Audio Provenance Analysis in Heterogeneous Media Sets

Milica Gerhardt, Luca Cuccovillo, and Patrick Aichroth
Fraunhofer Institute for Digital Media Technology IDMT
Ehrenbergstraße 31, 98693 Ilemanu, Germany
{ milica gerhardt, luca.cuccovillo, patrick.aichroth}@idmt.fraunhofer.de

Abstract

This paper introduces a framework for Audio Provenance Analysis, addressing the complex challenge of analyzing heterogeneous sets of audio items without requiring any prior knowledge of their content. Our framework applies a novel approach that combines partial audio matching and phylogeny techniques. It constructs directed acyclic graphs to capture the origins and the evolution of content within near-duplicate audio clusters, identifying the least altered versions and tracing the reuse of content within these clusters. The approach is evaluated for two selected application scenarios, demonstrating that it can accurately determine the direction of content reuse and identify parent-child relationships, while also offering a dedicated dataset for benchmarking future research in this area.

1. Introduction

The ability to verify the reliability and origin, i.e. provenance, of audio files is crucial in combating disinformation and ensuring the integrity of media content. This need is particularly evident in fields such as journalism and law enforcement, where the validation of audio material can be pivotal in investigations or fact-checking efforts.

Journalists and law enforcement agencies often face the task of examining media files to trace their distribution and identify the earliest or least altered versions. This process is crucial for gathering information on the lifecycle of audio files in order to verify content authenticity, identify sources of information, and unravel distribution patterns. This analysis becomes increasingly complex when dealing with extensive sets of audio files, where manipulated or decontextualized materials may incorporate segments from genuine sources, and identical content may proliferate across multiple platforms. Therefore, being able to distinguish between derived and original or first-published versions, and to detect the transformations applied, is essential.

In this context, scientists have identified the need for a specialized field focused on tracing the lineage of transformations and identifying the source or the least transformed version within a set. For audio content, this field has been termed audio phylogeny, with all suggested methods requiring audio items to be part of a set of near-duplicates. However, challenges arise in more heterogeneous sets composed of media content from various internet sources or devices, and lacking detailed content information. In such cases, detecting content similarity and transformations remains an unresolved issue in the current state of the art.

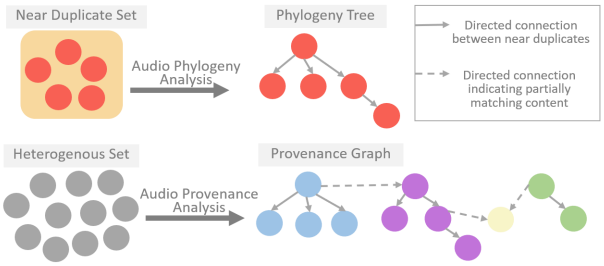

Figure 1. Audio Phylogeny vs Audio Provenance Analysis

This paper introduces an approach for *Audio Provenance Analysis* that addresses these challenges, aiming at mapping the directed relationships among media files by focusing on reused audio segments. The goal is to identify near-duplicate audio sets, reconstruct their lineage in directed acyclic graphs, and highlight partial content reuses that contribute to the creation of new compositions (see Fig. 1).

The proposed approach represents the first comprehensive effort to automate the process that takes a set of heterogeneous audio files as input, and provides provenance graphs as output. As such, it introduces the following contributions to the research domain:

- The definition of the novel task of audio provenance analysis, along with a cohesive workflow that integrates various analysis methods.
- A new clustering methodology that exploits the detailed output of partial audio matching for refined analysis.
- A novel graph-building approach that integrates audio phylogeny analysis with cross-cluster segment matching.
- The creation of a dataset for audio provenance evalua-

tion, demonstrating the effectiveness of our approach and establishing a foundation for future advancements.

2. Literature Review

2.1. Multimedia Phylogeny

Multimedia Phylogeny is a research field aiming to trace the origin and evolutionary pathways of closely related – i.e., near-duplicate – media documents, and thus to identify the source document from a set of near-duplicates and to map the genesis of each near-duplicate, typically visualizing the relations via a so-called phylogeny tree. The foundational principles of this field, especially in the context of images,were laid out by Dias *et al*. in [10], who elaborated on its methodologies and evaluative processes. This pioneering work was further expanded in order to apply multimedia phylogeny principles across various domains, including video [9], audio [12, 19, 24, 34], and even text [31].

Since then, the field has progressed to address more complex scenarios, such as Multiple Parenting Phylogeny. This innovative approach, originally explored for the image domain, extends beyond analyzing near-duplicate sets, to also address compositions derived from combining elements of two donor images. Oliveira *et al*. in [8, 26] delved into multiple parenting image phylogeny and introduced methodologies that facilitate the analysis of image compositions created from one alien and one donor image. However, their studies also highlighted limitations of current methodologies, including constraints on the number of donor images in a composition and dependence on phylogeny forest techniques for effective clustering.

2.2. Provenance Analysis

2.2.1 Image Domain

In response to the limitations in existing Multimedia Phylogeny approaches and to encourage further research into multi-asset forensic analysis, both the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) have played important roles [25]. They introduced new terminology, metrics, and datasets, thereby expanding the scope of phylogeny reconstruction into what is now referred to as Provenance Analysis. This term not only covers the reconstruction of derivation stories of assets, but also emphasizes the critical step of asset retrieval.

In the realm of image analysis, the transition to Image Provenance Analysis has witnessed significant advancements. Building upon the foundation set by de Oliveira *et al*. [8], Bharati *et al*. [3] broadened the framework to include images derived from multiple sources, thereby enhancing the methodology applicability and analytical depth by constructing an undirected graph. Another notable advancement is from Moreira *et al*. [23], who proposed a fully automated framework for image provenance analysis. This framework introduces key methodologies, including Provenance Image Filtering, designed to retrieve both directly and indirectly related images, and utilizes both global and localized dissimilarity metrics to analyze matching regions of images.

Given that Provenance Analysis has largely been confined to the image domain, exploring parallels to potential audio approaches becomes essential. Unlike images, conducting reverse audio searches across the internet for content retrieval is currently impractical, implying that provenance filtering must commence with a predefined set of audio files. Furthermore, there is a deviation from the query-centered approach recommended by NIST for image provenance, highlighting the unique challenges in audio provenance analysis where the application-driven goal is to reveal connections between all files in a set without assuming a predefined query. Despite these methodologies being tailored to the image provenance analysis [23, 36], the core strategy of detecting localized segment matches and constructing a tailored provenance graph bears resemblance to approaches that are applicable for audio file analysis, suggesting that the foundational principles of provenance analysis are potentially applicable across different media types.

2.2.2 Video Domain

Video provenance introduced additional challenges due to the temporal dimension of video content, adding complexities not found in still images. While a comprehensive approach targeting video provenance analysis does not yet exists, initial steps have been taken to develop some of the essential components of such a system. These include foundational approaches within video phylogeny field, applying image phylogeny methodologies in a frame-by-frame manner to video content [4, 5, 9]. Subsequent advancements included techniques for aligning videos by age metrics [22], or temporally aligning similar video sequences [17, 18], establishing the basis for future frameworks in video provenance analysis. These developments signal a growing acknowledgement of the importance of provenance analysis across various media types, each presenting unique challenges and necessitating specialized approaches.

2.2.3 Audio Domain

Similarly, the exploration of Audio Provenance Analysis is in its early stages, yet lacking holistic approach. Nonetheless, several key components that are needed for such a system have been addressed. Drawing inspiration from the image domain, research in Audio Provenance Analysis calls for a comprehensive strategy, including:

1) Provenance Clustering: This critical step involves detecting and localizing content reuse within a collection of audio files in order to form clusters of near-duplicates, identifying the cross-tree connections that indicate segment reuse. Similar to work in the domain of image provenance in [23, 36] and to related video approaches [17, 18], achieving accurate and reliable matching is crucial for the distinction of near-duplicates, and the detection and localization of potentially matching content segments. This requires a method capable of a) reliably detecting perceptually identical content that was created via transformations, b) detecting and localizing matching segments, and c) adapting to an unknown quantity of segments between two files.

Numerous audio matching techniques have been developed, primarily for audio identification purposes, where a query is matched against a database to find its origin. [1, 2, 6, 7, 11, 13, 14, 30, 32, 33, 35]. These methods focus on global similarity and are optimized for retrieval speed, crucial for databases with millions of songs. Content-Based Copy Detection (CBCD), in contrast, represents a more relevant use case for our purpose: In CBCD, the aim shifts towards localizing the match within the reference data, assuming that only a *portion* of the query might match the reference content. This requires retrieval algorithms capable of localization, addressed by several approaches [15, 16, 27–29] that introduced enhancements for detecting a sub-sequence of the query within reference files.

Despite these advancements, most methods struggle with handling multiple partial matches or achieving precise match localization, assuming at most one matching segment between every pair of files, and relying on near-duplicate search, voting, or counting strategies. These strategies, while effective, face limitations regarding granularity and accuracy of match localization. This limitation was addressed by our work in [20, 21], where we proposed an audio matching technique capable of detecting and localizing an unknown number of reused segments, making it an ideal choice for provenance clustering in our Audio Provenance Analysis framework.

2) Provenance Graph Building: The subsequent goal is to synthesize identified clusters and reused audio segments into a directed graph, clarifying the origins of the content and, in cases of partial reuse, the content creation history. The methodology for constructing this graph must be tailored to the audio domain, diverging from the query-centered approaches for image analysis.

As a key component of provenance graph building, audio phylogeny methods could be utilized to reconstruct directed graphs within clusters of near-duplicates. In the literature, audio phylogeny approaches include computationally intensive efforts of Nucci *et al*. in [24], more streamlined approaches leveraging transformation detection functions [19, 34], and the more recent approach we proposed in [12] that relies on Deep Neural Networks (DNN) for transformation detection, improving extensibility and computational efficiency. Indeed, extensibility is critical for real-world applications, allowing for the expansion of the considered transformations set. Contrary to manually engineered transformation detection functions, our DNN-based audio phylogeny approach offer straightforward extension via retraining of the network with appropriate data, bypassing the challenging process of feature engineering. Thus, the DNN-based audio phylogeny we proposed in [12] was chosen for graph reconstruction within near-duplicate content sets, becoming a key component of the provenance graph building process in our proposed Audio Provenance Analysis system.

3. Proposed Approach

In the previous section, we reviewed the most relevant works related to the domains of image, video, and audio provenance analysis. Upon recognizing the need for a comprehensive provenance analysis framework within the audio domain, we drew parallels and inspiration from established methodologies in the image domain. This comparative analysis led us to identify two critical tasks essential for an effective audio provenance framework: Provenance Clustering and Provenance Graph Building. We detailed the state-of-the-art approaches relevant to both tasks within the audio domain and justified our selection of existing audio matching and audio phylogeny methods as being well-suited for integration into our framework. In the following sections, we will discuss our audio provenance analysis framework, presenting how both adapted and novel methodologies for Provenance Clustering and Provenance Graph Building are implemented, in order to address the unique challenges of audio provenance analysis.

3.1. Provenance Clustering

The goal of the Provenance Clustering task, as illustrated in Figure 2, involves initially applying Partial Audio Matching to determine which audio items are related to each other, followed by a Near Duplicate Clustering process. This process aims to group near-duplicate items in clusters, and to identify the connections between non-near-duplicates.

The Partial Audio Matching step utilizes the audio matching approach we proposed in [20, 21]. This method introduces an advanced retrieval algorithm tailored to detect and localize reused audio segments as short as 3 seconds, meeting our requirements for precision and reliability in segment localization. As such, it is the ideal solution for our task of provenance clustering.

Given a set of audio files $\mathcal{A} = \{a_i\}$ let us denote with N their number (i.e., $i \in [1, N]$) and with L_i the length in seconds of the i-th file in the set. The partial matching requires extracting one audio fingerprint F_i for each file under analy-

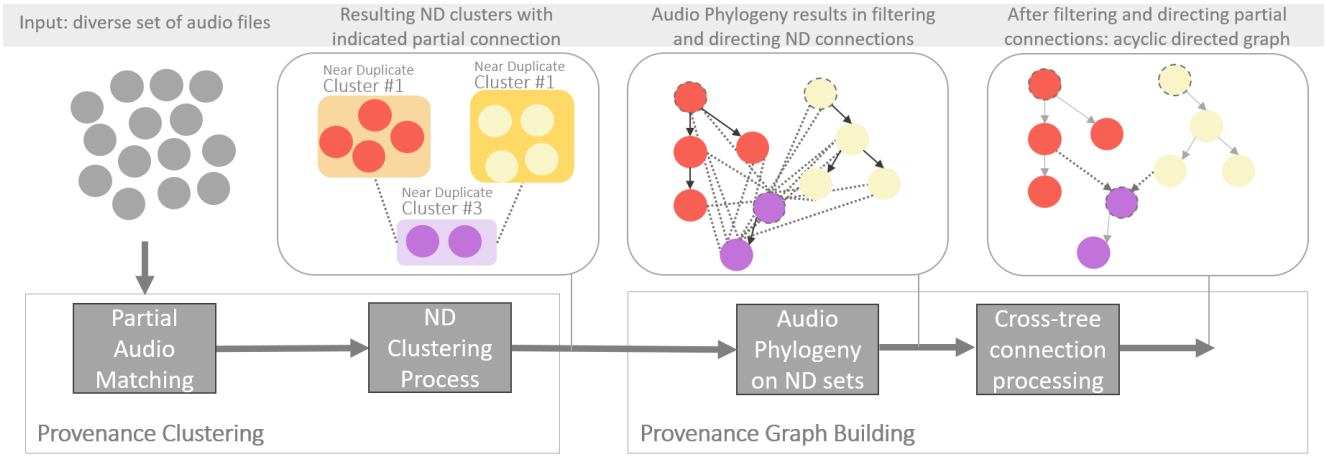

Figure 2. Proposed Audio Provenance Analysis workflow

sis, and performing a pairwise comparison between all pairs of existing fingerprints (F_i, F_j). The outcome of the comparison is a set of matching results, each one of the form

$$\{start\_i,\ start\_j,\ duration,\ confidence,\ i,\ j\}\,, \tag{1}$$

where *start_i* marks the start of the matching in audio item a_i, *start_j* the start of the match in audio a_j and *confidence* is the percent of matching sub-fingerprints over the *duration* of the detected match [20].

Modifying the original notation, we now define the k-th detected match between a file a_i and a file a_j as:

$$m_{ij}^{(k)} = \left(s_i^{(k)}, s_j^{(k)}, l_{ij}^{(k)}\right), k \in [1, K_{ij}] \tag{2}$$

where s_i and s_j indicate the start of the match in the respective file, l_{ij} its length, and K_{ij} the number of matches detected for the pair of files.[1]

The Near Duplicate Clustering process aims to discern which outputs of the partial matching component indicate the presence of a near-duplicate and which ones do not, and to group near-duplicate files into clusters. Let $\mathbf{D} = (D_{ij})$ represent a near-duplicates matrix, where

$$D_{ij} = \begin{cases} 1 & (a_i, a_j) \text{ are near-duplicates} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and let $\mathbf{P} = (P_{ij})$ denote a partial-duplicate matrix in which

$$P_{ij} = \begin{cases} 1 & (a_i, a_j) \text{ are partial-duplicates} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Given two files with at least one detected match ($K_{ij} > 0$), we classify them as near-duplicates if the total length of the respective matching segments exceeds the length of both files, or as partial-duplicates otherwise, i.e.

$$D_{ij} = \begin{cases} 1 & \sum_{k=1}^{K_{ij}} l_{ij}^{(k)} > \max(L_i, L_j) - \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$P_{ij} = \begin{cases} 1 & K_{ij} > 0 \wedge D_{ij} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where ϵ accounts for a small tolerance to minor localization inaccuracies due, for example, to background noise or encoding, and L_i, L_j represent the length in seconds of the two audio files.

3.2. Provenance Graph Building

The goal of the Provenance Graph Building task is depicted in Figure 2. In a first step, we transform the clusters of near-duplicates into set of phylogeny trees, i.e., of directed graphs in which the direction indicates provenance; In a second step, we process the cross-tree partial connections, i.e., the partial matching existing between disjoint phylogeny trees, to pinpoint the specific files acting as donors in the creation of derived content.

3.2.1 Audio Phylogeny

Audio phylogeny aims to detect the relationships and transformations within a set of near-duplicate audio items. This involves computing a dissimilarity matrix between each pair of near-duplicates, which is then transformed into a directed phylogeny tree using the Oriented Kruskal algorithm [10].

Among the few algorithms for audio phylogeny discussed in Section 2, we chose the one we presented in [12], which proposes the use of a neural network to detect the most probable transformation that occurred between every input pair of near-duplicates. This approach offers high computational efficiency, enables detection of specific

[1] The confidence value is not relevant in this context and was therefore omitted.

transformation between pairs of files, and allows for the expansion of the set of potentially detected transformations with relative ease.

As with partial matching, we will omit the specifics of the phylogeny analysis – the workflow of which is detailed in [12]. Instead, we focus on how to apply the output of the algorithm for our provenance analysis task.

Let us denote a single cluster of near-duplicates with $\mathcal{C}_n$. Before the phylogeny analysis, our only information is that for every pair (i, j) where $a_i, a_j \in \mathcal{C}_n$, both elements D_{ij}, D_{ji} in the near duplicate matrix $\mathbf{D}$ are equivalent and set to one.

The phylogeny analysis, in turn, yields an asymmetric matrix of edges $\mathbf{E} = (E_{ij})$, where

$$E_{ij} = \begin{cases} 1 & \text{if } \exists \mathcal{T}(\cdot) : a_j = \mathcal{T}(a_i) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

with $\mathcal{T}(\cdot)$ representing a content-preserving transformation. Hence, if $E_{ij} = 1$, then a_i is deemed a parent of a_j.

To disseminate this information, we impose $D_{ij} = E_{ij} \forall (a_i, a_j) \in \mathcal{C}_n$, repeat the operation for all identified near-duplicate clusters within our provenance graph, and reduce the number of relevant relationships from $|C_n|(|C_n| - 1)/2$ to $|C_n|$.

3.2.2 Processing of Cross-tree Connections

Following the phylogeny analysis, our original file set $\mathcal{A}$ is mapped to a forest of phylogeny trees represented by

$$T_i = (a_{i^*}, \mathcal{C}_i, \mathbf{E}_i), \tag{8}$$

with a_{i^*} indicating its root audio file, $\mathcal{C}_i$ the cluster of files in the tree, and $\mathbf{E}_i$ their adjacency matrix.

Hereon, we assume that if two trees (T_i, T_j) share partial connections and T_i was used to generate T_j, this happens because one audio file $\bar{a}_i \in C_i$ in the first tree was used to generate the root file a_{j^*} of the second tree. In more concise terms, we can identify a donor tree T_i, a donor audio $\bar{a}_i$, a composition root a_{j^*}, and a composition tree T_j.

For an arbitrary pair of trees (T_i, T_j) with partial connections we cannot tell in advance which tree is the composition, and which one is the donor: The goal of cross-tree connection processing is to identify the donor audio $\bar{a}_{(T_i,T_j)}$ for all pairs of phylogeny trees in the analysis set, and to prune all partial connections except the ones between $\bar{a}_{(T_i,T_j)}$ and either the root a_{i^*} or a_{j^*}.

The initial step in achieving this goal involves determining which segments of the audio files to analyze. We propose selecting the interval defined by the longest matching segment, thus

$$\arg\max_{ijk} \left(l_{ij}^{(k)} \right), \quad \forall m_{ij}^{(k)} : (i, j) \in \mathcal{C}_i \times \mathcal{C}_j. \tag{9}$$

Subsequently, we crop all files in $\mathcal{C}_i \cup \mathcal{C}_j$ to the corresponding optimal interval, effectively resulting in a set of near-duplicate partial files:

$$\mathcal{P} = \mathcal{P}_i \cup \mathcal{P}_j = \{p_i \in \mathcal{C}_i\} \cup \{p_j \in \mathcal{C}_j\} \tag{10}$$

where the root elements p_{i^*} and p_{j^*} are distinguishable, and the original cluster is known.

To identify the donor audio, we revisit the dissimilarity calculation process $\mathrm{diss}(\cdot)$ outlined in our audio phylogeny approach [12]. This process is applied to compare all elements of $\mathcal{P}_j$ against the root p_{i^*}, and all elements of $\mathcal{P}_i$ against the root p_{j^*}, thereby obtaining two dissimilarity vectors:

$$D_i^{(j)} = \mathrm{diss}(p_i, p_{j^*}) \quad \forall p_i \in \mathcal{C}_i \tag{11}$$

$$D_j^{(i)} = \mathrm{diss}(p_j, p_{i^*}) \quad \forall p_j \in \mathcal{C}_j \tag{12}$$

Donor audio, donor tree, composition root and composition tree can then be determined by identifying the donor file

$$\bar{a}_{(T_i,T_j)} = \begin{cases} a_{i^\dagger} \in \mathcal{C}_i & \text{if } \min(D_i^{(j)}) < \min(D_j^{(i)}) \\ a_{j^\dagger} \in \mathcal{C}_j & \text{otherwise} \end{cases} \tag{13}$$

where

$$i^\dagger = \arg\min_i (D_i^{(j)}) \tag{14}$$

$$j^\dagger = \arg\min_j (D_j^{(i)}) \tag{15}$$

i.e., by selecting the connection with the lowest dissimilarity.

Once the selection process is completed, all non-corresponding partial connections P_{ij} in the partial-duplicate matrix $\mathbf{P}$ are discarded unless they relate to the identified donor files $\bar{a}_{(T_i,T_j)}$ and the related composition roots a_{i^*} or a_{j^*}.

The outcome of the provenance graph building is a directed acyclic graph, exemplified at the top right of Figure 2: Each phylogeny tree can be identified by its cluster number (color in the figure), and within each tree every node is restricted to have a single parent, albeit possibly having multiple descendants. Between pairs of trees, only a unique directed connection is permitted from an arbitrary node of the donor tree to the root node of the composition tree, with the result that composition root trees may have multiple inbound partial connections.

4. Evaluation

In this section, we introduce a dataset designed for the evaluation of audio provenance analysis for selected usage scenarios, released together with this publication[2]. In addition, we elaborate on the evaluation metrics and summarize the results achieved by our proposed framework using the proposed dataset.

[2] https://doi.org/10.5281/zenodo.10960056

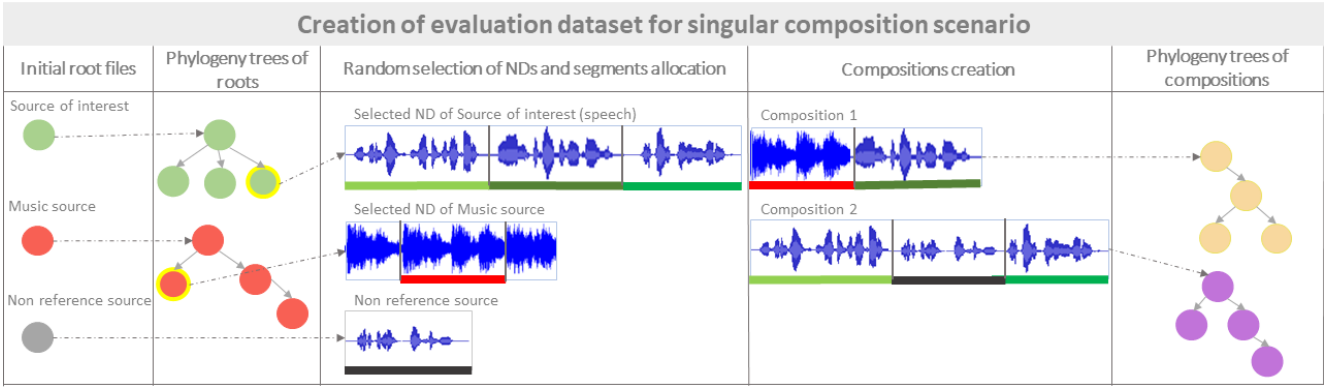

Figure 3. Dataset for the Singular Composition Scenario

4.1. Test Set Creation

In the following, we will outline two scenarios, Singular Composition and Multi-Source Composition, which reflect two common composition cases for audio/video media content.

Singular Composition (SC) outlines a scenario where a single source is segmented and integrated with other content. This happens, for instance, when fragments of interviews or statements are reused in various contexts. For this purpose, the creation process for the SC test dataset involves starting with a source of interest (SoI), a music source (MS), and a non-reference source (NS), as illustrated in Fig. 3. We construct a phylogeny tree from the SoI and the MS to reflect the various versions that might be discovered online or on different devices, and utilized in compilations, while the NS represents content not accessible for analysis, like restricted or private archive material. Two compilations are generated from randomly selected nodes of the SoI and MS phylogeny trees and a NS by merging their segments as illustrated in Fig. 3, where the first compilation contains one segment from SoI and the second includes two. The SC scenario does not aim at hiding the splicing of segments from different sources, hence audio segments are combined using simple concatenations with a 0.1-second crossfade to eliminate perceivable clicks. We create 40 sets under the SC scenario, featuring 80 audio files each.

Multi-Source Composition (MSC) introduces a second scenario where segments from two sources are utilized to create new content. This scenario draws inspiration from malicious content creation, such as a manipulated statement from a politician. The MSC test dataset creation involves two SoIs and a NS. Phylogeny trees are constructed for each SoI to represent possible variants used in compilations, with the NS representing synthetic or inaccessible content. Compositions in MSC involve merging segments from both SoIs with the NS, as specified in Fig. 4. Aligned with the malicious intentions of the creator in this scenario, we create compositions by first applying cross-fade and then introducing background environmental noise to conceal the splicing, thereby "pretending" an original recording. Moreover, all reused content (SoIs and non reference one) originates from the same speaker. Each MSC test dataset comprises three phylogeny trees with 20 nodes each, with roots in the two SoIs and the MSC compilation. We create a total of 40 of these datasets, with 60 audio files each.

The choice of these composition scenarios was influenced not only by their application relevance, but also by the goal to encompass a wide array of composition characteristics. This includes combining music and speech, utilising segments from the same source, merging segments from the same speaker but different sources, and mixing reference with non-reference material. The scenarios also differentiate between forests with two compositions (SC) and forests with a single composition (MS).

For the sake of a streamlined evaluation, the duration of all reused segments was set to 4 seconds. This duration ensures that the segments are suitable for audio phylogeny analysis while keeping the composed signals short. The generated phylogeny trees include 20 near duplicates nodes created by applying transformations such as MP3 and AAC encoding (320, 192 and, 128 kbps), and fading with a range of 0 to 3 seconds. Notably, unlike the transformation sets for generating audio phylogeny trees commonly referenced to in the literature, our selected set does not include trimming; the trim operation was omitted because it could be identified by the partial matching component, making it irrelevant for the reconstruction of phylogeny trees.

4.2. Evaluation metrics

Similar to the approach in [23], we utilize generalized F1 measures for the evaluation of both retrieved nodes and edges, named *Vertices Overlap* (VO) and *Edges Overlap* (EO), respectively. Moreover, acknowledging the importance of accurately detecting roots for analyzing cross-tree connections (as outlined in Sec. 3.2.2), we also examine

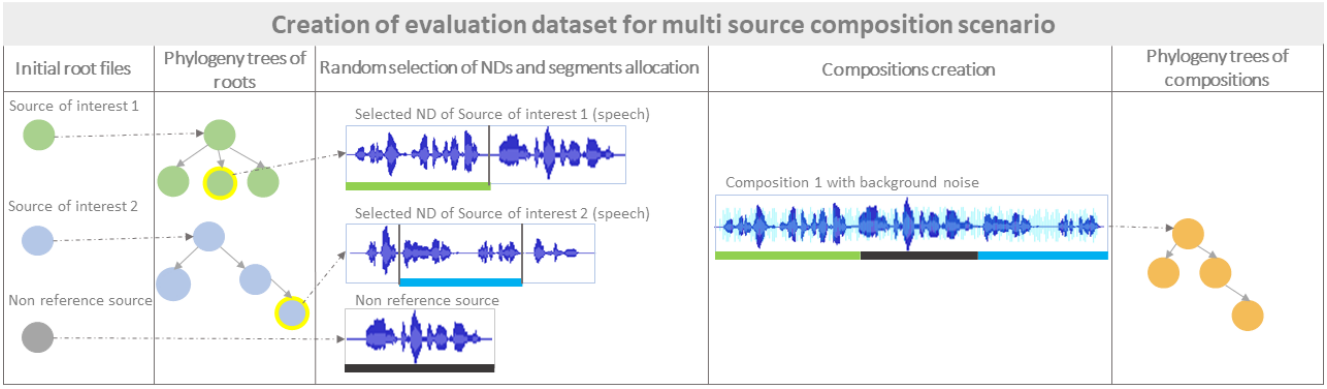

Figure 4. Dataset for the Multi Source Composition Scenario

whether the roots of all near-duplicate sets within an examined heterogeneous set were correctly identified (R). Thanks to these three measures, we can assess the efficiency of both the clustering process and the overall graph reconstruction (including near-duplicate and partial cross-tree edges). To compare the ground truth graph G and reconstructed graph $G^{'}$, we calculate:

$$EO = 2\frac{|E \cap E^{'}|}{|E| + |E^{'}|} \tag{16}$$

$$VO = 2\frac{|V \cap V^{'}|}{|V| + |V^{'}|} \tag{17}$$

$$R(G, G^{'}) = \begin{cases} 1, & \text{If } R = R^{'} \\ 0, & \text{Otherwise} \end{cases} \tag{18}$$

where V, E, R denote nodes, edges and root of ground truth graph respectively, and $V^{'}$, $E^{'}$, and $R^{'}$ denote nodes, edges and root of reconstructed graph.

Given the particular focus on cross-tree edges –three per test set in the SC scenario and two in the MSC scenario, which constitute only 0.03% of the total set of edges E– the effectiveness of reconstructing these crucial edges can be somewhat masked by the global measure EO. To mitigate this and provide a clearer insight into our framework capability in reconstructing cross-tree connections, we introduce additional metrics:

Accurate Partial Detection (PD)*:* This metric assesses the quantity of partial connections detected by the partial audio matching component relative to the ground truth number of connections. In test datasets outlined in Sec. 4.1, the ground truth number of connection is 3 in the SC scenario, while being 2 in the MSC scenario.

Accurate Directed Partial Detection (PD_{C_i,a_j^*})*:* This metric measures the quantity of total connections that are correctly oriented and link the correct donor cluster to the correct root node of a composition. It is relative to the ground truth number of connections.

	R	VO	EO	PD	PD_{C_i,a_j^*}	$PD_{a_{i\dagger},a_j^*}$
scenario SC	1	1	0.85	1	0.92	0.53
scenario SC (*)	1	1	0.89	1	0.925	0.51
scenario MSC	1	1	0.83	0.44	0.15	0.01
scenario MSC (*)	1	1	0.88	0.44	**0.47**	0.1

Table 1. Evaluation results averaged over 40 trees per each evaluated scenario of Single Composition (SC) and Multi Source Composition (MSC). Rows marked with (*) represent results obtained with retrained DNN for transformation detection.

Accurate Donor-Composition Connection ($PD_{a_{i\dagger},a_j^*}$)*:* Representing the most rigorous criterion, this metric calculates the quantity of partial connections that accurately link the correct donor node to the correct root of the composition cluster. Again, it is relative to the ground truth number of connections.

4.3. Results

In the table presented in Tab. 1, we summarize the evaluation results based on the criteria detailed in Sec. 4.2, for both scenarios examined in Sec. 4.1. The results reflect the average performance across all 40 heterogeneous audio sets per scenario, hence all metric scores range from zero to one, where one indicates optimal performance.

The optimal scores of VO and R in both scenarios demonstrate that all nodes (vertices) of ground truth graphs are included in reconstructed graphs. Furthermore, this confirms the accurate identification of the roots of each near-duplicate cluster within the reconstructions. The metrics evaluating correct edges overlap EO in reconstructed graphs also reveal high scores in both scenarios, although slightly lesser than those reported by our previous work in [12], likely due to the consideration of cross-tree connections in our evaluation.

In the SC scenario, the partial matching component effectively identified the correct number of partial cross-tree connections for every test set evaluated, achieving a PD

value of 1.0. Conversely, the MSC scenario posed a greater challenge. The application of cross-fading between content segments and background noise resulted in a PD value of 0.44, indicating that only 44% of the expected cross-tree matches were identified. This disparity significantly impacted the overall performance metrics for partial connections within the MSC scenario. In the SC scenario, a high value of 0.92 was noted for partial matches correctly connecting the donor cluster to the composition's root node (PD_{C_i,a_j^*}), while this metric dropped to 0.15 in the MSC scenario. We attribute this decline not only to the difficulty in detecting cross-tree connections, but also to an overlooked transformation: The addition of background noise, which was not included in the detectable transformations in the DNN used for transformation detection within the audio phylogeny method, as mentioned in [12].

Recognizing the necessity to incorporate this additional content transformation, we took advantage of the extensibility of the audio phylogeny approach by retraining the DNN to detect background noise addition as a transformation.

The results labeled with (*) in Table 1 show the improvements achieved by this adjustment. For the SC scenario, the retraining had minimal impact, as evidenced by the similar outcomes in both cases. However, for the MSC scenario, the retrained model's performance PD_{C_i,a_j^*} improved significantly to 0.47, indicating a substantial improvement in identifying correct cross-tree connections when the additional noise transformation was considered. This adjustment makes the performance of the MSC scenario comparable to the performance of the SC scenario for the metric PD_{C_i,a_j^*}: After retraining the DNN, the ratio between the number of accurately directed partial connections relative to the number of detected partial connections is 100% in both cases. This underscores the importance of comprehensive transformation detection for complex audio provenance scenarios.

The measure ($PD_{a_{i\dagger},a_j^*}$), which denotes whether the correctly directed cross-tree connection originates from the accurate node within the donor cluster, is indeed the most demanding test of our framework capabilities. For the SC scenario, this metric scores at 0.53, showing that over half of the cross-tree connections are correctly traced back to their correct origin within the donor cluster. Conversely, the MSC scenario, even after incorporating the retrained DNN, achieves only a score of 0.1. While this value may seem low, it is essential to acknowledge the inherent difficulty of this task. In scenarios where the task involves distinguishing between near-duplicate files, the challenge is intensified by the close resemblance among these files. Often, the only variances consist of subtle transformations like fade in/out, which, for segments extracted from the middle of a file, rarely alter the fundamental characteristics of the content.

5. Conclusion and Outlook

In this paper, we have presented an innovative audio provenance analysis framework, providing a novel solution within the current landscape for examining heterogeneous sets of audio files. Starting with a collection of media files lacking prior information on content similarity, our framework successfully generates an acyclic directed graph. This graph not only identifies sources –or the least changed versions– within near-duplicate clusters, but also maps outbound connections that indicate partial content reuse. Remarkably, for all detected partial connections, our system can accurately detect the direction of content reuse – distinguishing between donor and recipient clusters within these relationships.

However, pinpointing the exact node within a near duplicate cluster as the donor of a content segment to a composition node remains a substantial challenge. In the current system configuration, this precise identification was accomplished in 53% of all cases in the SC scenario and only in 10% of all cases in the more demanding MSC scenario. This issue constitutes a challenging area for future research, with several aspects of the current framework that could be optimized to enhance performance, such as refining the dissimilarity calculations between segments to be more nuanced or noise-aware.

Another challenge identified is the detection of partial matches if background noise is present, where performance significantly drops, as evidenced by the 44% success rate in the MSC scenario compared to the 92% in the absence of noise in the SC scenario. This disparity highlights the necessity for further refinement of the audio fingerprinting technique we originally proposed in [20], which, although effective for identifying perceptually identical content, exhibits limitations under noisy conditions. Future improvements could involve adjusting the existing fingerprint parameters or developing a new fingerprint more robust to background noise.

In conclusion, our work can be considered a significant advancement in the field of audio provenance analysis, tackling a previously uncharted task. We have not only developed a practical solution but also created and shared a dataset suitable for benchmarking. In our discussion of the evaluation results, we have highlighted the framework strengths and weaknesses, and related avenues for subsequent research to achieve further improvements.

Acknowledgments

This paper was supported by the EU Horizon Europe vera.ai project (grant no. 101070093), and by the BMBF newspolygraph project (grant no. 03RU2U151D).

References

[1] Xavier Anguera, Antonio Garzon, and Tomasz Adamek. Mask: Robust local features for audio fingerprinting. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 455–460, 2012. 3

[2] Shumeet Baluja and Michele Covell. Audio fingerprinting: Combining computer vision & data stream processing. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages II–213–II–216, 2007. 3

[3] A. Bharati, D. Moreira, A. Pinto, J. Brogan, K. Bowyer, P. Flynn, W. Scheirer, and A. Rocha. U-phylogeny: Undirected provenance graph construction in the wild. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1517–1521, 2017. 2

[4] Filipe de O. Costa, Silvia Lameri, Paolo Bestagini, Zanoni Dias, Stefano Tubaro, and Anderson Rocha. Hash-based frame selection for video phylogeny. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2016. 2

[5] F. O. Costa, S. Lameri, P. Bestagini, Z. Dias, A. Rocha, M. Tagliasacchi, and S. Tubaro. Phylogeny reconstruction for misaligned and compressed video sequences. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 301–305, 2015. 2

[6] Courtenay V. Cotton and Daniel P. W. Ellis. Audio fingerprinting to identify multiple videos of an event. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2386–2389, 2010. 3

[7] Michele Covell and Shumeet Baluja. Known-audio detection using waveprint: Spectrogram fingerprinting by wavelet hashing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages I–237–I–240, 2007. 3

[8] Alberto A. de Oliveira, Pasquale Ferrara, Alessia De Rosa, Alessandro Piva, Mauro Barni, Siome Goldenstein, Zanoni Dias, and Anderson Rocha. Multiple parenting phylogeny relationships in digital images. *IEEE Transactions on Information Forensics and Security*, 11(2):328–343, 2016. 2

[9] Zanoni Dias, Anderson Rocha, and Siome Goldenstein. Video phylogeny: Recovering near-duplicate video relationships. In *2011 IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2011. 2

[10] Zanoni Dias, Anderson Rocha, and Siome Goldenstein. Image phylogeny by minimal spanning trees. *IEEE Transactions on Information Forensics and Security*, 7(2):774–788, 2012. 2, 4

[11] Jacob George and Ashok Jhunjhunwala. Scalable and robust audio fingerprinting method tolerable to time-stretching. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 436–440, 2015. 3

[12] Milica Gerhardt, Luca Cuccovillo, and Patrick Aichroth. Advancing audio phylogeny: A neural network approach for transformation detection. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2023. 2, 3, 4, 5, 7, 8

[13] Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In *ISMIR International Conference on Music Information Retrieval*, pages 107–115, 2002. 3

[14] J. Haitsma and T. Kalker. Speed-change resistant audio fingerprinting using auto-correlation. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, pages IV–728, 2003. 3

[15] Maguelonne Héritier, Vishwa Gupta, Langis Gagnon, Gilles Boulianne, and Patrick Cardinal Samuel Foucher. Crim's content-based copy detection system for trecvid. In *NIST TREC Video Retrieval Evaluation (TRECVID) Conference*, 2009. 3

[16] Hervé Jégou, Jonathan Delhumeau, Jiangbo Yuan, Guillaume Gravier, and Patrick Gros. BABAZ: A large scale audio search system for video copy detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2369–2372, 2012. 3

[17] S. Lameri, P. Bestagini, A. Mellon, S. Milani, A. Rocha, M. Tagliasacchi, and S. Tubaro. Who is my parent? reconstructing video sequences from partially matching shots. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5342–5346, 2014. 2, 3

[18] Silvia Lameri, Paolo Bestagini, and Stefano Tubaro. Video alignment for phylogenetic analysis. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 2255–2259, 2016. 2, 3

[19] Milica Maksimovic, Luca Cuccovillo, and Patrick Aichroth. Phylogeny analysis for MP3 and AAC coding transformations. In *ICME*, 2017. 2, 3

[20] Milica Maksimović, Patrick Aichroth, and Luca Cuccovillo. Detection and localization of partial audio matches. In *International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2018. 3, 4, 8

[21] Milica Maksimović, Patrick Aichroth, and Luca Cuccovillo. Detection and localization of partial audio matches in various application scenarios. *Multimedia Tools and Applications*, 80(1):22619–22641, 2021. 3

[22] Simone Milani, Paolo Bestagini, and Stefano Tubaro. Video phylogeny tree reconstruction using aging measures. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2181–2185, 2017. 2

[23] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer. Image provenance analysis at scale, 2018. 2, 3, 6

[24] Matteo Nucci, Marco Tagliasacchi, and Stefano Tubaro. A phylogenetic analysis of near-duplicate audio tracks. In *MMSP*, 2013. 2, 3

[25] National Institute of Standards and Technology. Nimble challenge 2017 evaluation. https://www.nist.gov/itl/iad/mig/open-media-forensics-challenge, 2017. 2

[26] A. Oliveira, P. Ferrara, A. De Rosa, A. Piva, M. Barni, S. Goldenstein, Z. Dias, and A. Rocha. Multiple parenting identification in image phylogeny. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5347–5351, 2014. 2

[27] Chahid Ouali, Pierre Dumouchel, and Vishwa Gupta. A robust audio fingerprinting method for content-based copy detection. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2014. 3

[28] Chahid Ouali, Pierre Dumouchel, and Vishwa Gupta. Efficient spectrogram-based binary image feature for audio copy detection. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1792–1796, 2015.

[29] Ahmet Saracoglu, Ersin Esen, Tugrul K. Ates, Banu Oskay Acar, Unal Zubari, Ezgi C. Ozan, Egemen Ozalp, A. Aydin Alatan, and Tolga Ciloglu. Content based copy detection with coarse audio-visual fingerprints. In *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pages 213–218, 2009. 3

[30] Hendrik Schreiber and Meinard Müller. Accelerating index-based audio identification. *IEEE Transactions on Multimedia*, 16(6):1654–1664, 2014. 3

[31] Bingyu Shen, Christopher W. Forstall, Anderson De Rezende Rocha, and Walter J. Scheirer. Practical text phylogeny for real-world settings. *IEEE Access*, 6: 41002–41012, 2018. 2

[32] Joren Six and Marc Leman. Panako - a scalable acoustic fingerprinting system handling time-scale and pitch modification. In *International Society for Music Information Retrieval Conference*, 2014. 3

[33] Reinhard Sonnleitner and Gerhard Widmer. Robust quad-based audio fingerprinting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):409–421, 2016. 3

[34] Sebastiano Verde, Simone Milani, Paolo Bestagini, and Stefano Tubaro. Audio phylogenetic analysis using geometric transforms. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2017. 2, 3

[35] Avery Wang. An industrial strength audio search algorithm. In *ISMIR International Conference on Music Information Retrieval*, 2003. 3

[36] Xu Zhang, Zhaohui H. Sun, Svebor Karaman, and Shih-Fu Chang. Discovering image manipulation history by pairwise relation and forensics tools. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1012–1023, 2020. 2, 3