ROXSD: a Simulated Dataset of Communication in Organized Crime

Kvetoslav Maly[1], *Gerhard Backfried*[2], *Francesco Calderoni*[3], *Jan "Honza" Černocký*[4], *Erinç Dikici*[2], *Maël Fabien*[5,6], *Jan Hořínek*[7], *Joshua Hughes*[8], *Miroslav Janošik*[2], *Marek Kovac*[1], *Petr Motlicek*[5], *Hoang H. Nguyen*[9], *Shantipriya Parida*[5], *Johan Rohdin*[4], *Miroslav Skácel*[4], *Sergej Zerr*[10], *Dietrich Klakow*[11], *Dawei Zhu*[11], *Aravind Krishnan*[11]

[1]Phonexia s.r.o, Brno, Czech Republic
[2]HENSOLDT Analytics, Vienna, Austria
[3]Università Cattolica del Sacro Cuore and Transcrime, Milano, Italy
[4]Brno University of Technology, Brno, Czech Republic
[5]Idiap Research Institute, Martigny, Switzerland
[6]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[7]Police of the Czech Republic, National Drug Headquarters, Czech Republic
[8]Trilateral Research, London, UK
[9]Leibniz Universität Hannover, Hannover, Germany
[10]Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany
[11]Saarland University, Saarland Informatics Campus, Germany

kvetoslav.maly@phonexia.com

Abstract

Criminal investigations contain sensitive and confidential material and are nonpublic by nature. Access to investigation data is very limited and restricted to only selected groups of individuals. Even for research purposes, data typically cannot be accessed freely. Within criminal investigations, data is still processed manually to a large extent. Solutions provided for automation of this processing — or even of individual processing steps — can be assumed to have a significant impact on the work of Law Enforcement Agencies (LEAs). Automation may effectively be key to handle large and complex amounts of data in an efficient manner under the typical operating conditions of LEAs.

This paper introduces the ROXANNE Simulated Dataset (ROXSD), a dataset with unique properties prepared by the ROXANNE Project[1] with assistance from several LEAs, to facilitate the development and evaluation of novel tools and technologies for criminal investigations. ROXSD consists of a set of simulated intercepted telephone conversations in a variety of languages. The story follows a realistic setting and includes the conditions and constraints of a real investigation. The network topology corresponding to the conversations was created by partner LEAs to reflect various typical organized crime groups. Conversations have been transcribed carefully and annotated in the original language and in English. The dataset is expected to provide a sound basis for further research and is available to download for researchers under signed agreement.

Index Terms: criminal investigation, dataset, speaker identification, automatic speech recognition

1. Introduction

Investigations of organized crime cases are resource and labor intensive, since telephone conversations intercepted between wire-tapped suspects need to be appropriately processed, transcribed, potentially translated, and finally mapped into a knowledge graph. Few commercial tools and research projects tackle the process of automating such investigations. Although most of the individual technologies involved in this process (speaker identification, automatic speech recognition, named entity detection, etc.) are now mature enough to achieve good performance, the lack of relevant datasets to evaluate their accuracy in a criminal case remains a major obstacle in the further development and evaluation of tools by research communities.

Further, the legal framework for regulating how law enforcement agencies can process personal data from closed cases in research activities is still very unclear and differs in each country depending on how the GDPR and Law Enforcement Directive are implemented;[2] until this is resolved, alternative data sources not from real cases need to be made available to enable lawful research including highly-realistic synthetic datasets as in the present case.

In an effort to foster and boost the work on the support for automation of such investigations and to allow further contributions, we have created ROXSD, a simulated dataset representing an actual organized crime case in a realistic manner. It aims to meet the constraints of a real scenario and to realistically represent it at an appropriate level of complexity, while also remaining freely accessible for research.

Section 2 introduces existing datasets that satisfy part of the constraints of a real-case scenario, while Section 3 presents the ROXSD data, its characteristics, the scenario, and the data collection process. Finally, we discuss future directions to take.

[1]"Real-time network, text, and speaker analytics for combating organized crime" project has received funding from the European Union's Horizon 2020 Work Programme under grant agreement n°833635, 2019-2022)

[2]Stergios Aidinlis, David Barnard-Wills, Leanne Cochrane, Krzysztof Garstka, Agata Gurzawska, Joshua Hughes, 'Between GDPR and Law Enforcement Directive – The Legal Use of Real Law Enforcement Authority Data in Security Research' [Forthcoming]

2. Existing datasets

In an attempt to produce a realistic dataset which contains information regarding criminal activities, Gao et al. [1] matched the Enron e-mail database [2] with the Enron telephone call database. However, for privacy issues, most fraudulent conversations had to be removed from the Enron database. The resulting structure of the network does therefore not reflect all available activities.

Other notable databases which aim to provide a realistic scenario include the Ahumada III dataset in Spanish by Ramos et al. [3] which represents a public real casework and the NFI-FRIDA dataset in Dutch by Van Der Vloed et al. [4]. A precursor to the latter, NFI-FRITS [5], offers multilingual data of 600 speakers. However, the authors simulate forensic data by only selecting a subset of 10% of the original data. Since this dataset does not follow a realistic scenario, it remains doubtful whether the construction and analysis of networks might be possible or sensible. Furthermore, the dataset has not been released publicly. Similar conclusions are reached for the Forensic Voice Comparison Database [6].

In previous work [7], we used the Crime Scene Investigation (CSI) TV series as a candidate for mimicking criminal investigation data, as suggested in [8]. In CSI the number of characters, the sub-groups in the criminal network as well as the topics of the conversations match real-world conditions, and transcripts with time stamps are available for a number of episodes. However, this dataset also comes with limitations: the focus is set on the investigation team, the conversations are in English only, and they are not recorded over the telephone.

3. ROXANNE simulated dataset

The ROXANNE Simulated Dataset, ROXSD in short, aims to depict a realistic criminal investigation case while bringing together the strengths of existing databases by being multilingual, multimodal, and openly accessible for research purposes. In the following subsections, we present the steps taken in the preparation of this dataset.

3.1. Requirements

Real-world cross-border organized crime investigations frequently involve multimodal data. Investigators typically build the following knowledge around a case, answering the fundamental W's:

- who is involved: who spoke, to whom, what are their names, or phone numbers, what is each speaker's gender or age, what languages (dialects, accents) were spoken
- what was said: what was the topic of the conversation, who was mentioned, what places, dates, etc. were mentioned
- other information: which communication channel was used, when was the call made, where was the call made

The collected dataset should capture sufficient information in order to be able to answer these questions. Intercepted calls should also respect some specific constraints, including:

- be recorded from a telephone conversation
- have some realistic background noise
- be sampled at the appropriate bandwidth (8kHz)
- contain multi-lingual speech

The core of the ROXSD data is formed by a set of intercepted telephone calls taking into account these constraints. Two types of metadata files are provided along with the dataset. The police metadata contains a list of initial suspects together with their names and telephone numbers. In contrast, the research metadata contains a much broader set of information, such as the phone numbers, the age, gender, and native language of the speakers, their real names, the date/time and the language of the conversation, the transcript of the conversation in the original language, its translation into English where applicable, as well as a high-level topic label of the call content.

3.2. Scenario

The scenario involves the Prague anti-drug unit of the Czech police investigating three hypothetical cases at the same time: a first drug distribution case involving Czech and Russian students, named DDA (Drug Distribution A), a drug lab, ran by Vietnamese suspects, named DLA (Drug Lab A), and another drug distribution case which involves individuals speaking German, named DDB (Drug Distribution B).

In the DDA case, a university student in Prague, Kryštof (C01M), is suspected of selling drugs. The police have wiretapped two of his mobile phones. Kryštof's movements are mostly within Prague, with some occasional travels to Brno. The wiretaps have shown that Kryštof is in contact with other individuals who are either users or distributors of drugs. Communication is in Czech or Slovak. Most of the communication occurs at the point where the drugs change hands. The police also starts a wiretap on the mobile phones of his contacts Kristýna (C07F) and Horký (C04M). Kryštof often calls Sergej (R01M) speaking English. They have agreed on some larger transactions. The police consequently also wiretaps Sergej's telephone. In a call between Sergej and Oleg (R05M) in Russian, the tap catches information about the contact between Oleg and Kryštof. The police wiretaps the telephone of Oleg, and one conversation mentions a delivery of drugs to London.

In the DLA case, the police suspects that two Vietnamese individuals, Tuấn (V01M) and Hoàng (V02M) are dealing in large quantities of drugs and that Hoàng may operate a production site. Their telephones are wiretapped. They mostly speak Vietnamese and call each other several times, as well as calling a few other Vietnamese contacts. However, Tuấn also enters into contact with an unknown individual. Speaking English, they discuss a large delivery of drugs.

Finally, in the DDB case, the police investigates Max (G01M), an Austrian student at Charles University suspected of the distribution of drugs in the city center of Prague. Max's telephone is wiretapped. He is in contact with several unknown individuals, with whom he speaks German and English.

Please note that the current legislation and regulations prevent comparing voice-prints across cases, therefore, this scenario should be considered as R&D-only. However, we hope that the ROXSD data and experiments performed thereon might support adequate changes in the legislation.

3.3. Data Collection

Following this realistic scenario, ROXANNE Consortium partners recorded a total of 236 telephone calls on Twilio[3], as illustrated in Figure 1. The mechanism replicates the interception of telephone conversations put in place by investigators for a suspect's telephone numbers.

[3] https://www.twilio.com/

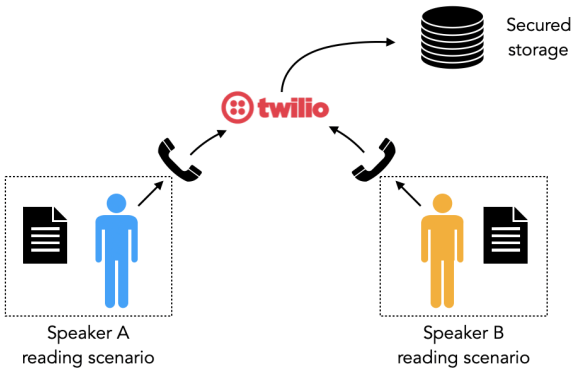

Figure 1: *Telephone call recording process.*

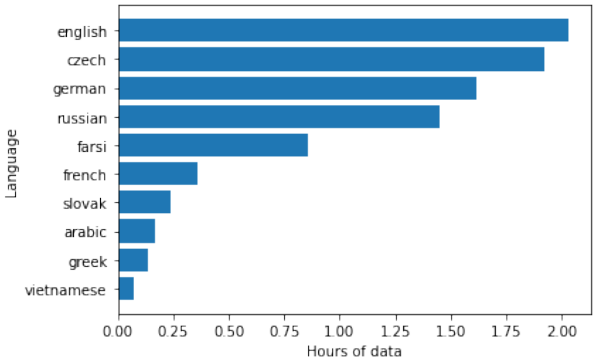

Figure 2: *Hours of audio in ROXSD per language.*

A total of 8.9 hours of conversations, including 6.6 hours of speech, were recorded. The duration of individual recordings ranges from a few seconds to several minutes, with an average length of 2m 25sec. Fifty speakers (36 males, 14 females, age 18-70 years) took part in the data collection assuming the roles of the involved characters, speaking ten different languages: English, French, Vietnamese, Arabic, Czech, Farsi, German, Greek, Russian and Slovak. The presence of multilingual speech, which is typical for organized crimes cases, makes our dataset unique and poses an additional challenge for speaker identification (SID) systems in performance degradation, as discussed in [9, 10]. The distribution of audio duration across languages is presented in Figure 2.

The calls in ROXSD are annotated as target or non-target calls. Approximately half of the conversations are target calls and involve criminal activities, with speakers following a prewritten scenario. Non-target calls, on the other hand, are longer side conversations captured between a suspect and some of their friends or family members. Guided by a high-level topic, they are improvised and constitute 83% of the overall conversation time. Transcripts are provided in the original language for all target calls, as well as for all English non-target calls. 24 speakers are only involved in non-target calls.

To realistically match a real-life scenario, the script was prepared jointly with LEAs. Each speaker could use several telephones (up to 3), and each telephone could be used by various speakers (at most 5), which justifies the need for speaker identification systems. Each conversation was recorded over two channels, leaving the choice for researchers to perform speaker diarization on mono-channel data or not.

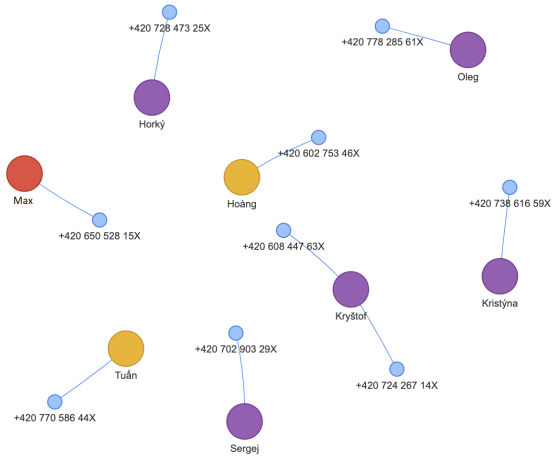

Figure 3: *Network representation of the initial wire-tapped numbers, with the 3 cases.*

3.4. Criminal network representation and analysis

Investigators suspect that two or more cases might be connected. But the links, if present, are like needles in a haystack. Investigators use SID to match the speakers across the cases. Eight speakers (the suspects) are initially wire-tapped, and all the incoming and outgoing telephone calls of their telephone numbers are intercepted. Figure 3 displays the initial knowledge of the investigators, which simply includes the eight characters, their attribution to each case (represented by the color of the node), and the fact that one of the characters, Kryštof, has two phone numbers.

If investigators do not leverage SID during this investigation, and only collect the links between phone numbers, the three cases would appear as completely disconnected components, as illustrated by Figure 4. The cases of the Vietnamese drug lab and the Austrian (German-speaking) drug distribution appear to be linked by a given phone number. We explicitly display known suspects with a green arrow.

Using information provided by SID, the network associates each unknown telephone number with a given speaker. Figure 5 displays the expected output network, built from metadata prepared for ROXSD. One character stands out as the central hub and builds a bridge between all cases. Finding this type of information without an automated system would require several days of manual work and collaboration between several interpreters. This setup provides a large potential for human error when faced with a pool of 50 speakers.

The content of the conversation can further be analyzed by Natural Language Processing (NLP) technologies. Often, named entities (person names, locations) are mentioned during telephone conversations. Such entities can form new nodes and edges in the graph and help find new associations between different actors and cases.

3.5. Research Ethics

All participants involved in the data collection for ROXSD were volunteers recruited from the ROXANNE partner organizations, including students and colleagues. All participants consented to the use of their data in the ROXANNE Project for the writing

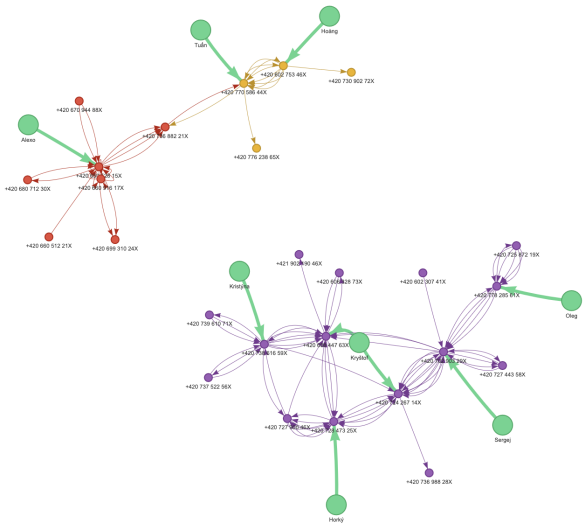

Figure 4: *Network representation of the simulated data, with the three cases.*

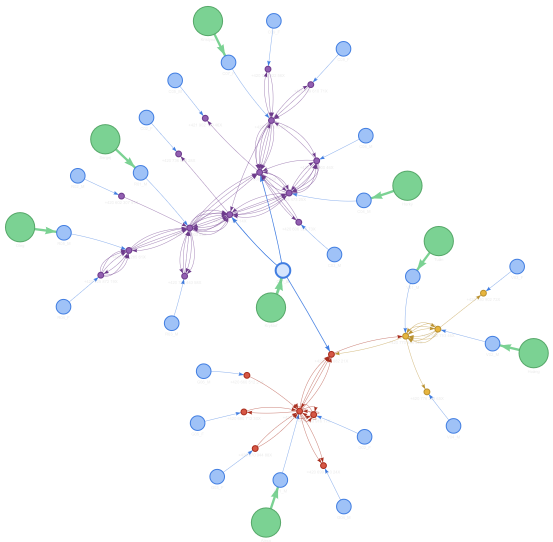

Figure 5: *Connection between the various cases after speaker identification, green nodes are wire-tapped numbers, blue nodes are identified speakers, and color nodes correspond to the different telephone numbers in the 3 cases.*

of scientific publications and for sharing with other researchers after the project's end. This involvement of human participants was approved by the research ethics committee of ROXANNE partners. With respect to data protection, all personal data was processed in compliance with the GDPR and respective national laws. The personal data of participants are included in the dataset within the speech recordings (i.e. their voice). However, most of the metadata relates only to the fictional characters portrayed in the scenarios.

3.6. Potential Use

Due to its rich set of data and associated meta-data, ROXSD lends itself to a multitude of applications and can be employed for a variety of evaluation (and development) purposes. This concerns a series of technologies from the field of NLP as well as from the field of network analysis. Regarding the NLP technologies, both audio-, and text-based aspects, can be addressed. Regarding network analysis, the previous technologies can be used to establish/hypothesize relations (and their evolution) between actors, locations, and conversational contents (like the mention of locations for handing over of substances, etc.). As such, they complement structural analysis from connection data and allow for a more holistic approach.

- Speech Processing: Speaker Clustering and Identification, Age and Gender Identification, Language and Dialect identification, Automatic Speech Recognition, Keyword-Spotting
- NLP: Named Entity Recognition (location, organizations, etc.), Topic Detection, Identification of unknown 2nd party in calls, Detection of mentions of 3rd parties
- Network Analysis: Network comparison metrics, Knowledge graph representation, Dynamic link prediction

3.7. Availability

ROXSD will be available to researchers and LEAs for research and non-commercial use with a signed agreement. The link to download the dataset will be provided soon on Zenodo, with a detailed access request and terms of use provided. Furthermore, we plan to make ROXSD available as a resource via the European Language Grid[4] (H2020 ELG funded by the European Union under grant agreement № 825627).

4. Discussion and Future Work

In this paper, we presented ROXSD, a simulated dataset representing communication within organized crime cases. It aims to be as realistic as possible and satisfies the constraints of a real criminal case. We presented our methodology for the data collection, the scenario, as well as the key features of the collected data. The current dataset already lends itself to a variety of evaluation and development tasks, both from an NLP as well as from a network analysis perspective.

ROXSD will be further expanded with additional data in the next releases. This will include short video calls between suspects, text messages, and images. The variety of collection modalities will help to enrich the cases further and make them as realistic as possible, while also providing even more challenging tasks to solve.

We hope that the initial release of the dataset will help stimulate work on the automation of criminal investigations and standardize the metrics considered for these tasks. There is important work to be done to enable lawful re-use of criminal case data in research projects that does not infringe on the rights of (suspected) criminals as data-subjects, and also their rights to privacy, fair trial, and non-discrimination, for example; additionally, with respect to reseach on criminal networks, freedoms of assembly and association also need to be considered. We will provide benchmark approaches and develop novel methods to build more reliable knowledge extraction systems. During the remainder of the ROXANNE project, several updates and extensions to ROXSD are planned, all of which will be released.

[4] https://www.european-language-grid.eu/

5. Acknowledgment

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

6. References

[1] N. Gao, G. Sell, D. W. Oard, and M. Dredze, "Leveraging side information for speaker identification with the Enron conversational telephone speech collection," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Okinawa, Japan: IEEE, Dec. 2017, pp. 577–583. [Online]. Available: http://ieeexplore.ieee.org/document/8268988/

[2] B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset for Email Classification Research," in *Machine Learning: ECML 2004*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3201, pp. 217–226. [Online]. Available: http://link.springer.com/10.1007/978-3-540-30115-8-22

[3] D. Ramos, J. González-Rodríguez, J. González Domínguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with Ahumada III: A public real-casework database in Spanish," in *Ninth Annual Conference of the International Speech Communication Association*, Sep. 2008, accepted: 2015-01-29T18:58:55Z Publisher: International Speech Communication Association. [Online]. Available: https://repositorio.uam.es/handle/10486/663472

[4] D. van der Vloed, J. Bouten, F. Kelly, and A. Alexander, "NFI-FRIDA–Forensically Realistic Inter-Device Audio database and initial experiments," in *27th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, 2018, pp. 25–27.

[5] D. v. Vloed, J. Bouten, and D. A. van Leeuwen, "NFI-FRITS: A forensic speaker recognition database and some first experiments," 2014.

[6] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155–167, Jun. 2012, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00450618.2011.630412. [Online]. Available: https://doi.org/10.1080/00450618.2011.630412

[7] M. Fabien, S. S. Sarfjoo, P. Motlicek, and S. Madikeri, "Improving Speaker Identification using Network Knowledge in Criminal Conversational Data," *arXiv:2006.02093 [cs, eess]*, Jun. 2020. [Online]. Available: http://arxiv.org/abs/2006.02093

[8] L. Frermann, S. B. Cohen, and M. Lapata, "Whodunnit? Crime Drama as a Case for Natural Language Understanding," *arXiv:1710.11601 [cs]*, Oct. 2017, arXiv: 1710.11601. [Online]. Available: http://arxiv.org/abs/1710.11601

[9] A. Misra and J. H. L. Hansen, "Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS Bi-Ling corpora," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2014, pp. 372–377.

[10] B. Ma and H. Meng, "English-Chinese bilingual text-independent speaker verification," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, May 2004, pp. V–293, iSSN: 1520-6149.