Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification

Christian Kraetzer, Andrea Oermann, Jana Dittmann and Andreas Lang
Research Group Multimedia and Security
Department of Computer Science
Otto-von-Guericke University of Magdeburg, Germany
{kraetzer, oermann, dittmann, alang}@iti.cs.uni-magdeburg.de

ABSTRACT

In this paper a first approach for digital media forensics is presented to determine the used microphones and the environments of recorded digital audio samples by using known audio steganalysis features. Our first evaluation is based on a limited exemplary test set of 10 different audio reference signals recorded as mono audio data by four microphones in 10 different rooms with 44.1 kHz sampling rate and 16 bit quantisation. Note that, of course, a generalisation of the results cannot be achieved. Motivated by the syntactical and semantical analysis of information and in particular by known audio steganalysis approaches, a first set of specific features are selected for classification to evaluate, whether this first feature set can support correct classifications. The idea was mainly driven by the existing steganalysis features and the question of applicability within a first and limited test set. In the tests presented in this paper, an inter-device analysis with different device characteristics is performed while intra-device evaluations (identical microphone models of the same manufacturer) are not considered. For classification the data mining tool WEKA with K-means as a clustering and Naive Bayes as a classification technique are applied with the goal to evaluate their classification in regard to the classification accuracy on known audio steganalysis features. Our results show, that for our test set, the used classification techniques and selected steganalysis features, microphones can be better classified than environments. These first tests show promising results but of course are based on a limited test and training set as well a specific test set generation. Therefore additional and enhanced features with different test set generation strategies are necessary to generalise the findings.

MM&Sec'07, September 20–21, 2007, Dallas, Texas, USA.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Signal Processing*

General Terms

Measurement, Reliability, Experimentation, Verification

Keywords

Digital Media Forensics, Multimedia Authentication

1. MOTIVATION

Today digital media forensics are becoming of increasing importance as information is rather communicated digitally than analogue. Many formats for digital information are existing while new formats are continuously emerging. Digital information is easily transformed and newly generated, which bears novel challenges of assuring the integrity and authenticity of that information. Digital media facilitate undetected forgeries and manipulations, and might thereby encourage criminals. Those threats need to be minimised. Thus, digital media forensics are established as a consequence of modern digital information technologies and communication systems such as for example the Internet and VoIP. Motivated from our experiences in steganalysis the work presented in this paper applies a specific steganalysis toolset for digital media forensics based on audio data as digital media. Our goal is to see if and how known features from the detection of hidden communications can help to classify the origin of audio streams. The idea is to extract known statistical features and to evaluate their discriminative power for microphone and environment classification. In this first practical evaluations on a small test set only inter-device evaluations are performed. This paper is a first but important contribution in the field of detecting forgeries in digital audio media.

Basically, forensics refers to the posterior detection and securing of information left behind at a crime scene. Digital media forensics further includes the assurance of integrity and authenticity of digital information, which in a digital world is rather problematic, due to the simple and fast reproduction of information and production of manipulations and forgeries. The focus of forensics is the evaluation of evidence to localise those manipulations, prove the authenticity and integrity of information as well as their origin. Considering digital media, such a proof can be the identification of sensors (devices) a digital information is created with. In this context, an approach for digital cameras from images based on the sensor's noise pattern is presented in [9] and [3]. An approach for determining the used digitiser tablets based on handwriting samples is introduced in [15] and further enhanced in [14]. An approach for identifying printers based on greylevel co-occurrence features is presented in [12], while another approach for the forensic identification of printers based on SVM techniques is presented in [11]. An overview of methods for forensic characterisation of devices in given in [6], where current forensic identification techniques for RF (radio frequency) devices, printers, and cameras are presented and examined, and a generalisation for the use with other devices is introduced. These approaches contain both, inter- as well as intra-device analysis.

This paper is based on our previous and recent research, whose fundamental concept is presented in [13]. A so-called "Verifier-Tuple" is developed in order to be able to structure and analyse information in detail and extract specific features for defined information layers. By applying the "Verifier-Tuple" five different information layers are classified which include a basic differentiation between syntax and semantic features. Thus, we are able to expose the connection between the different but interacting layers of information. Further in regard to digital media forensics, this "Verifier-Tuple" model enables a user to determine which parts of information are required in order to correctly derive additional information, which basically is not directly included in the available and analysable information - a phenomenon commonly occurring within forensics.

In this paper we evaluate the applicability of our audio steganalysis approach from [7] as "Verifier-Tuple" for audio forensics. This approach was used successful in audio steganalysis where the embedding for selected audio steganography algorithms was detected with up to 100% accuracy under certain assumptions introduced in [7]. The goal for the tests performed here was to evaluate whether the same features leading to good results in steganalysis are also applicable and useful in the identification of sensors.
So the main idea is that from syntactical audio features additional higher level semantical features can be derived up to microphone and environment classification. Based on this we propose three hypotheses and evaluate these hypotheses in the tests presented in this paper:

- Hypothesis I: Is it possible to correctly classify the used microphone for the generation of a recording?
  The evaluations for this hypothesis are split into two parts: a) for the general classification and b) for the classification of every single microphone.

- Hypothesis II: Is it possible to correctly classify the location where a recording was made?
  The evaluations for this hypothesis are split into two parts: a) for the general classification and b) for the classification of selected single rooms.

- Hypothesis III: Does feature selection (feature reduction) inprove the classification accuracy?

An exemplary test set of 10 different audio files recorded for the inter-device classification as mono audio data by four microphones from different manufacturers in 10 different rooms with 44.1 kHz sampling rate and 16 bit quantisation is considered for our evaluations to determine whether the audio steganalysis features from [7] are useful and relevant for media forensics. Note that these first tests (using $10*10*4 = 400$ audio files a 18.5 seconds = approximately 7400 seconds audio material for evaluation) can only give an indication since the used test set is relatively small. An generalisation of the approach is of course still an open question, since the number of reference files is significantly smaller than the number of features used in the evaluation, even when considering the fact that parts of the feature space are correlated.
All feature computations are done using AAST (AMSL Audio Steganalysis Toolset; version 1.03) [7] with its default window size of 1024 samples per window. Identical environmental signals are captured by all microphones due to the files are recorded syntactically synchronised for all four microphones at the same time for each room. Regarding the "Verifier-Tuple", the feature vectors computed from the recorded audio files are not semantically normalised by AAST, but syntactically as features are extracted parallel from a fixed number of windows a 1024 samples on the selected set of audio files.

This paper is structured as follows: In section 2 the feature computation derived from our audio steganalysis approach from [7] is briefly introduced and classification strategies are summarised. In section 3 the test scenario is presented which includes a description of the test sets, test set-up, test procedure as well as the precise test objectives. In section 4, the test results are presented and discussed. Finally, section 5 summarises and concludes the paper and outlines future work.

2. FEATURE COMPUTATION AND CLASSIFICATION

To address the three hypotheses identified in section 1, the AAST and the WEKA [17] data mining software (version 3.4.10) described in [18] are used to provide the feature vectors and results for classification. In the following sections the feature computation step, which outputs the feature vectors for classification, and the classification approaches used are explained in detail.

2.1 Feature Computation

The set of 63 statistical features ($sf_i \in \mathbb{FS}$) computed by AAST (version 1.03) for windows of the signal (intra-window) consists of 7 time domain based features and 56 mel-cepstral domain based features. It was decided to use initially the complete feature set of AAST in the classification and then use the feature selection function of WEKA to evaluate the impact of a reduced feature vector length on the classification accuracy.
The time domain based features computed by AAST are: sf_{ev} empirical variance, sf_{cv} covariance, $sf_{entropy}$ entropy, $sf_{LSB_{rat}}$ LSB ratio, $sf_{LSB_{flip}}$ LSB flipping rate, sf_{mean} mean of samples in time domain and sf_{median} median of samples in time domain. The computation of these features is described in detail in [2]. The 56 mel-cepstral domain based features sf_{mel_1}, ..., sf_{mel_C} and sf_{melf_1}, ..., sf_{melf_C}

($C = 28$ for CD-quality audio files) are described in [7]. For a complete list of the statistical features used see table 1. All feature computations are done using the default window size for AAST (1024 samples per window). The maximum number of windows to be computed for the tests is limited by the duration of the shortest file in the test set to 800 windows (file length = 18.65 seconds; $18.65s * 44100$ frames per second = 822272 frames; 822272/1024 frames per window = 803 windows). For a detailed description of the computed features and the usage of AAST we refer to [7]. The output feature vectors are not (semantically) normalised by AAST.

2.2 Classification Strategies

Classification is mainly applied in the research field of data mining and includes techniques for finding unknown patterns in large data sets [4]. Generally and according to the goal, classification is based on either supervised (classification by classifiers) or unsupervised (clustering) techniques. For classification by classifiers different models such as decision trees, regression analysis, support vector machines, or Naive Bayes classifiers can be applied, while each of them contains different specific algorithms. Most commonly the data set is separated in two parts: training data set and testing data set. The first part of the classification process is training by applying so-called supervised learning algorithms. Based on a set of mutually exclusive and predefined classes of classified (labelled) data, classifiers are built. As described in detail in the next section, the Naive Bayes algorithm is one of the most common applied algorithms for building a classifier. Based on the built classifier, unclassified (unlabelled) testing data gets assigned to a specific predefined class of the training data by considering the values of attributes (features) describing the class.
Clustering as the unsupervised classification strategy refers to the partitioning of data into groups (clusters) while the data's group affiliation is not known in advance. In this context, the data set is not separated. Data belonging to a group share common or similar traits, which are, for example by defined distance measures, determined as proximity. In the tests performed within this work the applicability of both techniques are evaluated by analysing and comparing the achieved results using one example classifier (Naive Bayes) and one clustering algorithm (K-means).

2.2.1 Classification

The Naive Bayes classifiers are a well-researched technique. They compute classifications using a probabilistic approach, i.e., they try to compute conditional class probabilities and then predict the most probable class. For a detailed description of the classification process of Naive Bayes classifiers see [1]. The Naive Bayes classifiers implemented in WEKA are chosen for this work because of three reasons:

Table 1: All single features sf_i in the feature space $\mathbb{FS}$ ($C = 28$ for audio files with 44.1 kHz sampling frequency)

domain	features
time domain	$\{sf_{ev}, sf_{cv}, sf_{entropy}, sf_{LSB_{rat}}, sf_{LSB_{flip}}, sf_{mean}, sf_{median}\}$
mel-cepstral domain	$\{sf_{mel_1}, ..., sf_{mel_C}, sf_{melf_1}, ..., sf_{melf_C}\}$

a) The problem at hand is a multi-class classification, which eliminates many other algorithms (e.g. most decision tree algorithms and SVM classification) from the set of easily usable classifiers.
b) They have a very low computational complexity when compared with other classification algorithms.
c) They did show good results in initial tests on all multi-class classifiers in WEKA.
In the tests performed the following parameterisations of WEKA's Naive Bayes classifier ($NaiveBayes$) are used: subset generation using percentual split (using the default 66%) or cross-validation (default 10-fold) with the default random seed for subset generation.

2.2.2 Clustering

The K-means algorithm [10] is one of basic unsupervised learning algorithms for solving the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The basic idea is to define k centroids, one for each cluster. Depending on the initial placement of these centroids the algorithm will return different clustering results. The general approach to address this problem is to place them randomly, but as far away from each other as possible. The second step after placing the centroids is to take each point belonging to a given data set and associate it to the nearest centroid. After all objects have been assigned, the positions of the k centroids are recalculated (the centroids change their location). For the k new centroids a new binding is computed on the data set points and the nearest new centroid. Theses steps, except the initial placement, are repeated until the centroids no longer move. Thus a separation of the objects into groups is achieved and from it the metric to be minimised can be calculated. Although it can be proven that the procedure will always terminate, the K-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is significantly sensitive to the initial placement of the centroids. A tutorial on K-means Clustering can be found in [5], for a more detailed description of the algorithm see [10].
In the clustering tests performed in this work, the following parameterisations of WEKA's implementation of the K-means algorithm ($SimpleKMeans$) are used: the seed for the placement of the k centeroids is set to 10 (default), k is set to the appropriate number of classes for each test. Also an equal number of samples for each class is ensured. The K-means clustering algorithm is chosen for the tests because it did show good results in initial tests on all clustering algorithms in WEKA.

2.2.3 Attribute Selection

To identify single features sf_i from the feature space $\mathbb{FS}$ which have a strong impact on the classification process, WEKA's attribute selection function is used. The attribute evaluator chosen for these tests is $CfsSubsetEval$ (evaluating the quality of a subset of attributes by considering the individual predictability of each feature as well as the degree of redundancy between them) using $BestFirst$ search (searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility [17]). The parameters for this search are left on their default settings. This evaluator / search pair is chosen because in initial tests

Table 3: The set of rooms evaluated ($\mathcal{R}$)

Rn	Room number	Description
R1	29R114	large office
R2	29R131	small office
R3	29R140	bathroom
R4	29R146	laboratory
R5	29R307	lecture hall
R6	audiobox	anechoic chamber
R7	outside1	quiet outside environment
R8	outside2	busy parking lot
R9	corridor	long and narrow corridor
R10	stairs	stone stairwell, strong echo

it resulted in a close match with the result of manual feature selection. The initial idea for the attribute selection was to identify all discriminative attributes and discard all others prior to the classification. The basic assumption for this idea was that a limitation to those features would improve the classification accuracy and at the same time reduce the computational complexity. But it was found in most of the tests performed that a reduction of the feature set would lead to a decreasing classification accuracy. Therefore, the idea to use the attribute selection as a pre-processing step was discarded and in the rest of the work it is only used to identify the most significant features for selected tests.

3. TEST SCENARIO

In this section the test sets, the set-up, the test procedure and the test goals are introduced for the evaluations of the three test hypotheses defined in section 1.

3.1 Test Sets and Test Set-up

In this section the following test sets are defined in regard to the practical evaluations: the set of microphones to be evaluated (see table 2), the set of rooms ($\mathcal{R}$; see table 3) and the set of test files (see table 4). The following criteria are applied in the selection of the test sets:

For the set of recording devices a selection of four microphones was randomly chosen from the set of microphones available at the AMSL (Advanced Multimedia and Security Lab of the Department of Computer Science, Otto-von-Guericke University of Magdeburg, Germany). The number was limited to four since only four microphone preamplifiers have been available for synchronous evaluations. Table 2 identifies all microphone and preamplifier combinations used.

For the set of the $Rn \in \mathcal{R}; n \in \mathbb{N}, n \in \{1, 2, ..., 10\}$ a number of 10 rooms was selected in the building 29 of the Otto-von-Guericke University of Magdeburg, Germany. The rooms chosen cover 10 different types of rooms (large and small office spaces, a bathroom, a noisy laboratory, a lecture hall, an anechoic chamber, a quiet and a busy outside location, a long and narrow corridor and a stairwell). For reasons of computational complexity in the tests it was decided to limit the number of rooms to these 10 representative classes. A complete listing of the rooms with their description is given in table 3.

In the set of the test files a number of 10 files (see table 4) from the AMSL audio test set described in [8] were chosen. These 10 files represent ten different classes of audio material (music (metal, pop, techno), noise (MLS and white noise), digital silence, a pure sine at 440Hz, recorded speech (male and female speaker) and one sample from the SQAM files (Sound Quality Assessment Material; see [16])). All material is provided in 44.1 kHz sampling frequency, 16 Bit quantisation, stereo and PCM coded.

The files are played in every room in $\mathcal{R}$ using a notebook computer and a Yamaha MSP 5 monitor speaker and the sound was recorded simultaneously by the four microphones (which were 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; recording parameters: 44.1 kHz sampling frequency, 16 Bit quantisation, mono and PCM coded). Using this procedure one sound file for each room, microphone and test file combination was generated (resulting in $10 * 4 * 10 = 400$ recorded files). The automated playback and recording routine used by the notebook guarantees that the corresponding files for all four microphones are synchronous (files are syntactical synchronised; environmental noise is encountered by all microphones at the same position in the recording).

The same 400 recorded files are used for the evaluations of hypothesis I (microphone classification) and II (room classification). In particular for hypothesis I the set of 400 files is divided into 10 subsets (with 10 files recorded by each microphone this results for 4 microphones in 40 files overall), each subset contains only the signals recorded within one particular room Rn, $n \in \mathbb{N}, 1 \leq n \leq 10$. Additional to the 40 files we add for each room the original reference files to simulate a 5th, lossless recording, this results in 50 files overall per set. Here we assume that a nearly lossless recording might be possible, while no similar proposition in the evaluations for hypothesis II could be found.

For hypothesis II the set of our 400 recorded files is split into the four microphone subsets of 100 files, each subset containing only the signals recorded by one of the four microphones for all rooms. On each of the 14 subsets generated (10 for hypothesis I and 4 for hypothesis II) a set of feature vectors is computed using AAST as it is described in section 2.1, resulting in a feature vector set for each subset of audio material.

The Bayesian classification as a supervised classification technique requires a training set for model generation and a test set for model evaluation. Here two different modes of this classifier are used. In the first mode the training and test set are split by WEKA by the ratio 66% for training and 34% for testing. In the second mode WEKA uses the10-fold cross-validation. The K-means clustering as unsupervised technique does not require a set splitting operation.

Note that, of course, a generalisation of the results cannot be achieved do to the small number of reference audio signals, which is far smaller than the dimensionality of the feature vectors computed. Furthermore the approach does not exclude yet the possibility of content dependent classification. Here we would need independent training and test sets (see the section on future work).

3.2 Test Procedure and Test Objectives

In the test procedure for hypothesis I concerned with microphone detection the influence of the room has to be minimised. Therefore in the evaluations 10 tests are performed, i.e. for each of the 10 rooms it is evaluated separately how good the microphones can be classified using different classifiers and different classifier modes (options). While the room

Table 2: The set of microphones used

Device short name	Manufacturer	Model	Used Pre-Amp.
AKG	AKG	SE 300 B	Millenium Mic 1
Headset	TerraTec	HeadsetMaster	Creative Sound Blaster USB
SM58	Shure	SM58	Creative Sound Blaster USB
Tbone	T.bone	MB45	Millenium Mic 1

Table 4: The set of reference files used

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

Rn is kept constant for each test the five different versions of each file (the original and the four recordings by the four microphones) are classified using the Naive Bayes classifier and clustered using the K-means algorithm.
The test goal for this tests is to determine how precise a known microphone can be identified using recordings from a set of known locations. Furthermore, the impact of the number of feature vectors on the classification is evaluated.
In contrast to the evaluations on hypothesis I, in the tests for hypothesis II, which is concerned with room detection, the influence of the used microphones has to be minimised. For the evaluations for the second test hypothesis four tests have to be performed, each using exactly one of the four microphones in the test set. In each test the used microphone has to be constant, while the signals recorded in the 10 rooms are classified. The goal for this tests is to determine how precise a known location can be identified using recordings by a set of known microphones. Here also the impact of the number of feature vectors on the classification is evaluated.
For hypothesis III the results for both classification strategies chosen (Naive Bayesian classification and K-means clustering) are compared and the relevance of single features sf_i from $\mathbb{FS}$ is evaluated. The goal is to identify the impact of feature selection strategies (i.e. the removal of non-significant features from the feature vector) has on the classification accuracy.

4. TEST RESULTS

In this section the test goals derived in section 3.2 from the three test hypotheses are addressed, using the test sets, the set-up and the test procedure described in section 3.

4.1 Results for the Evaluation of Hypothesis I

In the following the results of the evaluations for hypothesis I a) and b) ("Is it possible to correctly classify the used microphone for the generation of a recording?", a) for the general classification and b) for the classification of every single microphone) are discussed. These evaluations include results for the scaling of the classification accuracy with increasing number of feature vectors per file, which is summarised first for the complete set of rooms and then is discussed exemplary for one room ($R1$).

4.1.1 Comparison of the Results for Every Room and a Fixed Number of Vectors Computed per File

The table 5 averages the classification results for all rooms, a fixed number of feature vectors per file (800) and all chosen classifiers ($NaiveBayes$ with percentual split (66%) and 10-fold cross-validation as well as clustering using K-means) for the evaluation of hypothesis I a).
For the Bayesian classifiers the results for the microphone classification are in the range [61.37,75.99%], depending on the room. Both cases of Bayesian classification show very similar results in any test case. With 61.37% the lowest accuracy in the Bayesian microphone classification is found in the case of $R1$ (which is a large office). Noisy environments like $R4$ and $R8$ (a noisy lab and a busy outside parking lot) seem to have a positive effect on the classification result (second and third highest results). The best microphone classification (with 75.99% accuracy) was 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 [30.13,43.57%]. These results are lower than the ones achieved with Bayesian classification but nevertheless they are still by far better than "guessing" at the result (for five classes (recordings from the four microphones and the original data): 20%), therefore they are still considered significant within this work. In this case the worst result with 30.13% is computed for the evaluation in $R10$ (the stairwell, which was the best case in the Bayesian classifications). With 41.57% the best result is given for $R5$ (a large lecture hall). Since the results for the clustering are generally worse for the microphone detection than the results for Bayesian classification, the following discussions on hypothesis I are limited to the tests performed using Bayesian classification.

Table 9 at the end of the document shows the normalised confusion matrixes for all microphones, two different numbers of vectors computed per file (100 and 800) and the Naive Bayes classifier (using percentual split (66%) for training and testing set generation) for the evaluation of hypothesis I b). The results of the original confusion matrixes have been normalised to the range [0,1] to provide for a better comparability. The following facts are gained from the results displayed in table 9: The AKG microphone shows a very low classification accuracy when compared to the other microphones. In only two of the 10 rooms ($R3$ and $R7$) it is above the average classification accuracy identified in table 5 for the case of 800 vectors computed per file. In nine out of the 10 rooms the largest number of misclassified vectors for the AKG are classified as $SM58$ instead. In two cases the difference between the correctly as AKG classified vectors and the cases misclassified as $SM58$ is just 6% ($R2$ and $R5$).

Table 5: Average classification results (percentage of correctly classified feature vectors) for all rooms Rn (numbers of vectors computed per file is 800) and the different classificators applied

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
NaiveBayes; Perc. Split (66%)	61.37	68.84	62.05	73.34	68.25	69.23	69.44	74.76	71.97	75.65
NaiveBayes; 10xCross-Validation	61.76	69.63	62.61	74.38	68.63	69.31	70.29	75.14	72.50	75.99
SimpleKMeans(class)	36.53	34.68	38.07	33.49	41.57	36.97	38.12	43.57	33.39	30.13

The $Headset$ microphone shows generally a very good quality of classification results. For all 20 evaluations listed in table 9 it shows results above the average. With results between 78 and 96% in the case of 800 vectors computed per file, the $Headset$ can be very reliably identified in all rooms evaluated. In seven rooms out of the 10 tests for 800 vectors per file the $SM58$ shows the closest proximity to the $Headset$ but the minimal distance between the two is 59%.
The $SM58$ shows a very inhomogeneous behaviour with regard to the classification accuracy. In five out of the 10 tests for 800 vectors per file the $SM58$ performs above the average ($R2$, $R6$, $R8$, $R9$ and $R10$). The results for these tests range from 46 to 91% percent. The closest proximity in the results are found in seven cases to the $Tbone$ and in three cases to the $Headset$ (minimum distance here 15%).
The $Tbone$ shows for 800 vectors per file only in one case an accuracy above the average ($R4$). Nevertheless it has a smaller range ([55,86%]) of classification accuracies than the AKG. In eight cases the closest proximity in the classification can be found to the $SM58$ microphone (closest distance 23%).
In the evaluations for 800 vectors per file the *original* material was in four rooms ($R3$, $R7$, $R8$ and $R10$) classified with an accuracy below the average for the corresponding room as shown in table 5. The strongest numbers of misclassified vectors were assigned in six cases to the AKG (in five of this cases with a number of misclassified vectors between 22 and 28%).
When considering in the classification only the number of vectors falsely classified as originating from *original* material it can also be seen that this number is reduced to about 0% in the case of 800 vectors per file.
Figure 1 shows a histogram of the best classified microphone per room. In the case of 100 vectors per file in one case two microphones ($Headset$ and $SM58$) achieve the same maximum result ($R9$ and $R10$). It figure 1 it can be seen that the $Headset$ was the microphone detected best.

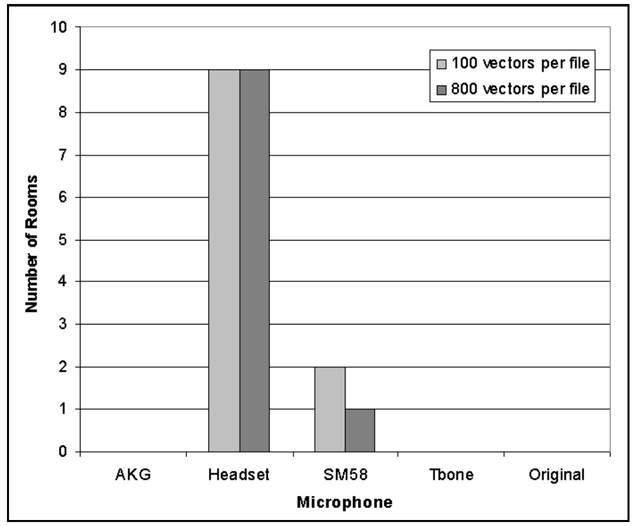

Figure 1: Histogram of the best classified microphones

4.1.2 Scaling of the Classification Accuracy with Increasing number of Feature Vectors

Table 6 shows exemplary the impact of the scaling of the number of input feature vectors on the classification accuracy for room $R1$. In the two cases of Bayesian classification the increasing of the number of feature vectors per file results in an increasing classification accuracy on the microphones. The best result is found with 61.76% in the case of $NaiveBayes$ with 10 fold cross-validation and 800 vectors per file. As already seen in section 4.1.1 above, both cases of Bayesian classification (percentual split (66%) and 10 fold cross-validation) show very similar results.
The results of the clustering using $SimpleKMeans$ are with a maximum of 40.33% lower than the results from the Bayesian classification but still far higher than a random classification on five equally distributed classes (which would be 20%). 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.
The other nine rooms show the same behaviour as $R1$ in the scaling tests, therefore a detailed description of the results for each room will be avoided here. The general results for all rooms are summarised and compared in section 4.1.1.
Concluding the results of the Bayesian classification it can be stated that, even when in selected cases the tests performed for 100 vectors per file return significantly better results than for 800 vectors per file (e.g. table 9 the $SM58$ in $R1$), the average results increase with the increasing number of vectors per file (see table 6).

4.2 The Evaluation of Hypothesis II

In the following the results of the evaluations for hypothesis II a) and b) ("Is it possible to correctly classify the location where a recording was made?", a) for the general classification and b) for the classification of selected single rooms) are discussed. For hypothesis II b) the $Headset$ microphone is chosen due to its performance in the evaluations for hypothesis I.
Table 7 summarises the tests performed for the evaluation of hypothesis II. In this table the rate of correct classificationa of all rooms is given for the material recorded by each microphone. Since a set of 10 rooms is considered here, a random correct classification of the rooms would occur with a likelihood of 10%. The results in table 7 are in the range of [23.97,41.54%] for the Bayes classificators and in the range of [10.99,26.49%] for the clustering using $SimpleKMeans$.

Table 6: Classification results (percentage of correctly classified feature vectors) for room $R1$ for different numbers of vectors computed per file and the different classificators applied

	100	200	300	400	500	600	700	800
NaiveBayes; Perc. Split (66%)	57.29	57.47	58.53	60.25	60.93	60.77	61.50	61.37
NaiveBayes; 10xCross-Validation	54.62	57.44	58.52	59.78	60.33	61.11	61.57	61.76
SimpleKMeans(class)	38.32	36.89	40.33	36.81	37.94	35.94	37.67	36.53

Table 7: Classification results (percentage of correctly classified feature vectors) for all four microphones and for different numbers of vectors computed per file and the different classificators applied

AKG	100	200	300	400	500	600	700	800
NaiveBayes; Perc. Split (66%)	30.76	23.97	24.42	25.77	25.43	25.23	24.80	26.08
NaiveBayes; 10xCross-Validation	31.59	24.95	24.93	25.09	25.00	25.11	25.56	25.87
SimpleKMeans(class)	23.72	19.55	15.04	13.33	16.22	13.06	16.23	16.59
Headset	100	200	300	400	500	600	700	800
NaiveBayes; Perc. Split (66%)	40.41	34.87	35.63	36.28	36.08	36.31	37.34	37.42
NaiveBayes; 10xCross-Validation	41.54	36.58	35.91	36.08	36.27	36.63	36.95	36.92
SimpleKMeans(class)	24.43	19.69	18.48	19.33	20.13	19.50	22.33	20.70
$SM58$	100	200	300	400	500	600	700	800
NaiveBayes; Perc. Split (66%)	30.74	27.84	28.68	28.51	28.61	28.54	28.23	28.84
NaiveBayes; 10xCross-Validation	32.23	28.31	28.20	28.35	28.58	28.65	28.77	28.91
SimpleKMeans(class)	26.49	16.13	17.30	16.88	15.90	15.23	15.79	16.43
Tbone	100	200	300	400	500	600	700	800
NaiveBayes; Perc. Split (66%)	37.26	29.44	30.09	29.55	28.95	28.40	29.24	29.26
NaiveBayes; 10xCross-Validation	38.57	31.10	29.78	29.48	29.17	28.98	29.33	29.62
SimpleKMeans(class)	18.30	14.18	12.66	13.72	10.99	12.37	11.29	12.58

When analysing the results for the Bayes classifier in the evaluations for hypothesis II a) it is obvious that the test files recorded using the *Headset* microphone allow for the most precise classification of the rooms. Here the highest accuracy of 41.54% (in the case of the Naive Bayes 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 accuracy 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 accuracy drops between 3 and 8%, if the number of feature vectors considered is increased further, no significant change in the accuracy can be noticed (i.e. it stays roughly constant).
The results for the clustering are generally lower than the results for the Bayes classification. While the best result of 26.49% classification accuracy for the room detection ($SM58$ with 100 vectors per file) is still considered significant, the lowest result of 10.99% is very close to "guessing" at the room. If the scaling of the classification accuracy with the increasing number of feature vectors per file is evaluated, it is noticed that in the case of only 100 vectors per file considered the classification accuracy for all four microphones is at its maximum. With increasing number of feature vectors the accuracy seems to drop for each microphone to an average rate about 7 to 10% below the maximum.

When comparing the results for all four microphones in the evaluations for hypothesis II a) it can be seen that in case of the Bayes classification the *Headset* shows generally the best performance in the room detection. For the clustering the $SM58$ shows the best performance. Since the results for the clustering are generally worse for the room detection than the results for Bayesian classification, the following discussions on hypothesis II are limited to the tests performed using Bayesian classification.

When comparing in the evaluations for hypothesis II b) the results for one microphone in all 10 rooms for the example of *Headset* with 100 vectors per file and Bayesian classification using percentual split (see table 8) it can be seen that the results for a correct classifications (principal diagonal of the confusion matrix in table 8) are very inhomogeneous for the rooms. The results are in the range of [12,63%] with an average of 40.41%. In seven of the ten cases the highest classification result is achieved for the room the recording was made in. In the other tree cases ($R5$, $R6$ and $R10$) a different room wrongly achieved the highest number of classifications. From table 8 also rooms showing mutually similar and dissimilar behaviour can be deduced. An example for mutually similar rooms is found with $R3$ and $R7$ where on one hand the vectors recorded in $R3$ are classified 55% belonging to $R3$ and 18% belonging to $R7$ while on the other hand the vectors recorded in $R7$ are classified 47% belonging to $R7$ and 34% belonging to $R3$. An example for a set of dissimilar rooms is composed e.g. by $R1$ and $R8$. There the number of vectors recorded in $R1$ and falsely classified as recorded in $R8$ is equal 0% and vice versa.

Figure 2 shows a histogram of the correct classified rooms (i.e. the number of correctly classified vectors is larger than the number attributed to every other room). When comparing the results for 100 and 800 vectors per file it seems that an increase of the number of vectors considered in the evaluations also increases the number of correctly classified rooms. The figure 2 shows that only in $R2$ and $R4$ the highest number of vectors is classified correctly.

Table 8: Normalised confusion matrixes for the $Headset$ microphone, 100 vectors computed per file and the Naive Bayes classifier using percentual split (66%) for set generation

$Headset$	100 vectors per file									
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	0.63	0.10	0.08	0.04	0.00	0.00	0.15	0.00	0.01	0.00
R2	0.03	0.49	0.01	0.17	0.08	0.05	0.01	0.11	0.04	0.02
R3	0.13	0.08	0.55	0.04	0.00	0.00	0.18	0.00	0.01	0.00
R4	0.00	0.24	0.00	0.52	0.04	0.03	0.00	0.08	0.07	0.02
R5	0.00	0.25	0.00	0.12	0.24	0.13	0.00	0.15	0.07	0.04
R6	0.01	0.40	0.00	0.19	0.05	0.12	0.01	0.13	0.07	0.02
R7	0.14	0.04	0.34	0.00	0.00	0.00	0.47	0.01	0.00	0.00
R8	0.00	0.05	0.00	0.13	0.08	0.03	0.00	0.54	0.13	0.05
R9	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

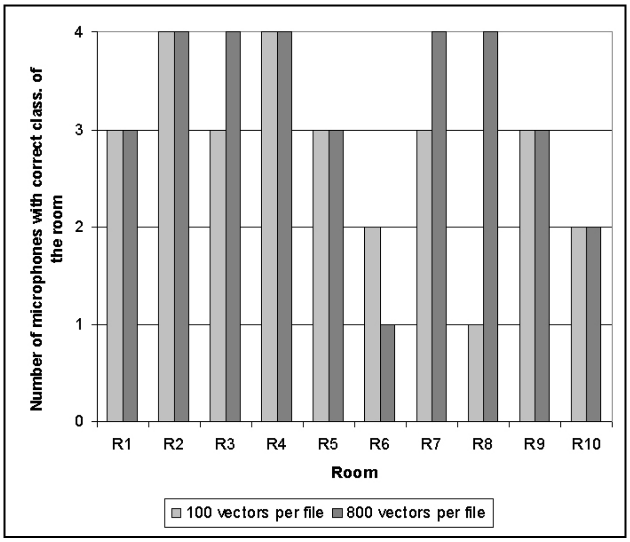

Figure 2: Histogram of the correct classified rooms

4.3 The Evaluation of Hypothesis III

The results from the Bayesian classifications and the clustering algorithm presented in sections 4.1.1, 4.1.2 and 4.2 do not only lead to different classification accuracies, they also show a different scaling behaviour. In the tests performed here the applied clustering algorithm with its parameterisation performed in every test worse than the Bayesian classification but still better than a “guessing” in the corresponding tests would perform.

What still needs to be evaluated in the scope hypothesis III is the impact of a feature selection on the classification accuracy of the selected classification strategies: In the tests described for $R1$ in section 4.1.2 the attribute selection function of WEKA identified the following features sf_i as significant (ordered by importance): sf_{ev}, $sf_{entropy}$, $sf_{LSB_{flip}}$, sf_{mean}, sf_{melf_1}, sf_{melf_3}, sf_{melf_4}, sf_{melf_8} and sf_{mel_1} for the tests with 100, 200, 300 and 400 feature vectors per file and sf_{ev}, $sf_{entropy}$, $sf_{LSB_{flip}}$, sf_{mean}, sf_{melf_1}, sf_{melf_3}, sf_{melf_4}, sf_{melf_8} for the tests with 500, 600, 700 and 800 feature vectors computed per file. It seems that the most significant features remain constant when increasing the data set. When expanding this evaluation to all rooms the result can be summarised with the fact that for all Rn the attribute selection function identifies sf_{ev}, $sf_{entropy}$, $sf_{LSB_{flip}}$, sf_{mean} and sf_{melf_1} as the five most significant features.

If the initial idea of using the attribute selection function for a reduction of the data set (in terms of features used for the classification/clustering) then it is noticed that the classification accuracy is reduced by this operation (e.g. from 57.29 to 50.52% in the case of evaluating hypothesis I with $NaiveBayes$ with percentual split (66%) in $R1$ using 100 feature vectors per file, and 38.32 to 29.64% in $R1$ with $SimpleKMeans$ using 100 feature vectors per file). In the tests performed it was noted that every reduction of the feature space reduces the classification accuracy.

When performing the same analysis on the relevance of single features $sf_i \in \mathbb{FS}$ for hypothesis II as done above for hypothesis I, for the room classification using the AKG microphone, in the case of 100 feature vectors per file the most relevant features are identified as: sf_{ev}, sf_{cv}, $sf_{entropy}$, sf_{mean} and sf_{melf_2}. If the number of feature vectors per file increases to 200-400 the number of relevant features drops to three (sf_{ev}, $sf_{entropy}$ and $sf_{melf_{22}}$). For even higher numbers of feature vectors per file the set of relevant features identified by WEKA's attribute selection function is limited to sf_{ev} and $sf_{melf_{22}}$.

For the room classification using the $Headset$ microphone, in the case of 100 feature vectors per file the most relevant features are identified as: sf_{ev}, $sf_{entropy}$, sf_{cv}, sf_{mean}, sf_{median} and sf_{melf_1}. For higher number of feature vectors per file the list shrinks to: sf_{ev}, sf_{cv}, sf_{mean}, sf_{median} and sf_{melf_1}.

For the room classification using the $SM58$ microphone, in the case of 100 feature vectors per file the most relevant features are identified as: sf_{ev}, $sf_{entropy}$ and $sf_{melf_{22}}$, above 100 features per file this list is reduced to: sf_{ev}, $sf_{entropy}$.

For the room classification using the $Tbone$ microphone, in the case of 100 feature vectors per file the most relevant features are identified as: sf_{ev}, $sf_{entropy}$, sf_{melf_1}, sf_{mel_2} and sf_{mel_3}, above 100 features per file this list changes to sf_{ev} and $sf_{mel_{12}}$.

Concluding the results for the impact of applying feature selection, it can be seen, that by increasing the number of feature vectors the number of features considered relevant by WEKA's attribute selection function decreases. All feature reductions performed within these tests had a negative impact to the classification accuracy. When comparing the tests for hypothesis I and II a very strong relevance of the feature sf_{ev} for all tests performed here has to be noted.

5. SUMMARY AND CONCLUSION

The goal of this paper was to present a first practical evaluation of microphone and environment classification as an approach for digital audio forensics. The evaluation was based on an exemplary test set of 10 different audio reference files synchronously recorded as mono audio data by four microphones in 10 different rooms with 44.1 kHz sampling rate and 16 bit quantisation. Motivated by the syntactical and semantical analysis of information and audio steganalysis, specific features were selected for classification. These features were computed for the introduced set of audio signals with the number of feature vectors defined by the shortest file. Our tests focussed on an inter-device analysis considering different device characteristics. WEKA's implementations of K-means clustering and the Naive Bayes classifier have been applied for classification.

The overall goal of our tests was to evaluate the classification performance in regard to the classification accuracy on known audio steganalysis features. Based on the test hypotheses defined in section 1, specific test goals have been defined in section 3.2. These three test goals are: determine how precise a known microphone can be identified using audio samples recorded in a set of known locations, determine how precise a known location can be identified using audio samples recorded by a set of known microphones and evaluate the impact of feature selection on the classification accuracies achieved.

Furthermore, the impact of the number of feature vectors on the classification and the relevance of single features sf_i from $\mathbb{FS}$, are evaluated. Our test results from section 4 can be summarised as follows:

For the evaluation of hypothesis I, i.e. the classification of the microphones for all rooms and a fixed number of vectors per file, the best results for the Bayesian classification (75.99% $Headset$ microphone) and K-means clustering (41.57%) are far above the percentage "guessing" the class would return (20%). Therefore, we consider the evaluations for hypothesis I as successful to that extent that even if an absolutely correct classification could not be achieved, this first evaluation shows promising results for the used feature set known from steganalysis.

If the results are evaluated exemplary for one room the scaling of the classification accuracy for the Bayesian classifiers indicated a linear connection between the number of feature vectors used from each file for the classification and the classification accuracy. Here the results of the clustering algorithm seem to be independent of the number of feature vectors supplied for the tests.

The evaluations for hypothesis II, i.e. the room classification, showed less impressive results than the microphone classification evaluated in hypothesis I. The best result here was found with 41.54% accuracy in the case of Bayesian classification, the $Headset$ and 100 vectors computed per file. The clustering with K-means resulted generally in worse accuracies than Bayes classification (about 15% worse in the maximum case; $SM58$ with 100 vectors per file). Nevertheless, these tests also lead to results which confirm hypothesis II. It is not surprising that the microphone identification is more accurate than the room identification, given that the filtering effect of a microphone is probably stronger and more unique than the filtering effect of a room environment. That both can be detected at greater than random chance show that even the room identification can be possible in practice.

For hypothesis III the initial idea of using the attribute selection function for a reduction of the data set (in terms of the number of single features sf_i from $\mathbb{FS}$ used for the classification/clustering) had to be discarded after it was noticed in the tests that the average classification accuracy is reduced by removing features which were identified as non-relevant by WEKA. Summarising the results for hypothesis III, the achieved results based on the used test sets have shown that the classifiers have performed best in all tests if the feature vectors were not reduced by a feature selection method.

In summary, our results show that for our test sets, the used classification techniques and selected steganalysis features microphones can be better classified than environments. An generalisation of the approach is of course still an open question, since the number of reference files is significantly smaller than the number of features used in the evaluation and the training and test set originate from the same very small set of audio signals.

In regard to the "Verifier-Tuple", the considered and extracted features for our tests belong to a particular information layer, the executive semantics. It is assumed to achieve better results when considering additional features of other information layers or when normalising differently. Additionally the impact of the context of the audio files (such as noise, music, etc) on the classification should be evaluated, because we can not exclude for our tests that the classification performance is influenced by pure context features.

Our actual work is concentrated on the enhancement of the test set by using more microphones and recording more samples for the tests, considering other training and test set generation strategies as well as applying other classification and clustering algorithms. The small size of the test set in regards to number of rooms, microphones, test files and classifiers implies a limited significance for the results achieved, a generalisation based on these results is not possible. Nevertheless these first evaluations demonstrate the applicability of the introduced audio steganalysis approach for audio forensics. An important step for further research is the consideration of other training and test set generation strategies (e.g. using only silence or white noise in the training of the classifiers and then classify other signals like music, thereby decorrelating the training and test sets) which would reduce the impact of eventual context dependencies and at the same time allow for a better generalisation of the classification results. Also interesting would be intra-room classifications in room identification, when the recordings are made at different locations within a room. Finally, it needs to be evaluated whether the proposed approach can be transferred to other application fields or other media with similar and/or different features.

6. ACKNOWLEDGMENTS

The authors want to thank Christian Zeitz for his work on the AAST, including the adaption to WEKA's input format. Furthermore we want to thank Rene Schuldt for the discussions on the classification problem at hand, Jörg Wissen for recording the test material and Claus Vielhauer for his idea of transferring the Mel-cepstral based signal analysis from biometric speaker verification to the domain of steganalysis. The work about applied feature computation in this paper has been supported in part by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. This work has been partly supported by the German Bundesministerium für Bildung und Forschung (BMBF) within the project NESTOR II "Expertise Vertrauenswürdige und Abgesicherte Langzeitarchivierung Multimedialer Inhalte" where our special thanks belong to Susanne Dobratz. The contents of this publication is the sole responsibility of the authors and can in no way be taken to reflect the views of the German Government.

7. REFERENCES

[1] C. Borgelt, H. Timm, and R. Kruse. Probabilistic networks and fuzzy clustering as generalizations of naive bayes classifiers. In B. Reusch and K.-H. Temme, editors, *Computational Intelligence in Theory and Practice (Advances in Soft Computing)*, pages 121–138. Physica-Verlag, Heidelberg, Germany, Heidelberg, Germany, 2001.

[2] J. Dittmann, D. Hesse, and R. Hillert. Steganography and steganalysis in Voice over IP scenarios: Operational aspects and first experiences with a new steganalysis tool set. In E. J. D. III and P. W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VII*, Electronic Imaging Science and Technology, pages 607 – 618, San Jose, California, USA, 2005. SPIE and IS&T, SPIE. ISBN 0-8194-5654-3.

[3] J. Fridrich, J. Lukas, and M. Goljan. Digital camera identification from sensor noise. In *IEEE Transactions on Information Security and Forensics, June 2006, Vol. 1(2)*, pages 205–214. IEEE, 2006.

[4] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[5] 2007. Andrew Moore: K-means and Hierarchical Clustering - Tutorial Slides, http://www-2.cs.cmu.edu/ awm/tutorials/kmeans.html.

[6] N. Khanna, A. K. Mikkilineni, A. F. Martone, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp. A survey of forensic characterization methods for physical devices. In *Proceedings of the 6th Digital Forensics Research Workshop (DFRWS), Lafayette, Indiana, USA, August 2006, 17-28*, 2006.

[7] C. Kraetzer and J. Dittmann. Mel-cepstrum based steganalysis for voip-steganography. In E. J. Delp and P. W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents IX*, Electronic Imaging Science and Technology, SPIE Vol. 6505, San Jose, CA, USA, 2007. SPIE and IS&T, SPIE.

[8] C. Kraetzer, J. Dittmann, and A. Lang. Transparency benchmarking on audio watermarks and steganography. In *SPIE conference, at the Security, Steganography, and Watermarking of Multimedia Contents VIII, IS&T/SPIE Symposium on Electronic Imaging, 15-19th January, 2006, San Jose, USA*, 2006.

[9] J. Lukas, J. Fridrich, and M. Goljan. Determining digital image origin using sensor imperfections. In *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VII, IS&T/SPIE Symposium on Electronic Imaging, 16-20th January, 2005, San Jose, USA*, pages 249–260, 2005.

[10] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. Berkeley, University of California Press, 1967.

[11] A. Mikkilineni, O. Arslan, P.-J. Chiang, R. Kumontoy, J. Allebach, and G.-C. Chiu. Printer forensics using SVM techniques. In *Proceedings of the IS&Ts NIP21: International Conference on Digital Printing Technologies, Baltimore, USA, October 2005, Vol. 21*, pages 223–226, 2005.

[12] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T. C. Chiu, J. P. Allebach, and E. J. Delp. Printer identification based on graylevel co-occurrence features for security and forensic applications. In *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VII, IS&T/SPIE Symposium on Electronic Imaging, 16-20th January, 2005, San Jose, USA*, pages 430–440, 2005.

[13] A. Oermann, A. Lang, and J. Dittmann. Verifyer-tupel for audio-forensic to determine speaker environment. In *Proceedings of the ACM Multimedia and Security Workshop 2005*, pages 57 – 62, New York, USA, 2005. ACM.

[14] A. Oermann, C. Vielhauer, and J. Dittmann. Digitale Handschrift: Extraktion Gerätespezifischer Merkmale. In *Proceedings of D-A-CH Security 2007*, 2007.

[15] A. Oermann, C. Vielhauer, and J. Dittmann. Identifying pen digitizers by statistical multimedia signal processing. In *Proceedings of SPIE Electronic Imaging - Multimedia on Mobile Devices III*, number SPIE Vol. 6507 in Electronic Imaging Science and Technology, San Jose, California, USA, 2007. SPIE and IS&T, SPIE.

[16] 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.

[17] 2007. WEKA 3 http://www.cs.waikato.ac.nz/ml/weka/.

[18] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Table 9: Normalised confusion matrixes for all four microphones, two different numbers of vectors computed per file (100 and 800) and the Naive Bayes classifier using perc. split (66%)

R1	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.27	0.16	0.46	0.03	0.08	0.59	0.20	0.12	0.08	0.00
Headset	0.00	0.83	0.12	0.05	0.00	0.00	0.78	0.19	0.03	0.00
SM58	0.02	0.01	0.86	0.12	0.00	0.00	0.30	0.52	0.18	0.00
Tbone	0.19	0.05	0.51	0.24	0.01	0.05	0.24	0.16	0.55	0.00
Original	0.18	0.07	0.07	0.05	0.63	0.22	0.06	0.04	0.05	0.62

R2	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.27	0.10	0.47	0.08	0.08	0.39	0.13	0.33	0.15	0.00
Headset	0.01	0.94	0.00	0.04	0.00	0.01	0.92	0.05	0.03	0.00
SM58	0.03	0.04	0.80	0.12	0.00	0.01	0.02	0.70	0.27	0.00
Tbone	0.03	0.07	0.09	0.76	0.05	0.04	0.02	0.30	0.64	0.00
Original	0.03	0.01	0.11	0.05	0.80	0.06	0.01	0.08	0.05	0.81

R3	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.32	0.03	0.46	0.10	0.08	0.65	0.12	0.16	0.07	0.00
Headset	0.01	0.84	0.07	0.08	0.00	0.00	0.86	0.12	0.02	0.00
SM58	0.03	0.25	0.53	0.19	0.00	0.01	0.31	0.46	0.22	0.00
Tbone	0.16	0.21	0.37	0.24	0.02	0.08	0.24	0.14	0.53	0.00
Original	0.21	0.09	0.05	0.05	0.61	0.27	0.07	0.04	0.02	0.61

R4	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.61	0.01	0.18	0.17	0.03	0.49	0.01	0.25	0.24	0.00
Headset	0.00	0.91	0.08	0.01	0.00	0.00	0.92	0.07	0.01	0.00
SM58	0.03	0.06	0.62	0.29	0.01	0.09	0.06	0.59	0.26	0.00
Tbone	0.07	0.03	0.10	0.79	0.01	0.05	0.02	0.07	0.86	0.00
Original	0.03	0.01	0.03	0.10	0.84	0.05	0.01	0.03	0.09	0.82

R5	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.28	0.08	0.47	0.11	0.07	0.39	0.11	0.33	0.17	0.00
Headset	0.00	0.96	0.01	0.01	0.02	0.01	0.96	0.02	0.01	0.00
SM58	0.02	0.00	0.80	0.18	0.00	0.01	0.00	0.67	0.32	0.00
Tbone	0.05	0.03	0.11	0.76	0.05	0.04	0.01	0.35	0.60	0.00
Original	0.03	0.01	0.09	0.07	0.80	0.05	0.01	0.08	0.06	0.80

R6	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.45	0.08	0.28	0.12	0.07	0.42	0.09	0.33	0.15	0.00
Headset	0.01	0.94	0.01	0.02	0.02	0.01	0.89	0.06	0.02	0.01
SM58	0.01	0.01	0.81	0.17	0.00	0.01	0.00	0.74	0.25	0.00
Tbone	0.05	0.02	0.12	0.76	0.05	0.02	0.01	0.35	0.62	0.00
Original	0.03	0.01	0.09	0.07	0.79	0.07	0.01	0.08	0.04	0.79

R7	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.57	0.00	0.25	0.11	0.07	0.74	0.03	0.17	0.06	0.00
Headset	0.00	0.87	0.07	0.06	0.00	0.00	0.96	0.03	0.01	0.00
SM58	0.01	0.03	0.84	0.12	0.00	0.00	0.23	0.63	0.14	0.00
Tbone	0.09	0.15	0.40	0.31	0.04	0.05	0.20	0.20	0.56	0.00
Original	0.20	0.08	0.06	0.04	0.62	0.28	0.06	0.03	0.03	0.60

R8	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.45	0.07	0.26	0.17	0.06	0.68	0.06	0.13	0.12	0.00
Headset	0.03	0.94	0.00	0.01	0.01	0.03	0.93	0.02	0.02	0.00
SM58	0.00	0.01	0.89	0.10	0.00	0.00	0.00	0.90	0.10	0.00
Tbone	0.09	0.01	0.50	0.36	0.04	0.03	0.00	0.33	0.63	0.00
Original	0.17	0.01	0.08	0.11	0.63	0.24	0.01	0.07	0.08	0.60

R9	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.39	0.07	0.34	0.15	0.05	0.49	0.08	0.32	0.11	0.00
Headset	0.00	0.99	0.00	0.00	0.00	0.01	0.96	0.01	0.01	0.00
SM58	0.04	0.01	0.80	0.14	0.00	0.02	0.00	0.79	0.18	0.00
Tbone	0.09	0.04	0.20	0.64	0.04	0.06	0.01	0.35	0.58	0.00
Original	0.07	0.02	0.09	0.05	0.77	0.08	0.02	0.08	0.04	0.78

R10	100 vectors per file					800 vectors per file				
	AKG	*Headset*	*SM58*	*Tbone*	*Original*	*AKG*	*Headset*	*SM58*	*Tbone*	*Original*
AKG	0.39	0.06	0.47	0.02	0.06	0.60	0.06	0.29	0.04	0.00
Headset	0.06	0.89	0.00	0.04	0.01	0.04	0.88	0.01	0.06	0.01
SM58	0.04	0.02	0.89	0.05	0.00	0.01	0.01	0.91	0.07	0.00
Tbone	0.08	0.06	0.11	0.71	0.04	0.06	0.03	0.17	0.74	0.00
Original	0.15	0.01	0.08	0.03	0.73	0.23	0.01	0.08	0.04	0.65