OPEN-SET MICROPHONE CLASSIFICATION VIA BLIND CHANNEL ANALYSIS

Luca Cuccovillo *Patrick Aichroth*

Fraunhofer Institute for Digital Media Technology
Ehrenbergstr. 31, 98693 Ilmenau, Germany

ABSTRACT

In this paper, we present a new algorithm for open-set microphone classification, which is based on a pre-existing blind channel estimation approach. The proposed method achieves a Rand index above 93% for AAC, MP3 and PCM-encoded recordings from eight different mobile devices.

Index Terms— microphone classification, audio forensics, open-set classification.

1. INTRODUCTION

Microphone classification is becoming an increasingly relevant topic within the audio forensics domain. It can e.g. be used to support the process of authenticating recordings and detect manipulations, to verify claims of ownership, or for automatic annotation of metadata about acquisition devices in A/V archives.

Initial microphone classification approaches used to target a broad range of recording devices [1–4], while recent ones focus more on mobile devices [5–9]. This change of focus is related to changed requirements for investigations and journalism: User-generated recordings made with lower-quality mobile recording devices, then to be distributed and shared using social networks, are becoming increasingly relevant [10].

Overall, performance of State-of-the-Art algorithms for microphone classification has steadily improved over time: While the first proof of concept by Kraetzer et al. [1] reached a maximum overall accuracy of $\approx$ 76%, most recent approaches by Aggarwal [8] and Jahanirad [9] achieved an accuracy of $\approx$ 90% and $\approx$ 99%, respectively. However, all State-of-the-Art approaches are based on a "closed-set assumption": The classifiers can only handle content from *previously known* devices, i.e., they need to be trained on a predefined device set and are not suited to classify content from arbitrary, previously unknown recording devices. This limits applicability in many real-life usage scenarios.

In the following, we present an open-set classification approach, which is based on previous work on microphone classification and discrimination via blind channel estimation: In [7], we proposed a feature vector for SVM-based closed-set microphone classification, and successfully applied it to audio tampering detection. In [11], the feature vector was applied to the more general case of microphone discrimination, by using it to confirm or deny the existence of a device change at a specific location within the recording. This approach, which is suited for an open-set context in principle, was now improved to provide an unsupervised open-set approach for microphone classification. It is based on a space transform that significantly improves accuracy of the method described in [11]. The proposed approach is not bound to a training set, and it incrementally creates new device classes whenever content from previously unknown recording devices is detected.

The following paper is structured as follows: Section 2 describes a baseline and an enhanced algorithm for microphone discrimination. Section 3 presents the proposed open-set algorithm for microphone classification. The testing procedure and the results are outlined in Section 4. Section 5 concludes by summarizing the work and providing an outlook to possible future improvements.

2. ALGORITHM FOR MICROPHONE DISCRIMINATION

2.1. Baseline algorithm

The performance of the microphone classification approach described in Section 3 depends on the accurate discrimination of audio content recorded by different devices. A baseline approach for this was provided in [11], where microphone discrimination was used to enhance an audio tampering detection algorithm based on stable tone analysis. It is described in the following.

Let (x_1, x_2) denote a pair of input recordings, and (h_1, h_2) the frequency responses of the respective microphone devices (X_1, X_2).

The baseline algorithm for microphone discrimination relies on the assumption $X_1 = X_2 \iff h_1 = h_2$ in order to distinguish between the two cases $X_1 = X_2$ and $X_1 \neq X_2$:

1. Compute $(\hat{h}_1, \hat{h}_2)$ from (x_1, x_2) as described in [7, 11, 12]. $\hat{h}_1$ and $\hat{h}_2$ represent the estimates of the real frequency responses h_1 and h_2, respectively.

This work has been conducted within *AudioTrust+*, a research project partially funded by the Thuringian Ministry for Economic Affairs, Science and Digital Society (TMWWDG)

2. Derive from each channel estimate $\hat{h}_j$ a feature vector f_j, equivalent to the feature vector f defined in [7, 11]. f_j is an high-level representation of a microphone, and embeds both information about the estimate $\hat{h}_j$, and about the power spectrum of the input recording x_j.

3. Evaluate the estimated device similarity by computing the Parson's correlation coefficient $\rho(f_1, f_2)$ between the two vectors f_1 and f_2:

$$\rho(f_1, f_2) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{f_1(i) - \bar{f}_1}{s_{f_1}} \right) \left(\frac{f_2(i) - \bar{f}_2}{s_{f_2}} \right)$$

where s_{f_1} and s_{f_2} denote the standard deviations of f_1 and f_2, both of length n. $\bar{f}_1$ and $\bar{f}_2$ represent their mean values.

4. If $\rho(f_1, f_2) \geq \tau$, with τ being a user defined threshold, then x_1 and x_2 are considered as being recorded by the same microphone.

The described baseline algorithm is straightforward, and provides the advantage that each file can be processed independently and in parallel, up to the computation of $\rho(f_1, f_2)$.

If two different devices were used to record x_1 and x_2, however, a single comparison is not ideal: The analysis of segments containing content recorded by the two devices can provide useful information, and hence lead to better results, even if the frequency estimates of both files differ. This insight led us to the development of an enhanced discrimination approach described in the next Section 2.2.

2.2. Enhanced algorithm

The enhanced algorithm for microphone discrimination is also based on the blind channel estimation procedure proposed by Gaubitch [12, 13]. However, it is not applied to the two input files x_1 and x_2 independently, but to the file $x := x_1 \circ x_2$ that is obtained by splicing the two files. The base assumption is that the outcome of the channel estimation in [12, 13], if applied locally to separate overlapping intervals of the spliced file, will remain almost constant in the case of both files being recorded by the same device, and it will show significant deviations otherwise. The procedure can be formalized as follows:

1. Splice x_1 with x_2, to obtain a single input file x.

2. Divide x in several overlapping frames, using a fixed window length L and hop size D.

3. Compute one channel estimate $\hat{h}_j$ per each frame, obtaining the set $\hat{\mathcal{H}} = \{\hat{h}_1, \ldots, \hat{h}_M\}$. The order of the channel estimates $\hat{h}_j \in \hat{\mathcal{H}}$ is time-aware, i.e., $\hat{h}_{j-\delta_i}$ starts $\delta_i \cdot D$ seconds before $\hat{h}_j$.

4. Compute several vectors v_{δ_i}, storing the Parson's correlation coefficient $\rho(\cdot, \cdot)$ between every existing pairs of estimates with displacement equal to δ_i:

$$v_{\delta_i}(j) = \rho(\hat{h}_j, \hat{h}_{j+\delta_i}) \text{ , with } v_{\delta_i} \in \mathbb{R}^{M-\delta_i},$$

where $\delta_i = \delta_1, \ldots, \delta_\Delta$ is the displacement between the channel estimates, M is the number of channel estimates, and $1 \leq \delta_1 \leq \delta_\Delta \leq M-1$, with $\delta_i \in \mathbb{N}$. The length of each v_{δ_i} decreases as δ_i increases, since the number of available pairs is lower.

5. All the information in the set $\mathcal{V} = \{v_{\delta_1}, \ldots, v_{\delta_\Delta}\}$ can be reshaped by means of a space transformation

$$\mathbb{T} : \mathcal{V} \mapsto \mathbb{R}^{(M-\delta_1) \times (\frac{\Delta}{2} \cdot (\Delta - 1))},$$

where $\Delta = (\delta_\Delta - \delta_1 + 1)$. As also shown in Figure 1, $\mathbb{T}$ is a function that concatenates $Toeplitz$-matrices ($V \in \mathbb{R}^{(M-\delta_1) \times (\delta_i - \delta_1 + 1)}$) of vectors v_{δ_i} for all δ_i:

$$\mathbb{T} := V(v_{\delta_1}) \circ V(v_{\delta_2}) \circ \cdots \circ V(v_{\delta_\Delta}).$$

The $Toeplitz$-matrices are built as follows:

$$V_{i,j} = V_{i+1,j+1} := \begin{cases} v_{\delta_\cdot}(j-i), & \text{if } \exists v_{\delta_\cdot}(j-i) \; \forall (i,j) \\ 1, & \text{if } \nexists v_{\delta_\cdot}(j-i) \; \forall (i,j) \end{cases}$$

6. Compute the vector $y \in \mathbb{R}^{M-\delta_1}$, characterizing the inter-similarity between adjacent time intervals of the spliced input file:

$$y(i) = \frac{1}{\frac{\Delta}{2}(\Delta - 1)} \sum_{j=1}^{\frac{\Delta}{2}(\Delta-1)} Y(i,j) \text{ , with } Y = \mathbb{T}(\mathcal{V}).$$

Every element $y(i)$ describes the similarity between the estimated frequency response of two non-overlapping intervals with L seconds duration, which end and start, respectively, at the time instant $t = D \cdot (\delta_1 + i)$.

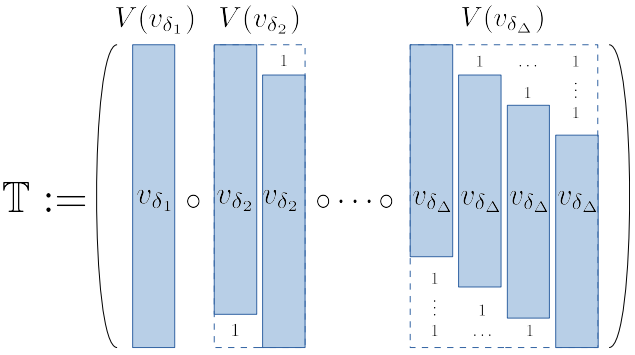

Fig. 1. Visualization of the space transform $\mathbb{T}(\mathcal{V})$

7. Evaluate the value $\bar{y}$ of the detection function $y(t)$ at the splicing point: If $\bar{y} \geq \tau$, with τ being a user defined threshold, then x_1 and x_2 are considered as being recorded by the same microphone.

Figure 2 provides a visual comparison between the proposed detection function $y(t)$ with $\delta_1 < \delta_\Delta$, the particular case of $y(t)$ computed with $\delta_1 = \delta_\Delta$ and the correlation coefficient $\rho(f_1, f_2)$ involved in the baseline algorithm from [11].

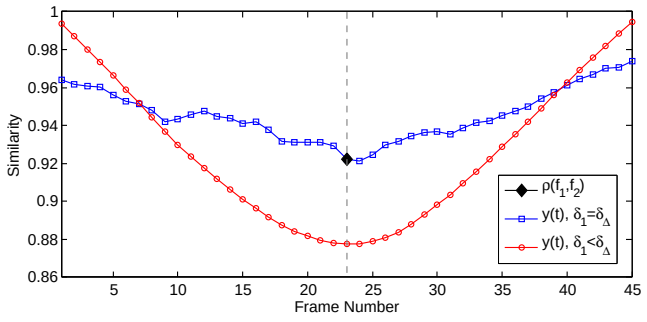

Fig. 2. Visual comparison of the detection functions

If $y(t)$ is computed in its degenerate form, i.e., $\delta_1 = \delta_\Delta = floor(\frac{L}{D})$, then $\bar{y} \simeq \rho(f_1, f_2)$, as evident by comparing the blue line with squares and the black diamond in Figure 2.

For $\delta_1 < \delta_\Delta$ the function $y(t)$ becomes smooth, and the value of $\bar{y}$ decrease significantly in presence of a different device. The red line with circles in Figure 2 was obtained by setting $L = 5$ s, $D = 0.25$ s, $\delta_\Delta = \delta_1 + 10$ and $\hat{h}_j \in \mathbb{R}^{512}$.

Due to the significant improvement regarding discrimination, this enhanced version of the algorithm was selected for the further work, despite being computationally more expensive than the baseline approach: Given two input files x_1 and x_2, both with at least L seconds duration, at least $floor(\frac{L}{D})$ frequency estimates need to be computed, while the baseline algorithm only requires 2 of them. However, the complexity increase is linear, which means that parallel computing could be effectively applied to reduce its impact.

3. OPEN-SET MICROPHONE CLASSIFICATION

The open-set classification approach was designed as follows: Let $\mathcal{X} = \{x^1, x^2, \ldots, x^N\}$ be a set of N unlabeled recordings from K different microphones $\{X_1, X_2, \ldots, X_K\}$.

The goal of a classic classification algorithm is to partition $\mathcal{X}$ into disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K$, so that the following conditions hold:

1. $\cup_{k=1}^{K} \mathcal{X}_k = \mathcal{X}$
2. $k \neq l \implies \mathcal{X}_k \cap \mathcal{X}_l = \emptyset$
3. $\mathcal{X}_k \neq \emptyset \quad \forall k \in [1, K]$
4. $x^i \in \mathcal{X}_k \iff X_k \rightarrow x^i$

where $X \rightarrow x$ implies that x was recorded using the microphone X.

Let $\bar{y}_{ij}$ denote the value of the detection function $y(t)$ evaluated at the splicing point of the recording y_{ij} created by splicing together the two recordings x^i and x^j, as described in Section 2. After fixing the *optimal* global threshold τ^*, an additional condition that reflects our specific system characteristics should be considered:

5. $\exists \tau^* \ : \ \bar{y}_{ij} \geq \tau^* \iff x^i, x^j \in \mathcal{X}_k$.

I.e., the partition is directly induced by the value of the detection function.

The proposed algorithm for open-set classification does not require any previous knowledge about the number of input recordings N or the amount of devices K, since both may increase over time. As a consequence, no training is required, and only one parameter, namely τ^*, must be set.

Starting from a completely new system, where no recording was ever annotated before, the complete procedure can be described as follows:

1. Label the first recording x^1 as coming from the device X_1: $\mathcal{X}_1 = \{x^1\}$.
2. Store the x^1 as the reference recording for class 1: $\mathbf{x}_1 = x^1$
3. For each following recording x^j:
   (a) $\bar{k} = -1$
   (b) for each pre-existing class k, uniquely identified by $\mathcal{X}_k$:
      i. Splice x^j with the reference recording $\mathbf{x}_k$
      ii. Compute the value $\bar{y}_{jk}$ of the detection function $y(t)$ at the splicing point
      iii. If $\bar{y}_{jk} < \tau^*$ set $k = k + 1$ and try with the following class
      iv. Else set $\bar{k} = k$, and stop the loop
   (c) If $\bar{k} = -1$ create a new class:
      i. $\mathcal{X}_{k+1} = \{x^j\}$
      ii. $\mathbf{x}_{k+1} = x^j$
   (d) Else update the $\bar{k}$-th class:
      i. $\mathcal{X}_{\bar{k}} = \mathcal{X}_{\bar{k}} \cup \{x^j\}$
      ii. $\mathbf{x}_{\bar{k}} = \text{select-reference-recording}(\mathcal{X}_{\bar{k}})$

Where the function select-reference-recording($\mathcal{X}$) is computed as follows:

1. $\bar{j} = -1, \bar{y} = -\infty$
2. for each recording $x_j \in \mathcal{X}$:
   (a) Compute $y(t)$ from x_j, and its mean value $\langle y(t) \rangle$

(b) If $\langle y(t) \rangle > \bar{y}$

i. $\bar{y} = \langle y(t) \rangle$

ii. $\bar{j} = j$

3. Return the best reference recording $\mathbf{x} = x_{\bar{j}}$

The algorithm is greedy: The first class fulfilling the requirement $\bar{y} \geq \tau^*$ is selected, following the assumption $\bar{y}_{ij} \geq \tau^* \iff x^i, x^j \in \mathcal{X}_k$.

This choice also reduces the average running time of the system: If K is the amount of already detected classes, the average amount of comparisons needed to classify a new file is equal to $(K+1)/2$.

4. RESULT ANALYSIS

The performance of the proposed approach was evaluated by using audio files recorded with 8 different devices and including 11 different codec and bitrate combinations, as shown in Table 1 and Table 2.

Table 1. Recording devices used

Label	Model
1	Dell Latitude D630, built-in microphone
2	Dell Latitude D630, headset
3	Google Phone G2, built-in microphone
4	Google Phone G2, headset
5	iPhone 4S, built-in microphone
6	iPhone 4S , headset
7	Samsung Galaxy S II, built-in microphone
8	Samsung Galaxy S II, headset

Table 2. Codecs / bitrates used

Encoding	Bitrate (kbps)
PCM	768 (48 kHz, mono, 16-bit)
MP3	256, 192, 128, 96, 64
AAC	192, 128, 96, 64, 32

19 test utterances of 10 s from four different speakers were recorded by each device in three different environments. Before testing, the global threshold parameter $\tau^* = 0.98$ used by the microphone discrimination algorithm (see Section 2.2) was determined for a small set of 8 original and unencoded recordings per device. In order to avoid any interdependency between the two phases, this tuning was performed by using recordings from different devices, not listed in Table 1.

The outcome of the algorithm was evaluated by using Rand Index (RI) [14], Normalized Mutual Information (NMI) [15], and F-measure (F_β) [16]. In order to penalize false negative errors, i.e., recordings from the same device being assigned to different classes, we set $\beta = 5$. The results achieved by the baseline and the enhanced algorithms, evaluated separately for every codec/bitrate combination, are shown in Table 3.

Table 3. Classification Results

Encoding	Baseline			Proposed		
	RI	NMI	F_5	RI	NMI	F_5
PCM	0.94	0.86	0.85	0.99	0.97	0.95
$MP3_{256}$	0.92	0.80	0.78	0.99	0.98	0.96
$MP3_{192}$	0.93	0.87	0.89	0.99	0.98	0.96
$MP3_{128}$	0.96	0.91	0.94	0.97	0.93	0.91
$MP3_{96}$	0.90	0.80	0.72	0.97	0.94	0.91
$MP3_{64}$	0.90	0.77	0.80	0.98	0.91	0.88
AAC_{192}	0.94	0.85	0.84	0.99	0.96	0.94
AAC_{128}	0.94	0.87	0.88	0.99	0.96	0.95
AAC_{96}	0.92	0.80	0.77	0.99	0.95	0.93
AAC_{64}	0.94	0.87	0.86	0.99	0.96	0.94
AAC_{32}	0.83	0.67	0.78	0.94	0.87	0.85

As evident from the results, the enhanced algorithm outperforms the baseline. The lower NMI of the baseline approach, in particular, is due to the incorrect assignment of recordings from different devices to the same classes. Both methods assigned only few recordings to isolated classes, which reflects directly on the F_5 scores. A general decrease of performance can be observed for lower bitrates, and especially so for AAC 32 kbps, but the accuracy of the partitioning produced by the proposed enhanced algorithm remains high, as reflected in the RI.

5. CONCLUSIONS AND OUTLOOK

To the best of our knowledge, the described work represents the first microphone classification approach for an open-set scenario. Being designed for user-generated content, as most other recent approaches in the field [5–9], the approach was tested using recordings from mobile devices, applying common codecs and bitrates.

While being computationally expensive, the proposed algorithm achieved a high RI $\geq 94\%$, independently from the involved scheme, despite using a feature with only one dimension. Scalability aspects with respect to the amount of input data were also taken into account - they are directly related to the amount of detected classes.

Future work will include the investigation of how the enhanced approach for microphone discrimination can be applied to audio tampering detection, and enhancement of channel estimation quality as e.g. suggested in [13]. Furthermore, the applicability of the proposed approach for microphone verification purposes, using a likelihood measure, will be investigated.

6. REFERENCES

[1] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," in *Proceedings of the 9th Workshop on Multimedia & Security*, New York, NY, USA, 2007, MM&Sec '07, pp. 63–74, ACM.

[2] R. Buchholz, C. Kraetzer, and J. Dittmann, "Microphone classification using fourier coefficients," in *International Workshop on Information Hiding*, Stefan Katzenbeisser and Ahmad-Reza Sadeghi, Eds., 2009, pp. 235–246.

[3] D. Garcia-Romero and C.Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 1806–1809.

[4] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," 2011, vol. 7880, pp. 78800P–78800P–15.

[5] C. Hanilci, F. Ertas, T. Ertas, and O. Eskidere, "Recognition of brand and models of cell-phones from recorded speech signals," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 625–634, April 2012.

[6] C. Kotropoulos, "Telephone handset identification using sparse representations of spectral feature sketches," in *Biometrics and Forensics (IWBF), 2013 International Workshop on*, April 2013, pp. 1–4.

[7] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept 2013, pp. 177–182.

[8] R. Aggarwal, S. Singh, A.K. Roul, and N. Khanna, "Cellphone identification using noise estimates from recorded audio," in *Communications and Signal Processing (ICCSP), 2014 International Conference on*, April 2014, pp. 1218–1222.

[9] M. Jahanirad, A.W. Abdul Wahab, N.B. Anuar, M. Yamani Idna Idris, and M.N. Ayub, "Blind identification of source mobile devices using voip calls," in *Region 10 Symposium, 2014 IEEE*, April 2014, pp. 486–491.

[10] P.B. Brandtzaeg, M. Lüders, J. Spangenberg, L. Rath-Wiggins, and A. Følstad, "Emerging journalistic verification practices concerning social media," *Journalism Practice*, pp. 1–20, March 2015.

[11] L. Cuccovillo, S. Mann, P. Aichroth, M. Tagliasacchi, and C. Dittmar, "Blind microphone analysis and stable tone phase analysis for audio tampering detection," in *Audio Engineering Society Convention 135*, Oct 2013.

[12] N.D. Gaubitch, M. Brookes, P.A. Naylor, and D. Sharma, "Single-microphone blind channel identification in speech using spectrum classification," in *Signal Processing Conference, 2011 19th European*, Aug 2011, pp. 1748–1751.

[13] N.D. Gaubitch, M. Brookes, and P.A. Naylor, "Blind channel magnitude response estimation in speech using spectrum classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2162–2171, Oct 2013.

[14] William M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[15] Tarald O. Kvalseth, "Entropy and correlation: Some comments," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 17, no. 3, pp. 517–519, May 1987.

[16] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.