A Context Model for Microphone Forensics and its Application in Evaluations

Christian Kraetzer, Kun Qian, Maik Schott and Jana Dittmann
Otto-von-Guericke University of Magdeburg, PO Box 4120, 39016 Magdeburg, Germany
Contact email: christian.kraetzer@iti.cs.uni-magdeburg.de

ABSTRACT

In this paper we first design a suitable context model for microphone recordings, formalising and describing the involved signal processing pipeline and the corresponding influence factors. As a second contribution we apply the context model to devise empirical investigations about: a) the identification of suitable classification algorithms for statistical pattern recognition based microphone forensics, evaluating 74 supervised classification techniques and 8 clusterers; b) the determination of suitable features for the pattern recognition (with very good results for second order derivative MFCC based features), showing that a reduction to the 20 best features has no negative influence to the classification accuracy, but increases the processing speed by factor 30; c) the determination of the influence of changes in the microphone orientation and mounting on the classification performance, showing that the first has no detectable influence, while the latter shows a strong impact under certain circumstances; d) the performance achieved in using the statistical pattern recognition based microphone forensics approach for the detection of audio signal compositions.

1. MOTIVATION AND INTRODUCTION

The past years have seen significant advances in digital image forensics. An overview of currently established authentication approaches for this domain is given by Hany Farid[5]. In contrast to image forensics, in the field of audio forensics so far only a limited number of approaches can be found, even though audio forensics can be considered to be very interesting for application scenarios where trust in authenticity and integrity of audio signals might be required, e.g. for evidences in court cases or in the ingest phase of secure digital long term archives. The currently existing approaches for microphone forensics (MF; a.k.a. recording forensics or recording source forensics) – as one of the most important sub-categories in audio forensics, can be classified into three classes:

ENF-based approaches: One quite mature, but physically complex approach found in literature (e.g. Grigoras[7]) is the usage of the electric network frequency (ENF) in recordings to evaluate digital audio authenticity. The complex electro-physical requirements for this approach are summarized by Grigoras et al.[8].

Time domain and local phenomena based evaluations: In 2010 Malik and Farid[2] describe a technique to model and estimate the amount of reverberation in an audio recording. Because reverberation depends on the shape and composition of a room, differences in the estimated reverberation can be used in a forensic setting for authentication. The usage of similar characteristics can be found in closely related research fields like e.g. in the works from Maher[9] on gunshot characterization. Yang et al.[6] introduced a format conversion dependent method for locating forgeries (insertions and deletions) in MP3 files by time domain based analyses of encoder frame offsets.

Statistical pattern recognition based approaches: We introduced in 2005 a statistical pattern recognition based approach for microphone identification[14] which was in 2007 substantiated by a fist practical evaluation[1]. In contrast to the other two classes of approaches this one is interesting for two reasons: on one hand it is not dependent on the existence of local phenomena (like e.g. reverberations), on the other hand it can actually generate domain knowledge and thereby can answer still open research questions on the signal under observation. Böhme and Westfeld[13] introduced a statistical pattern recognition based approach for the identification of the encoder used to generate an MP3 file based on features computed on the modelling layer of the file.

In this paper we extend the current state-of-the-art by investigations work described by Oermann et al.[14] and Kraetzer et al.[1]. As a first important step we design a suitable context model for microphone recordings, formalising and describing the involved 5-stage recording process pipeline. Second, we apply the context model to devise empirical investigations aiming at the generation of required domain knowledge.

These questions about the provenance, persistence and uniqueness of a sensor patterns in microphones are raised by previous work in this field[15]. The answers are generated within this paper by systematic empirical evaluations based on the introduced context model and suitable hardware setups. As feature extractor in our statistical pattern recognition based approach we use our AAFE (AMSL Audio Feature Extractor[11]) in its current version v.2.0.5. The extractor computes in this version 590 intra-frame features: 9 in time domain (zero crossing rate, energy, pitch, RMS-amplitude, entropy, LSB-ratio, LSB-fliprate, mean, median), 529 in frequency domain (spectral centroid, spectral roll-off, two

differently computed spectral bandwidths, spectral irregularity, spectral entropy, 11 formants and base frequency, as well as a 512 frequency component histogram) and 52 in Mel-cepstrum domain (MFCCs, FMFCCs, as well as 2nd order derivative MFCCs and FMFCCs, which have been added due to the good results achieved by Liu et al.[10] with the 2nd order derivative MFCCs). For the classification we use the renowned data mining suite WEKA[12] in version 3.6.1.
The generated domain knowledge (answers to open research questions) includes the answers to the following test goals:
a) **Identification of suitable classification algorithms for this forensics approach:** Here we show for a specific intra-class test set that 4 out of 74 classifiers from WEKAs (untuned) supervised classifiers can achieve classification accuracies between 80 and 82.5% and 27 further classifiers report accuracies between 60 and 80%. Considering the different classes of classifiers used, it can be summarized that the used meta-classifiers give the best results. Within the top 20 of the ranked classifiers only few tree-based classifiers or functions can be found. One further important observation made here is that the second main class of classification algorithms, the clustering algorithms (a.k.a. unsupervised classifiers) seem have absolutely no relevance for MF.
b) **Determination of suitable features for the statistical pattern recognition:** Here we achieve very good performance for the 2nd order derivative MFCC based features[10], as well as a good performance for selected time domain, frequency domain and FMFCC based features.
c) **Determination of the influence of changes in the microphone orientation and the microphone fixing on the classification performance and on the classification performance:** Our results for the investigations addressing this test goal show that the microphone orientation seems to have no impact to the classification behaviour, while a change of the mounting of a microphone can have a very strong influence if it affects its the reverberation behaviour.
d) **Accuracy achieved using the statistical pattern recognition based microphone forensics approach for detection of audio file compositions:** This question is investigated here for the different scenarios which might occur. Generally two distinct types of scenarios can be identified in this context: first, the audio data stream, into which other data is pasted into, originates from a known microphone and second, 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 MF, where we usually assume that we intend to verify the identity of a source microphone. Nevertheless the performance of the statistical pattern recognition based forensics approach used within this thesis on the less likely second scenario is also evaluated here to show its limitations. So, four different tests are performed in this composition detection evaluation:

1. Microphone recordings of one known microphone made in different locations composed into one stream
2. One known microphone pasted into a stream of completely different known microphone
3. One unknown microphone pasted into a stream of completely different known microphone
4. One unknown microphone pasted into a stream of completely different unknown microphone

The results achieved for this statistical pattern recognition based audio file composition show relatively good results on the first three tests and sub-optimal, but still significant results for the last test.

The rest of this paper is structured as follows: the second section describes the recording process pipeline used for generation of audio signals and derives a suitable context model. The section 3 uses the context model to devise the experiments necessary to answer the identified research questions, while section 4 presents the results for those experiments. In the final section the document is summarized and possible directions for future work are indicated.

2. A CONTEXT MODEL FOR MICROPHONE FORENSICS

A context model describing the audio recording process helps to understand the influences to the audio signal during these processes. As shown in Figure 1, an audio recording process within this thesis is described using a pipeline which consists of five segments.

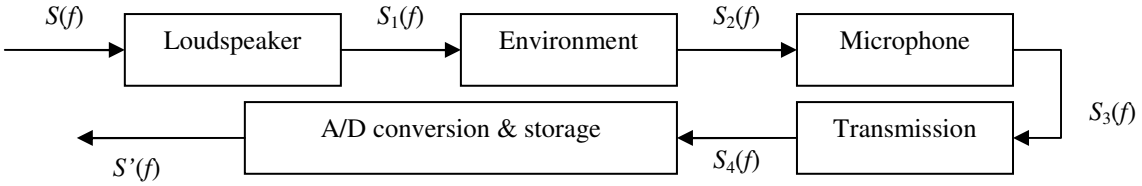

Figure 1: Recording process pipeline – context model

Audio signals can be considered as either continuous or discrete signals in frequency-domain. Let a function $S(t)$ denote the original audio signal in time-domain, thus $S(f)$, its representation in frequency-domain can be easily achieved by a Fourier transformation:

$$S(f) = FT(S(t)) \tag{1}$$

In Figure 1, $S_1(f)$, $S_2(f)$, $S_3(f)$ and $S_4(f)$ denote the analogue audio signals after each processing segment, while $S'(f)$ and its time domain counterpart $S'(t)$ computed via inverse Fourier transform as $S'(t) = \mathrm{FT}^{-1}(S'(f))$ denote the final audio signal as the result. Then a context model described as follows has to consider these processes.

$$S_1(f) = \sum_{N_{driver}} \int_l^u F_{driver}(f) S(f) df + N_{ls}(f) \tag{2}$$

A typical loudspeaker consists of multiple drivers[3], as individual electrodynamic drivers provide quality performance over at most about three octaves. Equation (2) simulates the process of a loudspeaker with multiple 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, 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.

$$S_2(f) = S_1(f) * F_{echo}(f) + N_{envi}(f) \tag{3}$$

There are mainly two aspects of possible distortions which occur to the played audio signal before it is collected by the microphone. In equation (3), a convolution of $S_1(f)$ and $F_{echo}(f)$, which describes the objects in the environment that reflect the played signal, is used to simulate the possible distortion caused by echoes or reverberations[4]. The consistency of this convolution of $S_1(f)$ and $F_{echo}(f)$ is the characteristic verified for the recording forensics approach by Malik et al.[2]. When the recording process is accomplished in an anechoic environment, then $F_{echo}(f)$ 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.

$$S_3(f) = \int_{spectrum} F_{mic}(f) S_2(f) df + N_{mic}(f) + N_{ENF}(f) \tag{4}$$

Equation (4) simulates the process of a microphone collecting the signal. $F_{mic}(f)$ 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 MF forensics).
We assume for our approach that the specificity of a microphone is decided by the characteristics (*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 O (orientation), M (mounting and A (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:

$$F_{mic}(f) = F_{\inf}(O, M, A) \cdot F_{membrane}(MembCharacteristics) \tag{5}$$

Usually $N_{mic}(f)$ can be considered as a constant as it contributes a rather minor influence on the recorded signal compared to that $F_{mic}(f)$ does.

$$S_4(f) = \int_{spectrum} F_{tran}(f) S_3(f) df + N_{tran}(f) \tag{6}$$

The component $F_{tran}(f)$ in equation (6) denotes the distortion during the transmission of the signal from the microphone to the A/D conversion device, while $N_{tran}(f)$ denotes the thermal noise coming from the transmission environment.

$$S'(f) = \int_0^{f_N} F_{samp}(f) S_4(f) df + N_{quan}(f) + N_{thernal}(f) \tag{7}$$

Equation (7) describes the process of storing the audio as an audio file. In the equation f_N denotes the Nyquist frequency, $N_{quan}(f)$ denotes the quantisation noise, and $N_{thermal}(f)$ the thermal noise of the A/D device.

3. TEST SETUPS FOR THE PRACTICAL INVESTIGATIONS

The practical investigations within this paper 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 (AAST / AAFE in version v.2.0.5; for a description of the feature extractor and the features see Kraetzer et al.[11]). All the 74 supervised classification techniques as well as the eight clustering algorithms provided by WEKA are evaluated within this paper using their default parameters (except for clusterers, where the number of clusters is adjusted in each experiment to the number of classes under observation). The chosen feature extractor AAST / AAFE v.2.0.5 is used within all practical investigations to compute a 590 dimensional segmental feature vector for each of 200 (except for the composition detection tests where only 100 windows are used for each mesh-up) consecutive windows per presented recording. The window size for these non-overlapping windows is per default set to 1024 samples, as the windowing function the Dirichlet window is used. All supervised classification experiments are either performed using 10-fold cross-validation or explicitly supplied training- and test sets, all clusterings are performed as classes to clusters evaluation.

Table 1 gives an overview over the 10 experimental setups used within this work. The goals of these experiments are either classifier benchmarking and selection, feature selection, orientation and mounting influence determination or audio file composition detection. **A more detailed description of the experimental setups is given in Table 9 at the end of the document!**

Table 1: Overview over the used experimental setups (detailed description in Table 9 at the end of the document)

Setup	Goal	Relevant components of the context model
Classifier-Benchmarking-RS4_Rode	classifier selection	*MembCharacteristics*
Classifier-Benchmarking-RS4_Beyer		
Feature-Selection	feature selection	*MembCharacteristics*
Classifier-Benchmarking-RS4_Rode-Best20Features-only	classifier / feature selection	*MembCharacteristics*
Orientation_Impact_RS7	orientation influence	*O*
Mounting_Impact_RS9	mounting influence	*M*
Composition-1	audio file composition detection	*MembCharacteristics*
Composition-2		
Composition-3		
Composition-4		

For most of the investigations the microphone (membrane) characteristic is the part of the context model under investigation, but two of the setups aim on determination of the influence of the orientation and the mounting performance of the considered MF approach.
In the evaluations performed the loudspeaker for the playback of $S(f)$ is kept constant, so that the impact of the amplifying functions $F_{driver}(f)$ as well as $N_{ls}(f)$ from equation (2) remain constant. Additionally a fixed set of reference signals $S(f)$ is used for all evaluations to keep the signal dependent driver response in equation (2) constant.
For the practical investigations fixed locations are used for the generation of the recording sets. Due to this fact we assume here the influences of $F_{echo}(f)$ and $N_{envi}(f)$ in equation (3) to be constant and negligible.
For storage we choose with PCM coded WAV files in CD quality (sampled with 44.1kHz and 16 Bit quantization) one of the most common audio signal format. This uncompressed format, on one hand combines a suitably high F_{samp} together with a well limited $N_{quan}(f)$ and on the other hand will allow us in future work to evaluate the impact of down-sampling or MP3 compression on the classification results. Again, the thermal noise of the A/D device ($N_{thermal}(f)$) is here assumed to be constant and negligible.

As **constraints** for the practical investigations, the allocated memory for all test on the test machine was set to 1.5 GByte RAM, additionally, strict timeout boundaries of 12h (=43,200s) for the clustering algorithms and 60h (=216,000s) for the (supervised) classification algorithms are defined for all tests.
The sets of **recorded material** used for the practical investigations within this paper are identified and described in Table 2 below (for a description of the recording environments / rooms used see Kraetzer et al.[1]).

Table 2: Microphone recording sets used in this paper

ID	description
RS2	***RS2*** is the recording set used for the generation of the test results in Buchholz et al.[16]. It contains seven different microphones which are used for sequential recording (all on a Sound Blaster USB as pre-amplifier): *M2* - Terratec Headset Master dynamic microphone, *M3* - Shure SM58 dynamic microphone, *M5* - PUX 70TX-M1 piezoelectric microphone, *M6* - T.bone MB45 dynamic microphone, *M7* - AKG CK93 condenser microphone, *M8* - AKG CK98 condenser microphone, and *M9* - T.bone SC600 condenser microphone
RS4	The ***RS4*** consists of two well distinct subsets *RS4_Rode* and *RS4_Beyer*, both containing 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 Røde NT6 condenser microphones (*M16*, *M17*, *M18*, and *M19*) while *RS4_Beyer* is a homogeneous set of four Beyerdynamic Opus 69 dynamic microphones (*M20*, *M21*, *M22*, and *M23*). The ten reference files used for the generation of RS2 and RS4 are described in detail by Kraetzer et al.[1]. Thereby, the tests performed on *RS4* will 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 *RS2*.
RS7	The recording set ***RS7*** is recorded in the anechoic chamber (room *R06*), using a Beyerdynamic Opus 69 microphone (*M22*), a t.bone SC600 spherical (*M9*) and a t.bone SC600 cardioid (*M9b*). The latter two microphones are the same device but with different directional characteristics. For each of those three microphones, two different reference sounds (a harmonic sinusoid at 440Hz and silence) are recorded in eight different microphone orientations, each 45° turned in the xy-plane from its predecessor, stating with the orientation directly towards the sound generating loudspeaker.
RS9	The recording set ***RS9*** is recorded the anechoic chamber (room *R06*) using a Beyerdynamic Opus 69 microphone (*M22*). With this microphone two different reference sounds (a harmonic sinusoid at 440Hz and silence) are recorded in eight different microphone mounting positions. The distance (50cm) and orientation to the loudspeaker are kept constant in these tests.

For the evaluation the classification performance, the following **metrics** are defined:
The **accuracy** in a classification is the number of correctly classified instances divided by the number of overall instances. If a classifier displays in an *n*-class classification problem an accuracy of 1/*n* (which is problem identical to the probability of guessing correctly) it can be considered to be completely unsuitable for the classification task.
The **classification gain** *cg* is derived from the accuracy, but takes the number of classes in the classification into account. It is computed for an *n*-class classification problem as $cg = (accuracy-(1/n)) / (1-(1/n))$.
The **runtime** of an experiment is in this paper expressed relative to the corresponding timeout boundary. Therefore the runtime for a clustering algorithm $t_{cu} = time$ / 43,200s and the run-time for a classification algorithm is $t_{cl} = time$ / 216,000s, with *time* being the time duration of the corresponding experiment on the test machine (for reasons of reliability measured using the Unix "time" command instead of WEKAs own time measurements).
The **classifier quality** is expressed in this paper as a distance from an optimal performance point. The optimal performance of a classifier would be obviously a perfect (100% *accuracy*) decision generated instantly (in no time). Here we transfer this principle by measuring for the supervised classifiers the Euclidean distance *optDist* from the optimum point in a runtime-*cg* diagram as: $optDist = \sqrt{t_{cl}^2 + (1-cg)^2}$
For **the performance of a single feature** from the feature vector no explicit metric is used. Here instead a two-stage ranking fusion is performed. The first fusion is done by performing and unweighted averaging for each feature on the ranked outputs of the five used evaluators (WEKAs ChiSquaredAttributeEval, FilteredAttributeEval, InfoGainAttributeEval, OneRAttributeEval, and SymmetricalUncertAttributeEval) per ranking block. Thereby a ranking block is composed from one recording set split into subsets for its individual recording locations.
The second fusion is done by unweighted averaging of the output of the individual ranking blocks processed in the first fusion step.
Within this paper the two intra-class recording sets *RS4_Beyer* and *RS4_Rode* are used for the feature performance evaluations. Each of those two sets will be considered in the feature ranking as one ranking block with ten individual recording locations for which the rankings are computed and then fused by averaging. In the second step the ranking lists of the two ranking blocks are averaged again to generate the final ranking.
For the audio file composition detection (see section 1) another set of metrics is defined with the **change rate** and **the average sequence length**. The change rate identifies how often in a pre-defined sequence of frames a classifier changes its opinion on the class of the audio material under observation. The average sequence length tells how long the average sequence is for which a classifier returns as an answer the same class.

4. RESULTS OF THE EXPERIMENTAL VALIDATION

Here the results for the practical investigations on the test goals defined in section 1 of this paper are performed. The subsection 4.1 focuses on classifier selection, 4.2 on feature selection, 4.3 on the influence of different orientations, 4.4 on the influence of different microphone mountings and 4.5 on the investigations on audio file compositions.

4.1. Classifier selection

The evaluations performed by Kraetzer et al.[1] indicated that (supervised) classification outperforms clustering for microphone classification. These observations are substantiated within this section, where all clustering and classification algorithms implemented in WEKA (v.3.6.1) are reviewed for their performance in microphone classification.

Tests performed for **clustering algorithms** in the preparation of this paper completely confirmed the findings of Kraetzer et al.[1]. In these tests all eight clustering algorithms provided by WEKA (v.3.6.1) were tested in setups identical to *Classifier-Benchmarking-RS4_Rode* and *Classifier-Benchmarking-RS4_Beyer* (see Table 9 at the end of the document), but using classes to clusters evaluation (with the number of clusters set to the number of classes in the training and test sets) instead of 10x stratified cross-validation. None of the clustering algorithms was able to show an accuracy better than 32.675% in these four-class problems (equivalent to cg=0.102). A reduction to the 20 most significant features (see section 4.2, Table 4) resulted in a further decrease of the achieved maximum classification accuracy to 27.6%.

Due to the large number of supervised classification algorithms in WEKA (74 in the current version v.3.6.1) it is hardly feasible to run all experiments within this paper with all classifiers. Therefore the experimental validations presented are usually carried out using only a subset of all available classification methods. To allow for any generalisation of those results nevertheless an extensive and comprehensive benchmarking of the performance of all those **(supervised) classification** algorithms has to be performed. To summarise these benchmarking efforts regarding the classification accuracy and run-time complexity of the classifiers is the task of this section.

To aim for maximum generalisability of the observations extensive intra-class practical evaluations close to the constraints boundary imposed by WEKA (maximum locatable memory within 32-Bit JAVA runtime environments) are performed using the complete 590 dimensional feature space provided by the segmental features (in preliminary tests the 17 global features also extracted by AAST / 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) in AAST / AAFE v.2.0.5. With the two audio recording sets *RS4_Rode* and *RS4_Beyer* introduced in section 3 two representative sets for classifier performance evaluation exist for usage within this paper. On these two sets all 74 classifiers available in WEKA v.3.6.1 are used in 10x 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 * 200 feature vectors per reference * 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 audio recording sets.

A timeout of 60 hours was defined at which a classifier is terminated if it has not finished until that point. The overall run time for these experiments on the test machine (an Intel Core 2 Duo E8400 CPU @3GHz with 4 GB RAM machine running WEKA v.3.6.1 with 1.5 GByte allocated RAM for each classifier.) 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 classification accuracy achieved on the material from *RS4_Beyer* imply that it proposes a somewhat easier intra-class classification problem than the microphone classification on *RS4_Rode*.

Classifier benchmarking for suitability in Microphone Classification on *RS4_Rode*: Summarising the benchmarking results on *RS4_Rode* generated by using the experimental setup *Classifier-Benchmarking-RS4_Rode*, it is shown in Table 3 (a) that the evaluated classifiers show a strong variation in their classification behaviour regarding the achieved accuracies.

In 18 out of the 74 cases the classification attempt terminated with an error. These errors can be summarised as being either timeouts, memory shortage, unsupported attribute types or missing helper data (e.g. cost files in case of cost sensitive classifiers).

For the 56 non-error cases shown in Table 3 (a) nine can be considered as "just guessing" at the true class of a sample (accuracy between 25% and 27% which would be equivalent to a classification gain close to 0) and therefore completely unsuitable for MF. The other classifiers in their default parameterisations show accuracies up to 75.88% (average over all ten rooms for meta.RotationForest).
If the run-time of the 56 non-error cases is considered, as shown in Table 3 (b), than it has to be admitted that the four "timeout" cases mentioned above constitute extreme outliers. While in those four cases the test is terminated after 60h (216000 seconds) none of the other classifiers took more than 93109 seconds (average for lazy.LWL).

Table 3: Accuracies and errors (a) as well as time measurements (b) for experiment *Classifier-Benchmarking-RS4_Rode*; Accuracies and errors (c) as well as time measurements (d) for experiment *Classifier-Benchmarking-RS4_Beyer*

	Average over all 10 rooms
Maximum achieved accuracy	75.88%
Time duration without timeouts (s)	361953.1
Duration including timeout test cases (s)	1225953.1
Errors	18
25<=accuracy<27%	9
27<=accuracy<40%	12
40<=accuracy<60%	11
60<=accuracy<80%	24
80<=accuracy<90%	0
accuracy>=90%	0

(a) accuracies and errors

time<10s	7
10<=time<100s	11
100<=time<1000s	12
1000<=time<10000s	20
10000<=time<100000s	6
>=100000s	0

(b) time measurements

	Average over all 10 rooms
Maximum achieved accuracy	82.51%
Time duration without timeouts (s)	280287.7
Duration including timeout test cases (s)	1144287.7
Error	18
25<=accuracy<27%	8
27<=accuracy<40%	8
40<=accuracy<60%	8
60<=accuracy<80%	28
80<=accuracy<90%	4
accuracy>=90%	0

(c) accuracies and errors

time<10s	3
10<=time<100s	16
100<=time<1000s	15
1000<=time<10000s	16
10000<=time<100000s	6
>=100000s	0

(d) time measurements

Classifier benchmarking for suitability in Microphone Classification on *RS4_Beyer*: Summarising the benchmarking results on *RS4_Beyer* generated by using the experimental setup *Classifier-Benchmarking-RS4_Beyer*, it can be seen that the experimental results are similar in distribution but marginally better than those discussed above for *Classifier-Benchmarking-RS4_Rode*. The achieved maximum classification accuracy averaged over all ten rooms is with 82.51% about 7% higher than for the *RS4_Rode* microphones.
Table 3 summarises in (c) the achieved classification accuracies for this experiment. It can be seen that not only the maximum achieved accuracy is higher but also more individual classifiers perform better, even with four classifiers in the range between 80% and 90% which could not be achieved on the *RS4_Rode* material even once. The erroneous behaviour of 18 classifiers is exactly the same (also for the same reasons) as for *Classifier-Benchmarking-RS4_Rode*.
If it comes to the run-time requirement of the classifiers (Table 3 part (d)), again the behaviour on *Classifier-Benchmarking-RS4_Beyer* is similar to the behaviour on *Classifier-Benchmarking-RS4_Rode*, even when the middle run-times are a little bit more dominant.
As described in section 3 the performance of a classifier is measured in this paper using the quality metric *optDist*. Comparing the classifier ranking for *Classifier-Benchmarking-RS4_Rode* (a detailed listing of the classifier ranking is presented as additional material to this paper on) http://omen.cs.uni-magdeburg.de/itiamsl/mitarbeiter/christian-kraetzer/publications.html) and *Classifier-Benchmarking-RS4_Beyer* it can be seen that 19 out of the top 20 are present in both rankings. The differences are *trees.FT* (2^{nd} for *Classifier-Benchmarking-RS4_Beyer*, but 32^{th} for *Classifier-Benchmarking-RS4_Rode*) and *trees.J48* (21^{th} in *Classifier-Benchmarking-RS4_Beyer* and 18^{th} in *Classifier-Benchmarking-RS4_Rode*). Interestingly both are decision tree classifiers, a class which contains 12 out of the overall

74 classifiers, but which shows no significant influence in the first ten ranks of the classifier ranking. The two most dominant classes of classifiers in this set are *meta* classifiers and *functions*, all other classes show only limited significance.

4.2. Feature selection

Regarding the question of suitable features for MF[1] 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 the microphone seems to have no positive impact on the classification accuracy, but it reduces computation times and generates domain knowledge. These first results on feature selection from Kraetzer et al.[1], which are based on the first inter-class statistical classification in this field, are further substantiated within this paper. 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 AAST / AAFE v.2.0.5 have been used as input for all 74 classifiers in WEKA and none of those features showed any significance in this application scenario. Therefore those global features are neglected for the MF observations in this paper.

Table 4: Best 30 features, based on the fused rankings computed on *RS4_Rode* and *RS4_Beyer* (setup *Feature-Selection*)

	RS4_Rode	***RS4_Beyer***		
Feature	Average Rank	Average Rank	**Arithmetic mean of the ranks**	**Final Rank**
FMFCC_D_1	2.08	1.06	1.57	1
FMFCC_D_2	2.6	2.72	2.66	2
FMFCC_D_13	4.18	11	7.59	3
FMFCC_D_10	14.42	9.06	11.74	4
FMFCC_D_3	18.18	5.48	11.83	5
FMFCC_D_5	14.7	9.74	12.22	6
FMFCC_D_4	18.26	6.34	12.3	7
FMFCC_D_11	18.02	6.68	12.35	8
FMFCC_D_12	19.22	5.94	12.58	9
FMFCC_D_9	19.98	12.46	16.22	10
FMFCC_D_6	20.24	13.04	16.64	11
FMFCC_D_8	22.34	12.02	17.18	12
FMFCC_D_7	22.62	12.08	17.35	13
FMFCC_3	16	27.94	21.97	14
FMFCC_12	15.9	28.32	22.11	15
SPEC_11	20.14	24.98	22.56	16
rms_amplitude	19.06	26.78	22.92	17
FMFCC_10	13.46	33.1	23.28	18
FMFCC_5	13.66	33.16	23.41	19
FMFCC_1	20.64	29.76	25.2	20
FMFCC_4	15.86	39.88	27.87	21
FMFCC_11	15.7	40.16	27.93	22
FMFCC_2	22.94	34.84	28.89	23
energy	19.54	38.44	28.99	24
FMFCC_9	18.1	41.76	29.93	25
spectral_entropy	14.72	46.76	30.74	26
FMFCC_6	18.54	42.98	30.76	27
SPEC_12	32.48	32.62	32.55	28
zcr	48.7	17.16	32.93	29
spectral_rolloff	22	49.1	35.55	30

Feature selection by feature ranking: The feature selection on segmental features for their suitability in MF is performed as described in section 3 above. As described there for the practical two-stage realisation of the test design, as two independent information sources the recording sets *RS4_Beyer* and *RS4_Rode* are chosen (see experimental setup *Feature-Selection*) because their material is best suited for generalisable intra-class evaluations. When this design is applied to the setup *Feature-Selection*, the 30 best segmental features are identified as shown in Table 4.

The results summarised in Table 4 imply that the second order derivative FMFCCs 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. It seems that these features are containing a complex but suitable description of the characteristics of microphone recordings that allow for their intra-class classification.

If only the 20 best features (see Table 4) are used in classification on the test material the classification accuracy is still significant for the application scenario but it drops in average for about 7.11% (see *Classifier-Benchmarking-RS4_Rode-Best20Features-only*) in comparison to the full feature set. The four classifiers which hit the 60 hour timeout boundary defined for *Classifier-Benchmarking-RS4_Rode* and *Classifier-Benchmarking-RS4_Rode* have no problem to keep below that boundary when using only 20 features instead of the full set of 590.
If the worst performing segmental features are considered, it can be summarised that especially the formants and many of the time domain features (e.g. the lsb_ratio) show absolutely no significance for MF.

Feature independency: In addition to the actual classification tests in Buchholz et al.[16] , a principle component analysis (PCA) is conducted there on the used feature vectors, to determine if the feature space used contains correlated features and could therefore be reduced resulting in a sped up classification. The analysis uncovered a strong correlation between the used features and that the classification could be sped up dramatically by feature selection without loosing much of the classification accuracy.
When the same PCA is conducted on the 590 dimensional feature vector generated by AAST / AAFE v.2.0.5 then 187 transformed components are identified as being responsible for 95% of the sample variance (on *RS4_Rode* in *R01*), which also implies strong potential for feature selection.

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 in average for about 7.11% in comparison to the full feature set. At the same time the average computation time is reduced by factor 32.7 (the feature space is reduced by factor 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) and the classifier quality value *optDist* improves in average by 0.094 due to the much stronger decrease of the run-times of the evaluations in comparison to the drop in classification accuracies.
It can be stated that the *optDist* of the classification using the 20 dimensional set is closer to the optimum (due to its faster classification), while the 590 dimensional set achieves higher classification gains (at the cost of dramatically increased costs in computation times).

4.3. Microphone orientation influence evaluation

To show how strong *O* (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 (setup *Orientation_Impact_RS7*). The eight different orientation recordings of the microphone *M22* used (see section 3 above) for the generation of *RS7* 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 two references (silence and a pure sinoid). The test hypothesis for both tests is: "The candidate material is recorded by *M22*." 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 classification *accuracy* of weka.classifiers.meta.RandomSubSpace for all ten reference signals in *RS4_Beyer* (*R06*; 590 dimensional feature vector) is 81.86% (*cg*=0.76). For the silence reference recorded in recording set *RS7* (and tested against the model generated from the corresponding *RS4_Beyer* material) an *accuracy* of 100% (*cg*= 1.0) is achieved. For the sinoid under the same conditions the *accuracy* is also 100% (*cg*=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.

4.4. Microphone mounting influence evaluation

To show how strong the influence *M* of microphone mounting changes is, in comparison to the inter-microphone distance of different microphones of the same brand and model, two experiments are constructed (both summarised in setup *Mounting_Impact_RS9*). The eight recordings in different mountings of the microphone *M22* 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). The test hypothesis for both tests is: "The candidate material is recorded by *M22*." 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 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) an *accuracy* of 100% (*cg*=1.0) is achieved. For the sinoid the *accuracy* is 87.5% (*cg*=0.86). The misclassifications are limited to mounting position 4 (all other positions using some sort of microphone clamp mounted on a tripod, while in 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). For this position all 200 corresponding feature vectors in this test are misclassified as origination from *M23* instead of *M22*.

4.5. Audio file composition detection

Here the results four the four different investigations on composition detection introduced in section 1 above are presented.

4.5.1. Composition Test 1

This test evaluates the case where microphone recordings of one known microphone made in different locations composed into one stream (experimental setup: *Composition-1*). In terms of the classification problem at hand this test is assumed to be the hardest problem in this MF approach. As shown by Kraetzer et al.[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 2 shows the results for this experiment (*Composition-1*) and the four exemplarily selected classifiers.

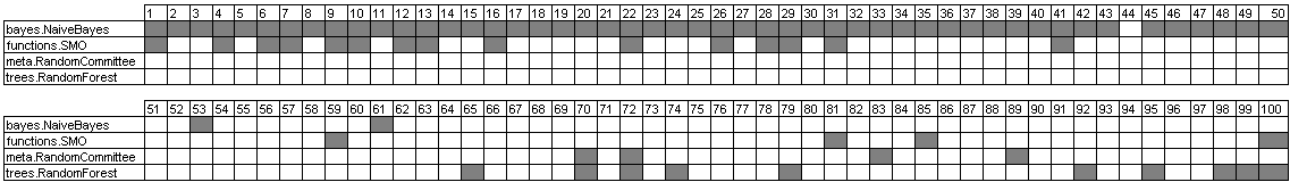

Figure 2: Mash-up1 test results (upper half original, lower half pasted in material)

The colour-coding in figure 2 has the following meaning: A white 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 *M22*. A dark grey 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 *M20*, *M21* or *M23*.
A white 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 *M22* (as it is registered in the model for the room / recording environment R01). A dark grey 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 *M22*.
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 white and the false classifications (FN and FP) in dark grey.

Table 5: Error rates for the exemplarily classifications on *Composition-1*

Classifier	*accuracy*	TP	FN	TN	FP
bayes.NaiveBayes	49%	2%	98%	96%	4%
functions.SMO	81%	70%	30%	92%	8%
meta.RandomCommittee	96%	100%	0%	92%	8%
trees.RandomForest	91%	100%	0%	82%	18%

As can be seen in figure 2 and in Table 5 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 accuracy achieved in the overall evaluations (see Table 5) could be used very well to rank the classifiers according to their performance.

4.5.2. Composition Test 2

This test evaluates the case where one known microphone pasted into a stream of completely different known microphone (experimental setup: *Composition-2*). 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 performance of the approach under this assumption.
Figure 3 shows the results for the experiment *Composition-2* and the four exemplarily selected classifiers. The same colour coding scheme is applied as in figure 2 above. Therefore all true classifications (TP and TN) are marked in white and the false classifications (FN and FP) in dark grey.

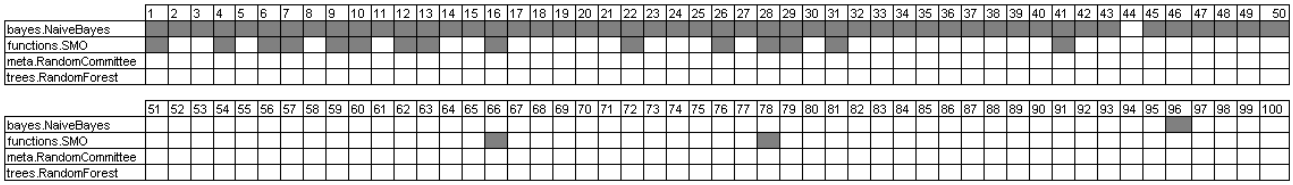

Figure 3: Mash-up2 test results (upper half original, lower half pasted in material)

As can be seen in figure 3 and in Table 6, the used classifiers on the original part show exactly the same performance as in figure 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.

Table 6: Error rates for the exemplarily classifications on *Composition-2*

Classifier	*accuracy*	TP	FN	TN	FP
bayes.NaiveBayes	50%	2%	98%	98%	2%
functions.SMO	83%	70%	30%	96%	4%
meta.RandomCommittee	100%	100%	0%	100%	0%
trees.RandomForest	100%	100%	0%	100%	0%

Differences can be seen in the performance on the impostor part (the lower half in figure 3; material from *M23* claimed to originate from *M22*). 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 Naïve Bayes classifier achieves an accuracy of 50% in this evaluation, which due to the fact that here a mere two-class evaluation is performed, would be equal to the probability of guessing correctly at the class.

4.5.3. Composition Test 3

This test evaluates the case where one unknown microphone pasted into a stream of completely different known microphone (experimental setup: *Composition-3*).
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.

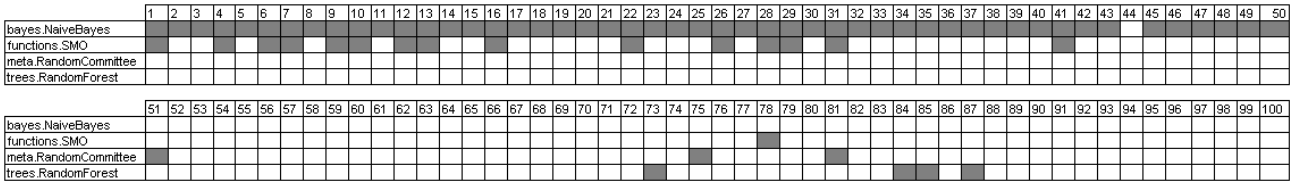

Figure 4: Mash-up3 test results (upper half original, lower half pasted in material)

Figure 4 shows the results for the experiment *Composition-3* and the four exemplarily selected classifiers. The same colour coding scheme is applied as in figure 2 above. Therefore all true classifications (TP and TN) are marked in white and the false classifications (FN and FP) in dark grey.

Table 7: Error rates for the exemplarily classifications on *Composition-3*

Classifier	***accuracy***	**TP**	**FN**	**TN**	**FP**
bayes.NaiveBayes	51%	2%	98%	100%	0%
functions.SMO	84%	70%	30%	98%	2%
meta.RandomCommittee	97%	100%	0%	94%	6%
trees.RandomForest	96%	100%	0%	92%	8%

The performance in the original part (see figure 4 upper half and Table 7) is exactly the same as for the previous tests. For the impostor part (the lower half in figure 4; material from *M8* claimed to originate from *M22*) a good to very good performance is achieved by all four classifiers. Nevertheless the Bayesian classifier achieves only an accuracy of 51% which would disqualify this classifier from practical application.

4.5.4. Composition Test 4

This test evaluates the case where one unknown microphone pasted into a stream of completely different unknown microphone (experimental setup: *Composition-4*). 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 5 shows the results for the experiment *Composition-4* and the four exemplarily 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 (*M20* = light hatching, *M21*=white, *M22*=dark hatching, and *M23*=dark grey) classification result for each of the 100 frames in the test material is marked in this colour coding.

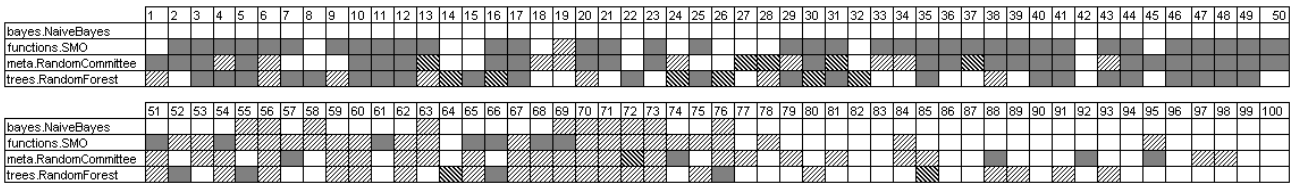

Figure 5: Mesh-up4 classification results (upper half *RS2 M2*, lower half *RS2 M3*)

In figure 5 the upper half is representing the first unknown or impostor microphone (*M2* from *RS2* room *R01*) and the second half is the other one (*M3* 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 MF approach. The higher the change rate and the shorter the average sequence length in the classifications, the better the classification under this circumstances.
For the upper half the Naïve Bayes classifier shows here a very bad performance. All frames are insistently classified as belonging to *M21*. 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 including the average classification accuracy of the used classifier into account, which for the Naïve Bayes is about 40% in the tests performed in section 4.1.

Table 8: Change rate and average sequence length for the experiments in *Composition-4*

		change rate	avg. sequence length
first half (M2)	bayes.NaiveBayes	0	50
	functions.SMO	20	2.38
	meta.RandomCommittee	26	1.85
	trees.RandomForest	36	1.35
second half (M3)	bayes.NaiveBayes	10	4.55
	functions.SMO	17	2.78
	meta.RandomCommittee	34	1.43
	trees.RandomForest	28	1.72

The RandomCommittee and RandomForest classifiers show a much better performance in these two tests presented here. They show in Table 8 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 Naïve Bayes classifiers, which implies that these (known good classifiers; see section 4.1) are run on impostor material.

5. SUMMARY AND CONCLUSIONS FOR FURTHER WORK

Summarising the results for **classifier selection** it can be said that for a test set generated by using four identical microphones in parallel recording and all 74 of WEKAs classifiers a highest classification accuracy of 82.5% is achieved in the tests. Four of the 74 non-tuned classifiers give results between 80 and 82.5% and 27 further classifiers report accuracies between 60 and 80%. Of the remaining 43 classifiers 16 perform between 27% and 60%, 8 are just "guessing" (at 25% in this four-class classification problem) and 19 returned errors (insufficient memory size at 1.5GByte, timeouts at 60h, missing cost files or wrong data format). The results for this test set are substantiated by similar findings on another set of four identical microphones. Considering the different classes of classifiers used, it can be summarized that the used meta-classifiers give the best results. Within the top 20 of the ranked classifiers only few tree-based classifiers or functions can be found. The clustering algorithms evaluated within this paper did show no significant results for MF.
Regarding the **feature selection** our results showed a very good performance of the 2nd order derivative MFCC based features (introduced by Liu et al.[10]), as well as a good performance for selected time domain, frequency domain and FMFCC based features. Furthermore it is indicated by an estimation of the true dimensionality of the feature space using a PCA that about 1/3 of the features are responsible for 95% of the of the sample variance. Based on this a strong feature selection down to 20 most significant features was performed. The classification result using this reduced set does show only small impact to the classification accuracy, but a strong influence to required runtime of the classifiers.
Based on the investigations presented in section 4.3 the **orientation** *O* 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.
The **mounting** (*M*) of a microphone 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 desk top. 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.

In the tests 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 exemplarily chosen classifiers for the evaluations here the SMO, which is performing quite well in all other MF evaluations (see e.g. section 4.1) shows a dissatisfactory performance. The 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 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.

ACKNOWLEDGEMENTS

The effort for this publication was in part sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-09-M3061. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.
The work in this paper on the application of source authentication in long-term archiving contexts has been supported in part by the European Commission through the FP7 ICT Programme under Contract FP7-ICT-216736 SHAMAN. 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.

REFERENCES

[1] C. Kraetzer, A. Oermann, J. Dittmann, A. Lang: *Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification*. Proc. ACM Multimedia and Security Workshop 2007, ACM, 2007.
[2] H. Malik, H. Farid: *Audio Forensics from Acoustic Reverberation*. Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASP), Dallas, TX, 2010.
[3] D. Davis, C. Davis: *Sound System Engineering*, 2nd ed., Focal Press, 1997
[4] T. Takala, J. Hahn: *Sound Rendering*, Proc. 19th annual conf. on Computer graphics and interactive techniques, 1992.
[5] H. Farid: *A survey of image forgery detection*, IEEE Signal Processing Magazine, vol. 2, no. 26, 2009.
[6] R. Yang, Z. Qu, J. Huang: *Detecting Digital Audio Forgeries by Checking Frame Offsets*. Proc. ACM Multimedia and Security Workshop 2008, ACM, 2008.
[7] C. Grigoras: *Application of ENF Criterion in Forensic Audio, Video, Computer and Telecommunication Analysis*. Forensic Science international, no. 167, 2007.
[8] C. Grigoras, A. Cooper, M. Michalek: *Forensic Speech and Audio Analysis Working Group - Best Practice Guidelines for ENF Analysis in Forensic Authentication of Digital Evidence*. European Network of Forensic Science Institutes, 2009.
[9] R. C. Maher: *Modeling and signal processing of acoustic gunshot recordings*. Proc. IEEE Signal Processing Society 12th DSP Workshop, 2006.
[10] Q. Liu, A. H. Sung, M. Qiao: *Novel stream mining for audio steganalysis*. Proc. ACM international conference on Multimedia. ACM, 2009.
[11] C. Kraetzer, J. Dittmann: *Improvement of information fusion-based audio steganalysis*. Proc. Multimedia on Mobile Devices 2010, Electronic Imaging Conference 7542, IS&T/SPIE, 2010.
[12] I. H. Witten, E. Frank: *Data Mining: Practical machine learning tools and techniques*. 2nd Edition. Morgan Kaufmann, 2005.
[13] R. Böhme, A. Westfeld: *Statistical Characterisation of MP3 Encoders for Steganalysis*. Proc. ACM Multimedia and Security Workshop 2004, ACM, 2004.
[14] A. Oermann, A. Lang, J. Dittmann: *Verifier-tuple for audio-forensic to determine speaker environment*. Proc. ACM Multimedia and Security Workshop 2005, ACM, 2005.
[15] C. Kraetzer, M. Schott, J. Dittmann: *Unweighted Fusion in Microphone Forensics using a Decision Tree and Linear Logistic Regression Models*. Proc. ACM Multimedia and Security Workshop, 2009.
[16] Buchholz, Robert, Christian Kraetzer and Jana Dittmann: *Microphone Classification Using Fourier Coefficients*. In Proceedings of 11th Information Hiding Darmstadt, LNCS 5806, Springer-Verlag Berlin Heidelberg, 2009.

Table 9: Test setup descriptions

Setup	Training material	Test material	Classifiers / clusterers	Feature set
Classifier-Benchmarking-RS4_Rode	•*RS4_Rode* (10 reference files) •200 feature vectors per file for all of the 4 microphones (*M16, M17, M18, M19*) and each of the 10 rooms •10x stratified cross-validation		all 74 in WEKA (v.3.6.1) implemented supervised classifiers in default parameterisations	all 590 segm.
Classifier-Benchmarking-RS4_Beyer	•*RS4_Beyer* (10 reference files) •200 feature vectors per file for all of the 4 microphones (*M20*, *M21*, *M22*, *M23*) and each of the 10 rooms •10x stratified cross-validation			all 590 segm.
Feature-Selection	•*RS4_Beyer* and *RS4_Beyer* (10 reference files) •200 feature vectors per file for all of the 2x4 microphones ((*M16, M17, M18, M19)* and (*M20*, *M21*, *M22*, *M23*)) and each of the 10 rooms		none	Feature select.
Classifier-Benchmarking-RS4_Rode-Best20Features-only	•*RS4_Rode* (10 reference files) •200 feature vectors per file for all of the 4 microphones (*M16, M17, M18, M19)* in *R01* 10x stratified cross-validation		all 74 in WEKA implemented supervised classifiers in default parameterisations	best 20
Orientation_Impact_RS7	•*RS4_Beyer* (2 reference files: silence and sine) •200 feature vectors per file for all of the 4 microphones (*M20*, *M21*, *M22*, *M23*) in *R06*	• *RS7* (2 reference files: silence and sine) • 8 mic. orientations 200 feature vectors per file and orientation for *M22* in *R06*	weka.classifiers: meta.RandomSubSpace (in default parameterisation)	all 590 segm.
Mounting_Impact_RS9		• *RS9* (2 reference files: silence and sine) • Eight mounting positions 200 feature vectors per file and mounting for *M22* in *R06*		
Composition-1	• *RS4_Beyer* (10 reference files) • First 200 feature vectors per file for all of the 4 microphones (*M20, M21, M22, M23*) in *R01*	• "original half": 50 feature vectors (disjunctive with training material) from *M22* in *R01* • "impostor half": 50 feature vectors from *M22* in *R06*	weka.classifiers: • bayes.NaiveBayes • functions.SMO • meta.RandomCommittee • trees.RandomForest (all in default parameterisation)	all 590 segm.
Composition-2		•"original half": 50 feature vectors (disjunctive with training material) from *M22* in *R01* •"impostor half": 50 feature vectors from *M23* in *R01*		
Composition-3		• "original half": 50 feature vectors (disjunctive with training material) from *M22* in *R01* • "impostor half": 50 feature vectors from *M8* (*RS2* in *R01*)		
Composition-4		• "first half": 50 feature vectors from *M2* (*RS2* in *R01*) • "second half": 50 feature vectors from *M3* (*RS2* in *R01*)		