappa values and statistical confidence introduced in section 4.1.4, these results range from a 'poor' to a 'fair to good' statistical confidence.
This investigation task is supposed to contain also an answer on what 'sufficient' means in terms of required training and testing (application / evaluation) set sizes. In regard to this question, only a concept for addressing this problem (see section 5.1.2) as well as some first estimations on required set sizes can be given. For the evaluations performed, only for a very narrow application scenario – the observation on speech channels – it was possible to estimate what sufficient 'sufficient' means. Here, a rough estimate is presented for five chosen IH algorithms at 400 feature vectors per reference file, a figure which would be much too large for the training multi-content or even general-purpose classifier models.
Regarding the tendency for overfitting, section 5.2.3 strongly implies that content dependency in the training and testing has a huge influence on the achieved classification accuracies. Like all other empirical results presented here for audio steganalysis, these facts would have to be verified in future work with a larger number of audio steganography algorithms to extend the degree to which these results can be generalised.

Based on the results of the performed audio steganalysis investigations, the summarising statement for investigation task A for this application scenario is: *The results presented here imply that statistical pattern recognition (SPR) based audio steganalysis is possible, if suitable features can be found that are affected by the embedding process of the audio steganalysis algorithm under investigation.*
The second part of this statement is motivated on the fact that for one of the algorithms in the test set it (A_{S1}) it is not possible to achieve a successful detection for multi-genre content, while this algorithm was detectable when embedding in specific content (here speech). Obviously, the algorithms to be detected have to registered (i.e. models have to be trained for them), otherwise the detection approach used here would not work sufficiently. As shown in section 5.2.1, this training might be performed for an embedding domain or strategy, rather than the steganographic algorithm using this embedding strategy. The progress made within this thesis on that regard has to be substantiated in future work.
Another important point made on investigation task A is the fact that in section 5.2.4 it is shown that a discussion of two-class vs. multi-class setups for steganalysis is indeed necessary. In practice, steganalysis would be a multi-class detection problem (one class for each possible embedding method) trying to answer a two class decision problem (steganographic communication present or not). Even though the evaluations on multi-class realisations of audio steganalysis performed within this thesis are of very limited nature, they show that such a realisation might be possible or even required in practice. This is contradicting the current trend in the state-of-the-art to trivialise steganalysis as a two-class detection problem.
With this statement and its counterpart for microphone forensics in section 6.5.1, research objective 1 (resp. research challenge (a)) is answered positively.
The investigations performed within thesis show significant influences from parametrisations of the components of the statistical pattern recognition (SPR) pipeline, the setup of the steganalysis application scenario as well as from potential post-processing operations. These influences are discussed in the summaries on investigation tasks B and C, which both focus on the question how adequately the application scenarios can be implemented with the introduced approach (research challenge c)), below.

The investigations on the impact of application scenario specific intrinsic influences to the statistical pattern recognition (SPR) process (**investigation task B**) look into the influences arising from different instantiations of the SPR pipeline. This application task is mainly fuelled by two realisations: first, that the process of statistical pattern recognition (SPR) is a powerful but complex method, and second, that many different classification algorithms (as core component of the SPR process) exist, which allow for a successful detection of steganographic embedding into audio material.

One important statement is made in this thesis on the nature of the classification problem encountered in steganalysis. While most publications in the state-of-the-art model it as a two-class problem, a small number of publications strongly argue that steganalysis is a two-class decision problem build upon a multi-class detection problem (see e.g. [Provos02]). In this thesis both general modelling approaches (two-class and multi-class) are implemented. The results imply for both detection performances significantly better than guessing. Nevertheless, future work would have to be invested into a more detailed investigation on the pros and cons of both modelling approaches.

The complex SPR process can be considered as a four component processing pipeline (see section 2.4). In the following the evaluated influences to those four components are discussed:

- Pre-processing: the pre-processing operations have been restricted in this thesis to the absolute minimum (mostly windowing with a fixed window size)

- Feature extraction: the features are the enabling part of the pattern recognition method. If they allow the distinction between pattern and background and between different patterns, then a successful application of this method is possible. Here, with the AAFE and known good audio feature extractor is chosen for the most part of the investigations performed.

- Feature selection: this component complements the feature extraction by identifying the significant features and therefore allowing for the removal of the insignificant ones. The feature selection concept presented in this thesis is considered significant as well as representative for audio steganalysis because it is applied to a large multi-genre audio test set used as basis for a range of different information hiding (IH) algorithms covering the different embedding domains (time-, frequency- and wavelet domain). For each of the evaluated algorithms, the feature selection identified different segmental features as being relevant.
  The results of a PCA performed on the feature space of 590 segmental features computed by AAFE version 2.0.5 imply that for audio steganalysis within this feature space only about 150 independent dimensions exist (see section 5.1.5). For the global features the PCA identifies 11 independent dimensions in the 17 different features. A reduction of the feature space to uncorrelated features would significantly reduce the runtime of the classifiers.

- Classification: As stated above on the methodology and solution concept used in this thesis (see e.g. section 3.1.3), the choice of classifiers is for this thesis is restricted to the application of already existing classifiers as implemented in WEKA (version 3.6.1), presumably showing very different performance in terms of classification accuracy achieved and computation time requirements. Here, an application specific benchmarking scheme for existing classifiers is introduced, aiming at the identification of suitable classifiers for the audio steganalysis application scenario. The results of this classifier selection are presented in detail in section 5.1.4. Results for the application of clustering-driven classification show that this approach is outperformed by the supervised techniques. Supervised classification can be successfully used for audio steganalysis but so far no feature extractor / single classifier combination has been found that wields perfect results (a detection performance of $\kappa = 1$, preferably at a low computational run-time).

The summarising statement for investigation task B is: *The results for the supervised classification evaluations presented in section 5.1.4 show that: The detectability of an information hiding algorithm in audio steganalysis strongly depends of the availability of suitable features. For some of the evaluated algorithm (especially A_{W1}) the used feature set allow for the reliable detection of the impact of the embedding function. For other algorithms (especially A_{S1}) no such features are currently implemented in the used feature extractor.*

This implies, together with the context dependency identified, that the approach would have to be adapted and optimised (in terms of channel and steganographic embedding strategy assumptions) prior to any field application.

The investigations on influences outside the statistical pattern recognition (SPR) process on the performance of the scheme (**investigation task C**) focus on selected, common audio signal post-processing operations. Regarding these common audio signal post-processing operations, the investigations performed in section 5.3 show a negative effect on the SPR-based audio steganalysis approach. Therefore, if its application as a specialised integrity verification mechanism is considered, the implementation of the mechanism should undergo extensive plausibility evaluations against other audio signal modifications (encoding, re-sampling, etc.) that are likely in the considered application field or counter-forensics that might be used by the operators of a steganographic channel.

The summarising statement for investigation task C is: *The SPR-based audio steganalysis approach introduced here is not only sensitive to the steganographic message embedding but potentially also to other signal modifications (non-malicious and malicious alike). Since the approach is following the same methodology as the majority of approaches in the state-of-the-art in this field, it might be important to evaluate such sensitivity and plausibility evaluations also for other approaches found in the literature.*

With the summarising statements for investigation tasks A, B and C for both exemplary selected application scenarios (audio steganalysis and microphone forensics), part of the question raised by research challenge (c) on how adequately the application scenarios can be implemented with the introduced approach is answered. The other part of this answer is given in the comparison with the state-of-the-art in both application scenarios in chapter 7.

5.4.2 Reflection on the evaluation criteria derived from the Daubert standard

The summary table presented below is derived from table 3.2 in section 3.3. Here, the progress made within this thesis in the application scenario of audio steganalysis in regard to the criteria derived from the FRE rule 702 and the Daubert standard is summarised.

Table 5.23: Progress made in this thesis for audio steganalysis – projection onto the Daubert criteria

Criterion	Description / Progress made
FREC0	**Description** ([LLI10a]): "*the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue*" **Progress made**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.
FREC1	**Description** ([LLI10a]): the investigation (which leads to the corresponding expert testimony) is "*based upon sufficient facts or data*" **Progress made**: Case specific, the only significance arises due to the fact that the term "*sufficient*" has to be manifested into training and testing (application / evaluation) set sizes. Here, a rough estimate is presented for five chosen IH algorithms and the application specialisation on speech content. With those restrictions a model size of 400 feature vectors per reference file seems to be sufficient. Unfortunately this figure cannot be generalised due to the strong content dependency of the introduced approach. Therefore, the question what "*sufficient*" means for the composition of the training data for audio sets that are not limited to speech data is still an open question to be answered by future research.
FREC2	**Description** ([LLI10a]): the investigation is based upon "*reliable principles and methods*", preferably scientific methodology and knowledge **Progress made**: Chapter 2 as well as sections 3.1, 3.2 and chapter 4 of this thesis are dedicated to establish the fact that the two exemplary selected audio forensic methods are implemented as deterministic processes using the decades old and well accepted methodology of statistical pattern recognition (SPR). The fact that the audio steganalysis application scenario can indeed be solved by SPR is successfully addressed in section 5.1.1.
FREC3	**Description** ([LLI10a]): the (forensic) methods are applied "*reliably to the facts of the case*" **Progress made**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.

Continued on Next Page. . .

Table 5.23 – Continued

Criterion	Description / Progress made
DC1	**Description** ([LLI10b]): "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"; summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*" **Progress made**: This criterion imposes the most important task to the practical investigations performed within this thesis: The main part of chapter 5 is dedicated to exactly this goal, trying to establish within which limits the proposed media forensic methods can give plausible results. It has to be admitted that the size of the experiments performed might still lack generalisability but the methodology and evaluation concepts show that there are ways for objectively challenging the introduced approach.
DC2	**Description** ([LLI10b]): "*whether the technique or theory has been subject to peer review and publication*"; with "*publication*" meaning 'open publication' **Progress made**: This criterion is not translated into tasks but instead requires the author to interact with the scientific community relevant for the chosen application scenario. To address this criterion this thesis is submitted for (peer) review, as have been the accompanying journal, conference and workshop papers on the work on the audio steganalysis application scenario. The reviewer comments received have helped shaping the described approach as well as its evaluations.
DC3	**Description** ([LLI10b]): "*the known or potential rate of error of the technique or theory when applied*" **Progress made**: It has to be admitted that the size of the experiments performed might still lack generalisability, but the detection performances achieved in evaluations on two-class as well as multi-class setups for audio steganalysis against multiple steganographic tools are promising. Nevertheless, they would have still have to be improved to achieve detection performances and corresponding error rates that are fit for application in court cases. Additionally, any application in court might require the implementation of the ideal forensic steganalysis process as it is described in section 3.1.1, explicitly binding the (digital piece of) evidence to the court case.
DC4	**Description** ([LLI10b]): "*the existence and maintenance of standards and controls*" **Progress made**: The task that would be derived from this criterion would be the compilation of the work into standards together with or within a standardisation body. This complex process is outside the scope of this thesis, no progress made in this regard.
DC5	**Description** ([LLI10b]): "*whether the technique or theory has been generally accepted in the scientific community*" **Progress made**: This criterion is similar to DC2 in its meaning and in the fact that it is not translated into tasks, no progress made in this regard.

6 Investigations for Application Scenario 2: Microphone Forensics

This chapter is dedicated to the experimental evaluation of the performance of an instantiation of the introduced general-purpose SPR-based audio forensics approach for microphone forensics. It is structured along the investigation tasks A) to C) defined in section 3.3.

The **empirical ground truth** requested in investigation task A) is established for the microphone forensics approach developed in this thesis in section 6.1. For the performed investigations, this section shows the statistical relevance of the introduced solution approach as well as provide required knowledge for the following evaluations.
In section 6.2 the **impact of application scenario specific intrinsic influences to the statistical pattern recognition (SPR) process** is considered by investigations on the different degrees of freedom in the recoding process identified above in section 3.2.3 (recording environment, microphone orientation in reference to the source of sounds[64], the mounting of the microphone and content influences).
The investigation task C) (**Influences to the performance of the scheme, which are outside the SPR process**) is split for microphone forensics into two parts: In section 6.3 influences from assumedly non-malicious modifications (here normalisation, MP3 conversion and de-noising as well as playback recording) are considered, while section 6.4 considers with audio file composition an attack scenario for microphone forensics.
Finally, the major results of the investigations performed within this chapter are summarised in section 6.5, including a mapping of the progress made on this application scenario to the Daubert criteria as specified in section 2.2 and its subsections.

As usual for a dissertation project in the field of computer science, in compliance with Daubert criterion DC2 (*"whether the technique or theory has been subject to peer review and publication"* [LLI10b]) and to give other researchers / reviewers the chance to dispute the theory and its application (Daubert criterion DC5 "*whether the technique or theory has been generally accepted in the scientific community*" [LLI10b]), parts of the results presented in this chapter have been previously published in workshop and conference proceedings. The corresponding papers are (in chronological order):

- **2007**:
  - [Kraetzer07c] presented at the 9th ACM Workshop on Multimedia and Security 2007 in Dallas, Texas, USA, September 20th-21st, 2007.
- **2009**:
  - [Kraetzer09b] presented at the 11th ACM Workshop on Multimedia and Security 2009 in Princeton, NJ, USA, September 7th-8th, 2009.
  - [Buchholz09] presented at the 11th Information Hiding Conference 2009 in Darmstadt, Germany, June 7th-10th, 2009.

[64] Only singular sound sources are used here, the evaluation of the impact of multiple sound sources to microphone forensics is reserved for future work.

- **2011**:
  - [Kraetzer11] presented at the Media Watermarking, Security, and Forensics XIII, IS&T / SPIE Electronic Imaging 2011 in San Francisco, CA, USA, January 23th-27th, 2011.
- **2012**:
  - [Kraetzer12b] presented at the Media Watermarking, Security, and Forensics XIV, IS&T / SPIE Electronic Imaging 2012 in San Francisco, CA, USA, January 22th-26th, 2012.

The major results from these publications are recapitulated in the following sections, where they are further substantiated and accompanied by additional investigations as necessary.

6.1 Establishing some empirical ground truth for the used microphone forensics approach

Prior to the extensive investigations performed in the following sections, some empirical basis has to be established here. On one hand, this is done to show that microphone forensics can actually be solved by statistical pattern recognition and that the performed empirical evaluations are of statistical significance, on the other hand, it allows the following investigations to be accelerated (e.g. by using only classifiers or features that have been identified as suitable in the performed classifier and feature selection).

6.1.1 Intra-class (intra microphone class) classifications

One of the most fundamental required basic investigations is the observation of the detection performance of an extensive number of classifiers in intra-class classifications on sets of identical microphones. One of these two evaluations is using the recording set of condenser microphones *RS4_Rode* and the other one using the set of dynamic microphones *RS4_Beyer*. By the usage of these two sets, representative candidates of these two most prominent microphone classes are selected for evaluation in an intra-class setup considering sets of identical microphones. The experimental setups[65] *Mic-01* and *Mic-02* (see table 11.1 in appendix C (starting on page 201)) summarise the practical setups for these evaluations.

Excerpts out of the results for these experiments are presented in table 6.1 and table 6.2 – the complete summaries are presented in section 6.1.3.
From the overall set of 74 (supervised) classification algorithms, for the set of four identical Rode condenser microphones experimental setup *Mic-01* a maximum Kappa value of $\kappa = 0.678$ (equals a classification *accuracy* of 75.88% in this four-class classification problem) is achieved under the given constraints[66]. In summary, over all ten considered recording environments (rooms), 23 of the classifiers perform with an average detection performance of $\kappa > 0.467$ (*accuracy*$>$ 60% in this 4-class problem).

Table 6.1: Detection performances better than $\kappa = 0.467$ (*accuracy*$>$ 60%) for the experimental setup *Mic-01*

	Average over all 10 recording environments
Maximum achieved κ value	0.678
Maximum achieved *accuracy*	75.88%
Classifiers with: $0.467 \leq \kappa < 0.733$ (60 $\leq$*accuracy*$<$ 80%)	23
Classifiers with: $0.733 \leq \kappa < 0.867$ (80 $\leq$*accuracy*$<$ 90%)	0
Classifiers with: $\kappa \geq 0.867$ (*accuracy*$\geq$ 90%)	0

The results for the set of four identical Beyerdynamics dynamic microphones in the similarly conducted experiment *Mic-02* are even better: a maximum Kappa value of $\kappa = 0.767$ (classification *accuracy* of

[65] The experimental setups used in chapters 5, 6 and 8 are identified by underlined and italic font setting (e.g. *Mic-01*) – they are resolved in appendix B (audio steganalysis) and C (microphone forensics).

[66] Some classifiers are excluded due to their computation time behaviour (timeout set at 60h) – nevertheless these classifiers have shown in some preliminary test significant accuracies.

82.51%) is achieved under the given constraints and 31 of the classifiers perform over all ten considered recording environments (rooms) with an average detection performance of $\kappa > 0.467$ (*accuracy*$> 60\%$ in this 4-class problem). Four of the classifiers perform better than $\kappa = 0.867$ (an *accuracy* better than 80%). More details on these experiments are discussed in section 6.1.3.

Table 6.2: Detection performances better than $\kappa = 0.467$ (*accuracy*$> 60\%$) for the experimental setup *Mic-02*

	Average over all 10 recording environments
Maximum achieved κ value	0.767
Maximum achieved *accuracy*	82.51%
Classifiers with: $0.467 \leq \kappa < 0.733$ ($60 \leq$*accuracy*$< 80\%$)	27
Classifiers with: $0.733 \leq \kappa < 0.867$ ($80 \leq$*accuracy*$< 90\%$)	4
Classifiers with: $\kappa \geq 0.867$ (*accuracy*$\geq 90\%$)	0

Résumé for this section: The investigation results presented in this section can be summarised as follows: Notwithstanding the fact that all classifiers are used in default parametrisation – which has to be assumed to be sub-optimal (a fact which would require more detailed considerations on classifier optimisation and -generation, which are outside the scope of this thesis) – the detection performance results achieved in the intra-class recording classifications can be considered significant. If the agreement-to-statistical-confidence mapping introduced in section 4.1.4 is used, the best results presented above would be in the ranges of substantial or almost perfect agreement (cf. [Landis77]), equivalent to fair to good or even good statistical confidence. As a result it seems to be possible to distinguish between recordings made by different microphones of the same brand and model and in the experiments a sufficiently large number of different classifiers are capable of doing so.

6.1.2 Number of feature vectors in training and the detection performance

Addressing the question of statistical significance for the performed experimental validations here first the impact of the number of feature vectors per used file is evaluated. Table 6.3 (adapted from [Kraetzer07c]; experiment *Mic-Kraetzer2007ACM* – see table 11.1 in appendix C (starting on page 201)) shows exemplary the impact of the scaling of the number of input feature vectors on the classification *accuracy* for room *R01*. In the two cases of Bayesian classification the increasing of the number of feature vectors per file results in increasing classification *accuracy* on the microphones. The best result is found with $\kappa = 0.522$ in the case of *weka.classifiers.bayes.NaiveBayes* with 10-fold stratified cross-validation and 800 vectors per file. Both cases of Bayesian classification (percentual split (66%) and 10-fold cross-validation) show very similar results.

Table 6.3: Detection performance κ for room *R01* for different numbers of vectors computed per file used for training and testing for three exemplary selected classifiers (adapted from experimental setup *Mic-Kraetzer2007ACM* as used in [Kraetzer07c])

	Number of feature vectors							
Classifier	100	200	300	400	500	600	700	800
NaiveBayes (66%)	0.466	0.468	0.482	0.503	0.512	0.510	0.519	0.518
NaiveBayes (10-fold cross-valid.)	0.433	0.468	0.486	0.497	0.504	0.514	0.520	0.522
SimpleKMeans	0.229	0.211	0.254	0.210	0.224	0.199	0.221	0.207

The other nine rooms in experimental setup *Mic-Kraetzer2007ACM* show the same behaviour as *R01* in the scaling tests, therefore a detailed description of the results for each room is omitted here. Concluding the results of the Bayesian classification, it can be stated that even for small numbers of feature vectors for training significant detection performance results are achieved. Since the average results increase with the increasing number of vectors per file, future work will have to be investigated into investigations on optimal training set sizes.

The results of the clustering using *SimpleKMeans* in experimental setup *Mic-Kraetzer2007ACM* are with a maximum κ of 0.254 lower than the results from the Bayesian classification but still far higher than a random classification on five equally distributed classes (i.e. $\kappa = 0$). If their scaling behaviour is evaluated for increasing numbers of feature vectors used per file they show, in contradiction to the Bayesian classification, no increase in the classification *accuracy*.

A similar test run on *RS4_Beyer* (experimental setup Mic-03 – see table 11.1 in appendix C (starting on page 201)) for recording location *R01* shows similar results for the scaling of the *accuracy* of supervised classification techniques. Table 6.4 and table 6.5 show the scaling behaviour of the classification results achieved by four randomly selected classifiers in 10-fold stratified cross-validation on vector fields containing between 100 and 800 feature vectors per file recorded on basis of the reference set *ref10* (equals overall numbers between 4000 and 32000 vectors in 10-fold stratified cross-validation).

Table 6.4: Scaling of the detection performance κ and detector runtime of four selected classifiers (default parametrisations, 10-fold stratified cross-validation) on *RS4_Beyer* (*R01*) with 100, 200, 300 and 400 vectors per reference file; exp. setup *Mic-Kraetzer2007ACM*

	Number of feature vectors			
Classifier	100	200	300	400
bayes.NaiveBayes	0.157 $(20s)$	0.165 $(40s)$	0.174 $(61s)$	0.173 $(83s)$
functions.Logistic	0.632 $(1372s)$	0.680 $(2530s)$	0.704 $(3708s)$	0.710 $(5852s)$
meta.RandomSubSpace	0.641 $(352s)$	0.685 $(773s)$	0.723 $(1188s)$	0.728 $(1650s)$
trees.RandomForest	0.552 $(38s)$	0.594 $(83s)$	0.615 $(127s)$	0.639 $(173s)$

Table 6.5: Scaling of the detection performance κ and detector runtime of four selected classifiers (default parametrisations, 10-fold stratified cross-validation) on *RS4_Beyer* (*R01*) with 500, 600, 700 and 800 vectors per reference file; exp. setup *Mic-Kraetzer2007ACM*

	Number of feature vectors			
Classifier	500	600	700	800
bayes.NaiveBayes	0.178 $(106s)$	0.184 $(136s)$	0.187 $(171s)$	0.187 $(216s)$
functions.Logistic	0.718 $(7608s)$	0.728 $(10797s)$	0.726 $(13850s)$	0.731 $(16948s)$
meta.RandomSubSpace	0.740 $(2131s)$	0.742 $(2638s)$	0.747 $(3201s)$	0.762 $(3738s)$
trees.RandomForest	0.653 $(226s)$	0.661 $(278s)$	0.662 $(332s)$	0.676 $(385s)$

The results presented in table 6.4 and table 6.5 show, apart from some small glitches, a steadily increase of the detection performance achieved. The strongest increase can be seen for all classifiers when the input vector field size is increased from 100 to 200 feature vectors per reference file. At the same time, the duration of the evaluations increases linearly. For the classifier with the longest duration (*functions.Logistic*) the test duration increases from 1372 seconds for 100 feature vectors per reference file to 16948 seconds at 800 feature vectors per reference file.

Résumé for this section: While the increasing model size slightly increases the detection performance achieved, it also shows a strong impact to the run-times required. Already with small set sizes of 100 feature vectors per reference file significant results are presented by the exemplary chosen classifiers in 10-fold stratified cross-validation. By increasing the set size from 100 feature vectors per reference file to 200 the detection performance achieved shows a significant increase, which is less strong for the following increases. At the same time the required computation increases strongly. From the tests performed here, it is assumed, based on the results shown above, that a training set size of 200 feature vectors per reference file (resulting for the 10 chosen references in about in 2000 representative feature vectors per microphone) is suitably enough for the evaluations. At that size for each of the four microphones of *RS4_Beyer* and with a dimensionality of 590 attributes per vector WEKAs implementation of a multilayer-perceptron returned $k > 0.8$ (*accuracy* of more than 85%) at an inacceptable high computation time of more than 100 hours on the test machine[67]. Any further increase results in only slightly better classification accuracies and strongly worse run-times. Future work will have to be investigated into investigations on optimal (regarding detection performance as well as throughput considerations) training (and test) set sizes (see section 8.2.1).

[67] A Intel Core 2 Duo E8400 CPU 3GHz with 4 GB RAM machine running Microsoft Windows XP, WEKA v.3.6.1 on Java SE 6 (32-bit Windows) with 1.6 GByte allocated RAM for each WEKA instance (i.e. classifier, clusterer, PCA or feature selector).

6.1.3 Application scenario specific classifier selection for microphone forensics

Here, a summarising review on the detection performance of existing classification algorithms on selected microphone forensics tasks is given for this thesis. The practical observations are limited to the performance of the classification algorithms currently implemented in the renown data mining suite WEKA (v.3.6.1) and one feature extractor (AAFE).

In general with (supervised) classification and clustering two different approaches to classification exist (see section 2.4.4). It has been established in [Kraetzer07c] that both approaches can be applied in microphone forensics, but to a different extend of success. The evaluations performed in [Kraetzer07c] indicate that (supervised) classification outperforms clustering for microphone as well as room / recording environment classification. These observations are substantiated within this thesis – see below where all clustering and classification algorithms implemented in WEKA (v.3.6.1) are reviewed for microphone classification and section 6.2.1, where the performance of a selected clusterer and selected classification algorithms are reviewed as representative candidates for room / environment classification.

The table 6.6 below averages the detection performance results taken from [Kraetzer07c]. The κ values are computed in the experimental setup *Mic-Kraetzer2007ACM* for all ten recording locations evaluated in this experiment, a fixed number of feature vectors per file (800) and using representative candidates for supervised classification (*NaiveBayes*) and clustering (*K-means*).

Table 6.6: Detection performance (κ values) for all ten recording locations, NaiveBayes classification and KMeans clustering applied (experimental setup *Mic-Kraetzer2007ACM*, adapted from [Kraetzer07c])

	R01	*R02*	*R03*	*R04*	*R05*	*R06*	*R07*	*R08*	*R09*	*R10*	**Avg.**
NaiveBayes (66%)	0.517	0.611	0.526	0.667	0.603	0.615	0.618	0.685	0.650	0.696	0.619
NaiveBayes (10-fold cross-valid.)	0.522	0.620	0.533	0.680	0.608	0.616	0.629	0.689	0.656	0.700	0.625
SimpleKMeans	0.207	0.184	0.226	0.169	0.270	0.212	0.227	0.295	0.167	0.127	0.208

For the Bayesian classifier the Kappa values achieved are in the range $[0.517, 0.700]$, depending on the recording location. Both cases of Bayesian classification show very similar results. With $\kappa = 0.517$ the lowest detection performance in the Bayesian microphone classification is found in the case of *R01* (which is a large, quiet office). Noisy environments, like *R04* and *R08* (a noisy lab and a busy outside parking lot), seem to have a positive effect on the classification performance (second and third highest results). The best microphone classification in this experiment with $\kappa = 0.7$ is achieved on the material recorded in a small stone stairwell with a strong echo (*R10*).

For the clustering using *SimpleKMeans*, the results are in the range $[0.127, 0.295]$. These results are much lower than the ones achieved with Bayesian classification on the same test material, but nevertheless they are still by far better than 'guessing' at the result ($\kappa = 0$), therefore they are still considered significant within this work. In this case the worst result with $\kappa = 0.127$ is computed for the recordings made in *R10* (the stairwell, which was the best case in the Bayesian classifications). With $\kappa = 0.295$ the best result is given for *R05* (a large lecture hall).

Within this thesis an additional evaluation on this matter is performed, to verify the basic assumption established in [Kraetzer07c] that the detection performance achieved in **clustering** is by far outperformed by the accuracies achieved in supervised classifications.

For these conclusive evaluations on clustering subsets from the *RS4_Rode* and *RS4_Beyer* test sets are evaluated because the corresponding intra-class evaluations propose the biggest challenge in microphone identification and therefore should allow giving an estimation on the worst case performance of the clustering methods as mechanisms applicable in the field. The following representative setups *Mic-Clustering-RodeR01* and *Mic-Clustering-BeyerR01* are used. The setup uses a smaller number of feature vectors per reference (only 200 instead of the 800 in *Mic-Kraetzer2007ACM* – for a discussion of the impact of training/test set sizes see section 6.1.2) but the vectors are of much higher dimensionality (590 instead of 63), since a newer version of the AAFE is used (v.2.0.5 instead of v.1.0.3).

Table 6.7 summarises the detection performance achieved in the application of all of WEKAs eight clustering algorithms using the experimental setup *Mic-Clustering-RodeR01*.

Table 6.7: Detection performance on *RS4_Rode* in *R01* with 200 features vectors of 590 dimensions per file (*ref10*) for all 8 clustering algorithms in WEKA (v.3.6.1); experimental setup *Mic-Clustering-RodeR01*

Clustering algorithm	κ value
cobweb	n.a. – timeout (12h)
DBScan	-0.057
EM	0.059
FarthestFirst	0.044
FilteredClusterer	-0.049
MakeDensityBasedClusterer	0.024
OPTICS	n.a. – crash error
SimpleKMeans	0.031

In two out of the eight test cases the (for *weka.clusterers.cobweb* and *weka.clusterers.OPTICS*) the evaluations are terminated with an error. In the first case the defined timeout of 12 hours is hit, in the second case WEKA crashed with an 'error 0' error at 1.5 GB allocated RAM. For the other six cases the results shown confirm the observations made in [Kraetzer07c] on the detection performance of clusterers: the κ values achieved in this application scenario and on the used test material / feature set combination is much too close to 0 (i.e. the probability of guessing) to be of any use.

Table 6.8 shows the in general same performance for *Mic-Clustering-BeyerR01* as table 6.7 does for *Mic-Clustering-RodeR01*: the same two clustering algorithms terminate with an error and the rest performers in average slightly, but not much, above the probability of guessing (i.e. $\kappa = 0$).

Table 6.8: Detection performance on *RS4_Beyer* in *R01* with 200 features vectors of 590 dimensions per file (*ref10*) for all 8 clustering algorithms in WEKA (v.3.6.1); experimental setup *Mic-Clustering-BeyerR01*

Clustering algorithm	κ value
cobweb	n.a. – timeout (12h)
DBScan	-0.021
EM	0.102
FarthestFirst	0.033
FilteredClusterer	0.075
MakeDensityBasedClusterer	0.090
OPTICS	n.a. – crash error
SimpleKMeans	0.056

The same evaluations are run again in this thesis, using a different feature set (experimental setup *Mic-Clustering-RodeR01-selectedfeatures* – table 11.1 in appendix C (starting on page 201)) to substantiate the results and to eliminate the presumption that a too high dimensional feature representation might influence the classification negatively[68]. From the recording set *RS4_Rode* the room *R01* is selected as a representative material. From the 590 dimensional segmental features set computed by AAFE v.2.0.5 the 20 most significant ones are selected (see section 6.1.4).

Table 6.9: Detection performance on *RS4_Rode* in *R01* with 200 features vectors of 20 dimensions per file (*ref10*) for all 8 clustering algorithms in WEKA (v.3.6.1); experimental setup *Mic-Clustering-RodeR01-selectedfeatures*

Clustering algorithm	κ value
cobweb	0
DBScan	0
EM	0.035
FarthestFirst	0.007
FilteredClusterer	0.024
MakeDensityBasedClusterer	0.027
OPTICS	n.a. – crash error
SimpleKMeans	0.024

The results presented for the reduced feature set shown in table 6.9 indicate the same outcome as the observations presented above. The application of clustering algorithms as classification mechanisms seems to have no benefit for the microphone forensics application scenario considered within this thesis.

[68] Also known as 'curse of high dimensionality' [Bellman61].

Due to the large number of **supervised classification** algorithms in WEKA (74 in the used version 3.6.1) it is hardly feasible to run all experiments within this thesis with all classifiers. Therefore, the experimental validations presented are usually carried out using only a subset of all available classification methods. Considerations on the required application scenario specific classifier selection are presented below, to act as a point of reference for the other microphone forensics evaluations within this thesis. First ideas for a benchmarking of classification approaches, which outside the actual scope of this thesis, are presented in section 8.2.1.

To allow for some generalisability of the observations made here, extensive intra-class practical evaluations close to the constraints imposed by WEKA (maximum locatable memory within the used 32-Bit Java runtime environment) are performed using the complete 590 dimensional feature space provided by the segmental features[69] in AAFE v.2.0.5. With the two audio recording sets *RS4_Rode* and *RS4_Beyer* introduced in section 4.3.1 two representative sets for classifier performance evaluation exist for usage within this thesis. On these two sets all 74 classifiers available in WEKA v.3.6.1 are used in 10-fold stratified cross-validation to determine those who are most suitable for the microphone forensics approach pursued here. The required experimental tests are run on vector fields with 8000 feature vectors (4 microphones times 200 feature vectors per reference file times 10 references) with a dimensionality of each feature vector of 590. Each test is run on 10 sets of recordings (one for each recording location) for each of the two microphone sets.
The Unix 'time' command is used to measure the time duration of each combination of model generation and classification. A timeout of 60 hours is defined at which a classifier is terminated if it has not finished until that point. The overall run time for these experiments on the reference machine for the test set *RS4_Rode* was about 3405 hours including timeouts or 1005 hours without the classifiers which resulted in timeouts. The overall run time for the *RS4_Beyer* was with 3179 respectively 779 hours shorter. This fact and the higher detection performance achieved on the material from *RS4_Beyer* imply that it proposes a somewhat easier intra-class pattern recognition problem than the microphone classification on *RS4_Rode*.

Classifier selection for suitability in microphone forensics on RS4_Rode

Summarising the results on *RS4_Rode* generated by using the experimental setup *Mic-01* (see table 11.1 in appendix C (starting on page 201)), it is shown in table 6.10 that the evaluated classifiers show a strong variation in their classification behaviour regarding the achieved accuracies.

Table 6.10: Comparison of the detection performances achieved with the AAFE v.2.0.5 segmental features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *Mic-01* – see table 11.1 in appendix C (starting on page 201))

	Average over all 10 recording locations
Maximum achieved κ value	0.678
Maximum achieved accuracy	75.88%
Time duration without timeouts (s)	361953.1
Duration including timeout test cases (s)	1225953.1
Performance histogram:	
Errors	18
$0.00 \leq \kappa < 0.04$ ($25 \leq$*accuracy*$< 28\%$)	9
$0.04 \leq \kappa < 0.20$ ($28 \leq$*accuracy*$< 40\%$)	12
$0.20 \leq \kappa < 0.40$ ($40 \leq$*accuracy*$< 55\%$)	6
$0.40 \leq \kappa < 0.60$ ($55 \leq$*accuracy*$< 70\%$)	11
$0.60 \leq \kappa < 0.80$ ($70 \leq$*accuracy*$< 85\%$)	17
$0.80 \leq \kappa < 1.00$ ($85 \leq$*accuracy*$< 100\%$)	0

In 18 out of the 74 cases the classification attempt terminated with an error. The following erroneous behaviours are observed: In two cases (*bayes.BayesNet*, *trees.BFTree*) the error is of the type 'memory error', in three cases (*bayes.ComplementNaiveBayes*, *bayes.NaiveBayesMultinomial*, and *bayes.NaiveBayesMultinomialUpdateable*) 'Numeric exception', in one case (*functions.LibLINEAR*) it is

[69] In preliminary tests the 17 global features also extracted by AAFE v.2.0.5 have been used as input for all 74 classifiers in WEKA and none of those features showed and significance in this application scenario.

'liblinear', in one case (*functions.LibSVM*) it is 'libsvm', in four cases (*functions.MultilayerPerceptron*, *rules.DTNB*, *trees.LMT* and *trees.NBTree*) 'timeout', in two cases (*meta.CostSensitiveClassifier* as well as *meta.MetaCost*) it is of type 'Cost file', in one case (*meta.GridSearch*) 'Unsupported Attribute' and in four cases (all four *weka.classifiers.mi.** classifiers) 'Format exception'.

For the 56 non-error cases summarised in table 6.10 nine can be considered as 'just guessing' at the true class of a sample (κ at about 0) and therefore completely unsuitable for microphone forensics. The other classifiers in their default parametrisations show detection performances up to $\kappa = 0.678$ (average over all ten rooms for *weka.classifiers.meta.RotationForest*). The top 20 of the classifiers for the experimental setup *Mic-01* (see table 11.1 in appendix C (starting on page 201)) are identified in table 6.11.

Table 6.11: Ranking by κ value of the best 20 classifiers for experiment *Mic-01* – see table 11.1 in appendix C (starting on page 201))

Ranking	Classifier	κ value	Avg. runtime (s)
Best	*meta.Decorate*	0.694	51399.65
2nd	*meta.RotationForest*	0.678	14012.96
3rd	*rules.PART*	0.65	6994.95
4th	*meta.EnsembleSelection*	0.649	33735.9
5th	*trees.J48graft*	0.646	2083.03
6th	*trees.RandomForest*	0.641	129.4
7th	*rules.JRip*	0.637	4163.79
8th	*meta.MultiClassClassifier*	0.634	1491.73
9th	*trees.J48*	0.634	1832.22
10th	*trees.SimpleCart*	0.629	2061.17
11th	*functions.SimpleLogistic*	0.627	20974.18
12th	*trees.REPTree*	0.619	395.28
13th	*meta.RandomSubSpace*	0.617	2018.56
14th	*functions.Logistic*	0.616	1726.59
15th	*meta.Bagging*	0.611	3123.13
16th	*meta.END*	0.611	9900.37
17th	*functions.SMO*	0.605	2289.37
18th	*meta.ClassificationViaRegression*	0.598	3137.65
19th	*meta.Dagging*	0.557	169.91
20th	*meta.RandomCommittee*	0.534	164.61

While the classifiers presented in table 6.11 show similar κ values, their runtime strongly varies. This fact motivates the first considerations on a benchmarking strategy for SPR-driven audio forensics presented in section 8.2.1.

Classifier selection for suitability in microphone forensics on RS4_Beyer

Summarising the benchmarking results on *RS4_Beyer* generated by using the experimental setup *Mic-02* (see table 11.1 in appendix C (starting on page 201)), it can be seen (cf. table 6.10 and table 6.12) that the experimental results are similar in distribution but marginally better than those discussed above for *Mic-01*.

Table 6.12 summarises the achieved classification accuracies for this experiment. It can be seen that not only the maximum achieved detection performance is higher but also more individual classifiers perform better, with 23 classifiers in the range $0.6 \leq \kappa < 0.8$, which is achieved on the Rode material only in 17 cases. The erroneous behaviour of 18 classifiers noted is exactly the same (also for the same reasons) as discussed in detail for *Mic-01*.

Table 6.12: Comparison of the detection performances achieved with the AAFE v.2.0.5 segmental features – overview over all 74 WEKA (v.3.6.1) classifiers (using experimental setup *Mic-02* – see table 11.1 in appendix C (starting on page 201))

	Average over all 10 recording locations
Maximum achieved κ value	0.767
Maximum achieved accuracy	82.51%
Time duration without timeouts (s)	280287.7
Duration including timeout test cases (s)	1144287.7

Continued on Next Page. . .

Table 6.12 – Continued

	Average over all 10 recording locations
Performance histogram:	
Errors	18
$0.00 \leq \kappa < 0.04$ ($25 \leq$*accuracy*$< 28\%$)	8
$0.04 \leq \kappa < 0.20$ ($28 \leq$*accuracy*$< 40\%$)	8
$0.20 \leq \kappa < 0.40$ ($40 \leq$*accuracy*$< 55\%$)	8
$0.40 \leq \kappa < 0.60$ ($55 \leq$*accuracy*$< 70\%$)	7
$0.60 \leq \kappa < 0.80$ ($70 \leq$*accuracy*$< 85\%$)	23
$0.80 \leq \kappa < 1.00$ ($85 \leq$*accuracy*$< 100\%$)	0

The top 20 of the classifiers for the experiment *Mic-02* (see table 11.1 in appendix C (starting on page 201)) are identified in table 6.13.

Table 6.13: Ranking by κ value of the best 20 classifiers for experiment *Mic-02* (see table 11.1 in appendix C (starting on page 201))

Ranking	**Classifier**	**κ value**	**Avg. runtime (s)**
Best	*meta.RotationForest*	0.780	8788.1
2nd	*meta.EnsembleSelection*	0.73738333	12684.6
3rd	*trees.FT*	0.73615	2561.5
4th	*functions.SimpleLogistic*	0.736	23375
5th	*meta.MultiClassClassifier*	0.72748333	3509.7
6th	*functions.Logistic*	0.72196667	2459.8
7th	*meta.RandomSubSpace*	0.7213	759.6
8th	*meta.Bagging*	0.7181	1261.6
9th	*meta.END*	0.71738333	4879.3
10th	*meta.Decorate*	0.71366667	27355.2
11th	*meta.ClassificationViaRegression*	0.71111667	3450.6
12th	*functions.SMO*	0.7077	3528.3
13th	*meta.Dagging*	0.66533333	384
14th	*meta.RandomCommittee*	0.65845	93.9
15th	*rules.PART*	0.64321667	3305.4
16th	*trees.RandomForest*	0.64181667	564.6
17th	*trees.J48graft*	0.63236667	900.5
18th	*rules.JRip*	0.63045	3869.1
19th	*trees.SimpleCart*	0.61851667	1986.2
20th	*trees.REPTree*	0.61408333	152

Comparing table 6.13 to the classifier ranking presented in table 6.11 for *Mic-01* it can be seen that 18 out of the top 20 are present in both tables. The differences are *trees.FT* (3rd for *Mic-02*, but 32th for *Mic-01*) and *trees.J48* (21th in *Mic-02* and 9th in *Mic-01*). Interestingly both are decision tree classifiers, a class which contains 12 out of WEKAs overall 74 classifiers, but which shows no significant influence in the first ten ranks of the classifier rankings presented above. Like in the case of the Rode microphones, the runtime of the classifications shows extreme differences between the different classifiers.

Résumé for this section: The performed practical investigations imply that WEKAs clustering algorithms are of no use for the microphone forensics presented here. Even if the number of clusters (here the number of microphone classes) for the audio material is known in advance (an unlikely scenario in practice) the detection performance achieved here is barely above $\kappa = 0$ and thereby much lower than the performance of the supervised classification approaches on the same material. Since the results for the clustering are generally worse for the microphone detection than the results for supervised classification, the discussions in [Kraetzer07c] and all our other papers on microphone forensics as well as within this thesis, are limited to the usage of supervised classification. As a result of this section, for all further experimental validations conducted within this thesis the table 6.11 and table 6.13 can be used as a reference for choosing suitable classifiers. In those two tables the two most dominant classes of WEKA classifiers in this set are *meta.** classifiers and *functions.**, all other classes show only limited significance.

6.1.4 Feature selection for microphone forensics

Regarding the question of usable features [Kraetzer07c] does show with its achieved results for inter-device analysis (for the used test set, classification techniques and selected audio features) that feature selection in microphone forensics seems to have no positive impact on the achieved detection performance, but it reduces computation times and generates domain knowledge.
In addition to the actual classification tests in [Buchholz09], a principal component analysis (PCA) is conducted in this paper to establish suitable parameters for the pre-processing performed (a threshold for silence detection). The outcome of the PCA shows a significant redundancy in the feature set used, implying that a feature space reduction might be possible without decreasing the classification performance. If the same PCA is conducted on the 590 dimensional feature vector generated by AAFE v.2.0.5 then 187 transformed components are identified as being responsible for 95% of the sample variance (on the *RS4_Rode* subset for recording location *R01*). A reduction of the dimensionality of the feature space would result in a significant decrease of the computation power required for the classification and would therefore strongly beneficial for the introduced approach.
Within this thesis, these first results on feature selection and feature (in-)dependency from [Kraetzer07c] and [Buchholz09], are further substantiated. To do so, two sets of intra-class classifications are performed and suitable features identified by feature ranking.

Segmental versus global features

In preliminary tests the 17 global features also extracted by AAFE v.2.0.5 have been used as input for all 74 classifiers in WEKA and none of those features showed any significance (i.e. κ values larger than 0) in this application scenario. Therefore those global features are neglected for the microphone forensics observations in the rest of this thesis.

Feature selection by feature ranking - the generation of domain knowledge

The feature selection on segmental features for their suitability in microphone forensics is performed as described in section 4.1.2. For the practical realisation of this feature selection process, as two independent information sources the recording sets *RS4_Beyer* and *RS4_Rode* are chosen (see experimental setup *Mic-Feature-Selection*) because their material is best suited for generalisable intra-class evaluations.
In table 6.14 the 30 best segmental features are identified as the output of this feature selection procedure.

Table 6.14: Best 30 segmental features, based on the fused rankings computed on *RS4_Rode* and *RS4_Beyer* (see experimental setup *Mic-Feature-Selection*)

Feature	*RS4_Rode* **Average rank**	*RS4_Beyer* **Average rank**	**Arithmetic mean**	**Final rank**
$sf_{d2FMFCC\_1}$	2.08	1.06	1.57	1
$sf_{d2FMFCC\_2}$	2.6	2.72	2.66	2
$sf_{d2FMFCC\_13}$	4.18	11	7.59	3
$sf_{d2FMFCC\_10}$	14.42	9.06	11.74	4
$sf_{d2FMFCC\_3}$	18.18	5.48	11.83	5
$sf_{d2FMFCC\_5}$	14.7	9.74	12.22	6
$sf_{d2FMFCC\_4}$	18.26	6.34	12.3	7
$sf_{d2FMFCC\_11}$	18.02	6.68	12.35	8
$sf_{d2FMFCC\_12}$	19.22	5.94	12.58	9
$sf_{d2FMFCC\_9}$	19.98	12.46	16.22	10
$sf_{d2FMFCC\_6}$	20.24	13.04	16.64	11
$sf_{d2FMFCC\_8}$	22.34	12.02	17.18	12
$sf_{d2FMFCC\_7}$	22.62	12.08	17.35	13
sf_{FMFCC_3}	16	27.94	21.97	14
sf_{FMFCC_12}	15.9	28.32	22.11	15
sf_{spec_11}	20.14	24.98	22.56	16
$sf_{RMS\_amplitude}$	19.06	26.78	22.92	17
sf_{FMFCC_10}	13.46	33.1	23.28	18
sf_{FMFCC_5}	13.66	33.16	23.41	19

Continued on Next Page...

Table 6.14 – Continued

Feature	*RS4_Rode* **Average rank**	*RS4_Beyer* **Average rank**	**Arithmetic mean**	**Final rank**
sf_{FMFCC_1}	20.64	29.76	25.2	20
sf_{FMFCC_4}	15.86	39.88	27.87	21
sf_{FMFCC_11}	15.7	40.16	27.93	22
sf_{FMFCC_2}	22.94	34.84	28.89	23
sf_{energy}	19.54	38.44	28.99	24
sf_{FMFCC_9}	18.1	41.76	29.93	25
$sf_{sp\_entropy}$	14.72	46.76	30.74	26
sf_{FMFCC_6}	18.54	42.98	30.76	27
sf_{spec_12}	32.48	32.62	32.55	28
$sf_{zero\_cross\_rate}$	48.7	17.16	32.93	29
$sf_{sp\_rolloff}$	22	49.1	35.55	30

The results summarised in table 6.14 imply that the second-order derivative filtered MFCCs clearly outperform every other class of features. The 13 features within this class occupy the 13 highest ranks within the fused ranking, followed by 10 further FMFCC-features within the next 17 ranks. This implies that the microphone influence in recording – the intrinsic pattern – manifests itself the strongest in higher order cepstral-domain features, while time-domain features, which are the natural domain of the audio signal, do not play a strong role in the classification, since they are influenced to strongly by the recorded content.

If only the 20 best features are used in classification on the four-class evaluations on this test material (experimental setups *Mic-01* vs. *Mic-RS4_Rode-Best20Features-only*), the classification *accuracy* drops in average for about 7.11% in comparison to the full feature set. Four classifiers, which originally hit the 60 hour timeout boundary defined for *Mic-01*, have no problem to keep below that boundary when using only 20 features instead of the full set of 590.

In contrast to the best ranking features presented above, table 6.15 bellow identifies the 25 worst performing segmental features. Here all formants as well as the two LSB features are found in the list of the least contributing segmental features.

Table 6.15: Worst 25 segmental features, based on the fused rankings computed on *RS4_Rode* and *RS4_Beyer* (see experimental setup *Mic-Feature-Selection*)

Feature	*RS4_Rode* **Average rank**	*RS4_Beyer* **Average rank**	**Arithmetic mean**	**Final rank**
sf_{MFCC_10}	502.9	554.22	528.56	566
sf_{spec_4}	520.9	545.14	533.02	567
sf_{spec_22}	537.48	537.04	537.26	568
sf_{spec_24}	528.22	548.46	538.34	569
sf_{median}	533.76	554.58	544.17	570
sf_{spec_1}	551.12	544.16	547.64	571
sf_{spec_5}	540.3	560.26	550.28	572
sf_{sp_bw}	551.62	559.3	555.46	573
sf_{mean}	564.08	552.8	558.44	574
$sf_{sp\_centriod}$	569.3	551.68	560.49	575
$sf_{sp\_irregularity}$	567.48	560.1	563.79	576
$sf_{sp\_base\_freq}$	573.74	556.14	564.94	577
sf_{LSBrat}	564.1	570.08	567.09	578
$sf_{formant\_Singer}$	578.74	566.64	572.69	579
$sf_{formant\_A1}$	578.76	571.9	575.33	580
$sf_{LSBflip}$	573.26	577.68	575.47	581
$sf_{formant\_I2}$	579.08	576.34	577.71	582
$sf_{formant\_E2}$	578.66	578.18	578.42	583
$sf_{formant\_U2}$	576.88	581.7	579.29	584
$sf_{formant\_E1}$	576.62	584.32	580.47	585
$sf_{formant\_O2}$	578.68	583.8	581.24	586
$sf_{formant\_A2}$	578.58	584.34	581.46	587
$sf_{formant\_O1}$	579.26	584.38	581.82	588
$sf_{formant\_U1}$	579.4	584.24	581.82	589
$sf_{formant\_I1}$	581.04	586.42	583.73	590

The results in table 6.15 indicate a very bad performance of the formant features. All 11 of them are found in this list. Also, a large percentage of the time-domain features is present in this set, including the two LSB-based features. This supports the assumption formulated above that time-domain features, which are computed in the natural domain of the audio signal, do not play a strong role in the classification, since they are influenced to strongly by the recorded content.

Impact to classifier run-time

As mentioned above, if only the 20 best features are used in classification on the test material the classification accuracy drops slightly in comparison to the full feature set. By the same feature space reduction, the average computation time is reduced by factor 32.7 (the feature space is reduced at the same time by $\frac{590}{20} = 29.5$, so a simple estimation would assume a roughly linear dependent relationship between the decrease of the dimensionality of the vector space and the decrease in required computation power).

Résumé for this section: Findings presented in [Kraetzer07c] on the general impact of feature selection are confirmed here. The feature selection indeed allows reducing the computation cost required in this microphone forensics approach while achieving similar classification accuracies.
The feature ranking presented in this section is considered significant as well as representative for microphone forensics because it was performed independently on two intra-class test sets of statistically significant sizes, composed from microphones of the two most common microphone classes (dynamic and condenser microphones). I.e. by the selection of representative and statistically significant training and test sets a generalisable answer is derived for the whole microphone forensics problem.
In the comparison of segmental versus global features it shows that the global features provided here by the used feature extractor AAFE v.2.0.5 are of no use for the introduced microphone forensics approach, while the segmental features achieve significant classification accuracies for selected classifiers.
Regarding the feature independency of the segmental features the investigations performed here show that they are strongly correlated, which implies strong potential for feature selection. In the feature selection evaluations performed this strong potential is confirmed: the impact to classifier runtime is rather dramatic – the classification process is speed up by a factor of about 30 while keeping the classification accuracies achieved on a nearly constant level.
Since the procedure ranks the features by their benefit in classification, this ranking then can be used to derive domain knowledge about the considered classification problem. The evaluations imply that the microphone influence in recording – the intrinsic pattern – manifests itself the strongest in higher order cepstral-domain features, while time-domain features do not play a strong role in the classification, since they are influenced to strongly by the recorded content.

6.2 Evaluation of different influences (degrees of freedom) in the recording process

One of the most important tasks for the microphone forensics approach introduced within this thesis is to determine which of the degrees of freedom in the recording of a sound has an impact to the achievable detection performances and therefore directly on the practicability of the approach. Following the design for the evaluations introduced in section 4.3.1, here the influences of the room / recording environment, the orientation and mounting as well as the recorded content are considered.

6.2.1 Influence of the recording environment

Regarding the room / recording environment identification question [Kraetzer07c] does show with its achieved results for inter-device analysis (for the used test set, classification techniques and selected steganalysis features) that room / recording environment classification is also possible but with a much smaller accuracy than microphone classification. The results that lead to this observation are recapitulated here and supplemented by additional experiments. Since it was shown in section 6.1.3 that clustering and (supervised) classification show different performance, the evaluations in this section also consider these two classification approaches differently.

Clustering

Table 6.16 summarises the tests performed in experiment *Mic-Kraetzer2007ACM* for the *SimpleKMeans* clustering. In this table, the average detection performance κ is computed in recording environment classification is given for the material recorded by each of the four microphones. The resulting κ values are in the range of $[0.011, 0.183]$ for the clustering using *SimpleKMeans*. The results are in general significantly lower than the results presented below for exactly the same test setup and supervised classification in table 6.17.

Table 6.16: Detection performance (κ value) for all four microphones in *RS1* and for different numbers of vectors computed per file and the *SimpleKMeans* clusterer (experimental setup *Mic-Kraetzer2007ACM*; adapted from [Kraetzer07c])

	Vectors per recording							
Microphone	100	200	300	400	500	600	700	800
M_1	0.152	0.106	0.056	0.037	0.069	0.034	0.069	0.073
M_2	0.160	0.108	0.094	0.104	0.113	0.106	0.137	0.119
M_3	0.183	0.068	0.081	0.076	0.066	0.058	0.064	0.071
M_4	0.092	0.046	0.030	0.041	0.011	0.026	0.014	0.029

(Supervised) Classification

Table 6.17 summarises the tests performed for the evaluation of the supervised classification part of the experiment *Mic-Kraetzer2007ACM*. In this table the κ value for all rooms is given for the material recorded by each microphone. The results in table 6.17 are in the range of $[0.155, 0.350]$ for the Bayesian classifier.

Table 6.17: Detection performance (κ value) achieved in room classification for all four microphones and for different numbers of vectors computed per file and Bayesian classification with 10-fold cross-validation and percentage split (66 to 34%) (experimental setup *Mic-Kraetzer2007ACM*; results adapted from [Kraetzer07c])

		Vectors per recording							
Microphone	**Mode**	100	200	300	400	500	600	700	800
M_1	perc. split 66%	0.231	0.155	0.160	0.175	0.171	0.169	0.164	0.179
	cross-validation	0.240	0.166	0.166	0.168	0.167	0.168	0.173	0.176
M_2	perc. split 66%	0.338	0.276	0.285	0.292	0.290	0.292	0.304	0.294
	cross-validation	0.350	0.295	0.288	0.290	0.292	0.296	0.299	0.299
M_3	perc. split 66%	0.230	0.198	0.208	0.206	0.207	0.206	0.203	0.209
	cross-validation	0.247	0.203	0.202	0.204	0.206	0.207	0.209	0.210
M_4	perc. split 66%	0.303	0.216	0.223	0.217	0.211	0.204	0.214	0.214
	cross-validation	0.317	0.234	0.220	0.216	0.213	0.211	0.215	0.218

When analysing the results for the Bayesian classifier, it can be seen that the test files recorded using the M_2 microphone (a simple headset) allow for the most precise classification of the rooms. Here, with $\kappa = 0.350$ the best detection performance (in the case of the *NaiveBayes* classifier using 10-fold cross-validation) is found when using 100 feature vectors from each file. An interesting observation in the Bayesian tests is the scaling of the detection performance with increasing number of feature vectors per file. For all tests if the number of feature vectors considered is increased from 100 to 200 the detection performance drops slightly, if the number of feature vectors considered is increased further, no significant change in the detection performance can be noticed (i.e. it stays roughly constant).

When comparing in the evaluations on experimental setup *Mic-Kraetzer2007ACM* the results for one microphone in all 10 rooms for the example of M_2 with 100 vectors per file and Bayesian classification using percentual split (see table 6.18) it can be seen that the results for a correct classifications (principal diagonal of the confusion matrix in table 6.18) are very inhomogeneous for the rooms.

Table 6.18: Normalised confusion matrix for the M_2 microphone, 100 vectors computed per file and the *NaiveBayes* classifier using percentual split (66%) for set generation (experimental setup *Mic-Kraetzer2007ACM*)

	R01	*R02*	*R03*	*R04*	*R05*	*R06*	*R07*	*R08*	*R09*	*R10*
R01	0.63	0.10	0.08	0.04	0.00	0.00	0.15	0.00	0.01	0.00
R02	0.03	0.49	0.01	0.17	0.08	0.05	0.01	0.11	0.04	0.02
R03	0.13	0.08	0.55	0.04	0.00	0.00	0.18	0.00	0.01	0.00
R04	0.00	0.24	0.00	0.52	0.04	0.03	0.00	0.08	0.07	0.02
R05	0.00	0.25	0.00	0.12	0.24	0.13	0.00	0.15	0.07	0.04
R06	0.01	0.40	0.00	0.19	0.05	0.12	0.01	0.13	0.07	0.02
R07	0.14	0.04	0.34	0.00	0.00	0.00	0.47	0.01	0.00	0.00
R08	0.00	0.05	0.00	0.13	0.08	0.03	0.00	0.54	0.13	0.05
R09	0.00	0.07	0.00	0.10	0.17	0.13	0.00	0.20	0.27	0.07
R10	0.00	0.15	0.00	0.09	0.20	0.10	0.00	0.17	0.09	0.19

The results on the main diagonal are in the range of $[0.12,\ 0.63]$ with an average of 0.404. In seven of the ten cases the highest classification result is achieved for the room the recording was made in. In the other three cases (*R05*, *R06* and *R10*) a different room wrongly achieves the highest number of classifications. From table 6.18 also rooms showing mutually similar and dissimilar behaviour can be deduced. An example for mutually similar rooms is found with *R03* and *R07* where on one hand the vectors recorded in *R03* are classified in 55% of the cases belonging to *R03* and 18% belonging to *R07* while on the other hand the vectors recorded in *R07* are classified 47% belonging to *R07* and 34% belonging to *R03*. An example for a set of dissimilar rooms is composed e.g. by *R01* and *R08*. There the number of vectors recorded in *R01* and falsely classified as recorded in *R08* is equal 0% and vice versa.

To substantiate the results for room / recording environment identification another set of experiments is run in this thesis, using different test recordings and an extended feature set (see the setup *Mic-Room-Classification-RS4-Selections* in table 11.1 in appendix C (starting on page 201)). From the list of best microphone forensics classifiers (see section 6.1.3) one Bayesian, one function, one meta-classifier and one tree are chosen for these classifications[70]. The test is run on 200 feature vectors per reference file.
The results for these evaluations, using one microphone out of the two recording sets *RS4_Rode* (microphone M_{16}) and *RS4_Beyer* (M_{20}), are presented in table 6.19.

Table 6.19: Detection performance (κ values) for room classification using microphones M_{16} and M_{20} and four selected classifiers (experimental setup *Mic-Room-Classification-RS4-Selections*)

	Source microphone	
Classifier	M_{16}	M_{20}
weka.classifiers.bayes.NaiveBayes	0.125	0.132
weka.classifiers.functions.SMO	0.354	0.404
weka.classifiers.meta.RandomCommittee	0.316	0.363
weka.classifiers.Trees.RandomForest	0.298	0.351

The results presented for room classification in table 6.19 can be summarised as follows: the detection performance achieved is significant, showing in some cases a Kappa value of $\kappa > 0.35$ (equivalent to accuracies of more than 40% in this 10-class classification problem). Nevertheless, the results achieved with the extended feature set of 590 dimensions (of AAFE v.2.0.5) are nearly in the same range as the results shown in table 6.17 for the lower dimensional feature set (of AAFE v.1.0.3).

Verification of a microphone against a model generated for a different room

To show how strong the influence of the room really is in the microphone classifications, it has to be evaluated what happens when the initial hypothesis about the used room / recording environment in such a test is wrong.

[70] For many of WEKAs classifiers this set size with its 1 microphone times 200 vectors times 10 references times 10 rooms times 590 (dimensionality of the feature vector) is already to large at 1.5 GB allocated main memory, therefore classifiers had to be selected which were still capable of performing the task at hand.

Table 6.20 summarises the results of tests on the recording environment (or room) influence using experimental setup *Mic-Room-Classification-RS4-WrongRoom* (see table 11.1 in appendix C (starting on page 201)) with a wrong and a true hypothesis on the used room / recording environment. In case of the wrong hypothesis the recordings are made in *R01* and tested against material recorded in *R06*. For the true hypothesis on the room both, training- and test material have been recorded in *R06*.

Table 6.20: Detection performance (κ values) for microphone classification – selected classifiers under a wrong and a true hypothesis on the used room / recording environment (experimental setup *Mic-Room-Classification-RS4-WrongRoom*)

Classifier	Model: *R01*	Model: *R06*
Classifier	Test material: *R06*	Test material: *R06*
weka.classifiers.bayes.NaiveBayes	-0.128	0.218
weka.classifiers.functions.SMO	0.565	0.760
weka.classifiers.meta.RandomCommittee	0.469	0.695
weka.classifiers.Trees.RandomForest	0.379	0.683

The results in table 6.20 show a significant drop in the detection performance achieved in case of the wrong hypothesis on the recording location. For the four selected classifiers the strongest drop occurs for *weka.classifiers.bayes.NaiveBayes* where κ drops from 0.218 down to -0.128, which is even lower than the probability of guessing correctly (i.e. $\kappa = 0$). For the other three classifiers the drop is less severe but they also show a strong decrease in their detection performance.

Résumé for this section: The evaluations on room / environment classification presented first in [Kraetzer07c] have been substantiated here by further experimental validation. Room classification is obviously possible, but with the current feature extractor and classifiers available it seems to be performing less good than microphone classification. Further work should extend the observations in this direction.
If a microphone is verified against a model generated in a wrong (different) room then the detection performance drops significantly as shown above. It can therefore be stated that the room / recording environment is a very strong influence on the recording behaviour of a microphone.

6.2.2 Orientation influence testing

For the experimental observations on the influence of the impact of the orientation of a microphone two different recording sets (*RS7* and *RS8* – see experimental setups *Mic-Orientation_Impact_RS7* and *Mic-Orientation_Impact_RS8* in table 11.1 in appendix C (starting on page 201)) are generated especially for this purpose. Here, first *RS7* is used to evaluate, for the applied method of statistical pattern recognition based microphone classification using segmental features, the impact of rotating a microphone by steps of 45° in the xy-plane. Then *RS8* is used to describe the impact of rotating a microphone by 180° in yz-plane.

To show how strong the influence of microphone orientations is, in comparison to the inter-microphone distance of different microphones of the same brand and model, two simple experiments are constructed (both described in experimental setup *Mic-Orientation_Impact_RS7*). For *RS7*, the recordings made at eight differently orientated positions with the microphone M_{22} are used in evaluations as test material against a model generated by a selected classifier (*weka.classifiers.meta.RandomSubSpace*) on *RS4_Beyer* in *R06* as well as on the same references (*ref2*, silence and a pure sinoid). The test hypothesis for both tests is: 'The candidate material is recorded by M_{22}.' If the accuracy achieved is equal or better than the results achieved by the classifier in the intra-class evaluations on *RS4_Beyer*, it can be assumed that the orientation is of limited impact to the microphone classification.

The average detection performance of *weka.classifiers.meta.RandomSubSpace* for all ten reference signals in *RS4_Beyer* (*R06*; 590 dimensional feature vector) is $\kappa = 0.76$ (*accuracy*$=81.86\%$). For the silence reference recorded in recording set *RS7* (and tested against the model generated from the corresponding *RS4_Beyer* material) a detection performance of $\kappa = 1.0$ is achieved. For the sinoid the detection performance is also $\kappa = 1.0$. The orientation seems to have no influence on the microphone classification problem, since the inter-microphone difference, even for microphones of the same brand and model, is higher than the differences between the recordings of one microphone in different orientations.

Another fact is highlighted by these results: *RS4* and *RS7* use the same microphones and hardware setup (room (*R06*), reference sounds, loudspeaker and soundcard) but between the times of recording lies a temporal distance of one year. Based on the perfect classification results achieved, it can be deduced from those evaluations that the statistical patterns which allow for the classification of the microphones show for this time span no aging behaviour / no significant change over time.

To show how strong the influence of microphone orientations is in *RS8*, the same experiments on the inter-microphone distance of microphones of the same brand and model versus different orientations described above for *RS7* are also run on *RS8* (see *Mic-Orientation_Impact_RS8*). The two different orientations used in the recording are 'head up' and 'head down', i.e. standing upright and a rotation by 180° in yz-plane.

Like in the tests on *RS7*, for both references recorded in recording set *RS8* (and tested against the model generated from the corresponding *RS4_Beyer* material) a detection performance of $\kappa = 1.0$ is achieved.

Résumé for this section: The orientation seems to have no influence to the microphone classification problem, since the inter-microphone difference, even for microphones of the same brand and model, is higher than the differences between the recordings of one microphone in different orientations.
Based on the perfect classification results achieved with a recently recorded test set on a training set recorded one year ago, it can be assumed from those evaluations that the statistical patterns which allow for the classification of the microphones show for this time span no aging behaviour / no significant change over time. Nevertheless, long term observations on this matter would be required using time spans of at least 5 to 10 years to allow for any generalisation on this fact.

6.2.3 Mounting influence testing

To show how strong the influence of microphone mounting changes is, in comparison to the inter-microphone distance of different microphones of the same brand and model, two simple experiments are constructed (both described in experimental setup *Mic-Mounting_Impact_RS9*). The eight recordings in different mountings of the microphone M_{22} used for the generation of *RS9* are used in these tests as test material against a model generated by a selected classifier (*weka.classifiers.meta.RandomSubSpace*) on *RS4_Beyer* in *R06* and on the same references (silence and sinoid; *ref2*). The test hypothesis for both tests is: 'The candidate material is recorded by M_{22}.' If the detection performance achieved is equal or better than the results achieved by the classifier in the intra-class evaluations on *RS4_Beyer*, it can be assumed that the mounting is of limited impact to the microphone classification.

For the silence reference recorded in recording set *RS9* (and tested against the model generated from the corresponding *RS4_Beyer* material) a detection performance of $\kappa = 1.0$ is achieved. For the sinoid the detection performance is $\kappa = 0.86$. The misclassifications are limited to exactly one mounting position (position 4, the microphone lying flat on the table). For this position all 200 corresponding feature vectors in this test are misclassified as originating from M_{23} instead of M_{22}. Mounting position 4 is the position where physically the assumedly strongest impact to the vibration behaviour of the microphone occurs – all other positions use some sort of microphone clamp mounted on a tripod, while for mounting position 4 the microphone is lying on a table which can be assumed to vibrate with the reference signal if this is strong enough, which is very well illustrated in figure ash which shows the intra-class distribution for the feature sf_{energy} for M_{22}. The figure 6.1 shows that the energy values for mounting position 4 are much larger than for all other mounting positions, which implies (due to the fact that no other environmental change happened between the recordings in the used anechoic chamber) that the table reverberates with the reference signal and that the microphone absorbs part of this reverberation into the recorded signal.

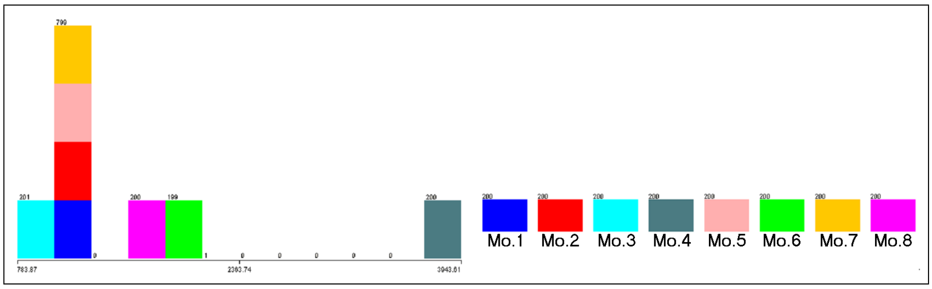


Figure 6.1: Distribution of values for the feature sf_{energy} on the harmonic (sine) signal in *RS9*

Résumé for this section: The mounting only seems to have influence in specific cases, where the vibration behaviour of the microphone is strongly influenced, like in the case where a microphone lies directly on a vibrating surface like a table. Otherwise, the inter-microphone difference, even for microphones of the same brand and model, seems to be higher than the differences between the recordings of one microphone in different mountings.

6.2.4 Classification using content selection as well as content dependent and independent training and testing

The experiments in [Kraetzer07c] are conducted using time-, frequency- and cepstral-domain features on context insensitive audio content evaluations. Another approach using content selection was examined in our laboratory with a master thesis [Dohnal08] using the test set *RS2* containing seven different microphones and using only Fourier coefficients as frequency-domain features. It aims mainly at the noise component $N_{mic}(f)$ in equation 2.4 in the context model for microphone recordings (see section 2.3.2). The template matching used in [Dohnal08] turned out to be inadequate for the classification of high-dimensional feature vectors[71], but these first results were a motivation to conduct further research on the evaluation of content selection as pre-processing for feature extraction and classification. In [Buchholz09] we use the Fourier features and silence threshold strategy introduced in [Dohnal08] together with WEKAs advanced classification techniques. The results are presented and discussed in detail in [Buchholz09] and are summarised here.

The classification results for $coef$ =2048 (experimental setup *Mic-BKD2009* – see table 11.1 in appendix C (starting on page 201)) are given in table 6.21, while the results for the evaluations using $coef$ =256 frequency coefficients are given in table 6.22. The second column in the two tables gives the percentage of the recorded audio material which is considered in the feature extraction (due to the fact that it is below the defined threshold based content selection). With an increasing (silence) threshold $thresh$ it is obvious that the percentage of the material considered (because the contents of the corresponding window is completely below $thresh$) increases (e.g. form 47.5% of the overall material for $coef$ =2048 and $thresh$ =0.0 to 67.6% for the same $coef$ and $thresh$ =0.025).

[71] The best reported detection performance in [Dohnal08] was $\kappa = 0.536$ (*accuracy* $= 60.26\%$) for a window size of 2048 samples and an amplitude silence-threshold of 0.25 (for a detailed discussion of the window size and silence-threshold parameters see [Buchholz09]). Even for that parameter combination, our best classification result is $\kappa = 0.862$ (an *accuracy* of 88.2%), while our optimal result is obtained with a silence threshold of 0.35 (indicating that the new approach is less context sensitive) with $\kappa = 0.924$ (*accuracy* of 93.5%).

Table 6.21: Detection performance (κ value) for $coef$ =2048 (experimental setup *Mic-BKD2009*; adapted from [Buchholz09])

Threshold $thresh$	**Percentage of samples**	*NaiveBayes*	*SMO*	*SimpleLogistic*	*J48*	*IB1*	*IBk* (2-nearest neighbour)
0.010	47.5%	0.257	0.473	0.474	0.364	0.455	0.438
0.025	67.6%	0.334	0.609	0.638	0.424	0.619	0.591
0.050	78.6%	0.367	0.724	0.736	0.538	0.704	0.663
0.100	86.8%	0.354	0.762	0.778	0.551	0.766	0.705
0.250	97.3%	0.381	0.862	0.862	0.634	0.798	0.769
0.350	99.7%	0.250	0.890	0.924	0.669	0.865	0.830
0.400	100.0%	0.259	0.861	0.908	0.697	0.868	0.833
0.500	100.0%	0.210	0.804	0.851	0.729	0.862	0.830
1.000	100.0%	0.210	0.804	0.851	0.729	0.862	0.830

Table 6.22: Detection performance (κ value) for $coef$ =256 (experimental setup *Mic-BKD2009*; adapted from [Buchholz09])

Threshold $thresh$	**Percentage of samples**	*NaiveBayes*	*SMO*	*SimpleLogistic*	*J48*	*IB1*	*IBk* (2-nearest neighbour)
0.010	64.6%	0.290	0.447	0.596	0.409	0.540	0.496
0.025	80.8%	0.339	0.575	0.720	0.523	0.707	0.671
0.050	87.5%	0.302	0.577	0.736	0.545	0.707	0.665
0.100	95.2%	0.293	0.677	0.804	0.556	0.712	0.676
0.250	99.6%	0.306	0.698	0.854	0.670	0.808	0.745
0.350	99.7%	0.292	0.701	0.868	0.629	0.802	0.741
0.400	100.0%	0.272	0.705	0.890	0.690	0.850	0.795
0.500	100.0%	0.218	0.652	0.816	0.701	0.846	0.811
1.000	100.0%	0.218	0.652	0.816	0.701	0.846	0.811

The first fact to be observed is that the number of samples for which not a single window falls within the amplitude threshold and which thus can only be classified by guessing is quite high even if the threshold $thresh$ used is set as low as 0.1 of the maximum amplitude – a value at which the audio signal definitely still contains a high portion of audible audio signal in addition to the noise. For all classifiers, the classification accuracy drops sharply when choosing even lower thresholds. This result is to be expected since with decreasing threshold, the number of audio samples without any acceptable windows at all increases sharply and thus the classification for more and more samples is based on guessing alone.

For most classifiers, the optimal classification results are obtained with a threshold that is very close to the lowest threshold at which features for all recordings in the test can be extracted (i.e. each recording has at last a single window that lies completely below the threshold). This, too, is reasonable. For a lower threshold, an increasing number of samples can only be classified by guessing. And for higher thresholds, the amount of signal in the FFT results increases and the amount of noise decreases. Since our classification in these observations is based on analysing the noise spectrum $N_{mic}(f)$ in equation 2.4, this leads to lower classification accuracy as well. However, the decline in accuracy even with a threshold of 1 (i.e. every single sample window is considered) is by far smaller than that of low thresholds.

The classification results for the two window tested sizes do not differ much. The *NaiveBayes* classifier yields better results when using smaller windows and thus fewer attributes. For all other classifiers, the results are usually better for the bigger window size, owing to the fact that a bigger number of attributes allows for the samples to differ in more ways.

The overall best classification results are obtained with the *SimpleLogistic* classifier, with $\kappa = 0.924$ (at $coef$ =2048) and $\kappa = 0.890$ (for $coef$ =256). However, for very high thresholds that allow a louder audio signal (as opposed to noise) to be part of the extracted features, the *IB1* classifier performs better than the *SimpleLogistic* one.

One notable odd behaviour is the fact that for very small thresholds the percentage of correctly classified samples exceeds the percentage of samples with valid windows, i.e. samples that can be classified. This is due to the behaviour of the classifiers to in essence guess the class for samples without valid attributes. Since this guess is likely to be correct with a probability of one seventh for seven microphones, the mentioned behaviour can indeed occur.

Inter-Microphone Differences

To analyse the differences in microphone detection performance between the individual microphones the detailed classification results for the test case with the most accurate results (*SimpleLogistic*, $coef$ =2048, threshold $thresh$ =0.35) are shown in a confusion matrix in table 6.23.
The results are rather unspectacular. The detection performance varies only slightly, between $\kappa = 0.879$ and 0.964 (number of correct classifications, i.e. detection performance, in table 6.23 between 0.861 and 0.960). Similar microphones from the same manufacturer (M_7 and M_8) even get mixed up less often as is the case with other microphone combinations. The only anomaly is the frequent misclassification of the M_9 as the M_2. This may be attributed to these two microphones sharing the same transducer technology, because otherwise, their purpose and signal quality differ considerably.

Table 6.23: The confusion matrix for the test case *SimpleLogistic*, $coef = 2048$, $thresh = 0.35$

	M_2	M_5	M_3	M_6	M_7	M_8	M_9
M_2	0.875	-0.320	-0.333	-0.333	-0.333	-0.333	-0.221
M_5	-0.320	0.933	-0.320	-0.320	-0.320	-0.320	-0.333
M_3	-0.292	-0.333	0.861	-0.263	-0.320	-0.333	-0.320
M_6	-0.333	-0.333	-0.320	0.960	-0.320	-0.333	-0.320
M_7	-0.320	-0.333	-0.320	-0.304	0.917	-0.320	-0.320
M_8	-0.333	-0.333	-0.333	-0.333	-0.320	0.931	-0.277
M_9	-0.305	-0.333	-0.305	-0.333	-0.320	-0.320	0.917

The results of this content selection approach can be directly compared to the results presented for single classifier classifications without content selection on the same test material (*RS2*) presented in table 6.24 below. Table 6.24 summarises the classification results from [Kraetzer09b] achieved with experimental setup *Mic-Kraetzer2009ACM-single-classifier* on *RS2* (see table 11.1 in appendix C (starting on page 201)). In these evaluations 200 feature vectors of 98 dimensions (AAFE v.1.0.3, see experimental setup *Mic-Kraetzer2009ACM-single-classifier*) are computed per reference file.

Table 6.24: Detection performance (κ values) of two selected classifiers on *RS2* (see experimental setup *Mic-Kraetzer2009ACM-single-classifier*) – adapted from [Kraetzer09b]

	R01	*R02*	*R03*	*R04*	*R05*	*R06*	*R07*	*R08*	*R09*	*R10*
SimpleLogistic avg. κ	0.704	0.659	0.771	0.736	0.698	0.780	0.707	0.742	0.741	0.735
J48 avg. κ	0.727	0.704	0.850	0.762	0.706	0.808	0.713	0.768	0.820	0.770

Table 6.24 shows for general classification on *RS2* using the *functions.SimpleLogistic* classifier a best Kappa value of $\kappa = 0.78$ (with an average of $\kappa = 0.727$). The content selection approach limiting its observations to silent parts shows for the two evaluated window sizes and the *functions.SimpleLogistic* classifier $\kappa = 0.924$ and $\kappa = 0.890$ respectively (see table 6.21 and table 6.22).
In this case the content selection based frequency-domain approach seems to perform better than the general approach.

A different picture can be seen if the performance of the *trees.J48* classifier is evaluated. Here, table 6.24 shows for the general classification on *RS2* a best detection performance of $\kappa = 0.820$ (with an average of $\kappa = 0.763$) while the observations restricted to parts show only $\kappa = 0.701$ and $\kappa = 0.729$ for the two evaluated window sizes (see table 6.21 and table 6.22).

To clearly answer the question whether the content selection based microphone forensics using only frequency-domain features introduced in [Buchholz09] technically outperforms the general-purpose approach considered within this thesis, three further sets of investigations have to be performed to consider the two different sources of influence to the classification: first, the impact of **content selection** in content dependent training, and second, the generalisation (**content dependency**) of models generated for one specific type of content in classification on another type of content. After these two content related questions are answered, as a third point the **performance of frequency domain features** (which are the only type of features considered in [Buchholz09]) in comparison to a mixed set of time-, frequency- and cepstral-domain features is investigated.

For the first of those three required evaluations, two experiments are run to establish empirical knowledge about the **content influence on the classification**. In the experiments *Mic-Content-Selectivity-01* and *Mic-Content-Selectivity-02* (see table 11.1 in appendix C (starting on page 201)) for material from *RS4_Rode* and *RS4_Beyer* recorded in room *R01* the classification performance on silence and speech references are separately measured using three classifiers from the top 20 classifiers (see section 6.1.3, from each of the most significant classifier classes one representative candidate is chosen). Here, the performances of content-adapted classifier models on similar content are evaluated. The tests are run in 10-fold stratified cross-validation on 200 and 800 feature vectors per reference and using the 20 most significant features only (see section 6.1.4).

The results achieved with experimental setup *Mic-Content-Selectivity-01* on *RS4_Rode* in *R01* (see table 6.24) show, for the silence reference, very similar results for all three classifiers and the tested vector field sizes. All those κ values are in the range $[0.669,\ 0727]$. On the speech material the classifications show a tendency towards better classification results for larger vector fields. The most significant observation in this experiment is nevertheless the rather dramatic difference between the two content classes. While the speech results are barely above the probability of guessing correctly (i.e. $\kappa = 0$), the silence results are much better.

Table 6.25: Detection performance (κ value) for silence and speech signals recorded with *RS4_Rode* in *R01*, best 20 features, three selected classifiers (out of the top 20, all using 10-fold stratified cross-validation; experimental setup *Mic-Content-Selectivity-01*)

Content	**Feature vectors per file**	**Feature set**	**Classifier**	**κ value**
Silence	200	best 20	*functions.Logistic*	0.727
	800	best 20	*functions.Logistic*	0.700
	200	best 20	*meta.dagging*	0.698
	800	best 20	*meta.dagging*	0.669
	200	best 20	*trees.RandomForest*	0.703
	800	best 20	*trees.RandomForest*	0.704
Speech	200	best 20	*functions.Logistic*	0.016
	800	best 20	*functions.Logistic*	0.032
	200	best 20	*meta.dagging*	0.026
	800	best 20	*meta.dagging*	0.043
	200	best 20	*trees.RandomForest*	-0.002
	800	best 20	*trees.RandomForest*	0.057

When the experiment described above is repeated on material from *RS4_Beyer* in *R01* similar tendencies are shown in table 6.26. The rather dramatic difference between the two content classes remains.

Table 6.26: Detection performance (κ value) for silence and speech signals recorded with *RS4_Beyer* in *R01*, best 20 features, three selected classifiers (out of the top 20, all using 10-fold stratified cross-validation; experimental setup *Mic-Content-Selectivity-01*)

Content	**Feature vectors per file**	**Feature set**	**Classifier**	**κ value**
Silence	200	best 20	*functions.Logistic*	0.618
	800	best 20	*functions.Logistic*	0.610
	200	best 20	*meta.dagging*	0.558
	800	best 20	*meta.dagging*	0.567
	200	best 20	*trees.RandomForest*	0.573
	800	best 20	*trees.RandomForest*	0.613
Speech	200	best 20	*functions.Logistic*	0.115
	800	best 20	*functions.Logistic*	0.131
	200	best 20	*meta.dagging*	0.082
	800	best 20	*meta.dagging*	0.082
	200	best 20	*trees.RandomForest*	0.187
	800	best 20	*trees.RandomForest*	0.217

The results presented for the content selection can be summarised therefore as follows: the content seems to have a strong influence on the detection performances achieved, even when using classifier models trained on the same type of content. Therefore, a content identification in pre-processing and a corresponding model selection might strongly improve the performance (in terms of accuracy) of a microphone identification scheme.

For the second of those three required evaluations also two experiments are run to establish **empirical knowledge about the content dependency between training and testing**. In the experiments *Mic-Content-Independency-01* and *Mic-Content-Independency-02* (see table 11.1 in appendix C (starting on page 201)) for material from *RS4_Rode* and *RS4_Beyer* recorded in room *R01* the classification performance of models generated on one specific content class (silence) in classification of completely different content (speech), and vice versa, are evaluated. For reasons of comparability the same three classifiers and the same features are used as for *Mic-Content-Selectivity-01* and *Mic-Content-Selectivity-02*. The tests are run in with selected training sets on 200 and 800 feature vectors per reference.

The results achieved on *RS4_Rode* in *R01* (see table 6.27) show for the content independent training and testing devastating results: nearly all achieved detection performances are as low as the probability of guessing (i.e. $\kappa = 0$). Only the result for the *Trees.RandomForest* classifier trained on speech and tested on silence achieves a detection performance which is with $\kappa = 0.298$ significantly better than guessing. A verification of the results with 800 feature vectors per reference for both recording sets also returned very similar results than in the case of 200 feature vectors per reference file.

Table 6.27: Detection performance (κ value) for *RS4_Rode* in *R01*, best 20 features, three selected classifiers (out of the top 20, specified test sets) 200 feature vectors per file; experimental setup *Mic-Content-Selectivity-01*

Training material	**Test material**	**Feature set**	**Classifier**	**κ value**
Silence	Speech	best 20	*functions.Logistic*	0.034
		best 20	*meta.dagging*	0.033
		best 20	*trees.RandomForest*	0.028
Speech	Silence	best 20	*functions.Logistic*	0.000
		best 20	*functions.Logistic*	0.000
		best 20	*meta.dagging*	0.298

When the experiment described above is repeated on material from *RS4_Beyer* in *R01* similar tendencies are shown in table 6.28. A verification of the results with 800 feature vectors per reference for both recording sets also returned very similar results than in the case of 200 feature vectors per reference file.

Table 6.28: Detection performance (κ value) for *RS4_Beyer* in *R01*, best 20 features, three selected classifiers (out of the top 20, specified test sets) 200 feature vectors per file; experimental setup *Mic-Content-Selectivity-01*

Training material	**Test material**	**Feature set**	**Classifier**	**κ value**
Silence	Speech	best 20	*functions.Logistic*	0.063
		best 20	*meta.dagging*	0.052
		best 20	*trees.RandomForest*	0.045
Speech	Silence	best 20	*functions.Logistic*	0.007
		best 20	*functions.Logistic*	0.000
		best 20	*meta.dagging*	0.180

These results can be summarised therefore as follows: if the content used for training and testing is completely different in its characteristics (completely independent content classes are used), then the classification model will result in very low detection performances. The same conclusion for the training material selection can be drawn as for the content selectivity considerations above: a content identification in pre-processing and a corresponding model selection might strongly improve the performance (in terms of κ values) of a microphone identification scheme.

The third question to be evaluated for the approach from [Buchholz09] is the **choice of features** (frequency-domain features versus mixed domains feature sets). Here, just a brief summary of the results of the feature selection for microphone forensics performed in section 6.1.4 is given: from the 590 features in the segmental feature vector computed by AAFE v.2.0.5 530 are computed in frequency-domain, including a 512 Fourier coefficient spectrogram similar to those used as feature set in [Buchholz09], but in the top 30 of these 590 features only two frequency-domain features are found. This mismatch in those numbers speaks strongly against a paramount performance of frequency-domain features.

Résumé for this section: The investigations on **content selection** in content dependent training and content independent training and testing performed within this section show that the introduced approach might benefit from a content identification in pre-processing and a corresponding model selection. The implementation of such a pre-processing scheme, which involves the solution of another audio pattern recognition task – reliable content classification – is outside of the scope of this thesis and reserved for future research.

Notwithstanding the quite good detection performances achieved with the context sensitive approach from [Buchholz09] and the additional observations on the impact different classes of content (here speech vs. silence), the idea of performing content removal (e.g. by silence detection) is discarded for the remaining investigations in this thesis. Technically such a content selection (e.g. by silence thresholding) might improve the detection performance achieved, but the restriction imposed thereby on the audio signals that can undergo such microphone forensics is considered to be too severe to be useful in field applications. With these restrictions, the introduced general-purpose approach would lose the feature which distinguished it from the majority of the approaches in the current state-of-the-art: its general applicability (see the description of the shortcomings of the current state-of-the-art in section 2.6.2. I.e. for authentication purposes such a silence thresholding would prevent the processing of recordings were no such 'silence' would be present in sufficient quantities. For composition detection purposes in integrity verification, a context removal seems even more contra productive because normally loud parts are 'mixed in' to change the meaning of an audio data stream and not silence. Therefore, further investigations on content removal are abandoned for the rest of this thesis and reserved for future work.

6.3 Persistence of the microphone pattern under selected post-processing operations and playback recording

As mentioned in the design considerations on the required evaluations in microphone forensics in section 4.3.1 selected, relevant signal modifications as common post-recording influences to the recorded signals are evaluated here. Further investigations summarised in this section address the question whether the microphone response function based traces used here for microphone forensics survive a playback recording.

6.3.1 Normalisation

Normalisation is a fairly common audio signal post-processing operation. Since it is an amplitude scaling operation in time domain, normalisation is considered here to be a representative for all such operations. The models used in the evaluations performed here are trained on original recordings and then used to classify material that underwent normalisation (with the normalisation factor computed independently for each file). The experimental setup *Mic-Normalisation-RS4_Beyer* describes those tests (see table 11.1 in appendix C (starting on page 201)). A second set of practical evaluations is performed using the same procedure and the *RS4_Rode* recording set (see *Mic-Normalisation-RS4_Rode*).

For the experiment *Mic-Normalisation-RS4_Beyer*, from each recorded and normalised file the first and second set of 200 feature vectors are computed and used as two distinct test sets. Table 6.29 compares the detection performances achieved on these test sets against the accuracy achieved in 10-fold cross-validation on the first set of 200 feature vectors per file on original data.

Table 6.29: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and normalised audio data for *RS4_Beyer* in *R01* (experimental setup *Mic-Normalisation-RS4_Beyer*)

Classifier	**κ on original data**	**κ on normalised data (1st test)**	**κ on normalised data (2nd test)**
bayes.NaiveBayes	0.175	0.061	0.090
functions.SMO	0.680	0.576	0.562
meta.RandomCommittee	0.637	0.329	0.279
trees.RandomForest	0.615	0.332	0.285

The results in table 6.29 show a rather strong impact of the normalisation operation to the achieved detection performances. The result remains for all four exemplary tested classifiers above the probability of guessing (i.e. $\kappa > 0$), but for the *NaiveBayes* and the *SMO* the accuracy drops slightly and for the *RandomCommittee* and *RandomForest* rather strongly.
To verify those results the same test is run again on the recordings of the four Rode microphones in *RS4_Rode* (see *Mic-Normalisation-RS4_Rode*).

Table 6.30: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and normalised audio data for *RS4_Rode* in *R01* (experimental setup *Mic-Normalisation-RS4_Rode*)

Classifier	**κ on original data**	**κ on normalised data (1st test)**	**κ on normalised data (2nd test)**
bayes.NaiveBayes	0.201	0.058	0.070
functions.SMO	0.583	0.490	0.513
meta.RandomCommittee	0.515	0.245	0.250
trees.RandomForest	0.493	0.204	0.229

The results in table 6.30 show a similar impact of the normalisation of the test material as in table 6.29 on the classification performance achieved in this microphone forensics classification. Like for the *RS4_Beyer* recordings, the detection performances achieved after normalisation are much lower than the results achieved by these four exemplary selected classifiers on original material.

Résumé for this section: In the classification using models trained on original recordings against test material which underwent normalisation, a negative impact of this post-processing procedure on the detection performance can be observed. Nevertheless, the classification with the exemplary chosen classifiers after this time-domain modification still leads to detection performances which are significantly better than the probability of guessing correctly. This implies that frequency-domain and cepstral-domain features, which are rather unaffected by time-domain normalisation, contribute strongly to the classification behaviour, an observation which is consistent with the observation on the best performing features in microphone forensics (see section 6.1.4).

6.3.2 MP3 conversion

The MP3 conversion is one of the most widely used audio signal post-processing operations. Here, it is applied with a common bit rate of 128kBit/s (using the LAME codec[72]) to show the impact of this data reduction to the classification performance achieved in microphone forensics. The models used in those evaluations are trained on original (never-compressed) recordings and then used to classify material that underwent MP3 conversion. The experimental setup *Mic-MP3conversion-RS4_Beyer* describes those tests (see table 11.1 in appendix C (starting on page 201)). For verification of the results, a second set of practical evaluations is performed using the same procedure and the *RS4_Rode* recording set (see *Mic-MP3conversion-RS4_Rode*).

[72] http://lame.sourceforge.net/

For the experiments from each recorded and MP3 encoded file the first and second set of 200 feature vectors are computed and used as two distinct test sets. Table 6.31 compares the classification accuracies achieved on these test sets against the accuracy achieved in 10-fold cross-validation on the first set of 200 feature vectors per file on original data.

Table 6.31: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and MP3 encoded audio data for *RS4_Beyer* in *R01* (experimental setup *Mic-MP3conversion-RS4_Beyer*)

Classifier	**κ on original data**	**κ on MP3 data (1st test)**	**κ on MP3 data (2nd test)**
bayes.NaiveBayes	0.175	0.165	0.209
functions.SMO	0.680	0.665	0.659
meta.RandomCommittee	0.637	0.679	0.615
trees.RandomForest	0.615	0.651	0.585

The results in table 6.31 show no negative impact of the MP3 conversion of the test material on the classification performance achieved in this microphone forensics classification. The detection performance achieved after MP3 conversion is very close to the result achieved by these four exemplary selected classifiers on original material.
To verify those results the same test is run again on the recordings of the four Rode microphones in *RS4_Rode* (see *Mic-MP3conversion-RS4_Rode*).

Table 6.32: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and MP3 encoded audio data for *RS4_Rode* in *R01* (experimental setup *Mic-MP3conversion-RS4_Rode*)

Classifier	**κ on original data**	**κ on MP3 data (1st test)**	**κ on MP3 data (2nd test)**
bayes.NaiveBayes	0.201	0.187	0.219
functions.SMO	0.583	0.567	0.606
meta.RandomCommittee	0.515	0.546	0.553
trees.RandomForest	0.493	0.552	0.541

The results in table 6.32 show also no negative impact of the MP3 conversion of the test material on the classification performance achieved in this microphone forensics classification. Like for the *RS4_Beyer* recordings, the detection performance achieved after MP3 conversion is very close to the result achieved by these four exemplary selected classifiers on original material.

Résumé for this section: In the investigations using models trained on original recordings against test material which underwent MP3 conversion, no negative impact of this encoding can be observed. The conversion seems to have no impact on the patterns which are used in this statistical pattern recognition (SPR) based approach for microphone forensics. This observation is consistent with the observation on the best performing features in microphone-forensics, which seem to be (see section 6.1.4) higher-order cepstral-domain features, which are rather unlikely to be changed by MP3 conversion.

6.3.3 De-noising by re-quantisation

With the impact of de-noising (here by re-quantisation) another common signal modification for recorded audio signals is evaluated for its impact on the classification performance achieved by the introduced statistical pattern recognition (SPR) based microphone forensics approach. Like for the impact observations on normalisation and MP3 conversion, the models used in those evaluations are trained on original recordings and are then used to classify material that underwent targeted signal modifications (here de-noising by re-quantisation). The experimental setup *Mic-Denoise-RS4_Beyer* (see table 11.1 in appendix C (starting on page 201)) describes those tests. A second set of practical evaluations is performed using the same procedure and the *RS4_Rode* recording set (see *Mic-Denoise-RS4_Rode*).

For the experiments from each recorded and re-quantised file the first and second set of 200 feature vectors are computed and used as two distinct test sets. Table 6.33 compares the classification accuracies achieved on these test sets against the accuracy achieved in 10-fold stratified cross-validation on the first set of 200 feature vectors per file on original data.

Table 6.33: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and re-quantised data for *RS4_Beyer* in *R01* (exp. setup *Mic-Denoise-RS4_Beyer*)

Classifier	**κ on original data**	**κ on test data (1st test)**	**κ on test data (2nd test)**
bayes.NaiveBayes	0.175	0.192	0.232
functions.SMO	0.680	0.195	0.209
meta.RandomCommittee	0.637	0.518	0.472
trees.RandomForest	0.615	0.436	0.397

The results in table 6.33 vary rather strongly for the four exemplary chosen classifiers. For the *NaiveBayes* no negative impact by this signal modification is observed. Instead its relatively low detection performance seems to be improved by the re-quantisation. For the other three classifiers a strong decline of the detection performance is observed.

For verification purposes the same test is run again on the recordings of the four Rode microphones in *RS4_Rode* (see *Mic-Denoise-RS4_Rode*).

Table 6.34: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and re-quantised data for *RS4_Rode* in *R01* (exp. setup *Mic-Denoise-RS4_Rode*)

Classifier	**κ on original data**	**κ on test data (1st test)**	**κ on test data (2nd test)**
bayes.NaiveBayes	0.201	0.216	0.237
functions.SMO	0.583	0.303	0.351
meta.RandomCommittee	0.515	0.365	0.389
trees.RandomForest	0.493	0.339	0.378

If the results presented in table 6.34 are compared to those in table 6.33 the same overall trend can be observed: for the Bayesian classifier no negative impact can be seen while the degradation in classification accuracies for the other three exemplary chosen classifiers is strongly visible.

Résumé for this section: In the classification, using models trained on original recordings against test material which underwent de-noising by re-quantisation, the impact appears to be classifier dependent. For the four exemplary chosen classifiers, one did show no negative impact by this signal modification, for the other three a strong degradation of the achieved classification accuracy is noticeable. It can be assumed that the latter behaviour is more characteristic for this modification. This would be consistent with the observation on the best performing features in microphone forensics, which seem to be (see section 6.1.4) higher-order cepstral-domain features. The de-noising by re-quantisation has a rather strong impact on the spectrogram of the audio signal and this influence also influences those higher-order cepstral-domain features derived from the frequency-domain representation.

6.3.4 An exemplary combination of signal processing operations

After the impact of the single signal processing operations is investigated in the previous three subsections, here an exemplary selected combination of these processing operations and its impact is evaluated.

The models used in those evaluations are trained on original recordings and then used to classify material that underwent first de-noising, then normalisation and third MP3 conversion. The experimental setup *Mic-MultiProcessing-RS4_Beyer* (see table 11.1 in appendix C (starting on page 201)) describes those tests. For verification purposes, a second set of practical evaluations is performed using the same procedure and the *RS4_Rode* recording set (see *Mic-MultiProcessing-RS4_Rode*).

For the experiments from each recorded and re-quantised, normalised and MP3 encoded file the first and second set of 200 feature vectors are computed and used as two distinct test sets. Table 6.35

compares the classification accuracies achieved on these test sets against the accuracy achieved in 10-fold cross-validation on the first set of 200 feature vectors per file on original data.

Table 6.35: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and re-quantised, normalised and MP3 encoded data for *RS4_Beyer* in *R01* (exp. setup *Mic-MultiProcessing-RS4_Beyer*)

Classifier	**κ on original data**	**κ on test data (1st test)**	**κ on test data (2nd test)**
bayes.NaiveBayes	0.175	0.023	0.045
functions.SMO	0.680	0.379	0.446
meta.RandomCommittee	0.637	0.165	0.195
trees.RandomForest	0.615	0.155	0.198

The results in table 6.35 show, in comparison to the results presented for single processing operations, that for most cases the influence of the combined signal processing (de-noising, normalisation and MP3 conversion) imposes a stronger disturbance to the microphone-intrinsic recording-pattern. The only exception is the *SMO* classifier where the detection performance after the combined signal processing is better than after a de-noising only (compare *SMO* results in table 6.33 and table 6.35).
For verification purposes the same test is run again on the recordings of the four Rode microphones in *RS4_Rode* (see *Mic-MultiProcessing-RS4_Rode*).

Table 6.36: Detection performance (κ value) for four exemplary selected classifiers and models generated on original material run against original data (in 10-fold cross-validation) and re-quantised, normalised and MP3 encoded data for *RS4_Rode* in *R01* (exp. setup *Mic-MultiProcessing-RS4_Rode*)

Classifier	**κ on original data**	**κ on test data (1st test)**	**κ on test data (2nd test)**
bayes.NaiveBayes	0.201	0.039	0.044
functions.SMO	0.583	0.357	0.412
meta.RandomCommittee	0.515	0.119	0.140
trees.RandomForest	0.493	0.108	0.138

If the results presented in table 6.36 are compared to those in table 6.35, the same overall trend can be observed: the combined signal processing in the majority of the cases disturbs the microphone intrinsic recording pattern classified here stronger than as in case of single signal processing operations.

Résumé for this section: Even after the rather strong impact of the combined signal processing, the detection performance remains for three of the four tested classifiers (the *SMO*, *RandomCommittee* and *RandomForest*) significantly better than the probability of guessing correctly (i.e. $\kappa = 0$).
Further investigations on signal post-processing influences are strongly recommended for future research to establish a close estimation of the robustness of the microphone-intrinsic recording-pattern used in the introduced approach for the microphone authentication and thereby of the plausibility of the whole approach.

6.3.5 Playback recording

In [Kraetzer12b] the influence of playback recording on the introduced microphone forensics approach is investigated (see table 11.1 in appendix C (starting on page 201); experimental setup *Mic-SPIE2012-Double-Recording*). Summarising this paper, it has to be stated first that it presents a very high detection performance for microphone classification in **single recording evaluations** on the used recording set *RS16*. For the recordings with six different microphones (two of them being of the same brand and type) the detection performance on recorded speech is in average about $\kappa = 0.989$ (equivalent to an accuracy of 99.14% or an error rate of 0.86%).
This result is much better that the performance shown in our previous papers (see e.g. table 6.13 in section 6.1.3 where a set of four identical microphones achieves a best value of $\kappa = 0.780$ in classifier selection). The very good detection performance is assumed here to be due to the rather strong recording characteristics of the microphones used in *RS16* (mostly headsets and low-quality devices), the perfect recording conditions (*R06*; a sound-proof, anechoic chamber) and the limitation of the recorded

content to human speech (instead of multi-genre audio). If normalisation is applied to the recorded content as audio signal post-processing (prior to training and testing), the detection performance decreases to $\kappa = 0.981$ (an average error of 1.44%) – confirming the fact established in section 6.3.1 that normalisation influences the introduced statistical pattern recognition (SPR) based microphone forensics approach.

For the **investigations of playbacks**, the results presented in [Kraetzer12b] imply that the traces left by both microphones involved in the recording process are still to some extend detectable for our approach. If the **same microphone** is used for the initial and playback recording, the achieved detection performance for the six microphones set decreases to $\kappa = 0.841$ (equivalent to an error rate of 11.91%). If the audio material is normalised between initial and secondary recording, the detection performance drops to $\kappa = 0.677$ (an error rate of about 24.24%). Here, the misclassifications are more or less randomly distributed. The playback recording with the same microphone seems to somehow decrease the clarity of the microphone pattern as perceived by using our audio feature set.

Regarding the investigations on the audio material that is initially recorded by one microphone, played back via loudspeaker and is then re-recorded with a **different microphone**, we see an even stronger influence of normalisation: Without normalisation the results show a strong tendency (267 out of 300 test cases; 89.00% – see [Kraetzer12b]) for a positive indication on the correct two microphones. Only 11 cases (3.67%) give two microphones of which one was actually not involved and only 22 cases (7.33%) make completely wrong assignments. With the normalisation involved, only 108 out of 300 test cases (36.00%) for this test set show a positive indication on the correct two microphones. With 152 cases (50.67%) the classifiers indicate two microphones, with one of them actually not involved and only 40 cases (13.33%) make completely wrong assignments. Based on these results, it has to be assumed that the normalisation (with a different factor for each file) makes the classifier model more complex, i.e. decreases its detection performance.

Résumé for this section: The results presented for the investigations on a test set of six microphones recording human speech (and its playback) imply that playback detection by the introduced microphone forensics approach is possible. Without post-processing of the recorded sound prior to playback, in the tests 89.00% positive indications are achieved on the two correct microphones. If post-processing is applied in the form of normalisation, this percentage significantly drops to 36.00% while another 50.67% of the tests indicate two microphones, of which one has actually not been involved in the recording and playback recording process.

For the overall application scenario of microphone forensics, these results imply that a playback attack has a strong influence on the detection performance, if it is not detected prior to the performed microphone authentication. Therefore, future work should be invested into an updated design for a microphone forensics scheme which incorporates context analysis like playback detection and content class (i.e. speech, music, etc.) analysis prior to the actual source authentication operations.

6.4 Investigations on composition detection

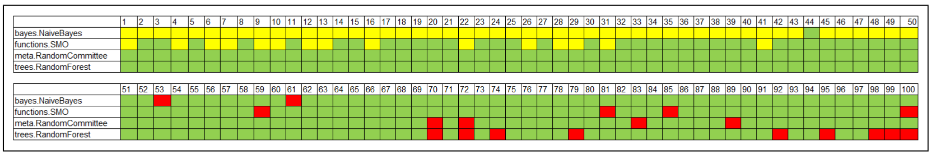
The evaluation design for microphone forensics in section 4.3.1 addresses the question of composition or mesh-up detection. Generally, two distinct types of scenarios are here identified in this context: a) the audio data stream, into which other data is pasted into, originates from a known microphone and b) the audio data stream, into which other data is pasted into, originates from an unknown microphone. The first scenario is the more likely one in microphone forensics, where we usually assume that we intend to verify the identity of a source microphone. Nevertheless, the performance of the statistical pattern recognition (SPR) based forensics approach used within this thesis on the less likely second scenario is also evaluated here to show its limitations.
Four different tests are performed in this composition detection evaluation:

- Microphone recordings of one known microphone made in different locations composed into one stream
- One known microphone pasted into a stream of completely different known microphone
- One unknown microphone pasted into a stream of completely different known microphone
- One unknown microphone pasted into a stream of completely different unknown microphone

The evaluations performed here extend our initial considerations on playback recording influences in microphone forensics as published in [Kraetzer11].

Test 1: Microphone recordings of one known microphone made in different locations composed into one stream

This experiment is implemented by experimental setup *Mic-Composition-1* (see table 11.1 in appendix C (starting on page 201)). In terms of the classification problem at hand, this test is assumed to be the hardest problem in this microphone forensics approach. As shown in section 6.2.1 the room / recording environment has a strong influence on the classification performance, nevertheless the microphone used for the 'patched-in' material is the same as the one for the original recordings. Figure 6.2 shows the results for this experiment (*Mic-Composition-1*) and the four exemplary selected classifiers.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
bayes.NaiveBayes	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟨	🟨	🟨
functions.SMO	🟨	🟩	🟩	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟨	🟩	🟨	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
meta.RandomCommittee	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
trees.RandomForest	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
bayes.NaiveBayes	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
functions.SMO	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
meta.RandomCommittee	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
trees.RandomForest	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟥

Figure 6.2: *Mic-Composition-1* test results (upper half original, lower half pasted in material)

The colour-coding in figure 6.2 has the following meaning: A green field in the upper half of the figure (the 'original' half) denotes a true positive (TP) in the classification of this feature vector, i.e. the feature vector is classified correctly as belonging to the microphone M_{22}. A yellow field in the upper half of the figure corresponds to a false negative (FN), i.e. the feature vector is classified wrongly as belonging to M_{20}, M_{21} or M_{23}.

A green field in the lower half of the figure (the 'patched-in' or 'impostor' half) denotes a true negative (TN) in the classification of this feature vector, i.e. the feature vector is rejected correctly as not belonging to the microphone M_{22} (as it is registered in the model for the room / recording environment *R01*). A red field in the lower half of the figure denotes a false positive (FP), i.e. the feature vector is classified wrongly as belonging to the microphone M_{22}.

The intention of this colour coding (which is consistently used for all three evaluations where the audio data stream, into which other data is pasted into, originates from a known microphone) is to mark all true classifications (TP and TN) in green and the false classifications in yellow (FN) and red (FP).

As can be seen in figure 6.2 and in table 6.37 the performance of the used classifiers strongly differs. The *NaiveBayes* classifier produces an extremely high (98% on the original half) false negative rate on the original half of the test material, while its FP rate on the impostor part is with only 4% very good. The *SMO* shows in comparison to the Bayesian classifier a better, but still imperfect, FN rate. The *RandomCommittee* and *RandomForest* classifiers achieve in this test a perfect classification behaviour on the original half (100% TP) with FP-rates of 8% and 18% respectably on the impostor part. The detection performance achieved in the overall evaluations (see table 6.37) could be used very well to rank the classifiers according to their performance.

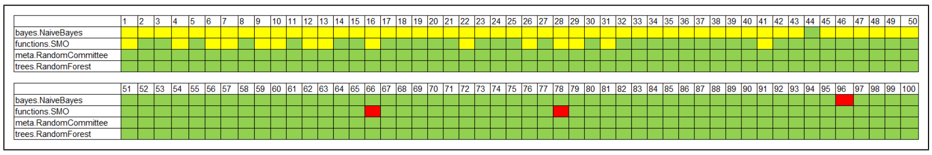
Table 6.37: Detection performance (κ value) and statistical errors for the exemplary classifications on *Mic-Composition-1*

Classifier	κ value	TP	FN	TN	FP
bayes.NaiveBayes	-0.02	2%	98%	96%	4%
functions.SMO	0.62	70%	30%	92%	8%
meta.RandomCommittee	0.92	100%	0%	92%	8%
trees.RandomForest	0.82	100%	0%	82%	18%

Test 2: One known microphone pasted into a stream of completely different known microphone

This experiment is implemented by experimental setup *Mic-Composition-2* (see table 11.1 in appendix C (starting on page 201)). The setup of this second test on the 'mesh-up' detection seems to be quite unlikely, for it assumes that the microphone which recorded the material to be inserted into an audio data stream is also registered in the classification models, something that an attacker/manipulator would try to avoid. Nevertheless, this test is performed to evaluate the detection performance of the approach under this assumption.

Figure 6.3 shows the results for the experiment *Mic-Composition-2* and the four exemplary selected classifiers. The same colour coding scheme is applied as in figure 6.2 above. Therefore all true classifications (TP and TN) are marked in green and the false classifications in yellow (FN) and red (FP).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
bayes.NaiveBayes	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟨	🟨	🟨
functions.SMO	🟨	🟩	🟩	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟨	🟩	🟨	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
meta.RandomCommittee	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
trees.RandomForest	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
bayes.NaiveBayes	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩
functions.SMO	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
meta.RandomCommittee	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
trees.RandomForest	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 6.3: *Mic-Composition-2* test results (upper half original, lower half pasted in material)

As can be seen in figure 6.3 and in table 6.38, the performance of the used classifiers on the original part shows exactly the same performance as in figure 6.2 above. This is due to the fact that the same models are used here for the classification of the same material – all four classifiers work in a deterministic way.

Differences can be seen in the performance on the impostor part (the lower half in figure 6.3; material from M_{23} claimed to originate from M_{22}). Here, the performance is strongly increased, as had to be expected for this rather unlikely scenario. The *RandomCommittee* and *RandomForest* classifiers achieve perfect classification performance on original as well as impostor material while the *SMO* classifier returns significant but less than optimal results. The *NaiveBayes* classifier achieves a detection performance of $\kappa = 0$ in this two-class evaluation.

Table 6.38: Detection performance (κ value) and statistical errors for the exemplary classifications on *Mic-Composition-2*

Classifier	κ value	TP	FN	TN	FP
bayes.NaiveBayes	0.00	2%	98%	98%	2%
functions.SMO	0.66	70%	30%	96%	4%
meta.RandomCommittee	1.00	100%	0%	100%	0%
trees.RandomForest	1.00	100%	0%	100%	0%

Test 3: One unknown microphone pasted into a stream of completely different known microphone

This experiment is implemented by experimental setup *Mic-Composition-3* (see table 11.1 in appendix C (starting on page 201)). The setup for this evaluation would be the rather most likely in recording authentication: material originating from an unknown source is pasted into an audio data stream generated by a registered microphone.

Figure 6.4 shows the results for the experiment *Mic-Composition-3* and the four exemplary selected classifiers. The same colour coding scheme is applied as in figure 6.2 above. Therefore all true classifications (TP and TN) are marked in green and the false classifications in yellow (FN) and red (FP).

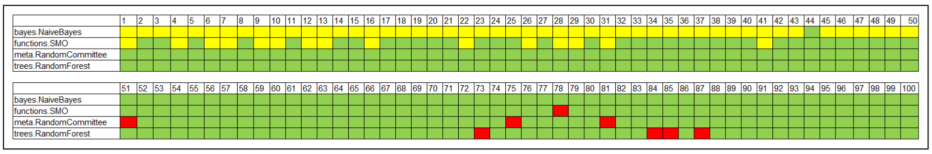


Figure 6.4: *Mic-Composition-3* test results (upper half original, lower half pasted in material)

The performance in the original part (see figure 6.4 upper half and table 6.39) is exactly the same as for the previous tests. For the impostor part (the lower half in figure dfghs; material from M_8 claimed to originate from M_{22}) a good to very good detection performance is achieved by all four classifiers. Nevertheless, the Bayesian classifier achieves only a κ value of 0.02 which would disqualify this classifier from practical application in composition detection.

Table 6.39: Detection performance (κ value) and statistical errors for the exemplary classifications on *Mic-Composition-3*

Classifier	κ value	TP	FN	TN	FP
bayes.NaiveBayes	0.02	2%	98%	100%	0%
functions.SMO	0.68	70%	30%	98%	2%
meta.RandomCommittee	0.94	100%	0%	94%	6%
trees.RandomForest	0.92	100%	0%	92%	8%

Test 4: One unknown microphone pasted into a stream of completely different unknown microphone

This experiment is implemented by experimental setup *Mic-Composition-4* (see table 11.1 in appendix C (starting on page 201)). Like the setup of the second test on the 'mesh-up' detection, this setup seems to be rather unlikely; nevertheless, this test is performed to evaluate the performance of the approach. Here it is assumed that material should be verified for mesh-ups for which the sensor is not registered. This situation would be avoided by a person performing sensor forensics – in this field it is generally assumed that the sensor to be authenticated is available to the examiner.

Figure 6.5 shows the results for the experiment *Mic-Composition-4* and the four exemplary selected classifiers. Here, a different colour coding has to be applied than in the previous mesh-up tests. All four microphones in the used classification model are assigned one colour (M_{20} = blue, M_{21}=orange, M_{22}=magenta, and M_{23}=cyan) classification result for each of the 100 frames in the test material is marked in this colour coding.

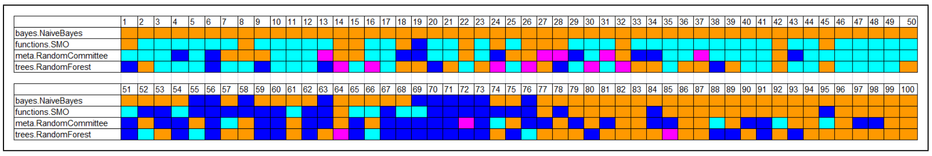


Figure 6.5: *Mic-Composition-4* test results (upper half original, lower half pasted in material)

In figure 6.5 the upper half is representing the first unknown or impostor microphone (M_2 from *RS2* room *R01*) and the second half is the other one (M_3 from *RS2* room *R01*). Since the classification model used was trained on different recording material from completely different microphones here in this evaluation the stability of the decisions can be used to evaluate the performance of the microphone forensics approach. The higher the change rate and the shorter the average sequence length in the classifications, the better the classification under these circumstances.

For the upper half the *NaiveBayes* classifier shows here a very bad performance. All frames are insistently classified as belonging to M_{21}. Here a correct classification was of course not possible since the correct microphone was not available in the model but this insistence is implying a wrong certainty of the classifier. This wrong sense of certainty can be eliminated by taking the average classification detection performance of the used classifier into account, which for the *NaiveBayes* is about $\kappa = 0.4$ in the tests performed in section 6.1.3.

The *RandomCommittee* and *RandomForest* classifiers show a much better performance in these two tests presented here. They show in table 6.40 a change rate of more than 25 out of 50 with an average sequence length of smaller than two consecutive frames. These values are much closer to maximum entropy than the values for the *SMO* or *NaiveBayes* classifiers, which implies that these (known good classifiers) are run on impostor material.

Table 6.40: Change rate and average sequence length for the experiments in *Mic-Composition-4*

	Classifier	change rate	average sequence length
first half (M2)	*bayes.NaiveBayes*	0	50
	functions.SMO	20	2.380952381
	meta.RandomCommittee	26	1.851851852
	trees.RandomForest	36	1.351351351
second half (M3)	*bayes.NaiveBayes*	10	4.545454545
	functions.SMO	17	2.777777778
	meta.RandomCommittee	34	1.428571429
	trees.RandomForest	28	1.724137931

Résumé for this section: In the evaluations performed here on composition detection, it is shown for strongly limited setups that the mixing of audio recordings into another recorded audio signal can be very well detected by some classifiers. Of interest is the fact that from the four exemplary chosen classifiers for the evaluations here the *SMO*, which is performing quite well in all other microphone forensics evaluations (see e.g. section 6.1.3) shows a dissatisfactory performance. The *RandomCommittee* and *RandomForest* classifiers do show here a very good performance if the microphone used in the generation of the audio data stream into which other data is pasted into is registered in the classification models. In fact, it seems not to matter much for their performance whether the material pasted into the original stream originates from a registered or unknown microphone.
In case none of the two sources for a mesh-up is registered, it is shown here for a small example that the change rate and average sequence length can be used to tell that a wrong model is used and, since a different tendency for classification of the individual feature vectors can be observed, that a composition is likely. Nevertheless, these facts should be subjected to further research to substantiate these findings.

6.5 Summary of the findings for microphone forensics

In section 3.3 the tasks for the practical investigations performed within this thesis are defined. In this summarising section, the results for the microphone forensics application scenario are first projected onto these investigation tasks. In the second step performed here, the results achieved are reflected under consideration of the evaluation criteria for forensic investigations derived within this thesis from the Daubert standard (see section 2.2 and its subsections).

6.5.1 Projection of the results onto the defined investigation tasks

The first step required in the investigations is to establish some empirical ground truth (**investigation task A**, as an précising statement for research objective 1 – see section 3.3) to show that the application scenario of microphone forensics (as it is considered within this thesis) can actually be solved by statistical pattern recognition (SPR).
The fact that microphones can be classified with this SPR-based approach and with an detection performance much better than the probability of guessing correctly was first demonstrated by us in [Kraetzer07c] for inhomogeneous recording sets and verified here with intra-class classifications on sets of identical microphones in section 6.1.1. The best results achieved have show a detection performance of κ close to 1 on strongly inhomogeneous microphone sets of low quality microphones ($\kappa = 0.989$ on *RS16*, see section 6.3.5). In more realistic setups for sets of four identical microphones, detection

performances of $\kappa = 0.678$ for the Rode condenser microphones and $\kappa = 0.767$ for the Beyerdynamics dynamic microphones (*RS4_Rode* and *RS4_Beyer*; see section 6.1.1) are achieved. Using the mapping between Kappa values and statistical confidence introduced in section 4.1.4, these results represent a 'fair to good' statistical confidence.

This investigation task is supposed to contain also an answer on what 'sufficient' means in terms of required training and testing (application / evaluation) set sizes. Here, due to the limited amount of available recording hardware, only a concept for answering this question (see section 6.1.2) as well as first estimations on required set sizes can be given. For the evaluations performed, a training set size of 200 windows (computed over 1024 audio sample values each) per considered recorded content (genre) already achieves close to optimal results[73] for most classifiers used here. If the amount of recorded material used in training is more than these (about) 5 seconds audio material per considered genre, the tests performed show only small improvements on the classification performance, while the computation time for model generation and classification strongly increases.
Regarding the tendency for overfitting, section 6.2.4 strongly implies that content dependency in the training and testing has a huge influence on the achieved classification accuracies. Like all other empirical results presented here for microphone forensics, these facts would have to be verified in future work with much larger microphone sets to extend the degree to which these results can be generalised.

Based on the results of the performed microphone forensics investigations, the summarising statement for investigation task A for this application scenario is: *The results presented here imply that statistical pattern recognition (SPR) based microphone identification is possible, if the degrees of freedom in the recording process can be controlled.*

The second part of this statement is based on the fact that certain degrees of freedom in the recording process (especially the recording environment, the mounting of the microphone and content influences) have to be considered in the generation of the registration data (training data set and / or reference model) for microphone authentication. Obviously, microphones to be identified have to registered (i.e. models have to be trained for them), otherwise the identification approach used here, like many other reference-based authentication approaches, would not work.
With this statement and its couterpart for audio steganalysis in section 5.4.1, research objective 1 (resp. research challenge (a)) is answered positively.
The investigations performed within thesis show significant influences from parametrisations of the components of the statistical pattern recognition (SPR) pipeline, the different degrees of freedom in the recording process as well as from potential post-processing operations. These influences are discussed in the summaries on investigation tasks B and C, which both focus on how adequately the application scenarios can be implemented with the introduced approach (research challenge c)), below.

The investigations on the impact of application scenario specific intrinsic influences to the statistical pattern recognition (SPR) process (**investigation task B**) look into the influences arising from different instantiations of the SPR pipeline.
This application task is mainly fuelled by two realisations: first, that the process of statistical pattern recognition (SPR) is a powerful but complex method, and second, that many different classification algorithms (as core component of the statistical pattern recognition (SPR) process) exist, which allow for a successful classification of recorded audio samples as belonging to a source microphone.
The complex SPR process can be considered as a four component processing pipeline (see section 2.4). In the following the evaluated influences to those four components are discussed:

[73] From the tests performed here it is assumed, that a training set size of 200 feature vectors per reference file (resulting for 10 chosen references in about in 2000 representative vectors per microphone) is suitably enough for the evaluations. At that size for each of the four microphones of *RS4_Beyer* and with a dimensionality of 590 attributes per vector WEKAs implementation of a multilayer-perceptron achieved an detection performance of $\kappa > 0.8$ at an inacceptable high computation time of more than 100 hours on the used test machine (a Intel Core 2 Duo E8400 CPU 3GHz with 4 GB RAM machine running Microsoft Windows XP, WEKA v.3.6.1 on Java SE 6 (32-bit Windows) with 1.6 GByte allocated RAM for each WEKA instance). Any further increase results in only slightly better (but not perfect) classification accuracies and strongly worse run-times.

- Pre-processing: the pre-processing operations have been restricted in this thesis to the absolute minimum (mostly windowing with a fixed window size)

- Feature extraction: the features are the enabling part of the pattern recognition method. If they allow the distinction between pattern and background and between different patterns, then a successful application of this method is possible. Here, with the AAFE and known good audio feature extractor is chosen for the most part of the investigations performed.

- Feature selection: this component complements the feature extraction by identifying the significant features and therefore allowing for the removal of the insignificant ones. The feature selection concept presented in this thesis is considered significant as well as representative for microphone forensics because it is applied independently on two intra-class test sets of statistically significant sizes, composed from microphones of the two most common microphone classes (dynamic and condenser microphones).
  The results of a PCA performed on the feature space of 590 segmental features computed by AAFE version 2.0.5 imply that for microphone forensics within this feature space only 187 independent dimensions exist (see section 6.1.4). By the reduction of the feature space to the 20 most relevant ones, the classification process is speed up by a factor of about 33 while keeping the detection performance achieved on a nearly constant level. Since the used procedure ranks the features by their benefit in classification, this ranking then can be used to derive domain knowledge about the considered classification problem. The evaluations imply that the microphone influence in recording – the intrinsic pattern – manifests itself the strongest in higher order cepstral-domain features, while time domain features, which are closest to the natural domain of the audio signal, do not play a strong role in the classification, since they are influenced to strongly by the recorded content. The global features computed by AAFE version 2.0.5 do not show any significance in this application scenario.

- Classification: As stated above on the methodology and solution concept used in this thesis (see e.g. section 3.1.3), the choice of classifiers is for this thesis is restricted to the application of already existing classifiers as implemented in WEKA (version 3.6.1), presumably showing very different performance in terms of classification accuracy achieved and computation time requirements. Here, an application specific classifier selection for existing classifiers is introduced, aiming at the identification of suitable classifiers for the microphone forensics application scenario. The results of this classifier selection are presented in detail in section 6.1.3. They show that clustering-driven classification seems to be of no use for the microphone forensics approach used here. Even if the number of clusters (here the microphone classes) for an audio data stream is known in advance (an unlikely scenario in microphone identification as well as composition detection) the detection performance is barely above $\kappa = 0$ (i.e. the probability of guessing correctly) and much lower than the performance of the supervised classification approaches on the same material. Supervised classification can be successfully used for microphone identification but so far no feature extractor / single classifier combination has been found that wields perfect results (a detection performance of $\kappa = 1$, preferably at a low computational run-time).

The summarising statement for investigation task B is: *The results for the intra-class classifications on sets of identical microphones in section 6.1.3 show clearly that: Notwithstanding the fact that all classifiers are used in default parametrisation – which has to be assumed to be sub-optimal (a fact which would require more detailed considerations on classifier optimisation and -generation, which are outside the scope of this thesis) – the results achieved in the intra-class recording classifications can be considered significant. I.e. it is not only possible to distinguish between recordings made by different models of the same brand and model, but also a sufficiently large number of different classifiers are capable of doing so with the evaluated setup for the SPR pipeline.*
Nevertheless, the different specific intrinsic influences to the SPR process have a strong impact on the achieved detection performance and would have to be adapted and optimised prior to any field application.

The investigations on influences outside the statistical pattern recognition process on the performance of the scheme (**investigation task C**) focus on two distinct sets of influence factors: the degrees of freedom in the recording process and selected, common audio signal post-processing operations.

Regarding the **degrees of freedom in the recording process**, the evaluations on room / recording environment classification presented first in [Kraetzer07c] are substantiated in this thesis by further experimental validation. Recording environment classification is obviously possible, but with the current feature set and classifiers combinations it seems to be performing less good than microphone authentication. If a microphone is run against a model generated in a wrong (different) room then the detection performance decreases. It can therefore be stated that the room / recording environment is a very strong influence on the recording behaviour of a microphone.
As the second degree of influence investigated, the microphone orientation seems to have no influence to the microphone classification problem, since the inter-microphone difference, even for microphones of the same brand and model, is higher than the differences between the recordings of one microphone in different orientations. It can therefore be assumed that the influence of the orientation factor *Or* introduced for the recording context model in section 2.3.2 is indeed very small (or since this influence is modelled as a multiplicative influence it is assumed to be rather close to 1 with a small variance). Based on the perfect classification results achieved with a recently recorded test set on a training set recorded one year ago, it can be assumed from those evaluations that the statistical patterns which allow for the classification of the microphones show for this time span no aging behaviour, i.e. no significant change over time. It can therefore be assumed that, like for the orientation influence, the aging factor *Age* introduced for the recording context model is also very small. Nevertheless, long term observations on this matter would be required using time spans of at least 5 to 10 years to allow for any generalisation on this fact.
From the degrees of freedom, the mounting only seems to have influence only in specific cases, where the vibration behaviour of the microphone is strongly influenced, like in the case where a microphone lies directly on a vibrating surface like a the top of the desk on which also the playback loudspeaker is standing. Otherwise the inter-microphone difference, even for microphones of the same brand and model, is higher than the differences between the recordings of one microphone in different mountings. The mounting factor *Mount* for the context model would therefore be modelled with a mean of 1 but with a larger variance than the orientation and aging factors.
As the last degrees of freedom investigated, the content dependency shows the following results: Notwithstanding the quite good classification accuracies achieved with the context sensitive approach from [Buchholz09] and the additional observations on the impact different classes of content (here speech vs. silence) the idea of performing content removal (e.g. by silence thresholding) is discarded for this thesis. Technically such a content selection might improve the detection performance achieved, but the restrictions imposed thereby are considered to be too severe to be useful in field applications. For identification purposes such a thresholding would prevent identification of recordings were no such 'silence' would be present in sufficient quantities. For composition detection purposes a context removal seems even more contra productive because normally loud parts are 'mixed in' to change the meaning of an audio data stream and not silence. Therefore, this approach is abandoned for the rest of this thesis instead the context-insensitive, general-purpose approach introduced by [Kraetzer07c] is pursued further.
Nevertheless, the tests on content selection in content dependent training as well as content independent training and testing performed within this thesis show, that the introduced approach might benefit from a content identification in pre-processing and a corresponding model selection. The implementation of such a pre-processing scheme, which involves the solution of another audio pattern recognition task – reliable content classification – is outside of the scope of this thesis and reserved for future research.

Regarding the selected, common **audio signal post-processing operations** considered within this thesis for their influence on the introduced microphone forensics approach, the investigations performed in section 6.3 show that the microphone forensics approach remains useful even after those post-processing operations. Neither the individual operations (normalisation, MP3 conversion and de-noising) nor the chosen exemplary combination tested are able to completely remove the microphone characteristics

used which lead in this approach to a correct classification.
The results presented for the investigations on playback detection imply that the introduced approach is to some extend resilient to this attack. A playback attack seems to have a strong negative influence on the detection performance but the tests performed still show a tendency to detect that two microphones actually involved in the recording process. Here, in many cases at least one of the microphones actually involved in the recording process is correctly identified. If normalisation of the signal is performed in between playback and recording, this additional influence has a significant negative impact on the detection performance. For the overall application scenario of microphone forensics, these results imply that a playback attack has a strong disturbing impact on the detection performance if it is not detected prior to the performed microphone authentication.

Regarding the composition detection investigations performed in section 6.4 it can be stated that it seems to be possible to detect a composition of recorded signals, as long as the microphone used for the recording of the major part in the recording is registered in the classifier model. It is shown here, that for strongly limited setups the mixing of audio recordings into another recorded audio signal can be very well detected by some classifiers. Of interest is the fact that from the four exemplary chosen classifiers for the evaluations the *SMO*, which is performing quite well in all other microphone forensics evaluations (see e.g. section 6.1.3) shows here a dissatisfactory detection performance. From the other exemplary evaluated classifiers, *RandomCommittee* and *RandomForest* do show here a very strong performance if the microphone used in the generation of the audio data stream into which other data is pasted into is registered in the classification models. In fact it seems not to matter much for their performance whether the material pasted into the original stream originates from a registered or unknown microphone.
In case none of the two sources for a composition (mesh-up) is registered, it is shown here for a small example that the change rate and average sequence length can be used to tell that a wrong model is used and, since a different tendency for classification of the individual feature vectors can be observed, that a composition is likely.

The summarising statement for investigation task C is: *There are a lot of influence factors to microphone forensics that are located outside the statistical pattern recognition (SPR) pipeline. To address the questions of plausibility and generalisibility these influence factors have to be identified and their impact has to be investigated. The context model for the microphone recording process presented in this thesis in section 2.3.2 is an important first step in these regards. It has to be fine-tuned in future work and accompanied by attack models for non-malicious or malicious (anti-forensic) modifications to the recorded audio signal.*

With the summarising statements for investigation tasks A, B and C for both exemplary selected application scenarios (audio steganalysis and microphone forensics), part of the question raised by research challenge (c) on how adequately the application scenarios can be implemented with the introduced approach is answered. The other part of this answer is given in the comparison with the state-of-the-art in both application scenarios in chapter 7.

6.5.2 Reflection on the evaluation criteria derived from the Daubert standard

The summary table presented below is derived from table 3.2 in section 3.3. Here, the progress made within this thesis in the application scenario of microphone forensics in regard to the criteria derived from the FRE rule 702 and the Daubert standard is summarised.

Table 6.41: Progress made in this thesis for microphone forensics – projection onto the Daubert criteria

Criterion	Description / Progress made
FREC0	**Description** ([LLI10a]): "*the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue*" **Progress made**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.
FREC1	**Description** ([LLI10a]): the investigation (which leads to the corresponding expert testimony) is "*based upon sufficient facts or data*" **Progress made**: Case specific, the only significance arises due to the fact that the term "*sufficient*" has to be manifested into training and testing (application / evaluation) set sizes. Here, the experiments performed imply that a small amount of audio material (a few seconds of recordings) seem to be enough for verification (see section 6.1.2). Nevertheless, the question what "*sufficient*" precisely means for the composition of the training data for arbitrary size multi-class problems is still an open question to be answered by future research.
FREC2	**Description** ([LLI10a]): the investigation is based upon "*reliable principles and methods*", preferably scientific methodology and knowledge **Progress made**: Chapter 2 as well as sections 3.1, 3.2 and chapter 4 of this thesis are dedicated to establish the fact that the two exemplary selected audio forensic methods are implemented as deterministic processes using the decades old and well accepted methodology of statistical pattern recognition (SPR). The fact that the microphone forensics application scenario can indeed be solved by SPR is successfully addressed in section 6.1.1.
FREC3	**Description** ([LLI10a]): the (forensic) methods are applied "*reliably to the facts of the case*" **Progress made**: Since this criterion is case specific for the law case at hand, the only thing that can be done within this thesis is to raise the awareness for its existence.
DC1	**Description** ([LLI10b]): "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"; summarised more precisely in [USC93] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*" **Progress made**: This criterion imposes the most important task to the practical investigations performed within this thesis: The main part of chapter 6 is dedicated to exactly this goal, trying to establish within which limits the proposed media forensic methods can give plausible results. It has to be admitted that the size of the experiments performed might still lack generalisability but the methodology and evaluation concepts show that there are ways for objectively challenging the introduced approach.
DC2	**Description** ([LLI10b]): "*whether the technique or theory has been subject to peer review and publication*"; with "*publication*" meaning 'open publication' **Progress made**: This criterion is not translated into tasks but instead requires the author to interact with the scientific community relevant for the chosen application scenario. To address this criterion this thesis is submitted for (peer) review, as have been the accompanying conference and workshop papers on the work on the microphone forensics application scenario. The reviewer comments received have helped shaping the described approach as well as its evaluations.
DC3	**Description** ([LLI10b]): "*the known or potential rate of error of the technique or theory when applied*" **Progress made**: It has to be admitted that the size of the experiments performed might still lack generalisability, but the detection performances achieved in evaluations on sets of identical microphones are promising. Nevertheless, they would have still have to be improved to achieve detection performances and corresponding error rates that are fit for application in court cases. Additionally, the combination with other techniques (e.g. the ENF-based approach discussed in section 2.6.1) might raise the chance to successfully pass a Daubert hearing.
DC4	**Description** ([LLI10b]): "*the existence and maintenance of standards and controls*" **Progress made**: The task that would be derived from this criterion would be the compilation of the work into standards together with or within a standardisation body. This complex process is outside the scope of this thesis, no progress made in this regard.
DC5	**Description** ([LLI10b]): "*whether the technique or theory has been generally accepted in the scientific community*" **Progress made**: This criterion is similar to DC2 in its meaning and in the fact that it is not translated into tasks, no progress made in this regard.

7 General Results – Comparison of both Application Scenarios

The solution of decision problems by statistical pattern recognition (SPR) requires mainly three things: first, features which allow the distinction between the possible pattern classes, second, a classification mechanism that can perform this distinction based on the features and a predefined rule set (the model), and third, a number of training samples that describe the decision problem sufficiently well to act as basis of reference for the generation of the model.

This thesis proposes a general-purpose audio forensics approach that provides solutions for the first and second point, together with the means necessary (here the feature and classifier selection strategies discussed) for the adoption to the two exemplary selected application scenarios. The third point cannot be addressed by any generalised concept. Here, an application scenario specific modelling of the decision problem, the classification problem, possible influence factors in the application field as well as potential countermeasures has to be performed.

For the evaluation of the performance of a solution approach it actually means, on one hand, the introduction of a suitable performance metric (as done within this thesis in accordance with the specified research objective 2 by application of the Kappa statistics used extensively in chapters 5 and 6 for detection performance evaluations) and, on the other hand, the definition of suitable training and test sets, the definition of the setup of the SPR-pipeline, the execution of the evaluations and considerations on the plausibility of the results.

The prospects of the introduced general-purpose audio forensics approach are obvious: as long as the patterns to be distinguished leave a discernible impact in the used feature space, it can easily be adapted to any new problem. The consequences of the adaptation are the description of the classification problem in terms of training sets, the application as a classification or decision engine, the performance and plausibility evaluations and finally the addressing of the Daubert criteria (including publication, discussion in the community, standardisation efforts, etc.).

Limitations to the applicability of the general-purpose approach might arise from different aspects. These include potential issues like:

- The modelling of the dimensionality of the classification problem to be solved (i.e. all potential classes of patterns have to be known in advance and training sets have to sufficiently describe the problem)
- Lack of features allowing to distinguish the patterns sufficiently (leading to too high error rates)
- Lacking plausibility (influence factors outside the application scenarios have an influence to the detection process, i.e. actual influences to the process are not sufficiently modelled)
- Lacking performance (detection performance or system throughput to low for the intended practical application)

Research objective 4 (as formulated in section 1.3) as well as the investigation task D) derived from this objective (see section 3.3) aim at showing the prospects and current limitations of the introduced

general-purpose audio SPR forensics approach. The question behind research objective 4 is: How large are steps required to adapt the general-purpose approach to a specific application scenario, like the two exemplary chosen for this thesis?
The actual adaptations of the introduced general-purpose approach have been described in chapters 4, 5 and 6 of this thesis. Below the outcome of both application scenario specific adaptations is compared: first, to show that both adaptations are adequate (in terms of similar detection performance as the state-of-the-art in the corresponding fields; research challenge (c) as defined in section 1.2), and second, to perform a summary and comparison of the results achieved (addressing research objective 4 as well as investigation task D)).

7.1 Detection performance and plausibility achieved

The detection performance remains the most important evaluation criterion for any statistical pattern recognition (SPR) approach. Here, first the detection performances are compared separately for the two application scenarios with the corresponding state-of-the-art. In section 7.1.3 the detection performances are compared between both addressed application scenarios to address investigation task D) as defined in section 3.3.

7.1.1 Comparison to the results presented in the state-of-the-art in audio steganalysis

First of all, it has to be mentioned that it is difficult to compare the performance of different audio steganalysis schemes. The corresponding research community in audio watermarking, steganography and steganalysis seems to be reluctant to exchange their algorithms (or even audio test sets, sets of stego objects or marked files). There are some benchmarking activities like StirMark Benchmark for Audio (SMBA)[74] which are capable of performing statistical analyses on audio material that could be used for audio steganalysis, but these tools lack practical acceptance by the community. In general, equivalents to open completions like BOWS, BOWS2 or BOSS (see section 2.5.1) hosted by researchers and organisations active in image data hiding are still missing in the research field of audio data hiding to which audio steganalysis belongs.

For the reasons mentioned above, a fair comparison with other publications in this field is hardly possible. If the mere detection performances are compared (on the basis of strongly differing audio sets) then the detection performance achieved here (between $\kappa = 0$ and $\kappa = 1$; equivalent to a range of detection accuracies between 50% and 100% in an equally distributed two-class problem) would be in a similar range as the results presented by [Özer03], [Altun05] or [Liu08].
What differentiates the results presented in this thesis from most publications in the state-of-the-art are the considerations on: required model sizes, the content dependency in training and testing, feature and classifier selection strategies, plausibility against some common audio signal modifications, comparisons of two-class and multi-class setups and, last but not least, the simple fact that here a general-purpose tool set is used instead of a highly specialised steganalysis detector.

7.1.2 Comparison to the results presented in the state-of-the-art in microphone forensics

The comparison with the state-of-the-art in this application scenario has to be performed based on the three categories of approaches identified in section 2.6.1: the electric network frequency (ENF) based approach, the time-domain local phenomena (reverberations) based approach and the microphone response based pattern recognition approach.
For the ENF approach, as the most mature in the set Grigoras et al. demonstrate extremely low error rates on impressively large data sets. Therefore, it is assumed by Grigoras that for ideal circumstances

[74] http://wwwiti.cs.uni-magdeburg.de/~alang/smba.php#smba_get; [Lang07]

the error rate for authentication is known. Other authors like Brixen argue that under non-ideal circumstances (see e.g. [Brixen08b]) or even under the assumption of counter-forensics the rates are not known and that reliable means for integrity verification based on the ENF still have to be devised.
A direct comparison to the time-domain local phenomena (reverberations) based approach of [Malik10] is impossible, since that paper lacks an evaluation of the approach introduced. Nevertheless, the main problem of that approach is its requirement for the existence of usable reverberations in the audio signal. Without such reverberations in the audio signal it will completely fail, a weakness that is not shared by the microphone forensics approach introduced in this thesis.
Within category of the microphone response based pattern recognition approaches, which the work introduced here shares with other authors like [Garcia-Romero10] and [Malik12], all performed evaluations show similar setups (especially in terms of the number of microphones used, which usually varies between four and eight) and detection performances achieved in controlled condition, closed-set experiments. Even though the results for those small sets of recordings cannot directly compared[75], it has to be assumed that the performance of the approach introduced here and the works described in [Garcia-Romero10] and [Malik12] show equivalent error rates, as long as the rather severe requirements for the two template matching based approaches (speech signals or silence, see section 2.6.2) are met. The approach introduced here does not impose such severe requirements on the recorded content. It even survives to some extend influences imposed by common audio signal post-processing operations (like normalisation, blind de-reverberation, MP3 conversion, etc.), a fact that distinguishes the introduced approach not only from its template matching counterparts but also from the ENF-based approach.

In summary, it can be said that the results presented in this thesis are differentiated from the publications in the state-of-the-art especially by the considerations on: different influences in the recording process, required model sizes, content dependent and independent evaluations, feature and classifier selection strategies, the plausibility against some common audio signal modifications and, last but not least, the simple fact that here a second instantiation of the same general-purpose tool set is used as in the steganalysis application scenario.

7.1.3 Comparison between the two application scenarios

A comparison between audio steganalysis, which is a research field strongly researched since the 1990s, and the much younger field of digital microphone forensics is hard. Both application scenarios have completely different goals if it comes to the security aspect primarily considered (integrity for audio steganalysis versus authenticity for microphone forensics).
What they share is the fact that they can be solved by means of pattern recognition. But also in this solution approach we see strong differences: steganalysis is in most works of the state-of-the-art modelled as a two-class decision problem[76], while microphone forensics is most commonly agreed upon as a multi-class decision problem aiming at microphone identification.

The main contribution of this thesis is the introduction of a general-purpose statistical pattern recognition (SPR) based audio forensics approach, which is capable of addressing both application scenarios successfully. While the results achieved with the introduced approach are well within the ranges shown in the corresponding state-of-the-art (see sections 7.1.1 and 7.1.2), the variance in the results achieved in the results is much higher in audio steganalysis. This implies that this application scenario provides the much tougher classification problem – which is easily understandable, considering the fact that audio steganalysis (or information hiding in general) aims at transparent and undetectable modifications of the cover objects.

[75] It has to be assumed that the detection performance is strongly influenced by the quality of the used microphones. Low quality microphones have a stronger noise footprint and should therefore simplify the pattern recognition problem.

[76] It has to be acknowledged that there are some authors (like Provos et al. in [Provos02] who argue in favour of a multi-class modelling of the practical steganalysis problem.

Regarding the evaluations performed here, the evaluation set sizes as well as the evaluation setups vary between both application scenarios: For audio steganalysis the evaluation set size is strongly influenced by the small number of audio steganography algorithms freely available. The evaluated influences are focussing on steganography specific aspects like key scenario or used embedding domain.
For microphone forensics the most severe limitations to the evaluations performed within this thesis are imposed by the extent of the recording setups that could be implemented. Like in the state-of-the-art in this field (especially [Garcia-Romero10] and [Malik12]) the used set sizes do not exceed eight microphones, which is a physical boundary imposed by current soundcards. Application scenario specific aspects for microphone forensics are the recording process specific influence factors as well as the composition detection considerations.

The main point here is that, even if a general-purpose audio forensic approach can be applied without strong changes to different application scenarios, the evaluation performed has to be adapted to the specifics of applications field. Furthermore, it will not show the same performance in all application scenarios, which is within this thesis well illustrated with the detection performance for the steganalysis on the algorithm A_{S1}, which only for specific cover signals achieved a Kappa value of $\kappa > 0$. The results presented for the performance in both application scenarios highlight the importance of two issues that are ignored in many of the state-of-the-art publications on the chosen application scenarios: the fitting (the correlation between material for training and the application or classification) and the plausibility (the resilience against non-malicious signal modifications as well as potential counter-forensics).

7.2 Achieved forensic compliance

Obviously, a single PhD thesis cannot achieve the tasks of making a forensic method Daubert compliant and acceptable in court cases. The goals pursued within this thesis in terms of forensic compliance have are: one hand, to raise the awareness within the corresponding research communities, and on the other hand, to investigate into the current degree of maturity within both application scenarios.
For the application scenario of **audio steganalysis**, the number of publications in the state-of-the-art is too large to perform a one-on-one comparison as done for microphone forensics in table 2.1 (section 2.6.2). The main points that distinguish the work performed on audio steganalysis in this thesis can be summarised on basis of section 5.4.2 as:

- The discussion of steganalysis as a forensic problem (which therefore has to acknowledge the criteria of the Daubert standard).
- Investigations on training set sizes required to address the pattern recognition problem sufficiently (criterion FREC1).
- An in-depth discussion of the complete statistical pattern recognition (SPR) pipeline as methodological basis and different instantiations thereof as evaluation designs (criterion FREC2).
- Practical investigations on two-class and multi-class setups for audio steganalysis (criterion DC1).
- Detection performance and plausibility considerations (criterion DC3).

Since the number of approaches in the state-of-the-art in **microphone forensics** is currently very limited, the complete one-to-one comparison performed in table 2.1 in section 2.6.2 can be extended to compare the introduced approach with the forensic performance of the existing approaches. The result of this extension is shown in table 7.1.

Table 7.1: Using the Daubert criteria (see section 2.2) for comparison of the existing microphone forensics approaches and the results achieved for this application scenario within this thesis

	electric network frequency (ENF)	**time-domain local phenomena (reverberations)**	**microphone response based pattern recognition**		
	Initial publication: [Grigoras03]	Initial publication: [Malik10]	Initial publication: [Garcia-Romero10]	[Malik12]	Initial publication: [Oermann05] (theory) & [Kraetzer07c] (solution approach)
FREC0: the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue	criteria, that cannot be answered in general, because they are related to the specific court case under consideration				
FREC1: the investigation is based upon sufficient facts or data					
FREC2: the investigation is based upon reliable principles and methods, preferably scientific methodology and knowledge	Rather mature, considered in [Bijhold07]	Only a concept not tested	Pattern recognition / template matching	Pattern recognition / template matching	Pattern recognition / SPR
FREC3: the methods are reliably applied to the facts at hand	criterion, that cannot be answered in general, because it is related to the specific court case under consideration				
DC1: "*whether the expert's technique or theory can be or has been tested*"	yes, large scale tests (see e.g. [Grigoras05])	no	yes, limited closed-set experiments (two sets with eight microphones each)	yes, limited closed-set experiments (one set with eight microphones)	yes, limited closed-set experiments (sets between 4 and 7 microphones, incl. setups with devices of the same brand and model)
DC2: "*whether the technique or theory has been subject to peer review and publication*"	Publication count: >20	Publication count: 1	Publication count: 1	Publication count: 1	Publication count: 6

Continued on Next Page...

Table 7.1 – Continued

	electric network frequency (ENF)	**time-domain local phenomena (reverberations)**	**microphone response based pattern recognition**		
DC3: “*the known or potential rate of error of the technique or theory when applied*”	for ideal circumstances the error rate for authentication is known, for integrity verification it is not known under non-ideal circumstances (see e.g. [Brixen08b]) or even under the assumption of counter-forensics the rates are not known	no	only for two small sets and under ideal (speech only) circumstances	only for a small set and under ideal (environmental noise generated with a 12-inch fan only) circumstances	only for two small sets, considering different influences in the recording process, different recording contexts as well as selected signal post-processing operations
DC4: “*the existence and maintenance of standards and controls*”	European Network of Forensic Science Institutes (ENFSI) Forensic Speech and Audio Analysis Working Group (FSAAWG) guidelines on ENF analysis in forensic authentication of digital evidence [Grigoras09]	no	no	no	no
DC5: “*whether the technique or theory has been generally accepted in the scientific community*”	supporting arguments: large number of publications and citations, applied also for video recordings, document [Bijhold07] compiled by forensic experts from different police forces for an INTERPOL Forensic Science Symposium – opposing arguments: context dependency, does not work for DC powered devices, [Brixen08b]: “*In praxis, approximately 40-60% of the digital recordings in question contain traceable ENF*”	supporting arguments: none known – opposing arguments: context dependency, [Gupta12]: “*Currently, this measure has been successfully applied to synthesized audio with assumptions that cannot be fulfilled by most real-world signals*”. [REW11]: hindered by common signal post-processing operations (e.g. blind de-reverberation) performed in many application scenarios, like audio / video conferencing, hands-free telephone, etc	supporting arguments: none known – opposing arguments: context dependency (speech only)	supporting arguments: none known – opposing arguments: only tested with one kind of recording content (environmental noise generated with a 12-inch fan)	supporting arguments: none known – opposing arguments: detection performance strongly dependant on the availability of suitable models (i.e. strong correlation between training and test material required)

The main points that distinguish the work performed on microphone forensics in this thesis can be summarised on basis of section 6.5.2 and table 7.1 as:

- The discussion of steganalysis as a forensic problem (which therefore has to acknowledge the criteria of the Daubert standard) – this consideration was so far restricted to the electric network frequency (ENF) approach developed by Grigoras et al.
- Investigations on training set sizes required to address the pattern recognition problem sufficiently (criterion FREC1),
- An in-depth discussion of the complete statistical pattern recognition (SPR) pipeline as methodological basis and different instantiations thereof as evaluation designs (criterion FREC2),
- Practical investigations on different influence factors to the recording process as well as on the context (in)dependency of the SPR-based approach (criterion DC1),
- Detection performance and plausibility considerations (criterion DC3)

If the progress made in **both application scenarios** is compared, then the reflections on the evaluation criteria derived from the Daubert standard (see sections 5.4.2 and 6.5.2 respectively) show strong similarities between both application scenarios. Nevertheless, the progress is assumedly higher for the microphone forensics application scenario which is still much younger, in term of scientific maturity, than the audio steganalysis counterpart.

For the former, the main achievement is the introduction of a solution method which shows less severe recording content restrictions than the current state-of-the-art and the consideration of post-processing operations is rather new to the application field. For the latter, the consideration as a forensic mechanism is a new angle onto steganalysis. The practical results presented here outdo most of the state-of-the-art, because they integrate plausibility as well as two-class versus multi-class considerations. Nevertheless, the work performed here is hardly capable of moving the whole application scenario closer to the requirements of the ideal forensic audio steganalysis process (see section 3.1.1) – which would be required to get steganalysis results accepted as relevant for the case at hand by a judge presiding any court case.

For microphone forensics the main achievement is the introduction of a solution method which shows less restrictions than the state-of-the-art in terms of content restrictions. Even though it is more tolerant in theory, the evaluations performed show that the SPR-approach strongly relies on the availability of suitable models (or training material).

8 Summary, Conclusions, Ongoing and Future Work

This chapter presents in section 8.1 a summary of the results of this thesis and the conclusions that can be drawn from these results. To avoid redundancy, a detailed re-iteration of the generalisation of the results (as performed in chapter 7) as well as a summary of the main contributions of the thesis (presented in section 1.4) is omitted here. Instead the considerations include a projection of general lessons learned onto similar application scenarios in the image domain (here, image steganalysis and camera forensics).

Section 8.2 discusses, with the benchmarking of statistical pattern recognition (SPR) based approaches and information fusion, two topics from ongoing research work that are based on the outcome of this thesis.

In section 8.3 possible directions for future work are summarised.

8.1 Summary and conclusions

In this thesis a general-purpose statistical pattern recognition (SPR) based audio forensics approach is introduced. It is implemented using a new designed, high dimensional set of easy to compute audio features and existing classification techniques from the renowned data mining suite WEKA [Hall09].

The answers generated by the solving the investigation tasks derived in section 3.3 from the research challenges and objectives for this thesis (see sections 1.2 and 1.3) can be summarised into short statements as follows:

For **research challenge (a)** – 'Existence of a generalised SPR approach for audio forensics' – the successful application of the introduced approach to the exemplary selected scenarios of audio steganalysis and microphone forensics shows that a generalised statistical pattern recognition (SPR) approach as an adaptable solution concept for media forensics is indeed possible.

Regarding **research challenge (b)** – 'Applicable performance measures' – the considerations made within the thesis on detection performance evaluations using Kappa statistics already provide a metric that outperforms the dominant choice in the state-of-the-art in both application scenarios (i.e. the detection accuracy) in terms of interpretability and comparability. In section 8.2.1 below, one possible extension of this metric towards a benchmarking scheme is discussed as current, ongoing work.

For **research challenge (c)** – 'Adequate implementation of mechanisms for the chosen application scenarios' – the evaluation results presented in chapters 5, 6 and 7 show that the developed instantiations are capable of performing adequately (with similar detection performance as the state-of-the-art approaches in the corresponding research fields) in both application scenarios (see sections 7.1.1 and 7.1.2 respectively).

In the performed evaluations specific influence factors to the performance of the approach are considered. These include approach intrinsic influences (e.g. from audio features computed or classification algorithms used) as well as external influences (like the embedding domain or key scenario used in steganography or specific influences in the recording process in microphone forensics). One special class of external influences considered are the resilience of the solutions for steganalysis and microphone forensics against selected, common audio signal post-processing operations. These evaluations

aim specifically at the plausibility of the approach. They establish the fact that such post-processing operations can indeed influence the detection performance achieved and have to be evaluated and compensated prior to any intended field application of such a SPR-based security mechanism.
A further important point within this thesis is the application of the Daubert standard as a projection surface for the estimation of the forensic conformity of the introduced approach. The goal here has never been to try to make the solutions Daubert compliant – this is not achievable with one PhD thesis. Instead it is shown here how the criteria of the standard can be used to estimate the maturity of a forensic approach. Despite the fact that neither the universal audio steganalysis nor the microphone forensics approach introduced within this thesis will be able to pass a Daubert hearing anytime soon, the work performed within this thesis makes important steps into this direction.

Lessons learned that would similarly apply to similar application scenarios considering other media types (e.g. in image steganalysis and camera forensics)

Audio signal analysis receives in the field of media forensics much less attention than image analysis. This can be very well illustrated by comparing audio steganalysis and its image counterpart or microphone forensics to digital camera forensics. In both compared examples the numbers of publications are much higher for the image domain, strongly implying a higher maturity of the approaches for that domain (see the analyses on the state-of-the-art in section 2.5 and 2.6 where also image domain counterparts are briefly reflected).
Nevertheless, there some lessons learned in this thesis, that would similarly apply to image steganalysis or camera forensics, despite the higher maturity of the approaches in these fields:

- **In general:**
  - Statistical pattern recognition (SPR) is a complex process that does not only consist of the feature extraction and classification components. Here, also other process components should receive attention in research. This is especially true for feature selection, which would allow the deduction of information about a signal modification (and the pattern it imposes to the signal) from the data under analysis.
  - Any pattern recognition (PR) approach is strongly influenced by the used training and testing material (and the contextual correlation between both sets). Here, statistically significant, representative and openly available sets would be required for any fair comparison of results.
  - Fair evaluation would also require metrics that allow a better comparison of the results of different tests. A first step into this direction could be the abandoning of the classification accuracy for something like the Kappa statistics used within this thesis.
  - For any forensic application, the impact of malicious (counter-forensics) and non-malicious signal post-processing should be integrated into plausibility considerations for any analysis approach. This is already done in some publications but yet not in general.
  - The Daubert criteria are of importance for anyone doing research in forensics, including media forensics.
- **For image steganalysis:**
  - There exists more than one good classification approach. Right now SVMs dominate the research in image steganalysis, here it might be beneficial to investigate also other classifiers.
  - Any steganalysis attempt should be aware of the two-class vs. multi-class discussion raised by Provos et al. in [Provos02] as well as the high probability that an decryption problem is attached to the steganalysis (see the discussion of the ideal forensic audio steganalysis process in section 3.1.1).

- **For camera forensics:**
  - The currently most mature[77] approach in this field is the PRNU-based approach introduced by Jessica Fridrich and her group (see e.g. [Goljan09]). It might be possible that even such a rather mature approach could be enhanced in its performance (e.g. numbers of cameras that can be distinguished, resilience to changing environmental conditions, etc.) by the application of SPR instead of distance based template matching.

8.2 Selected topics from ongoing research work

In [Kraetzer09a] it is shown that some of WEKAs classifiers are able to outperform the SVM classifier *libSVM* (which is in many publications on steganography considered to be the expert classification engine for this application scenario) in terms of practically achieved accuracies. This points out why it is so important to run application scenario based **benchmarking statistical pattern recognition (SPR) based mechanisms**. First ideas on such benchmarking can be derived from the research work presented for both application scenarios in chapters 5 and 6. These first ideas are presented in section 8.2.1. The main concept is the extension of classifier selection criteria from classification *accuracy* based evaluations to something more practically relevant, in [Kraetzer10] we introduce a first single quality function driven benchmarking based on *accuracy* and runtime. This very first quality function simply computes the quotient of accuracy and runtime. It is within [Kraetzer12a] substituted by the more complex (and fairer) quality function, which is, with slight modifications, here also adapted for microphone forensics. Within this thesis, these benchmarking efforts for different instantiations of the introduced general-purpose, statistical pattern recognition (SPR) driven audio forensics approach are summarised and the required next steps for future work are outlined.

Information fusion, as the second topic in the ongoing work, is focusing on combination strategies for the output of individual expert systems info a combined statement. During the research work on this thesis, initial considerations on information fusion in the application scenarios of audio steganalysis and microphone forensics have been published by the author in a small number of papers like [Kraetzer09a], [Kraetzer10] and [Kraetzer09b]. To completely cover the immensely complex topic of information fusion for SPR-based security mechanisms is outside the focus of this thesis. Nevertheless, the author considers this topic to be of huge importance for this research field, motivating the presentation of corresponding research results in section 8.2.2.

8.2.1 Throughput analysis as a step towards benchmarking

Benchmarking, focusing on making systems comparable, focuses on the projection of system properties onto certain predefined indicators. Intuitive approaches result in one identifier describing the system performance. In this case a quality function is used to generate from the (complex) system description, presented by weighted performance indicators, the one figure that typifies its performance rating. Together with a comparison scheme description (e.g. ‘the bigger the figure, the better’), this figure can then be used to compare multiple systems or to determine the discrepancy between a system and a defined goal.

The alternative to this intuitive single-figure approach is a multi-figure benchmarking scheme. Here, either due to the system complexity or to a lacking precise application goal definition, the system performance is projected by a set of functions onto a tuple of predefined indicators. For this alternative approach, a more complex comparison scheme description might be necessary for the identification of the best candidate in a number of alternative approaches that undergo comparison in benchmarking.

[77] In the Daubert hearings of the law case *United States of America v. Nathan Allen Railey* (United States District Court for the Southern District of Alabama, August 2nd, 2011), the method got accepted for the first time as forensic evidence. The FBIs Forensic Audio, Video, and Image Analysis Unit (FAVIAU) established in the Daubert hearings that this approach meets all necessary criteria and the presiding judge furthermore decided that this evidence (or more precisely the FBI expert testimony based on this media forensic analysis) could be accepted into the trial.

Despite the fact that the compliance with the Federal Rules of Evidence (FRE) and Daubert standard compliance can be considered to be the general aim for any forensic technique, these criteria are not the only performance indicators that determine whether a specific mechanism will ever be applied in an investigation. Further characteristics of a technique also determine its suitability for practical application and have therefore to be integrated in a **benchmarking methodology**. In this thesis, the focus is set on already measurable technical aspects (like detection performance and runtime requirements), reserving organisational aspects, like e.g. costs estimation, for future considerations.

As stated in the analysis of the state-of-the-art in research on the two chosen application scenarios in chapter 2, the classification $accuracy$ (the ratio between true classifications and all classification attempts in a supervised classification) is currently in both fields the dominating performance metric. To overcome the limitations of the $accuracy$ in terms of comparativeness, a naive design for a gain-to-cost-ratio based **metric** between detection performance (as gain) and $runtime$[78] required (runtime complexity as cost), which has to be considered to be fairer than only the $accuracy$, could look for a two-class setup with equally distributed classes like (see [Kraetzer12a]):

$$q = \begin{cases} \frac{accuracy}{runtime} & if \quad accuracy > 0 \\ 0 & if \quad accuracy = 0 \end{cases} \tag{8.1}$$

If the $accuracy$ of the classifier is better than guessing (i.e. 50% in this equally distributed two-class problem), then its classifier throughput performance q is determined by the $accuracy$ achieved on a fixed sized classification problem divided by the classifiers $runtime$ (combined training and testing times) on this problem for a selected test machine[79]. The measurement unit of this computation would be percentage of true (positive and negative) classifications per second, which is, for the standardised set sizes, a simplified version of the more intuitive 'percentage of correctly classified files per second' ratio.

Nevertheless, it has to be admitted that this simple metric is hardly applicable in practice, unfair and its result hard to interpret. We have to consider it hardly applicable because it assumes, on one hand, that we have always equally distributed classes, and, on the other hand, that no classification performance worse than guessing (in this example 50% $accuracy$) is possible. These two assumptions are rather unlikely in practice. From a scientific point this naive metric is also unfair, because it does not compare the classification algorithms but instead compares their implementations. Therefore, a rather well suited algorithm implemented in an interpreted language might be ranked lower than a less suitable algorithm implemented directly in machine code, only because the latter can be executed much faster. For the same reason, results achieved on different computers would not be directly comparable. From the practical point of view these two points, which would be considered as unfair by scientists, would be a desired characteristic of the detection system. The person wanting to install a steganographic channel detector to observe communications or data exchanges would exactly look for the fastest implementation as well as the most suitable (in most cases the fastest) computer to run the detector.

Another point, which makes this concept not exactly unfair but instead inept to handle certain benchmarking problems, is the fact that the $accuracy$, if used directly, is not suitable for comparisons between different classification problem classes. For example the direct comparison of the classification performance in a two-class classification problem (i.e. the classical hypothesis testing for a assumable steganographically modified channel) and a 4-class problem (i.e. steganographic algorithm identification on a set of three algorithms (plus unmodified covers) that might have been applied, see e.g. the work of Provos and Honeyman summarised in section 2.5.2) would lead to completely misleading results, because in the equally distributed two-class problem the probability of guessing correctly is two times higher (i.e. 50%, while in an equally distributed 4-class problem an $accuracy$ of 50% would already be

[78] In theory, the $runtime$ of a forensic method could be ignored. As long as it achieves its goals, we would not care whether it takes seconds, minutes, hours, days, or in some cases even years, to accomplish the task. In practice, nevertheless fast forensic methods are preferred for obvious reasons.

[79] The $runtime$ is the execution time of the classifier on a given classification problem (training and testing) measured in seconds (for this thesis using the Unix time() command [The Open Group08]).

a rather good indicator, being 25% away from the probability of guessing correctly in this case). This point basically implies a strong need for normalisation of results.

For the interpretability of the results, the $accuracy$ is expressed as a percentage between 0 and 100% and the $runtime$ is given in seconds and is not bounded. Therefore the result is not normalised in any way so that the actual distance from an 'optimal' performance is hard to figure out. Also, the notion of the $runtime$ used here combines the training and the testing times (while in a field application the models would be in many cases assumed to have been trained in advance) of a classifier. Since the ratio between training and testing times varies strongly between individual classifiers, the usage of this combined time might be enormously unfair for application scenarios where the classifier can be trained in advance, i.e. where the characteristics of the expected cover objects and steganographic embedding techniques are known a priori and appropriate training material can be supplied for training. In other application scenarios, where the models could not be trained in advance (due to a lack of knowledge regarding the cover material and/or techniques to be expected or if appropriate training material is missing – see e.g. [Özer03] where the 'unmarked cover' version of an audio file is estimated/predicted by using de-noising on the assumed stego object), this modelling of the $runtime$ would be the only suitable approach.

The points mentioned above lead to a redesign of our naive gain-to-cost-ratio based metric q for benchmarking purposes. In the modified metric we still use for the run-time the time required for the classifier (because the ultimate goal would be the practical application in tools and in this case a faster implementation of an algorithm is better than a slower implementation). Additionally, we introduce a fixed timeout-boundary, after which a classifier working on a problem is automatically considered unfit for this problem independent of the classification accuracy he might have achieved in the end. This $timeout$ serves two purposes: first, it makes algorithm benchmarking evaluations more feasible by faster removing candidates which would in any case unsuitable for practical application, and second, it allows to generate a normalised runtime description.

$$time = \frac{runtime}{timeout} \tag{8.2}$$

Equation 8.2 shows the normalised runtime description ($time$) used for an improvement of the quality function for the throughput analysis. The $runtime$ is the execution time of the classifier on a given classification problem (training and testing) measured in seconds (for this thesis using the Unix time() command [The Open Group08]). The $timeout$ is the timeout-boundary predefined for this investigation. Since the execution of the classifier is terminated at $timeout$, the resulting $time$ is a variable devoid of a unit in the range $[0,1]$.

For the measurement of the classification gain for fair performance evaluation within this thesis it is proposed to use the **Kappa statistics** instead of the accuracy (see section 4.1.4). By using Kappa statistics, it is possible to construct for classification-based investigations a degree of closeness of measurements of a quantity to its actual (true) value that is exempt from the influence of the probability of guessing correctly. To construct the new quality metric q_{new} for the benchmarking work in this thesis, the (normalised) Euclidean distance between $time$ and an inverted κ is computed. This inversion has to be performed since the $time$, as introduced in equation 8.2, is a 'the-bigger-the-worse' and metric and the Kappa statistics κ is be a 'the-bigger-the-better' metric. The metric q_{new} is therefore computed as:

$$q_{new} = \frac{1}{\sqrt{2}}\sqrt{time^2 + (1-\kappa)^2} \tag{8.3}$$

Since $time$ is bounded in the range $[0,1]$ and κ is assumed to yield results in the range $[0,1]$, the Euclidean distance has to be normalised with $\sqrt{2}$. The result of this computation q_{new} is, like $time$, a 'bigger-the-worse' metric in the range $[0,1]$. It describes the distance of a current performance from the 'optimal' point, which would be a forensic decision machine that gives a perfect classification ($\kappa = 1$) in an extremely short time-span ($time = 0$). Therefore a classification result which is very bad (equal to the probability of guessing, $\kappa = 0$) and finishes only shortly before the $timeout$-boundary ($time = 1$) would be as far as possible from this optimal point with $q_{new} = 1$ in this case. The threshold for suitable classifiers is moved by the normalisation performed to the value of $\frac{1}{\sqrt{2}}$, i.e. classifiers that only guess at the result but do so very fast are located exactly at this boundary. In case of extremely unlikely Kappa values (i.e. $\kappa < 0$) the metric still works well, resulting in values for q_{new} that could be larger than 1.

Summarising the benefits of this new performance metric q_{new}, it can be said that:

- It takes the runtimes of the classifier/detector implementations into account, which is closer to the practical requirements for such a system (i.e. faster implementations would be preferred over slower implementations with the same detection power).
- It efficiently removes classifiers that are per definition unsuitable from the list of candidates by defining a timeout boundary for the execution time.
- It allows for an intuitive performance description by using as a metric a normalised distance from an easy to understand 'optimal' operation point.
- It allows for a direct comparison between classifications of different class-sizes (e.g. two-class problems and 4-class problems).
- It is independent of the composition of the training and test set sizes (i.e. they do not have to be equally distributed)

The drawbacks of this metric can be summarised as follows:

- It is dependent of the machine it is run on. This drawback could easily be compensated by computing a time correction factor between different machines to make their $runtime$ results directly comparable.
- It is strongly dependent of the training and test sets used, because they directly influence the runtime as well as the Kappa values achieved – it has to be made sure that these sets are representative for the application scenario. As a result, the values computed for q_{new} are only comparable within one application scenario and still lack comparability between different application scenarios.
- For the selection of methods for the implementation of a security mechanism, it would have to be accompanied by another value or set of values for precise throughput description (e.g. the processing speed in feature vectors per second – which could be given separately for training and testing in case the training can be performed a priori).

For the integration of the metric q_{new} into the **benchmarking methodology**, a precise transfer function $SFQ_{Classifier}()$ has to be defined for every detector in the benchmark. Examples for such transfer functions are described in the following for the considered application scenarios of audio steganalysis and microphone forensics.

Application of the proposed benchmarking scheme for audio steganalysis

In this application scenario specific discussion on the classifier benchmarking methodology four different instantiations of the transfer function $SFQ_{Classifier}()$ are compared. These four example instantiation differ in the feature space used in the classification (AAFE v.2.0.5 all 17 global features vs. all 590 segmental features) and in the test strategy (10-fold stratified cross-validation vs. training and testing with independent sets).
All four instantiations use in this application scenario a timeout of 12 hours (43200s) and the large-scale multi-genre set *aats389* designed by Lang et al. ([Lang07]) especially for audio benchmarking purposes.

Example instantiation 1: Based on experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*, for the three exemplary selected IH algorithms and all 74 supervised classifiers implemented in WEKA (v.3.6.1) a benchmarking value is computed as:

$$q_{new_{classifier}} = SFQ_{classifier}(q_{new}, Alg, GF_{all}, timeout = 12h, cv, aats389) \tag{8.4}$$

The index for the transfer function is the classifier (with its parametrisation, which is here kept for all classifiers to the default setting). The metric used by the transfer function is q_{new} (as introduced with its Euclidean distance measurement above). Additional parameters supplied to the transfer function are the IH algorithm Alg, the feature set used (here GF_{all} – all 17 global features of AAFE v.2.0.5), the defined $timeout$ boundary (here 12h) and the test method for the classifier (here cv, 10-fold stratified cross-validation on the audio test set $aats389$).
Figure 8.1 shows, for the three exemplary selected IH algorithms A_{S1}, A_{S3} and A_{W1} and all 74 classifiers the achieved benchmarking performances in a time-duration vs. κ diagram. The point $(0{,}1)$ in this diagram is the reference for the optimal performance. A classifier close to this point would deliver a close to perfect detection performance in a very short time-duration.
As can be seen, in this test only four classifiers exceed the 100s time-duration mark. It is also evident in this figure that the detection performance achieved by the classifiers strongly differs between the three IH algorithms. An summarising value for all 74 classifiers could be expressed by the centre of gravity of the achieved κ values for each algorithm).

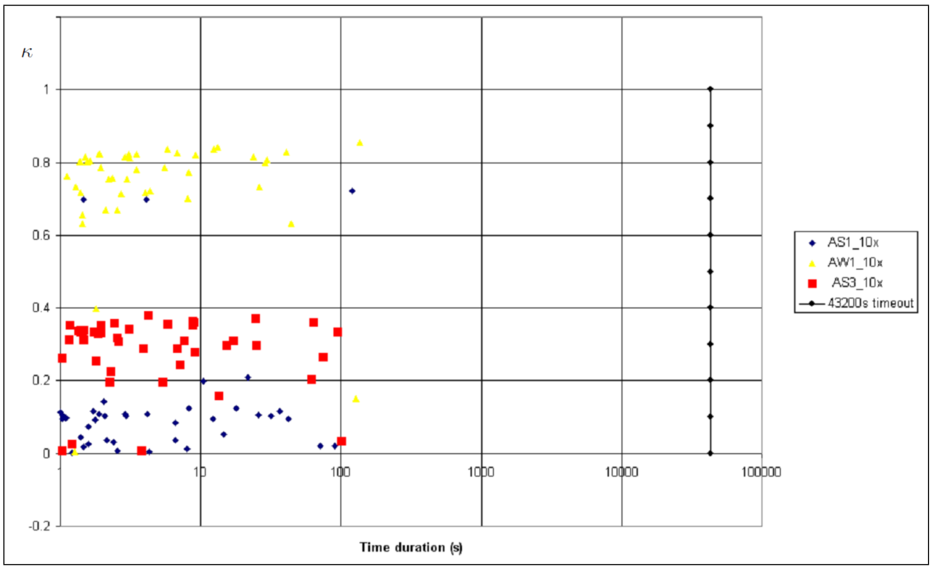


Figure 8.1: Time duration vs. κ diagram for all 74 classifiers and the 17 global features in AAFE v.2.0.5 in 10-fold stratified cross-validation (experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*)

Example instantiation 2: Based on experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*, for the three exemplary selected IH algorithms and all 74 supervised classifiers implemented in WEKA (v.3.6.1) a benchmarking value is computed as:

$$q_{new_{classifier}} = SFQ_{classifier}(q_{new}, Alg, GF_{all}, timeout = 12h, trte, aats389 + testset24) \tag{8.5}$$

The metric used, the IH algorithms, the feature set and the $timeout$ boundary for this instantiation are inherited from instantiation 1 above. Only the test method is changed to independent training and testing ($trte$) with the audio test set $aats389$ for training and $testset24$ for testing.
Figure 8.2 shows the same diagram-like figure 8.1 for the test case of two-set training and testing. In comparing the two figures two things are obvious: on one hand the average computation time required for the classifications decreases in this test mode, on the other hand the centres of gravity of the achieved κ values for the three algorithms move away from the optimum point in the upper left corner, as a result of the decreased classification performance in this test cases.

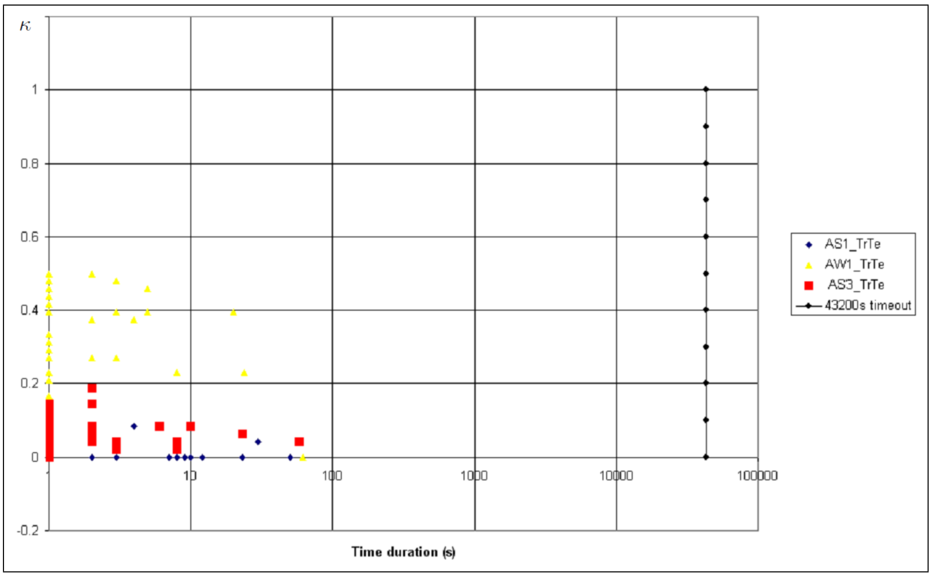


Figure 8.2: Time duration vs. κ diagram for all 74 classifiers and the 17 global features in AAFE v.2.0.5 in independent training and testing (experimental setup *AS-Kraetzer2010SPIE-GF-singleClass-summary*)

Table 8.1 identifies (based on *AS-Kraetzer2010SPIE-GF-singleClass-summary*) the best five classifiers for each algorithm and the global features. In contrast to most of the state-of-the art in research in this field, the best algorithms are here not determined by looking only on the classification accuracies achieved but instead by the ratio between the detection performance of a classifier and the time it requires for the decision. This is based on the observation that in practical application a faster decision is sometimes much more valuable than a much slower but slightly more accurate answer.

Table 8.1: Identification of the five classifiers for each IH algorithm with the lowest distance from the optimal performance, 10-fold stratified cross-validation as well as two-set training and testing using global features (the value in brackets identify the distance value q_{new})

10-fold stratified cross-validation			
	A_{S1}	A_{W1}	A_{S3}
Best classifier	*lazy.Kstar* (0.198)	*trees.LMT* (0.106)	*trees.LADTree* (0.438)
2nd	*lazy.Ibk* (0.212)	*meta.RotationForest* (0.113)	*meta.Decorate* (0.445)
3rd	*lazy.IB1* (0.212)	*trees.FT* (0.113)	*trees.BFTree* (0.453)
4th	*trees.RandomForest* (0.488)	*functions.MultilayerPerceptron* (0.113)	*meta.ClassificationViaRegression* (0.453)
5th	*rules.Nnge* (0.559)	*meta.EnsembleSelection* (0.120)	*trees.LMT* (0.453)

two-set training and testing			
	A_{S1}	A_{W1}	A_{S3}
Best classifier	*rules.OneR* (0.601)	*trees.RandomForest* (0.354)	*meta.ClassificationViaRegression* (0.573)
2nd	*trees.RandomTree* (0.651)	*meta.Bagging* (0.354)	*meta.RandomCommittee* (0.601)
3rd	*meta.LogitBoost* (0.665)	*meta.RotationForest* (0.354)	*trees.FT* (0.601)
4th	*trees.RandomForest* (0.665)	*meta.ClassificationViaRegression* (0.368)	*trees.RandomForest* (0.615)
5th	*rules.ConjunctiveRule* (0.679)	*trees.SimpleCart* (0.368)	*rules.OneR* (0.615)

In case of the 10-fold stratified cross-validation, the best classifier for A_{S1} (*lazy.Kstar*) achieves with $q_{new} = 0.198$ about half the quality rating as the best classifier for A_{W1} (*trees.LMT*, $q_{new} = 0.106$), simply because it takes about two times the time to reach a decision with a similar κ value.

Example instantiation 3:
Based on experimental setup *AS-Kraetzer2010SPIE-SF-singleClass-summary*, for the three exemplary selected IH algorithms and all 74 supervised classifiers implemented in WEKA (v.3.6.1) a benchmarking value is computed as:

$$q_{new_{classifier}} = SFQ_{classifier}(q_{new}, Alg, SF_{all}, timeout = 12h, cv, aats389) \tag{8.6}$$

The setup is nearly identical to the one described in equation 8.1 for instantiation 1. The only deviation is the choice of segmental features instead of global features. Here the feature set SF_{all} is used, containing all 590 segmental features computed by AAFE v.2.0.5.

Figure 8.3 shows the time-duration vs. κ diagram for all 74 classifiers and the three IH algorithms A_{S1}, A_{S3} and A_{W1}. Again, the point $(0,1)$ in this diagram is the reference for the optimal performance. A classifier close to this point would deliver a close to perfect κ value in a very short time-duration.

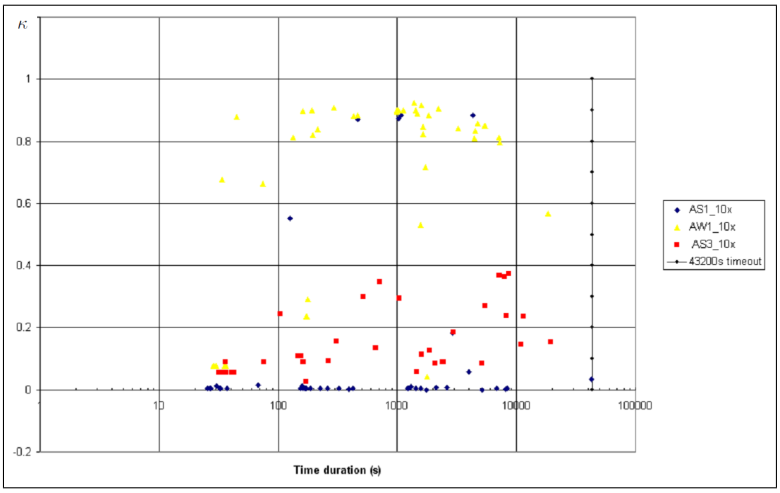


Figure 8.3: Time duration vs. κ diagram for all 74 classifiers and the 590 segmental features in AAFE v.2.0.5 in 10-fold stratified cross-validation (experimental setup *AS-Kraetzer2010SPIE-SF-singleClass-summary*)

Example instantiation 4: Based on experimental setup *AS-Kraetzer2010SPIE-SF-singleClass-summary*, for the three exemplary selected IH algorithms and all 74 supervised classifiers implemented in WEKA (v.3.6.1) a benchmarking value is computed as:

$$q_{new_{classifier}} = SFQ_{classifier}(q_{new}, Alg, SF_{all}, timeout = 12h, trte, aats389 + testset24) \tag{8.7}$$

This setup mimics the one in instantiation 3 but exchanges the evaluation strategy to $trte$ – independent training (on *aats389*) and testing (on *testset24*).

Figure 8.4 shows the time-duration vs. κ diagram for this instantiation.

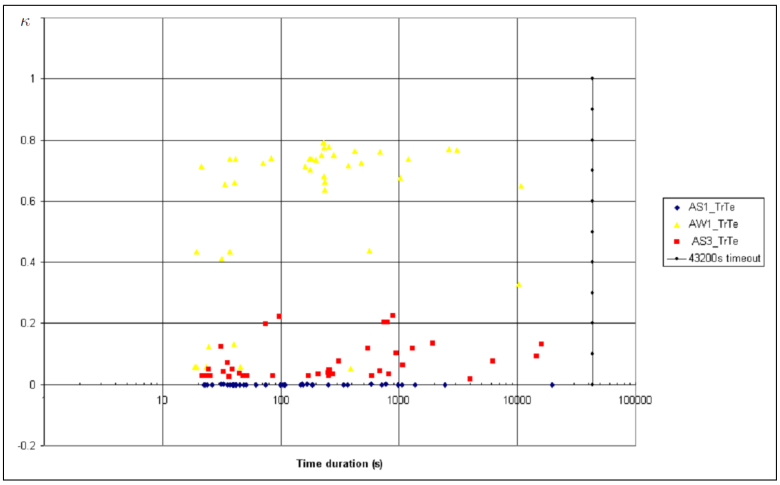


Figure 8.4: Time duration vs. κ diagram for all 74 classifiers and the 590 segmental features in AAFE v.2.0.5 in independent training and testing (*AS-Kraetzer2010SPIE-SF-singleClass-summary*)

Like in the examples for the global features presented above, the direct comparison between the 10-fold stratified cross-validation and the two-set training and testing (figures 8.3 and 8.4) shows faster but less accurate performance in the two-set training and testing.

Table 8.2: Identification of the five classifiers for each IH algorithm with the lowest distance from the optimal performance, 10-fold stratified cross-validation as well as two-set training and testing using segmental features (the value in brackets identify the distance value q_{new})

10-fold stratified cross-validation			
	A_{S1}	A_{W1}	A_{S3}
Best classifier	*lazy.Ibk* (0.085)	*meta.RandomSubSpace* (0.057)	*trees.J48graft* (0.445)
2nd	*trees.RandomForest* (0.092)	*rules.PART* (0.057)	*meta.OrdinalClassClassifier* (0.445)
3rd	*trees.RandomTree* (0.092)	*trees.REPTree* (0.064)	*trees.J48* (0.445)
4th	*meta.RandomCommittee* (0.106)	*meta.Bagging* (0.071)	*meta.END* (0.453)
5th	*rules.OneR* (0.318)	*trees.J48graft* (0.071)	*meta.RandomCommittee* (0.460)

two-set training and testing			
	A_{S1}	A_{W1}	A_{S3}
Best classifier	*rules.OneR* (0.706)	*functions.Logistic* (0.148)	*trees.J48graft* (0.552)
2nd	*bayes.NaiveBayesUpdateable* (0.706)	*meta.MultiClassClassifier* (0.148)	*meta.RandomCommittee* (0.552)
3rd	*trees.RandomForest* (0.706)	*meta.RandomSubSpace* (0.156)	*meta.OrdinalClassClassifier* (0.559)
4th	*meta.Dagging* (0.706)	*rules.PART* (0.559)	*trees.J48* (0.559)
5th	*lazy.Ibk* (0.706)	*functions.SMO* (0.559)	*meta.END* (0.559)

Table 8.2 identifies, similar to table 8.2 above, the best five classifiers for each algorithm and the segmental features in AAFE v.2.0.5. Since the classification times for the segmental features are – due to their much larger and more numerous feature vectors – higher than for the global features used, the achieved distances from the optimum are much higher, even if similar κ values are. In the case of A_{S1} and the training and testing with independent sets, no classifier achieves κ values significantly larger than 0. Therefore, all corresponding benchmarking values are close to $\sqrt{2}$ or higher (see figure 8.4 where these values are placed on the x-axis of the diagram).

Application of the proposed benchmarking scheme for microphone forensics

In this application scenario specific discussion on the classifier benchmarking methodology introduced above, one instantiation of the transfer function $SFQ_{Classifier}()$ is presented for the microphone forensics application scenario. This example uses the AAFE v.2.0.5 set of 590 segmental features and 10-fold stratified cross-validation as summarised in experimental setup Mic-01. For the experiment on the *RS4_Rode* set, containing recordings made with four identical microphones, a timeout of 60 hours (216,000s) is defined. Based on experimental setup Mic-01, for all 74 supervised classifiers implemented in WEKA (v.3.6.1) a benchmarking value is computed as:

$$q_{new_{classifier}} = SFQ_{classifier}(q_{new}, SF_{all}, timeout = 60h, cv, RS4\_Rode) \tag{8.8}$$

The index for the transfer function is the classifier (with its parametrisation, which is here kept for all classifiers to the default setting). The metric used by the transfer function is q_{new} (as introduced with its Euclidean distance measurement above). Additional parameters supplied to the transfer function are the feature set used (here SF_{all}, all 590 segmental features of AAFE v.2.0.5), the $timeout$ boundary (here 60h) and the test method for the classifier (here cv; 10-fold stratified cross-validation on the audio test set *RS4_Rode*).
Table 8.3 summarises the top 20 of the classifiers for the experiment Mic-01, ordered by q_{new}.

Table 8.3: Ranking by quality of the best 20 classifiers for experiment Mic-01

Ranking	Classifier	κ	Avg. runtime (s)	q_{new}
Best	*meta.RotationForest*	0.678	14012.96	0.23226358
2nd	*meta.MultiClassClassifier*	0.634	1491.73	0.25884715
3rd	*meta.RandomSubSpace*	0.617	2018.56	0.2709025
4th	*meta.EnsembleSelection*	0.649	33735.9	0.27165665
5th	*functions.Logistic*	0.616	1726.59	0.27158783
6th	*functions.SimpleLogistic*	0.627	20974.18	0.27254168
7th	*meta.Bagging*	0.611	3123.13	0.27525448
8th	*meta.END*	0.611	9900.37	0.27696737
9th	*functions.SMO*	0.605	2289.37	0.27940771
10th	*meta.ClassificationViaRegression*	0.598	3137.65	0.28444245
11th	*meta.Dagging*	0.557	169.91	0.3132488
12th	*meta.RandomCommittee*	0.534	164.61	0.3295122
13th	*rules.PART*	0.650	6994.95	0.24854449
14th	*meta.Decorate*	0.694	51399.65	0.27409998
15th	*trees.J48graft*	0.646	2083.03	0.25040867
16th	*trees.RandomForest*	0.641	129.4	0.25385169
17th	*rules.JRip*	0.637	4163.79	0.25704143
18th	*trees.J48*	0.634	1832.22	0.25887058
19th	*trees.SimpleCart*	0.629	2061.17	0.26242338
20th	*trees.REPTree*	0.619	395.28	0.26941079

In comparison to a classifier selection strategy that is based only on the detection performance (i.e. κ value; cf. table 6.11 in section 6.1.3), here classifiers that achieve a good detection performance with a short runtime are rising higher in the ranking. A good example is the classifier *meta.MultiClassClassifier* which, due to its rather short runtime is ranked 2nd best in table 8.3, while achieving only rank 8 in table 6.11. This ranking seems to be closer to the requirements for practical implementation of security mechanisms and is therefore considered within this thesis to be a first step in the direction of a benchmarking scheme.

Feature selection down to the 20 most significant features is applied (as performed in section 6.1.4), the classifier quality q_{new} improves for the experimental setup *Mic-RS4_Rode-Best20Features-only* in average by 0.094, due to the much stronger decrease of the *runtime* of the evaluations in comparison to the drop in the κ values.

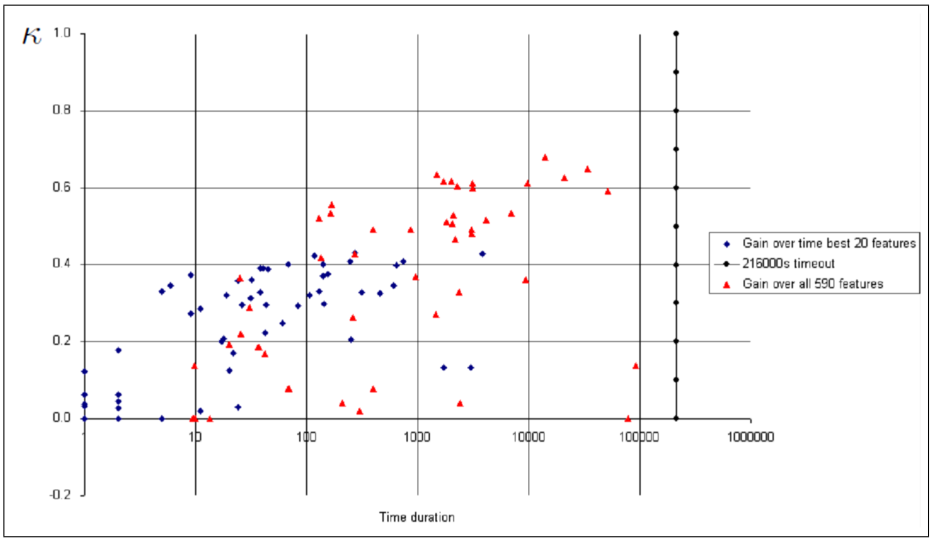


Figure 8.5: Time duration vs. κ diagram for *RS4_Rode* 10-fold cross-validation with all 74 classifiers in WEKA v.3.6.1; red: all 590 segmental features (averages over *R01*-*R10*); blue: best 20 segmental features only (*R01*; experimental setup *Mic-RS4_Rode-Best20Features-only*)

Figure 8.5 shows the time duration vs. κ diagram for the complete set of 590 features (*RS4_Rode* average for 10-fold cross-validations in *R01-R10*; experimental setup *Mic-01*) and the best 20 features (*RS4_Rode* average for 10-fold cross-validation in *R01*; *Mic-RS4_Rode-Best20Features-only*). The results for the 590 dimensional set are marked in red, the results for the 20 dimensional set in blue. It can be clearly seen that the centre of gravity for the 20 dimensional feature space is closer to the optimum (the point $(0,1)$, i.e. the upper left corner in this diagram), while the 590 dimensional set achieves higher κ (at the cost of dramatically increased costs in computation times).

Future directions on benchmarking

In the theoretical considerations at the begin of this section, the drawbacks of the benchmarking metric q_{new} introduced here have been summarised as: the dependence on the computation power of the machine the benchmark is run on, the dependence on the used training and test sets (size, representativeness) as well as the lacking support for the direct identification of fast mechanisms.
Regarding the first of these points, one possible solution might be the introduction of a descriptor for the machine speed. For the second point, future work will have to be invested into investigations on optimal (regarding detection performance as well as throughput considerations) training (and test) set sizes. This is due to the fact that the training and test set composition and size directly influence both indicators used in the benchmarking approach introduced above. For fair benchmarking, the system performance would have to be measured under predefined set sizes. Furthermore, the definition of standard set sizes would be an important step to make the benchmarking results comparable between different application scenarios. Regarding the third of the drawbacks mentioned above, the solution alternative would be to move away from single-figure benchmarking to more complex multi-figure benchmarking, as introduced by Lang in [Lang07] for audio watermarking benchmarking.

Further research in this field should include considerations on benchmarking strategies for classifiers which are not implemented into the same run-time environment. Furthermore, other performance indicators, besides the detection performance and the runtime, should be integrated in the benchmarking scheme. Good first candidates for such integration would be indicators on plausibility and forensic compliance, like the ones considered within this thesis. For each of these indicators, a corresponding transfer function would be required for the integration into a benchmarking metric. For some indicators, like e.g. the Daubert-compliance, the design of such a transfer function will be a hard research challenge.

8.2.2 Information fusion

Information fusion has the goal to determine the best set of experts (or expert systems) in a given problem domain and devise an appropriate function that can optimally combine the decisions rendered by the individual experts (cf. [Ross06], [Kuncheva04]). Information fusion is a science that is also known by other names. The most prominent are: evidence pooling, ensemble methods, expert combination or classifier combination. Within the focus of this thesis there exist two different approaches to information fusion that have to be distinguished: the signal processing approach (i.e. practical considerations motivated by application examples of fusion in other research fields) and the decision theory approach. In the following two subsections initial considerations are presented or both approaches to show how the work presented within this thesis could benefit from information fusion.

Fusion in the SPR process and post classification fusion

One of the research fields where applied signal processing and statistical pattern recognition methods are extensively employed in combination with information fusion is the fields of biometrics. Having emerged in the 1960s and early 1970s (see e.g. Atal [Atal74] for biometric speaker verification/identification), biometrics achieved a level of maturity from which other (similar) pattern recognition problems could benefit. The idea of a knowledge transfer from biometrics to the application scenarios considered within this thesis is not a new one. One early attempt on the transfer of the fusion concept into steganalysis is presented by Kharrazi et al. [Kharrazi06].

From the numerous fusion concepts, known in biometrics for pattern recognition processes, two different ones shall be briefly considered here. The first one was presented by **Sanderson and Paliwal** [Sanderson02] in 2002 and is used in a simplified version in the considerations by Kharrazi et al. [Kharrazi06]. It uses a model which distinguishes into pre-classification and post-classification information fusion. Pre-classification fusion refers in this context to combining information prior to the application of any classifier (or matching algorithm), while in post-classification the information is combined after the decisions of the classifiers have been obtained.
In [Kharrazi06] Kharrazi et al. limit themselves to three different operations: First, the transfer of the aforementioned fusion model from biometrics to the image steganalysis domain, second the practical evaluation of the impact of a fusion of three different steganalysers (two universal, one algorithm specific) on the classification performance for two image steganography techniques (with the fusion results presented ranging from worse than the best individual technique to better than all techniques – depending on the tested algorithm), and third, the question whether the fusion of steganalysers might lead to the same classification results as a truly 'global' universal steganalyser (trained with a training set containing samples for all available steganographic techniques). In the test results of the third presented evaluation a reduction of the classification result by choosing an universal or fused detector instead a of specific one is seen (results achieved are between 3 and 7% worse), while at the same time it is indicated that the scalability of the steganalysis increases (complexity decreases).

The second fusion approach to be mentioned here is the one used by **Ross, Nandakumar, and Jain** [Ross06]. In this approach a five level fusion model (sensor-, feature-, match-, rank- and decision-level fusion) is employed. This latter fusion approach by Ross et al. has a finer granularity and incorporates (amongst other benefits) a more appropriate model for dynamic classifier selection. This fusion approach is formalised and visualised for the field of biometric research by Oermann et al. in [Oermann06]. It is enhanced here by adding the corresponding signal processing operations between the fusion levels (see figure 8.6).

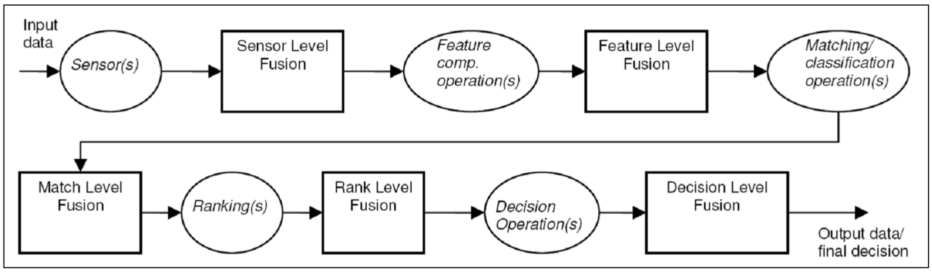


Figure 8.6: Overview of the five signal processing steps and the five different fusion levels (based on Oermann et al. [Oermann06])

The five fusion levels [Ross06] (sensor-, feature-, match-, rank-, and decision-level fusion) used in this model with their corresponding signal processing operations (signal acquisition at sensor level, feature computation, classification/matching, ranking and decision making) can be summarised as follows (note: detailed examples on how fusion on these levels is performed in biometrics are presented by Ross et al. [Ross06]):

- **Sensor-level fusion:** Entails the consolidation of evidence presented by multiple sources of raw data before they become subject to feature extraction.

- **Feature-level fusion:** Involves consolidating the evidence presented by different feature sets of the same source. The following requirements for feature-level fusion have to be considered: The features have to be related (i.e. really belong to one source), they must be of the same type (e.g. a variable length and a fixed length feature set should not be joined), and should be considered under the knowledge of the course-of-dimensionality problem (i.e. the number of samples for a training set has to reflect the number of features).

- **Match-level fusion (also known as score-level fusion or classification-level fusion):** A fusion on matching score level implies a consolidation of matching scores (respectively classification results) gained from separate comparisons/classification of reference data and test data for each source. Because fusion on this level is the most commonly applied technique it biometrics it incorporates a separate chapter in the work of Ross et al. [Ross06].
- **Rank-level fusion:** Which is of importance especially for identification problems, has the goal to consolidate the ranked outputs of individual classification systems in order to derive a consensus rank for each identity known.
- **Decision-level fusion:** if a fusion is applied on decision-level then each subsystem draws completely autonomous decisions, which are then combined. The operator for this decision combination could be Boolean functions (like logical AND or OR), (weighted) majority voting, Bayesian decisions, etc.

Within the work on this thesis, this second fusion approach has been used for initial considerations on the integration of information fusion into the introduced general-purpose statistical pattern recognition (SPR) approach for audio forensics. The work of Ross et al. in [Ross06] is favoured in these considerations over the work by Sanderson and Paliwal [Sanderson02] applied by previously by Kharrazi et al. in [Kharrazi06] for steganalysis, because its finer granularity allows for a more flexible process integration.

Application of information fusion for audio steganalysis

In [Kraetzer10] a limited-scale experiment on **sensor-level fusion** is performed. In the performed experiment, a second audio input (a second sensor) is emulated by de-noising the original. The fusion operation is then the subtraction of both sensor signals. Thus, the emulated sensor fusion outputs the 'noise' component of the original audio signal after content removal (for details see [Kraetzer10]).
The results presented in [Kraetzer10] (using all of WEKAs 74 classifiers, global as well as segmental features computed by AAFE v.2.0.5 and 10-fold stratified cross-validation on *aats389* as well as training with *aats389* and testing with *testset24*) show, for the three investigated information hiding algorithms (A_{S1}, A_{S3} and A_{W1}), similar detection performances as the evaluations without the (content removal performing) sensor-level fusion. The basic idea of using such a content removal for content influence elimination in audio steganalysis seems to be a very promising one.

In [Kraetzer09a] match-level and decision-level fusions are performed for the introduced statistical pattern recognition (SPR) based audio steganalysis approach. The results presented in this paper show a small increase in detection performance by the performed match-level fusion, while the decision-level fusion-results are identical to those without fusion (for details see [Kraetzer09a]). This work is extended in [Kraetzer10] by further considerations on match-level fusion for segmental features and a new decision-level fusion for global features. While the results of the on match-level fusion for segmental features confirm the small improvement in the detection performance already demonstrated in [Kraetzer09a]. The new decision-level fusion for global features alone shows no positive impact, but if both are combined in an additional mixed-level fusion, the detection performance is again increased by a small amount (for details see [Kraetzer10]).

Résumé for the fusion considerations on audio steganalysis is: The sensor-level fusion presented in [Kraetzer10] seems to be a good way of reducing the contend dependability in steganalysis. After the (rather crude) sensor-level fusion performed in the tests, the content influence is reduced dramatically, while the subsequent classifications show similar classification accuracies. Here further and more sophisticated methods for content removal/de-noising should be tested in future research. Since the sensor-level fusion tested here is basically a content removal or anonymisation of the data, the benefit of this operation would be that it would pave the road for the outsourcing of the computationally burdensome tasks of feature extraction, model generation and classification to a third party (e.g. a commercial service with more computational power).

The practical results achieved in fusion with the introduced statistical pattern recognition (SPR) based audio steganalysis approach show little to no improvement in the actual detection performance. This is consistent with the results on fusion obtained by Kharazzi et al. in [Kharrazi06], where the fusion accuracies are also only in few test cases better than the non-fused results. Nevertheless, future research should be invested into the application of fusion techniques in steganalysis, because it might allow for an efficient combination of application-specific detectors into a scheme for universal steganalysis.

Application of information fusion for microphone forensics

In [Kraetzer09b] we focussed on the benefits of post-classification fusion for the discussed instantiation of the general-purpose audio forensics approach for microphone forensics. In that paper we introduced a set of potential fusion operators for match-, rank- and decision-level fusion.
The table 8.4 summarises the investigation results presented in [Kraetzer09b] on the influence of information fusion to the **detection performance** of the used statistical pattern recognition (SPR) based microphone forensics approach.

Table 8.4: Average κ values for microphone forensics for selected classifiers without and with fusion (averages over the results computed for the ten recording locations *R01-R10*; for a precise description of the used experimental setups see [Kraetzer09b])

		$RS1$ (average κ over all 10 recording locations)	$RS2$ (average κ over all 10 recording locations)
without fusion	*SimpleLogistics*	0.784	0.727
	J48	0.829	0.763
Match-level fusion		0.691	0.614
Rank-levelfusion		1.000	1.000
Decision-level fusion		0.967	0.966

Summarising the investigation results for all fusions tested in [Kraetzer09b], it can be said that the rank-level fusion performed there shows the strongest impact on the detection performance achieved. It increases the performance for both test sets to $\kappa = 1$. The decision-level fusion performs not as good as the rank-level fusion, but with an average $\kappa > 0.96$ still better than the corresponding classifications without fusion. The performed match-level fusion reduces the classification performance in comparison to the single classifiers, which can be, at least partially, contributed to the modelling of the fusion operator as a majority voting decision with equal weights. Here, in case of different statements for the two used classifiers, the fusion decision deadlocked, which illustrated very well the need for appropriate fusion operator designs.

Besides the detection performance considerations, in [Kraetzer09b] also **confidence estimation** functions for post-classification fusion are introduced. An accuracy of 100% or a κ value of 1 for a fusion-based system does not tell much about its applicability in real world investigations. Due to the implications of the Daubert standard, not only its detection performance would have to be known, but also a confidence has to be determined as a measure how far the fusion decision is away from the complex decision boundary. In [Kraetzer09b] some prototypical confidence functions are introduced for the tested fusion operators. Our results in that paper show that for test cases with similar detection performances different system confidences could be achieved.

Résumé for the fusion considerations on microphone forensics is: In the exemplary selected investigations on post-classification fusion presented in detail in [Kraetzer09b] and summarised here, the following three facts are shown: a) a fusion operation designed to suite the classification problem at hand can indeed increase the detection performance achieved in microphone forensics; b) the estimation of the confidence/trust put into a fusion decision is a non-trivial problem which still holds a lot of potential for future research; c) post-classification fusion can dramatically increase the complexity of the classification process in microphone forensics, which opens opportunities for future research on low complexity fusion operators and the balancing between classification accuracy increase on one hand and complexity scaling on the other hand.

Alternative directions on information fusion

The progress made in the field of information fusion in recent years is very well illustrated by text-books on this topic (like e.g. [Kuncheva04]) as well as high-quality survey-papers in well established journals (like e.g. [Atrey10]). One prominent current trend in this field is the application of the Dempster-Shafer evidence theory [Shafer76] for fusion considerations (cf. [Atrey10]). This theory generalises Bayesian theory to relax its restriction on mutually exclusive hypotheses. Furthermore, it uses belief and plausibility values to represent the evidence and their corresponding uncertainty. In [Fontani11] Fontani et al. demonstrate the application of the Dempster-Shafer evidence theory for the combination of manipulation detection tools in media forensics. The analysis of this technique leads to the realisation that it relies in the modelling of the plausibility values strongly on assumptions for the certainty of the systems that are to be combined in the fusion. In [Fontani11] the authors simply state on that fact: "[...] *we also assume to have some information (possibly image dependent) about tools reliability (for instance such an information could derive from experimental evidence)*". In the opinion of the author this assumption would have to be substantiated by the existence of suitable benchmarking strategies for such tools, imposing to the benchmarking considerations presented in section 8.2.1 of this thesis an additional requirement, extending the detection performance driven considerations.

8.3 Possible directions for future work

Two major points for future work on the general-purpose statistical pattern recognition (SPR) based audio forensics approach can be identified: On one hand, this would be the **transfer to other application scenarios**, to investigate to which extent it can be adapted to research fields like e.g. voice recognition, speech recognition, speaker recognition, audio coder verification, gunshot characterisation, audio signal quality verification, etc. On the other hand, it would have to be accompanied by **detailed juristic analyses**. Since the author possesses absolutely no legal training, all legal considerations made within this thesis (especially on the Daubert standard) are therefore layman's interpretation of freely available material, which are made to the best of the author's knowledge. The intention behind the work on forensic compliance performed here is to derive a performance metric for research in the field of audio forensics. If the content of this thesis is intended to be used in any legal proceedings, the reader must consult appropriate legal counsel for the corresponding jurisdiction.

Accompanying these two rather broad important points, in the following more specific potential extensions are presented. For reasons of accessibility they are structured into remarks on possible extensions of the introduced approach, potential alternatives in methodology and concepts in audio steganalysis and microphone forensics as well as benchmarking considerations:

Possible extensions of the introduced approach

- **Pre-processing:** The usage of more sophisticated (instead of simple windowing) pre-processing methods should increase **the distance between the patterns to be detected and the background signal**. In the cases of the two application scenarios considered within this thesis an inverted form of noise removal (as briefly addressed in section 6.2.4 with silence detection) might be used to deduce the influence of the cover or recorded signal in the classification.

- **Feature extraction:**

  - The features are the main issue in any pattern recognition approach. The introduced approach might strongly benefit from an **extension of the feature space** by **additional context insensitive features** (e.g. from [Peeters04] or [Mathieu10]). Also, **adapted versions of existing features** might be used to improve the detection performance on specific classes, e.g. if the features are shaped to detect specific characteristics of a certain pattern like LSB-features that acknowledge the fact that the steganographic algorithm Publimark only embeds its payload into every third LSB.

  - Additionally, the implementation of **higher-level (semantical) content analysis**[80] into audio features would assumedly increase the detection performance significantly. The downside is that the time and computation power required in feature extraction might also show a strong increase.

- **Feature selection:** An extension of the work on feature selection performed within this thesis would be the usage of methods form **analytical statistics** (e.g. variance analysis or factor analysis) to determine the relationships within the feature space as well as the exact influence of each feature to the patterns detected. This would for example help to improve the context model for the recording process discussed in section 2.3.2.

- **Classification:** Within this thesis the classifier **parameter optimisation** has been omitted. It is obvious that such parameter optimisation (e.g. by grid search through the parameter space) will be able to improve the detection performance (at the cost of immense computation power spent for the determination of the optimal parameter settings). Also the design of new, **specialised classification algorithms** for the discussed classification problems might be a promising research field.

- **Evaluation:**
  - The results achieved for audio steganalysis as well as microphone forensics would have to be verified in future work with **larger evaluation sets**. These extended evaluations would have to verify the statements on the detection performance and (re-)address the questions of **sufficient model** sizes and **scaling behaviour** of the introduced approach.
  - The **integration of content analysis** should be considered to enable an automatised shift from cover independent to cover type dependent training and testing.
  - The number of evaluations performed for **plausibility investigations** is limited within this thesis to a practically feasible number. The idea is here to establish the concept for both application scenarios and reserve more detailed analyses with more common audio signal post-processing operations for future work.
  - The determination of the **precise error rates** (as required by the Daubert standard) and the estimation of the achievable security levels for the introduced audio forensics methods require extensive benchmarking (see the statements on benchmarking of SPR-based security mechanisms below).

Alternatives in methodology and concepts in audio steganalysis

- **Signal preparation:** For future research on this approach, the consideration of further information hiding algorithms as well as further embedding parameters (e.g. embedding strength) might be important to enhance the generalisability of the statements made.

- **Creation of multi-level detection schemes:** In trying to get closer to Daubert compliance, an important step would be to combine networks of two-class classifiers with multi-class classifiers, e.g. to use the first ones to identify the embedding method/domain (e.g. LSB) and the latter ones to identify the actual tool that was used to embed the data (i.e. perform multi-class steganalysis with the goal of algorithm identification). This could prepare the systems for an attempt to extract (and potentially decrypt) the message embedded and thereby create the binding to the case at hand that would be required for an admission as evidence in a law case.

Alternatives in methodology and concepts in microphone forensics

- **Signal preparation:** For future research on this approach, the consideration of larger recordings sets, created under strongly varying influence factors, would be important to enhance the generalisability of the statements made.

[80] Like provided by the Freesound audio analysis API hosted by the Universitat Pompeu Fabra in Barcelona, Spain. See: http://www.freesound.org/docs/api/analysis_index.html

- **Extension of the context model for the recording process:** The context model for the recording process introduced in section 2.3.2has to be extended, fine-tuned in future work and accompanied by attack models for non-malicious or malicious (anti-forensic) modifications to the recorded audio signal.

- **Recording integrity verification:** It should be established how reliable means for audio signal integrity verification can be achieved using the SPR-based approach proposed in this thesis in **combination (fusion) with other approaches** from the state-of-the-art in this field (especially the ENF-based work from [Rodríguez10]). The idea is that this combination with other approaches for editing operation detection (see [REW11]) can be used to perform a 'scene-analysis' (i.e. scene segmentation) on edited audio signals prior to authentication of the sources used for the scenes in a potential composition.

Benchmarking of statistical pattern recognition (SPR) based security mechanisms

- The basis for any advanced benchmarking considerations would be a complete mathematical formalisation of the media forensic process considered. Some of the work performed within this thesis (e.g. the context model for the recording process, see section 2.3.2) are already backed by such a formalisation. These parts could be used as initial points for the development of a general formalisation and process model.

- The next step to be considered in future work on benchmarking of statistical pattern recognition (SPR) based audio security mechanisms should be the extension of the considerations on benchmarking metrics into a fully developed and fair benchmarking scheme for practical application. Such benchmarking would be a necessity basis for large-scale usage in communication security. Similar fields of research on communication security already have benchmarking methods in place. Two examples for such initiatives to be mentioned here are the National Institute of Standards and Technology's (NIST, see http://www.nist.gov/itl/biometrics/index.cfm) work on Biometrics as well as the European Institute for Computer Antivirus Research (EICAR, see e.g. www.eicar.org/) with its work on malware detection. Alternatives for performance metrics, to be used in this context, should extend the work described in sections 4.1.4 and 8.2.1, e.g. by incorporating different weights for the false positive and false negative error rates into a fair benchmarking scheme, because their proportion is in most cases equivalent to the security (false negative ratio) and usability (false positive ratio) of the approach.

9 Appendix A: Audio Features used in AAFE

This appendix gives detailed descriptions on the features used in the composition of the three different versions of the AMSL Audio Feature Extractor (AAFE) used throughout the thesis. An overview of the features and the mapping to the three major releases AAFE (versions 1.0.3, 1.0.4 and 2.0.5) is performed in section 4.1.1.
The mathematical feature descriptions are based on the formalisation of audio signal representations given in section 2.3.1.

9.1 Segmental features

All **segmental features** are computed for a window with index i (and channel k) of the sampled, quantised and windowed digital audio signal S_i^k (see the formalisation of digital audio signals in section 2.3.1).

Empirical variance (time-domain feature):
In [Dittmann07] the empirical variance (sf_{ev}) is described as the statistical dispersion of the samples in a window of size w of the audio signal. The feature indicates how the sample values are spread around the arithmetic mean. It is computed for the samples $s_{i,j}^k$ of a given window with index i and a specific channel k of the sampled, quantised and windowed digital audio signal S_i^k as:

$$sf_{ev} = \frac{1}{n}\sum_{j=1}^{w}\left(s_{i,j}^k - sf_{mean}\right)^2 \tag{9.1}$$

Where w is the number of samples in the window, j is the sample-in-the-frame index ($j \in \mathbb{N}; 1 \leq j \leq w$) for the i-th window in this channel of the stream. The feature sf_{mean} represents the arithmetic mean of the samples in this window (see below).
The feature is included in the feature vectors of AAFE versions 1.0.3 and 1.0.4.

Covariance (time-domain feature):
In [Dittmann07] the covariance is described as follows: "*In probability theory and statistics, covariance is the measure of how much two random variables vary together (as distinct from variance, which measures how much a single variable varies). If two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be above its expected value too), then the covariance between the two variables will be positive. On the other hand, if when one of them is above its expected value, the other variable tends to be below its expected value, then the covariance between the two variables will be negative.*"
Here the covariance feature sf_{cv} is computed for $\frac{w}{2}$ pairs of samples from one window S_i^k. For the expected value, the arithmetic mean of all values in the window (sf_{mean}) is used:

$$sf_{cv} = \frac{1}{n}\sum_{j=1}^{w}\left(s_{i,j}^k - sf_{mean}\right)\cdot\left(s_{i,j-1}^k - sf_{mean}\right) \tag{9.2}$$

The feature is included in the feature vectors of AAFE versions 1.0.3 and 1.0.4.

Entropy (time-domain feature):

In information theory, the Shannon entropy or information entropy is a measure of the uncertainty associated with the development of a random variable. The feature $sf_{entropy}$ is computed in [Dittmann07] as follows: For each window S_i^k of the sampled, quantised and windowed digital audio signal a histogram H of the occurring values is generated. In the next step, the value for each position in the histogram is divided by the window size resulting in a new histogram entry $\bar{H}_{histpos}$. Equation 9.3 shows the computation of $sf_{entropy}$ from the $\bar{H}_{histpos}$ values.

$$sf_{entropy} = - \sum_{\bar{H}_{histpos}} \bar{H}_{histpos} \log_2 \left(\bar{H}_{histpos} \right) \tag{9.3}$$

The feature is included in this form in the feature vectors of AAFE versions 1.0.3, 1.0.4 and 2.0.5.

LSB ratio (time-domain feature):

The feature sf_{LSBrat} describes the ratio between the value '0' and value '1' least significant bit (LSB) values of a sample within a window of the audio material. It is computed for a window S_i^k with samples $s_{i,j}^k$, where j is the sample-in-the-frame index ($j \in \mathbb{N}$; $1 \leq j \leq w$), i is the frame-index and k is the channel, as:

$$sf_{LSBrat} = \frac{\sum_{j=1}^{w} \mathsf{LSB}_0(s_{i,j}^k)}{\sum_{j=1}^{w} \mathsf{LSB}_1(s_{i,j}^k)} \tag{9.4}$$

The functions $\mathsf{LSB}_0(s_{i,j}^k)$ and $\mathsf{LSB}_1(s_{i,j}^k)$ are used to count the occurrences of '0's and '1's in the LSB values. They are defined as:

$$\mathsf{LSB}_0(s_{i,j}^k) = \begin{cases} 1 & if \quad \mathsf{LSB}(s_{i,j}^k) = 0 \\ 0 & if \quad \mathsf{LSB}(s_{i,j}^k) = 1 \end{cases} \tag{9.5}$$

$$\mathsf{LSB}_1(s_{i,j}^k) = \begin{cases} 0 & if \quad \mathsf{LSB}(s_{i,j}^k) = 0 \\ 1 & if \quad \mathsf{LSB}(s_{i,j}^k) = 1 \end{cases} \tag{9.6}$$

The helper function LSB() in equations 9.5 and 9.6 returns the LSB-value of a sample with sample-in-the-frame index j (frame-index i and channel k).

The feature is included in this form in the feature vectors of AAFE versions 1.0.3, 1.0.4 and 2.0.5.

LSB flipping rate (time-domain feature):

The feature $sf_{LSBflip}$ counts the number of flips of the least significant bit (LSB) values within the window S_i^k (window size w, samples $s_{i,j}^k$ with sample-in-the-frame index j ($j \in \mathbb{N}$; $1 \leq j \leq w$), frame-index i and channel k). It is computed as:

$$sf_{LSBflip} = \sum_{j=1}^{w-1} \left| \mathsf{LSB}(s_{i,j}^k) - \mathsf{LSB}(s_{i,j+1}^k) \right| \tag{9.7}$$

Here, LSB() is again the helper function already used in equations 9.5 and 9.6 above.

The feature is included in this form in the feature vectors of AAFE versions 1.0.3, 1.0.4 and 2.0.5.

Mean of samples in time domain (time-domain feature):

The feature sf_{mean} computes the arithmetic average for the samples in the window. It is computed for the window S_i^k (window size w, samples $s_{i,j}^k$ with sample-in-the-frame index j ($j \in \mathbb{N}$; $1 \leq j \leq w$), frame-index i and channel k) as:

$$sf_{mean} = \frac{1}{w} \sum_{j=1}^{w} s_{i,j}^k \tag{9.8}$$

The feature is included in this form in the feature vectors of AAFE versions 1.0.3, 1.0.4 and 2.0.5.

Median of samples in time domain (time-domain feature):
In case that the window size w is odd, the feature sf_{median} returns the value of the $\frac{w+1}{2}$-th element of an ordered array of size w, containing all samples of the window S_i^k. In case w is even, the arithmetic mean of the values of two elements in the middle of the array (indices $\frac{w}{2}$ and $\frac{w+1}{2}$) is returned.

$$sf_{median} = \begin{cases} \text{arraySort}_{\frac{w+1}{2}}(S_i^k) & if \quad \text{length of } w \text{ is odd} \\ \frac{1}{2}\left(\text{arraySort}_{\frac{w}{2}}(S_i^k) + \text{arraySort}_{\frac{w+1}{2}}(S_i^k)\right) & if \quad \text{length of } w \text{ is even} \end{cases} \tag{9.9}$$

The helper function arraySort$_{pointer}$() sorts the samples $s_{i,j}^k$ in the window S_i^k by their value and allows access to the entries in the sorted array by using the index $pointer$.
The feature is included in this form in the feature vectors of AAFE versions 1.0.3, 1.0.4 and 2.0.5.

Zero-cross-rate (time-domain feature):
The feature $sf_{zero\_cross\_rate}$ computes the zero-crossing-rate for a window S_i^k (window size w, samples $s_{i,j}^k$ with sample-in-the-frame index j ($j \in \mathbb{N}$; $1 \leq j \leq w$), frame-index i and channel k) of the audio signal, i.e. how often the signed time-domain value changes from values above '0' to values below '0'. It is computed as:

$$sf_{median} = \frac{1}{2}\sum_{j=2}^{w}\left|\text{sgn}(s_{i,j}^k) - \text{sgn}(s_{i,j-1}^k)\right| \tag{9.10}$$

$$sgn(s_{i,j}^k) = \begin{cases} 1 & if \quad s_{i,j}^k \geq 0 \\ -1 & if \quad s_{i,j}^k < 0 \end{cases} \tag{9.11}$$

The helper function sgn() identifies the sign of an element and maps it to '1' for positive values and '-1' for negative values.
This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Energy (time-domain feature):
The feature sf_{energy} estimates the energy in a window of the sampled, quantised and windowed digital audio signal S_i^k (window size w, samples $s_{i,j}^k$ with sample-in-the-frame index j ($j \in \mathbb{N}$; $1 \leq j \leq w$), frame-index i and channel k) as:

$$sf_{energy} = \sqrt{\frac{\sum_{j=1}^{w}(s_{i,j}^k)^2}{w}} \tag{9.12}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

RMS amplitude (time-domain feature):
The feature $sf_{RMS\_amplitude}$ root mean square (RMS) amplitude is a more traditional description of the energy in an audio window of the sampled, quantised and windowed digital audio signal S_i^k (window size w, samples $s_{i,j}^k$ with sample-in-the-frame index j ($j \in \mathbb{N}$; $1 \leq j \leq w$) than sf_{energy}. Both features are strongly correlated. The RMS amplitude is computed as:

$$sf_{RMS\_amplitude} = \frac{1}{w}\sqrt{\sum_{j=1}^{w}(s_{i,j}^k)^2} \tag{9.13}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Formants (frequency-domain features):
[Kraetzer08a] describes the set of formant features as follows: "*Distinguishing frequency components of human speech and of singing are called formants. The information required by humans to distinguish between vowels can be represented purely quantitatively by the frequency characteristics of the vowel sounds* [Duncan88] *such as provided by the so called formants. Most often the two first formants are enough to disambiguate the vowel* [Duncan88]. *These two formants are primarily determined by the*

position of the tongue.
With the measurements of the spectral energy in the frequency bands representing the first two formants for each English vowel as well as the so called singer formant, a first frequency domain based feature set is added to the AAFE v.1.0.4."
The feature $sf_{formant\_*}$ computes the arithmetic mean of all frequency bins in the frequency range $[low, up]$ describing the formant. Table 9.1 shows the frequency bounds for the 11 formants (five times two for the five vowels plus the singer formant) computed here. Based on these frequency bounds, the AAFE features $sf_{formant\_*}$ are computed window-wise as:

$$sf_{formant\_*} = \frac{1}{\text{num\_coef}(low, up, coef)} \sum_{low \leq h \leq up} y_{i,h}^{k} \tag{9.14}$$

In equation 9.14 $y_{i,h}^{k}$ (channel k, frame-index i and coefficient-in-the-frame index h ($h \in \mathbb{N}$; $1 \leq h \leq coef$)) are the frequency coefficients in the spectrogram Y_i^k resulting from a FFT on an audio window of the sampled, quantised and windowed digital audio signal S_i^k and low and up are the lower and upper frequency bound of the considered formant (see table 9.1). The function num_coef() computes the number of frequency bins in the range $[low, up]$ by considering the FFT size $coef$.

Table 9.1: Frequency ranges for the vowel and singer formants (taken from [Kraetzer08a])

Formant	**Lower frequency bound** (low) [Hz]	**Upper frequency bound** (up) [Hz]
$sf_{formant\_A1}$	800	1200
$sf_{formant\_A2}$	1300	1500
$sf_{formant\_E1}$	400	600
$sf_{formant\_E2}$	2200	2600
$sf_{formant\_I1}$	200	400
$sf_{formant\_I2}$	3000	3500
$sf_{formant\_O1}$	400	600
$sf_{formant\_O2}$	900	1100
$sf_{formant\_U1}$	200	400
$sf_{formant\_U2}$	700	900
$sf_{formant\_Singer}$	2800	3400

These features are included in this form only in the feature vectors of AAFE versions 1.0.4 and 2.0.5.

Bark scale spectrogram (frequency-domain features):
The Bark scale is computed by Zwicker [Zwicker61] by the projection of the normal spectrogram Y_i^k onto a non-linear frequency domain scale consisting of 24 frequency bands spectrogram B_i^k (unit [Bark]). The 24 features sf_{Bark_1} to sf_{Bark_24} are each computed as the arithmetic mean of all frequency bins $y_{i,h}^{k}$ of Y_i^k between two consecutive Bark in B_i^k. The computation of this mean is equivalent to the one performed for the formants as described above (for details see [Kraetzer08a]).
These features are included in this form only in the feature vectors of AAFE version 1.0.4. In AAFE version 2.0.5 the Bark scale spectrogram is replaced by a higher resolution linear-scale spectrogram.

Spectral centroid (frequency-domain feature):
The feature $sf_{sp\_centroid}$ describes the mass center of a spectrum. Here the method from [Eisenberg08] for the computation of this feature is applied:

$$sf_{sp\_centroid} = \frac{\sum_{h}^{coef} y_{i,h}^{k} \cdot \text{fr}(h)}{\sum_{h}^{coef} y_{i,h}^{k}} \tag{9.15}$$

In this equation, fr(h) is the frequency belonging to coefficient-in-the-frame index h ($h \in \mathbb{N}$; $1 \leq h \leq coef$).
This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral flux (frequency-domain feature):
The feature sf_{sp_flux} (a.k.a. spectral fluctuation) represents the variation of the spectrum between consecutive windows. It is computed as:

$$sf_{sp\_flux} = \left| \sum_{h=1}^{coef} y_{i,h}^{k} - y_{i-1,h}^{k} \right| \tag{9.16}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral roll-off (frequency-domain feature):
The spectral roll-off $sf_{sp\_rolloff}$ is described (e.g. in [Smaragdis09] and [Lerch08]) as the frequency band at which the cumulated spectral energy exceeds a specified threshold. In [Lerch08] this threshold is defined as 85% of the overall spectral energy, in other publications (e.g. [Smaragdis09]) this value is defined as 90% or even 95% of the overall spectral energy. Here, 85% of $\sum_{h=1}^{coef} y_{i,h}^{k}$ is used for each window with window index i in channel k.
This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral bandwidth (frequency-domain feature):
The spectral bandwidth sf_{sp_bw} describes the width of the frequency range in which contains 90% of the overall spectral energy ($\sum_{h=1}^{coef} y_{i,h}^{k}$) of a window with index i in channel k. The upper and lower boundaries of this range are iteratively determined to exclude 5% of the spectral energy on each of the two ends of the spectrum.
This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral smoothness (frequency-domain feature):
The spectral smoothness $sf_{sp\_smoothness}$ describes the continuity (or discontinuity) of the spectrum of a window of the audio signal. It is computed here as:

$$sf_{sp\_smoothness} = \sum_{h=2}^{coef-1} \left| y_{i,h}^{k} - \frac{y_{i,h-1}^{k} + y_{i,h}^{k} + y_{i,h+1}^{k}}{3} \right| \tag{9.17}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral irregularity (frequency-domain feature):
Like $sf_{sp\_smoothness}$, the spectral irregularity $sf_{sp\_irregularity}$ is describing the continuity of the spectrum. Here the definition of [Luck08] is used for the computation:

$$sf_{sp\_irregularity} = \frac{\sum_{h=2}^{coef} \left(y_{i,h}^{k} - y_{i,h-1}^{k} \right)^{2}}{sf_{RMS\_amplitude}} \tag{9.18}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Spectral entropy (frequency-domain feature):
For the computation of the entropy of the spectrum ($sf_{sp\_entropy}$) the description given in [Luck08] us used:

$$sf_{sp\_entropy} = -\frac{\sum_{h=1}^{coef} y_{i,h}^{k} \cdot \ln \quad y_{i,h}^{k}}{\ln \quad coef} \tag{9.19}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Base frequency (frequency-domain feature):
The feature $sf_{sp\_base\_freq}$ computes the base frequency for the signal within one window of the audio material. Here an extremely simple approach to base frequency determination is used:

$$sf_{sp\_base\_freq} = \max(\{y_{i,1}^{k}, \cdots, y_{i,coef}^{k}\}) \tag{9.20}$$

This feature is included in this form only in the feature vectors of AAFE version 2.0.5.

Linear spectrogram (frequency-domain features):

The majority of the frequency-domain features computed in AAFE v.2.0.5 are presenting a 512-bin[81] linear spectrogram sf_{spec_*} (features $sf_{spec\_1}, \cdots, sf_{spec\_512}$). The features represent the energy within the corresponding frequency bin:

$$sf_{spec\_h} = y^k_{i,h} \tag{9.21}$$

This set of features is included in this form only in the feature vectors of AAFE version 2.0.5. It replaces the lower resolution Bark scale spectrogram from AAFE version 1.0.4.

Mel-Frequency Cepstral Coefficients (MFCCs) (cepstral-domain features)

An extremely detailed description of the computation of the MFCC computation for AAFE version 1.0.3 is given in [Dittmann07].

Originally, the cepstrum was defined by Bogert, Healy and Tukey in 1963 [Bogert63]. The term itself is an anagram of the word spectrum. Basically, a cepstrum is the result of taking the Fourier transform (FT) or short-time Fourier analysis [Allen77] of the decibel spectrum as if it were a signal. The cepstrum can be interpreted as information about the rate of power change in different spectrum bands. The cepstrum K^k_i for is computed here from a window S^k_i as:

$$K^k_i = \text{FT}\left(\log\left(\text{FT}\left(S^k_i\right)\right)\right) \tag{9.22}$$

For the work described in this thesis, only a real cepstrum [Dittmann07] is considered since the feature of invertability provided by a complex cepstrum is not required in the analysis performed. A modified version of the cepstrum, the Mel-cepstrum is considered by [McEachern94] as an excellent feature vector for representing the human voice and musical signals. This consideration led to the idea in [Dittmann07] to transfer corresponding Mel-cepstrum based features to speech steganalysis.

For the computation of the Mel-cepstrum, the spectrum is usually first transformed using the Mel frequency bands. The result of this transformation is called the Mel-spectrum and is used as the input of the second FT computing the Mel-cepstrum represented by the Mel frequency cepstral coefficients (MFCCs) which are used as sf_{MFCC_1} to sf_{MFCC_28} in AAFE v.1.0.3 and 1.0.4. The complete transformation for the input time-domain signal S^k_i is described in equation 9.23.

$$\{sf_{MFCC\_1}, \cdots, sf_{MFCC\_28}\} = \text{FT}\left(\text{MelScaleTransformation}\left(\text{FT}\left(S^k_i\right)\right)\right) \tag{9.23}$$

In equation 9.23 the helper function MelScaleTransformation() performs the required Mel-scale transform of the spectrum.

Figure 9.1 shows the complete transformation procedure for a FFT based Mel-cepstrum computation as introduced by Thrasyvoulou and Benton in [Thrasyvoulou03] in 2003. Alternative approaches found in literature use linear prediction based based Mel-cepstrum computation. A detailed discussion about which transformation should be used in which case is given by Thrasyvoulou et al. [Thrasyvoulou03]. From these discussions it is obvious that the FT based approach suffices the means of audio feature extraction for forensic purposes pursuit in this thesis.

[81] To be more precise, the feature extractor returns $\frac{coef}{2}$ frequency coefficients for a window of w samples, but since all considerations within this thesis have been made with the default window size of 1024 samples, this feature subset contains for all evaluations 512 features.

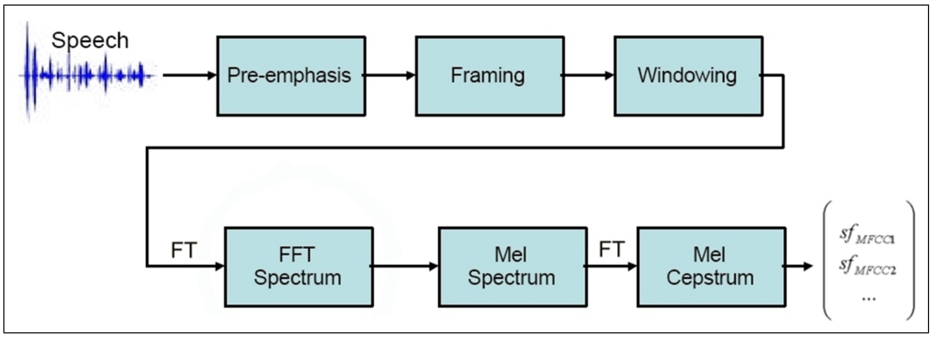


Figure 9.1: FFT-based Mel-cepstrum computation adapted from Thrasyvoulou et al. [Thrasyvoulou03]

For the AAFE versions 1.0.3 and 1.0.4 the pre-emphasis, framing and windowing, Fourier transform, filter-bank based Mel-transform and second Fourier transform are designed on basis of the concepts of Thrasyvoulou et al. [Thrasyvoulou03]. For precise implementation description for the MFCC computation used in AAFE version 1.0.3 and detailed information about alternatives in the computation of the cepstrum the author refers to [Dittmann07].

In AAFE version 2.0.5 the computation of the MFCCs is done by using the corresponding MATLAB function mfcc() from the Auditory Toolbox maintained by Malcolm Slaney[82]. This function returns for the provided audio data 13 MFCCs. To denote the differences in the MFCCs computed by the different versions of AAFE, the MFCCs extracted by the MATLAB-based version 2.0.5 are denoted as sf_{MFCC_1} to sf_{MFCC_13}.

Filtered Mel-Frequency Cepstral Coefficients (FMFCCs) (cepstral-domain features)
In [Kraetzer07a] a modification of the Mel-cepstral based signal analysis is introduced for AAFE version 1.0.3. It is based on the application scenario of VoIP steganalysis and the basic assumption that a VoIP communication consists mostly of speech communication between human speakers. This, in conjunction with the knowledge about the frequency limitations of human speech (see e.g. Fastl et al. [Zwicker90]), led to the idea of removing the speech relevant frequency bands (the spectrum components between 200 and 6819.59 Hz) in the spectral representation of a signal before computing the cepstrum.
This procedure returning the FMFCCs (filtered Mel frequency cepstral coefficients; sf_{FMFCC_1} to sf_{FMFCC_28}) is shown in figure 9.2.
This procedure, which enhances the computation described by equation 9.23 by a filter step, returns the FMFCCs (sf_{FMFCC_1} to sf_{FMFCC_28} in AAFE versions 1.0.3 and 1.0.4) and is expressed in equation 9.24.

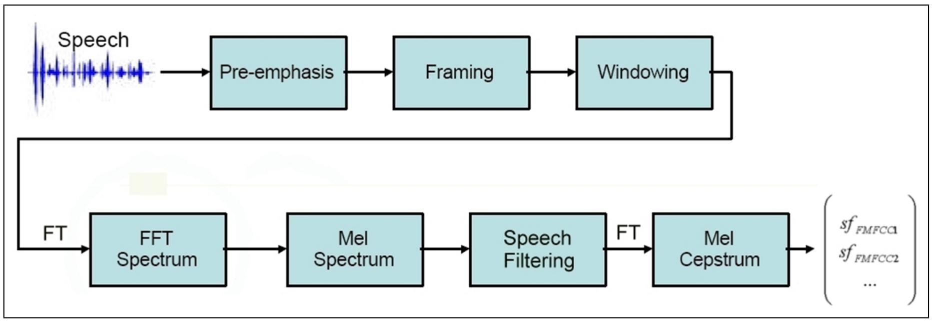


Figure 9.2: Computation of the FMFCCs (adapted from [Kraetzer07a])

$$\{sf_{FMFCC\_1}, \cdots, sf_{FMFCC\_28}\} = \mathrm{FT}\left(\mathrm{BandFilter}\left(\mathrm{MelScaleTransformation}\left(\mathrm{FT}\left(S_i^k\right)\right)\right)\right) \quad \boxed{9.24}$$

[82] See https://engineering.purdue.edu/~malcolm/interval/1998-010/

In equation 9.24 the helper function BandFilter() performs the required removal of the audio content between 200 and 6819.59 Hz.
For AAFE version 2.0.5 the original C/C++ implementation from version 1.0.3 is replaced by a MATLAB implementation. For this purpose the function mfcc() from the Auditory Toolbox maintained by Malcolm Slaney[83] is enhanced by a filtering operation in the frequency domain representation of the signal, prior to the Mel-scale filtering and the second Fourier transform. This computation returns for the provided audio data 13 FMFCCs, denoted sf_{FMFCC_1} to sf_{FMFCC_13}.

Second-order derivative MFCCs (cepstral-domain features)

In AAFE version 2.0.5, in addition to the normal MFCCs, also the second-order derivative MFCCs introduced in [Liu09] are computed (sf_{d2MFCC_1} to sf_{d2MFCC_13}):

$$\{sf_{d2MFCC\_1},\cdots,sf_{d2MFCC\_13}\} = \mathsf{FT}\left(\mathsf{MelScaleTransformation}\left(\mathsf{FT}\left(\frac{d^2}{d^2}\frac{S_i^k}{n\mathrm{T}}\right)\right)\right) \tag{9.25}$$

For a detailed description of the computation process the author refers to [Liu09].
This set of features is included in this form only in the feature vectors of AAFE version 2.0.5.

Second-order derivative FMFCCs (cepstral-domain features)

Equivalent to the second-order derivative MFCCs described above, also second-order derivative FMFCC are computed by AAFE version 2.0.5:

$$\{sf_{d2FMFCC\_1},\cdots,sf_{d2FMFCC\_13}\} = \mathsf{FT}\left(\mathsf{BandFilter}\left(\mathsf{MelScaleTransformation}\left(\mathsf{FT}\left(\frac{d^2}{d^2}\frac{S_i^k}{n\mathrm{T}}\right)\right)\right)\right) \tag{9.26}$$

This set of features is included in this form only in the feature vectors of AAFE version 2.0.5.

9.2 Global features

The **global features** introduced for AAFE v.2.0.5 consists of the total zero-crossing-rate gf_{zcr_total} for the complete audio signal and arithmetic averages of all considered windows for the following segmental features: the eight AAFE version 2.0.5 time-domain features ($sf_{entropy}$, sf_{LSBrat}, $sf_{LSBflip}$, sf_{mean}, sf_{median}, $sf_{zero\_cross\_rate}$, sf_{energy}, $sf_{RMS\_amplitude}$) as well as the spectral centroid, spectral rolloff, spectral bandwidth, spectral smoothness, spectral irregularity, spectral entropy, base frequency and spectral flux ($sf_{sp\_centroid}$, $sf_{sp\_rolloff}$, sf_{sp_bw}, $sf_{sp\_smoothness}$, $sf_{sp\_irregularity}$, $sf_{sp\_entropy}$, $sf_{sp\_base\_freq}$ and sf_{sp_flux}).
The elements of the global features vector are therefore denoted as: gf_{zcr_total}, $gf_{entropy\_AVE}$, gf_{LSBrat_AVE}, $gf_{LSBflip\_AVE}$, gf_{mean_AVE}, gf_{median_AVE}, $gf_{zero\_cross\_rate\_AVE}$, gf_{energy_AVE}, $gf_{RMS\_amplitude\_AVE}$, $gf_{sp\_centroid\_AVE}$, $gf_{sp\_rolloff\_AVE}$, $gf_{sp\_bw\_AVE}$, $gf_{sp\_smoothness\_AVE}$, as well as $gf_{sp\_irregularity\_AVE}$, $gf_{sp\_entropy\_AVE}$, $gf_{sp\_base\_freq\_AVE}$ and $gf_{sp\_flux\_AVE}$.
This set of features is included in this form only in the feature vectors of AAFE version 2.0.5.

[83] See https://engineering.purdue.edu/~malcolm/interval/1998-010/

10

Appendix B: Experimental Setups for the Audio Steganalysis Application Scenario

This appendix summarises the experimental setups used for the practical investigations on the audio steganalysis application scenario in chapter 5 and section 8.2.1.

Originally, this summary on the training and test data, classifiers and features used in every experimental setup[84], which is given in table 10.1, was part of section 4.2. It has been placed in this appendix to improve the accessibility of the core chapters of this thesis.

Table 10.1: Summary of the experimental setups for audio steganalysis

Setup	**Training material**	**Test material**	**Classifiers / clusterers**	**Feature set(s)**
AS-Kraetzer2007SPIE-summary	• Original material and material marked with all nine IH algorithms (default param.) • 64 feature vectors per file of the set *aats389*	• Original material and material marked with all nine IH algorithms (default param.) • 16 feature vectors per file (disjunctive with training material) of the set *aats389*	*libSVM* default parametrisation	all 63 segm. feat. of AAFE v.1.0.3 vs. only time-domain features vs. only time-domain features and the FMFCCs of AAFE v.1.0.3
AS-Kraetzer2007SPIE-longfile	• Original material and material marked with all nine IH algorithms (default param.) • 2200 feature vectors per file of the set *longfile*	• Original material and material marked with all nine IH algorithms (default param.) • 400 feature vectors per file (disjunctive with training material) of the set *longfile*		all 63 segm. feat. of AAFE v.1.0.3
AS-Kraetzer2010SPIE-GF-singleClass-summary	• Original material and material marked with A_{S1}, A_{S3} and A_{W1} • 10x stratified cross-validation vs. separate training and test set • 40 feature vectors per file of the set *aats389* for training and cross-validation • 40 feature vectors per file of the set *testset24* for testing		all 74 in WEKA (v.3.6.1) implemented supervised classifiers in default parametrisations	all 17 global feat. of AAFE v.2.0.5
AS-Kraetzer2010SPIE-SF-singleClass-summary				all 590 segm. feat. of AAFE v.2.0.5

Continued on Next Page...

[84] The experimental setups used in chapters 5, 6 and 8 are identified in those chapters by underlined and italic font setting (e.g. *Mic-01*).

Table 10.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
AS-Kraetzer2010SPIE-SF/GF-singleClass	• Original material and material marked with A_{S1}, A_{S3} and A_{W1} • Post-processing by MP3 conversion and denoising • 40 feature vectors per file of the set *aats389* for training • 40 feature vectors per file of the set *testset24* for testing		best 5 (for all feature set / IH algorithm combination) out all 74 in WEKA (v.3.6.1) implemented supervised classifiers in default parametrisations	all 590 segm. feat. vs. all 17 global features of AAFE v.2.0.5
AS-D-SF-scaling	• Original material and material marked with A_{S1}, A_{S3} and A_{W1} • 10x stratified cross-validation • 4, 8, 12, 16, 20, 24, 28, and 32 feature vectors per file of the set *aats389*		*weka.classifiers.**: • *trees.J48* • *functions.Logistic* • *rules.OneR* • *trees.DecisionStump* • *trees.RandomTree* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5
AS-D-SF-multiClass	• Original material and material marked with A_{S1}, A_{S3} and A_{W1} • 18 feature vectors per file of the set *aats389* for percentage split (66%/34%)		all 590 segm. feat. vs. all 17 global features of AAFE v.2.0.5	
AS-Kraetzer2007IH-scaling	• Original material and material marked with all nine IH algorithms (default param.) • 16, 64 and 256 feature vectors per file of the set *aats389*	• Original material and material marked with all nine IH algorithms (default param.) • 4, 16 and 64 feature vectors per file (disjunctive with training material) of the set *aats389*	*libSVM* default parametrisation	only time domain features and FMFCCs of AAFE v.1.0.3
AS-Kraetzer2007IH-scaling_VoIP	• Original material and material marked with all nine IH algorithms (default param.) • 400 and 2200 feature vectors per file of the set *longfile*	• Original material and material marked with all nine IH algorithms (default param.) • 400 and 2200 feature vectors per file (disjunctive with training material) of the set *longfile*		
AS-Kraetzer2007IH-unmarked	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file of the set *aats389*	• Original (i.e. unmarked) material • 64 feature vectors per file of the set *aats389*		
AS-Kraetzer2007IH	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file of the set *aats389*	• Original material and material marked with all nine IH algorithms (default param.) • 64 feature vectors per file (disjunctive with training material) of the set *aats389*		
AS-Kraetzer2007IH-CrossEval	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file of the set *aats389* • One model trained for every IH algorithm	• Original material and material marked with all nine IH algorithms (default param.) • 64 feature vectors per file (disjunctive with training material) of the set *aats389* • Testing of all models against any test data set		

Continued on Next Page...

Table 10.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
AS-Kraetzer2008SPIE-VoIP	• Original material and material marked with all nine IH algorithms (default param.) • 15000 feature vectors per file of the set *ahss1*	• Original material and material marked with all nine IH algorithms (default param.) • 1200 feature vectors per file (disjunctive with training material) of the set *ahss1*		all 98 segm. feat. of AAFE v.1.0.4
AS-Kraetzer2008SPIE-ContentDependent	• Original material and material marked with all nine IH algorithms (default param.) • speech data(*ahss1*) vs. multi-genre data (*aats389*)	• Original material and material marked with all nine IH algorithms (default param.) • speech data(*ahss1*) vs. multi-genre data (*aats389*) (disjunctive with training material)		
AS-Kraetzer2008SPIE-ContentInDependent	• Original material and material marked with all nine IH algorithms (default param.) • Percentage split (80% : 20%) • speech data(*ahss1*) vs. multi-genre data (*ref10*)			
AS-Kraetzer2008SPIE-ClassifierComparison	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file of the set *aats389*	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file (disjunctive with training material) of the set *aats389*	*libSVM* and *weka.classifiers.**: • *bayes.NaiveBayes* • *functions.MLRM* (all in default parametrisation)	
AS-Feature-Selection-GF	• Original material and material marked with all nine IH algorithms (default param.) • 256 feature vectors per file of the set *aats389*		none	feature selection on the 17 global features of AAFE v.2.0.5 (evaluators used: *ChiSquaredAttributeEval*, *FilteredAttributeEval*, *InfoGainAttributeEval*, *OneRAttributeEval*, *SymmetricalUncertAttributeEval* from WEKA version 3.6.1; search method used: *Ranker*
AS-Feature-Selection-SF				feature selection on the 590 segm. feat. of AAFE v.2.0.5 (evaluators used: *ChiSquaredAttributeEval*, *FilteredAttributeEval*, *InfoGainAttributeEval*, *OneRAttributeEval*, *SymmetricalUncertAttributeEval* from WEKA version 3.6.1; search method used: *Ranker*

Continued on Next Page. . .

Table 10.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
AS-Feature-Selection-SF/GF-PCA				PCA on the 17 global features vs. 590 segm. feat. of AAFE v.2.0.5
AS-KraetzerSPIE2009-KeyScen	• Original material and material marked A_{S1}, A_{S3} and A_{W1} (default param.) • 200 feature vectors per file of the set *aats389_Part1* • Two key scenarios ('fixed key' and 'variable key')	• Original material and material marked with all nine IH algorithms (default param.) • 200 feature vectors per file of the set *testset24*	*libSVM* and *weka.classifiers.**: • *bayes.NaiveBayes* • *functions.SimpleLogistics* • *lazy.ADABoost* • *trees.J48* (all in default parametrisation)	all 98 segm. feat. of AAFE v.1.0.4 and the global 19 features of *AudioRS*

11

Appendix C: Experimental Setups for the Microphone Forensics Application Scenario

This appendix summarises the experimental setups used for the practical investigations on the microphone forensics application scenario in chapter 6 and section 8.2.1.

Originally, this summary on the training and test data, classifiers and features used in every experimental setup[85], which is given in table 11.1, was part of section 4.3. It has been placed in this appendix to improve the accessibility of the core chapters of this thesis.

Table 11.1: Summary of the experimental setups for microphone forensics

Setup	**Training material**	**Test material**	**Classifiers / clusterers**	**Feature set(s)**
Mic-01	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) and each of the 10 rooms (*R01*, *R02*, ..., *R10*) • 10x stratified cross-validation		all 74 in WEKA (v.3.6.1) implemented supervised classifiers in default parametrisations	all 590 segm. feat. of AAFE v.2.0.5
Mic-02	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) and each of the 10 rooms (*R01*, *R02*, ..., *R10*) • 10x stratified cross-validation		all 74 in WEKA (v.3.6.1) implemented supervised classifiers in default parametrisations	all 590 segm. feat. of AAFE v.2.0.5
Mic-Kraetzer2007ACM	• *RS1* and original files from *ref10* (10 reference files) • 100, 200, 300, 400, 500, 600, 700 and 800 feature vectors per file for all of the 4 microphones (M_1, M_2, M_3, M_4) and each of the 10 rooms (*R01*, *R02*, ..., *R10*) plus original files • 10-fold stratified cross-validation and percentage split (66% to 34%) for *NaiveBayes* and classes to clusters evaluation for *SimpleKMeans*		*weka.classifiers.bayes.-NaiveBayes*, *weka.clusterers.SimpleKMeans* in default parametrisations	all 63 segm. feat. of AAFE v.1.0.3
Mic-03	• *RS4_Beyer* (10 reference files; set *ref10*) • 100, 200, 300, 400, 500, 600, 700 and 800 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) and each of the 10 rooms (*R01*, *R02*, ..., *R10*) • 10x stratified cross-validation		*weka.classifiers.**: • *bayes.NaiveBayes* • *functions.Logistic* • *meta.RandomSubSpace* • *trees.RandomForest* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5

Continued on Next Page. . .

[85] The experimental setups used in chapters 5, 6 and 8 are identified in those chapters by underlined and italic font setting (e.g. *Mic-01*).

Table 11.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
Mic-Feature-Selection	• *RS4_Rode* and *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 2x4 microphones (M_{16}, M_{17}, M_{18}, M_{19} and M_{20}, M_{21}, M_{22}, M_{23}) and each of the 10 rooms (*R01*, *R02*, ..., *R10*)		none	feature selection on the 590 segm. feat. of AAFE v.2.0.5 (evaluators used: *ChiSquaredAttributeEval*, *FilteredAttributeEval*, *InfoGainAttributeEval*, *OneRAttributeEval*, *SymmetricalUncertAttributeEval* from WEKA version 3.6.1; search method used: *Ranker*
Mic-RS4_Rode-Best20Features-only	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in $R01$ • 10x stratified cross-validation		all 74 in WEKA implemented supervised classifiers in default parametrisations	best 20 (see section 6.1.4)
Mic-Composition-1	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in $R01$	• "original half" 50 feature vectors (disjunctive with training material) from M_{22} in $R01$ • "impostor half" 50 feature vectors (disjunctive with training material) from M_{22} in $R06$	*weka.classifiers.**: • *bayes.NaiveBayes* • *functions.SMO* • *meta.RandomCommittee* • *trees.RandomForest* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5
Mic-Composition-2		• "original half" 50 feature vectors (disjunctive with training material) from M_{22} in $R01$ • "impostor half" 50 feature vectors (disjunctive with training material) from M_{23} in $R01$		
Mic-Composition-3		• "original half" 50 feature vectors (disjunctive with training material) from M_{22} in $R01$ • "impostor half" 50 feature vectors (disjunctive with training material) from M_8 (RS2 in $R01$)		
Mic-Composition-4		• "first half" 50 feature vectors (disjunctive with training material) from M_2 (RS2 in $R01$) • "second half" 50 feature vectors (disjunctive with training material) from M_3 (RS2 in $R01$)		
Mic-Denoise-RS4_Rode	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01*	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* after de-noising		

Continued on Next Page. . .

Table 11.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
Mic-Denoise-RS4_Beyer	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01*	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01* after de-noising		
Mic-MP3conversion-RS4_Rode	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01*	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* after MP3 conversion		
Mic-MP3conversion-RS4_Beyer	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01*	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01* after MP3 conversion		
Mic-Normalisation-RS4_Rode	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01*	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* after normalisation		
Mic-Normalisation-RS4_Beyer	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01*	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{22}) in *R01* after normalisation		
Mic-MultiProcessing-RS4_Rode	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01*	• *RS4_Rode* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* after de-noising, normalisation and MP3 conversion		
Mic-MultiProcessing-RS4_Beyer	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01*	• *RS4_Beyer* (10 reference files; set *ref10*) • 200 feature vectors per file (disjunctive with training material) for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01* after de-noising, normalisation and MP3 conversion		

Continued on Next Page...

Table 11.1 – Continued

Setup	Training material	Test material	Classifiers / clusterers	Feature set(s)
Mic-Content-Selectivity-01	• *RS4_Rode* 1 reference file silence vs. 2 ref. files speech from *ref10* • 200 vs. 800 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* • 10-fold stratified cross-validation		*weka.classifiers.*:* • *meta.dagging* • *functions.Logistic* • *trees.RandomForest* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5
Mic-Content-Selectivity-02	• *RS4_Beyer* 1 reference file silence vs. 2 ref. files speech from *ref10* • 200 vs. 800 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01* • 10-fold stratified cross-validation			
Mic-Content-Independency-01	• *RS4_Rode* 1 reference file silence vs. 2 ref. files speech from *ref10* • 200 vs. 800 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* • training on speech and testing on silence and vice versa			
Mic-Content-Independency-02	• *RS4_Beyer* 1 reference file silence vs. 2 ref. files speech from *ref10* • 200 vs. 800 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01* • training on speech and testing on silence and vice versa			
Mic-Kraetzer2009ACM-single-classifier	• *RS2* all 10 reference files from *ref10* • 200 feature vectors per file for all of the 7 microphones (M_2, M_5, M_3, M_6, M_7, M_8, M_9) in *R01* • Percentage split 80% / 20%		*weka.classifiers.*:* • *trees.J48* • *functions.simple-Logistics* (all in default parametrisation)	all 98 segmental features of AAFE v.1.0.4
Mic-BKD2009	• *RS2*, content: *ref10* without the two speech signals • Pre-processing: windowing ($coef = 2048$ vs. $coef = 256$), silence detection (silence thresholds $thresh$ tested: 0.01, 0.025, 0.05, 0.1, 0.2, 0.225, 0.25, 0.5 and 1 for $coef = 256$ and 0.01, 0.025, 0.05, 0.1, 0.25, 0.35, 0.4, 0.5, and 1 for $coef = 2048$) for all of the 7 microphones (M_2, M_5, M_3, M_6, M_7, M_8, M_9) in *R01* • Percentage split 80% / 20%		*weka.classifiers.*:* • *bayes.NaiveBayes* • *functions.SMO* • *functions.simple-Logistics* • *trees.J48* • *lazy.IB1* • *lazy.IBk* (all in default parametrisation)	spectrogram with $coef = 256$ vs. $coef = 2048$
Mic-Orientation_Impact_-RS7	• *RS4_Beyer ref2* references • 200 feature vectors per file and orientation for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R06*	• *RS7 ref2* references • 200 feature vectors per file for M_{22} in *R06* • Eight orientations with 45° offset in the xy-plane	*weka.classifiers.meta.-RandomSubSpace* in default parametrisation	all 590 segm. feat. of AAFE v.2.0.5
Mic-Orientation_Impact_-RS8		• *RS8 ref2* references • 200 feature vectors per file for M_{22} in *R06* • Two orientations with 180° offset in the yz-plane		
Mic-Mounting_Impact_-RS9	• *RS4_Beyer ref2* references • 200 feature vectors per file and orientation for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R06*	• *RS9 ref2* references • 200 feature vectors per file for M_{22} in *R06* • Eight different mountings/fixings for the microphone		
Mic-Clustering-RodeR01	• *RS4_Rode* all 10 reference files from *ref10* • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* • Classes to clusters evaluation		all 8 clustering algorithms in WEKA v.3.6.1 in default parametrisations	all 590 segm. feat. of AAFE v.2.0.5

Continued on Next Page. . .

Table 11.1 – Continued

Setup	**Training material**	**Test material**	**Classifiers / clusterers**	**Feature set(s)**
Mic-Clustering-BeyerR01	• *RS4_Beyer* all 10 reference files from *ref10* • 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01* • Classes to clusters evaluation			
Mic-Clustering-RodeR01-selectedfeatures	• *RS4_Rode* all 10 reference files from *ref10* • 200 feature vectors per file for all of the 4 microphones (M_{16}, M_{17}, M_{18}, M_{19}) in *R01* • Classes to clusters evaluation			$sf_{d2FMFCC\_1}$, $sf_{d2FMFCC\_2}$, $sf_{d2FMFCC\_13}$, $sf_{d2FMFCC\_10}$, $sf_{d2FMFCC\_3}$, $sf_{d2FMFCC\_5}$, $sf_{d2FMFCC\_4}$, $sf_{d2FMFCC\_11}$, $sf_{d2FMFCC\_12}$, $sf_{d2FMFCC\_9}$, $sf_{d2FMFCC\_6}$, $sf_{d2FMFCC\_8}$, $sf_{d2FMFCC\_7}$, sf_{FMFCC_3}, sf_{FMFCC_12}, sf_{spec_11}, $sf_{RMS\_amplitude}$, sf_{FMFCC_10}, sf_{FMFCC_5}, sf_{FMFCC_1}
Mic-Room-Classification-RS4-Selections	• parts of *RS4_Rode* and *RS4_Beyer* all 10 reference files from *ref10* • 200 feature vectors per file for all of the 2 microphones (M_{16} and M_{20}) in all 10 recording environments • 10x stratified cross-validation		*weka.classifiers.**: • *bayes.NaiveBayes* • *functions.SMO* • *meta.Random-Committee* • *trees.Random-Forest* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5
Mic-Room-Classification-RS4-WrongRoom	• *RS4_Beyer ref10* references • Training 1: 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R01* • Training 2: 200 feature vectors per file for all of the 4 microphones (M_{20}, M_{21}, M_{22}, M_{23}) in *R06*	• *RS4_Beyer ref10* references • 200 feature vectors (disjoint with training 2) per file for M_{22} in *R06*		
Mic-SPIE2012-Double-Recording	• *RS16_ProbM01* live speech references • 200 feature vectors per file for all of the 6 microphones (M_{33}, M_{34}, M_{35}, M_{36}, M_{37} and M_{38}) in *R06*	• *RS16_ProbM01_playback* (*RS16_ProbM01* after playback with a Yamaha MSP 5 high-quality monitor speaker) • with vs without normalisation • 200 feature vectors per file for all of the 6 microphones (M_{33}, M_{34}, M_{35}, M_{36}, M_{37} and M_{38}) in *R06*	*weka.classifiers.**: • *meta.RotationForest* • *meta.MultiClass-Classifier* • *meta.RandomSub-Space* • *meta.Ensemble-Selection* • *functions.Logistic* (all in default parametrisation)	all 590 segm. feat. of AAFE v.2.0.5

12 Bibliography

Acknowledgement: Regarding the bibliography, the author wishes to express his thanks to the maintainers of all these services that make compiling `bibtex` bibliographies much easier than it was ten years ago.
For this thesis the `bibtex` export functions of the following websites have been used extensively:

- ACM Digital Library (http://dl.acm.org)
- AES E-Library (http://www.aes.org/e-lib/)
- BibSonomy (http://www.bibsonomy.org/)
- CiteULike (http://www.citeulike.org/home)
- DBLP (http://dblp.uni-trier.de/)
- DFD the Digital Forensic Database maintained by Hany Farid (http://www.cs.dartmouth.edu/~farid/dfd/index.php/topics)
- Microsoft Academic Search (http://academic.research.microsoft.com/)
- SPIE Digital Library (http://spiedigitallibrary.org/index.aspx)

Bibliography

[Allen77] J. B. Allen and L. R. Rabiner. *A Unified Approach to Short-Time Fourier Analysis and Synthesis*. Proceedings of IEEE, vol. 65(11):pp. 1558–1564, 1977. 194

[Altun05] O. Altun, G. Sharma, M. Celik, M. Sterling, and M. Bocko. *Morphological Steganalysis of Audio Signals and the Principle of Diminishing Marginal Distortions*. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. 21–24. 2005. ISBN 0-7803-8874-7. 164

[Aoki08] N. Aoki. *A Technique of Lossless Steganography for G.711 Telephony Speech*. In *Proceedings of the 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, IIH-MSP '08, pp. 608–611. IEEE Computer Society, Washington, DC, USA, 2008. ISBN 978-0-7695-3278-3. doi:10.1109/IIH-MSP.2008.122. URL http://dx.doi.org/10.1109/IIH-MSP.2008.122. 45

[Atal74] B. S. Atal. *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*. In *J. Acoust. Soc. Amer., vol. 55, no. 6*, pp. 1304–1312. Acoust. Soc. Amer., 1974. 182

[Atrey10] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli. *Multimodal fusion for multimedia analysis: a survey*. Multimedia Syst., vol. 16(6):pp. 345–379, 2010. URL http://dblp.uni-trier.de/db/journals/mms/mms16.html#AtreyHEK10. 186

[Avcibas06] I. Avcibas. *Audio steganalysis with content-independent distortion measures*. IEEE Signal Processing Letters, vol. 13(3):pp. 92–95, 2006. 49

[Bandyopadhyay08] S. K. Bandyopadhyay, D. Bhattacharyya, D. Ganguly, S. Mukherjee, and P. Das. *A tutorial review on Steganography*. In *International Conference on Contemporary Computing (IC3-2008), Noida, India, August 79*, pp. 105–114. 2008. 44

[Bebis06] G. Bebis. *Introduction to Pattern Recognition*. In *Lecture CS479/679 Pattern Recognition (Spring'06)*. Reno, NV 89557, 2006. 31

[Bellman61] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Tech. rep., Princeton University Press, 1961. 37, 109, 132

[Bender96] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. *Techniques for data hiding*. IBM Syst. J., vol. 35(3-4):pp. 313–336, 1996. ISSN 0018-8670. doi:10.1147/sj.353.0313. URL http://dx.doi.org/10.1147/sj.353.0313. 44, 45

[Bersano-Begey97] T. F. Bersano-Begey and J. M. Daida. *A Discussion on Generality and Robustness and a Framework for Fitness Set Construction in Genetic Programming to Promote Robustness*. In J. R. Koza (ed.), *Late Breaking Papers at the 1997 Genetic Programming Conference*, pp. 11–18. Stanford Bookstore, Stanford University, CA, USA, 1997. ISBN 0-18-206995-8. 7

[Bijhold07] J. Bijhold, A. Ruifrok, M. Jessen, Z. Geradts, S. Ehrhardt, and I. Alberink. *Forensic audio and Visual Evidence 2004-2007: A Review*. In *15th INTERPOL Forensic Science Symposium*. Lyon, France, 2007. 24, 51, 57, 58, 59, 167, 168

[Bishop03] M. Bishop. *Computer Security: Art and Science*. Prentice Hall, 2003. ISBN 9780201440997. URL http://books.google.de/books?id=pfdBiJNfWdMC. 2

[Blackman59] R. B. Blackman and J. W. Tukey. *Particular Pairs of Windows*. In The Measurement of Power Spectra, From the Point of View of Communications Engineering, pp. 98–99, 1959. 27

[Blum06] A. Blum. *Random projection, margins, kernels, and feature-selection*. In *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection*, SLSFS'05, pp. 52–68. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3-540-34137-4, 978-3-540-34137-6. doi:10.1007/11752790_3. URL http://dx.doi.org/10.1007/11752790_3. 35

[Bogert63] B. Bogert, M. Healy, and J. Tukey. *The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking*. In M. Rosenblatt (ed.), *Proceedings of the Symposium on Time Series Analysis*. Wiley New York, USA, 1963. 194

[Böhme04] R. Böhme and A. Westfeld. *Statistical characterisation of MP3 encoders for steganalysis*. In *Proceedings of the 2004 workshop on Multimedia and security*, MM&Sec '04, pp. 25–34. ACM, New York, NY, USA, 2004. ISBN 1-58113-854-7. doi:10.1145/1022431.1022437. URL http://doi.acm.org/10.1145/1022431.1022437. 18

[Böhme08] R. Böhme. *Improved Statistical Steganalysis using Models of Heterogeneous Cover Signals*. Ph.D. thesis, University Dresden, Germany, 2008. 41

[Bolt74] R. Bolt, F. Cooper, J. Flanagan, J. McKnight, T. Stockham, and M. Weiss. *The Executive Office Building Tape of June 20, 1972: Report on a technical investigation*. Tech. rep., Advisory Panel on White House Tapes, United States District Court for the District of Columbia, 1974. URL http://www.aes.org/aeshc/docs/forensic.audio/watergate.tapes.report.pdf. 1, 18

[Bosi03] M. Bosi and R. E. Goldberg. *Introduction to Digital Audio Coding and Standards*. Springer, 2003. ISBN 978-1-4020-7357-1. 31

[Brixen07] E. B. Brixen. *Techniques for the Authentication of Digital Audio Recordings*. In *123rd AES Convention*. Audio Eng. Soc., 2007. 51

[Brixen08a] E. B. Brixen. *ENF; Quantification of the Magnetic Field*. In *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice*. 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14412. 52, 53, 56, 59, 73

[Brixen08b] E. B. Brixen. *How to Extract the ENF from Digital Audio Recordings*. In *ACFEI National Conference*. 2008. 55, 56, 57, 58, 59, 73, 165, 168

[Buchholz09] R. Buchholz, C. Kraetzer, and J. Dittmann. *Information Hiding*. chap. Microphone Classification Using Fourier Coefficients, pp. 235–246. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-642-04430-4. doi:10.1007/978-3-642-04431-1_17. URL http://dx.doi.org/10.1007/978-3-642-04431-1_17. 79, 99, 127, 136, 143, 144, 145, 146, 148, 160

[Cachin04] C. Cachin. *An information-theoretic model for steganography*. Inf. Comput., vol. 192(1):pp. 41–56, 2004. ISSN 0890-5401. doi:10.1016/j.ic.2004.02.003. URL http://dx.doi.org/10.1016/j.ic.2004.02.003. 40

[Carletta96] J. Carletta. *Assessing Agreement on Classification Tasks: The Kappa Statistic*. Computational Linguistics, vol. 22(2):pp. 249–254, 1996. URL http://dblp.uni-trier.de/db/journals/coling/coling22.html#Carletta96. 11, 89

[Chang11] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines*. ACM Trans. Intell. Syst. Technol., vol. 2(3):pp. 27:1–27:27, 2011. ISSN 2157-6904. doi:10.1145/1961189.1961199. URL http://doi.acm.org/10.1145/1961189.1961199. 38, 95

[Cheddad10] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt. *Review: Digital image steganography: Survey and analysis of current methods*. Signal Process., vol. 90(3):pp. 727–752, 2010. ISSN 0165-1684. doi:10.1016/j.sigpro.2009.08.010. URL http://dx.doi.org/10.1016/j.sigpro.2009.08.010. 44

[Clancy12] T. Clancy. *Dead or Alive*. Penguin Books Limited, 2012. ISBN 9781405913331. URL http://books.google.de/books?id=nI_4wRzgxcQC. 2

[Cohen60] J. Cohen. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, vol. 20(1):pp. 37–46, 1960. ISSN 0013-1644. doi:10.1177/001316446002000104. URL http://epm.sagepub.com/cgi/content/refs/20/1/37. 11

[Cole04] L. Cole, D. Austin, and L. Cole. *Visual Object Recognition using Template Matching*. In *Proceedings of Australian Conference on Robotics and Automation*. 2004. 32

[Cooper08] A. J. Cooper. *The Electric Network Frequency (ENF) as an Aid to Authenticating Forensic Digital Audio Recordings an Automated Approach*. In *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice*. 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14411. 52

[Cox08] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edn., 2008. ISBN 0123725852, 9780080555805, 9780123725851. 3, 5, 77

[Craver98] S. Craver. *On Public-Key Steganography in the Presence of an Active Warden*. In *Information Hiding*, pp. 355–368. 1998. 41

[Daeid10] N. N. Daeid and M. Houck (eds.). *Interpol's Forensic Science Review*. Taylor & Francis Inc, 2010. ISBN 978-1-4398-2658-4. 24

[D'Alessandro09] B. D'Alessandro and Y. Q. Shi. *Mp3 bit rate quality detection through frequency spectrum analysis*. In *Proceedings of the 11th ACM workshop on Multimedia and security*, MM&Sec '09, pp. 57–62. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-492-8. doi:10.1145/1597817.1597828. URL http://doi.acm.org/10.1145/1597817.1597828. 18

[Davis97] D. Davis and C. Davis. *Sound System Engineering*. Elsevier Science & Technology, 2 edn., 1997. ISBN 978-0240803050. 29

[Davis12] J. H. Davis and G. Stohr. *Republicans Tampered With Court Audio in Obama Attack Ad., Blomberg Businessweek*, 2012. URL: http://www.businessweek.com/printer/articles/32414?type=bloomberg. 2

[Dean91] D. Dean. *The relevance of replay transients in the forensic examination of analogue magnetic tape recordings*. Tech. Rep. 16/1991, Scientific Research and Development Branch, Home Office, British Government, London, UK, 1991. 18

[Dittmann00] J. Dittmann. *Digitale Wasserzeichen - Grundlagen, Verfahren, Anwendungsgebiete*. Xpert.press. Springer, 2000. ISBN 978-3-540-66661-5. 3

[Dittmann05] J. Dittmann and D. Hesse. *Network based intrusion detection to detect steganographic communication channels: on the example of audio data*. In *Proc. IEEE 6th Workshop on Multimedia Signal Processing 2004*, pp. 343–346. IEEE, 2005. ISBN 0-7803-8578-0. doi:10.1109/MMSP.2004.1436563. 49

[Dittmann06] J. Dittmann and C. Kraetzer. *D.WVL.10 Audio Benchmarking Tools and Steganalysis*. Public project deliverable, ECRYPT WAVILA, 2006. 92, 93

[Dittmann07] J. Dittmann and C. Kraetzer. *D.WVL.16 Report on Watermarking Benchmarking And Steganalysis*. Tech. rep., ECRYPT, European Network of Excellence in Cryptology, 2007. 103, 189, 190, 194, 195

[Dohnal08] M. Dohnal. *Forensische Analyse von Audiosignalen zur Mikrofonerkennung*. Master's thesis, Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany, 2008. 79, 143

[Duda01] R. Duda, P. Hart, and D. Stork. *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley Interscience, New York, 2 edn., 2001. ISBN 9780471056690. URL http://books.google.de/books?id=YoxQAAAAMAAJ. 5, 31, 32, 33, 34, 36, 37

[Dufaux01] A. Dufaux. *Detection and Recognition of Impulsive Sound Signals*. Ph.D. thesis, Institute of Microtechnology, University of Neuchatel, Switzerland, 2001. 54

[Duncan88] G. Duncan and M. Jack. *Formant estimation algorithm based on pole focusing offering improved noise tolerance and feature resolution*. Radar and Signal Processing, IEEE Proceedings, vol. 135(1), 1988. 191

[Eckert11] C. Eckert. *IT-Sicherheit: Konzepte - Verfahren - Protokolle*. Oldenbourg Wissenschaftsverlag, 2011. ISBN 9783486706871. URL http://books.google.de/books?id=x5Y8psQsdnIC. 2

[Eisenberg08] G. Eisenberg. *Identifikation und Klassifikation von Musikinstrumentenklängen in monophoner und polyphoner Musik*. Cuvillier, 2008. 192

[Erfani07] Y. Erfani, M. S. Moin, and M. Parviz. *New Methods for Transparent and Accurate Echo Hiding By Using the Original Audio Cepstral Content*. In *ACIS-ICIS*, pp. 1087–1092. IEEE Computer Society, 2007. URL http://dblp.uni-trier.de/db/conf/ACISicis/ACISicis2007.html#ErfaniMP07. 44

[Eugenio04] B. D. Eugenio and M. Glass. *The Kappa Statistic: A Second Look*. Computational Linguistics, vol. 30(1):pp. 95–101, 2004. 89, 90

[Farid01] H. Farid. *Detecting Steganographic Messages in Digital Images*. Tech. Rep. TR2001-412, Deparment of Computer Science, Dartmouth College, 2001. 46

[Farina00] A. Farina. *Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique*. In *Audio Engineering Society Convention 108*. 2000. URL http://www.aes.org/e-lib/browse.cfm?elib=10211. 54

[Flexner87] S. B. Flexner (ed.). *The Random House Dictionary of the English Language*. Random House, Inc., New York, 2 edn., 1987. 3

[Fontani11] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni. *A Dempster-Shafer framework for decision fusion in image forensics*. In *Proceedings of the 2011 IEEE International Workshop on Information Forensics and Security*, WIFS '11, pp. 1–6. IEEE Computer Society, Washington, DC, USA, 2011. ISBN 978-1-4577-1017-9. doi:10.1109/WIFS.2011.6123156. URL http://dx.doi.org/10.1109/WIFS.2011.6123156. 186

[Franz00] E. Franz and A. Pfitzmann. *Steganography Secure against Cover-Stego-Attacks*. In *Proceedings of the Third International Workshop on Information Hiding*, IH '99, pp. 29–46. Springer-Verlag, London, UK, UK, 2000. ISBN 3-540-67182-X. URL http://dl.acm.org/citation.cfm?id=647596.731697. 39

[Fridrich98] J. Fridrich. *Applications of data hiding in digital images*. Tech. rep., Tutorial for the ISPACS 1998 conference, Melburne, Australia, 1998. 39

[Fridrich00] J. Fridrich, R. Du, and M. Long. *Steganalysis of LSB Encoding in Color Images*. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 2000. 46

[Fridrich01] J. Fridrich, M. Goljan, and R. Du. *Reliable detection of LSB steganography in color and grayscale images*. In *Proceedings of the 2001 workshop on Multimedia and security: new challenges*, MM&#38;Sec '01, pp. 27–30. ACM, New York, NY, USA, 2001. ISBN 1-58113-393-6. doi:10.1145/1232454.1232466. URL http://doi.acm.org/10.1145/1232454.1232466. 94

[Fridrich06] J. Fridrich and D. Soukal. *Matrix embedding for large payloads*. Trans. Info. For. Sec., vol. 1(3):pp. 390–395, 2006. ISSN 1556-6013. doi:10.1109/TIFS.2006.879281. URL http://dx.doi.org/10.1109/TIFS.2006.879281. 43

[Fridrich09] J. Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, New York, NY, USA, 1st edn., 2009. ISBN 0521190193, 9780521190190. 2, 3, 18, 19, 28, 38, 39, 40, 41, 42, 44, 45, 51, 63, 64, 66

[Fridrich11] J. Fridrich, J. Kodovský, V. Holub, and M. Goljan. *Breaking HUGO: the process discovery*. In *Proceedings of the 13th international conference on Information hiding*, IH'11, pp. 85–101. Springer-Verlag, Berlin, Heidelberg, 2011. ISBN 978-3-642-24177-2. URL http://dl.acm.org/citation.cfm?id=2042445.2042454. 47

[Friedman99] M. Friedman and A. Kandel. *Introduction to pattern recognition – Statistical, Structural, Neural and Fuzzy Logic Approaches*, vol. 32 of *Machine Perception and Artificial Intelligence*. World Scientific Pub Co., 1999. ISBN 981-02-3312-4. 31, 32

[Fulero09] S. Fulero. *Admissibility of expert testimony based on the Grisso and Gudjonsson scales in disputed confession cases*. Open Access Journal of Forensic Psychology, (1):pp. 44–55, 2009. 24

[Garcia-Romero10] D. Garcia-Romero and C. Y. Espy-Wilson. *Automatic acquisition device identification from speech recordings*. In *ICASSP*, pp. 1806–1809. IEEE, 2010. ISBN 978-1-4244-4296-6. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2010.html#Garcia-RomeroE10. 53, 54, 55, 56, 57, 59, 69, 74, 165, 166, 167

[Gianvecchio07] S. Gianvecchio and H. Wang. *Detecting covert timing channels: an entropy-based approach*. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS '07, pp. 307–316. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-703-2. doi:10.1145/1315245.1315284. URL http://doi.acm.org/10.1145/1315245.1315284. 39

[Givner-Forbes07] R. Givner-Forbes. *Steganography: Information Technology in the Service of the Jihad*. Tech. Rep. 2, A Report for the International Centre for Political Violence and Terrorism Research, Singapore, 2007. Translated from Secret

Information: Hide Secrets Inside of Pictures – unknown authorship in The Technical Mujahid. 2, 48

[Goljan09] M. Goljan, J. J. Fridrich, and T. Filler. *Large scale test of sensor fingerprint camera identification*. In E. J. Delp, J. Dittmann, N. D. Memon, and P. W. Wong (eds.), *Media Forensics and Security I, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 19, 2009, Proceedings*, vol. 7254 of *SPIE Proceedings*, p. 72540. SPIE, 2009. doi:http://dx.doi.org/10.1117/12.805701. 13, 24, 173

[Goodman07] S. E. Goodman and H. S. Lin. *Toward a Safer and More Secure Cyberspace*. Tech. rep., Committee on Improving Cybersecurity Research in the United States, Computer Science and Telecommunications Board, National Academies Press, 2007. 5

[Grigoras03] C. Grigoras. *Digital Audio Recording Analysis - The Electric Network Frequency Criterion*. Tech. rep., Diamond Cut Productions, Inc., Hibernia, NJ, 2003. 11, 18, 52, 57, 167

[Grigoras05] C. Grigoras. *Digital Audio Recording Analysis: the Electric Network Frequency (ENF) Criterion*. Speech, Language and the Law, vol. 12(1):pp. 63–76, 2005. 52, 55, 57, 167

[Grigoras07] C. Grigoras. *Application of ENF Criterion in Forensic Audio, Video, Computer and Telecommunication Analysis*. Forensic Science international, vol. 167:pp. 136–145, 2007. doi:10.1016/j.forsciint.2006.06.033. 30, 52, 69

[Grigoras09] C. Grigoras, A. Cooper, and M. Michalek. *Forensic Speech and Audio Analysis Working Group - Best Practice Guidelines for ENF Analysis in Forensic Authentication of Digital Evidence*. Tech. rep., European Network of Forensic Science Institutes (ENFSI) - Forensic Speech and Audio Analysis Working Group (FSAAWG), 2009. 11, 25, 52, 53, 57, 59, 168

[Gruhl96] D. Gruhl, A. Lu, and W. Bender. *Echo Hiding*. In *Proceedings of the First International Workshop on Information Hiding*, pp. 293–315. Springer-Verlag, London, UK, UK, 1996. ISBN 3-540-61996-8. URL http://dl.acm.org/citation.cfm?id=647594.728894. 44

[Guillon02] P. Guillon, T. Furon, and P. Duhamel. *Applied public-key steganography*. In E. J. Delp and P. W. Wong (eds.), *Security and Watermarking of Multimedia Contents IV*, vol. 4675 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, January 21-24, 2002. 41

[Gupta12] S. Gupta, S. Cho, and C. C. J. Kuo. *Current Developments and Future Trends in Audio Authentication*. IEEE MultiMedia, vol. 19(1):pp. 50–59, 2012. ISSN 1070-986X. doi:10.1109/MMUL.2011.74. URL http://dx.doi.org/10.1109/MMUL.2011.74. 52, 53, 58, 59, 168

[Gwet12] K. Gwet. *Handbook of Inter-Rater Reliability (3rd Edition): The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Advanced Analytics, LLC, 2012. ISBN 9780970806277. URL http://books.google.de/books?id=Bx_h_GZe0uAC. 89

[Hall09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. *The WEKA data mining software: an update*. SIGKDD Explor. Newsl., vol. 11(1):pp. 10–18, 2009. ISSN 1931-0145. doi:10.1145/1656274.1656278. URL http://doi.acm.org/10.1145/1656274.1656278. 11, 36, 38, 76, 83, 171

[HC-STC05] HC-STC. *Forensic Science on Trial (HC 96-I), Seventh Report of Session 2004-05*, 2005. House of Commons, Science and Technology Committee (HC-STC). 7

[Herrera-Joancomarti07] J. Herrera-Joancomarti. *Information Hiding and Steganography*. In *Conference on Cryptology and Digital Content Security*. 2007. 48

[Hopper02] N. J. Hopper, J. Langford, and L. v. Ahn. *Provably Secure Steganography*. In *Proceedings of the 22nd Annual International Cryptology Conference on Advances in Cryptology*, CRYPTO '02, pp. 77–92. Springer-Verlag, London, UK, UK, 2002. ISBN 3-540-44050-X. URL http://dl.acm.org/citation.cfm?id=646767.704303. 41

[Hughes06] G. Hughes. *On the mean accuracy of statistical pattern recognizers*. IEEE Trans. Inf. Theor., vol. 14(1):pp. 55–63, 2006. ISSN 0018-9448. doi:10.1109/TIT.1968.1054102. URL http://dx.doi.org/10.1109/TIT.1968.1054102. 37

[Jackson08] A. Jackson and J. Jackson. *Forensic Science*. Pearson Prentice Hall, 2008. ISBN 9780131998803. 20

[Jayaram11] P. Jayaram, H.R.Ranganatha, and H. Anupama. *Information hiding using audio steganography - a survey*. International Journal of Multimedia & Its Applications, vol. 3(3):pp. 86–96, 2011. ISSN 0975-5934. 44

[Johnson98] N. F. Johnson and P. A. Sallee. *Detection of Hidden Information, Covert Channels and Information Flows Applications of data hiding in digital images*. Tech. rep., John Wiley & Sons, Inc., Published Online: 15 June 2009, New York, 1998. doi:10.1002/9780470087923.hhs427. 44

[Johnson05] M. K. Johnson, S. Lyu, and H. Farid. *Steganalysis of recorded speech*. In E. J. Delp and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents*, vol. 5681 of *Proceedings of SPIE*, pp. 664–672. SPIE, 2005. URL http://dblp.uni-trier.de/db/conf/sswmc/sswmc2005.html#JohnsonLF05. 49, 50, 51

[Johnson08] N. F. Johnson and P. A. Sallee. *Detection of Hidden Information, Covert Channels and Information Flows*. In J. G. Voeller (ed.), *Wiley Handbook of Science Technology for Homeland Security*. John Wiley & Sons, New York, 2008. 44

[Katzenbeisser00] S. Katzenbeisser and F. A. Petitcolas (eds.). *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Inc., Norwood, MA, USA, 1st edn., 2000. ISBN 1580530354. 39, 43, 44, 48

[Katzenbeisser02] S. Katzenbeisser and F. A. P. Petitcolas. *Defining security in steganographic systems*. In E. J. Delp and P. W. Wong (eds.), *Security and Watermarking of Multimedia Contents IV*, vol. 4675 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, January 21-24, 2002. 41

[Ker04] A. D. Ker. *Quantitative evaluation of pairs and RS steganalysis*. In E. J. Delp and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents*, vol. 5306 of *Proceedings of SPIE*, pp. 83–97. SPIE, 2004. 42, 43, 94

[Ker07a] A. D. Ker. *Batch steganography and pooled steganalysis*. In *Proceedings of the 8th international conference on Information hiding*, IH'06, pp. 265–281. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-74123-7. URL http://dl.acm.org/citation.cfm?id=1759048.1759068. 42, 67

[Ker07b] A. D. Ker. *The ultimate steganalysis benchmark?* In *Proceedings of the 9th workshop on Multimedia & security*, MM&Sec '07, pp. 141–148. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-857-2. doi:10.1145/1288869.1288889. URL http://doi.acm.org/10.1145/1288869.1288889. 40, 41, 43

[Kerckhoffs83] A. Kerckhoffs. *La cryptographie militaire*. Journal des Sciences Militaires, pp. 161–191, 1883. 39

[Kharrazi05] M. Kharrazi, H. T. Sencar, and N. D. Memon. *Benchmarking steganographic and steganalysis techniques*. In E. J. Delp and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents*, vol. 5681 of *Proceedings of SPIE*, pp. 252–263. SPIE, 2005. URL http://dblp.uni-trier.de/db/conf/sswmc/sswmc2005.html#KharraziSM05. 47

[Kharrazi06] M. Kharrazi, H. T. Sencar, and N. D. Memon. *Improving Steganalysis by Fusion Techniques: A Case Study with Image Steganography*. vol. 4300:pp. 123–137, 2006. URL http://dblp.uni-trier.de/db/journals/tdhms/tdhms1.html#KharraziSM06. 47, 48, 50, 51, 182, 183, 184, 185

[Kiltz08] S. Kiltz, A. Lang, and J. Dittmann. *Taxonomy for Computer Security Incidents*. Premier Reference Series. Information Science Reference, 2008. ISBN 9781591409915. URL http://books.google.com/books?id=6CJ-aV9Dh-QC. 2, 3

[Kiltz09] S. Kiltz, T. Hoppe, and J. Dittmann. *A new forensic model and its application to the collection, extraction and long term storage of screen content off a memory dump*. In *Proceedings of the 16th international conference on Digital Signal Processing*, DSP'09, pp. 1135–1140. IEEE Press, Piscataway, NJ, USA, 2009. ISBN 978-1-4244-3297-4. URL http://dl.acm.org/citation.cfm?id=1700307.1700495. 17

[Klein90] D. Klein. *Foiling the Cracker: A Survey of, and Improvements to, Password Security*. In *2nd USENIX Security Workshop*, pp. 5–14. 1990. 46

[Koenig90] B. Koenig. *Authentication of Forensic Audio Recordings*. J. Audio Eng. Soc., vol. 38(1/2), 1990. 18

[Koenig09] D. S. Koenig, Bruce E.; Lacey. *Forensic Authentication of Digital Audio Recordings*. J. Audio Eng. Soc, vol. 57(9):pp. 662–695, 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=14836. 51

[Kraetzer06a] C. Kraetzer and J. Dittmann. *Früherkennung von verdeckten Kanälen in VoIP-Kommunikation*. In *Proceedings of the BSI-Workshop IT-Frühwarnsysteme*. BSI, Bonn, Germany, 2006. 103

[Kraetzer06b] C. Kraetzer, J. Dittmann, and A. Lang. *Transparency benchmarking on audio watermarks and steganography*. In E. J. Delp and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents VIII*, vol. 6072 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, 15 - 17 January, 2006. doi:10.1117/12.642805. 92, 93, 97

[Kraetzer06c] C. Kraetzer, J. Dittmann, T. Vogel, and R. Hillert. *Design and evaluation of steganography for voice-over-IP*. In *ISCAS*. IEEE, 2006. URL http://dblp.uni-trier.de/db/conf/iscas/iscas2006.html#KratzerDVH06. 103

[Kraetzer07a] C. Kraetzer and J. Dittmann. *Mel-cepstrum-based steganalysis for VoIP steganography*. In E. J. Delp and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*, pp. 650505–650505–12. San Jose, CA, USA, 28 January - 1 February, 2007. doi:10.1117/12.704040. 49, 86, 87, 92, 93, 104, 105, 107, 118, 122, 195

[Kraetzer07b] C. Kraetzer and J. Dittmann. *Pros and Cons of Mel-cepstrum Based Audio Steganalysis Using SVM Classification*. In T. Furon, F. Cayre, G. J. Dorr, and P. Bas (eds.), *Information Hiding*, vol. 4567 of *Lecture Notes in Computer Science*, pp. 359–377. Springer, 2007. ISBN 978-3-540-77369-6. URL http://dblp.uni-trier.de/db/conf/ih/ih2007.html#KratzerD07. 104, 107, 108, 110, 116

[Kraetzer07c] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang. *Digital audio forensics: a first practical evaluation on microphone and environment classification*. In *MM&Sec '07: Proceedings of the 9th workshop on Multimedia & security*, pp. 63–74. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-857-2. doi: http://doi.acm.org/10.1145/1288869.1288879. 28, 97, 98, 99, 127, 129, 131, 132, 135, 136, 138, 139, 141, 143, 157, 160, 167

[Kraetzer08a] C. Kraetzer and J. Dittmann. *Cover Signal Specific Steganalysis: the Impact of Training on the Example of two Selected Audio Steganalysis Approaches*. In E. J. Delp, P. W. Wong, J. Dittmann, and N. D. Memon (eds.), *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, January, 2008. doi:10.1117/12.766419. 86, 87, 92, 104, 108, 109, 110, 111, 113, 118, 119, 191, 192

[Kraetzer08b] C. Kraetzer and J. Dittmann. *Impact of feature selection in classification for hidden channel detection on the example of audio data hiding*. In *Proceedings of the 10th ACM workshop on Multimedia and security*, MM&Sec '08, pp. 159–166. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-058-6. doi:10.1145/1411328.1411356. URL http://doi.acm.org/10.1145/1411328.1411356. 104, 105, 113, 117

[Kraetzer09a] C. Kraetzer and J. Dittmann. *The Impact of Information Fusion in Steganalysis on the Example of Audio Steganalysis*. In E. J. Delp, J. Dittmann, N. D. Memon, and P. W. Wong (eds.), *Media Forensics and Security XI*, vol. 7254 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, January, 2009. doi: 10.1117/12.805884. 92, 104, 105, 117, 173, 184

[Kraetzer09b] C. Kraetzer, M. Schott, and J. Dittmann. *Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models*. In *Proceedings of the 11th ACM workshop on Multimedia and security*, MM&Sec '09, pp. 49–56. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-492-8. doi:10.1145/1597817.1597827. URL http://doi.acm.org/10.1145/1597817.1597827. 127, 145, 173, 185

[Kraetzer10] C. Kraetzer and J. Dittmann. *Improvement of information fusion-based audio steganalysis*. In R. Creutzburg and D. Akopian (eds.), *Multimedia on Mobile Devices 2010*, vol. 7542 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, January, 2010. doi:10.1117/12.838869. 86, 87, 104, 105, 106, 109, 111, 120, 121, 173, 184

[Kraetzer11] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann. *A Context Model for Microphone Forensics and its Application in Evaluations*. In N. D. Memon, J. Dittmann, A. M. Alattar, and E. J. Delp (eds.), *Media Watermarking, Security, and Forensics XIII*, vol. 7880 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Francisco, CA, USA, 24 - 26 January, 2011. doi:10.1117/12.871929. 19, 28, 29, 128, 154

[Kraetzer12a] C. Kraetzer and J. Dittmann. *Plausibility Considerations on Steganalysis as a Security Mechanism - Discussions on the Example of Audio Steganalysis.* T. Data Hiding and Multimedia Security, vol. 8:pp. 80–101, 2012. URL http://dblp.uni-trier.de/db/journals/tdhms/tdhms8.html#KraetzerD12. 104, 173, 174

[Kraetzer12b] C. Kraetzer, K. Qian, and J. Dittmann. *Extending a context model for microphone forensics*. In N. D. Memon, A. M. Alattar, and E. J. D. III (eds.), *Media Watermarking, Security, and Forensics XIV*, vol. 8303 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Francisco, CA, USA, 2012. ISBN 9780819489500. doi:10.1117/12.906569. 28, 29, 128, 152, 153

[Kruus03] P. Kruus, C. Scace, M. Heyman, and M. Mundy. *A survey of steganographic techniques for image Files*. Adv. Secur. Res. J., vol. 5(1):pp. 41–51, 2003. ISSN 0018-8670. 45

[Kullback59] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959. 40

[Kumar10] S. Kumar, K. Singh, and R. Saxena. *Analysis of Dirichlet and Generalized Hamming-window functions in the fractional Fourier transform domains*. Signal Processing, 2010. ISSN 01651684. doi:10.1016/j.sigpro.2010.04.011. URL http://dx.doi.org/10.1016/j.sigpro.2010.04.011. 27

[Kuncheva04] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. ISBN 0471210781. 182, 186

[Lally03] S. Lally. *What tests are acceptable for use in forensic evaluations? A survey of experts*. Professional Psychology: Research and Practice, (34):pp. 491–498, 2003. 24

[Landis77] J. Landis and G. Koch. *The measurement of observer agreement for categorical data*. Biometrics, vol. 33:p. 159174, 1977. 13, 90, 129

[Lang06] A. Lang and J. Dittmann. *Profiles for Evaluation and their Usage in Audio WET*. In P. W. Wong and E. J. Delp (eds.), *Security and Watermarking of Multimedia Content VIII*, vol. 6072 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, January, 2006. 93

[Lang07] A. Lang. *Audio Watermarking Benchmarking – A Profile Based Approach*. Ph.D. thesis, Dept. of Computer Science, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany, 2007. 92, 164, 176, 182

[Lerch08] A. Lerch. *Software-Based Extraction of Objective Parameters from Music Performances*. GRIN, 2008. 193

[Li11] B. Li, J. He, J. Huang, and Y. Q. Shi. *A Survey on Image Steganography and Steganalysis*. Information Hiding and Multimedia Signal Processing, vol. 2(2), 2011. ISSN 2073-4212. 44, 45, 51

[Liu08] Y. Liu, K. Chiang, C. L. Corbett, R. Archibald, B. Mukherjee, and D. Ghosal. *A Novel Audio Steganalysis Based on High-Order Statistics of a Distortion Measure with Hausdorff Distance.* In T.-C. Wu, C.-L. Lei, V. Rijmen, and D.-T. Lee (eds.), *ISC*, vol. 5222 of *Lecture Notes in Computer Science*, pp. 487–501. Springer, 2008. ISBN 978-3-540-85884-3. URL http://dblp.uni-trier.de/db/conf/isw/isc2008.html#LiuCCAMG08. 164

[Liu09] Q. Liu, A. H. Sung, and M. Qiao. *Novel stream mining for audio steganalysis.* In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pp. 95–104. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-608-3. doi:10.1145/1631272.1631288. URL http://doi.acm.org/10.1145/1631272.1631288. 49, 86, 196

[Liu11] Q. Liu, A. H. Sung, and M. Qiao. *Derivative-based audio steganalysis.* ACM Trans. Multimedia Comput. Commun. Appl., vol. 7(3):pp. 18:1–18:19, 2011. ISSN 1551-6857. doi:10.1145/2000486.2000492. URL http://doi.acm.org/10.1145/2000486.2000492. 86

[LLI10a] LLI. *Federal Rules of Evidence - FRE702*, 2010. URL http://www.law.cornell.edu/rules/fre/rules.htm#Rule702. Legal Information Institute, Cornell Law School (LLI). 20, 21, 81, 124, 162

[LLI10b] LLI. *Federal Rules of Evidence - Notes on FRE702*, 2010. URL http://www.law.cornell.edu/rules/fre/ACRule702.htm. Legal Information Institute, Cornell Law School (LLI). 21, 24, 81, 125, 127, 162

[LLI11] LLI. *Federal rules of evidence overview)*, 2011. URL http://www.law.cornell.edu/rules/fre/rules.htm. Legal Information Institute, Cornell Law School (LLI). 7

[Lu07] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian. *Feature selection using principal feature analysis.* In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pp. 301–304. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-702-5. doi:10.1145/1291233.1291297. URL http://doi.acm.org/10.1145/1291233.1291297. 35

[Luck08] G. Luck and P. Toiviainen. *Exploring Relationships between the Kinematics of a Singer's Body Movement and the Quality of Their Voice.* Journal of interdisciplinary music studies, vol. 2(0821211):pp. 173–186, 2008. 193

[Maher06] R. C. Maher. *Modeling and signal processing of acoustic gunshot recordings.* In *Proc. IEEE Signal Processing Society 12th DSP Workshop*. 2006. 18

[Maher07] R. C. Maher. *Acoustical Characterization of Gunshots.* In *Proc. SAFE07*. USA, 2007. 18

[Maher10] R. C. Maher. *Overview of Audio Forensics.* In H. T. Sencar, S. A. Velastin, N. Nikolaidis, and S. Lian (eds.), *Intelligent Multimedia Analysis for Security Applications*, vol. 282 of *Studies in Computational Intelligence*, pp. 127–144. Springer, 2010. ISBN 978-3-642-11754-1. URL http://dblp.uni-trier.de/db/series/sci/sci282.html#Maher10. 1, 51, 52

[Malik10] H. Malik and H. Farid. *Audio forensics from acoustic reverberation.* In *ICASSP*, pp. 1710–1713. IEEE, 2010. ISBN 978-1-4244-4296-6. 30, 53, 54, 55, 56, 57, 58, 59, 74, 165, 167

[Malik12] H. Malik and J. W. Miller. *Microphone Identification Using Higher-Order Statistics.* In *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*. 2012. URL http://www.aes.org/e-lib/browse.cfm?elib=16333. 54, 55, 56, 57, 59, 69, 74, 165, 166, 167

[Mathieu10] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. *YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software*. In J. S. Downie and R. C. Veltkamp (eds.), *ISMIR*, pp. 441–446. International Society for Music Information Retrieval, 2010. ISBN 978-90-393-53813. URL http://dblp.uni-trier.de/db/conf/ismir/ismir2010.html#MathieuEFPR10. 186

[Matsuoka06] H. Matsuoka. *Spread Spectrum Audio Steganography Using Sub-band Phase Shifting*. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, IIH-MSP '06, pp. 3–6. IEEE Computer Society, Washington, DC, USA, 2006. ISBN 0-7695-2745-0. doi:10.1109/IIH-MSP.2006.157. URL http://dx.doi.org/10.1109/IIH-MSP.2006.157. 45

[McEachern94] R. McEachern. *Hearing it like it is: Audio signal processing the way the ear does it*. Tech. rep., DSP Applications, 1994. 194

[Meghanathan10] N. Meghanathan and L. Nayak. *A review of the audio and video steganalysis algorithms*. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, pp. 81:1–81:5. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0064-3. doi:10.1145/1900008.1900118. URL http://doi.acm.org/10.1145/1900008.1900118. 44, 45, 47

[Meyers04] M. Meyers and M. Rogers. *Computer Forensics: The Need for Standardization and Certification*. International Journal of Digital Evidence, vol. 3(2), 2004. 24

[Miche06] Y. Miche, B. Roue, A. Lendasse, and P. Bas. *A feature selection methodology for steganalysis*. In *Proceedings of the 2006 international conference on Multimedia Content Representation, Classification and Security*, MRCS'06, pp. 49–56. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 978-3-540-39392-4. doi:10.1007/11848035_9. URL http://dx.doi.org/10.1007/11848035_9. 48, 49

[Molero01] S. Molero. *Establishment of the Individual Characteristics of Magnetic Recording Systems for Identification Purposes*. Tech. rep., Problems of Forensic Sciences, Second EAFS meeting, Volume XLVII, Cracow, Poland, 2001. 18

[Moncrieff06] S. Moncrieff, S. Venkatesh, and G. West. *Unifying Background Models over Complex Audio using Entropy*. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 04*, ICPR '06, pp. 249–253. IEEE Computer Society, Washington, DC, USA, 2006. ISBN 0-7695-2521-0. doi:10.1109/ICPR.2006.1141. URL http://dx.doi.org/10.1109/ICPR.2006.1141. 54

[Nelson10] B. Nelson, A. Phillips, and C. Steuart. *Guide to computer forensics and investigations*. Course Technology, Boston, MA, 4 edn., 2010. ISBN 978-1-435-49883-9. 24

[Neustein11] A. Neustein and H. A. Patil. *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. Springer Publishing Company, Incorporated, 2011. ISBN 146140262X, 9781461402626. 18

[Nian06] G.-j. Nian, Z. Xie, and S.-x. Wang. *Audio Watermarking Based on Reverberation*. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, IIH-MSP '06, pp. 37–40. IEEE Computer Society, Washington, DC, USA, 2006. ISBN 0-7695-2745-0. doi:10.1109/IIH-MSP.2006.62. URL http://dx.doi.org/10.1109/IIH-MSP.2006.62. 45

[Nicolalde09] D. P. Nicolalde and J. A. Apolinario. *Evaluating digital audio authenticity with spectral distances and ENF phase change*. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pp. 1417–1420. IEEE Computer Society, Washington, DC, USA, 2009. ISBN 978-1-4244-2353-8. doi:10.1109/ICASSP.2009.4959859. URL http://dx.doi.org/10.1109/ICASSP.2009.4959859. 52

[Nissar10] A. Nissar and A. H. Mir. *Classification of steganalysis techniques: A study*. Digit. Signal Process., vol. 20(6):pp. 1758–1770, 2010. ISSN 1051-2004. doi:10.1016/j.dsp.2010.02.003. URL http://dx.doi.org/10.1016/j.dsp.2010.02.003. 42, 63

[Nosrati12] M. Nosrati and R. K. an Mehdi Hariri. *Audio Steganography: A Survey on Recent Approaches*. World Applied Programming (WAP), vol. 2(3):pp. 202–205, 2012. ISSN 2222-2510. 44

[Oermann05] A. Oermann, A. Lang, and J. Dittmann. *Verifier-tuple for audio-forensic to determine speaker environment*. In A. M. Eskicioglu, J. J. Fridrich, and J. Dittmann (eds.), *MM&Sec*, pp. 57–62. ACM, 2005. ISBN 1-59593-032-9. 3, 28, 53, 69, 167

[Oermann06] A. Oermann, T. Scheidat, C. Vielhauer, and J. Dittmann. *Semantic fusion for biometric user authentication as multimodal signal processing*. In *Proceedings of the 2006 international conference on Multimedia Content Representation, Classification and Security*, MRCS'06, pp. 546–553. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3-540-39392-7, 978-3-540-39392-4. doi:10.1007/11848035_72. URL http://dx.doi.org/10.1007/11848035_72. 183

[Orsdemir08] A. Orsdemir, H. O. Altun, G. Sharma, and M. F. Bocko. *Steganalysis aware steganography: statistical indistinguishability despite high distortion*. In E. J. Delp, P. W. Wong, J. Dittmann, and N. D. Memon (eds.), *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, January, 2008. 41, 50, 73

[Owen88] T. Owen. *Forensic Audio and Video-Theory and Applications*. J. Audio Eng. Soc, vol. 36(1/2):pp. 34–41, 1988. URL http://www.aes.org/e-lib/browse.cfm?elib=5164. 51

[Özer03] H. Özer, I. Avcibas, B. Sankur, and N. D. Memon. *Steganalysis of audio based on audio quality metrics*. In *Security and Watermarking of Multimedia Contents*, Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series. 2003. doi:10.1117/12.477313. 49, 164, 175

[Palmer01] G. L. Palmer. *A Road Map for Digital Forensics Research - Report from the First Digital Forensics Research Workshop (DFRWS) (Technical Report DTR-T001-01 Final)*. Tech. rep., Air Force Research Laboratory, Rome Research Site, Utica, NY, 2001. 5, 17

[Pawera03] N. Pawera. *Microphone Practice: Tips and Tricks for Stage and Studio: Equipment, Acoustics and Recording Practice for Instruments and Vocals*. A @book from PPVMedien. PPV Medien GmbH, 2003. ISBN 9783932275630. URL http://books.google.de/books?id=YQNu0gAACAAJ. 29, 30

[Peeters04] G. Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. rep., CUIDADO I.S.T. Project Report, 2004. 186

[Pekalska05] E. Pekalska and R. P. W. Duin. *The dissimilarity representation for pattern recognition - Foundations and Applications*, vol. 64 of *Machine Perception and Artificial Intelligence*. World Scientific Pub Co., 2005. ISBN 981-256-530-2. 32

[Pellicano90] A. Pellicano. *Tape recordings as evidence*. California Lawyer, 1990. 18

[Pinkl09] P. Pinkl. *Alter (ver)messen? Gesellschaftspolitische Anwendungszusammenhänge wissenschaftlicher Vermessungstechniken zur chronologischen Lebensalterbestimmung im österreichischen Kontext*. Ph.D. thesis, University of Vienna, Austria, 2009. 24

[Pressnitzer00] D. Pressnitzer and S. McAdams. *Acoustics, psychoacoustics and spectral music*. Contemporary Music Review, vol. 19(2):pp. 33–60, 2000. 26

[Provos02] N. Provos and P. Honeyman. *Detecting Steganographic Content on the Internet*. In *NDSS*. The Internet Society, 2002. ISBN 1-891562-14-2, 1-891562-13-4. 46, 47, 48, 50, 51, 63, 65, 73, 74, 78, 119, 120, 123, 165, 172

[Reichow11] B. Reichow. *Evidence-Based Practices and Treatments for Children with Autism*. SpringerLink. Springer, 2011. ISBN 9781441969750. URL http://books.google.de/books?id=znCCkL6XvFAC. 90

[Rekhis07] S. Rekhis. *Theoretical Aspects of Digital Investigation of Security Incidents*. Ph.D. thesis, Engineering School of Communications, Tunisia, 2007. 17

[REW11] *The REWIND Project REVerse engineering of audio Visual content Data: Deliverable D3.1 State-of-the-art on multimedia footprint detection*. Tech. rep., The REWIND Project, 2011. 53, 56, 58, 59, 73, 168, 188

[Robin00] M. Robin and M. Poulin. *Digital television fundamentals: design and installation of video and audio systems*. McGraw-Hill Video/Audio Engineering Series. McGraw-Hill, 2000. ISBN 9780071355810. URL http://books.google.com/books?id=BkC0d_d8_u0C. 31

[Rodríguez10] D. P. N. Rodríguez, J. A. Apolinário, and L. W. P. Biscainho. *Audio authenticity: detecting ENF discontinuity with high precision phase analysis*. Trans. Info. For. Sec., vol. 5(3):pp. 534–543, 2010. ISSN 1556-6013. doi:10.1109/TIFS.2010.2051270. URL http://dx.doi.org/10.1109/TIFS.2010.2051270. 13, 52, 54, 188

[Ross06] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387222960. 3, 182, 183, 184

[Ru05] X.-M. Ru, H.-J. Zhang, and X. Huang. *Steganalysis of audio: Attacking the steghide*. In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, pp. 3937–3942. 2005. 49

[Rumsey08] F. Rumsey. *Forensic Audio Analysis*. J. Audio Eng. Soc, vol. 56(3):pp. 211–217, 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14382. 51

[Sanderson02] C. Sanderson and K. Paliwal. *Information Fusion and Person Verification Using Speech and Face Information*. In *Technical ReportIDIAP 02-33*. Martigny, Switzerland, 2002. 183, 184

[Santhi12] B. Santhi, G. Radhika, and S. R. Rek. *Information Security using Audio Steganography - A Survey*. Research Journal of Applied Sciences, Engineering and Technology, vol. 4(14):pp. 2255–2258, 2012. ISSN 2040-7467. 44

[Shafer76] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976. 186

[Shannon49] C. E. Shannon. *Communication in the Presence of Noise*. Proceedings of the Institute of Radio Engineers, IRE, vol. 37(1):pp. 10–21, 1949. 26, 30

[Shyu98] C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick. *Local versus global features for content-based image retrieval*. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 30–34. IEEE, Santa Barbara, CA, USA, 1998. ISBN 0-8186-8544-1. 34

[Sim05] J. Sim and C. C. Wright. *The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements*. Physical Therapy, vol. 85(3):p. 257268, 2005. 90

[Smaragdis09] P. Smaragdis, R. Radhakrishnan, and K. W. Wilson. *Context Extraction Through Audio Signal Analysis*. In A. Divakaran (ed.), *Multimedia Content Analysis*, Signals and Communication Technology, pp. 1–34. Springer US, 2009. ISBN 978-0-387-76567-9. doi:10.1007/978-0-387-76569-3_1. URL http://dx.doi.org/10.1007/978-0-387-76569-3_1. 193

[SWGFAST11] SWGFAST. *The Fingerprint Sourcebook*. National Institute of Justice, U.S. Department of Justice, 2011. ISBN 9781477664766. URL http://books.google.de/books?id=Z8dyLwEACAAJ. Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). 3, 4, 20, 22, 23, 25

[Takala92] T. Takala and J. Hahn. *Sound rendering*. SIGGRAPH Comput. Graph., vol. 26:pp. 211–220, 1992. ISSN 0097-8930. doi:http://doi.acm.org/10.1145/142920.134063. URL http://doi.acm.org/10.1145/142920.134063. 30

[The Open Group08] The Open Group. *IEEE Std 1003.1-2008*. The Open Group Base Specifications Issue 7, The IEEE and The Open Group, NY, 2008. 174, 175

[Thrasyvoulou03] T. Thrasyvoulou and S. Benton. *Speech parameterization using the Mel scale Part II*. Tech. rep., 2003. 194, 195

[Tian09] H. Tian, K. Zhou, H. Jiang, J. Liu, Y. Huang, and D. Feng. *An M-sequence based steganography model for voice over IP*. In *Proceedings of the 2009 IEEE international conference on Communications*, ICC'09, pp. 683–687. IEEE Press, Piscataway, NJ, USA, 2009. ISBN 978-1-4244-3434-3. URL http://dl.acm.org/citation.cfm?id=1817271.1817399. 45

[Tibbitts09] J. Tibbitts and Y. Lu. *Forensic applications of signal processing*. IEEE Signal Processing Magazine, vol. 26:pp. 104–111, 2009. doi:10.1109/MSP.2008.931099. 51

[U.S. Congress23] U.S. Congress. *Frye v. United States, 293 F. 1013 (D.C. Cir.)*, 1923. 22

[U.S. Congress05] U.S. Congress. *Planning for Library of: Sustainability of Digital Formats – PCM*. Tech. rep., Congress Collections, 2005. URL http://www.digitalpreservation.gov/formats/fdd/fdd000016.shtml. 26

[U.S. Congress10] U.S. Congress. *Federal Rules of Criminal Procedure. (Amended by the United States Supreme Court Dec. 1st, 2010, (eff. Dec. 1st, 2010.))*, 2010. URL http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/. 22

[U.S. Congress11] U.S. Congress. *Federal Rules of Evidence. (Amended by the United States Supreme Court Apr. 26, 2011, (eff. Dec. 1, 2011.))*, 2011. URL http://federalevidence.com/downloads/rules.of.evidence.pdf. 21, 22, 23, 43, 62

[USC93] USC. *United States Court (USC) 509 U.S. 579*, 1993. Daubert v. Merrell Dow Pharmaceuticals, Inc. 7, 19, 20, 22, 23, 24, 62, 68, 72, 78, 81, 103, 125, 162

[USC99] USC. *United States Court (USC) 526 U.S. 137, 119 S.Ct. 1167*, 1999. Kumho Tire Co. v. Carmichael. 20

[USCA95] USCA. *United States Court of Appeals (USCA), Ninth Circuit. No. 90-55397. Argued and Submitted March 22, 1994. Decided January 4, 1995*, 1995. Daubert, William and Joyce Daubert, individually and as Guardians Ad Litem for Jason Daubert, (a minor); Anita De Young, individually, and as Guardian Ad Litem for Eric Schuller, Plaintiffs-Appellants, vs. Merrell Dow Pharmaceuticals, Inc., a Delaware corporation, Defendant-Appellee. 7, 19

[Vielhauer05] C. Vielhauer. *Biometric User Authentication for IT Security: From Fundamentals to Handwriting (Advances in Information Security)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 038726194X. 3, 6, 33

[Waters88] G. T. Waters. *Sound Quality Assessment Material Recordings for Subjective Tests*. Users' handbook for the EBU - SQAM compact disc, European Broadcasting Union, Avenue Albert Lancaster 32, 1180 Bruxelles (Belgique), 1988. 92, 97

[Westfeld99] A. Westfeld and A. Pfitzmann. *Attacks on Steganographic Systems*. In A. Pfitzmann (ed.), *Information Hiding*, vol. 1768 of *Lecture Notes in Computer Science*, pp. 61–76. Springer, 1999. ISBN 3-540-67182-X. 42, 45

[Wilkins41] J. Wilkins. *Mercury: Or the Secret and Swift Messenger: Shewing, How a Man May With Privacy and Speed Communicate His Thoughts to a Friend at Any Distance*. Printed by I. Norton for John Maynard and Timothy Wilkins, London, 1641. 18

[Winkler11] A. Winkler. *Advances in Syndrome Coding based on Stochastic and Deterministic Matrices for Steganography*. Ph.D. thesis, University of Dresden, Germany, 2011. URL http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-84141. 31, 39

[Witten05] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edn., 2005. 35, 88, 89

[Wölfel11] U. Wölfel. *Efficient and Provably Secure Steganography*. Ph.D. thesis, University of Lübeck, Germany, 2011. 2, 18

[Yang08] R. Yang, Z. Qu, and J. Huang. *Detecting digital audio forgeries by checking frame offsets*. In *Proceedings of the 10th ACM workshop on Multimedia and security*, MM&Sec '08, pp. 21–26. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-058-6. doi:10.1145/1411328.1411334. URL http://doi.acm.org/10.1145/1411328.1411334. 18

[Yang09] R. Yang, Y.-Q. Shi, and J. Huang. *Defeating fake-quality MP3*. In *Proceedings of the 11th ACM workshop on Multimedia and security*, MM&Sec '09, pp. 117–124. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-492-8. doi:10.1145/1597817.1597838. URL http://doi.acm.org/10.1145/1597817.1597838. 18

[Yang10] R. Yang, Y. Q. Shi, and J. Huang. *Detecting double compression of audio signal*. In N. D. Memon, J. Dittmann, A. M. Alattar, and E. J. Delp (eds.), *Media Forensics and Security II*, vol. 7541 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Electronic Imaging Conference Series*. San Jose, CA, USA, January, 2010. doi:10.1117/12.838695. 18

[Yang12] R. Yang, Z. Qu, and J. Huang. *Exposing MP3 audio forgeries using frame offsets*. ACM Trans. Multimedia Comput. Commun. Appl., vol. 8(2S):pp. 35:1–35:20, 2012. ISSN 1551-6857. doi:10.1145/2344436.2344441. URL http://doi.acm.org/10.1145/2344436.2344441. 18

[Zwicker61] E. Zwicker. *Subdivision of the audible frequency range into critical bands (Frequenzgruppen)*. Journal of the Acoustical Society of America, vol. 33(2):pp. 248–249, 1961. 192

[Zwicker90] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer-Verlag, 1990. 195