D5.3 FINAL SET OF ROXANNE SPEECH/NLP/VIDEO TECHNOLOGIES FOR NETWORK ANALYSIS

Grant Agreement:	833635
Project Acronym:	ROXANNE
Project Title:	Real time network, text, and speaker analytics for combating organised crime
Call ID: **Call name:**	H2020-SU-SEC-2018-2019-2020, Technologies to enhance the fight against crime and terrorism
Revision:	V1.0
Date:	13 January 2023
Due date:	31 October 2022
Deliverable lead:	BUT
Work package:	WP5
Type of action:	RIA

Disclaimer

The information, documentation and figures available in this deliverable are written by the "ROXANNE - " Real time network, text, and speaker analytics for combating organised crime" project's consortium under EC grant agreement 833635 and do not necessarily reflect the views of the European Commission.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the ROXANNE project partners. © 2019 – All rights reserved.

The European Commission is not liable for any use that may be made of the information contained herein.

Copyright notice

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Nature of deliverable:	OTHER	
Dissemination Level		
PU	Public	☒
CO	Confidential, only for members of the consortium (including the Commission Services)	☐
EU-RES	Classified Information: RESTREINT UE (Commission Decision 2015/444/EC)	☐
* R: Document, report (excluding the periodic and final reports) DEM: Demonstrator, pilot, prototype, plan designs DEC: Websites, patents filing, press & media actions, videos, etc. OTHER: Software, technical diagram, etc.		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635. No part of this document may be used, reproduced and/or disclosed in any form or by any means without the prior written permission of the ROXANNE project partners. © 2019 – All rights reserved.

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
V1.0	26 Sep. 2022	Johan Rohdin	All	First version of executive summary and Section 1. Structure and comments for the other sections, adding old material as a starting point.
	13 Oct. 2022	Denis Marraud	Introduction, Video technos	Description of developed and integrated video technologies
	15-16 Oct. 2022	Johan Rohdin	All	Update VAD, SID section. General check and comments.
	17 Oct. 2022	Johan Rohdin	Phone number metadata	Added this section.
	17 Oct. 2022	Erinc Dikici	2.1, 7	Updated ROXSD description and stats, added introductory paragraph to ASR.
	17 Oct. 2022	Johan Rohdin	All	Fixes, clean-up, addressing some comments before the WP wide proofreading.
	18 Oct. 2022	Marek Kovac	1, 2.1, 15	Proofreading, 2.1 cleaning
	19 Oct. 2022	Johan Rohdin	Software, (currently 1.4) Data, (2.2) Spk. Rec. (5.5)	Added software section. Changes related to the real case.
	20 Oct 2022	NFI	11 & 12	Made some corrections and added comments
	21 Oct 2022	Maria Jofre	All	Proofreading
	24 Oct 2022	Johan Rohdin	1-4,6-15	Addressing some comments, proofreading
	25 Oct 2022	Nikos Nikolaou, Vassilis Chatzigiannakis	Chapter 15 - Video	Description of the multi-modal query (15.5)
	25 Oct 2022	Johan Rohdin	17	Addressing some comments, proofreading
	31 Oct 2022	Petr Motlicek	All	Proofreading, adding SID, ASR, etc.
	31 Oct 2022	Johan Rohdin	Conclusion, All	Updating conclusion. Misc. editorial fixes.

	16 Dec 2022	Petr Motlicek	All	Proofreading, providing the comments back to authors, identifying missing parts
	10 Jan 2023	Jakub Tkaczuk	9.	Adding results for German ASR
	10 Jan 2023	Driss khalil	5.4	Adding telephone network initialization part
	12 Jan 2023	Jakub Tkaczuk	11.3	Adding NER results for German and English baseline and boosting
V1.0	13 January 2023	Petr Motlicek	All	Final proof-reading

Executive summary

This deliverable, D5.3 *Final set of ROXANNE speech/NLP/video technologies for network analysis* summarizes the final set of speech, NLP and video technologies available at the end of the project. The emphasis is on the customizations made for improving the analysis of criminal networks. The deliverable is presented as an update of deliverable D5.1, which introduced the speech, text and video technologies employed in the ROXANNE platform and deliverable D5.2, which described the interaction of these technologies with network analysis. The technologies are demonstrated on several datasets available to the project partners. The primary data set is the ROXANNE Simulated Dataset (ROXSD) that encompass many different modalities and thus is suitable for analysing the interaction of the different technologies. Further, real data provided by a LEA partner are also included in the analyses in this report.

Table of contents

1. Introduction

1.1. Background

ROXANNE aims to enhance criminal investigations through the extraction of auxiliary information from multiple facets of data. WP5 concentrates the efforts dedicated to the construction of the core multilingual speech, language and video technology components within the platform. The objectives of WP5 are outlined in the grant agreement (GA) as:

- *Construct the core multilingual speech, language and video technology components;*
- *Make technologies operate in the environment of network analysis (NA): adapt them to serve the goals of NA and improve their performances based on NA outputs;*
- *In Speaker Identification (SID), the transition from the nowadays classical i-vector technology to fully DNN-based systems and augment SID by the information coming from NA;*
- *Advance diarization, to cope with mono recordings (two speakers in one channel) in realistic scenarios;*
- *In automatic transcription, focus on the vocabulary that changes over time on a speaker and group level, turning the out-of-vocabulary (OOV) problem to our advantage; Further, working on a multilingual input and allowing to boost the apriori known set of words (uni-grams or n-grams) provided by a user in the automatic transcript;*
- *Advance video and metadata processing (i.e., Geolocation, textual input associated with audio or video source) to provide contextual information and to support complex speech and video data mining cases;*

The first deliverable from WP5 (D5.1) focused on the development of several individual technological components. Table 1 provides a summary of the individual technologies we support in the final platform. More details are given the in following sections.

		Main contributing partner							
	Technology	IDIAP	BUT	PHO	HENS	US AAR	AIR-BUS	NFI	MOPS-INP
Speech	Voice activity detection	🟩	🟩	🟩					
	Diarization		🟩	🟩					🟩
	Speaker recognition	🟩	🟩	🟩					🟩
	Gender classification	🟩							
	Language recognition	🟩							
	Speech recognition	🟩			🟩				
Text	Topic detection	🟩							
	Entity detection	🟩				🟩			
	Relation extraction					🟩			
	Co-reference					🟩			
	Authorship attribution	▨						▨	
Video	Face matching						🟧		
	Scene matching						🟧		
Meta data	Geolocation							▨	

Table 1: Technologies developed for the final ROXANNE Platform. Green squares are finalized in the platform by the time of submitting the deliverable. Orange squares are expected to be finalized in the platform by the end of the project. Grey squares are available as stand-alone modules.

The second deliverable from WP5 (D5.2) presented preliminary work focused on methods for combining the individual technologies as well as their combination with social network analysis. In particular:

- prior knowledge (e.g. initial graph built from available metadata and/or knowledge of police investigators on the analysed case) combined with speaker identification and network analysis.
- natural language processing (NLP), in particular, named entity recognition (NER) enhanced by so called mention network and co-reference network to automatically extract person identities from case textual data (e.g. manual transcripts obtained from wire-tap recordings), further combined with network analysis and audio-based speaker identification.
- speech-to-text engines that specifically transcribe highly informative words (e.g. names, nicknames, places, etc.) using enhanced methods, in an automatic manner, together with enhanced kinds of language models (more precisely reflecting language use in criminal cases.)
- combination of Automatic Speech Recognition (ASR) outputs with subsequent NLP technologies (i.e. as mentioned above, NER followed by mention and co-reference networks).
- usage of face and scene characterization in images and videos to automatically enrich the speaker network with new edges and nodes derived from visual extracted information.

1.2. Purpose and scope

This report is the third and final deliverable from WP5 (following D5.1 and D5.2), which is responsible for the provision of speech, Natural Language Processing (NLP) and video technologies in ROXANNE. Since D5.2, project partners have worked on adapting the technologies to the ROXANNE datasets and scenarios. This deliverable (D5.3) describes the final set of speech, NLP and video technologies in the ROXANNE platform[1] and their application to criminal network analysis. The experimental results are mainly reported on the ROXANNE Simulated Dataset (ROXSD), which is an in-house developed dataset that simulates communication in the domain of a drug dealing case-work. Details on the development and deployment of the dataset are provided in report D4.2. Furthermore, real-case data partially available (i.e. only results from the analysis or pre-processed/pseudonymized data are shared with the technology partners) from one of LEA partner are also analysed in this document. Compared to D5.1 and D5.2, the main technology improvement and additions described in this report are:

- Improved performance of the core technologies, for example through *boosting* in speech recognition.
- The technologies can handle more generic scenarios, for example speaker clustering can be done with some enrolled speakers.
- More modalities and meta information can be taken into account, e.g., linguistic information for speaker identification as well as geolocation and video analysis.

1.3. Software

The main software output from the project is a platform currently referred to as A*utocrime* or the *ROXANNE platform.* It will be available to, among others, LEAs from EU countries for free. The platform is described in detail in D7.1 - D7.5. In addition, there is code that has not been integrated in the platform. For example, implementations of methods that, though giving promising research results, not yet considered robust enough for use in real operational scenarios. Such code may also be obtained based on agreement with

[1] Please note that since the submission of the last deliverable D5.2, the Consortium has decided to switch to an entirely different platform architecture based on an open-source, Python-driven backend and a modular and responsive docker-based frontend (graphical user interface), and some of the technological components described here had to be adapted to this new environment. The new ROXANNE platform is called “Autocrime”.

the relevant partners. For further information about the software, please contact the ROXANNE consortium at https://roxanne-euproject.org/contact.

1.4. Document structure

This report is structured as follows. Section 2 describes the datasets used to demonstrate the technologies. Sections 4 to 9 describe the speech technologies, Sections 10 to 14 describe the NLP technologies, Section 15 describes the Video technologies, Section 16 is related to geolocation aspects implemented by ROXANNE. Each of these sections describes both core technology and its interaction with other technologies, including network analysis. Finally, Section 17 concludes this deliverable.

2. Datasets

This section briefly introduces the datasets used in the experiments reported in this document. These datasets will be described in more detail in D4.3.

2.1 ROXANNE Simulated Dataset (ROXSD)

Data is crucial for research, development and demonstration activities in the project as Law Enforcement Agencies (LEAs) face considerable obstacles in delivering real data to technical partners for multiple reasons, including legal, security and ethical. Therefore, the consortium proceeded to work with the best alternative by designing and recording its own dataset, called ROXSD in short. ROXSD_v3.0 is the latest version of the dataset as of the writing of this deliverable, and has originated from previous versions of dataset (v1.0 and v2.0) and Task 4.6.

The main advantage of the ROXSD dataset is the fact that the collected data and scenarios are as realistic as possible compared to investigated cases. Still the data are simulated (people/speakers act under their fake role, similarly as actors). This makes the dataset unique, hence suitable to help LEAs to test their technologies and the research entities to test the LEAs requirements.

The task T4.6 benefited from the presence of the Police of the Czech Republic (PCR), namely of the representatives of its National Drug Headquarters (NPC) in the ROXANNE consortium. The original draft of scenarios for recording was defined by PCR based on their professional experience and without revealing real cases. It does not exactly match any of the real cases, but is inspired by them[2]. The case involves a group of criminals communicating over the telephone (i.e., the "target" calls). The wire-tapped data includes also a number of "innocent" people communicating with the criminals and with each other (i.e., the "non-target" calls). The offenders speak in Czech internally (planning local criminal activities) and heavily accented English when planning transnational activities.

The ROXSD was built on this core scenario and was then extended with two extra data collection stages, following up on the main story. This included additional calls and the inclusion of other modalities, such as text messages between persons, GPS location of phones, photos and videos, data that are also simulated.

[2] To get an insight about the range of criminal activities dealt with by NPC, please refer to NPC's public annual reports, available in Czech and English. https://www.policie.cz/clanek/vyrocni-zpravy-annual-reports-jahresbericht.aspx

Its current version, v3.0 (since finalised, from now called as ROXSD dataset), consists of the following subsets:

- audio recordings, simulating the phone calls intercepted by the police.
- video recordings, simulating video messages being sent on a messaging platform or found in a seized mobile device.
- ROXHOOD posts, simulating the content posted on a fictional online discussion site (the "ROXHOOD") where people exchange text messages, share images and videos.

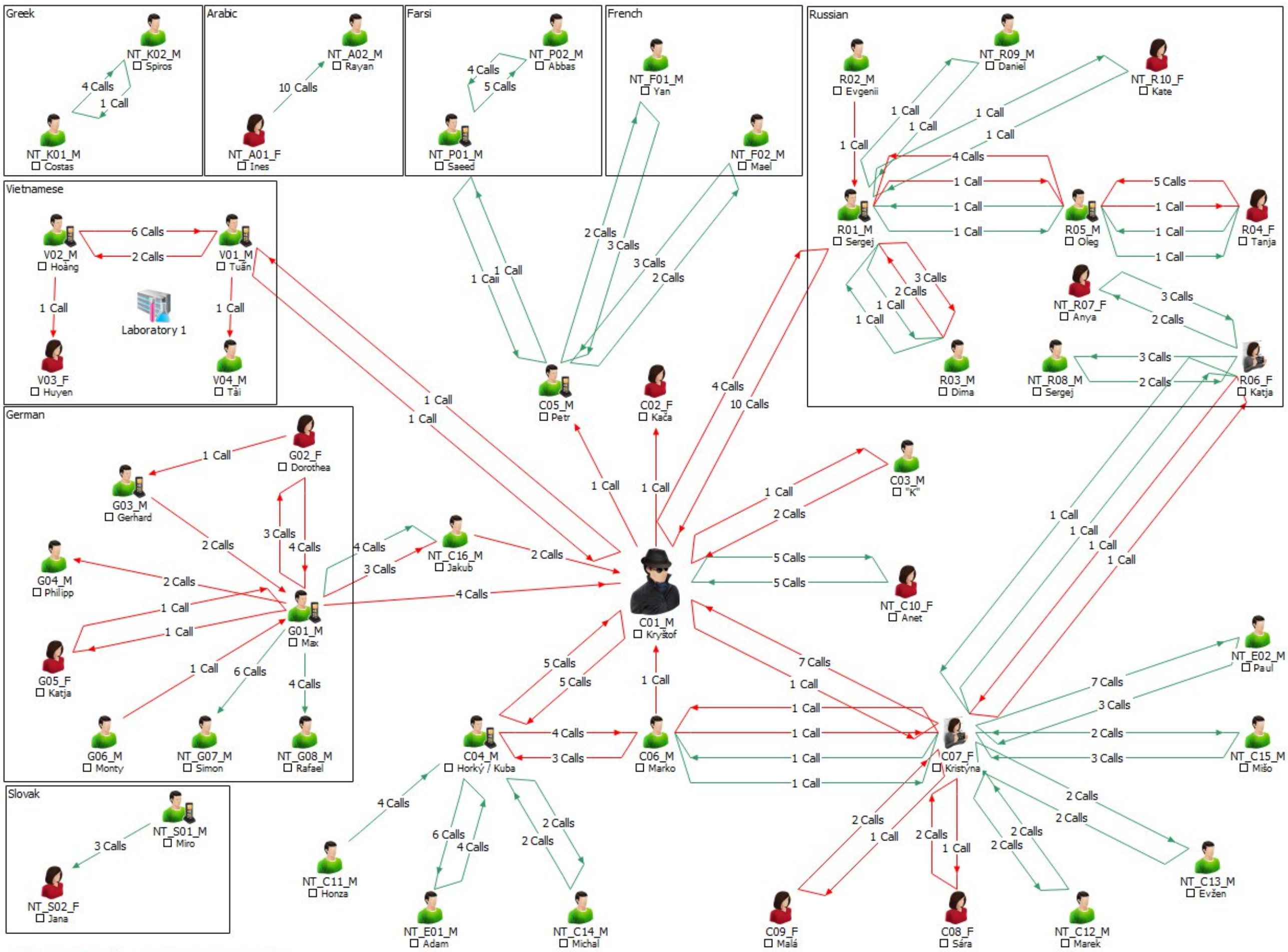
There are 481 recordings inside the phone calls subset of ROXSD. Out of these recordings, 444 contain intelligible speech, while the rest are either failed or interrupted calls. The total audio time of the recordings is 18 h 28 min 7 sec. The total length of two concatenated channels almost doubles this amount, resulting in 36 h 55 min 38 sec. The total time of speech activity (based on an automatic voice activity detector) is 19 h 34 min 21 sec.

The total number of speakers in the calls subset is 103. There may be one or more speakers in a single channel of a conversation. The dataset contains conversations in 15 languages (Arabic, Croatian, Czech, English, Farsi, French, German, Greek, Polish, Romanian, Russian, Slovak, Spanish, Swedish, Vietnamese), some being multilingual.

The recordings are encoded in 8kHz 16-bit wave files. Out of the 481 files, 476 are stereo whereas 5 are mono (single channel).

Ground Truth

As can be seen in Figure 1, the scheme of calls within the criminal group is highlighted in red (“target calls”) and green lines represent telephone calls between speakers who are not suspects (“non-target”) or the call does not contain valuable information for an investigator in the simulated case. The arrow direction on the lines between speakers shows the initiator of the call and points to the receiver of the call. The numbers displayed on the edges is the number of calls realized by each pair of speakers. The edges represent the links between the speakers with attached speaker label and name used in the fictional case. The ground-truth network was manually created at Brno University of Technology with the tool IBM i2 Analyst’s Notebook. The picture represents main actors (“targets” and “non-targets”) of the investigated case in ROXSD dataset. It does not include all actors (“non-target”) calls.



red lines = target calls green lines = non-target calls

Note: It is difficult to distinguish between a target and non-target call so this is just a subjective designation.

Figure 1: Structure of target and non-target calls in network **in ROXSD dataset.**

For the illustration of the whole ROXSD dataset, we add also the scheme (Figure 2) with all speakers. Each node represents one speaker. The colour represents the group in the story (separately organised group suspected for drug dealing). The long name refers to enrolled speaker, the short number refers to unknown speakers or cluster of audios with same speaker.

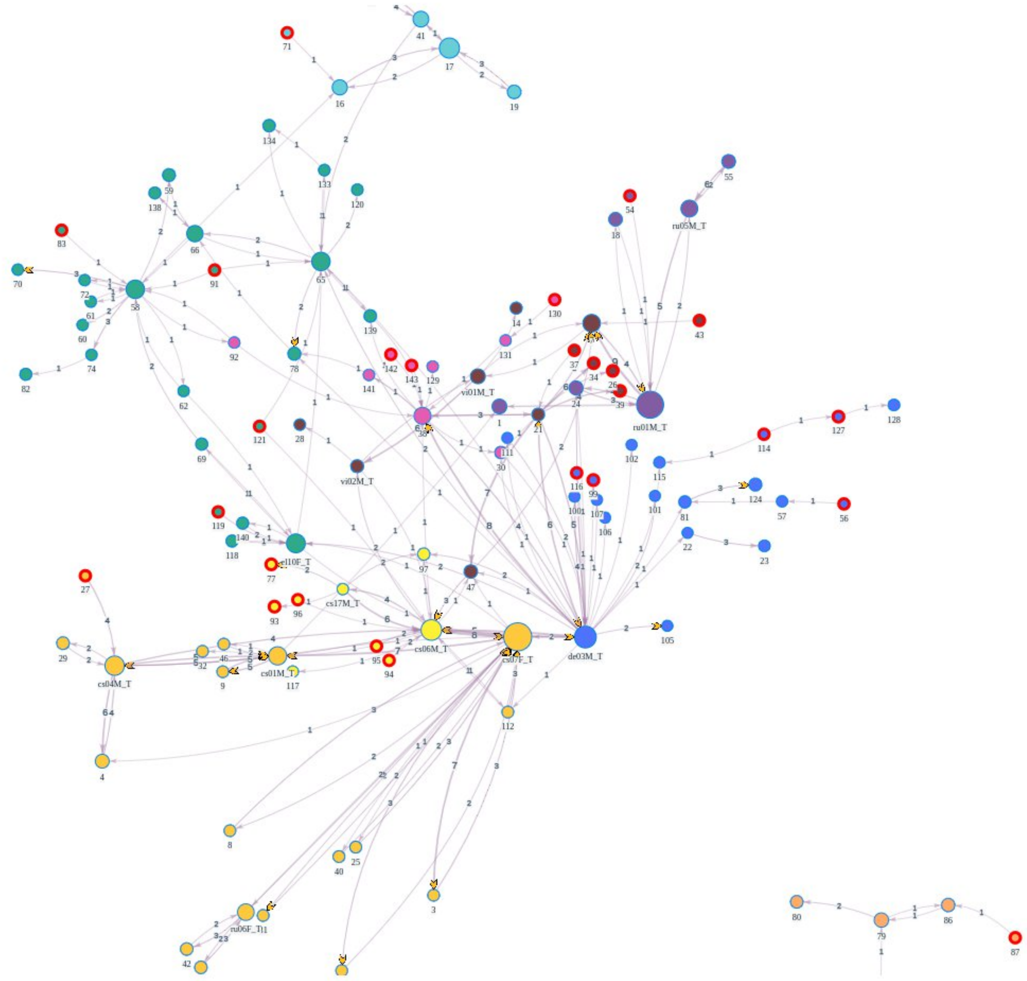


Figure 2: The structure of target and non-target calls in network that can be seen in SW Autocrime.

The following metadata are available for all of the recordings. These data contain partially fictional elements and pseudonymized personal data of the speakers, mainly:

- Case (LEA label of specific investigation)
- Intercepted number
- Owner (telephone number)
- Speaker label
- Speaker story name
- Gender
- Age
- Call information
- Call ID
- Connected numbers
- Date
- Time
- Audio length
- Audio file name
- Call transcription (target calls only + English non-target calls)
- Call translation to English (target calls only)

- Call Status and Notes
- Topics and Keywords (target calls only)
- Caller Environment (target calls only)
- Receiver Environment (target calls only)

The existence of ground-truth, complete speaker labelling, audio transcription and translation provides researchers the opportunity to test their new technologies and their interaction while enabling them to measure changes in the accuracy of results.

2.2 Real Case data corpus

One of the ROXANNE LEA partners have provided data from a real (closed) criminal case for testing within the project. The case contains information about approximately 40K interactions. Some of them are phone calls and some of them are SMS. For the phone calls, the speakers have been identified manually. For both phone calls and SMS, information about the source and target number as well as the start and end time for the call is available. Among the phone calls, around 200 were of special interest for the LEA and were transcribed manually by them Figure 3Figure 3: Phone number network. Each node is a phone number and the links are calls. There are 86 phone numbers in the network. shows the social network with phone numbers as nodes for the ~200 calls of special interest. Note that the speaker labels may have occasional errors.

Due to the sensitive nature of the data, audio, transcriptions, identity of speakers and phone numbers have not been transferred to the project partners. Instead, speaker embeddings (a.k.a. *voiceprints*[3]) from which a speaker's identity cannot be inferred were extracted from the recordings by the LEA on their site. The speaker embeddings of the 200 calls of special interest together with anonymized speaker labels and phone numbers were provided to the other project partners. This allows for experiments and analysis on speaker recognition and network analysis as well as their combination. On the other hand, experiments and analysis on ASR and NLP have to be run on the LEA premises without the need to share the data.

[3] Extracted with the VBx software available at https://github.com/BUTSpeechFIT/VBx.

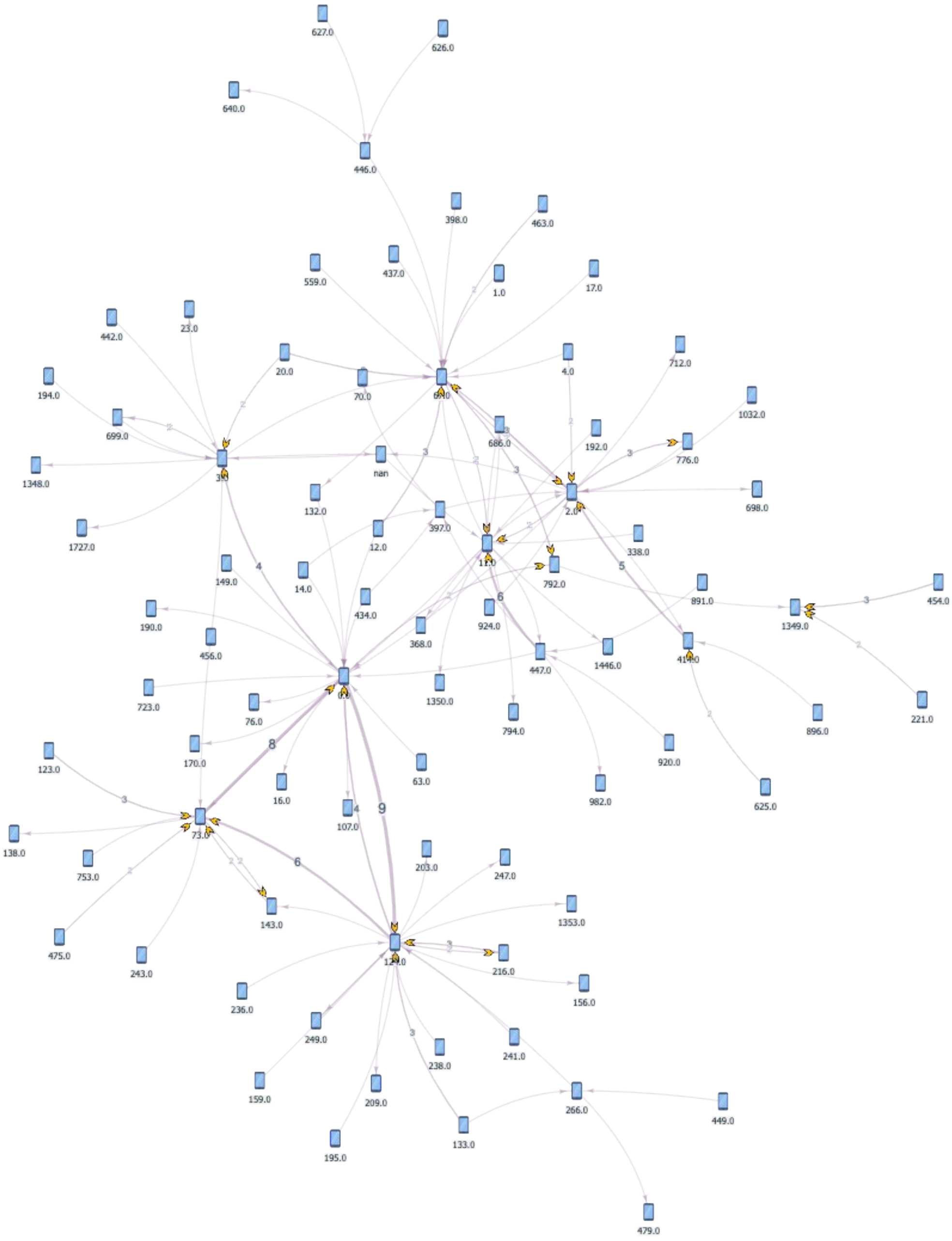


Figure 3: Phone number network. Each node is a phone number and the links are calls. There are 86 phone numbers in the network.

2.3 FRIDA - Forensically relevant inter-device audio database

FRIDA[4] is a database of forensically relevant speech recordings that were acquired simultaneously by multiple recording devices. The telephone conversations are in Dutch between 250 speakers. Particular

[4] D. van der Vloed, et al., “Exploring the Effects of Device Variability on Forensic Speaker Comparison Using VOCALISE and NFI-FRIDA, A Forensically Realistic Database.” In Odyssey 2020 The Speaker and Language Recognition Workshop, 402–7. ISCA: ISCA.

care was taken to recruit speakers from different socio-economic backgrounds. Each speaker has 16 conversations with another participant for approximately five minutes. Approximately half of the conversations, i.e., around eight conversations per speaker, have been orthographically transcribed by three native Dutch speakers following a transcription protocol. This resulted in the transcribed data from 223 participants.

This dataset has been used on text-based speaker recognition, as described in Section 14.1 - Adding Linguistic Features for Speaker Comparison.

3. Network analysis

The key novelty of the ROXANNE platform is the interaction of speech, NLP and video technologies with network analysis (NA). This deliverable together with D5.1 and D5.2 focus on speech, NLP and video technologies in ROXANNE. The NA technologies are described in detail in D6.1 - D6.4. In this section we, however, provide a very brief introduction to network analysis in order for the reader to understand the motivation for the various features of the speech, NLP and video technologies described in this deliverable.

3.1 Criminal network analysis in ROXANNE

In general, network analysis is the process of uncovering patterns in networks using a wide range of computational and statistical methods, which regard the behaviours and relations among individuals in the networks. Those patterns could be the distribution of relationships among the individuals, the underlying factors determining the links or cohesive groups of individuals with dense connections. Besides various domains of daily life, including economics, biology and sociology, the NA methods are also applied in security and criminology. In particular, NA is suitable for ROXANNE's objectives and applications since the members of a criminal network and their interactions form a “social” network as illustrated in Figure 1-4. Several existing methods for NA functionalities relevant to ROXANNE's goals include social influence analysis, community detection, link prediction, outlier detection, network embedding, and cross-network entity matching. There are some practical applications of these methods in ROXANNE. First, social influence analysis methods are employed to quantify the influence of individuals within a social network and identify the most influential individuals in a criminal group or measure the centrality of criminal organizations. Second, community detection methods can assist in identifying individuals' frequent cohesive subgroups and provide added value to LEAs. Third, link prediction methods are suitable for uncovering the missing or hidden, unobserved interactions and predicting the most likely links to be formed shortly. This way, they provide previously undetected but potentially useful links for investigations. Next, outliers detection identifies individuals who exhibit abnormal attributes or interactions/relations with other individuals and strange behaviors that certain individuals do not often perform. Finally, the cross-network entity matching, also called “Cross-Domain Entity Resolution” or “Entity Linkage,” finds graph nodes that indicate the same entities across different graphs.

4. Speech: Voice activity detection

The objective of Voice Activity Detection (VAD) technologies is to distinguish *speech* from *non-speech* in an audio signal. Voice Activity Detection methods are generally designed to be a language, domain and channel independent technology. The output of a VAD module can be used to decide which parts of a recording that should be processed (has enough of speech) by other technologies, such as speech recognition or speaker recognition.

In the ROXANNE platform, two modules for VAD are available. One simple energy-based VAD module and one more advanced VAD module based on neural networks. The latter is proprietary software by Phonexia for which a license must be acquired.

For the project integration platform, IDIAP is in a process to pursue third version of VAD, to be especially robust for noisy speech and used mainly in speech-to-text engine.

4.1 Energy based VAD

Energy based VAD approaches assume that regions of the signal with high energy are speech and regions with low energy are non-speech. Our implementation is a translation of the C++ implementation of the Kaldi toolkit[5] into Python. In this method, frames of the signal are extracted every 10ms. The log energy of each frame is then calculated. Finally, for each frame, the percentage of the frames within a context (configurable) of the current frame that has energy higher than a threshold (configurable) is calculated. If this percentage is high enough (configurable), the frame is judged to be speech. This way, frequent changes in the speech/non-speech decision are prevented.

Obviously, this approach (as other energy based VAD approaches) is not robust against noise. Whether this lack of robustness is problem depends on the downstream task and system. For example, for speaker recognition it may not matter much as long as the speaker recognition system has been trained on data processed with the same VAD because then it will learn to ignore noise that the VAD typically fails to reject. Nevertheless, we also include a state-of-the art neural network based VAD described in the next subsection.

4.2 Neural network based VAD — "GENERIC_3"

We choose to use a VAD model called "GENERIC_3" (version 3.0.2) for integration into the AutoCrime platform. This model labels "speech" and "non-speech" in the signal (as stand-alone task) in a way that is optimal for automatic speech recognition rather than speaker recognition (i.e., the VAD tends to detect fairly long speech segments). Using this model for speaker recognition results in minor degradations in the speaker recognition performance compared to using a VAD model that is optimized for speaker recognition. However, for the simplicity of having a common VAD module for all tasks, we accept this degradation.

The VAD (GENERIC_3 model) is based on a neural network. The network was trained on audio data, always consisting of a pair of clean and noisy or reverberated recordings. The training data were carefully labelled for speech and non-speech segments by human operators.

The model was trained on 8000+ hours of audio (34+ datasets) consisting of various types of channels (mainly telephony), languages and environments.

It is evaluated on the complete ROXSD dataset against the ground truth manual segmentation. The frame level evaluation gives the total detection error rate of VAD equal to 1.0%. False alarms are detected for 0.9% of frames and the miss detections sum up to 0.17%. After automatically segmenting the English part of ROXSD, the Word Error Rate of transcripts is calculated and compared against WER obtained from manually segmented data:

[5] https://github.com/kaldi-asr/kaldi/blob/master/src/ivectorbin/compute-vad.cc

- manual segmentation: 28.4%
- automatic segmentation: 32.9%

4.3 Multilingual-based VAD

To better model the contextual information and increase the generalization ability of VAD system, this approach leverages a multi-lingual ASR system to perform voice activity. Sequence-discriminative training of acoustic model using Lattice-Free Maximum Mutual Information (LF-MMI) loss function, effectively extracts the contextual information of the input acoustic frame. Multi-lingual acoustic model training causes the robustness to noise and language variabilities. The index of maximum output posterior is considered as a frame-level speech/non-speech decision function. Majority voting and logistic regression are applied to fuse the language-dependent decisions. The multi-lingual ASR is trained on 18 languages of BABEL datasets and the built VAD is internally evaluated on 3 different languages (see Interspeech 2021 paper[6]). On out-of-domain datasets, the proposed VAD model reveal significantly better performance with respect to baseline models.

W.r.t. ROXSD data, the performance of the technology is very accurate. We are able to reach the detection error-rates around 1.18% (i.e. a frame-based error for correct or wrong voice activity classification, compared to 4.1% error of WebRTC engine[7]). For the case of analysing the error rate on segment level (i.e., if the whole speech segment is correctly or incorrectly classified, taking into account the center of each segment as the point of interest), we are able to reach 20.6% error, while for the WebRTC open-source engine the error-rate was around 37.4%).

In analogy to 4.2, the multilingual-based VAD is evaluated on the ROXSD dataset. The total detection error rate of this VAD is 0.70%. The false alarm are detected for 0.54% of frames. Miss detection is equal to 0.16%. Even though the presented numbers are slightly better compared to those discussed in 4.2, the influence on the automatic speech recognition system is marginal. The WER for segments obtained with this VAD is equal to 32.8%.

5. Speech: Speaker recognition

Speaker recognition (SR) refers to the process where a machine infers the identity of a speaker by analysing his/her speech. Speaker recognition is an integral part of the ROXANNE platform because the identities of speakers in recordings from criminal investigations are usually not known. There are five sub-problems of speaker recognition that are commonly studied, namely:

- **speaker diarization** Speakers are detected in one mono recording and individual speakers' segments are clustered.
- **speaker verification** The speech from one specified *enrolled* (registered) speaker is compared with the speech of a *test* utterance and the system should decide whether the test utterance is spoken by the specified speaker or not.
- **speaker identification** Similar to speaker verification but the test utterance is compared to many enrolled speakers. The system should then tell which of the enrolled speakers speak in the test utterance. Depending on the nature of the task, the system may also have the option to decide the speaker of the test utterance is not any of the enrolled speakers.
- **speaker clustering** Based on a set of unlabelled recordings, we try to infer the amount of speakers and attribute them to the calls.

[6] http://publications.idiap.ch/index.php/publications/show/4570

[7] https://webrtc.org

- **speaker search** One or several speakers are searched in a large quantity of data.

All of the above tasks are relevant to ROXANNE and can be performed with the technologies implemented in the platform. In fact, a more general usage which we refer to as "clustering with enrollments" is supported, which is the most relevant scenario in criminal investigations. The speaker recognition technologies and the customizations we have done are described in the following subsections.

5.1 Core speaker recognition technology

The speaker recognition system in the ROXANNE platform follows a state-of-the art approach[8,9]. In this approach, a sequence of feature vectors are extracted from the signal, each representing few milliseconds of speech. Feature vectors from non-speech according to a voice activity detection module are then discarded. The remaining sequence of feature vectors are then converted by a neural network into a fixed size vector representation usually referred to as a *voiceprint* or *embedding.* Typical neural network architectures first process every feature plus some context individually, then calculate e.g. the mean and the standard deviation as a pooled representation for the utterance. The pooled representation is then usually processed with a few more dense layers to produce the embeddings. A light introduction to voiceprints and their properties is provided in our blog[10]. The ROXANNE partners have experimented with several state-of-the-art voiceprint extractors. The one integrated in the platform is based on a ResNet[11,12] architecture and available as part of the VBx recipe[13] described in the next section.

The embeddings are then modelled by a generative model called Probabilistic Linear Discriminant Analysis (PLDA)[14]. The first four of the above mentioned problems can then be solved directly with the PLDA model and rules of probability.

5.2 Diarization

Given a recording where several speakers are present, the diarization task is to segment the recording into regions such that only one person speaks in a region and cluster the regions according to speaker identity. See figure 4.

[8] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification" in Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[9] D. Snyder et al., "Deep neural network embeddings for text-independent speaker verification", Proc. Interspeech, 2017.

[10] Shttps://roxanne-euproject.org/news/voiceprints-and-their-properties

[11] K. He, X. Zhang, S Ren, J .Sun, *Deep Residual Learning for Image Recognition*. *2016 IEEE CVPR*. Las Vegas, NV, USA: IEEE. pp. 770–778

[12] H. Zeinali et al., "BUT system description to VoxCeleb speaker recognition challenge 2019," in *VoxSRC Challenge workshop*, 2019

[13]F. Landini, J. Profant, M. Diez, L. Burget: Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks, Computer Speech & Language, Volume 71, January 2022, https://github.com/BUTSpeechFIT/VBx

[14] S. Ioffe, "Probabilistic Linear Discriminant Analysis", in ECCV 2006.

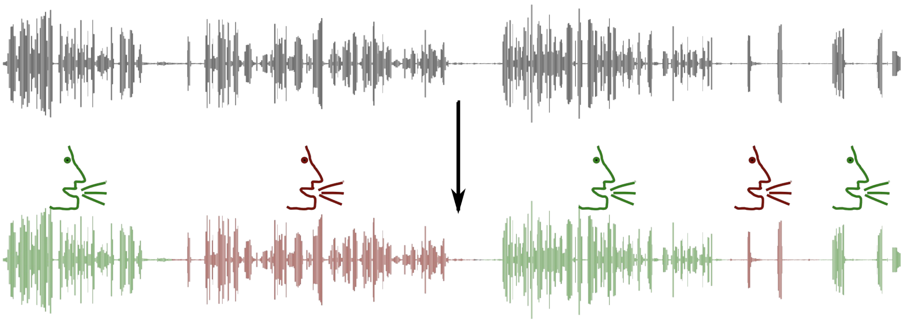


Figure 4: Speaker diarization. An utterance with that may contain several speakers is input to the diarizarion system (top). The diarization system detect that there are two speakers and where they speak (bottom).

Diarization is important in ROXANNE because:

- Telephony data is often stored as mono which means that the two speakers in the call are in one channel.
- There could be more than one speaker per side in the call. For example, if the phone is handed over to another speaker.
- Audio from video or virtual meetings often contain several speakers.

After diarization, one can extract one speaker embedding as usual.

The most common approach to diarization is to divide the utterance into segments that are so short that they generally contain only one speaker and then cluster them. Typically (unweighted pair group method with arithmetic mean) Agglomerative Hierarchical Clustering (AHC)[15] is used. In ROXANNE, we use this approach plus an additional step based on Variational Bayes Hidden Markov Models (VBx) that refines the result[16]. Both AHC and VBx relies on the PLDA model for comparing speaker embeddings. VBX is, in short, a first order hidden Markov model for transitions between speakers where the output probabilities are modelled by probabilistic linear discriminant analysis PLDA.

Constraining the number of speakers

To correctly detect the number of speakers in a recording is difficult so in scenarios where constraints on the number of speakers can be assumed (as is the case in many of the ROXANNE scenarios), there is potential to greatly improve the diarization. The standard VBx recipe, however, does not provide a mechanism for constraining the number of detected speakers. Given a *minimum* and *maximum* on the number of speakers, we therefore modify the recipe as follows:

- The AHC used for initialization is stopped if the minimum number of speakers is reached. As usual, it is also stopped if a threshold in the similarity score for the two most similar clusters is reached.
- The VB diarization then outputs the best solution that does not have less than the minimum number of required speakers.
- In a post-processing step, if there are more than the maximum number of required speakers, the speakers with most segments are selected and the remaining segments are reassigned to one of these speaker based on the posterior probability of them belonging to the different speakers.

Note that AHC always reduces the number of speakers in every iteration and that the VB diarization cannot increase the number of speakers to more than what is available in its initialization. In our experiments, both the minimum and maximum number of speakers are set to two which means that exactly two speakers will

[15] https://en.wikipedia.org/wiki/UPGMA

[16] M. DIEZ Sánchez, et al. "Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, vol. 28, no. 1, 2020, Software: https://github.com/BUTSpeechFIT/VBx

be obtained. The resulting diarization (DER)[17,18] error rate is 14.6%. A more meaningful result from the application point of view is however how mono data (which requires diarization) compares to stereo data for speaker clustering. Such analysis is given in the next subsection.

5.3 Speaker recognition in ROXANNE — Speaker clustering with enrollments

To facilitate the needs in criminal investigations, we developed a speaker recognition system that support a very general scenario and includes the verification, identification and clustering as special cases. The scenario is the standard clustering but with the possibility for the user to specify the identity of the speakers in some recordings (may be known due to manual identification or other external information). When some identities are specified, the system will cluster the data with the following constraints:

- Recordings with the same ID must be grouped together.
- Recordings with different ID must not be grouped together.

In addition, there is also the constraint that one speaker cannot be present on both sides of a call. The platform uses AHC for speaker clustering and the constraints are implemented by setting the relevant entries in the score matrix to a very high or low value: a very high value enforces the corresponding two recordings to end up in the same cluster and a very low value prevents the corresponding two recordings to end up in the same cluster. The process is illustrated in Figure 5.

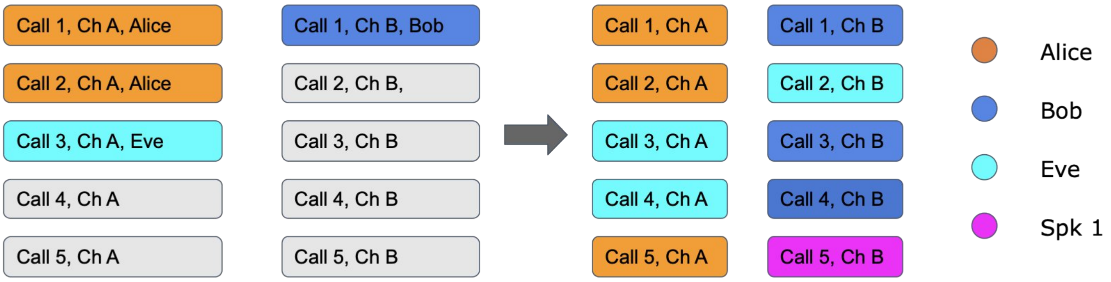


Figure 4: Speaker clustering with enrollments. Five calls (ten recordings) are to be clustered. Identities have been provided for four recordings, i.e, these speakers are enrolled. The remaining recordings will be assigned either to one of the enrolled speakers or to a new speaker as in the case with *Spk 1*.

Note that it is possible to force the system not to detect more speakers than those enrolled, which leads to the standard (closed set) *identification* task.

Results

In order to evaluate a clustering system, one needs a mapping between the speaker labels assigned by the system and the ground truth labels. We obtain the mapping by using the Hungarian algorithm to maximize the clustering accuracy, which is defined as percentage of recordings that are correctly recognized. To illustrate this, let "R*" denote a reference label and "A*" denote an automatic label. Standard

[17] In the evaluation we used a *collar* of 0.25s. See https://pyannote.github.io/pyannote-metrics/reference.html for a description of *diarization error rate* and *collar*. The reference was created by applying voice activity detection to the two channels of the stereo recordings (which we have for the ROXSD data) and assuming each channel had exactly one speaker.

[18] Note that since the references were created as described in the previous footnote, they have always exactly two speakers which may not be the true situation. The DER obtained by constraining the number of speakers to two may therefore be overoptimistic.

brackets “()” refer to an ordered set, while curly brackets “{}” to unordered set. As an example, we then have have:

	Reference	Automatic
Call 1	(R1, R3)	(A6, A2)
Call 2	(R1, R3)	(A7, A6)
Call 3	(R2, R3)	(A1, A7)

In the case of stereo, if we map R1=A6, R3=A7, R2=A2, None=A1, we get:

Call 1	(R1, R3)	(A6, A2)	1 error
Call 2	(R1, R3)	(A7, A6)	2 error
Call 3	(R2, R3)	(A1, A7)	1 error

The accuracy is 2/6.

The best mapping is R1=A6, R2=A1, R3=A2, None_0 A7 which gives

Call 1	(R1, R3)	(A6, A2)	0 error
Call 2	(R1, R3)	(A7, A6)	2 error
Call 3	(R2, R3)	(A1, A7)	1 error

The accuracy is 3/6.
To be precise there are more than one mapping that could give the same accuracy.

In the case of mono recordings, both speakers are in the same recording so if we map R1=A6, R3=A7, R2=A2, None= A1, we get:

Call 1	{ R1, R3 }	{ A6, A2 }	1 error
Call 2	{ R1, R3 }	{ A7, A6 }	0 error
Call 3	{ R2, R3 }	{ A1, A7 }	1 error

So the “call accuracy” is 4/6.

The base mapping is R1=A6, R2=A1, R3=A2, None_0=A7, which results in:

Call 1	{ R1, R3 }	{ A6, A2 }	0 error
Call 2	{ R1, R3 }	{ A7, A6 }	0 error
Call 3	{ R2, R3 }	{ A1, A7 }	1 error

So the “call accuracy” is 5/6.

Note that mapping is only used when evaluating the system against a ground truth reference and not under real operation where the ground truth is not known. Using the mapping that gives the best results[19] is reasonable because it cannot remove the two errors that can happen in clustering, namely:

- recordings of two different speakers are incorrectly placed in the same cluster.
- recordings of one speaker is divided into more than one clusters.

In many real data sets it may not be exactly one speaker per side in the call. Sometimes the phone might be handed over to a second person so that there are two speakers in one side. And sometimes, there is no one speaking on one side. In this situation, too many or to few speakers will be considered errors. For example, if the system predicts two speakers but there are only one in the reference, then the extra speaker will be counted as an error.
These are the errors we wish to measure with cluster accuracy. For an analysis of different metrics that reflects other aspects of the system output important for a criminal investigation, please see D6.3.

[19] Note that this is also done in the calculation of diarization error rate. Similarly, in the calculation of word error rate in speech recognition, we use the alignment of reference and system output that give the fewest errors.

The clustering accuracy is 89.5%. When enrolling 13 speakers (the criminals in ROXSD) with one recording each, the result remains the same. When enrolling 97 speakers with one recording each, the results reduces to 88.9%. It may seem counter-intuitive that the accuracy reduces when we enroll more speakers since we provide more information. The reason is that when we enroll speakers, the evaluation mapping procedure constrains the mapping so that a cluster containing an enrolled file keeps the correct speaker ID and this may not be the optimal mapping. For example if the recordings of one speaker are assigned to one small and one large cluster it is better to map the large cluster to the speaker ID in evaluation. However, if one of the files in the small cluster is enrolled, we cannot do this. To evaluate the performance of diarization when the downstream task is clustering, we merge the two channels of each call in the ROXSD data into one mono channel. We apply diarization to separate the speakers and the clustering as usual. This result in an accuracy of 84.1% which can be compared to 89% if stereo data is used but evaluated as mono.

Closed set identification, i.e., each recording must be identified as one of the enrolled speakers, has an accuracy of 98.5% (13 speakers, 461 recordings) and 96.1% (97 speakers, 896 recordings). Note that in the closed set identification experiments, we only consider recordings (one side of a call) which has exactly one speaker and where this speaker is among the enrolled speakers.

Set \ # enrollments	0	13	97
Open set	89.5%	89.5%	88.9%
Closed set		98.5%	96.1%

Table 1: Identification accuracies.

5.4 Research

This section describes two research directions we have been working on for speaker recognition in ROXANNE. The first method, *mono enrollments*, is likely to be integrated in the platform before the end of the project. The second method, *network structure,* is likely to need some more work before it is robust enough for being part of the platform.

Mono enrollments

When the data is stored as mono, it becomes more complicated to enroll a speaker. Suppose the user knows that speaker A speaks in a call and wishes to enroll this speaker. There is one more speaker in the recoding. Diarization can separate the two speakers but obviously it cannot tell who is speaker A. The problem can of course be solved with some manual work. For example, the user can listen to the original recording and mark a segment where speaker A speaks or listen to the two outputs from diarization and inform the system who is speaker A.

We have developed three approaches to avoid manual work when enrolling speakers from mono recordings[20]. The assumptions for these approaches is that we have several calls for the enrollment speaker where it can be assumed that the conversation partners are different in the different calls. The results are shown in Table 2. See also some results in Section 10.5.

Method \ # enrollments	2	4	6	8
Median	12.22	5.00	5.00	12.22
Intersection	4.44	3.33	2.78	2.78
Cluster	19.44	5.00	1.67	1.67

Table 2 EER (%) for distinct enrollment and verification combinations using multisession PLDA by-the-book scoring.

[20] *Yosef Solewicz*, *Noa Cohen*, *Johan Rohdin*, *Srikanth Madikeri*, *Jan "Honza" Cercnocky*, Speaker recognition on mono-channel telephony recordings, Odyssey 2022, https://www.isca-speech.org/archive/pdfs/odyssey_2022/solewicz22_odyssey.pdf.

Network structure

The speaker recognition scenarios in criminal investigations differ from the traditional speaker recognition scenarios in that all the recordings from a criminal case are related to each other and therefore should be analyzed jointly. Most importantly, many people are present in more than one recording. Some of them naturally occur more often than others and this should be taken into account when modelling. Moreover, people typically do not call all other people in the network equally often but rather they tend to call some people more often than others. The expected structure of a criminal network should therefore be taken into account as *prior information* when applying speaker recognition to the recordings of a criminal investigation.

In previous work[21] it has been proposed to scale up the speaker recognition score based on the frequency of the participants appearing in previous communications. The ROXANNE partners developed this method further to be able to use it for multiparty conversation and applied it to the CSI[22] and ROXSD data[23]. The speaker identification accuracies[24] improved from 89.5% to 90.6% and 88.1% to 89.2% on the CSI and ROXSD datasets respectively. The method developed relied on re-ranking of the most likely pairs. For example, if two speakers have similar likelihood ratios for one conversation, but another speaker is clearly identified for the second one, the structure of the existing network can help us identify the most likely pair. An average score, weighted by the strength of the link between the two potential speakers, is then computed, and the most likely pair is selected. This can be seen in the Figure 7 below:

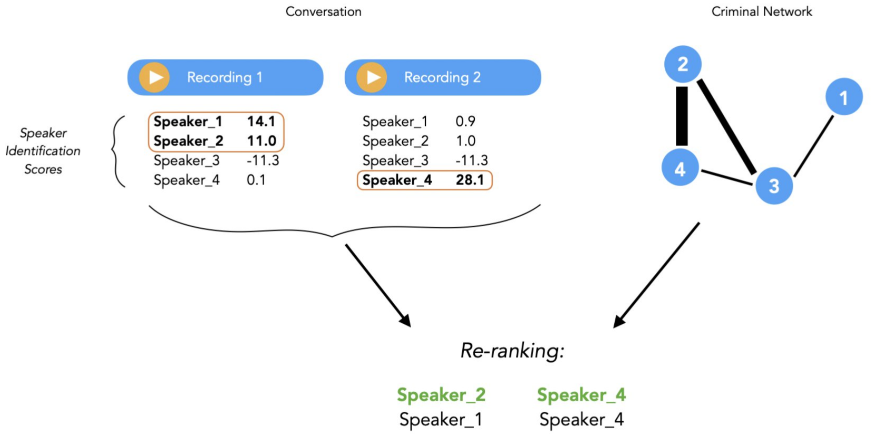


Figure 5: Graph2Speak re-ranking process.

The partners have also started to work on a probabilistic framework for combining speaker recognition and social network structure. In the following description, “L” denotes the label or the speaker whereas “X” denotes the acoustic evidence. In the probabilistic framework, our assumption about the structure of a criminal network is formulated as prior probability, $P(L)$, for the network structure. Note that in this context,

[21] Ning Gao,Gregory Sell,Douglas W. Oard, Mark Dredze *LEVERAGING SIDE INFORMATION FOR SPEAKER IDENTIFICATION WITH THE ENRON CONVERSATIONAL TELEPHONE SPEECH COLLECTION,* ASRU 2017

[22] Lea Frermann et al., Whodunnit? Crime Drama as a Case for Natural Language Understanding, Transactions of the Association for Computational Linguistics, Volume 6, 2018

[23] Mael Fabien, Seyyed Saeed Sarfjoo, Petr Motlicek, Srikanth Madikeri, *GRAPH2SPEAK: IMPROVING SPEAKER IDENTIFICATION USING NETWORK KNOWLEDGE IN CRIMINAL CONVERSATIONAL DATA, SPSC 2021*

[24] In this task, each speech segment should be assigned to an enrolled speaker. Speaker identification accuracy then refers to the percentage of speech segments that have been assigned to the correct assigned speaker.

network structure means information about who is calling who, i.e. a prior for the speaker labels of the calls is equivalent to a prior of the network structure. The prior, $P(L)$, is then used together with the likelihood of the acoustic evidence given the labels, $P(X \vee L)$, to compute the posterior probability of labels given the acoustic evidence, $P(L \vee X)$, via Bayes' theorem: $P(L \vee X) = \frac{P(X \vee L) P(L)}{P(X)}$.

Telephone number as prior clustering

In order to use the telephone numbers as prior knowledge for speaker verification we introduce two parameters that needs to be specified by the user based on their expectation of the case. The parameters are

- $P\left(samespeaker \vee samephonenumber\right)$
- $P\left(samespeaker \vee differentphonenumber\right)$

Where P denotes *probability*. The parameters are then used to modify the speaker verification scores based on probabilistic rules.

Evaluating on ROXSDV1[25], without telephone prior we had 443 out 472 correct classifications. With telephone prior we get 446 out of 472 correct classifications with optimal choice of the above parameters. Obviously this effectiveness of this method depends on the situation of the case. If the speakers always switch phone numbers, it is not helpful. On the other hand, if the speakers never switch phone number it results in perfect accuracy (speaker recognition would not be needed because we could rely completely on the phone number as the speaker ID.) If the speakers sometimes switch phone numbers, the method may improve speaker recognition.

Telephone network initialization

It’s also possible to use the telephone numbers in another way using the AHC algorithm. Instead of using the number of embeddings as an initialization for the algorithm, we initialize the cluster with the phone number network. Clustering starts with an assumption of a single phone number per cluster. A speaker can use multiple phone numbers. Evaluating this on ROXSD, we created 4 different scenarios:

A. 2 speakers with 2 phone numbers
B. 7 speakers with 2 phone numbers
C. 2 speakers with 4 phone numbers
D. 7 speakers with 4 phone numbers

Scenario	Initialization accuracy	Clustering accuracy
Scenario A	91%	98%
Scenario B	82%	98%
Scenario C	86%	97%
Scenario D	74%	99%

It was assumed that each phone number was only used by 1 speaker, and we can remark that the clustering accuracy gives an accuracy around 98% even with more complex scenarios. The initialization accuracy represents the accuracy obtained from the initialization of the clusters (1 cluster represents 1 phone number). This initialization accuracy depends on the complexity of the scenario.

Future work

[25] Note that the method presented here can only be used when we have a mapping between phone number and and channel in the intercepted telephone call. This information is only available for ROXSD V1. We are currently developing methods for finding such mappings automatically based on various heuristics which are likely to be added to the final version of the platform.

It should be noted that there are more complex aspects regarding the network structure than how frequently different speakers occur in the calls. For example, assume there is telephone communication between Alice and Donald and between Bob and Donald, in other words both Alice and Bob know Donald. Since *Alice and Bob are linked via Donald*, there is a reasonable chance that they know each other. Accordingly, if there is a phone call between Alice and unknown person we shall be more inclined to believe that the unknown person is in fact Bob than if there was no link between Alice and Bob. This is an example of *link prediction* and relates also to community detection, which has been studied in WP6. Obviously not all problems can be tackled within the timespan of the project. The partners aim to study this and other more complex aspects of network structures and their interaction with speaker clustering in future work.

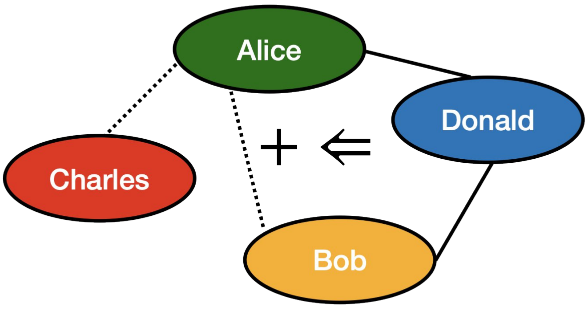


Figure 6: A simple example of *link prediction* and its effect on speaker recognition. There are existing links between Alice and Donald and between Donald and Bob. Accordingly, there is an increased probability for a link between Alice and Bob since they have a common friend.

5.5 Speaker Recognition on a real case

In this section we describe experiments and analysis on data from the real criminal case described in Section 2.2 Real Case data corpus.

The data from the real case is stored as one mono recording to save data storage space. This means that the recordings need to be diarized before clustering can be done.

Speaker recognition and analysis

Although speaker clustering is the task of interest, it is beneficial to evaluate the *speaker verification* for better understanding of the performance of the system. Since all speakers in a call (usually two) have been mixed into a mono channel, we cannot unfortunately evaluate the standard speaker verification scenario (two recordings with one speaker each are compared and the system should tell whether it is the same speaker in both recordings). The reference information tells us the ID of two speakers that are in each call, but of course we do not know which one of the detected speakers in the diarization corresponds to which ID so we cannot create the correct reference labels for the standard speaker verification scenario.

Instead, we can look at two calls, A and B, and ask the system whether any of the speakers in Call A are the same as any of the speakers in Call B. We will refer to this task as *call verification*. For this we know the reference. A simple heuristic for calculating the "score" for this task is to compare all speaker embeddings from Call A with all embeddings from Call B and then take the maximum of the scores as the final score for the call verification score. That is, in the case of two speakers per call, denoted A1, A2, B1, B2, there are 4 comparisons [A1,B1], [A1,B2], [A2,B1], [A2,B2].

Before analysing the performance on the real case, we performed several experiments on ROXSD[26] where we have audio recordings in stereo and therefore can evaluate standard speaker verification as well as better analyse the call verification task. The summary of the results is:

- **Standard speaker verification:** Experiments are done by extracting embeddings from each channel in the mono recording. All possible trials that can be created from the data are used for evaluation. The equal error rate (EER)[27] is 3.9%.
- **Call verification:** Extract one embedding from the mono of the mixed channels, which can be seen as the most naive and probably worst possible way to do call verification. EER is 16.3%.
- **Call verification.** Cheating[28] by extracting one embedding from each channel. Then use the *max approach* described above. EER is 7.7%. Note that this is worse than the results of the standard speaker verification experiment, probably because "non-target" trials are more easily taken for target trials since out of the four scores in the trials it is sufficient that one of them is large by chance.
- **Call verification:** Embedding extracted from diarized segments (i.e. no cheating). Max approach for scoring. EER is 14.2%.

In conclusion, the EER for call verification is approximately two times higher for call verification than for standard speaker verification if diarization is perfect. If real diarization is used, the result is approximately two times worse. We suspect that this may be more because of how the speaker embeddings are created from the different time segments resulting from diarization rather than because of the performance of the diarization system, but this needs further analysis.

The call verification on the real case with real diarization results in an EER of 26.0%, which should be compared to the EER of 14.2% for the ROXSD data. There are several possible reasons why the result for the real case is worse than for the ROXSD data, including:

- The speaker recognition system is trained mainly with English data.
- There are some technical signals (e.g., tones) appearing in the real case data which we are not yet dealing with properly.
- The conversation lengths are not uniform and we not compensate the thresholds for that.
- The very informal language and lazy talking.

Mono enrollments

Table 3 shows evaluation results for the three proposed methods, including Median, Intersection and Clustering enrollment with three different scoring methods, namely PLDA as well as simple cosine similarity scoring optionally followed by t-norm[29]. The results clearly indicate the superiority of the Intersection enrollment method in combination with t-norm. The explanation seems to be that scores obtained for the segmented speakers in the testing conversation are differently biased and it is important to reduce this effect before the max operation discussed in the previous subsection. For other scoring methods (Intersection and Clustering enrollment), we observe that they are quite comparable.

	Cosine, No T-norm	Cosine, T-norm	PLDA
Median + cosine	20.6	17.4	18.3

[26] These experiments were done on ROXSD V1 which is more suitable for this analysis because it always contains one speaker per side in the call.

[27] Equal error rate is the error rate obtained by adjusting the decision threshold so that the false acceptance rate and the false rejection rates are equal.

[28] We assume perfect diarization (i.e. no errors due to splitting the speakers).

[29] R. Auckenthaler, Mi. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, no. 1, pp. 42–54, 2000.

Intersection + cosine	15.6	9.3	16.8
Cluster + cosine	19.2	15.4	15.6

Table 3. EER (%) for distinct enrollment and verification combinations using cosine similarity and PLDA for scoring.

Speaker Clustering

Finally, we performed clustering on the speaker embeddings obtained from diarization of the real case data. We used the standard AHC algorithm for this end. The algorithm takes as argument the score threshold that determines when to stop the clustering process so by adjusting the threshold, one can adjust the number of obtained speakers. Figure 7 shows the network from clustering with threshold that resulted in 88 clusters (i.e. unique speakers) and a clustering accuracy of 61.4%

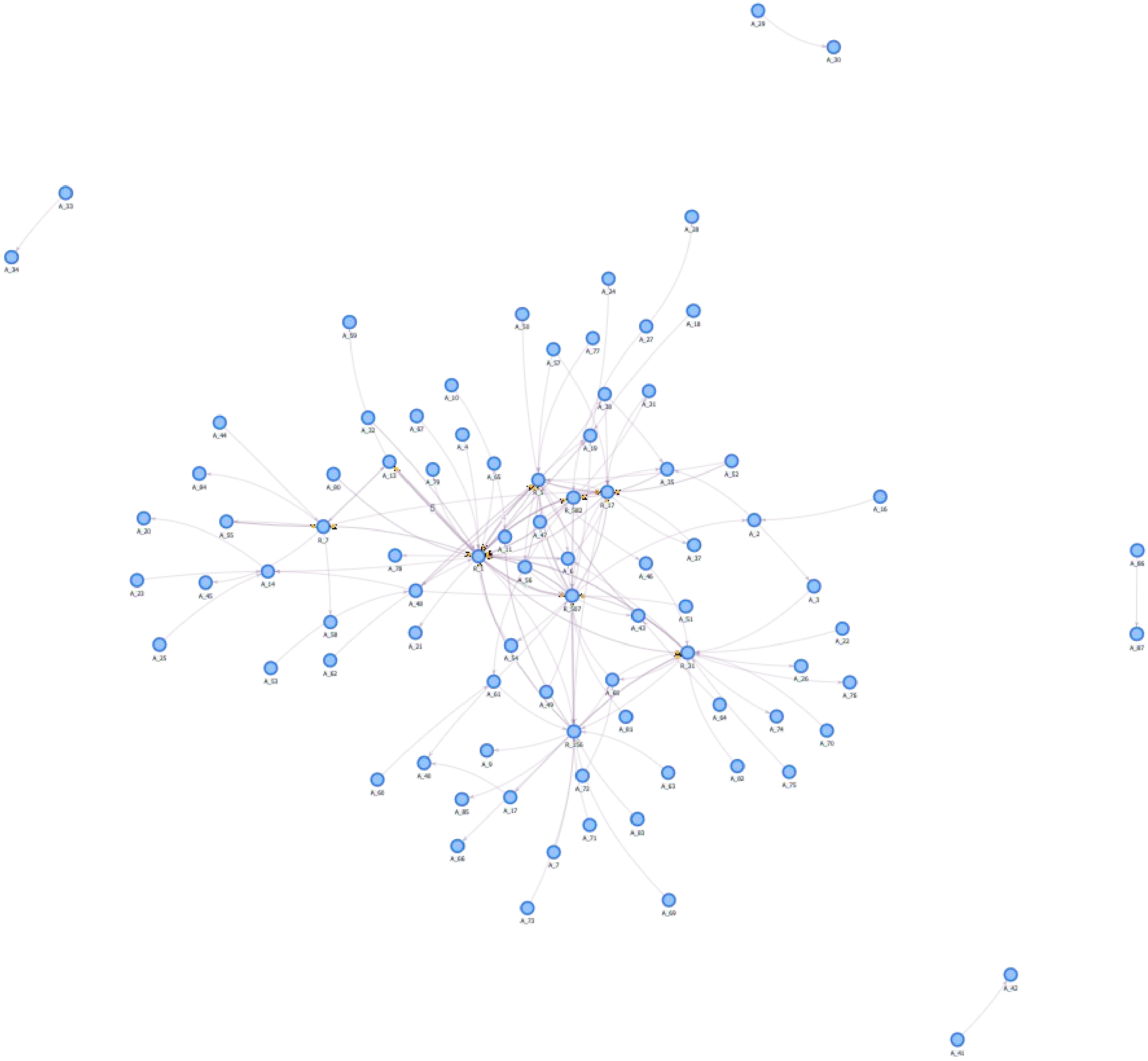


Figure 7: Network based on automatic speaker clustering of the real case. There are 88 detected speakers in the network.

6. Speech: Gender classification

Current gender classification engine integrated into the Autocrime platform is built around two Gaussian mixture models, each representing one of the class. The observation vectors are formulated as standard MFCCs. This type of modeling provides a gender classification accuracy above 90% on ROXSD test data. For the final project Autocrime platform, we aim to replace the code by a modern based x-vector extractor classifier, ideally adapted on ROXSD development data.

7. Speech: Age recognition

Age Estimation (AGE) speech technology automatically estimates the age group of a speaker.
Technology Details are as follows:

- Estimates a person's age with an accuracy of ±10 years,
- Trained with an emphasis on spontaneous telephone conversations,
- Is language-, accent-, text-, and channel-independent,
- Applies state-of-the-art channel compensation techniques.

The technology has not been validated within ROXANNE project as we did not get access to labelled data for the technology assessment.

8. Speech: Language recognition

Language recognition module is developed to automatically detect a target language given the test segment.

We have developed and deployed a suitable language recognition module system deploying x-vector extractor followed by the PLDA classification module, similar to speaker recognition pipeline. The development of the module has been following the work from M. Grisard summarised here[30], where so called bottleneck features (partially similar to well-known x-vectors) are used where the final layer of the neural network represents the language targets.

The most recent work, which is being integrated to the autocrime project platform follows the challenges of language recognition evaluation organised by NIST LRE 2022[31]. The module will be able to reliably detect the language for of the length between 3s and 30s of speech (as determined by an automatic speech activity detector). We assume that each segment comprising the speech Each segment contains one of the target languages only to be used for classification. The module is trained on large variety of languages (including NIST LRE 2017 data[32]), nevertheless, the classification module is aimed to be modular (depending on the languages requested by the police users). Our work on LRE22 consisted on three back-end systems, two of which use the kaldi-based x-vector models and the the ECAPA-TDNN model from speechbrain trained with Voxlingua 107 data for language identification task. We also investigated employing relatively simple classification systems such as Random Forest (RF) and Support Vector Machine (SVM) for language identification instead of the PLDA classifier.

Two experimental setup were used in our work, the first one consist on using only the data provided by the LRE22 for training our models. The second one consist on a open condition where we used the following datasets for training: LRE 17 train data is used for training the x-vector system, LRE22 dev data and BABEL datasets are then used to train two separate PLDA models. The train split of LRE22 dataset is additionnaly augmented by adding reverbation and noise with the Musan corupus.

Our primary system consisted of a score level fusion (linear combination of scores) of RF and SVM classifiers.

Our alternative system was a fusion between the kaldi-based x-vector-PLDA trained with kaldi and a Kaldi-based PLDA trained on the pre-trained model's embeddings.

For the open condition we used two kaldi based x-vectore PLDA systems with score-level fusion.

[30] https://publications.idiap.ch/attachments/papers/2019/Grisard_TSD2019_2019.pdf

[31] https://lre.nist.gov

[32] https://www.nist.gov/system/files/documents/2017/09/29/lre17_eval_plan-2017-09-29_v1.pdf

All systems are evaluated on the NIST'S test split of the LRE22 development set. The same test set is used to tune the fusion weights. The evaluation metric used is actual cost and minimum cost generated by the NIST scoring toolkit.
For the fixed training setup, we got an actual cost of 0.52 using the fusion of RF and SVM, and an actual cost of 0.6 using the alternate system.
For the open training setup, we got an actual cost of 0.6 using the fusion of PLDA with LRE22 dev and PLDA with babel as described above.

9. Speech: Speech Recognition (ASR)

By submitting deliverable D5.2, two principal partners (IDIAP and HENS) supported the ASR technology by providing their models and components for the platforms being developed in the project. Emphasis was placed on the English language (with different models achieving a similar performance on the same test set). However, several other languages were also supported via basic/prototype models, including Greek, Hebrew, German, Albanian and Russian. In line with the decisions given towards selecting the platform (i.e., Autocrime) and open-sourcing all components in it, in February 2022, the Consortium has agreed to continue the ASR work with the models and the component provided by IDIAP. Consequently, among the two innovative methods described in D5.2, the "Boosting of specific words" was favoured instead of the "Vocabulary and language model adaptation". The following subsection describes the results of the work carried out with the latest and improved ASR models and the outcomes of applying the boosting method.

Many research and development activities were made for automatic speech recognition in the ROXANNE project. Practically two independent automatic speech recognition (ASR) approaches have been leveraged from past work: (i) multilingual approach and (ii) XLSR-based approach (see details below).

These two approaches mentioned above are built on three research areas which were pursued in the ROXANNE project: (a) multilingual recognition, (b) iterative learning using partially transcribed data, and (c) information boosting.

9.1 Multilingual speech recognition models

Multilingual acoustic model training combines data from multiple languages to train an automatic speech recognition system. Such a system is beneficial when training data for a target language is limited. Lattice-Free Maximum Mutual Information (LF-MMI) training performs sequence discrimination by introducing competing hypotheses through a denominator graph in the cost function. The standard approach to training a multilingual model with LF-MMI is to combine the acoustic units from all languages and use a standard denominator graph. The resulting model is either used as a feature extractor to train an acoustic model for the target language or directly fine-tuned. A scalable approach to train the multilingual acoustic model is used with a typical multitask network for the LF-MMI framework. A set of language-dependent denominator graphs is used to compute the cost function. The proposed approach is evaluated under typical multilingual ASR tasks using GlobalPhone and BABEL datasets. Relative improvements up to 13.2% in WER are obtained compared to the corresponding monolingual LF-MMI baselines. The implementation is made available as a part of the Kaldi speech recognition toolkit.

Two models are trained with the cost function described above. The first one, the purely Kaldi-based Time Delay Neural Network (TDNN) is built on a multilayer neural network architecture and models context at each layer of the network. A neuron in the TDNN network receives input from activations at the layer below and a pattern of unit output with its context. For time signals, such as speech, each unit receives the activation patterns over time from units below as input.

The second one, built on wav2vec 2.0 is trained by solving a contrastive task over masked latent speech representations and jointly learns a quantization of the latents shared across languages. It is a pretrained model, fine-tuned to the conversational speech recognition.

For both models described above, the acoustic models are bound to the language models during decoding. The acoustic model converts audio into probabilities of characters/words, whereas the language model helps to turn these probabilities into words of a coherent language. The language model assigns probabilities to words and phrases based on statistics from training data, i.e., the more common a phrase is, the higher probability it has when scoring with the language model. All the language models used in Autocrime are the n-gram language models (3- and 4-grams). They estimate the probability of the last word of an 3/4-gram given the previous words.

The XLSR-LFMMI model is developed for both English and German, whereas the TDNN model is available only for English. The lexicon of the English 4-gram model for XLSR consists of more than 1`023`000 words. The German lexicon for the 3-gram model has more than 640`000 words. For the English TDNN acoustic model, the language model is 3-gram and the lexicon consists of only 47`000 words.

The performance of both models is evaluated on the English and German subsets of ROXSD. The XLSR-LFFMI model for English shows the word error rate of 28.4%, whereas for German - 36.2%. The TDNN kaldi-based model has WER equal to 41.9% on English subset of ROXSD.

Boosting of specific (i.e. highly informative) words

This technology aims to significantly improve recognition of highly informative words (or word sequences) identified by police so that automatically transcribed spoken data (e.g. wiretap recordings) will be more beneficial for subsequent NLP tasks. It has already been applied in different domains by partners IDIAP and BUT (see, for instance, the paper on out-of-vocabulary word recognition problems or boosting of contextual information in ASR for air-traffic communication). A similar approach, allowing for the dynamic boosting of highly important words (or word sequences) in ASR transcripts, is also aimed to be applied in ROXANNE. The technology itself is language-independent, although it depends on the availability of ASR (which is typically developed for a particular language). The technology can work in a dynamic mode, i.e. words to be boosted can be dynamically added to a list specified by end-users. Using background knowledge and case-specific context, highly informative words important for investigators, such as names, locations, addresses, and places, can be boosted. ASR systems often misrecognize these words, and boosting them increases the probability of correct recognition, as presented in the table below. The results in the table are obtained by boosting 55 informative monograms, often appearing in ROXSD. The F1-score is calculated by direct transcript analysis, without the NLP module running on top of them.

Language	Accuracy measure	Baseline ASR	Lattice rescoring
English	F1-score (person)	15.4%	20.6%
	F1-score (location)	44.0%	47.9%
	F1-score (time)	78.4%	78.8%
	Average F1-score	**45.9%**	**49.1%**
German	F1-score (person)	43.3%	54.6%
	F1-score (location)	61.7%	61.7%
	F1-score (time)	63.3%	68.5%
	Average F1-score	**56.1%**	**61.6%**

As a further extension, we plan to exploit the boosting technology (applied to ASR) in bi-directional mode with subsequent NLP technologies (specifically for NER followed by the mention network and co-reference resolution).

10. NLP: Topic detection

The topic detection module uses the cross-encoder classification model and solves the sentence-pair regression tasks, such as the semantic textual similarity. This model is, in principal, multilingual and can be used to compute sentence / text embeddings for more than 100 languages. The topic classifier performs a zero-shot classification for given set of labels. The labels are the expected topics of conversations taken as an input to the classifier.

The topic detection module in Autocrime is evaluated on English subset of ROXSD data for six topics the conversations are supposed to be classified: drugs, meeting, money, work conversation, family conversation, other.

The results are presented below for both the manual, ground truth transcriptions and the automatic transcriptions of English XLSR model.

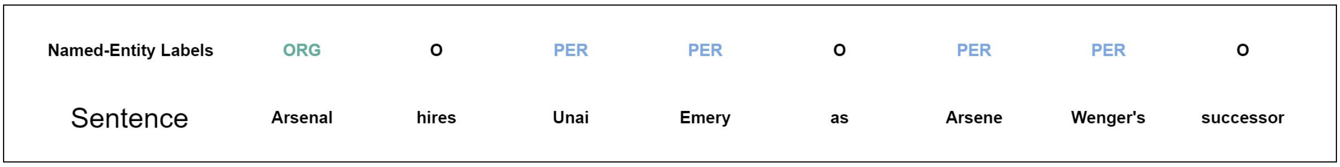
	ground truth transcripts	automatic transcripts
F1-score (drugs)	N/A	N/A
F1-score (meeting)	49%	35%
F1-score (money)	0%	0%
F1-score (work)	37%	22%
F1-score (family)	N/A	N/A
F1-score (other)	39%	34%
Weighted average F1-score	39%	30%
Accuracy	28.7%	21.3%

11. NLP: Named-Entity Recognition

This section introduces the Named-Entity Extraction (NER) module developed during the ROXANNE project. The NER module is one of the key components in the Natural Language Processing (NLP) toolbox in ROXANNE. It aims to automatically extract useful information – the named entities in particular – from documents (e.g., speech transcripts). This should not only assist LEAs to quickly focus on the informative pieces of text from vast amounts of textual data, but also allow to enhance other components like Network Analysis.

11.1 Introduction

A large volume of textual data needs to be investigated during criminal investigations to track and analyze illegal activities. Inspecting documents manually is time-consuming and error-prone. We developed the NER module to accelerate and improve the text reading and comprehension. It automatically detects important entities appearing in the text (e.g., person names, locations, times). A visualization of NER is shown in Figure 8.



Named-Entity Labels	ORG	O	PER	PER	O	PER	PER	O
Sentence	Arsenal	hires	Unai	Emery	as	Arsene	Wenger's	successor

Figure 8: A visualization of the NER task. The named-entities are shown above every word in the sentence. The ORG label stands for organization (Arsenal is an organization). The PER label stands for person. The O label stands for "not an entity".

Often, information-rich pieces of text come with entities. Our NER module will highlight the detected entities in any document (through the AutoCrime platform) so that the LEA investigators can directly jump into informative pieces of text without being distracted by non-relevant text parts. This can significantly accelerate text comprehension during the investigation.

While the NER module can be used as a stand-alone component, we further connect it with other components developed during the ROXANNE project to excavate its potential further. For example, we add detected entities to enrich the communication network between potential criminals hence assisting network analysis.

11.2 Model Design

Detecting the named-entities in a text is not a trivial task for computers, as it is hard to define all entities beforehand. Also, a noun can belong to different types of entities depending on its context. For example, the word “Washington” can be a person’s name or a location. Rule-based systems will often miss entities or assign wrong entity types to the words. Wrong predictions may mislead and slow down investigations, which contradicts our aim. Hence, we leverage contemporary deep neural networks (DNNs) to detect entities quickly and accurately. In particular, we leverage the state-of-the-art transformer-based DNN architecture[33] to pursue the best possible solution.

Currently, we support three types of entities: PERSON, LOCATION and TIME. But our developing pipeline is flexible and can support additional entities on demand.

We use a pre-trained language model (PLM) called RoBERTa[34] as our base component in NER. PLMs like RoBERTA are trained on a vast amount of textual data (160GB text data in our case), so they can model human language very well. We further fine-tuned the RoBERTa model on different NER datasets to specialize it for NER tasks.

We have faced two additional challenges during the NER development in ROXANNE. First, the PLM was trained mostly on formal text. However, in ROXANNE we often process web text or transcripts from phone calls containing mostly informal language. Second, the input text to our NER module is often an ASR’s output which contains errors. We found that a NER module obtained by standard training pipelines is vulnerable to text in the ROXANNE case, so we adjusted the training pipeline. In particular, we trained the model to focus less on the orthography (e.g., spelling, casing, punctuation) and grammar of the text, and more on the high level semantic. This adaptation significantly increases our NER model's performance and outperforms standard NER models by a large margin (up to 30% more entities detected).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin (2017). Attention is All you Need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (pp. 5998–6008).

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

Our final NER module contains roughly 354 million parameters, yet it is very fast at detecting entities in texts. On average, it requires only 0.4 seconds to detect all entities in an entire conversation (about 500 words) in the ROXSD dataset.

11.3 Model Evaluation

We evaluated our NER module on three test scenarios: ROXSD-Phone-Calls, ROXHOOD-Videos and ROXHOOD-Text.

ROXSD-Phone-Calls contain phone calls in different languages. We evaluated our NER module on the transcripts of 164 English phone calls. We evaluated the NER module on both manual transcripts and ASR transcripts. The manual transcripts contain less mistakes through transcription, while ASR transcripts may come with word errors from the ASR modules. The performance on the manual transcripts should offer a fair evaluation of our model as a stand-alone module, while the evaluation on ASR offers the performance indicators of the ASR+NER pipeline.

ROXHOOD-Videos is a test scenario similar to ROXSD-Phone-Calls, where the transcripts come from the video recordings instead of the phone calls.

ROXHOOD-Text contains 299 English posts from the ROXHOOD forum. We evaluated the NER module on ROXHOOD-Text to examine the NER performance on the text data from forums and social media.

A simple data statistic of all three test scenarios can be found in Table 4.

Evaluation Case	ROXSD-Phone-Calls	ROXHOOD-Videos	ROXHOOD-Text
Content	164 phone calls 4856 utterances 1278 entities	23 videos 107 utterances 91 entities	299 posts 110 entities
Entity Distribution	487 PERSON entities 301 LOCATION entities 309 TIME entities	44 PERSON entities 26 LOCATION entities 21 TIME entities	21 PERSON entities 49 LOCATION entities 40 TIME entities

Table 4. Data statistic of the three evaluation scenarios.

We evaluate our NER model using F1 scores. The F1 score is a commonly used evaluation metric for classification tasks. It is the harmonic mean of precision and recall. A higher precision rate indicates that the model predictions are often true, while a higher recall rate means that the model can find more entities appearing in the text. Ideally, a good model should attain both high precision and recall rates. In practice, however, a model with higher recall often comes with some sacrifice on precision and vice versa. Consequently, F1 score is considered to be a reasonable weighting strategy of the two scores. The F1 score lies in the range of 0 to 100; the higher, the better.

The performance of the NER model can be found in Table 5.

Evaluation Case	ROXSD-Phone-Calls	ROXHOOD-Videos	ROXHOOD-Text
Performance (Manual Transcripts)	82.60	92.27	77.76
Performance (ASR Transcripts)	39.92	39.76	Not Applicable
Performance (boosted ASR Transcripts)	43.21	52.44	Not Applicable

Table 5. Model Performance in F1 score under the three evaluation scenarios. The performance on manual transcripts is in general high, while the performance with ASR transcripts has significant drop due to word errors in the transcripts. Still, compared

with standard NER modules, our model achieves a better F1 score on ASR. When the ASR module gets improved, the NER performance on ASR transcripts will increase automatically.

Our NER development pipeline is automatic can can be easily extended to support other languages. As an example, we also developed a German version of our NER module. We showcase its performance in Table 6.

Evaluation Case	ROXSD-Phone-Calls (German)	ROXHOOD-Videos (German)
Performance (Manual Transcripts)	70.19	81.43
Performance (ASR Transcripts)	35.56	42.85
Performance (Boosted ASR Transcripts)	40.53	47.31

Table 6. Model Performance in F1 score on German phone calls from ROXSD. Like the English case, the performance on manual transcripts high, while the performance is lower. When the ASR module gets improved (e.g., when using the boosted version of the ASR transcripts), the NER performance will also increase.

12. NLP: Co-reference Resolution

As discussed in 15, the NER model extracts all persons mentioned in a telephone conversation. However, the persons mentioned in the call can either be third parties (when the speakers talk about a third person not taking part in the call), or it can be one of the parties in the call (this mention usually appears when the speakers greet each other). An important step is therefore to disambiguate these mentions into “Third Party” or “Party” before the Phone Network is modified. We call this “Mention disambiguation”. The disambiguation module is built on co-reference resolution (see below in Section 12.2). This helps the system weed out noisy edges. Apart from extracting third parties, such a module is also helpful to extract the names of parties involved in the call for disambiguation purposes.

12.1 Introduction

Identification of authors (so called authorship attribution) from written text usually involves analysing and mapping the writing style of an individual from one block of text and detecting similar patterns in another span of text. It is intuitable how spoken language can also be mapped in a similar fashion. However, this kind of analysis usually requires a large volume of text authored by an individual. The reader is directed to 18 for more information about these scenarios. A similar kind of analysis has limited application when data is scarce and consequently, we turn to other patterns in conversation style to extract person information. For conversations over the phone in particular, we note that it is generally considered social etiquette to introduce oneself at some point in the conversation. Although this has become less frequent with the advent of caller-ID, it is highly plausible that either parties address each other by name at some point in the conversation. If these “participant mentions” can be isolated from mentions of third parties, this can help determine the speakers in the conversations through the transcripts exclusively. Figure 9. shows a block level diagram of a Mention disambiguation system.

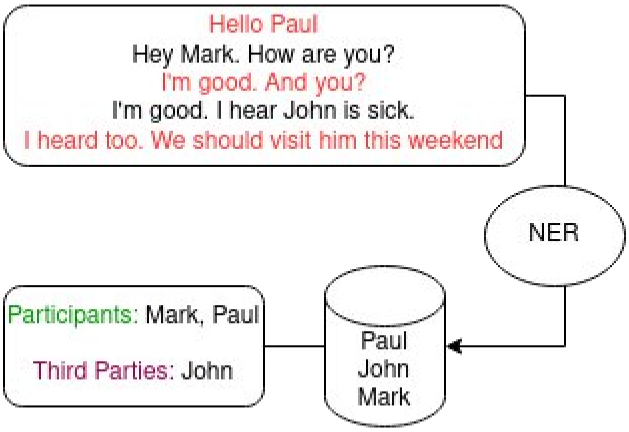
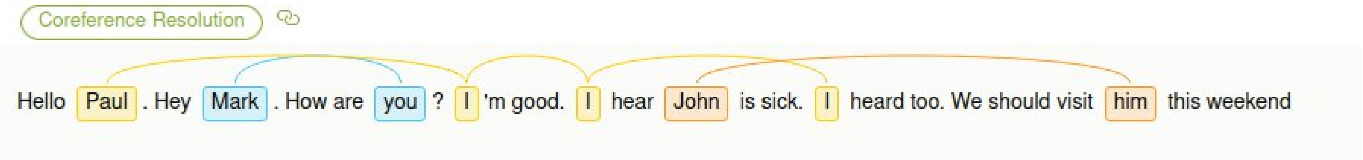


Figure 9: Proposed mention disambiguation system for resolving third parties and parties in the conversation.

12.2 Co-reference Resolution

Co-reference resolution is the task of linking all linguistic expressions (like pronouns and referrals) in a span of text to the original entities that they refer to. This is an important task in Natural Language Processing and is a key component used in the development of complex models that understand written language. The first stage of Co-reference Resolution is usually an NER module. Then, the module looks for pronouns/mentions within the text. When it finds one, it extracts hand-crafted information (i.e closest entity token, position of other entities, surrounding words etc.) from the surrounding context. A neural network is trained to predict co-reference scores for mentions/pronouns and named entities. A simple ranking algorithm is used to find the best matching entity. An example output of a co-reference model is shown in Figure 10.



Hello Paul . Hey Mark . How are you ? I 'm good. I hear John is sick. I heard too. We should visit him this weekend

Figure 10: Example output of a co-reference resolution model on a span of text. All personal pronouns are linked back to their original entities, as can be seen.

In the case of telephone conversations, it is observed that third parties are often referred to with third party pronouns like “him”, “her”, “she”, “he” and so forth. Similarly, participants in the conversation usually are linked to pronouns like “you”, “I”, “my”, etc. Therefore, a co-reference resolution module working on top of the Mention Network can theoretically disambiguate all entities detected into participants and third parties.

As has been shown in Table 4, the English subset of ROXSD contains 487 mentions in overall, with 289 unique utterances (avoiding repetitive utterances in a single call). Among them, 107 are mentions of third parties and 182 are mentions of one of the two parties engaged in the call. Out of a total of 164 phone calls, it is observed that at least one party was addressed by name in 98-calls in the network. This amounts to about 59.7% of calls where at least one participant in the conversation can be identified exclusively with the transcripts. The co-reference model is evaluated against both third parties and participants in the conversation, i.e., it is expected to recognize both classes and segregate the mentions appearing in each conversation transcript. Any entity that links to third party pronouns is treated as a third person and entities that link to first party pronouns are listed as parties in the call.

12.3 Rule-Based and Hybrid approaches

In many conversations in the ROXSD data, it is observed that there are entities in the text that often do not link to any pronoun. An example is provided below:

" Hi, so what? How are you? Can I go with you? I have talked to Andrej, we're going to buy some things... I'm taking a bag, so there will be enough of it. "

The co-reference resolution model consequently fails to detect any reference to the entity, since its performance is exclusively based on the usage of referral expressions. A rule-based model is introduced to tackle such cases where additional pronouns are not used to refer to the original entity. Rule-based models are defined using the position of appearance of the named entity in the course of the conversation. They are described using a threshold *t* as : "If Named Entity *E* occurs before the *t*-th sentence in the conversation, it is considered a participant/party in the conversation." This is based on the intuition that the participants introduce themselves in the initial stage of the conversation and there is decreasing probability that a named entity mentioned in a later stage of the conversation is a party in the call. This intuition is verified by analysing the positions of mentions in the ROXSD dataset. This is plotted in Figure 11. It can clearly be seen that mentions of parties in the call happen early on in telephone conversations.

However, the performance of the rule-based approach is heavily dependant on the threshold defined and is observed to decrease exponentially with the increase of the threshold. Therefore, a hybrid model is created using a combination of rule-based model and co-reference resolution model. The hybrid model emphasizes the rule-based approach towards the initial portion of the conversation and moves onto a co-reference based model in the latter sections.

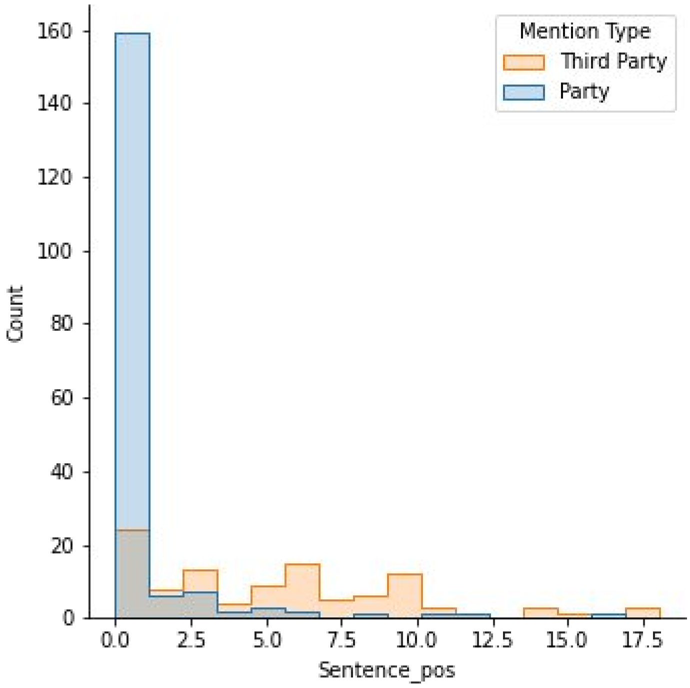


Figure 11: Mention positions in the ROXSD dataset. Mentions are plotted against the position (sentence-wise) of their appearance in the conversation.

12.4 Results

The performance of the mention disambiguation module is measured in terms of its ability to separate mentions. Since it is a downstream task of Named entity recognition, the performance is measured in terms of the Accuracy of Mention Labelling (AML). Specifically, the system performance is measured as the percentage of detected entities that are correctly identified as Parties or Third Parties. Any Person entities missed out by the NER module or in an ASR system cannot be retrieved/processed using the disambiguation module. As such, these entities are avoided when computing the performance metric. The disambiguation model evaluations are tabulated in Table 7: Performance of the Mention disambiguation component on ROXSD. Note that the ASR results are computed for the entities detected from the ASR transcripts..

Data	Rule-Based Model		Hybrid Model	
	AML	Entities Detected	AML	Entities Detected
ROXSD - Manual Transcripts	74.81%	454	80.37%	454
ROXSD - ASR Transcripts	70.66%	75	76.00%	75
ROXSD - Boosted ASR Transcripts	72.63%	95	75.7%	95
ROXSD Videos - Manual Transcripts	90.45%	43	92.30%	43
ROXSD Videos - ASR Transcripts	100%	6	100%	6
ROXSD Videos - Boosted ASR Transcripts	88.23%	17	94.11%	17

Table 7: Performance of the Mention disambiguation component on ROXSD. Note that the ASR results are computed for the entities detected from the ASR transcripts.

Below is a short explanation on achieved results from Table 7:

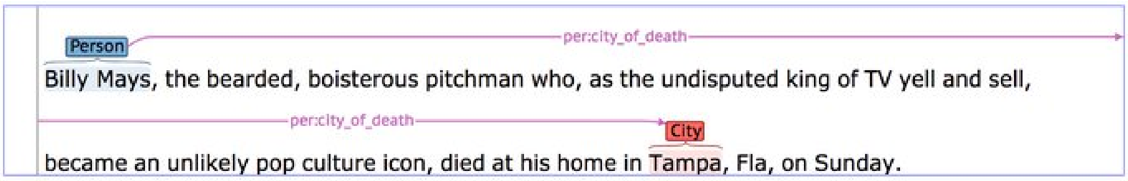
Given there are 100 PER entities in the analysed document, and the ASR/NER pipeline detects 50 correctly, the Mention-module will assign labels (Party/Third) only to the 50 detected entities.
Then the AML score is calculated as AML~=(#correct labels[PER/THIRD])/50. In case we had detected 60 entities in the ASR+NER pipeline, AML is recomputed as AML~=(#correct labels[PER/THIRD])/60 .
The idea here is to convey how good the Mention-module is. The reason we don't include a [ASR+NER+Mention Network] Metric in this section is to keep scored quantized. The mention-network section aims to focus on evaluation of the mention-labeling algorithm. Losses from the ASR+NER pipeline thus do not contribute to the metrics and comparisons made here. Otherwise, the mention-module will always be upper bounded by the ASR+NER performance (i.e the best Mention-model score can only be 50/100 or 60/100 in the examples above). Therefore, when the AML is higher, it does not mean that

more entities are detected. It simply means that the Mention module is doing its job correctly: i.e labelling the detected entities efficiently.

13. NLP: Relation Extraction

Relation Extraction in NLP is generally seen as the task of extracting semantic relationships between tokens within the text. Relation extraction is seen as the natural extension of the Mention detection/NER module. As discussed, the Mention Network module extracts entities from the transcripts from the phone network. Then, it focuses on the people mentioned in the calls and supplements the Phone Network with more edges between parties. It is also intuitive that the conversations contain information about places and organizations, which would be unavailable in the Phone Network. A Relation Extraction component could pick out such additional information from the transcripts.

A Relation Extraction module is an extension of an NER system. While a NER system detects and picks out Named Entities from a text, a Relation Extraction module discovers the relationship between a pair of entities, if one such exists. An example of the same is provided below:



Billy Mays, the bearded, boisterous pitchman who, as the undisputed king of TV yell and sell,
became an unlikely pop culture icon, died at his home in Tampa, Fla, on Sunday.

Figure 12: Relation between Named Entities in a span of text. The Person Entity “Billy Mays” is seen to be related to the Location Entity “Tampa” with the relation —> "city of death”.

Information of this nature is likely to be useful for LEAs. The Relation Extraction module is expected to work on top of the Mention Network described in the previous sections. With the introduction of these two NLP technologies working together, the original network is enriched significantly by the introduction of additional nodes and definition of new edges. A projection of the Relation Extraction module as applied on the example defined in Section 12.2 is depicted in Figure 13. The Relation Extraction module would create a new edge between “Micheal” and “Plaza”, since the conversation revolves around Micheal’s being at the Plaza.

13.1 Model creation and Results

With ROXSD, one important aspect of extracting high level information like relations is defining what relations are of interest. Since ROXSD involves inter-criminal conversations, we assume that LEAs have a general interest in their whereabouts. Specifically, we define two relations in this context:

1. The “Current Location” relation: This relation gives information about the current wherabouts of people or groups. Examples include “I am in Brno”, “We are in Paris today”, “Michael is here in London”.

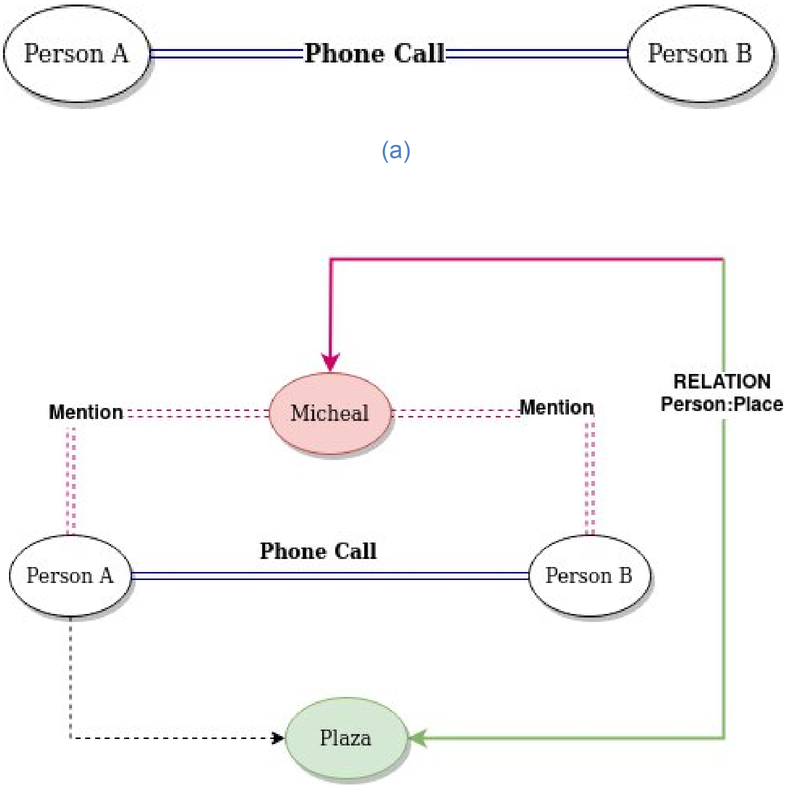
2. The "Movement" relation: This relation tracks the movement of people from one location to another. Specifically, we constrict it to the "going to" cases. Examples include "I'm going to Paris". "Mike is going to London tomorrow"

Our Relation extraction module scans segments of texts and produces Entity-Location annotated with one of the relations described above (or produces empty output if no targeted relation types are detected) . A sample input output pair is:

Text (Input) : I'm in Munich right now.

Relation (Output) : <sub> I <loc> Munich <rel> in

The ROXSD Manual transcripts are found to have about 71 occurrences of such relations. Out of these relations, 52 are "Current location" relations and 19 are "Movement" relations. Our Relation extraction module is seen to detect relations with an average F1 score of 70% on the Manual Transcripts. This drops to 33% for ASR transcripts and 36% for boosted ASR.



(b)

Figure 13: The application of NLP modules on Phone conversations. (a) depicts the original Phone Network and (b) depicts the final network after integration with the Mention Network and Relation Extraction component. It can be observed that a significant amount of additional information is made available through the interaction of the Social Network component and the NLP component.

14. NLP: Authorship attribution (text-based speaker recognition)

14.1 Adding Linguistic Features for Speaker Comparison

State-of-the-art automatic speaker comparison systems rely on voice characteristics to verify if two speech segments originate from the same speaker or not. It has been shown, though, that the inclusion of linguistic features, such as phones and words, improves the performance of automatic speaker identification. This is especially the case if there is a mismatch in recording circumstances or other situations where acoustic features are less identifiable.

We investigate whether linguistic information carried in a conversation can be exploited in forensic speaker comparison casework. Frequent words are particularly attractive as a type of linguistic feature for their statistical independence of voice characteristics and its sensitivity to speaker style, but not to the topic of the conversation. The NFI has developed a method within the likelihood ratio framework for court applicability. This approach takes into account both similarity and typicality by employing a novel method using percentile rank for feature extraction.

In our first experiments, we applied our method on the forensically relevant dataset FRIDA, achieving good results even when a limited amount of data is available. The average telephone conversation in FRIDA consists of approximately 5 minutes or 600 words, yielding an EER of 12 percent. With less speech available, the method is less discriminative, but still achieves an EER of 25% for conversations of 200 words. We are now looking at other datasets for which transcripts are available which are larger and cover different languages,including Fisher[35] and Callhome[36].

Our results are complementary to research on automatic speaker comparison, which concentrates on acoustic features. They can also be combined with other features, such as network information or geolocation. Within the likelihood ratio framework, combining modalities is trivial as long as they are statistically independent. Whether this independence between modalities holds should be subject of future research.

So far, our use-case focused on speaker verification. This task is highly demanding with respect to the quality and calibration of the outcome, but it requires transcripts to be made hence mostly applicable in small scale settings. However, results are promising and the technique itself is scalable. A further step should be the application of our method to large scale investigations, in combination with automatic speech recognition.

15. Video Processing

15.1 Introduction

Video is another modality in which speech, NLP, image and network analysis can be combined to support investigator needs. One typical pain point stressed by LEAs is the difficulty to process large volumes of video or image data resulting from seized phones or computers during an investigation. A typical scenario consists in identifying related documents (images, videos) across multiple devices (phones, computers) to discover a potential link between these devices or their owners. ROXANNE partners have investigated the interest of automatic image and video processing technologies to enrich speaker or device networks

[35] https://catalog.ldc.upenn.edu/LDC2004T19

[36] https://catalog.ldc.upenn.edu/LDC97S42

with additional edges and nodes to support investigations. Identifying related visual documents without any support would require investigators to watch all images and videos, most of them being irrelevant, with a high probability to miss the related documents.

To support these needs, ROXANNE partners developed and integrated the following technologies in the ROXANNE Autocrime platform:

- Face automatic detection, clustering, face cluster summarization and face similarity evaluation
- Scene or object characterization, clustering, scene cluster summarization and scene similarity evaluation

The following sections describe the technologies that were developed and evaluated. We then detail how they are used in practice to enrich an initial speaker network built from tapped calls.

15.2 Face characterization technology

To take into account the Roxanne Ethics board recommendations in terms of facial analytics dangers and the necessity to protect personal biometric information, Autocrime platform uses face related technologies with parsimony.

Firstly we did not train any new model but used an existing and open source state of the art face detection and embedding extraction model[37] as a basis for face characterization. The selected model is a Pytorch implementation of the Arface[38] paper, trained by the authors on the MS-Celeb-1M dataset consisting of more than 3M faces from more than 85k identities and achieves around 98% of accuracy on the MegaFace[39] dataset consisting of 1M faces from 600k identities.

Secondly, all the images or videos observing faces which are included in the ROXSDV3 dataset or were used for evaluation or demonstration purposes in Roxanne were pseudonymized using face swapping technologies. In practice, we adapted the SimSwap framework[40] that we used to pseudonymize a subset of selfie videos from the “Realtime Selfie Video Stabilization” dataset[41]. The SimSwap framework takes as input videos in which the faces have to be swapped and pictures of new faces to be inserted in the videos. In practice, we took as new faces, fake faces generated by the StyleGAN2 model[42] thus resulting in videos observing people who do not exist.

Finally, face matching is only performed between a limited set of manually enrolled face pictures corresponding to main suspects or victims which are then searched in all ingested images and videos. This allows to limit the usage of facial technology only on persons related to the case (victims, suspects), thus answering to the need to only use face technology in a proportionate manner, as recommended by our internal and external ethics boards. In particular, the complete search of all face matches across all ingested images or videos (independently of specific face enrolments) is not supported.

To check that our image and video pseudonymization process we evaluated the cosine similarities between the following sets of face embeddings:

[37] https://github.com/foamliu/InsightFace-PyTorch

[38] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

[39] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, E. Brossard, “The MegaFace Benchmark: 1 million faces for recognition at scale”, *2016 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[40] https://github.com/neuralchen/SimSwap

[41] https://github.com/jiy173/selfievideostabilization

[42] https://github.com/NVlabs/stylegan2

- As a reference, between matching and non matching original faces in original selfie videos (ensuring that query and matching faces are always from different videos)
- To check the face swapping consistency, between matching and non matching swapped faces in pseudonymized videos
- To check the pseudonymization process, between original and corresponding swapped / pseudonymized faces

The results, evaluated on a total of 137 short selfie video clips are given in the following figure.

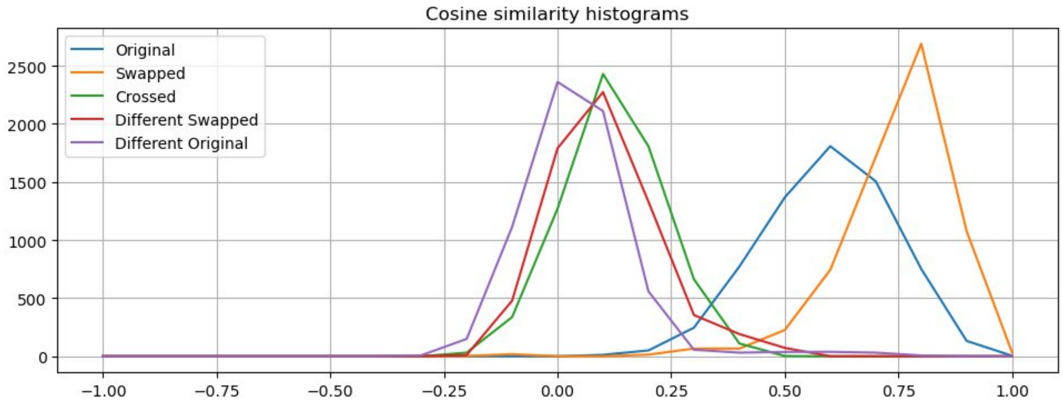


Figure 14. Cosine similarity histograms between matching original faces (blue), matching swapped faces (orange), corresponding original and swapped faces (green), different original faces (purple), different swapped faces (red)

These histograms show that the images were correctly pseudonymized (green curve representing the cosine similarities between original and swapped faces almost aligned with red and purple curves of cosine similarities between non matching swapped and original faces).

They also show that the face swapping is consistent from one video to another: the orange histogram shows good similarity measures between swapped faces of a same identity in different videos. The sharper histogram of cosine similarities between matching swapped faces than between original faces denotes some probable loss of observation diversity after the pseudonymization process. The resulting histogram still exhibits some diversity that could be representative of a more effective face recognition system thus still being valid for our needs.

In order to optimize the matching performances we developed an additional face clustering and face cluster summarization algorithm enabling to summarize each input video in a set of “identities”, each identity being described by their K most relevant face observations. In practice, we use the DBSCAN[43] clustering algorithm (Density Based Spatial Clustering for Applications with Noise) which is well suited to gather in a same cluster the various poses of a same face observed in a video sequence. DBSCAN iteratively gathers sample points in clusters through an incremental search for nearest neighbours.

This behaviour enables DBSCAN to gather in a same cluster the successive observations of a rotating face as illustrated at the following picture.

[43] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining, 1996

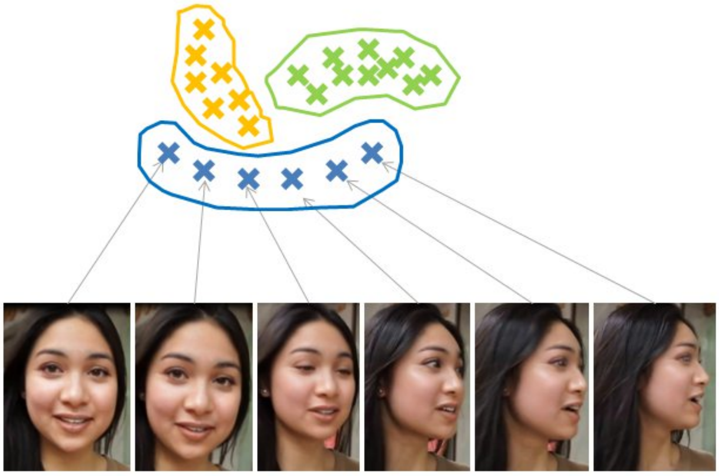


Figure 15. Clusters created by DBSCAN (pseudonymized video)

If the video processing rate is not too low, successive face observations will be very close in the embedding space, and thus added to the same cluster by DBSCAN, even if the head pose between the first image of the sequence and the last image of the sequence differ significantly as shown in the previous figure.

In practice, a video processing frame rate of 5 FPS was found as sufficient to enable a good clustering of faces despite head rotations. At this frame rate however, the resulting cluster may consist of hundreds of observations for each person observed longer than a few seconds. This is why we also put in place a cluster summarization consisting in selecting, among all the observations gathered in a same face cluster, the five most useful representations for the subsequent face matching phase. For this, we first select the face picture and corresponding embedding the closest to the cluster's centroid and then iteratively choose the four furthest embeddings from already selected embeddings. This way, we end up with five most dissimilar face observations for a specific identity, which can be advantageously used for finding this identity in other documents, with more robustness to varying head poses.

The following figure illustrates the typical five face representatives found on the selfie video from which the previous sequence was extracted.

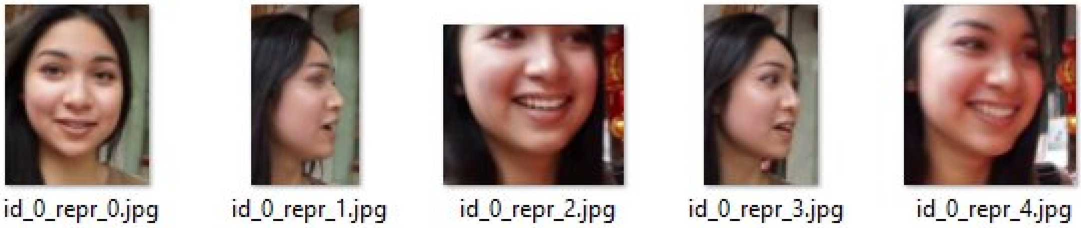


Figure 16. Automatically extracted five representative observations of one identity cluster found in a pseudonymized selfie video.

To evaluate the integrated face characterization module with representative videos we used as test dataset the 137 pseudonymized selfie videos complemented by a few additional images and videos specifically built for demonstration purposes and linked to the ROXSDV3 story. These videos are representative of video chat posts or conversations that could be found on a seized phone.

For each of the 47 pseudonymized persons found in the Selfie video dataset, we selected the first video as source for face enrolment and the one or two remaining videos for ingestion and search. We ran our face cluster summarization process on the first videos and selected the second representative face as enrolment face to make the face recognition task harder. The second representative face corresponds for a given cluster, to the face observation the furthest away from the cluster centroid:

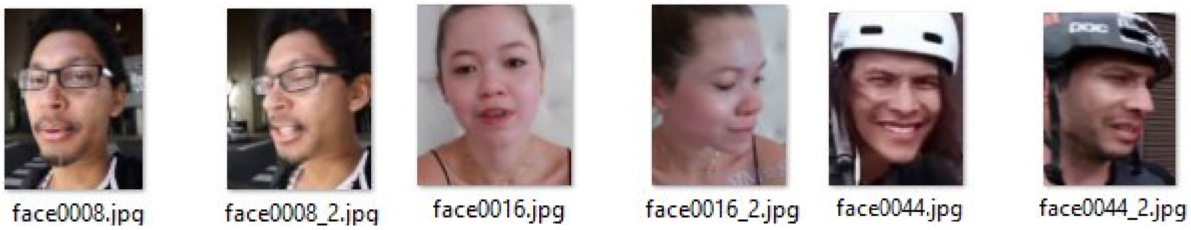


Figure 17. Examples of face observations selected for enrolment: for each person, the first picture corresponds to the cluster centroid, the second picture to the face observation, in the same cluster, the furthest away from the centroid. This is the one selected for enrolment.

Finally, we also evaluated the impact of heavy compression on face recognition performances. For that we re-encoded all videos with the mpeg4 codec and re-evaluated the performances (with the same not compressed enrolment images).

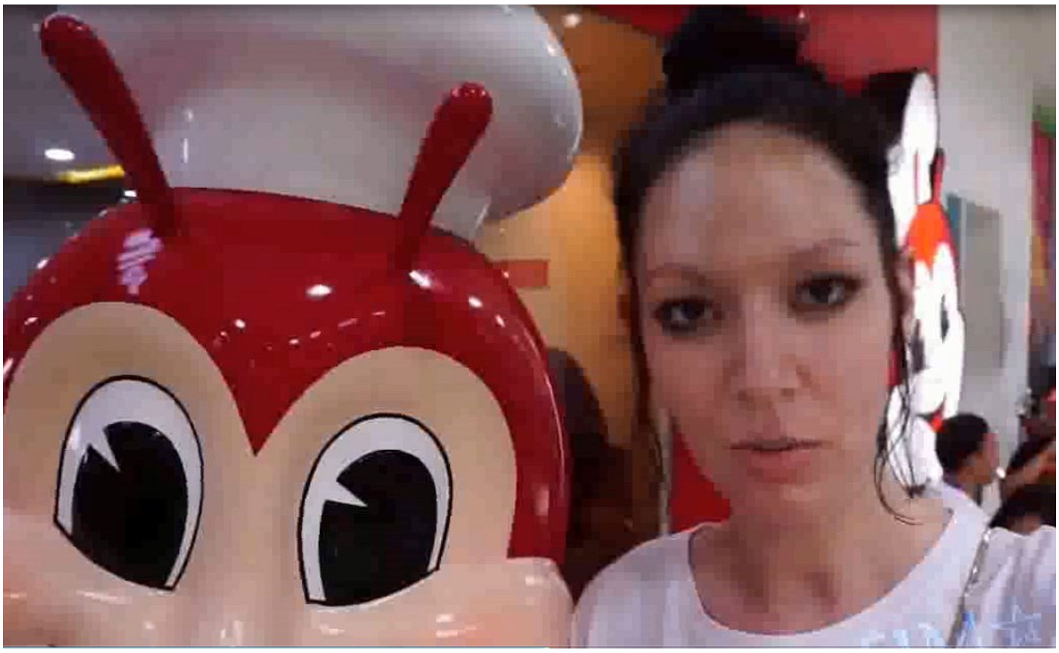


Figure 18. Example of heavily compressed video

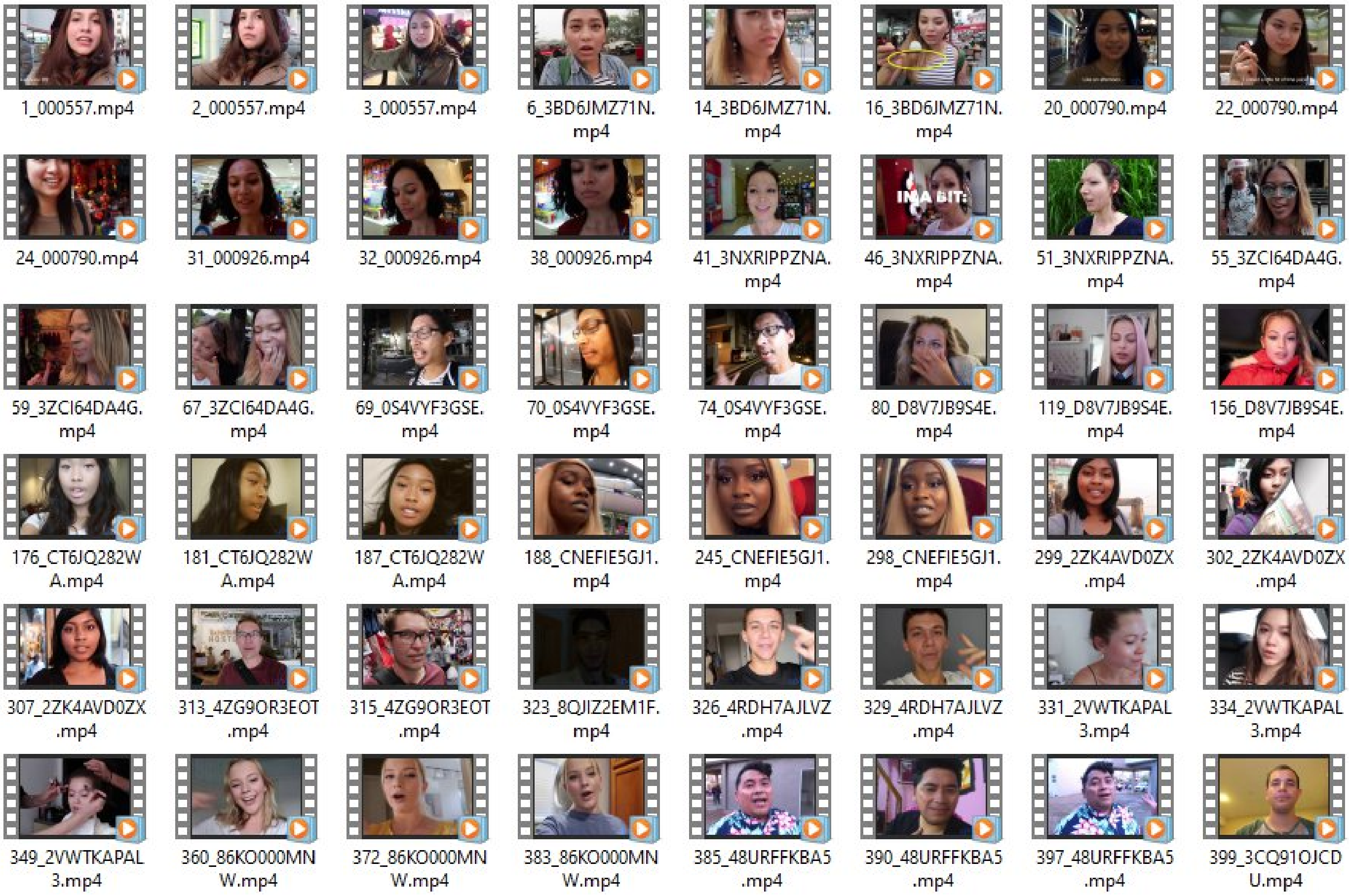
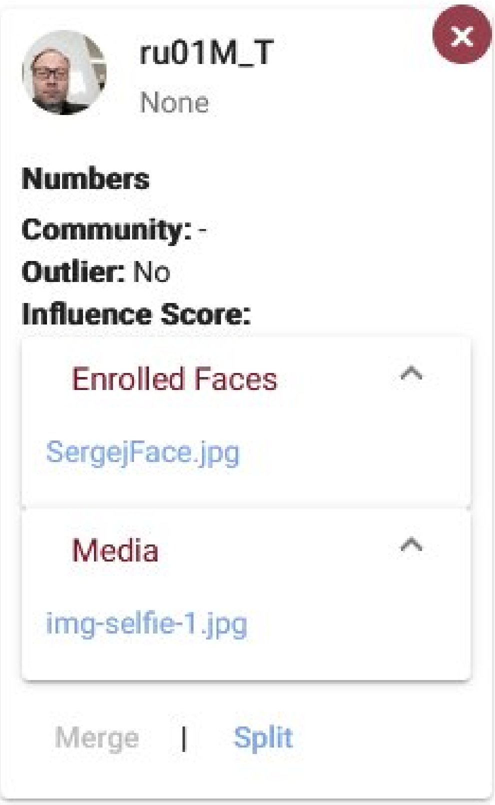


Figure 19. Snapshots of a subset of pseudonymized videos (naming convention: <nb-of-original-video>_<fake-face-id>.mp4)

We thus enrolled 47 face pictures from the Selfie Video dataset + 4 additional faces from ROXSD scenario (Kristina, Marko, Sergej, Samuel). The enrolled face pictures are always taken from other videos or images than the videos or images used for the test dataset. Enrolling a face in Autocrime means associating one or several face pictures to a speaker node (for instance the node of the main suspect whose phone is tapped and who is known from the police). The face may also be enrolled to a not yet existing speaker node that will then be created in the network (e.g. node of the main victim who is never heard in the calls but could be observed in some images or videos).

We then evaluate the ability of Autocrime platform to automatically associate all the images or videos observing enrolled persons to each node in which enrolled face pictures were declared:



ru01M_T

None

Numbers

Community: -

Outlier: No

Influence Score:

Enrolled Faces

SergejFace.jpg

Media

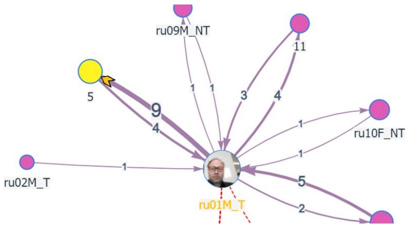
img-selfie-1.jpg

Merge | Split

Example of face enrolment in Autocrime

The “Enrolled Faces” widget lists the face pictures which were manually enrolled for the speaker node “ru01M_T”. They should obviously all correspond to the same person. These are the face observations which will be searched in all other ingested images or videos to try to find “ru01M_T” in other documents.

If at least one face picture has been enrolled in a speaker node, the first enrolled face picture is used as icon for this node in the network:



Images or videos in which a face match has been found with one of the enrolled pictures are associated to this node (added in the “Media” list of the node, here img-selfie-1.jpg):

The performances are evaluated in terms of document association recall (detection rate) and precision:

- **Recall**: Among all image or video files that observed at least one enrolled person and should thus have been associated to speaker nodes, percentage of images or videos that were indeed associated to their right node
- **Precision**: Among all image and video associations to nodes that were automatically found by Autocrime ingestion process, percentage of those that were correctly associated to the right nodes

	Original compression		Heavily compressed	
Document association recall from faces	98%	96/98	93%	93/98
Document association precision from faces	100%	96/96	100%	93/93

These performances were obtained with a cosine similarity threshold of 0.6. This parameter can be adjusted in Autocrime configuration file (parameter tech.video.face_similarity_threshold) to favour either high detection rates (lower threshold) or high precision rates (higher threshold).

It should be noted that these good performances may not be representative of the real performances obtained on real data. Indeed several parameters tend to contribute to the good performances related here:

- Only faces which are sufficiently large (70 pixels) are taken into account
- The number of different faces for the evaluation is limited (51 faces)
- The pseudonymization process tends to reduce the variability of the face observations as witnessed by the increased similarity score displayed in Figure 14

In any case, face similarity results shall not be taken for granted and shall always be checked and validated or invalidated by the analyst.

15.3 Scene characterization technology

Location is another key parameter in investigations. This is why we included a scene characterization and matching capability consisting in describing the whole image with a signature (embedding) whose objective is to encode the discriminative visual features which makes a given scene different from another one.

In practice, we use AIRBUS place embedding training pipeline based on a ResNet backbone and an ArcFace[44] module originally designed for face recognition but successfully applied here to scene embedding extraction. The embedding extraction pipeline is trained with a loss function aiming at grouping together embeddings extracted on different observations of a same location, while pushing away from each other embeddings extracted on observations of different locations. The embedding pipeline was trained on the Google Landmark Dataset[45] consisting - after cleaning - of 1.8 million images and 120k classes (i.e. different locations).

The corresponding model achieved state of the art performances on standard place retrieval datasets such as Oxford5k, Paris6k, Revisited Oxford and Revisited Paris datasets[46].

It was first qualitatively tested on video sequences extracted from CSI TV show to assess its ability to retrieve scenes observed at different times or across episodes as illustrated below.

[44] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[45] T. Weyand*, A. Araujo*, B. Cao, J. Sim, “Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval” , Proc. CVPR'20

[46] Radenović F., Iscen A., Tolias G., Avrithis Y., Chum O. , “Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking”, CVPR 2018

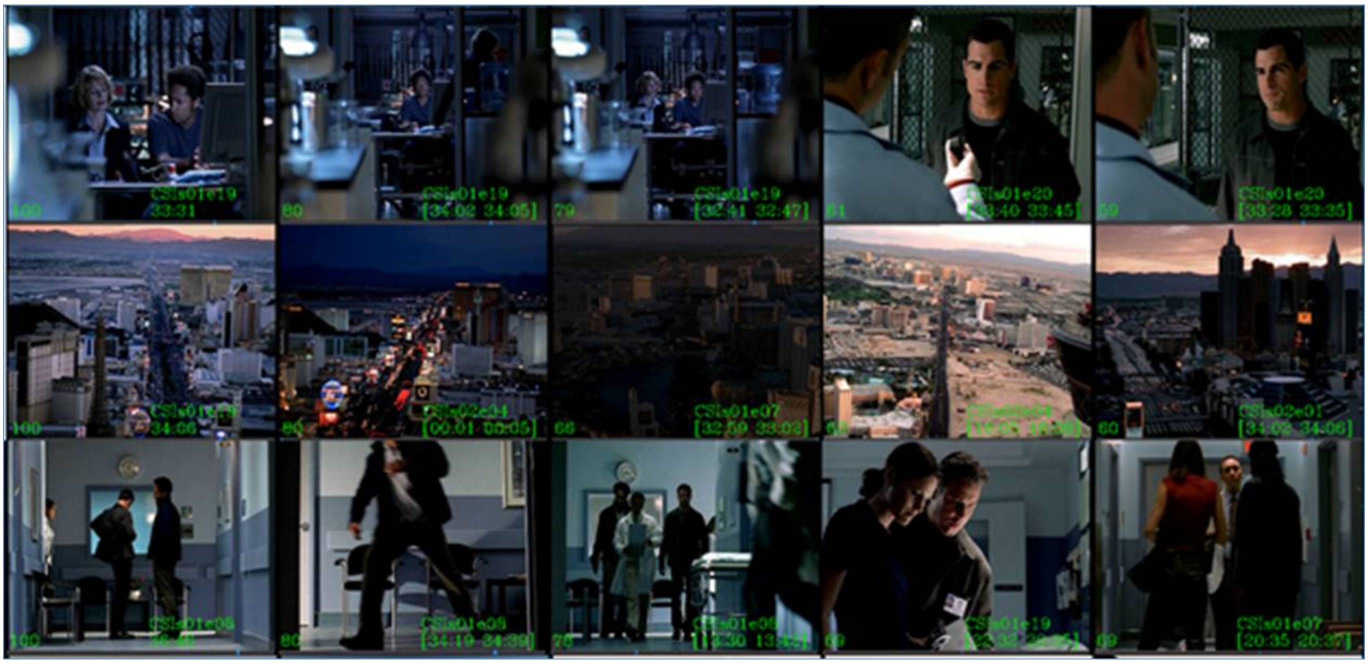


Figure 20. Examples of scenes retrieved at a different time in a same episode or across episodes (first column is the query, other columns are the retrieved shots)

To evaluate our scene characterization and matching module on more representative videos that could be found on a seized smartphone, we captured 154 videos from 15 different smartphone devices observing 8 different locations. These videos additionally include the voices of characters from the ROXSDV3 dataset enabling their multi-modal exploitation (voices and places).

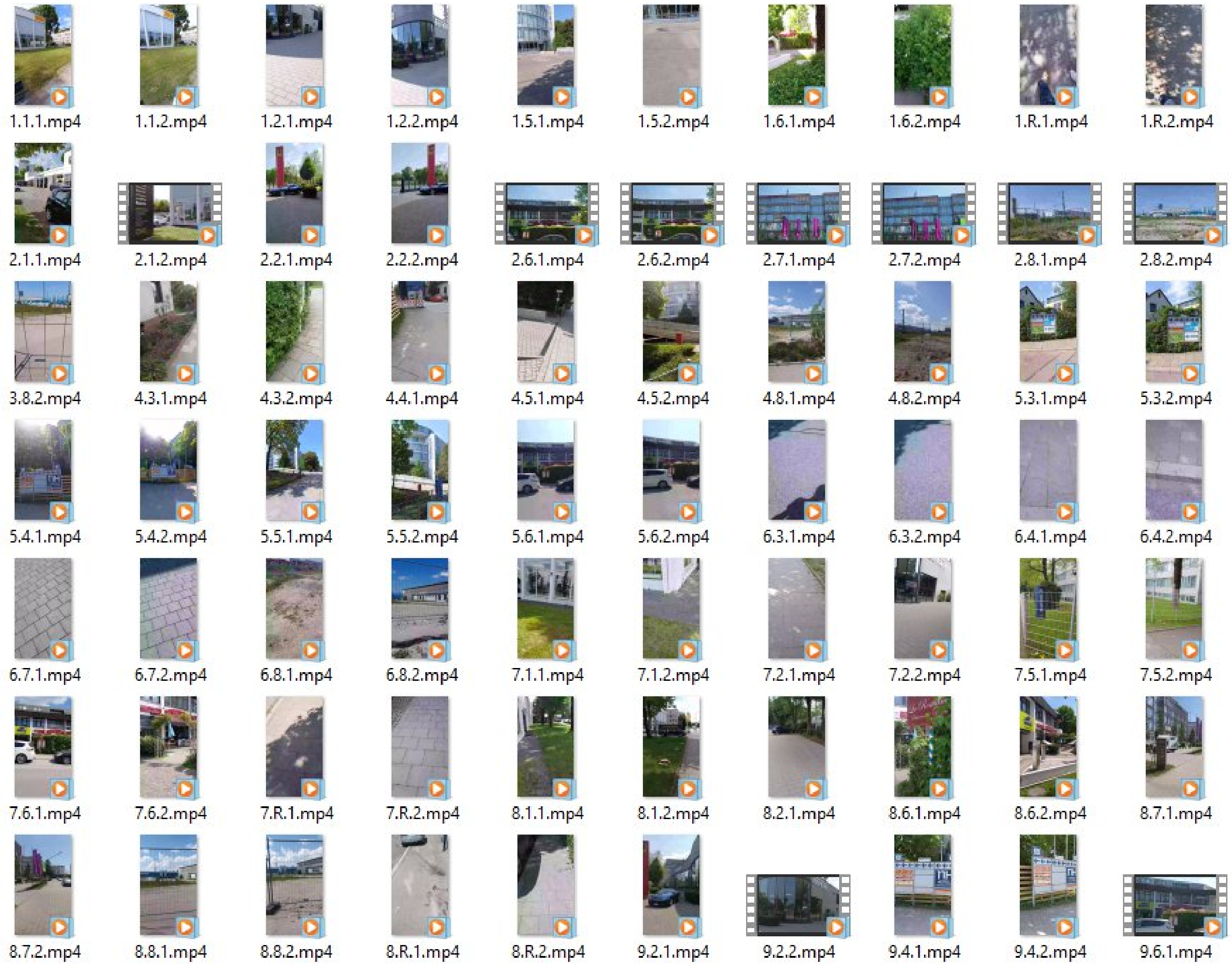
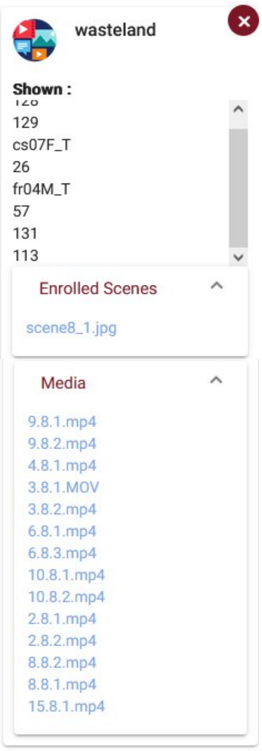


Figure 21. Snapshots of a subset of scene videos (naming convention: <speakerId_locationId_videoId>.mp4)

Scenes are used in a similar way as faces in the Autocrime platform: pictures of locations or objects of interest can be enrolled either in an existing speaker node (meaning that this location is related to this person), or in a new “media” type node which will be automatically added to the network. The “media” nodes can be used to automatically find links between speakers and this location even if no prior knowledge allows to associate this location to any speaker.

Each time an enrolled scene is found in an image or a video, the image or video name is associated with the node containing the enrolled scene (file name added in its “Media” list):


wasteland

Shown :

128
129
cs07F_T
26
fr04M_T
57
131
113

Enrolled Scenes

scene8_1.jpg

Media

9.8.1.mp4
9.8.2.mp4
4.8.1.mp4
3.8.1.MOV
3.8.2.mp4
6.8.1.mp4
6.8.3.mp4
10.8.1.mp4
10.8.2.mp4
2.8.1.mp4
2.8.2.mp4
8.8.2.mp4
8.8.1.mp4
15.8.1.mp4

<u>Example of scene enrolment in Autocrime</u>

The “Enrolled Scenes” widget lists the scene pictures which were manually enrolled for a new created node of type “Media” called “wasteland”. They should obviously all correspond to the same location. These are the scene observations which will be searched in all other ingested images or videos to try to find this location in other documents.

If at least one scene picture has been enrolled in a speaker or media node, the first enrolled scene picture is used as icon for this node in the network:

wasteland
2.

Images or videos in which a scene match has been found with one of the enrolled pictures are associated to this node (added in the “Media” list of the node, here videos *.8.*.mp4).

The performances are evaluated in terms of document association recall (detection rate) and precision:

- **<u>Recall</u>**: Among all image or video files that observed at least one enrolled location and should thus have been associated to a speaker or media node, percentage of images or videos that were indeed associated to their right node
- **<u>Precision</u>**: Among all image and video associations to nodes that were automatically found by Autocrime ingestion process, percentage of those that were correctly associated to the right nodes

As for face recognition evaluation, we evaluated the performances on the original compression setting and an additional heavy compression setting (reducing by almost a factor of 2 file sizes).

	Original compression		Heavily compressed	
Document association recall from scenes	70%	105/150	62%	93/150
Document association precision from scenes	87%	105/121	90%	93/103

These performances were obtained with a cosine similarity threshold of 0.7. This parameter can be adjusted in Autocrime configuration file (parameter tech.video.scene_similarity_threshold) to favour either high detection rates (lower threshold) or high precision (higher threshold).

15.4 Speaker network refinement from image and video processing

As explained in the previous sections, face and scene pictures can be enrolled either on existing or new speaker nodes of Roxanne speaker network, or, for scene pictures, on independent “media” nodes. Enrolled faces and scenes are then automatically searched in all ingested images and videos whose set of summarized face and scene embeddings are extracted at document ingestion. If the cosine similarity between an enrolled picture and summarized embeddings is higher than the thresholds defined in

Autrocrime configuration file, a match is found and the corresponding image or video is added in the "Media" list of the corresponding nodes.

Documents can then be visualized in Roxanne graphical interface thus already providing a beneficial support to investigators who have directly access, in a given node details view, to all visual documents related in some way to this person. This constitutes an advantageous focus of attention preventing the user from having to watch a large set of irrelevant videos before finding the ones that may be of interest.

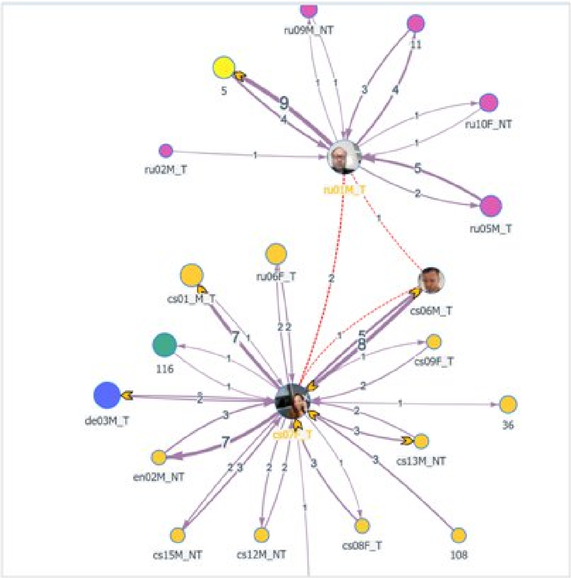
A visual document (image or video) can be associated to a "speaker" type node for four reasons:

- Either because it observes at some point a face corresponding to enrolled faces in this node (automatic association)
- Either because it observes at some point a scene corresponding to enrolled scenes in this node (automatic association)
- Either because the speaker's voice is heard at some point in the video. Note that unlike for faces and scenes, the speaker does not need to be enrolled (in the sense his name does not need to be known) for this association to take place: speaker clustering is performed on the audio track of videos as it is done on tapped calls (automatic association)
- Either because the image or video was manually associated to this node (for instance because this image or video was found on this person's phone) (manual association)

A visual document (image or video) can be associated to a "media" type node for two reasons:

- Either because it observes at some point a scene corresponding to enrolled scenes in this node (automatic association)
- Either because the image or video was manually associated to this node (for instance to group together all images and videos seized on a given website) (manual association)

From this association logic, a same image or video may be manually or automatically associated to several "speaker" or "media" nodes. For instance an image in which two enrolled persons are recognized, will be associated to the two corresponding nodes. Or a video in which a speaker is heard, an enrolled location is observed, and a third person's face is recognized, will be associated to the three corresponding nodes in Roxanne speaker network. This allows to automatically generate new "image" type edges between nodes sharing common associated visual documents. The edge name is then the visual document relating the two nodes.



Examples of added "image" type edges

In the example on the left, three additional "image" edges were automatically added after image and video ingestion (displayed in red):

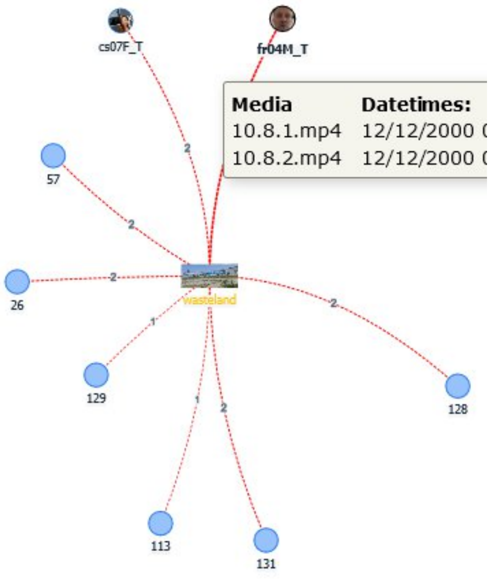
- The edge between cs07F_T and ru01M_T results from their both faces detected in a common image img-selfie-1.jpg:

- The edges between cs06M_T and ru01M_T and between cs06M_T and cs07M_T result from the fact that the image img-selfie-1.jpg was found on cs06M_T's phone and thus manually associated to this node at ingestion time. As this image is found in the three nodes 'Media' list, edges between these three nodes are created.

In the example above, an unknown link between characters cs07F_T and ru01M_T who never called each other in the available tapped calls, is discovered thanks to a picture involving both characters found on cs06M_T's seized phone.

Edges can also be automatically created between a "media" type node representing a location of interest and speaker nodes:

Examples of added "image" type edges



Media	Datetimes:
10.8.1.mp4	12/12/2000 (
10.8.2.mp4	12/12/2000 (

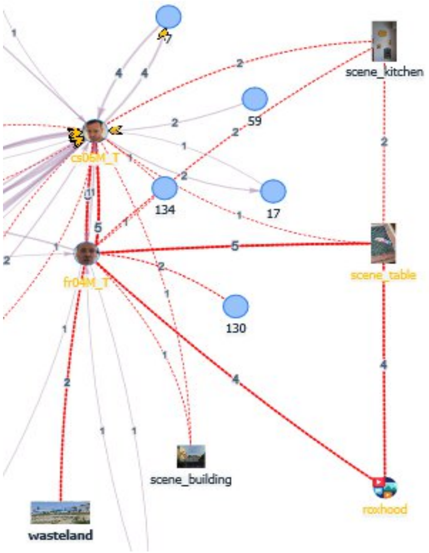
In this other example, a location of interest was enrolled in a "media" type edge ("wasteland")

New edges were automatically created between the node containing the enrolled scene ("wasteland") and speaker nodes whose voice has been recognized in videos observing the enrolled scene.

This is for instance the case of node fr04M_T whose voice is heard in videos 10.8.1.mp4 and 10.8.2.mp4 which indeed observe the "wasteland" area. When looking at all other found links, we note that they all correspond to files ?.8.?.mp4, which all observe scene 8 which is indeed the wasteland area.

This capability allows investigators to rapidly find suspect speakers or victims somehow related to a given location.

Finally, edges can be automatically created between a "media" type node representing a source of data (e.g. a website) and speaker nodes:



Examples of added "image" type edges

In this other example, all images and videos grabbed on the ROXHOOD darknet website were manually associated to the "roxhood" node (bottom right).

New edges were automatically created between the node representing this data source ("roxhood") and speaker nodes whose voice has been recognized in these videos (namely fr04M_T) enabling to identify roxhood user "Roxbinhood" as fr04M_T. Besides, the enrolled picture of the table on which Roxbinhood advertises his pills is also found in a video found on cs06M_T phone, thus making an indirect link between cs06M_T and the Roxhood forum.

This capability allows investigators to rapidly find suspect speakers or victims somehow related to a given data source independent from a given person (e.g. a website).

As a conclusion, integrated video technologies enable the exploitation of additional images or videos that may be found during an investigation. Based on enrolled scenes or faces, images or videos potentially relevant to the case are automatically added in the "media" lists of speaker nodes of suspects and/or victims, thus accelerating the discovery of potential new information.

15.5 Multi-modal query

This section is to describe how the Multimodal Query, developed by ITML, can be combined or stand as additional element to the workflow of Video Processing technologies and more specific to the (a) Face automatic detection, clustering, face cluster summarization and face similarity evaluation and (b) Scene or object characterization, clustering, scene cluster summarization and scene similarity evaluation, as described in the previous subsections. More details regarding the technical description of the Multimodal Query are included in the relevant deliverable of *WP7 – "Integration and visualization of results*", which is the *D7.5* – *"Data visualization V2, ROXANNE platform V2"*. This component stands as an outcome of and is also directly linked to the activities of *T6.1 - "Fusion of information from component technologies for network analysis".* In this task the goal is the aggregation of data coming from the WP5 analysis components. In our use case, the component is the Video Processing from AIRBUS and the connections and relations was examined through the Multi-modal query.

Due to the fact that Video Processing complements existing Speech and NLP technologies, for instance when deriving network graph(s), there is ground to work and provide the multimodal retrieval capability. This is because Autocrime platform has additional processing capabilities on additional document types such as videos. When the number of videos and images (relevant and non-relevant to a specific case) are very large, then a tool or a technology to support LEAs, in order to accelerate the process of the identification of relevant documents in an ethical way is a must. The Multimodal Query offers the capability of multimodal document similarity evaluation. In this approach, the added value coming from Video Processing, which is the enhancement of ROXSD dataset and existing network based on phone calls is being combined with the added value coming from the Multimodal Query, which is to search for specific modalities such as voice, scene, face and any combination of them. To deliver such queries the approach is: (a) to ingest embeddings of one or several entities; (b) to retrieve the info related to face/scene; face/scene uid; the max N main face/scene clusters found in document <doc_id> and (c) to submit a ranking list of documents characterized as relevant, via a ranking strategy. For this ranking strategy the steps are: [1] to retrieve the N most similar embeddings., [2] to eliminate retrieved embeddings below threshold, [3] to score the relevant documents. The multimodal retrieval capability is being demonstrated over the modalities mentioned below through a process starting with the selection of a number of retrieved documents and the type of retrieved documents. The following step was to adjust the importance of given modalities via a ed optimization technique and finally offer the ranking strategy which includes: [1] how to eliminate (per modality) documents below threshold, [2] how to get for each modality a score, and [3] how to get the overall score.

The requirements for this Indexing and Retrieval technology regarding the Ingestion mechanism were that:

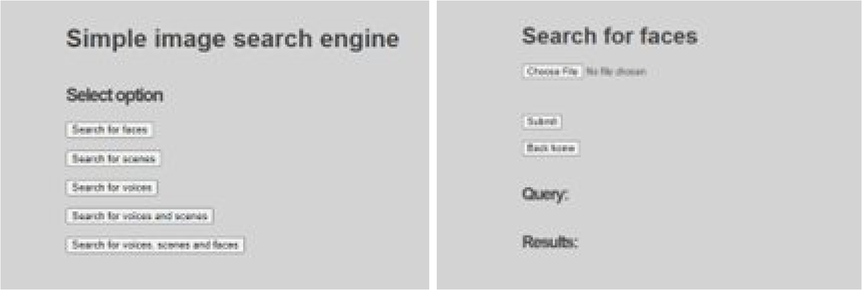
- This component to index all the embeddings characterizing the person and place entities found in a document - (voice embeddings), images (face and or scene embeddings) and videos (voice, face and or scene embeddings).
- This component to associate each indexed embedding with the document in which it appears, and, when available, at which list of timestamps.
- This component to keep the correspondence between each face embedding and the main representative picture of this face cluster to be able to serve it on request.

- This component to keep the correspondence between each scene embedding and the main representative picture of this scene cluster to be able to serve it on request.

while the requirements for the Query itself where that:

- The component can be used to list documents (calls, videos, images) most similar to a query.
- The component to enable to make queries on each modality individually.
- The component to enable multi-modal queries on a user defined list of modalities among voice/face/scene.

For the single modality query, either searching for faces or scenes a dynamic search engine interface has been prepared, as shown in Figure 21.



Simple image search engine

Select option

Search for faces
Search for scenes
Search for voices
Search for voices and scenes
Search for voices, scenes and faces

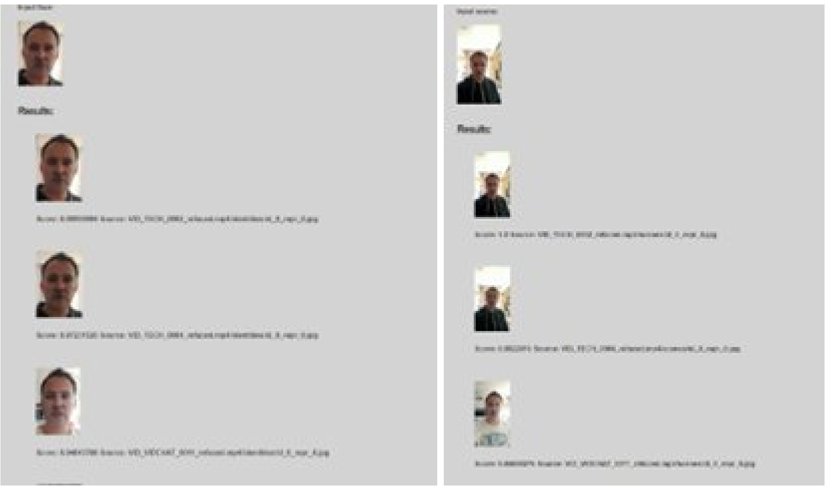
Search for faces

Query:

Results:

Figure 22. Dedicated endpoint to choose from a list of options to proceed with the search process; Search for faces

The results of each query, including the relevant document and the extracted score is shown in Figure 22.

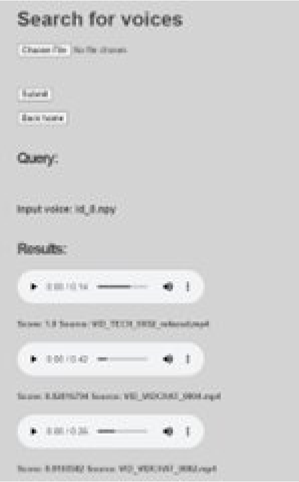


Results:

Results:

Figure 23. Dedicated endpoint for the results of the query “Search for faces”

Similarly, for the query prepared only for Voice modality the dynamic search engine interface and the results of the query (document and score) along with the play/save button are shown in Figure 23.



Search for voices

Query:

Input voice: id_8.npy

Results:

Score: 1.0 Source: VID_TECH_[illegible].mp4

Score: [illegible] Source: VID_[illegible].mp4

Score: [illegible] Source: VID_[illegible].mp4

Figure 24. Dedicated endpoint for the query “Search for voices” and Results

Finally, when it comes to the multi-modal query, where voices, scenes and faces are ingested through the dedicated dynamic search interface, the user may enter a specific audio file, scene file and face file, and enter the voice, scene, face weights. Then search engine executes the query and provides a list of results as shown in Figure 24.

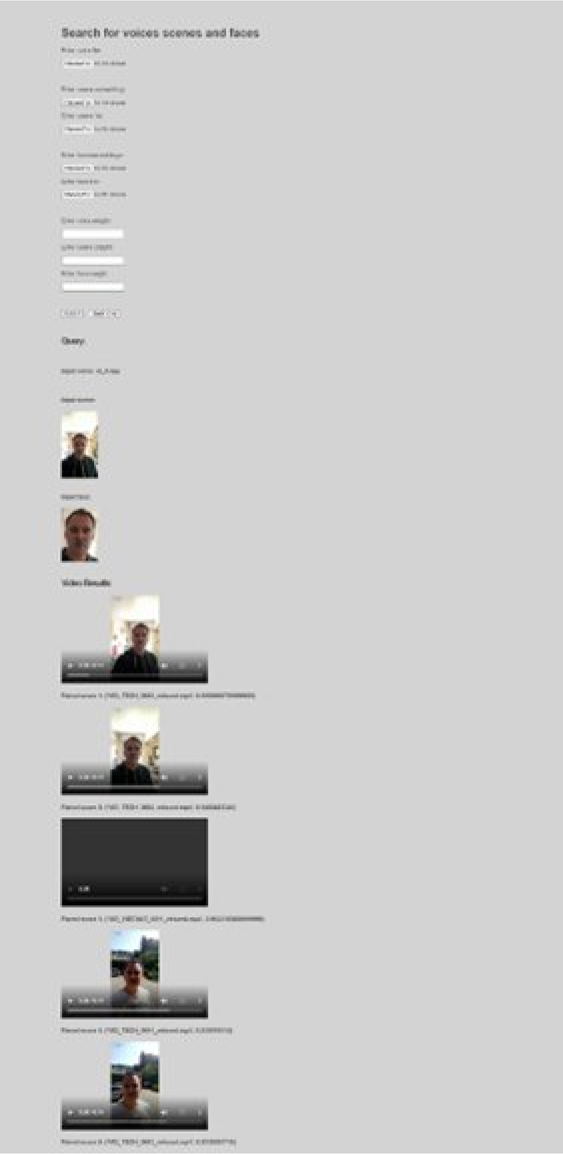


Figure 25. Dedicated endpoint for the query “Search for voices, scenes and faces” and Results

For a list of indexed documents (e.g. vid1, vid2, … vidN) three (3) lists of embeddings are being prepared and ingested to the Multi-modal Query: voice embeddings, face embeddings and scene embeddings. The Multi-modal Query is handling these lists of embeddings through three (3) indices: a voice index, a face index and a scene index. Regarding the Voice embedding results, these are sorted by decreasing similarity in a table which includes the Embedding ID, the degree of similarity (max 1), the Source Video ID and the Video voice similarity scores (threshold 0.7). An identical workflow is being delivered for the scenes where inputs are the Embedding ID and Source Video ID while the outputs are the degree of Similarity (max 1) and Video scene similarity scores (threshold 0.5). The following step is the Score Fusion (voice weight: 1.0, scene weight: 2.0). Finally, the final document ranking (decreasing order of fused document score) is being executed.

To conclude, the entire workflow for the face/scene/voice search (single modality) is

(i) input face/scene (front-end), (ii) input embedding (backend), (iii) cosine-similarity (input embedding - face-embeddings from database (Elasticsearch)) and (iv) return top 20 results with min-score 85%. The high-level visualization is presented in Figure 25.

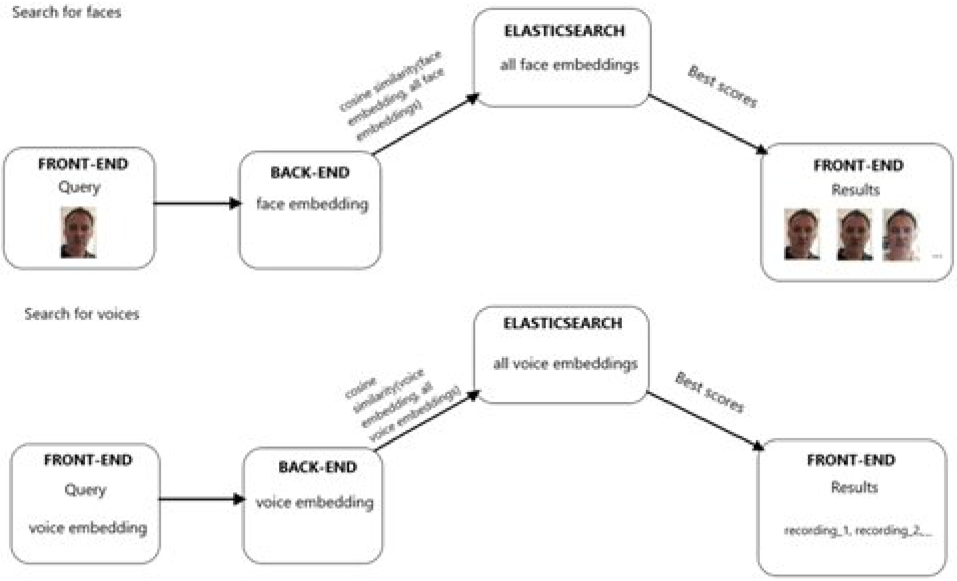


Figure 26. High-level visualization for the queries “Search for faces” and “Search for voices”

Similarly, the workflow for the complete Multi-modal Query is: (i) inputs: voice, scene and face & weight(s) (frontend), (ii) embeddings of inputs (backend), (iii) search algorithm based on cosine-similarity for voices scenes and faces and (iv) return top 20 results with lower min-score. The high-level visualization is presented in Figure 26.

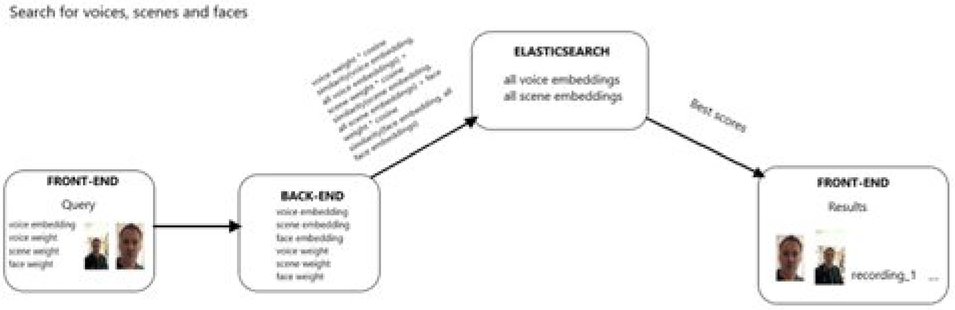

Search for voices, scenes and faces
FRONT-END
Query
voice embedding
voice weight
scene weight
face weight
BACK-END
voice embedding
scene embedding
face embedding
voice weight
scene weight
face weight
voice weight * cosine similarity(voice embedding, all voice embeddings) + scene weight * cosine similarity(scene embedding, all scene embeddings) + face weight * cosine similarity(face embedding, all face embeddings)
ELASTICSEARCH
all voice embeddings
all scene embeddings
Best scores
FRONT-END
Results
recording_1 ...

Figure 27. High-level visualization for the query “Search for voices, scenes and faces”

To conclude, the main technologies behind this Multi-modal query component are:

- The Elasticsearch engine
- The Python programming language
- The Flask application.

Elasticsearch is a distributed, free and open search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured. Elasticsearch was used for adding new data, more specific about the multi-modal data ingestion initial step. These data, but also metadata coming from the video data processing pipeline, were ingested after a detailed design process regarding the main data ingestion pipeline. Elasticsearch also provided the capabilities of data manipulation, data integrity and data flow. In our approach Elasticsearch was used both for data ingestion and data storing. In our case, Elasticsearch was used also for performing and combining many types of searches and analyses. More specific, regarding the procedure to measure the similarity between two sequences of numbers, standing as the ID of metadata information (voice, scene, face embedding) our data analysis was based on cosine similarity offered through Elasticsearch. Relying on Elasticsearch allows to scale to large cases and to ease the extension of the Multi-modal query component to textual data if needed in the future. Finally, the

Python programming language was used in order to develop a web API. For this activity Flask was used[47]. Flask is one of the customizable Python frameworks and is designed as web framework for RESTful API development. Through Flask the control of how the accessing to data is taking place was executed.
The final step is evaluate the performance of the Multimodal query. During the calculation of cosine similarity at Elasticsearch a score field is returned which represents the similarity between the input dense vector and the dense vectors stored in Elasticsearch and is expressed as accuracy. Accuracy values range between 0-1; 0 being the lowest (worst) value and 1 being the highest (best) value. In addition, every value under 0.65 is considered as “arbitrary” and every value over 0.85 is considered as “similar”. Values between the range of 0.65 and 0.85 need to be examined by the web interface in order to be determined.

For every use case, corresponding to a single modality query for faces, scenes or voices, a basic evaluation average accuracy was used. The range of values was between 0-1. To provide these average accuracy values a python script was developed which takes a random image (in fact the dense vector which represents the image) as input and compares the cosine similarity with the already ingested images (dense vectors) in Elasticsearch. When tested the evaluation results were: (1) for a random *face image* - average accuracy of 0.91; (2) for a random *scene image* – average accuracy of 0.78 and (3) for a random *voice* – average accuracy of 0.77.

Furthermore, for the multimodal queries: (i) *modality-2 - voices and scenes*, (ii) *modality-3 - voices, scenes and faces* another calculation of average accuracy was conducted this time using input weights. The range of values was 0-1. To provide these values of average accuracy a python script was also used for this case. Therefore, when tested with the above inputs for *modality-2: voice and scenes* and with optimized weights of 0.35 and 0.65, respectively, the result of the average accuracy was equal to 0.76. For *modality-3 - voices, scenes and faces* having optimized weights of 0.30, 0.35, 0.35 the average accuracy was calculated 0.73.

To conclude, the evaluation process can be characterized as complete only when the results are being evaluated with metrics such as *Recall* and *Precision* based on the ground truth. Typically, *Recall* and *Precision* values range from 0 to 1 with 1 being the optimal. Moreover, the ideal approach would be that every result that was calculated, to be evaluated only once in order to ensure the highest accuracy possible for all results. Any results for both metrics over 0.75 would be considered satisfactory.

16. Meta data: Geolocation

16.1 Geolocation

As mobile phones have become established in everyday communication, so has their use in criminal activity. Criminals often use cheap, prepaid action phones or crypto phones to communicate about criminal activities. While criminal users often employ various strategies to obfuscate their communication, every cell phone leaves location traces, which are accessible to law enforcement. Most notably, such location traces are included in intercepted communication that can be derived from call detail records (CDRs). The latter are typically stored by operators and historical data can be requested for up to several months in the past. Additionally, as CDRs contain only meta data, these data are less intrusive compared to alternatives. These two aspects make CDRs relatively accessible to law enforcement and therefore interesting for criminal investigation and evidence evaluation.

[47] https://flask.palletsprojects.com/en/2.2.x/ and https://www.elastic.co/what-is/elasticsearch

Although CDRs are widely used in criminal investigations, the location data must be interpreted with care. The location of the device cannot be read directly from the data, but only the data of the cell tower it connects to. This means that the device is in the cell's service area at the time of measurement.

In ROXANNE, we developed technology to help an investigator interpret these traces. In particular, we use geolocation for phone user identification and network analysis.

16.2 Using location traces for phone user identification

For any two cell phones, a typical set of hypotheses to evaluate is whether (hypothesis 1) the phones were used by the same individual during a time period; or (hypothesis 2) the phones were used by two individuals. In forensic examinations, it is common practice to evaluate the location traces within the likelihood ratio framework. Even in investigations, where different quality requirements apply, it may be recommended to use the likelihood ratio framework nonetheless. This not only makes it easier to apply results more broadly, but it also makes it trivial to combine the evidential strength of multiple observations, as long as the observations are statistically independent. Statistical independence between two observations means that one observation does not yield any information regarding the likelihood of observing the other.

In ROXANNE we have developed a novel method for evaluating the strength of evidence from call detail records that any pair of phones were carried by the same person. The method produces a score for pairs of registrations, which lead to a score for any pair of phones for a given period. A calibration step follows, in which the score is converted into a likelihood ratio (LR) between both hypotheses. Using data from field experiments and actual phone usage we have assessed the performance of the method and also the impact of a range of model and data changes. Our findings show that the method performs well under different modelling choices and that it is robust under lower quantity or quality of data.

Our method achieves a log likelihood ratio cost (Cllr) of 0.42. Figure 30 shows the distribution of LRs under both hypotheses. The average minimum and maximum LRs found are 1/78 and 57. Performance of the more sophisticated models is mixed. In particular, gradient boosting gives better results on the test data. As expected, performance increases when more data points are available per track pair.

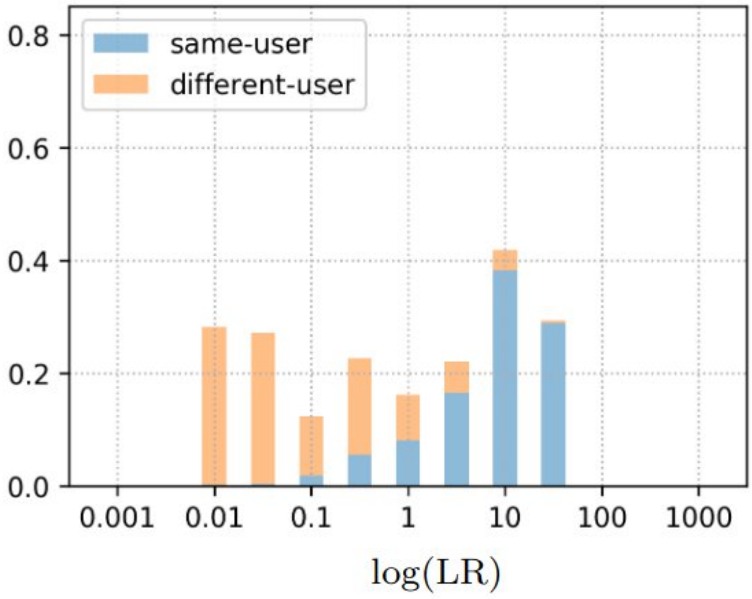


Figure 30 Distribution of log10 LR values for (blue)same-user and (orange)different-user track pairs.

Location traces are often of value for phone user identification. However, there may be other information to answer the same question, such as speech or text messages. In practice, no single information source provides conclusive answers, but the various sources of information may be combined to make a better assessment of a particular set of hypotheses.

We expect that travel behaviour as revealed from location traces is statistically independent from these other modalities. This is important because it dictates how output from different modalities can be combined. Whether this assumption holds is subject to future research.

16.3 Using location traces for network analysis

In large investigations, there may be dozens or more devices involved. It may be laborious to search for a potential user of a device in a pool of suspects. When people exchange or replace devices, this becomes even more cumbersome. ROXANNE can help in this process by integrating the phone user identification technology in network analysis. This way, the investigator has access to the joint network with other communication data.

17. Conclusions

This deliverable intends to inform the reader about the various speech, Natural Language Processing and video technologies that has been developed as a part of the ROXANNE project. The deliverable is and extension and update of previous deliverables on these topics. As such, it describes the proposed methods that LEAs can use to improve current technological baselines. When describing the methods, emphasis is put on how the extracted information from multi-modal sources reveals the structure of criminal network analysis. The deliverable also briefly presents the status of integration of these technologies.

The primary aim of the report is to present a streamlined pipeline that exploits multi-modal information, in particular, from phone calls but also in e.g. videos. We first extract metadata such as phone numbers and gelocation which can provide an initial social network structure assuming there is a one-to-one mapping between person and phone numbers. The audio processing starts with voice activity detection that separates speech from non-speech in the audio. If needed, diarization may then be used to speparate multiple speakers in the recording. Speaker recognition can then be used to produce a more accurate network than that based on phone numbers. Depending on circumstances, the phone number information can also be used to improve the speaker recognition performance. Speech recognition is then used to produce a transcript of the speech. In this process it is possible to boost the prediction of important words specified by the user. We then build a “mention network” from the transcripts using entities mentioned in conversations. Thereafter, we process the transcripts to find third parties mentioned in the network and integrate it into the network. Thus, we investigate social, textual and acoustic information from a network to provide LEAs with meta-information about the crime. We also present the work done in integrating video technologies into this pipeline. An exploratory analysis is also made into using geolocation information in the paradigm proposed. This should provide the reader with insights into the potential of previously untapped information resources and display the extent to which the retrieval and analysis of this information can be automated.

The deliverable also describes research results as well as directions for future work that seem promising according to the members of the consortium.