Now ForReal: Towards Practical Forensic Room Identification with Real-World Audio Data

Denise Moussa, Leon Huber, Germans Hirsch and Christian Riess
Friedrich-Alexander University Erlangen-Nürnberg, Germany
{denise.moussa, christian.riess}@fau.de

***Abstract*—Recovering the place of origin of, e.g., a phone call can aid the reconstruction of events in a criminal case. For audio forensics, identifying the recording location exclusively from an audio signal still poses a challenge. While various works address this task, they evaluate on semi-synthetic reverberant speech data in a supervised setting. Thus, there barely exist any empirical insights on practical forensic recording environment identification, i.e., the handling of real-world audio data from case-dependent locations that are unknown to a tool in advance.**

In this work, we take a first step towards such a practical scenario. We collect a set of real-world speech from several rooms under varying recording parameters. In forensic cases, audio evidence usually stems from uncontrolled sources, such that factors like the recording position, speaker or microphone can be unknown and reverberation characteristics are of mixed quality. The influence of such factors for room identification is analysed in detail, with several results. For example, we find that prior knowledge about the recording position strongly aids classification, and that characteristics of a speaker's voice notably impact performance. Instructions on how to obtain the data set are online: https://faui1-gitlab.cs.fau.de/mmsec/forreal

***Index Terms*—Room Identification, Audio Forensics**

I. Introduction

Nowadays, speech data from all kind of sources, like phone calls or voice messages, can be subject to criminal investigations as evidence. One audio forensic task is the identification of the recording location from the audio signal. This can uncover valuable cues about events in a criminal case, *e.g.*, in which room of a house some recording was made.

To this end, several works propose tools that assign some given speech recording to one location from a candidate set [1]–[4]. However, existing tools have only been evaluated on semi-synthetic reverberant speech data, so far (*cf*. Sec. II). Hereby, anechoic speech is convolved with separately measured acoustic impulse responses (AIRs) from different places to produce reverberant signals. As a result, it is up to now unclear, how well real-world signals of different properties can be automatically assigned to their respective recording location. This is nevertheless important to be able to provide reliable and robust forensic tools for practical use-cases.

In this work, we thus want to give first insights into this matter. To this end, we record the, to the best of our knowledge, first data set of reverberant speech under varying recording parameters from five different rooms. We adopt few-shot classification to be able to classify query samples from rooms not seen in training with the help of few collected reference audio recordings from candidate locations. Training is conducted on artefact-free semi-synthetic data. Our approach is motivated by practice, since a supervised setting would result in tedious data collection and retraining for each set of candidate locations per forensic case.

Our study addresses practical forensic questions. First, we investigate general aspects, *i.e.*, the number of reference audio signals needed to stably identify the recording location, and the influence of the samples' reverberation characteristics. Second, we analyse in detail the specific influence of recording parameters. After all, the recording situation of given audio material in forensic cases is often unknown. So, the question arises as to how reference samples should be collected from candidate locations. We here focus on the factors speaker, microphone and recording position, and, among others, investigate the impact of individual microphone and speaker types, as well as differing parameter settings in query and reference samples.

We hope that our findings on forensic real-world location identification provides valuable leads for the development of practically applicable forensic tools and raises the awareness for the need of realistic evaluation datasets.

Section II summarises existing work for recording environment identification. Section III describes our collected dataset, Sec. IV presents the experiments on our real-world dataset, and Sec. V concludes our work.

II. Related Work

The large majority of existing works adopt supervised learning to train and test a classifier on a known closed set of recording environments [1]–[3], [5]–[8]. Some approaches rely on either analytical acoustic features [1], [2], [5], [6] or deep features [7] that are input to simple classifiers like a Support Vector Machine (SVM) [2], [6], [7], Gaussian-based models [1], [5] or a classification tree [6]. Also, end-to-end trainable deep learning (DL) methods are increasingly explored for the task [3], [8]. Here, both a convolutional recurrent neural network (CRNN) classifier and convolutional neural network (CNN) architectures operating on frequency representations of speech samples were proposed [3], [8].

Nevertheless, in criminal investigations, the set of potential recording locations is case-dependent and thus dynamic instead of static. Therefore, a traditional closed-set classifier has to be retrained upon every change to the set of candidate environments which results in a high effort for data collection. To avoid this limitation, metric learning strategies like few-shot classification can be utilised to handle locations not previously seen during training [4]. Hereby, given few recording samples for each candidate location, the system is able to assign some given query sample to one of those candidates (*cf*. Sec. IV-A).

Due to the lack of reverberant speech corpora from annotated environments, related works simulate reverberant samples according to the acoustic sound model

$$x(t) = r(t) * s(t) + n(t) \ , \tag{1}$$

where $*$ denotes the convolution operator [1]–[6]. A speech signal $s(t)$ is assumed to travel through a reverberant environment which is described by its characteristic AIR $r(t)$. Background noise from the environment is modelled by the additive signal $n(t)$, where many works assume noiseless reverberant speech only and thus set $n(t) = 0$ [1]–[3], [6]. Otherwise, synthetic or real-world noise is used [4], [5].

For speech signal $s(t)$, related work uses openly available resources, *e.g*., the TIMIT [9] or the LibriSpeech [10] corpus [2], [3], [6]. Some works only include artefact-free anechoic data from the ACE [11] or TSP [12] set [4], [5]. As for AIR measurement $r(t)$, the ACE [11] database of seven rooms with sizes in $[47.3\,\mathrm{m}^3, 371.5]\,\mathrm{m}^3$ is widely used [2]–[6]. Other works include the Aachen Impulse Response database (AAIR) [13] of five rooms with volume range $[11.9, 370.8]\,\mathrm{m}^3$ [4], the Queen Mary Univ. of London set (QMUL) [14] of three rooms up to $9500\,\mathrm{m}^3$ [1], [6], the OpenAIR database [15] of large and open-air spaces [1], [4], and the MIT [16] set with 271 AIRs from diverse places [4]. Also, the source image method [17] is used to simulate AIRs [4], [18]. To our knowledge, only one work tests supervised room classification on in-house real-world samples of the word ‘Alexa’ from one speaker in several rooms [8].

III. Collecting Reverberant Real-World Audio

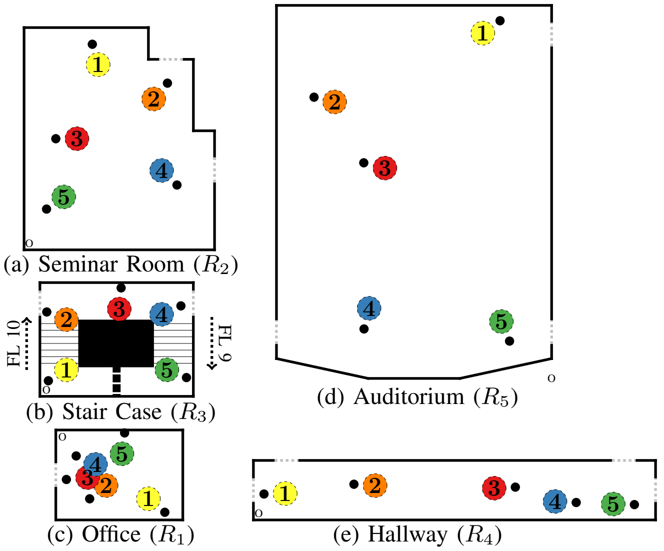
Our collected set consists of in total 6.2 hours of recordings, in detail 600 German single speech samples of 6 speakers recorded with 4 microphones at 5 positions in 5 rooms.

The set enables for investigating real-world room identification under different recording setups. We avoid overly noisy environments, however, distinct background noise and sound events are tolerated. After all, in forensic cases, a given audio evidence usually stems from uncontrolled sources. This includes, by example, exterior noise from cars or rain, room and interior noise from the building like falling doors, footsteps or indistinct babble noise.

The few-shot method is exclusively trained on artefact-free semi-synthetic data and then evaluated on our real-world set. Contrary to supervised training, this prevents short-cut learning [19] which could by example be caused by background noise like church bells that extends over consecutive recordings. While an influence on query and reference embeddings cannot be ruled out, we did not observe biased decisions even for rooms with characteristic clock ticking (Sec. IV-D).

A. Dataset Parameters and Acquisition

In the following, the properties of our set are described. **Selected Rooms** Figure 1 shows the layout and Tab. I the specifications of the 5 used enclosed spaces at the campus of our university. They provide different characteristics:



(a) Seminar Room (R_2)
(b) Stair Case (R_3)
(c) Office (R_1)
(d) Auditorium (R_5)
(e) Hallway (R_4)

Fig. 1: Layout of rooms with dimensions true-to-scale.

- *Office:* a rectangular room with synthetic hard-flooring, concrete walls, one front of windows, three desks, four chairs, two shelves and one clock.
- *Seminar Room:* an 8-sided room with equal floor and wall materials as the office, 2 window fronts, several rows of desks and chairs and two clocks.
- *Stair Case:* a windowless, rectangular stair case encompassing 12 floors, with stone flooring and concrete walls.
- *Hallway:* a rectangular hallway with synthetic hard-flooring, one long-sided brick and concrete wall, respectively, four glass doors and multiple glass cabinets.
- *Auditorium:* a windowless 6-sided room with synthetic hard-flooring, concrete walls, several seat rows and computer equipment that hummed at times during recordings.

Speakers and Text The set includes 3 female and male German speakers, denoted as $F[1,3]$ and $M[1,3]$. Per recording, one speaker reads aloud the German version of the short text ‘the north wind and the sun’ as transcribed in the handbook of the International Phonetic Association and often used in linguistics to illustrate the phonetic structure of languages [20]. The speakers read aloud the text in their natural speaking rate. Hereby, slight variations like slips of the tongue or hesitation markers are tolerated as natural elements of speech.

Recording Equipment We select 4 multi-purpose microphones $C_{[1,4]}$ with different polar patterns capturing sound from different directions (*cf*. Tab. II). They are fixated in row on a tripod at a height of 130 cm and are each cable-connected to a ‘Zoom F4 MultiTrack Field Recorder’ to record each speaker in parallel on 4 separate channels. The sensitivity of each device is set in pretests such that volume peaks of speech are at circa $-3\,\mathrm{db}$. All recordings are stored as signed 24-bit PCM encoded WAV files with a sample rate of 96 kHz.

Recording Procedure Per room, we mark 5 recording

TABLE I: Room specifications and (x, y)-coordinates of recording positions $P_{[1,5]}$, measured from the origin denoted by O in Fig. 1. For non rectangular rooms, the maximum dimensions are given ($*$). The height of the auditorium is an approximate average since it varies along the room (†) and is used to compute an approximate volume (‡).

Room	Label	Width [m]	Length [m]	Height [m]	Volume [m^3]	P_1 [m]	P_2 [m]	P_3 [m]	P_4 [m]	P_5 [m]
Office	R_1	3.4	4.8	3.2	52.2	$(2.6, 3.5)$	$(2.1, 1.9)$	$(1.8, 1.2)$	$(1.3, 1.5)$	$(0.9, 2.5)$
Seminar Room	R_2	7.2*	8.4*	2.9	158.4	$(2.8, 7.0)$	$(4.9, 5.7)$	$(2.0, 4.2)$	$(5.2, 3.0)$	$(1.2, 2.0)$
Stair Case	R_3	4.3	5.8	-	-	$(1.0, 1.0)$	$(2.9, 1.0)$	$(3.3, 3.0)$	$(3.1, 4.6)$	$(1.0, 4.8)$
Hallway	R_4	2.3	15.2	3.1	108.4	$(1.0, 1.2)$	$(1.3, 4.6)$	$(1.2, 9.1)$	$(0.7, 11.4)$	$(0.6, 13.6)$
Auditorium	R_5	10.4*	13.7*	4.13†	588.4‡	$(2.6, 12.3)$	$(8.2, 9.7)$	$(6.3, 7.2)$	$(6.9, 1.9)$	$(1.9, 1.4)$

TABLE II: Microphone Specifications

Label	Brand	Model	Polar Pattern
C_1	AKG	C1000S	♡ Cardioid
C_2	Sennheiser	MD421	♡ Cardioid
C_3	Superlux	E304	⌓ Semi-Omnidirectional
C_4	AKG	C4000B	○ Omnidirectional

positions $P_{[1,5]}$ as defined in Tab. I and shown in Fig. 1. The positions are arbitrarily chosen under the constraints imposed by furniture. For each P_n, we set up the microphone tripod and, one after another, record each speaker. The speakers are positioned 80 cm away from the tripod (*cf.* black circles in Fig. 1) and read aloud facing the microphones.

Post-Processing We cut silent signal portions at the beginning and end of each recording using Audacity[1]. This yields 600 speech samples of on average 36.44 seconds. In total, our set consists of 6 hours, 8 minutes and 57 seconds of speech.

IV. Experiments

We investigate several practical scenarios for forensic few-shot recording location identification in a few-shot setting.

A. Few-Shot Room Classification

For our forensic few-shot classification scenario, we adopt a method based on Prototypical Networks [21] which we previously investigated [4]. Here, a semantic meaningful metric embedding space is trained to cluster samples belonging to the same class. This way, the classifier can generalise to case-dependent locations not seen in training given only few reference recordings from each. In detail, the reference samples are projected to the embedding space and then averaged to form one prototype for each candidate location. Some given input query sample can then be attributed to one candidate by selecting the closest prototype in the metric space. For more details on Prototypical Networks we refer to Snell *et al.* [21].

We train the few-shot classifier exclusively on semi-synthetic artefact-free data simulated as described in Eq. 1 and test on our real-world set. All samples are 3 s long. In total, we select AIRs from 291 environments for training. This includes 6 rooms from each the ACE [11], AAIR [13] and REVERB [22] dataset, 26 environments from the OpenAIR [15] set and 247 environments from the MIT [16] set. Furthermore, 239 anechoic speech snippets are selected from 4 female and 7 male speakers from the ACE [11] set and 93 snippets are selected from the anechoic TSP [12] set from 2 male and 2 female speakers. All AIRs are convolved with all speech snippets to obtain 96 612 training samples in total (Eq. 1). Per sample, additive Gaussian noise is randomly sampled from a signal-to-noise interval of $[-10, 50]$ db.

B. Experimental Protocol and Practical Assumptions

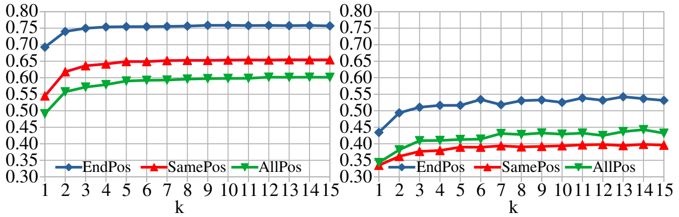
Per experiment, audio samples can qualify as query s_q and/or reference samples $s_r \in \mathcal{S}_R$. However, we enforce $s_q \notin \mathcal{S}_R$ per few-shot classification step. For each s_q, we perform 20 classification steps. Per step, k reference samples are randomly sampled from all 5 rooms to yield 5 prototypes, and s_q is then assigned to the closest prototype. In total, 5 experiment runs are conducted with different seeds and we report averaged scores with standard deviations over all runs.

In our study, we investigate a more and less informed classification scenario, *i.e.*, where sampled references either share or don't share recording setup settings with the query. An audio sample's recording setup is defined by the speaker $S_n \in \{F_n, M_n\}$, the recording position P_n (Tab. I) and the microphone C_n (Tab. II). The first WITHOL , *i.e.*, *with overlap* sampling scenario allows recording parameter values of the query to be present in the references. Thus, reference samples with the same speaker, position or microphone are not banned which simulates the adoption of parameters in references for cases where certain query properties are known. In the more challenging NOOL , *i.e.*, *no overlap* sampling scenario, we assume no knowledge about the query sample's recording setup and therefore only allow reference samples with different values in S_n and C_n and forbid the same P_n for the query room's prototype. Note that the same position identifiers share no common properties across rooms (*cf*. Fig. 1).

C. Reverberation Properties of Speech Snippets

First, we want to gain some insights as to how many references are needed to yield stable prototypes, as well as the influence of speech content within samples. The results over the whole set are shown in Fig. 2a for WITHOL sampling and in Fig. 2b for the more challenging NOOL sampling. We test $k \in [1, 15]$ reference samples per prototype and evaluate 3 different speech snippet variants. This includes 3 seconds of speech from, (1), different time positions (AllPos, green), *i.e.* different speech content in s_q and $\mathcal{S}_R$, (2), some randomly

[1] https://www.audacityteam.org/



(a) WITHOL sampling (b) NOOL sampling

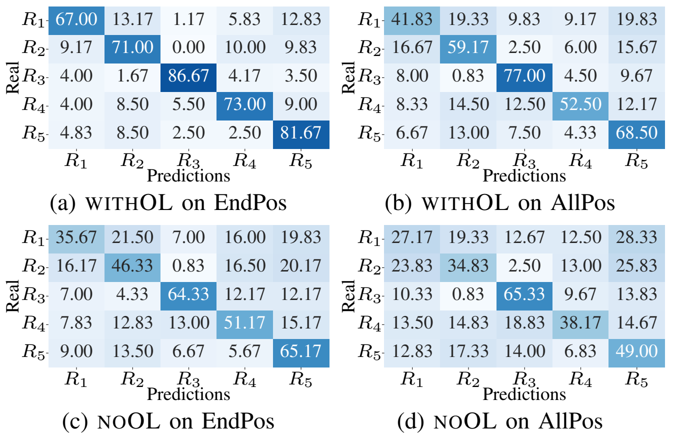
Fig. 2: Accuracy for $k \in [1, 15]$ reference samples.

chosen identical time snippet (SamePos, red) for both, and (3), the end of recording (EndPos, blue) that always contains reverberation not superimposed by successive speech. Note that identical time snippets in different recordings contain similar but not necessarily identical speech content. We do not align the recordings as to not distort reverberation information.

Several observations can be made from this experiment. First, NOOL sampling significantly hardens the task compared to the more relaxed WITHOL sampling strategy. Second, the accuracy increases significantly between $k = 1$ and $k = 3$ and only minor increases can be observed for $k > 5$ for all runs. We set $k = 10$ for our following experiments, as no consistent gains in performance can be observed afterwards. Third, the performance depends on the speech signal characteristic of the samples. Identical versus arbitrary time snippets (red vs. green) in query and references perform similarly. In detail, SamePos-snippets perform slightly better for WITHOL (Fig. 2a), but not for NOOL sampling (Fig. 2b). Since each speaker is recorded in parallel with 4 microphones per position, WITHOL sampling may cause the same sound emissions captured by different devices to be in s_q and $\mathcal{S}_R$. The resulting similar reverberation presumably acts as subtle side-channel for classification, however, such a high degree of similarity can hardly be adopted in practice. More details on the influence of the microphone are discussed in Sec. IV-E. Contrarily, the EndPos-snippets work best for both WITHOL and NOOL sampling with an accuracy of 75.87% and 52.53% for $k = 10$. We thus hypothesize that this is indeed due to the non-superimposed reverberation signal at the end. In our following experiments, we include both the well reverberant EndPos-snippets and more challenging AllPos-snippets with varying reverberation qualities.

D. Classification Performance per Room

Table III shows the mean F1-score, precision and recall with std. devs. per room. In line with previous results, the scores for each room are better for WITHOL prototype sampling than for NOOL sampling and better for EndPos than for AllPos-snippets. Samples from the staircase (R_3) are identified with the highest scores. This location is indeed relatively strongly reverberant which is audible from the recordings. For WITHOL sampling on EndPos, the F1-score is 88.51% and the score is still 73.57% for AllPos. For the especially challenging NOOL sampling the F1-score for EndPos and AllPos decreases to 67.07% and 61.26%. The hallway (R_4) and auditorium (R_5) perform similarly and second best. At maximum, they are identified with F1-scores around 75% for WITHOL sampling on EndPos, while the lowest scores for AllPos and NOOL sampling drop to around 42%. The office (R_1) and the seminar room (R_2) are most difficult to identify. The F1-score is around 70% for WITHOL sampling on EndPos, but decreases strongly to 37.17% for the seminar room and to 28.92% for the office for AllPos and NOOL sampling.



(a) WITHOL on EndPos (b) WITHOL on AllPos

(c) NOOL on EndPos (d) NOOL on AllPos

Fig. 3: Normalised confusion matrices for $k = 10$ in [%].

In addition, Fig. 3 shows the confusion matrices for all rooms. The worst performing office (R_1) is most likely confused with the seminar room (R_2) and the auditorium (R_5), especially for NOOL sampling (*cf*. Fig. 3c–3d.) Similarly, the office and auditorium are the most likely confusion candidates for the seminar room. Note that the office and seminar room are from the same building, share the same floor and wall materials and contain differently many clocks (*cf*. Sec. III-A). However, no such obvious properties are shared with the auditorium which is nevertheless a likely confusion candidate for both. On the contrary, the seminar room and stair case are most seldomly confused with each other (*cf*. Fig. 3b–3c).

E. Influence of Individual Recording Parameters

In a second experiment, we quantify the influence of identical and different recording parameters, *i.e.*, speaker, recording position and microphone, in query and references. We here differentiate between a *match* and *mismatch* of a specific parameter and otherwise pose no constraints on the remaining parameters. To isolate a parameter p, we also investigate an *exclusive match* w.r.t. p, where all other parameters are enforced to have different values in s_q and $\mathcal{S}_R$.

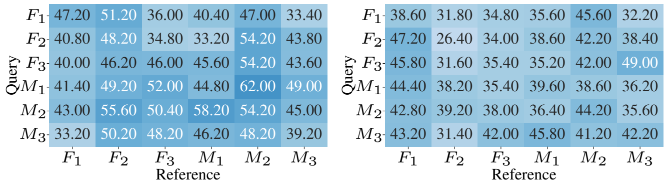
Table IV shows the results per parameter p as one of recording position P_n, speaker S_n and microphone C_n. The table allows to compare the influence of different aspects of sampling, namely both the type of audio snippets, *i.e.*, EndPos or AllPos, and the sampling strategies *match*, *mismatch* and *exclusive match*. Note that a *complete mismatch*, i.e., only different parameters in query and references, could be seen as a fourth sampling case. This case is essentially the same for all p and achieves 42.9% for AllPos and 52.53% for EndPos.

TABLE III: Mean F1-Score, Precision and Recall with std. devs. for the individual rooms [%].

Room	WITHOL prototype sampling						NOOL prototype sampling					
	EndPos-Snippets			AllPos-Snippets			EndPos-Snippets			AllPos-Snippets		
Label	F1	P	R	F1	P	R	F1	P	R	F1	P	R
R_1	70.89±1.52	75.29±1.48	67.00±1.94	46.07±1.83	51.30±1.26	41.83±2.20	40.60±1.28	47.18±1.43	35.67±0.96	28.92±1.54	30.96±2.21	27.17±1.86
R_2	70.00±0.82	69.06±0.71	71.00±1.78	57.21±0.78	55.39±0.81	59.17±0.91	46.66±1.45	47.07±2.39	46.33±1.53	37.17±1.62	39.89±2.95	34.83±2.36
R_3	88.51±0.41	90.44±0.85	86.67±0.00	73.57±0.76	70.43±0.87	77.00±0.67	67.07±0.78	70.05±1.33	64.33±1.07	61.26±1.98	57.71±1.55	65.33±1.38
R_4	74.68±0.49	76.44±0.54	73.00±0.85	59.48±0.74	68.65±0.75	52.50±1.39	50.78±0.84	50.43±1.55	51.17±0.86	42.36±3.09	47.60±2.76	38.17±2.86
R_5	75.33±0.43	69.91±0.89	81.67±0.53	60.67±0.78	54.45±0.93	68.50±0.97	56.06±1.01	49.21±1.78	65.17±0.85	42.31±1.23	37.25±0.82	49.00±0.92

TABLE IV: Mean accuracy and std. devs. in [%] for query-reference matches and mismatches w.r.t. parameter p.

p	Match		Mismatch		Exclusive Match	
	EndPos	AllPos	EndPos	AllPos	EndPos	AllPos
P_n	93.83±0.41	65.33±0.41	61.67±0.53	47.33±0.82	66.00±0.62	57.17±1.35
S_n	85.00±0.63	42.00±0.89	58.20±0.75	53.20±0.98	46.00±0.63	35.40±0.80
C_n	73.20±0.27	57.07±0.53	74.00±0.42	53.87±0.27	57.33±2.56	44.53±1.40



Query \ Reference	F_1	F_2	F_3	M_1	M_2	M_3
F_1	47.20	51.20	36.00	40.40	47.00	33.40
F_2	40.80	48.20	34.80	33.20	54.20	43.80
F_3	40.00	46.20	46.00	45.60	54.20	43.60
M_1	41.40	49.20	52.00	44.80	62.00	49.00
M_2	43.00	55.60	50.40	58.20	54.20	45.00
M_3	33.20	50.20	48.20	46.20	48.20	39.20

(a) EndPos-Snippets

Query \ Reference	F_1	F_2	F_3	M_1	M_2	M_3
F_1	38.60	31.80	34.80	35.60	45.60	32.20
F_2	47.20	26.40	34.00	38.60	42.20	38.40
F_3	45.80	31.60	35.40	35.20	42.00	49.00
M_1	44.40	38.20	35.40	39.60	38.60	36.20
M_2	42.80	39.20	38.00	36.40	44.20	35.60
M_3	43.20	31.40	42.00	45.80	41.20	42.20

(b) AllPos-Snippets

Fig. 4: Accuracy in [%] for speaker pairs and NOOL sampling.

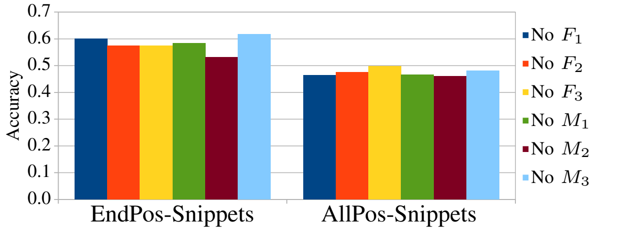


Fig. 5: Impact of individual speakers being removed from the references under NOOL sampling.[2]

Recording Position The same position P_n in query and references strongly benefits classification. By example, for EndPos, an *exclusive match* of P_n outperforms both a simple *mismatch*, where other parameters can be identical and a *complete mismatch* of all parameters with an accuracy of $66\% > 61.67\%$ and $66\% > 52.53\%$. The same accounts for AllPos with in total lower scores. For both EndPos and AllPos, a *match* of P_n with allowed overlaps in other parameters is understandably best with 93.83% and 65.33%. Still, note that results for *match* on EndPos are optimistic due to the subtle side-channel from parallel recording (*cf*. Sec. IV-C).

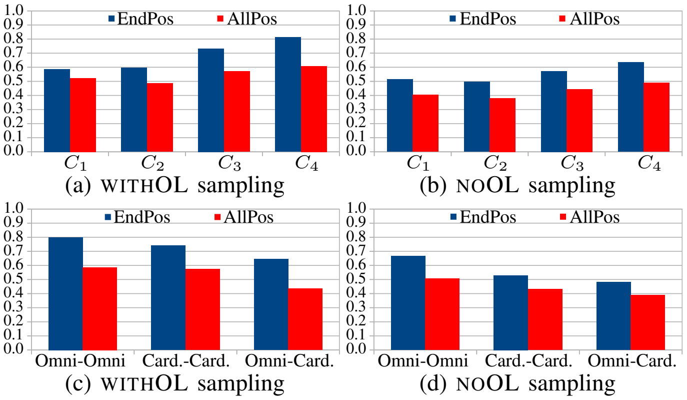
Speaker Sampling snippets from the same speaker S_n in query and references shows no useful practical advantage, but speakers generally perform differently well as references. In detail, apart from the (from a practical perspective too optimistic) results for *match* and EndPos, different speakers in s_q and $\mathcal{S}_R$ are always better, *e.g.*, with $52.53\% > 46\%$ for a *complete mismatch* versus an *exclusive match* (EndPos), and $53.2\% > 42\%$ for *mismatch* versus *match* (AllPos).

A more detailed analysis investigates specific pairs of query and reference speakers. Figure 4a and 4b show the accuracy for all 36 possible combinations for NOOL sampling. Sampling from the same speaker indeed shows no notable benefit. Surprisingly, on average, sampling from speaker M_2 in query or references exhibits the best accuracy scores with 40.83% on AllPos and 52.18% on EndPos-snippets. This also holds for more diverse prototypes of multiple speakers. Figure 5 shows the results for prototypes including all but the query and one additional speaker under NOOL sampling. Clearly, removing speaker M_2 from the references (brown) results in the strongest decrease in performance. Also, multi-speaker prototypes perform better on average than prototypes from one speaker (*cf*. Fig. 5 vs. 4). This indicates that collecting recordings from several speakers might result in information-wise richer prototypes that can facilitate the classification task.

We additionally investigate the impact of collecting references from a speaker of the same gender as in the query, since one could expect that this benefits classification. By example, properties like the frequency range of a voice correlates with gender, and we can indeed report higher frequencies for female than male voices in our set using the PYIN method [23]. However, separating gender groups has no general advantage. Instead, the results always slightly improve the more male speakers are present in query or references, also when excluding the anomalously strong performing speaker M_2. Then, the advantage of only males over only females ranges between around 1% and 5% for AllPos and EndPos and NOOL sampling. Overall, these results motivate further research into the properties of good reference speakers.

Microphone The same or a similar microphone C_n in query and references improves classification results in practical scenarios. A *match* of C_n performs similar to a *mismatch* with $73.2\% \approx 74\%$ for EndPos, but this is attributed to the side-channel effect of parallel capturing, where the microphone itself has no significant impact (*cf*. Sec. IV-C). In the other cases, a *match* of C_n is superior to a *mismatch*, and an *exclusive match* is better than a *complete mismatch*. Nevertheless, a matching microphone seems to be only of secondary importance. A *mismatch* of C_n, which still allows matches for other parameters, notably the position, is thus better than an *exclusive match* of C_n with $53.87\% > 44.53\%$.

[2] Numerical values for plots: https://faui1-files.cs.fau.de/public/mmsec/moussa/2024WIFS/supplemental.pdf



(a) WITHOL sampling

(b) NOOL sampling

(c) WITHOL sampling

(d) NOOL sampling

Fig. 6: Averaged accuracy for sampling within individual microphones (a-b) and w.r.t. polar patterns (c-d).[2]

A more detailed analysis on the individual devices shows that C_4 works best and sampling within C_3 or C_4 is superior to sampling within C_1 or C_2 (*cf*. Fig. 6a–6b). This is probably due to the individual polar patterns of the microphones (*cf*. Tab. II). In detail, C_4 captures sound from all directions (omnidirectional), and C_3 still covers the upper hemisphere (semi-omnidirectional). In contrast, C_1 and C_2 have cardioid polar patterns and thus record sound primarily from in front of the microphone which consequently captures less reverberation information from surroundings.

Furthermore, it showed that using reference microphones of the same polar pattern as in the query is beneficial, even if the devices differ. Figure 6c–6d show the accuracy for testing within the (semi-)omnidirectional and cardioid microphones, as well as testing across both types. The results show that testing (semi-)omnidirectional microphones against each other works best, followed by cardioid microphones, while a mismatch in polar patterns is worst. Interestingly, testing samples from C_1 and C_2 against each other works better than sampling either within C_1 or C_2, however the difference is small.

V. Conclusion

We present first insights into forensic few-shot recording location identification from real-world audio snippets. Our findings show that generalising from simulated data to real-world samples from unseen rooms is possible. The success rates however vary strongly depending on the recording settings and reverberation characteristics of audio material. While, *e.g.*, reference samples from the query recording location strongly aid classification, using the same or a similar microphone provides less benefit, and w.r.t. to speakers, even no such benefit could be observed. We see our study as a first step towards designing robust tools for practical use and hope to motivate more research in this direction.

VI. Acknowledgements

We kindly thank our colleagues from the Federal Criminal Police Office of Germany (BKA), and Alexander Eichmüller of the FAU media studio for the support in this project.

References

[1] N. Peters, H. Lei, and G. Friedland, “Name That Room: Room Identification Using Acoustic Features in a Recording,” in *International Conference on Multimedia*. ACM, 2012, pp. 841–844.

[2] M. Baum, L. Cuccovillo, A. Yaroshchuk, and P. Aichroth, “Environment Classification via Blind Roomprints Estimation,” in *International Workshop on Information Forensics and Security*. IEEE, 2022.

[3] C. Papayiannis, C. Evers, and P. A. Naylor, “End-to-end Classification of Reverberant Rooms Using DNNs,” *Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 3010–3017, 2020.

[4] D. Moussa, G. Hirsch, and C. Riess, “Can We Identify Unknown Audio Recording Environments in Forensic Scenarios?” *arXiv preprint arXiv:2405.02119*, 2024, under review.

[5] A. H. Moore, P. A. Naylor, and M. Brookes, “Room Identification Using Frequency Dependence of Spectral Decay Statistics,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 6902–6906.

[6] C. Papayiannis, C. Evers, and P. A. Naylor, “Discriminative Feature Domains for Reverberant Acoustic Environments,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 756–760.

[7] M. Azimi and U. Roedig, “Room Identification with Personal Voice Assistants,” in *European Symposium on Research in Computer Security*. Springer, 2021, pp. 317–327.

[8] ——, “Wake Word Based Room Identification with Personal Voice Assistants,” in *Workshop on Hot Trends in Embedded Systems Privacy*. ACM, 2022.

[9] J. S. Garofolo, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Linguistic Data Consortium, 1993.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR Corpus Based on Public Domain Audio Books,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 5206–5210.

[11] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “Estimation of Room Acoustic Parameters: The ACE Challenge,” *Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[12] P. Kabal, “TSP Speech Database,” *McGill University*, 2002.

[13] M. Jeub, M. Schafer, and P. Vary, “A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms,” in *International Conference on Digital Signal Processing*. IEEE, 2009.

[14] R. Stewart and M. Sandler, “Database of Omnidirectional and B-format Room Impulse Responses,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 165–168.

[15] D. T. Murphy and S. Shelley, “OpenAIR: An Interactive Auralization Web Resource and Database,” in *Audio Engineering Society Convention*. Audio Engineering Society, 2010.

[16] J. Traer and J. H. McDermott, “Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. 7856–7865, 2016.

[17] J. B. Allen and D. A. Berkley, “Image Method for Efficiently Simulating Small-Room Acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[18] P. Götz, C. Tuna, A. Walther, and E. A. Habets, “Contrastive Representation Learning for Acoustic Parameter Estimation,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023.

[19] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut Learning in Deep Neural Networks,” *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[20] I. P. Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[21] J. Snell, K. Swersky, and R. Zemel, “Prototypical Networks for Few-Shot Learning,” *Neural Information Processing Systems*, vol. 30, 2017.

[22] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, “The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech,” in *Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013.

[23] M. Mauch and S. Dixon, “PYIN: A Fundamental Frequency Estimator using Probabilistic Threshold Distributions,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 659–663.