Forensic Science International: Digital Investigation 49 (2024) 301765

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS USA 2024 - Selected Papers from the 24th Annual Digital Forensics Research Conference USA

Enhancing speaker identification in criminal investigations through clusterization and rank-based scoring

Antonio Artur Moura [a,*], Napoleão Nepomuceno [a], Vasco Furtado [a,b]

[a] University of Fortaleza, Graduate Program in Applied Informatics, Av. Washington Soares 1321, Fortaleza, 60811-905, Ceará, Brazil
[b] Empresa de Tecnologia da Informação do Ceará, Av. Pontes Vieira, 220, Fortaleza, 60.130-240, Ceará, Brazil

ARTICLE INFO

Keywords:
Digital forensic
Audio analytics
Speaker recognition

ABSTRACT

This paper introduces an approach that supports speaker identification in criminal investigations, specifically addressing challenges associated with large volumes of audio recordings featuring unknown speaker identities. Our approach clusters related recordings – potentially from the same person – based on representative voice embeddings extracted using the ECAPA-TDNN speaker recognition model. Grouping audio recordings from the same person enhances variability and richness in voice patterns, thereby improving confidence in automatic speaker recognition. We propose a combination of cosine similarity and a rank-based adjustment function to determine matches of audio clusters with individuals in an enrollment database. Our approach was validated through experiments on a Common Voice-based synthesized dataset and a real-life application involving cell phones seized in prisons, which contained thousands of conversational audio recordings. Results demonstrated satisfactory performance and stability, consistently reducing the pool of candidate speakers for subsequent analysis by a human investigator.

1. Introduction

The surge in digital services and the growing prevalence of digital device usage among individuals have led to an increase in the volume of data associated with human actions and interactions. In this context, there has been a fundamental transformation in criminal investigations in recent years (Quick and Choo, 2014, 2018). Particularly, the identification of speakers in audio recordings from various sources (e.g., cell phones, notebooks, etc.) is a non-trivial task – even when aided by computerized methods – that plays an important role in criminal investigations (Hansen and Bořil, 2018; Saleem et al., 2020; Basu et al., 2022; Guan, 2022). Electronic devices contain a high amount of unlabeled data, including a variety of audio recordings of interest, whose speakers' identities may not be known. Under these circumstances, speaker recognition becomes a vital element in identifying the speaker of a given audio recording through the analysis of voice patterns.

A common difficulty in employing speaker recognition in the context of criminal investigations is the need for an enrollment database, which is rarely available or often does not meet the appropriate level of quality. Common issues encountered include: (i) insufficient deployment of audio recording devices, (ii) low-quality microphones, (iii) high levels of background noise, (iv) varied audio levels across devices or locations, and (v) cross-talk and voice overlapping. Another technological challenge encountered when performing such a task is that, with the expansion of records in the enrollment database, the likelihood of false positives also rises. In other words, there is an increased occurrence of matches where the identity of the record retrieved from the enrollment database does not align with the actual identity of the speaker in the input audio.

Addressing these challenges poses a difficulty, demanding approaches primarily focused on reducing false positives to a manageable level without overlooking true positives. Ideally, reproducing the functionality of face recognition for speaker recognition would be a straightforward task. However, unlike facial expressions that generally remain stable, the voice is susceptible to rapid changes influenced by the person's behavior, intentions, health, and environmental factors such as interference and noise. This dynamic nature complicates the task of speaker recognition. While a speaker recognition system may not achieve the same accuracy as facial recognition, it remains valuable in eliminating potential suspects, increasing human investigator productivity, and validating existing evidence.

Speaker recognition systems differentiate between text-dependent,

* Corresponding author.
E-mail addresses: antonioartur@edu.unifor.br (A.A. Moura), napoleaovn@unifor.br (N. Nepomuceno), vasco@unifor.br (Vasco Furtado).

https://doi.org/10.1016/j.fsidi.2024.301765

Available online 5 July 2024

relying on specific word criteria, and text-independent approaches. Within these approaches, there are two types of solutions: closed-set and open-set. Closed-set solutions, tailored for specific speaker groups, are more restrictive, whereas open-set solutions offer greater flexibility, accommodating additional speakers. In real-world criminal investigations, where control over recording settings and speaker identity is limited, an open-set, text-independent approach is essential for effective speaker identification.

This paper introduces an approach aimed at equipping investigators with a text-independent, open-set speaker recognition system, capable of identifying speakers or narrowing down the pool of potential candidates. Our approach integrates the ECAPA-TDNN-based system (Desplanques et al., 2020) – which remains the state-of-the-art speaker verification framework in realistic scenarios to date (Sharma et al., 2024) – with variance mitigation strategies, ensuring good performance in the realm of automatic speaker identification within criminal investigations. To validate our approach, we conducted experiments on a Common Voice based synthesized dataset and applied it to a real-life scenario involving data from cell phones provided by the Public Ministry of the State of Ceará (MPCE) and the enrollment database supplied by the Penitentiary Administration Secretariat of the State of Ceará (SAP-CE). Government entities endowed with criminal investigation authority, such as MPCE, typically employ specialized software (e.g., Cellebrite UFED) to extract documents, images, videos, audio, and conversations from seized cell phones. Within this context, MPCE developed the *Digital Evidence Exploration Portal* (DEEP) with the goal of standardizing the design, implementation, and execution of criminal investigation procedures. Despite the incorporation of face recognition and natural language processing algorithms into DEEP, the potential of audio data has remained untapped, necessitating manual exploitation.

Our contribution is an approach for identifying speakers in audio recordings extracted from seized electronic devices by the analysis of voice patterns. The proposed speaker recognition system is self-tuning, able to handle a large enrollment database, producing relatively few false positives, and requiring minimal intervention from the investigator. It is important to highlight that the strategies employed also show promise for application in other pattern recognition scenarios where the data has a vector representation. In addition, affirmative responses are provided to the following research questions: (i) Can state-of-the-art speaker recognition models be applied to languages not originally considered in the model and (ii) Can a speaker recognition system support speaker identification in criminal investigations, addressing challenges associated with large volumes of audio recordings and handling enrollment databases with tens of thousands of individuals?

The remainder of this paper is structured as follows: Section 2 provides an overview of the current state of the art in speaker identification, focusing specifically on forensic speaker recognition. Section 3 introduces a speaker identification approach tailored for criminal investigations, incorporating clusterization and rank-based scoring strategies. In Section 4, we apply the proposed approach to a synthesized dataset for initial evaluation and further deploy it on real-world data within the context of criminal investigations. Finally, in Section 5, we conclude the paper with final remarks and outline perspectives for future work. A prototype application showcasing the practical implementation of the proposed approach for user interaction in the field of audio investigations is presented in Appendix A.

2. Related work

Speaker recognition, or voice biometrics, involves identifying or verifying individuals by analysing unique vocal characteristics, such as speech patterns, pronunciation, pitch, tone, and other acoustic features (Bricker and Pruzansky, 1976; Holmes, 1985; Peacocke and Graf, 1995; Jahangir et al., 2021). Historically, i-vector methods (Dehak et al., 2010; Kanagasundaram et al., 2011; Travadi et al., 2014) have been foundational in speaker recognition, building upon the Gaussian Mixture Model-Universal Background Model framework (GMM-UBM) (Reynolds et al., 2000) and incorporating probabilistic discriminant analysis (PLDA) (Cumani et al., 2013; Prince and Elder, 2007; Matějka et al., 2011; Garcia-Romero and Espy-Wilson, 2011) for classification. Recent years witnessed a shift to deep learning architectures (Variani et al., 2014; Chen et al., 2015; Li et al., 2017; Zhang et al., 2016; Sadjadi et al., 2016; Ravanelli and Bengio, 2018; Desplanques et al., 2020) such as Convolutional Neural Networks (CNNs) (Li et al., 2017) and Time Delay Neural Networks (TDNNs) (Desplanques et al., 2020). These models, accepting speech input in raw waveform, spectrogram, or Mel-Frequency Cepstral Coefficients (MFCCs) formats (Li et al., 2017), outperform traditional methods. Novel architectures like SincNet (Ravanelli and Bengio, 2018) further explore the landscape, enhancing performance on the i-vector benchmark.

Building on advancements in speaker recognition, practical applications are evident in forensic speaker recognition (FSR), where scientists focus on addressing challenges associated with identifying unknown speakers in audio recordings. The study by Saleem et al. (2020) aims to enhance FSR accuracy for short utterances by extracting accent and linguistic attributes. Another investigation by Cavalcanti et al. (2024) explores the use of fundamental frequency estimates to differentiate speakers within identical twin pairs and across different pairs. This study, involving 20 Brazilian Portuguese native speakers (10 male identical twins aged 19 to 35), reveals significant distinguishing patterns in fundamental frequency. It suggests the feasibility of constructing a reliable system even for speakers with closely resembling voice prints. However, despite the advancements in speaker recognition, existing forensic analysis software, such as the Peritus framework (de O. Cunha et al., 2020), are mainly suited for analyzing videos and images. This highlights the continuous need for further advancements in this field.

Although automatic speech recognition is currently not usually deemed admissible in courtrooms, a recent study (Basu et al., 2022) suggests that this technology already surpasses the performance of non-expert listeners. A thought-provoking research conducted by Youn et al. (2021) suggested that smart devices might be important digital observers at crime scenes, highlighting the increasing importance of audio digital evidence and the growing demand for automatic speaker recognition technology. Nevertheless, the quality of forensic phonetic features in voice comparison is greatly impacted by real-world settings, as demonstrated by Guan (2022). As a result, this directly affects the reliability of using speaker recognition in real-life situations. Our approach aims to address the difficulties encountered in practical situations characterized by subpar phonetic attributes.

3. Methodology

In this section, we introduce a speaker identification approach tailored for criminal investigations, adept at addressing challenges posed by large volumes of audio recordings featuring unidentified speakers. In summary, our approach systematically clusters related audio recordings presumed to emanate from a common speaker. Subsequently, it assigns scores to each audio cluster for every person enlisted in an enrollment database. Designed for utilization by human investigators, the approach aims to yield a discerning quantity of matches. It employs a thresholding procedure to formulate a candidate list of matches. This list, complemented by supporting evidence, helps the investigator in formulating a conclusive determination. The approach is outlined into two pipelines, as detailed in what follows.

3.1. Cluster-scoring pipeline

The cluster-scoring pipeline is built upon a target database (e.g., a cell phone) and an enrollment database (e.g., voice collection of the prison system). The primary objective is to assign a score to each tuple consisting of an individual from the target database and an individual

from the enrollment database, based on the similarity of their vocal patterns. As illustrated in Fig. 1, the cluster-scoring pipeline comprises five fundamental components: (i) feature extraction, (ii) scoring, (iii) clusterization, (iv) rank-based adjustment, and (v) cluster scoring.

3.1.1. Feature extraction

Preceding the process of audio scoring, a crucial step involves transforming the audio waveform into speaker embeddings, presented as a vector representation. The time complexity of this task is $O(m+n)$, where m is the number of audio recordings in the target database and n is the number of individuals in the enrollment database. The feature extraction component uses the ECAPA-TDNN speaker verification model (Desplanques et al., 2020), which incorporates components to address speaker verification. The training dataset includes speakers from over six different nationalities, suggesting potential robustness of the model across diverse languages. While originally designed for deployment in biometric systems, speaker verification models, such as ECAPA-TDNN, exhibit a tendency to produce a considerable number of erroneous identifications when applied to compare audio samples against an extensive database of registered speakers. Consequently, the outcomes yielded by these algorithms may be rendered ineffectual, as the individuals of interest become obscured amidst the multitude of (false) positive matches. To address this limitation, additional components are necessary in devising a method that proves valuable in supporting criminal investigations.

3.1.2. Scoring

The primary scoring mechanism in our approach relies on cosine similarity of audio embeddings, as recommended in the literature (Desplanques et al., 2020). Cosine similarity assesses the alignment of two vectors, measuring the degree to which they point in the same direction. Specifically, the cosine of the angle between two vectors A and B is computed by dividing their dot product by the product of their magnitudes, as expressed in (1).

$$CosineSimilarity(A,B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{1}$$

In the evaluation of a target audio against the enrollment database, the method calculates the cosine similarity between the target audio and each individual enlisted in the enrollment database. Overall, this task is performed in $O(mn)$. By setting a threshold or ranking individuals according to this metric, potential matches can be inferred. Nonetheless, as previously discussed, the straightforward application of the method anticipates a considerable number of false positives, especially when dealing with large datasets.

3.1.3. Rank-based adjustment

The substantial volume of individuals in the enrollment database, coupled with the often inadequate audio quality typical of real cases, leads to a considerable number of false positives. Consequently, the voice patterns of numerous individuals in the enrollment database may exhibit similarity to that of the target audio. In our approach, we deem a higher level of similarity as significant only when compared to others. We propose the incorporation of a score adjustment function, as expressed in (2), which considers the actual score, the relative rank, and an α parameter to control the decay rate based on the rank.

$$AdjustedScore_{i,s}(score_{i,s}, rank_{i,s}, \alpha) = score_{i,s} \cdot \underbrace{\frac{\alpha}{rank_{i,s}+\alpha}}_{\text{Adjustment factor}} \tag{2}$$

The value $score_{i,s}$ denotes the cosine similarity (1) between a target audio i and the audio associated with the speaker s in the enrollment database. By computing all scores, one can ascertain the ranking of each speaker in the enrollment database concerning the target audio. Let $rank_{i,s}$ denote the rank of the speaker s in the enrollment database, taking as reference the target audio i. The rank values span from 0 to n-1, where n represents the total count of enrolled individuals. The parameter α has a range from 0 to infinity. Fig. 2 illustrates the adjustment score function for ranks spanning from 1 to 10, considering different values of α. It is noteworthy that with an increase in α, the adjustment factor tends to converge to one for all ranks. In contrast, with a decrease in α, the adjustment factor approaches zero for all ranks except the highest-ranked individual. The time complexity of this task is $O(mn)$.

3.1.4. Clusterization

The distortion, arising from straightforward comparisons, is attributed to the significant variability among audio samples from the same person in the target database and the limited variability among certain individuals in the enrollment database. The rationale behind incorporating clusterization is rooted in the idea that jointly analyzing a set of audio recordings from the same speaker can diminish speech fluctuations and mitigate the impact of noise, thereby enhancing the capture of vocal patterns. Furthermore, clusterization offers investigators an extra advantage, allowing for analyses on a per-cluster basis and thereby notably diminishing the overall scope of the task. In our approach, we employ the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2013) to achieve this functionality. This task is performed in $O(m^2)$.

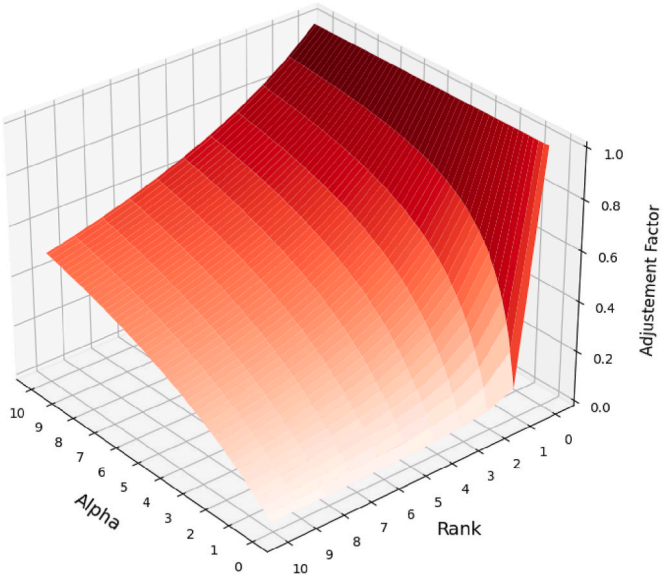


Fig. 2. Adjusted score for different rank and *alpha* values.

Fig. 1. Cluster-scoring pipeline.

3.1.5. Cluster scoring

In the conclusive phase of the cluster-scoring pipeline, the overall score between a cluster of audio recordings from the target database and an individual's audio enlisted in the enrollment database is determined. The incorporation of the score adjustment function, coupled with clusterization, leads to an outcome where the accumulation of scores for audio samples within a cluster yields a high cluster score exclusively for enrolled speakers consistently demonstrating elevated ranks and scores. To determine the score between a specific cluster c of audio recordings and a given speaker s, we propose calculating the mean of the adjusted scores, as defined in (3).

$$ClusterScore_{c,s}(c,s,\alpha) = \frac{1}{length(c)} \cdot \sum_{i \in c} AdjustedScore_{i,s}(score_{i,s}, rank_{i,s}, \alpha) \tag{3}$$

In essence, the method involves the summation of the adjusted scores, calculated with reference to a specific enrolled individual, for the audio recordings within a given cluster. Subsequently, the sum obtained is divided by the total number of audio recordings within the cluster. Alternatively, the cluster score can be computed by averaging the raw score without considering the adjustment function. Although the proposed approach does not incorporate such data aggregation, we will explore this alternative method in our empirical investigations. The time complexity of this task is $O(mn)$.

3.2. Speaker identification pipeline

The speaker identification pipeline, depicted in Fig. 3, is constructed upon the cluster-scoring pipeline. It employs a score-based threshold, both in relative and absolute terms, to pinpoint potential candidate individuals for further examination by a human investigator. This process is supplemented by external corroboration, facilitating an analysis before arriving at a final decision.

After computing scores for each audio cluster with respect to each individual in the enrollment database, the subsequent process is performed. For each target cluster, the method removes individuals from the candidate list if their scores in relation to the target cluster are below a specified threshold. Additionally, individuals with scores less than a certain percentage of the highest score associated with that cluster are excluded. This process leads to the compilation of potential candidates for each audio cluster. At times, audio recordings originating from the same individual may be subdivided into multiple clusters, introducing the potential for a single candidate to be identified as a potential match for multiple clusters. Conversely, there are instances where a cluster may be composed of audio recordings from distinct individuals. The final list of candidates will be consolidated with external corroboration, and it is the responsibility of the investigator to ascertain whether a candidate speaker is the source of the voice in the audio recording. We emphasize that the occurrence of individuals with similar vocal characteristics is not an uncommon phenomenon, and additional corroborating evidence remains essential.

4. Experiments

In this section, we employ the proposed approach on a synthesized dataset comprising voice data from various languages. The use of this dataset allows for the determination of appropriate parameters and an initial assessment of the approach. Following this, the approach is deployed on real-world data within the context of criminal investigation. To validate our findings, we conduct a comprehensive examination of external evidence, thereby reinforcing the robustness and reliability of our conclusions. To conduct our experiments, we used an Alienware M16 R1 machine equipped with 32 GB RAM, a 13th Gen Intel Core i9-13900HX processor, and an NVIDIA GeForce RTX 4070 with 8 GB VRAM, running Ubuntu 22.04.3 LTS. Feature extraction was carried out using the ECAPA-TDNN model implemented in the SpeechBrain library (Ravanelli et al., 2021), while clustering employed the HDBSCAN implementation of the cuML library (Raschka et al., 2020).

4.1. Common Voice experiments

The outlined speaker identification approach underwent an objective evaluation employing the Common Voice dataset (Ardila et al., 2019). To conform to the intended application's scale, we amalgamated data from four languages: Portuguese, French, Spanish, and Italian. For the purpose of monitoring and evaluating the approach's effectiveness, only utterances featuring a speaker identifier were included. Table 1 provides a comprehensive breakdown of the dataset composition.

In replicating a real-world scenario, we created an enrollment dataset by assigning a single utterance per speaker. Subsequently, we synthesized thirty cell phones using the remaining utterances, forming our target database. To prevent undue influence in analytical outcomes, the synthesis adhered to a constraint of 300 utterances per individual, crucial to address speakers with excessive contributions in the Common Voice dataset. The systematic allocation of speakers to cell phones, one by one, continued until the cumulative utterances surpassed 3000, ensuring a balanced distribution of speakers. Other than the speakers themselves, no additional features were considered when distributing the audio recordings.

4.1.1. Feature extraction

The initial stage in evaluating the speaker identification approach is to examine the feature extraction process employing the ECAPA-TDNN model. The resulting embeddings consist of 192-dimensional vectors, intended for subsequent procedures. A subset of 100,000 utterances was randomly selected, and leveraging data characterization, we created a visual representation of the embeddings using UMAP (McInnes et al., 2018), as depicted in Fig. 4.

The objective was to explore the gender and age distributions within the generated embeddings. Notably, gender-related differences in the utterances are apparent, whereas distinctions among different age groups appear less clear. In retrospect, determining a person's gender based on their voice is typically more straightforward than estimating

Table 1
Common Voice aggregated dataset.

Language	Utterances	Speakers	Mean duration (seconds)	Median Utterances/Speaker
Portuguese	145111	965	4.06	28
French	805495	4509	4.98	30
Spanish	1614602	5804	4.89	19
Italian	306182	1888	5.27	25

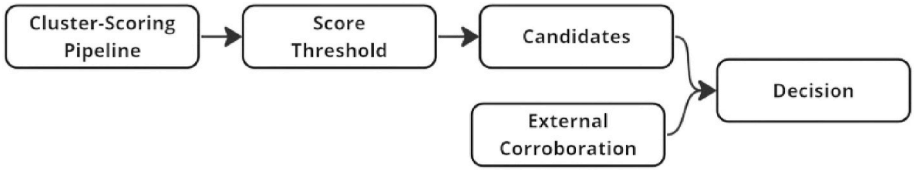
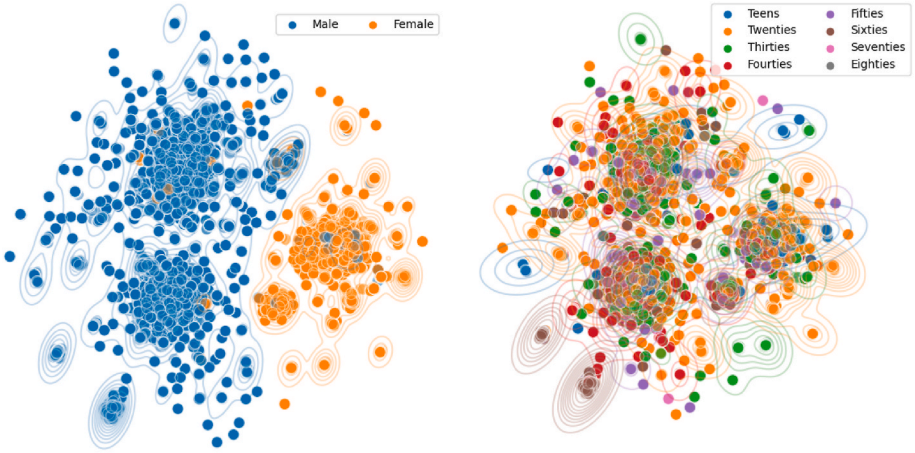


Fig. 3. Speaker identification pipeline.



(a) Gender (b) Age

Fig. 4. Distribution of audio embeddings classified by the (a) gender and (b) age of the speakers.

their age, aligning with common human perceptual patterns. As a result, the observed visualization pattern aligns with expectations.

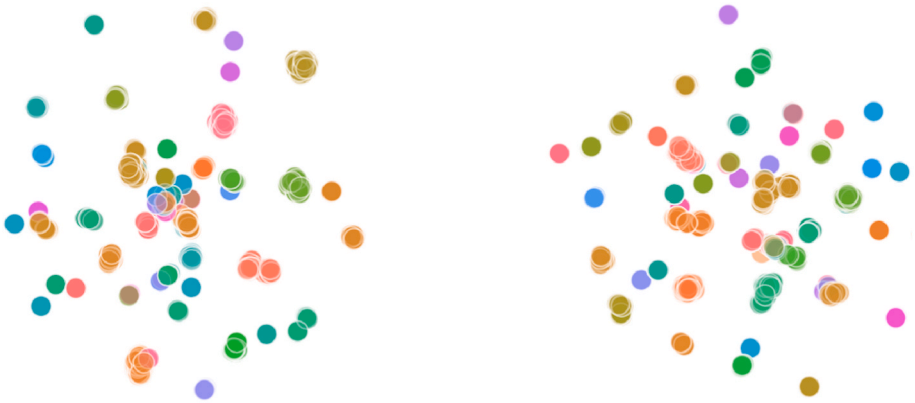
The plot depicted in Fig. 5 displays two of the synthesized cell phones. As an experimental step, the embeddings are mapped onto a two-dimensional surface using UMAP. In this representation, each utterance is assigned a color corresponding to its identified speaker. The ECAPA-TDNN architecture demonstrated significant proficiency in extracting distinctive feature crucial for speaker recognition. This claim is supported by the observation of numerous spatially distinct groups formed solely by utterances associated with a unique speaker within each group. The significance is heightened as the distribution involves a two-dimensional projection of 192-dimensional vectors.

4.1.2. *Clusterization*

The main goal of clusterization is to subdivide the target database into separate audio clusters, each linked to a specific speaker, making it easier to compare with the enrollment database later on. Optimally, in each cluster, the proportion of audio recordings ascribed to the primary speaker should approximate 100 %. Evaluation involves the Equal Error Rate (EER), which is a quantitative metric that measures error at the point in which the rates of false acceptance and false rejection are equivalent. Biometric systems demonstrate enhanced performance with lower EER values. Nevertheless, the utilization of the HDBSCAN clustering algorithm leads to a notable amount of audio recordings that are not assigned to any cluster, despite its effectiveness in reducing the EER. To tackle this issue, we investigated the arrangement of hyperparameters to attain a better balance between coverage (the proportion of audio recordings ascribed to any given cluster) and cluster predominance (the proportion of audio recordings inside a cluster that are attributed to its primary speaker). The balance between coverage and predominance is depicted in Fig. 6, where the minimum cluster size hyperparameter is adjusted during the execution of HDBSCAN.

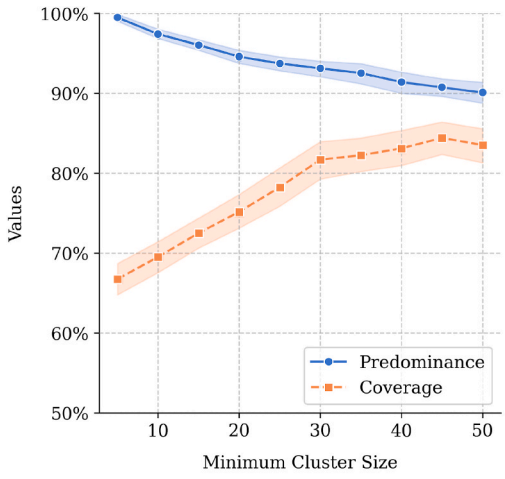
Exceeding a minimum cluster size of 30 results in diminishing returns on coverage. Nevertheless, cluster predominance remains notably high, rendering it an advantageous selection for this specific parameter. The evaluation of EER is also conducted for each minimum cluster size value, demonstrating a positive correlation, as depicted in Fig. 6. However, the marginal discrepancies in EER values lack sufficient justification to endorse the adoption of a minimum cluster size value below 30, potentially compromising the coverage criterion.

After establishing the minimum cluster size, we investigated the impact of clustering. In the baseline approach, each audio recording was treated as an independent cluster, with scoring based on cosine similarity between target database recordings and audio samples from the enrollment database. In the clustering approach, scores of clusters formed by the HDBSCAN algorithm for a specific enrolled individual were determined as the average of scores from audio recordings within each cluster. Audio recordings without a cluster were treated as a singular audio cluster. In both approaches, after determining matches, EERs were computed based on single audio recordings. However, in

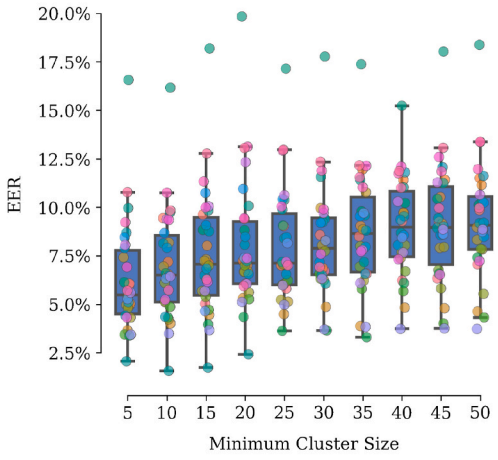


(a) Cell phone #1 (b) Cell phone #2

Fig. 5. Distribution of audio embeddings of (a) cell phone #1 and (b) cell phone #2 classified by the identification of the speakers.



(a) Predominance and coverage



(b) EER

Fig. 6. Minimum cluster size influence on (a) predominance, coverage, and (b) EER.

contrast to the baseline approach, which independently determines matches for each audio recording, the clustering approach provides matches for a cluster of audio recordings as a cohesive unit. The matches of the cluster are then replicated for each single audio recording within that cluster. Fig. 7 illustrates the boxplot showcasing results for 30 synthesized cell phones and their respective EERs under the alternative approaches. Notably, the sole integration of a clustering method does not show significant improvement. The plot also includes the results of clusterization combined with the proposed adjustment function, which will be further explored in the subsequent analysis.

4.1.3. *Rank-based adjustment*

The investigation entailed the assessment of the adjustment parameter α over a range of eight distinct values, ranging from 10^{-1} to 10^{3}. As detailed in the preceding section, a minimum cluster size of 30 was employed. The performance, illustrated in Fig. 8, demonstrates stability in the first five levels but exhibits a decline beyond an α value of 50. Based on these observations, subsequent experiments will adhere to an α value of 10. This specific value is deemed sufficient to accommodate minor variations in ranking while ensuring that the performance level has not undergone any decline thus far.

As depicted in Fig. 7, the strategy of combining clusterization with rank-based score adjustment resulted in a notable decrease in the EER.

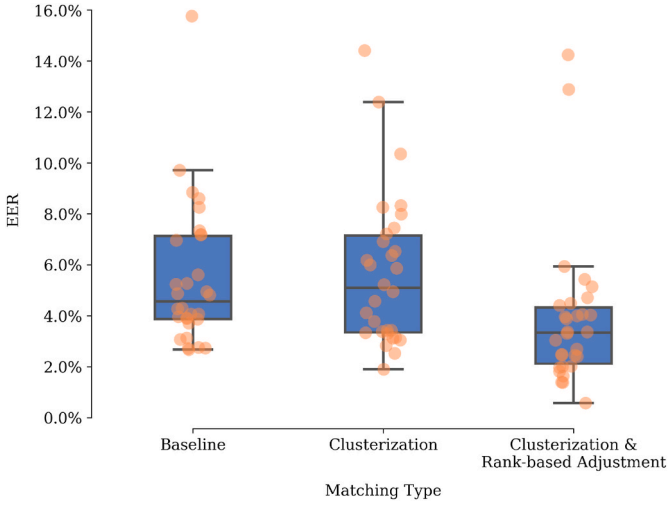


Fig. 7. Comparison of EERs across the different approaches.

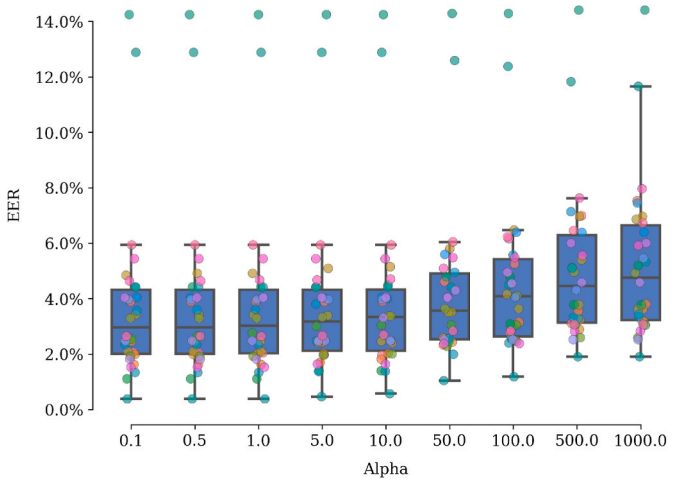


Fig. 8. The impact of α value on EER.

Despite the statistical analysis of the data in Table 2 failing to yield sufficient evidence to support the hypothesis that the sole utilization of a clustering method enhances the final outcome, a statistically significant improvement is observed when clusterization is coupled with rank-based score adjustment.

In the subsequent analysis, the approach is employed within an authentic context of criminal investigation.

4.2. *Applied experiments*

This research was supported by government institutions, namely the MPCE and the SAP-CE. Validation of our findings involves integrating external information with the audio data used in our approach. This process includes scrutinizing photographic and other documentary evidence related to specific individuals. The inclusion of external information, such as photographs, body markings, and names, contributes to the individual identification. However, it is important to note that such

Table 2
Statistical analysis of the different strategies.

	H0	H1	p-value
Experiment 1	$EER_{Baseline} = EER_{Clustering}$	$EER_{Baseline} > EER_{Clustering}$	0.615
Experiment 2	$EER_{Baseline} = EER_{Adjusted}$	$EER_{Baseline} > EER_{Adjusted}$	0.017

inclusion does not automatically serve as certification. We also acknowledge the ethical challenges posed by using speaker identification technology in criminal investigations, especially regarding privacy and potential bias. Our work is purely for research purposes. We collaborated with MPCE and SAP-CE, ensuring strict compliance with legal standards to safeguard individuals' rights and ensure accurate interpretation of results. We also recognize the importance of preventing potential misuse of this technology, underscoring the necessity for clear guidelines and ethical oversight in its application within the justice system.

4.2.1. Target database

The MPCE employs advanced software to extract audio data from electronic devices. This software additionally enables the retrieval of varied multimedia content, such as music, photographs, documents, and text messages, facilitating potential cross-referencing of findings. The provided target database from the MPCE includes information from 67 cell phones seized in prison units. Each cell phone is associated with a specific number of utterances and a count of individuals who have undergone external validation. The cumulative count of individuals subjected to external validation is 86 and the total number of utterances surpasses 400,000. Externally validated individuals are individuals who are connected to the specific seized cell phone through evidence other than audio. While this increases the likelihood of these individuals also having audio evidence in the seized cell phone, it does not provide a guarantee. Therefore, caution should be exercised when considering this number.

4.2.2. Enrolment database

The SAP-CE supplied the enrollment database, encompassing personal data, audio, and photo records of prisoners. Upon admission to the prison unit, each prisoner undergoes a registration procedure that encompasses recording their speech, along with collecting fingerprints, facial photographs, and images of body marks. Regrettably, the audio recordings exhibit imperfections such as secondary speakers, additional background noise, or even a lack of vocalization. Recognizing the impracticality of re-enrolling the entire prison system, it becomes necessary to acknowledge the limitations of this database. Nevertheless, the enrollment database has a significant amount of audio recordings that include verified speaker IDs, totaling 69,453 unique speakers.

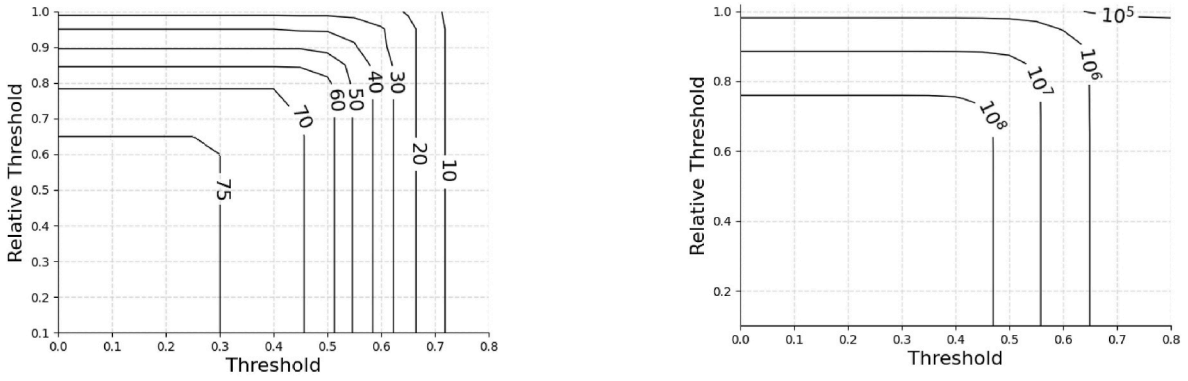
4.2.3. Evaluation

To enable comparative evaluation, the complete dataset was subjected to the baseline approach and our approach. As the number of potential candidates (matches) inferred by each approach varies depending on threshold settings, we conducted an analysis by manipulating two variables: the minimum score threshold and the threshold relative to the maximum observed score. Fig. 9 provides an overview of the results obtained using the baseline approach. Each curve represents the number of candidates for a combination of relative and absolute thresholds. The findings demonstrate a noteworthy number of identifications (75 out of 86 validated candidates) at specific absolute and relative threshold selections (e.g., 0.3 and 0.5 respectively). However, it is essential to recognize that an identification does not always ensure accuracy, as false positives may occur. Moreover, the feasibility of audio analysis is greatly impeded by the vast number of potential candidates, even when applying strict values like an absolute threshold of 0.8 and a relative threshold of 1.0. This emphasizes the need for a more sophisticated approach.

Our approach utilized a minimum cluster size of 30 and an α value of 10, as suggested by our preliminary analysis, resulting in the identification of 2211 clusters. In Fig. 10, it can be seen that our approach, despite employing tolerant limits, produces a limited number of identifications (30 out of 86 validated candidates). Consequently, the identification rate among the validated candidates is 34.9 %. However, it is essential to acknowledge a substantial decrease in the overall number of potential candidates, facilitating candidate verification by an investigator.

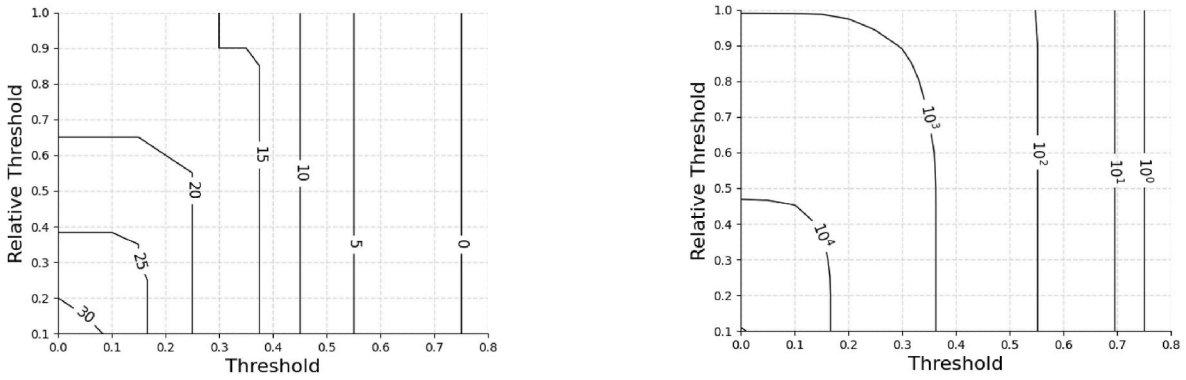
The disparity in the number of potential candidates can be partially attributed to the distinction in the size of the analyzed unit between the baseline approach (which considers single audio recordings) and our approach (which examines clusters of audio recordings). This variation inherently leads to a decrease in the overall pool of matches, as there is a substantial reduction in the total number of units subjected to analysis. To address this, evaluating the candidate quantity per unit in each approach is essential. Fig. 11 illustrates persistent discrepancies, with our approach yielding a significantly reduced list of candidates per analyzed unit compared to the baseline approach.

In an effort to provide a rough demonstration of the time-saving implications in an investigation, we will consider scenarios in which each approach successfully identifies 10 validated candidates. In this case, the baseline approach would yield over 300,000 potential candidates (i.e., matches between an audio recording and an individual from the enrollment database). In contrast, our approach would identify approximately 300 potential candidates (i.e., matches between an audio cluster and an individual from the enrollment database). Assuming that each match verification requires 1 min of an investigator's time, completing the task with the assistance of our method would require approximately 5 h of analysis. Conversely, employing the baseline method would demand over 200 uninterrupted days of analysis. Considering the execution time of our approach, the pipeline took less than 6 running hours, with the primary scoring task being the most time-consuming, taking around 2 h and 30 min to complete.



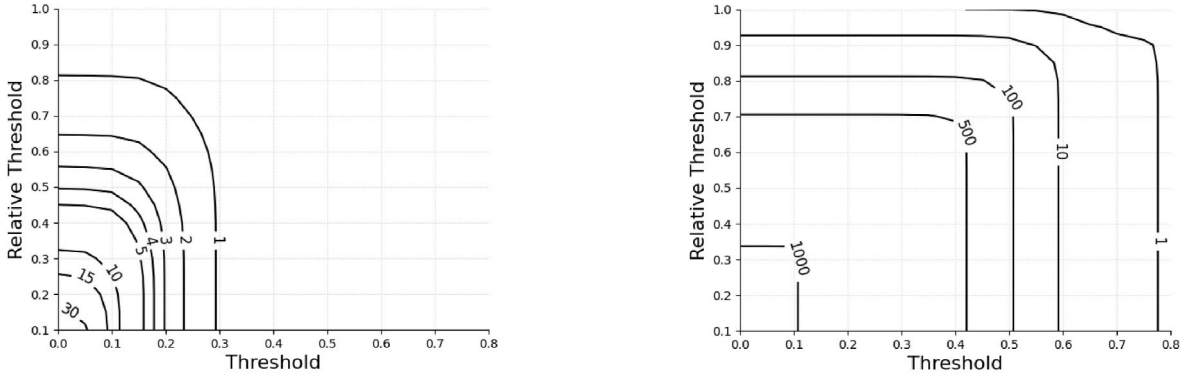
(a) Validated candidates (b) Potential candidates

Fig. 9. Examination of the count of (a) validated candidates and (b) potential candidates determined by the baseline approach under various threshold settings.



(a) Validated candidates (b) Potential candidates

Fig. 10. Examination of the count of (a) validated candidates and (b) potential candidates determined by our approach under various threshold settings.



(a) Our approach (b) Baseline approach

Fig. 11. Comparison of the number of potential candidates per analysed unit determined by (a) our cluster-based approach and (b) the audio-based baseline approach.

5. Conclusion

This section presents a comprehensive analysis of the research's findings, inferences, and potential avenues for future research. We employed experimental protocols to evaluate the proposed speaker identification approach, assessing its performance on the Common Voice dataset. Subsequent experiments confirmed the reliability and consistency of the approach when applied to empirical data from real investigations. Our findings demonstrate the efficacy of the proposed approach in clustering audio recordings and selecting potential candidates while maintaining a reasonable level of false positives.

The proposed speaker identification approach integrates clusterization and rank-based scoring strategies to form a practical system. It identifies clusters of audio recordings likely produced by a common speaker, suggesting potential candidates linked to each cluster. However, it's crucial to note that this approach doesn't offer unequivocal identifications and necessitates additional supporting information for substantial findings.

The research includes the development of a prototype to showcase the applicability of the proposed approach in real-world scenarios. The prototype covers various use cases, aiming to enhance investigators' capacity for insights extraction, helping in the exploration of audio data.

Two research questions were addressed.

- *Can state-of-the-art speaker recognition models be applied to languages not originally considered in the model?* This question is significant, given that the cutting-edge pre-existing model (Desplanques et al., 2020) was trained on languages distinct from Portuguese. In the context of synthesized cell phones, the model adeptly clusters speakers and captures peripheral characteristics, including gender and age. The results indicate satisfactory performance, even in the presence of language barriers.
- *Can a speaker recognition system support speaker identification in criminal investigations, addressing challenges associated with large volumes of audio recordings and handling enrollment databases with tens of thousands of individuals?* The second question focuses on reducing false positives in extensive comparisons. The baseline method generates an overwhelming number of candidates, posing a considerable burden for investigators. In contrast, our approach successfully reduced the candidate pool to a manageable number.

The responses to these research questions provide valuable insights for comprehending and improving speaker recognition technologies across a broader spectrum of linguistic contexts.

5.1. Limitations

The main limitation of our approach is its limited scope in the number of analyzed utterances. In the clustering process, certain utterances could not be assigned to any cluster, posing a challenge with low coverage. An alternative approach is to interpret unclustered audio recordings as a single utterance cluster, but this strategy would result in a loss of benefits by generating an excessive number of potential candidates.

5.2. Future work

Future advancements may involve utilizing speech-to-text models to extract textual content from audio recordings. One application involves automated verification of an individual's name within transcriptions, a scenario expected due to engaged dialogue with others possessing knowledge of their identity. Additionally, the application of topic modeling could aid investigators in focusing on relevant clusters, avoiding unnecessary time allocation to insignificant topics. Areas for improvement encompass integrating identified audio into the enrollment database for model training and addressing variations in data characteristics. Notably, dissimilarities in dataset collection may impact similarity score computation, given that embeddings are mapped onto distinct subspaces. An initial approach involving normalization of databases to establish a common mean and standard deviation showed promising outcomes. Nevertheless, further investigation and rigorous validation of this hypothesis are essential.

Acknowledgments

This work was supported by Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Funcap) [Grants 06401313/2020, 03063170/2023]. We acknowledge the Public Ministry of State of Ceará and the Penitentiary Administration Secretariat of the State of Ceará for the provision of data strictly designated for research purposes only.

Appendix A. Prototype

In this appendix, we present a prototype application demonstrating the practical implementation of the proposed approach for user interaction within the realm of audio investigations. This prototype serves as a showcase for user engagement and the operational effectiveness of the method in diverse investigative scenarios. The study outlines five specific utilization scenarios: open search, audio search, prisoner search, pairwise comparative analysis, and cell phone network analysis. Each use case is tailored to address unique investigative needs with distinctive characteristics and functionalities.

Open Search

The open search use case is designed for criminal investigators aiming to analyze audio data in a non-specific manner, and it executes the pipeline outlined in the paper. This scenario enables investigators to conduct a comprehensive search across diverse audio sources, thereby facilitating the exploration of potential connections. The interface specific to the open search use case is illustrated in Figure A12. It features a sidebar allowing users to adjust settings, including absolute and relative thresholds. Users can also select the target cell phone and cluster for analysis. After specifying the target cell phone, users can analyze audio recordings within a given cluster. The interface subsequently presents a compilation of potential candidates ranked according to computed scores relative to that cluster. The open search use case is a method for identifying valuable leads and potential associations between audio data and individuals, even in the absence of specific targets. Investigators can leverage this capability to unveil unforeseen connections and determine potential suspects, accomplices, or crime networks that might otherwise go undetected. Particularly valuable in situations with a lack of specific leads, this use case allows investigators to explore new avenues and draw evidence-based conclusions during their investigative endeavors.1

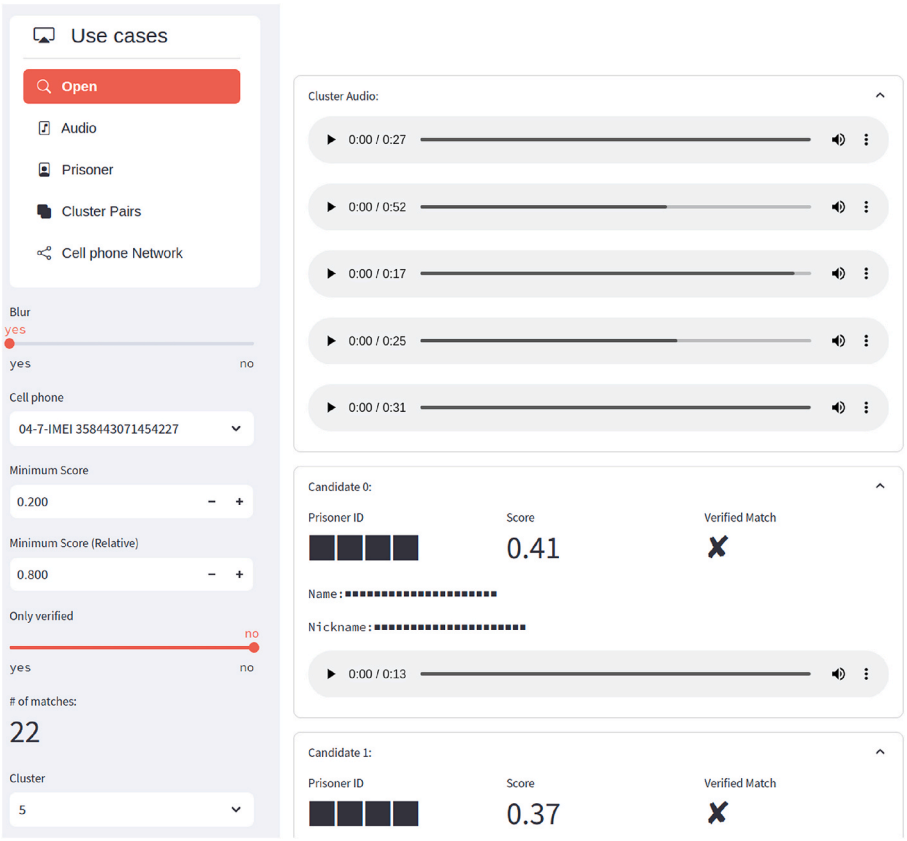


Fig. A.12. Open search interface.

Audio Search

The audio search use case refers to circumstances in which criminal investigators are interested in a specific audio sample and want to study similar audio within the same cluster in order to identify the speaker. This targeted approach allows investigators to concentrate on relevant evidence and potential matches, narrowing the scope of their investigation. Figure A13 illustrates the user interface for this use case, enabling users to select a particular audio file from an examined cell phone. The interface then presents other audio files categorized within the same cluster, supporting investigators in determining patterns, similarities, and correlations. Additionally, the prototype compiles potential candidates associated with the selected cluster. In cases where the chosen audio lacks a cluster association, the prototype employs a baseline method to generate results for that specific audio.11

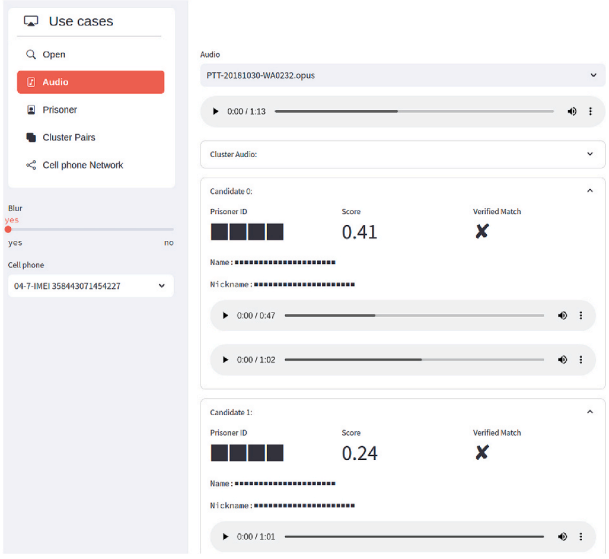


Fig. A.13. Audio search interface.

Prisoner Search

The prisoner search use case is a typical task done by criminal investigators while focusing on specific individuals. This case aims to thoroughly analyze seized cell phones to identify potential clusters associated with the individual under investigation, particularly someone currently or formerly incarcerated. The user interface, depicted in Figure A14, enables researchers to specify a particular prisoner and adjust thresholds associated with the approach, providing control over the leniency observed in search outcomes. The search results prominently present audio recordings of the individual sourced from the enrollment database, along with the clusters where recognition of this specific individual is likely. Conducting a prisoner search proves valuable in determining the involvement of incarcerated individuals in ongoing criminal activities. This process aids in resolving cold cases and acts as a preventive measure against potential criminal endeavors orchestrated from prison units.1

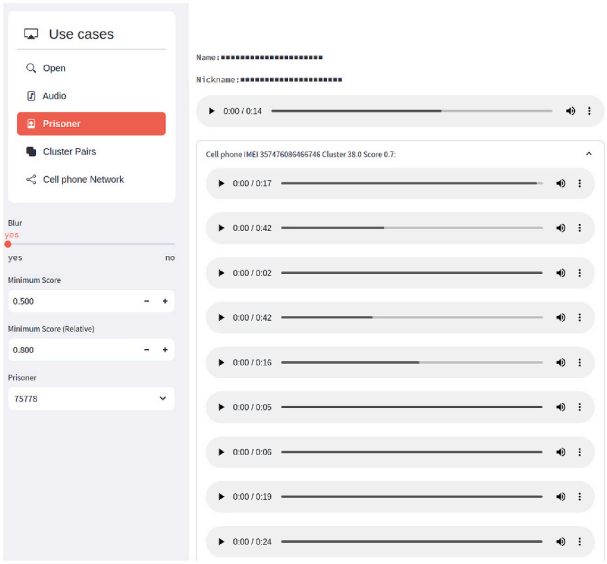


Fig. A.14. Prisioner search interface.

Pairwise Comparative Analysis

The pairwise comparative analysis use case assists criminal investigators in identifying audio clusters with significant similarity, indicative of a common speaker. Its objective is to conduct a comparative analysis of audio samples, potentially identifying the same speaker in two distinct electronic devices and creating correlations for effective case resolution. By verifying the common origin of multiple audio samples, investigators can establish connections between seemingly unrelated cases, unveil networks involved in organized criminal activities, and ascertain the identities of individuals engaged in serial offenses. The use case provides users with pairs of audio file clusters likely produced by the same person. The interface in Figure A15 allows users to traverse between pairs of clusters, evaluating similarity using the metric. The side-by-side format in presenting audio files facilitates understanding of similarities and differences, empowering investigators to make informed decisions and conduct thorough forensic analysis.1

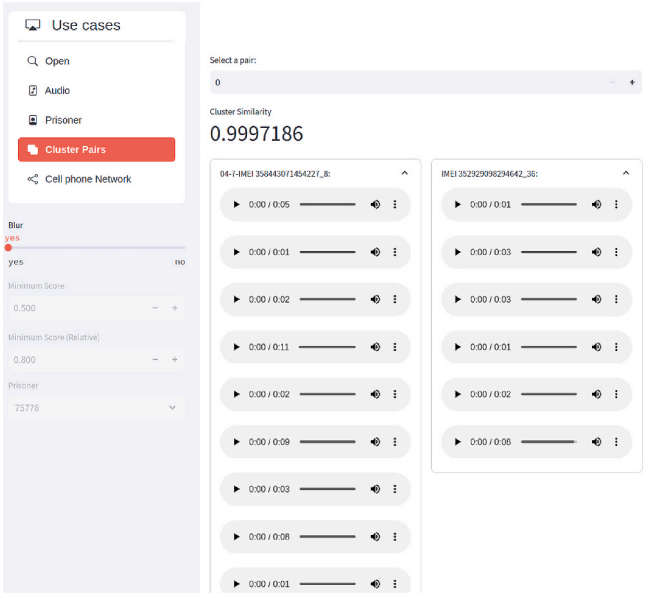


Fig. A.15. Pairwise comparative analysis interface.

Cell Phone Network Analysis

The cell phone network analysis use case focuses on examining associations among different cell phones, relying on audio similarity. The user interface in Figure A16 allows criminal investigators to set a threshold value for finding clusters with a common speaker. This feature assists in detecting interconnected cell phones within the network, generating a graph representation of connections with a circular arrangement. Each vertex represents a labeled cell phone, and edges indicate weights signifying similarity matches exceeding a predetermined threshold. Analyzing the cell phone network graph provides insights into connections and associations based on audio similarity. This visualization technique aids in detecting patterns, clusters, or connections within the network, offering valuable insights for further examination. The use case supports investigators in delineating criminal networks, identifying key individuals, and revealing concealed communication channels. Visualizing cell phone networks and analyzing audio similarities can potentially provide investigators with significant intelligence and crucial evidence, facilitating the successful apprehension of criminals and the dismantling of criminal organizations.

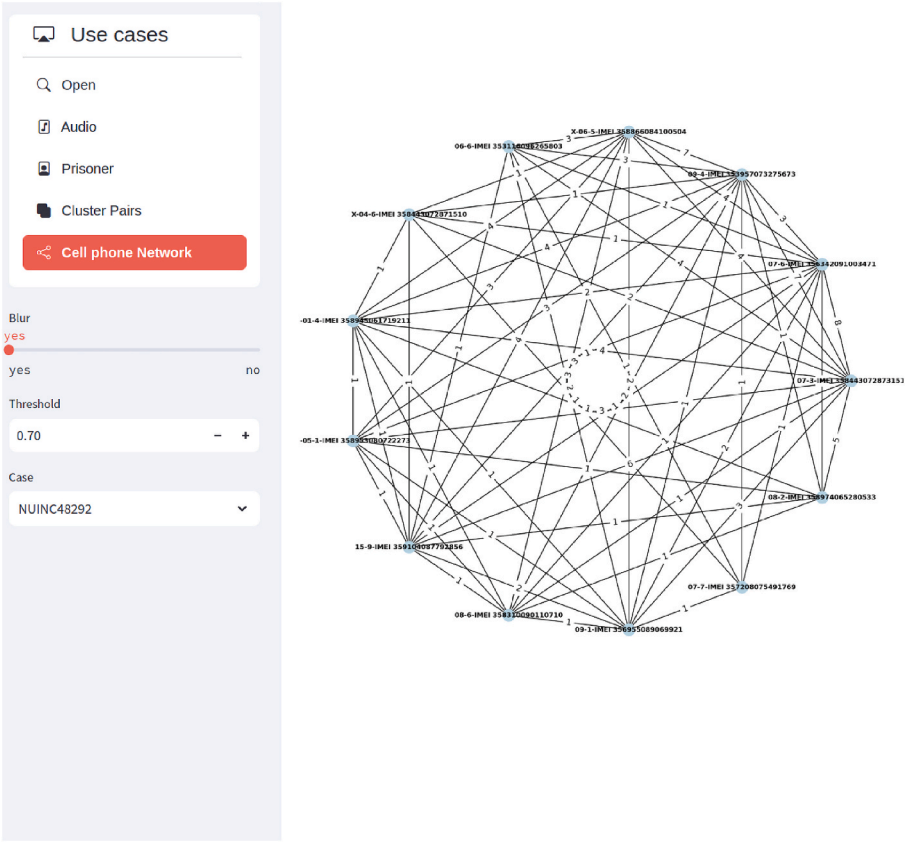


Fig. A.16. Cell phone network analysis interface.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2019. Common Voice: A Massively-Multilingual Speech Corpus arXiv preprint arXiv:1912.06670.

Basu, N., Bali, A.S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K.A., Morrison, G. S., 2022. Speaker identification in courtroom contexts – part i: individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. Forensic Sci. Int. 341, 111499 https://doi.org/10.1016/j.forsciint.2022.111499.

Bricker, P.D., Pruzansky, S., 1976. Speaker recognition. In: Lass, N.J. (Ed.), Contemporary Issues in Experimental Phonetics. Academic Press, pp. 295–326. https://doi.org/10.1016/B978-0-12-437150-7.50015-4.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172.

Cavalcanti, J.C., Eriksson, A., Barbosa, P.A., 2024. Multiparametric analysis of speaking fundamental frequency in genetically related speakers using different speech materials: Some forensic implications. J. Voice 38 (243), e11–e243.e29. https://doi.org/10.1016/j.jvoice.2021.08.013.

Chen, Y., Lopez-Moreno, I., Sainath, T.N., Visontai, M., Alvarez, R., Parada, C., 2015. Locally-connected and convolutional neural networks for small footprint speaker recognition. In: Proc. Interspeech 2015, pp. 1136–1140. https://doi.org/10.21437/Interspeech.2015-297.

Cumani, S., Plchot, O., Laface, P., 2013. Probabilistic linear discriminant analysis of i-vector posterior distributions. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 7644–7648.

Cunha, Daniel de O., Silva, Edmar A., Lambert, Jorge de A., Ribeiro, Rafael O., 2020. Peritus framework: towards multimedia evidence analysis uniformization in brazilian distributed forensic model. Forensic Sci. Int.: Digit. Invest. 35 (2666–2817), 301089 https://doi.org/10.1016/j.fsidi.2020.301089. https://www.sciencedirect.com/science/article/pii/S2666281720303917.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19, 788–798.

Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Proc. Interspeech 2020, 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650.

Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. Proc. Interspeech 2011, 249–252. https://doi.org/10.21437/Interspeech.2011-53.

Guan, X., 2022. An empirical study of the effects of pure real-world conditions on the reliability of forensic phonetic features. International Journal of Forensic Sciences. https://doi.org/10.23880/ijfsc-16000267.

Hansen, J.H., Bořil, H., 2018. On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. Speech Commun. 101, 94–108. https://doi.org/10.1016/j.specom.2018.05.004.

Holmes, J.N., 1985. Speech and speaker recognition. J. Phonetics 13, 359–362. https://doi.org/10.1016/S0095-4470(19)30766-1.

Jahangir, R., Teh, Y.W., Nweke, H.F., Mujtaba, G., Al-Garadi, M.A., Ali, I., 2021. Speaker identification through artificial intelligence techniques: a comprehensive review and research challenges. Expert Syst. Appl. 171, 114591 https://doi.org/10.1016/j.eswa.2021.114591.

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M., 2011. I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association. International Speech Communication Association, pp. 2341–2344.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., Zhu, Z., 2017. Deep Speaker: an End-To-End Neural Speaker Embedding System arXiv preprint arXiv:1705.02304 650.

Matějka, P., Glembek, O., Castaldo, F., Alam, M.J., Plchot, O., Kenny, P., Burget, L., Černocky, J., 2011. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4828–4831.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction arXiv preprint arXiv:1802.03426.

Peacocke, R.D., Graf, D.H., 1995. An introduction to speech and speaker recognition. In: BAECKER, R.M., GRUDIN, J., BUXTON, W.A., GREENBERG, S. (Eds.), Readings in Human–Computer Interaction. Morgan Kaufmann. Interactive Technologies, pp. 546–553. https://doi.org/10.1016/B978-0-08-051574-8.50057-1.

Prince, S.J., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE, pp. 1–8.

Quick, D., Choo, K.K.R., 2014. Impacts of increasing volume of digital forensic data: a survey and future research challenges. Digit. Invest. 11, 273–294. https://doi.org/10.1016/j.diin.2014.09.002.

Quick, D., Choo, K.K.R., 2018. Digital forensic intelligence: data subsets and open source intelligence (dfint+osint): a timely and cohesive mix. Future Generat. Comput. Syst. 78, 558–567. https://doi.org/10.1016/j.future.2016.12.032.

Raschka, S., Patterson, J., Nolet, C., 2020. Machine Learning in python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence arXiv preprint arXiv:2002.04803.

Ravanelli, M., Bengio, Y., 2018. Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 1021–1028.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y., 2021. SpeechBrain: A General-Purpose Speech Toolkit arXiv:2106.04624. arXiv:2106.04624.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10, 19–41.

Sadjadi, S.O., Ganapathy, S., Pelecanos, J.W., 2016. The Ibm 2016 Speaker Recognition System arXiv preprint arXiv:1602.07291.

Saleem, S., Subhan, F., Naseer, N., Bais, A., Imtiaz, A., 2020. Forensic speaker recognition: a new method based on extracting accent and language information from short utterances. Forensic Sci. Int.: Digit. Invest. 34, 300982 https://doi.org/10.1016/j.fsidi.2020.300982.

Sharma, R., Govind, D., Mishra, J., Dubey, A.K., Deepak, K.T., Prasanna, S.R.M., 2024. Milestones in speaker recognition. Artif. Intell. Rev. 57, 58. https://doi.org/10.1007/s10462-023-10688-w.

Travadi, R., Segbroeck, M.V., Narayanan, S.S., 2014. Modified-prior i-vector estimation for language identification of short duration utterances. Proc. Interspeech 2014, 3037–3041. https://doi.org/10.21437/Interspeech.2014-609.

Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4052–4056.

Youn, M.A., Lim, Y., Seo, K., Chung, H., Lee, S., 2021. Forensic analysis for ai speaker with display echo show 2nd generation as a case study. Forensic Sci. Int.: Digit. Invest. 38, 301130 https://doi.org/10.1016/j.fsidi.2021.301130.

Zhang, S.X., Chen, Z., Zhao, Y., Li, J., Gong, Y., 2016. End-to-end attention based text-dependent speaker verification. In: 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 171–178.