D5.1 INITIAL SPEECH/TEXT/VIDEO TECHNOLOGIES

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:	30 April 2020
Due date:	30 April 2020
Deliverable lead:	IDIAP
Work package:	WP5
Type of action:	RIA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

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.

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.

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
V1.0	16 April 2020	Honza Cernocky (BUT)	All	Structure defined
	23 April 2020	Gerhard Backfried		Speech Recognition
	29 April 2020	Denis Marraud (AIRB)		Video analysis updates
	24 April 2020	Marek Kovac (PHO)		Speech technologies
	27 April 2020	Johan Rohdin (BUT)		Speaker recognition
	28 April 2020	Honza Cernocky (BUT)		Corrections, updates of several sections
	29 April 2020	Marek Kovac (PHO)		Corrections at Speech Technologies
	29 April 2020	Dawei Zhu (USAAR)		Entitiy-Detection Technologies
	30 April 2020	Erinc Dikici (SAIL)		Speech Recognition
	30 April 2020	Petr Motlicek (IDIAP)		Proof-checking
	30 April 2020	Honza Cernocky and Johan Rohdin (BUT)		Final editing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Executive summary

This deliverable, D5.1 Initial speech/text/video technologies, summarizes the mentioned technologies available at the outset of the project. It first gives a non-expert overview of such technologies and concentrates on their use in LEA framework. It lists technologies available at partners, and presents experiments done so far in the ROXANNE project on three data-sets drawn from public sources: Crime Scene Investigation (CSI) series, National Institute of Standards and Technology (NIST) speaker recognition data, and ENRON data-set. In the following, it comments on preparing the speech, text and video technologies for integration into the ROXANNE platform, and mentions related work. The document concludes with drawing directions of future work.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

1. Introduction

This deliverable “D5.1 Initial speech/text/video technologies”, is the first set of software and report to be submitted as part of the “Speech, text and video data analysis” work package (WP5) of the ROXANNE project. In this introductory section, we present the definition of WP5, the purpose of this document and its scope within WP5, and the outline of this report.

1.1. Background

This Deliverable is the first official output of WP5, that is responsible for providing speech and text and video analysis technologies in ROXANNE. According to the project proposal and the Grant Agreement, WP5 has the following Objectives:

- *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 SID, 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 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;*
- *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;*

WP5 is in the core of the project and is technologically related to

- WP4 - heavy dependence on training and test data.
- WP6 - feeding the results to network analysis
- WP7 - speech, text and video analysis modules need to be integrated to ROXANNE platform.

WP5 however depends also on WP2 (the definition of requirements, scenarios, and languages relevant for participating LEAs), WP3 (all development, even of the basic data mining technologies, must be legal and respect ethical principles), WP8 (significant portion of the training will be concentrated on speech, text and video technologies) and WP9 (the R&D done in WP5 is not to be disseminated only as part of the integrated ROXANNE solution but also independently).

1.2. Purpose and scope

D5.1 is a public deliverable intended to provide (according to the Grant Agreement) "A set of software and associated report for rapid deployment of speech, NLP and video technologies for early integration and system testing.". The purpose of this document is primarily to provide a technically savvy LEA analyst with the know-how necessary to judge the relevance of speech, text and video technologies for LEA work, first alone, and later in combination with the network analysis (WP6). We took care to adapt the language and

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 833635.

style of the document to such readership - the necessary machine learning and mathematical background is mentioned in the footnote references rather than in the main text. The document does not detail the definition of metrics used for measuring the performance of basic technologies, it relies on already submitted deliverable D8.1 "Validation criteria list and performance test methodology". It also does not explicitly deal with the pertaining legal and ethical framework, and leaves it to respective WP3 (and additionally requested D10.*) deliverables.

The deliverable is mainly the output of, Task T5.1 Initial speech/NLP/video technologies, whose definition according to GA is as follows:

> *To enable a quick start of the project's integration activities, PHO and SAIL will deliver production grade speech technologies to partners. Similarly, USAAR will provide its existing NLP technologies addressing some initial issue relevant for the target domain, and AIRBUS will provide baseline video technologies. All these will be made available with easy-to-use interfaces (i.e. as Linux scripts for laboratory use or REST-API services. or command line interface for the production use in WP7). The initial ASR modules will be provided in 8 languages (section 1.3.3.4 at page 17) corresponding to the scenarios defined in T2.1; more languages will be dynamically added.*

But it also covers initial results in the follow-up Tasks T5.2-5.6, as a significant portion of the deliverable deals with experiments. As real LEA data is not yet available in the project, and the creation of simulated ROXANNE data is in progress (Task 4.6), experiments were done on three data-sets built on publicly available data. We are aware that these data-sets are not optimal but they were used to bootstrap the R&D collaboration in the project. The advantages and drawback of all Crime Scene Investigation (CSI) series, National Institute of Standards and Technology (NIST) speaker recognition data, and ENRON data-set, are discussed in the respective section.

1.3. Relation to ROXANNE first field-test event

The main objective of D5.1 is the preparation of technologies for the first field-test meeting of ROXANNE project.

More specifically, technologies selected for speech,text, and video processing are motivated by their use in the first field-test event, organised by ROXANNE project. The list of technologies has been carefully discussed with internal LEA partners of ROXANNE project, while following the Grant Agreement. The described technologies are therefore also prepared to be integrated into the ROXANNE platform, besides their objective evaluations. Rather than detailing data-flows, data exchange formats and APIs (that are part of WP7 documents), this deliverable deals with the actual capabilities (and limitations) of technologies prepared for the integration.

1.4. Document structure

In the following Section 2, structured per technology, general description of individual speech, text and video technologies is given. Section 3 is based on collection of inputs from partners in the beginning of the project and provides an overview of technologies available for ROXANNE. Section 4 describes initial tests on ROXANNE data-sets (CSI, NIST and ENRON data), also including comments on the choice of these data-sets. Section 5 mentions activities aiming at preparation of basic speech, text and video technologies for integration in the ROXANNE platform, and section 6 lists achievements not directly related to ROXANNE data or integration, but nevertheless important for the project. Finally, section 7 concludes the deliverable and draws direction of future WP5 work in the project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

2. Basics of speech / text and video technologies

This section provides basics of speech, text and video technologies used in the project. All used technologies build on the principles of machine learning (nowadays often called “artificial intelligence”) and generally fall into categories of detection (for example for speaker verification), identification (gender recognition, language identification, topic identification), regression (age estimation) and sequence processing (speech to text, diarization, video processing). Although their inputs vary from speech signal through text to sequences of video frames, almost all of them follow the structure depicted in Fig. 1:

.

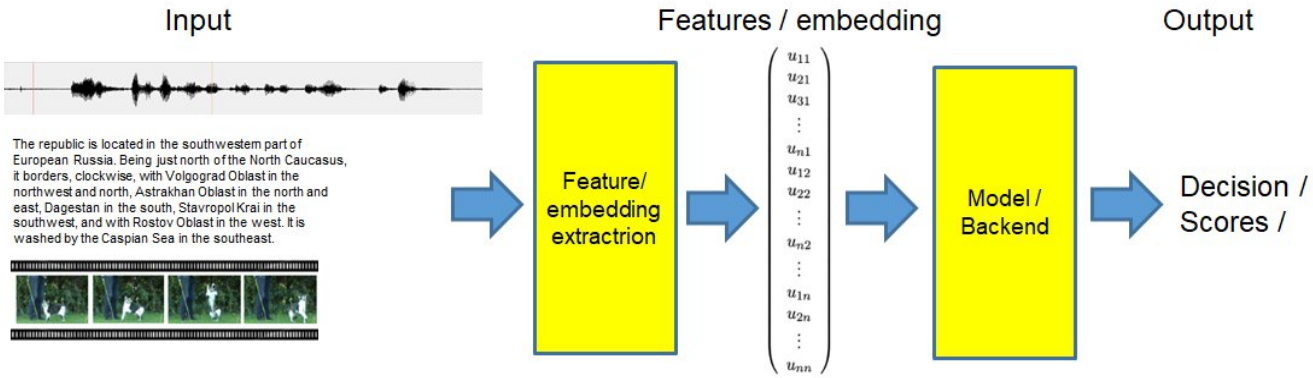

Figure 1: General structure of basic speech, text or video mining technology in ROXANNE .

The input is first processed into parameters (or features) whose task is to preserve the information useful for the processing goal while discarding the nuisance information. The output of this block can be a sequence of feature vectors as a function of time, but in many technologies, it is only one informative vector that is summarizing the information from a longer context (from syllable to whole audio file, from one word to whole document, and from a few seconds of video to the whole recording). In such a case, we often speak about an **embedding**. The following Model / Backend block processes the sequence of feature vectors or embeddings into the final result. In the past, both feature extraction and Model / Backend were usually hand-crafted and based on operations known from signal processing (such as frequency transforms or correlation), nowadays they both rely on trainable architectures based on artificial neural networks (NN) and their more recent and more complex variants, such as recurrent NNs (RNN), Long-Short Term Memory (LSTM) RNNs, NNs with attention mechanism, convolutional NNs (CNN) and others.

Both feature / embedding extraction and Model / Backend blocks need to be trained on **data**. Depending on the way they are trained, we usually distinguish **hybrid systems**, where the two blocks are trained separately, and **end-to-end** (e2e) model where the training proceeds straight from the input (speech, spectrogram, text video) to the final task. The e2e techniques are theoretically more powerful, but require significant amounts of training data (that is usually available only to big technology firms such as Google, Facebook, etc.) and significant computing infrastructure, therefore, practical systems (and most of the systems investigated in ROXANNE) are hybrid ones.

The following sections present individual technologies more in detail.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 833635.

2.1 Voice activity detection

Voice Activity Detection (VAD) identifies parts of audio recordings with speech vs. non-speech content. It is actually not speech data mining technology itself, but it is an important pre-processing step for many follow-up technologies. It can however be an extremely important and helpful tool for LEA users: When massive amounts of recordings should be processed in short period of time and provided to operators, detection of voice activity can be easily used for filtering out of those that do not contain human speech. That usually reduces overall load of recordings coming to further processing.

In case of clean speech signals (upper panel of Fig. 2), VAD can be built using a simple energy detector (speech signals have higher energy than background noise). However, for noisy signals from realistic conditions, it is necessary to resort to a trained VAD. Good results are usually obtained while training a simple NN on a sufficient amount of data containing typical background noises, music, etc.[1] If properly trained, the technology is language-, accent-, text-, and channel- independent.

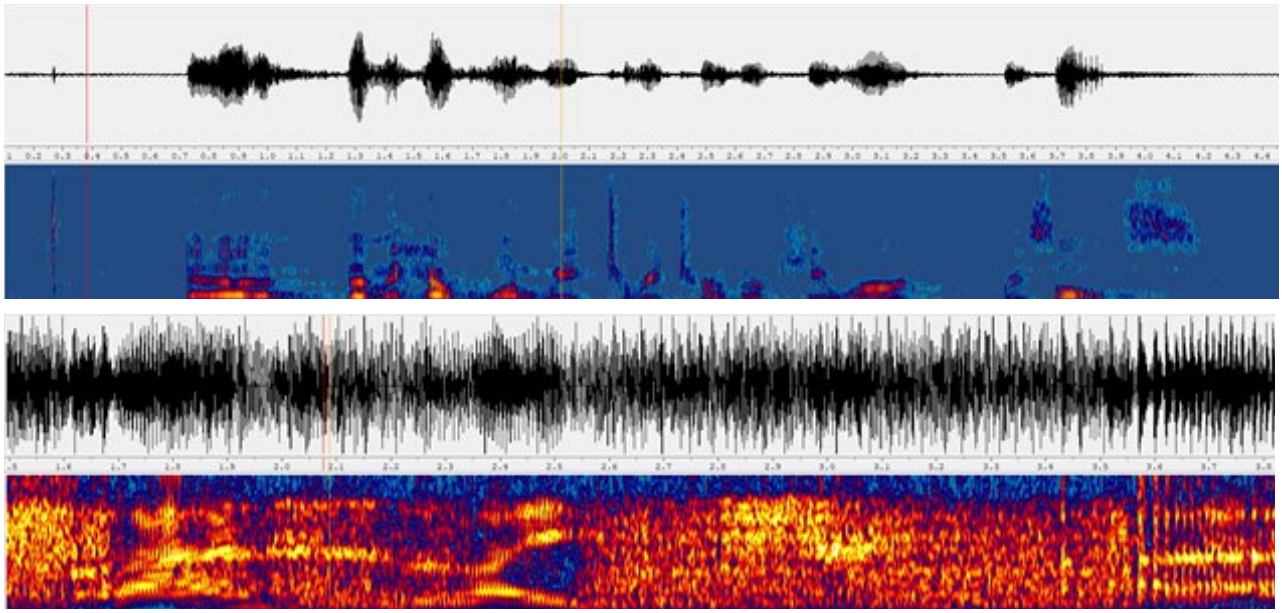

Figure 2: Examples of clean and noisy speech signals and their spectrograms.

The output of VAD is a file or stream with meta-information containing timing of speech and silence segments.

2.2 Diarization

Speaker Diarization (SD) is commonly defined as the task of answering the question “who spoke when?” (determining speaker turns in an utterance, see Figure 3). Although being apparently easy for humans, diarization is a highly challenging task for machines. SD deals not only with the already complex Speaker Recognition stage, but also faces the problem of having unknown number of speakers in utterances,

[1]NG Tim, ZHANG Bing, NGUYEN Long, MATSOUKAS Spyros, ZHOU Xinhui, MESGARANI Nima, VESELÝ Karel and MATĚJKA Pavel. Developing a Speech Activity Detection System for the DARPA RATS Program. In: Proceedings of Interspeech 2012. Portland, Oregon: International Speech Communication Association, 2012, s. 1-4. ISBN 978-1-62276-759-5. ISSN 1990-9772.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 833635.

segmentation of speech into speaker turns (finding boundaries between speakers), treatment of overlapped speech (cross-talk), etc.

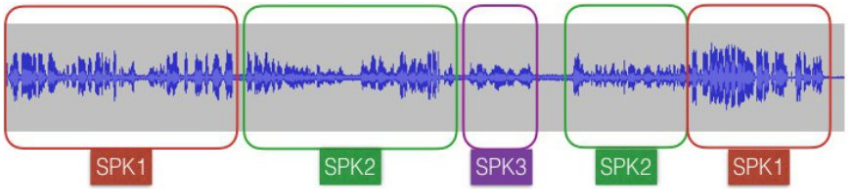

Figure 3: Speaker diarization.

State-of-the-art speaker diarization systems first "chop" the signal into small fragments and then cluster them according to the speaker identity. This is done by analysing the distribution of speech features in the segments and by comparing the segments with each other. At the same time, the system must infer the number of speakers in the utterance. The full process involves several steps as depicted in Figure 4. First, features are extracted from fixed intervals in the speech signal, then they are subject to a transformation in order to increase their speaker discrimination ability or to make them more appropriate for the later processing steps. Then the signal is segmented into uniform speech fractions. Next, statistics of the features for each segment are calculated. The previous versions of diarization used representations gathered using Gaussian models, so called i-vectors[2], newer versions resort to embeddings derived using NNs (so called x-vectors[3], see section 2.6 for more thorough explanation). As there is no enrolment data to build the speaker models in advance, unsupervised clustering is applied to assign the segments into classes (speakers), where the number of classes are unknown in advance. The last stages (segmentation, statistic computation and clustering) are frequently repeated, applying different approaches to attain optimal performance. Bayesian approaches are nowadays the breakthrough for SD tasks[4], however, diarization is still far from a mature technology, especially in challenging environments, and scenarios with lots of speaker overlap. Diarization needs speaker-labelled data in order to train the extractor of speaker embeddings.

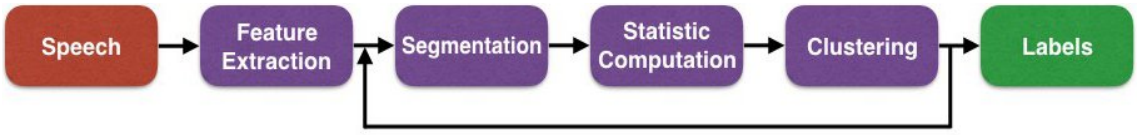

Figure 4: Scheme of a typical diarization system.

Diarization is an important technology for LEAs as it allows to process one-channel (mono) recordings which often is the only available option for telephone conversations (a lot of legacy equipment mixes sides of conversation to one channel to save space) as well as for room wire-tapping.

[2] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788-798.

[3] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333).

[4] DIEZ Sánchez Mireia, BURGET Lukáš, LANDINI Federico Nicolás a ČERNOCKÝ Jan. Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, roč. 28, č. 1, s. 355-368. ISSN 2329-9290.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

2.3 Gender identification

Gender identification (GID) technology allows to automatically distinguish whether a woman or a man is speaking in the segment or recording being examined. Because the basis for training this classifier is a large number of recordings of spontaneous speech in different languages, this technology can be considered independent of the language and text (speech content) used.

For LEA use, gender detection can bring a welcome narrowing of the search space, especially knowing that male subjects are mostly present in investigations into criminal and terrorist cases.

The error rate of gender recognition technology is in the range of 2-3% for spontaneous speech (telephone recordings) which makes it one of the most precise speech mining technologies.

The output is a "raw" score - the probability of the male/female gender. This raw score can be converted to another scale based on the needs of the user or integrator. As other detection systems, prior probabilities can be used, so that the technology can prioritize men (or women) as the final decision. This situation can occur, for example, if the examined recordings are from a speaker from the Asian region (men - Asians have the fundamental frequency of voice set higher) or come from acoustic channels with a limited frequency spectrum.

2.4 Age estimation

Age Estimation (AGE) enables users to estimate the age of a speaker. The technology is text, language, dialect, and channel independent, and performs a simple regression, usually from an embedding (i-vector or x-vector) trained for speaker recognition.

For LEA, this technology allows for dividing targets into age groups, so that LEAs can pre-filter recordings with the speakers of a specific age referring the ones in criminal groups.

The technology does not estimate the age precisely, in usual scenarios, it is necessary to count on +/-7 years precision of the estimate.

2.5 Language and dialect recognition

Language Identification (LID) helps distinguish the spoken language or dialect. It enables the system to automatically route valuable calls to experts in the given language or to send them to other software for analysis.

Historically, the LID systems started with standard spectral feature used by other speech data mining technologies, but soon moved to bottleneck features (BN)[5] that are obtained at a narrow (bottle-neck) layer of a neural network trained to recognize phonemes. A variant of BN-DNN, which used multilingual training, brought additional improvements over monolingual NNs[6]. The current research follows the general trend in speaker recognition and makes use of embeddings produced by a previously trained NN[7]. With embeddings, it is sufficient to use simple classifiers such as GLC (Gaussian Linear Classifiers) for

[5] P. Matejka et al: Neural Network Bottleneck Features for Language Identification. In: Proceedings of Odyssey 2014. Joensuu.

[6] R. Fer et al. 2017. Multilingually Trained Bottleneck Features in Spoken Language Recognition. Comp. Speech & Language, 46.

[7] A. Lozano-Diez et al.: DNN Based Embeddings for Language Recognition, In: Proceedings of ICASSP 2018, Calgary.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

estimating the probabilities of different classes (languages) of a speech recording. It is also possible to design a full end-to-end LID system[8].

As it is relatively easy to train LID (it is enough to have only recordings with language labels, no transcription or speaker labelling is necessary), companies and university laboratories have LID systems comprising almost 100 languages, and it is easy for users to train their own models. The technology is relatively independent on text or channel. However, calibration and fusion of systems, as well as adding new languages and dialects with limited amounts of (possibly badly labelled) data are still among the research issues.

For LEAs, LID has several possible uses:

- Preselecting multilingual sources and routing audio streams/files to language dependent technologies (transcribing, indexing, etc.)
- Analyzing network traffic media (language statistics)
- Routing particular calls (languages) to human operators (language experts)

2.6 Speaker recognition

Speaker recognition (SR) refers to the process where a machine infers the identity of a speaker by analyzing his/her speech. The basis of SR is the task of **speaker verification** (SV) that outputs a score (and eventually a hard decision) that in a pair of recordings, the same speaker speaks.

Traditionally, speaker recognition systems start with a stage of feature extraction, which allows to represent the speech segment with some features that already remove some of the unwanted information. This representation consists of a sequence of vectors, each representing few milliseconds of speech, and often based on acoustic features such as the well-established Mel Frequency Cepstral Coefficients (MFCC). This stage is followed by a modelling step, where different techniques try to remove the remaining nuisance variabilities and increase robustness to adapt it to real conditions, a main issue in research. One of the most common approaches that has been the state-of-the-art for many years is the model based on the Total Variability subspace, which broadly speaking, provides a fixed-length representation of an utterance that comprises the task-dependent and independent (variability) of the feature space, the so called i-vector[9].

Recently, the existing speech processing systems experienced a revolution in terms of performance thanks to the use of efficient deep learning algorithms. One successful approach for using deep neural networks (DNNs) in speaker recognition is to use them to extract feature vectors. The DNN is trained so that such features better reflect the phonetic content of the speech compared to traditional acoustic features. Current state-of-the-art systems for speaker verification use DNNs to, similarly to i-vector extractors, convert variable-length sequences of feature vectors into fixed-length vector representations. Such a representation is known as d-vector[10] or embedding (so called x-vector)[11]. Contrary to i-vector extractors, such DNNs are typically trained with an objective that emphasizes on the speaker discrimination. In order compare two utterance level representation such as i-vectors on DNN embeddings and judge whether the utterances are from the same speaker or not, simple cosine similarity often

[8] LOPEZ-MORENO Ignacio, GONZALEZ-DOMINGUEZ Javier, MARTÍNEZ González David, PLCHOT Oldřich, GONZALEZ-RODRIGUEZ Joaquin and MORENO Pedro. On the use of deep feedforward neural networks for automatic language identification. Computer Speech and Language, vol. 2016, no. 40, pp. 46-59. ISSN 0885-2308.

[9] N. Dehak et al., "Front-end factor analysis for speaker verification", IEEE Transactions on Audio, Speech & Language Processing, 2011.

[10] E. Variani et al., "Deep neural networks for small footprint text-dependent speaker verification", Proc. ICASSP, 2014.

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

produces good results. However, better results are usually obtained with probabilistic linear discriminant analysis[12,13] which is a probabilistic model trained on labelled data. The standard speaker recognition system therefore exactly matches the general scheme presented in Figure 1. Developments and comparison of classical and neural approaches to SR have been recently summarized in a journal article[14]. Note, that in the description of commercial systems, a business term "voiceprint" is used more frequently than "low-dimensional representation", "i-vector" or "x-vector".

A range of tasks can be derived from the basic speaker verification, all relevant to LEA use:

- **speaker identification** (SID), where registering a known speaker is called "enrolment" and computing the score of an unknown recording is often called "test". Based on the nature of the task, SID can be closed-set or open-set.
- **speaker search** (one or several speakers are searched in quantity of data)
- **speaker clustering** - based on a set of unlabelled recordings, we try to infer the amount of speakers and attribute them to the calls.
- **speaker diarization** (already discussed in section 2.2) where speakers are detected in one mono recording and individual speakers' segments are labelled.

For the Roxanne project, speaker recognition is a crucial technology, as it enables to discover missing links in the criminal network analysis and put names to speech recordings in case the subjects are intentionally obfuscating them (for example by the use of pre-paid SIM cards or stolen telephones).

2.7 Speech to text

The Speech-to-Text (S2T or Automatic Speech Recognition, ASR) component converts audio input into a sequence of words corresponding to the audio's transcript. It does so by first converting the audio input into a sequence of features and then by processing this sequence using an acoustic model (AM) as well as a language model (LM), producing the most likely sequence(s) of words corresponding to the input. Each word is output together with its exact timing (within the audio) and a confidence-score, indicating how *confident* the S2T engine was that the word is indeed the correct one. Depending on the scenario, ASR systems may output a single sequence of words or alternative sequences of words (typically referred to as *n-best*). In case of alternatives, these represent different possible transcripts of the audio. Furthermore, a richer graph-like structure, representing a network of possible alternatives (a so-called *lattice*) may be output, representing many more possibilities of which words could correspond to the audio. The latter may be searched for keywords and thus be employed in the scope of *keyword-spotting* applications.

The ASR engine itself (Figure 5) is typically language-agnostic. The models, however, are language- and domain-dependent. Both types of models - the AM as well as the LM - are trained from corpora which should resemble the kind of audio and speech encountered during actual application as closely as possible. Speech corpora typically contain samples of many different speakers, different genders and age groups in order to yield speaker independent models which can be employed for speaker independent ASR. The LM is trained from text which is typical of the domain and style of application. The vocabulary consists of words representing the language of interest as well as the particular domain.

For LEA use, ASR allows to tap into the content of conversations, and therefore is useful in multiple ways:

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

[13] P. Kenny, Bayesian speaker verification with heavy-tailed priors, in Odyssey 2010.

[14] MATĚJKA Pavel, PLCHOT Oldřich, GLEMBEK Ondřej, BURGET Lukáš, ROHDIN Johan A., ZEINALI Hossein, MOŠNER Ladislav, SILNOVA Anna, NOVOTNÝ Ondřej, DIEZ Sánchez Mireia and ČERNOCKÝ Jan. 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE. Computer Speech and Language, vol. 2020, no. 63, pp. 1-15. ISSN 0885-2308.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

1. producing transcriptions that can (to some extent) save police staff and analysts time otherwise needed for manual transcriptions.
2. use in standard indexing, search and relating of information, possibly enhanced by follow-up modules, such as named entity recognition and topic detection (see following sections).
3. providing material for relation, content and network analysis (as investigated in WP6).

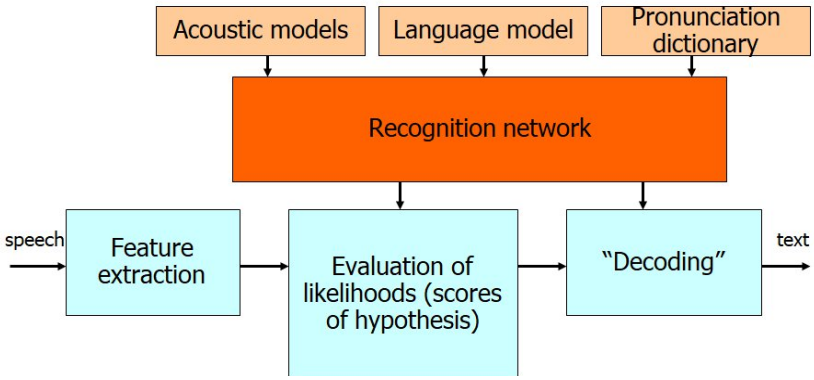

Figure 5: General scheme of speech to text.

ASR therefore forms an integral part of the suite of technologies employed in ROXANNE. Due to the specific application domains of ROXANNE, adaptation to specific audio conditions as well as language styles and domains are required and the use of ASR within ROXANNE poses several challenges with regard to these models:

AM: when training the acoustic model, the targeted domain and audio conditions need to be taken into account. This requires a specific corpus of training data, consisting of audio, corresponding transcripts and pronunciations for all words occurring within these transcript. The challenge within ROXANNE is the general absence of specific corpora for the field of organized crime. To account for this fact, corpora of similar audio contexts can be used. Likewise, methods to allow certain levels of adaptation of models built of different kinds of audio are employed.

LM and vocabulary: when training the LM and selecting a specific vocabulary, the targeted language, registers, style and domains need to be accounted for. Special terminology - which may be exhibited by a particular sector of organized crime, specific expressions and the use of language within criminal contexts may differ from every day use. Analysis of terminology pertaining to different kinds of crimes provides one way to adjust the models accordingly. The adaptation (or extension) of the vocabulary and LM can typically be performed by tools provided in combination with ASR systems. These tools allow domain experts to rapidly adapt an ASR system to a particular type of crime or even to a specific setting.

ASR-models are typically mono-lingual, which means that for every language a separate model needs to be created. Within ROXANNE, models will be provided for each language of interest. In case of audio containing multiple languages, it is foreseen to segment these audio files into language-coherent sections, each of which can be transcribed using one of the ASR components. As time-codes are preserved, the resulting sequence of words can be connected after processing.

2.8 Entity detection

The Named-Entity Recognition (NER) component extracts targeted named-entities (person names, location, organizations etc.) from a given text. A visualization of NER below:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

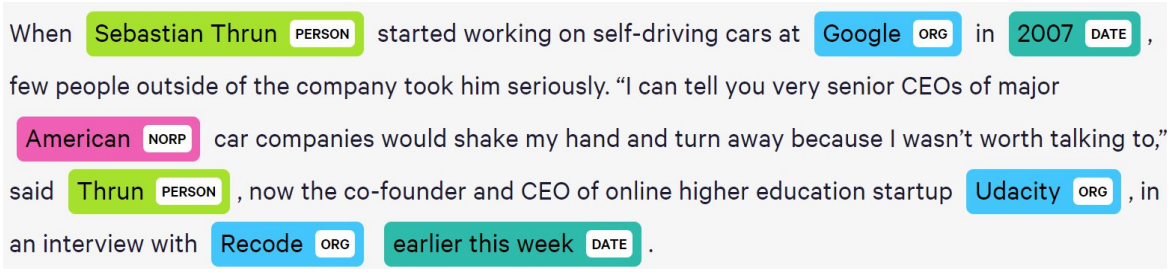

Figure 6: A visualization of NER.

Traditionally, one uses gazetteers, and grammatical rules to detect entities from text. This has the advantage of interpretability. However, many entities do not appear in the gazetteers (new entities appearing on the internet) or appear in different gazetteers. For example, The word “Washington” can be a location or a person name. So it can appear in two different gazetteers. Also, rules can be incomplete to detect all entities and/or be contradictory. Thanks to the rapid development of deep learning methods in recent years, we can build deep neural networks (DNNs) to detect named-entities, yielding state-of-the-art performance.

In the deep learning era, we detect a named-entity by considering the word and sentence features learned by the DNN. First, each word in the text is converted to a numerical representation in form of a vector, where similar words should have similar vectors. Hence, word vectors encode semantic and syntactic aspects of each word. Taking into account that each word can have different meanings in different sentences/contexts, we can even assign different vectors for one word depending on its context. Then, the sentence represented by its word vectors is processed by using a LSTM network (a deep learning model which considers the long-term dependencies of each word) or the BERT model (a recently developed deep learning model that scans a sentence using an attention mechanism). The output of a sentence consisting of N words are N named-entities labels, one for each word. Common labels are “PERSON NAME”, “LOCATION” and “O” (for a word that is not an entity).

For the LEAs, the NER model can speed up the text analysis. With proper visualization, all entities can be highlighted in the text. LEAs can promptly see them before having read the whole text and can decide whether to fully read on the text, or decide which part/text to focus on first.

The highlight of our NER technology is that it is language agnostic in the sense that it only requires training data for the target language, no language-specific rules are necessary. Furthermore, we can support new entities on-demand. For example, recognizing drug/weapon names appearing in the text could be important for LEAs to analyze the data. We can then adjust our model to recognize these entities. To achieve this, we only need a small amount of manually annotated data; much less than is usually necessary for these systems.

NER technology is trained on a text corpus annotated with named-entities.

2.9 Topic detection

The topic detection module utilizes transcript text as input and uses Concise Semantic Analysis (CSA) for inferring word representations. Thus, once the underlying semantics has been inferred, a small set of concepts is used to represent the input data. The intuition behind this approach is that highly abstract semantic elements (concepts) are good discriminators for clustering very short transcript texts that come

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 833635.

from a narrow (and noisy) domain. Once the relevant concepts are obtained, these are used for building a Bag-Of-Concepts (BoC). The advantages of BOC are :

- Able to address the deficiencies of traditional approaches, such as synonymity and polysemy.
- Semantic proximity is used to infer sets of terms that share a relationship.
- Is an unsupervised method, i.e, does not require prior information for finding topics.

The overall architecture of the topic detection module is shown in Figure 7.

Generally speaking, the idea is to first identify the underlying concepts contained in the dataset. For this, any semantic analysis (SA) approach for learning words representation can be employed; thus, learned representation allows to generate sets of semantically associated words. After obtaining the main concepts, documents are represented by a condensed vector, which counts for the occurrences of the concepts, i.e., a concept distribution vector. Finally, the build texts representation serves as the input to a clustering process, in this case, the K-means algorithm.

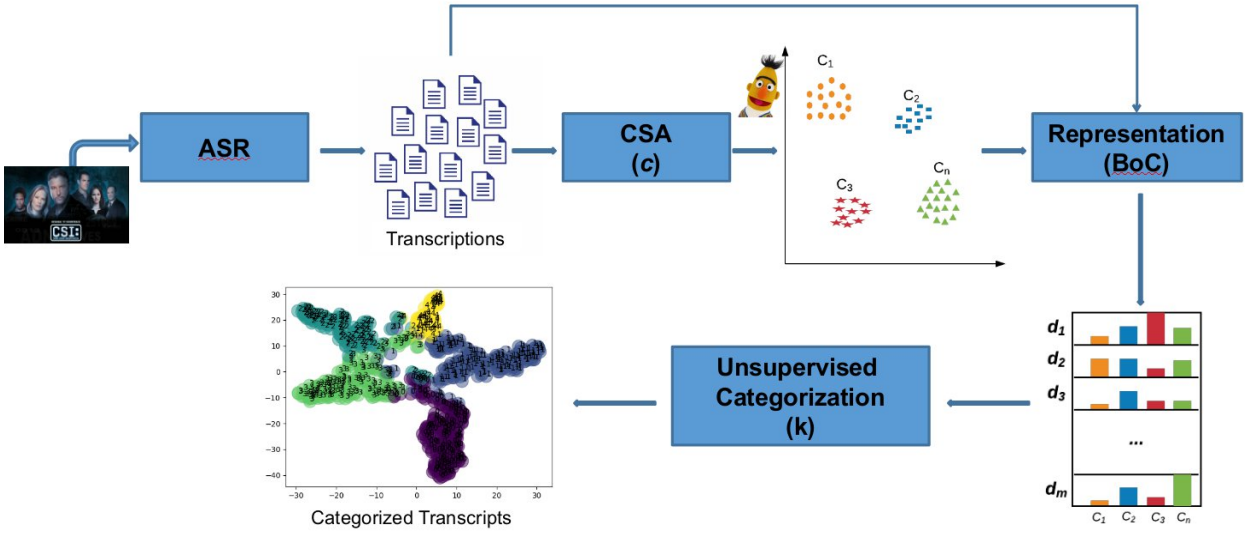

Figure 7: General framework to categorize spoken-documents using low-resolution concepts.

One crucial step of this approach is learning word representations, i.e., the semantic analysis process as shown in the general framework in Figure 7. For this, an important parameter is the resolution value (p), which indicates the number of concepts that will be employed for building the document-concepts matrix (BoC). Accordingly, four different methods (FastText, BERT, Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA) can be used for inferring the set of concepts). Overall, the developed method is able to find groups of documents that share the same amount of information about the same sub-set of concepts, resulting in a more coherent categorization of the documents.

The topic detection module can help LEAs to identify topics from a large collection of transcripts and it also helps to categorize them for more deep analysis. The topic detection module follows an unsupervised approach, hence it does not require much-labelled data to train the model and can perform efficiently in narrow domains with multilingual support and a flexible configuration change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

2.10 Video analysis

ROXANNE is focused on the exploitation of text, audio and network analysis techniques to support LEAs in the identification of speakers involved in criminal investigations. When the input documents are videos or when additional images are available, image or video analytics techniques can also provide useful information to populate or enrich the extracted networks.

With the very fast and successful development of learning techniques which leverage the availability of large amounts of annotated data, the performances and capabilities in the analysis of images or videos have significantly progressed in the last 8 years, providing unprecedented opportunities for supporting image analysts in their daily work. Video analytics techniques supporting forensics and investigations follow the three major following objectives:

- Helping analysts to cope with **very large amounts of data, most of which is not relevant** for the current investigation through filtering, searching or summarization capabilities.
- Helping analysts to **structure originally unstructured image or video data** through (1) the automatic extraction of semantic information about entities of interest (people, places, events, equipments, organizations) and (2) the identification of links between visual documents sharing similar or identical content.
- Supporting analysts in the **fine analysis of specific visual documents** through image enhancement techniques such as contrast enhancement, video stabilization, super-resolution, tampering detection, etc.

Part of these capabilities are already proposed in commercial softwares such as Analyze DI Pro from Griffeye[15], Amped FIVE[16], Magnet AXIOM, Cognitec TriSuite 64, Videntifier visual search engine or CameraForensics platform. While these tools may be very useful in terms of seamless import and sharing of information or documents from multiple sources in multiple formats, and in terms of image enhancement, their "intelligent analysis" capabilities are still limited. For instance:

- Analyze DI Pro integrates an automatic classification of child abuse content, face detection and recognition capabilities, specific object detection, near duplicates identification and an open API to integrate third parties analytics. Additional modules propose to detect and trace entities such as persons, vehicles, organizations or objects across the file database.
- Amped FIVE enables the search for duplicate images, supports linear and 3-dimensional measurements in images, various filters to enhance the video quality (sharpening, stabilization, contrast…) and video summarization based on motion detection.
- The Magnet AI module of the Magnet AXIOM solution identifies pictures containing weapons, extremist imagery, nudity and others.

In the following, we focus on the video analytics capabilities that could support speaker identification and network analysis from video files:

- **Face detection and recognition** are of course key capabilities for which industrial solutions are available and whose performances have significantly increased with the exploitation of very large training datasets (in particular produced by Chinese companies). If performances are very good for constrained setups, videos processed in real cases often exhibit poorer performances due to less optimal capture conditions in terms of viewpoint, resolution or image quality. For the Roxanne project, we rely on the state of the art open source model RetinaFace[17] for face detection.

[15] http://www.griffeye.com

[16] http://www.ampedsoftware.com

[17] *RetinaFace: Single-stage Dense Face Localization in the Wild*, J. Deng, J. Guo, Y. Zhou et al, arxiv, May 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- **People characterization**[18] includes all "soft biometrics" which can be automatically extracted from images such as gender, age class, skin color, face attributes or accessories (beard, moustache, glasses, hair cut and color). These attributes can be used to filter videos, locate specific video segments, assess a semantic distance between two detected persons, support re-identification when the image quality does not allow to use face recognition.
- **People or place diarization** identify in a video the segments related to a same person (even if its identity cannot be inferred) or to a same place. People or place diarization contributes to the summarization or structuring of long videos according to the places and people they observe.
- Finally, **object detection and semantic segmentation** can provide labels about specific objects or environments observed in videos.
- Finally, let's note that some techniques leverage both audio and video modalities to improve speaker identification[19] or to identify the active speaker[20].

In the Roxanne project, it is proposed to provide a place and face diarization service on videos, which could be fused with speaker diarization performed on the audio modality.

3. Speech / text and video technologies available for ROXANNE

This section summarizes speech, text and video mining technologies available at Roxanne's technical partners. For each, we mention the partner organization/company responsible for the technology, type of technology and its short description. We also find it useful to provide at least basic information about its training, as well as its "technology readiness" status (it is important to distinguish laboratory prototypes that are bunches of python scripts, executables and data, from professionally packaged production software). Where relevant, we also detail the languages the technology is available in.

partner: SAIL

- Technology: ASR (as part of MMI - Media Mining Indexer)
  - Description: https://www.isca-speech.org/archive/Interspeech_2019/abstracts/8029.html
  - Trained on: Various datasets + company-collected data + newswire
  - Technology status: Product
  - Language(s): 25 languages: Albanian, Modern Standard Arabic, Egyptian Arabic, Levantine Arabic, Mandarin Chinese, Dutch, International English, US English, Farsi, French, German, Greek, Hebrew, Bahasa Indonesia, Italian, Bahasa Malaysia, Norwegian, Pashto, Polish, Romanian, Russian, Spanish, Mexican Spanish, Turkish, Urdu

- Technology: Age&Gender Detection (as part of MMI - Media Mining Indexer)
  - Description: https://www.isca-speech.org/archive/Interspeech_2019/abstracts/8029.html
  - Trained on: company-collected data
  - Technology status: Product
  - Language(s): language independent

[18] *A survey of facial soft biometrics for video surveillance and forensic applications*, F. Becerra-Riera, A. Morales-Gonzales, J. Mendez Vasquez, in Artificial Intelligence Review, June 2019

[19] *Attention guided audio-face fusion for efficient speaker naming*, X. Liu, J. Geng, et al ,in Pattern Recognition, Vol 88, April 2019

[20] *AVA-ActiveSpeaker: An Ausio-Visual Dataset for Active Speaker Detection*, J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, C. Pantofaru, in arxiv May 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- Technology: Social Media Collector (as part of MMF - Media Mining Feeder)
    - Description: https://www.sail-labs.com/wp-content/uploads/2018/07/5_Media-Mining-Feeder-Indexer-Package-for-Internet-Content-Emails.pdf
    - Trained on: N/A
    - Technology status: Product
    - Language(s): language independent

Partner: IDIAP

- Technology: ASR
    - Description: A new work on iterative acoustic model training using untranscribed data and model adaptation using target data, language model extension using web crawled sources
    - Trained on: BABEL data
    - Technology status: Lab
    - Language(s): BABEL material (20+ languages)

- Technology: Speaker ID - text-independent
    - Description: An X-vector (text-independent) baseline[21]
    - Trained on: Speech Recognition Evaluation Data (Telephone, Microphone)
    - Technology status: Tech transfer release
    - Language(s): Language independent

- Technology: Speaker ID - text-dependent
    - Description: phonetic based text-dependent baseline
    - Trained on: Speech Recognition Evaluation Data (Telephone, Microphone)
    - Technology status: Lab
    - Language(s): majority of languages supported

Partner: Phonexia

- Technology: Speaker ID
    - Description: https://www.phonexia.com/en/product/speaker-identification
    - Trained on: Telephony data
    - Technology status: Product
    - Language(s): language independent

- Technology: Language ID
    - Description: https://www.phonexia.com/en/use-case/government
    - Trained on: Telephony data
    - Technology status: Product
    - Language(s): Afan Oromo, Albanian, Amharic, Arabic, Arabic Gulf, Arabic Iraqi, Arabic Levantine, Arabic Maghrebi, Arabic MSA, Azerbaijani, Bangla Bengali, Bosnian, Burmese, Chinese Cantonese, Chinese Dialects, Chinese Mandarin Creole, Croatian, Czech, Dari, English American, English British, English Indian, Farsi, French, Georgian, German, Greek, Hausa, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Khmer, Kirundi

[21] https://www.idiap.ch/en/tech-transfer/idiap_portfolio.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Kinyarwanda, Korean, Lao, Macedonian, Ndebele, Pashto, Polish, Portuguese, Punjabi, Russian, Serbian, Shona, Slovak, Somali, Spanish, Swahili, Swedish, Tagalog, Tamil, Thai, Tibetan, Tigrigna, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese

- Technology: Speech to text
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Product
  - Language(s): Czech (Czech Republic), Croatian, German (Germany), English (US), Spanish (Latin America) , French (France), Italian (Italy), Dutch (Netherlands), Polish (Poland), Russian (Russia), Slovak (Slovakia), Arabic, Chinese, Farsi (Iran)

- Technology: Keyword spotting
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Product
  - Language(s): Czech (Czech Republic), Croatian, German (Germany), English (US), Spanish (Latin America), French (France), Italian (Italy), Dutch (Netherlands), Polish (Poland), Russian (Russia), Slovak (Slovakia), Arabic, Chinese, Farsi (Iran), Turkish (Turkey)

- Technology: Speech quality estimation
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Product
  - Languages: Language independent

- Technology: Voice activity detection
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Product
  - Languages: Language independent

- Technology: Diarization
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Beta
  - Languages: Language independent

- Technology: Age estimation
  - Description: https://www.phonexia.com/en/use-case/government
  - Trained on: Telephony data
  - Technology status: Product
  - Languages: Language independent

- Technology: Denoiser

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- Description: https://www.phonexia.com/en/use-case/government
- Trained on: Telephony data
- Technology status: Beta
- Languages: Language independent

Partner: LUH

- Technology: Topic modelling
  - Description: https://github.com/smutahoang/ttm
  - Trained on: Short texts
  - Technology status: Prototype
  - Language(s): language-independent

Partner: USAAR

- Technology: Named-Entity extraction (Low-Resource)
  - Description: https://github.com/uds-lsv/noise-matrix-ner
  - Trained on: News text
  - Technology status: Lab
  - Languages: Language independent

- Technology: Automatic Named-Entity Annotation
  - Description: Under submission
  - Trained on: no training needed
  - Technology status: Prototype
  - Languages: Language independent

- Technology: Relation Extraction
  - Description: https://github.com/uds-lsv/relationfactory
  - Trained on: Wikipedia text
  - Technology status: Lab
  - Languages: Language independent

Partner: AIRBUS

- Technology: Video / Image processing: Image indexing and search optimized on places
  - Description: Deep Learning based signature extraction and indexing pipeline
  - Trained on: Cleaned Google Landmark Dataset. Tested on ROxford + RParis
  - Technology status: Prototype
  - Languages: N/A

Partner: BUT

- Technology: Speaker verification
  - Description: A large collection of speaker recognition systems, see [22] and [23]

[22] MATĚJKA Pavel, PLCHOT Oldřich, GLEMBEK Ondřej, BURGET Lukáš, ROHDIN Johan A., ZEINALI Hossein, MOŠNER Ladislav, SILNOVA Anna, NOVOTNÝ Ondřej, DIEZ Sánchez Mireia and ČERNOCKÝ Jan. 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE. Computer Speech and Language, vol. 2020, no. 63, pp. 1-15. ISSN 0885-2308.

[23] ALAM Jahangir, BOULIANNE Gilles, GLEMBEK Ondřej, LOZANO Díez Alicia, MATĚJKA Pavel, MIZERA Petr, MONTEIRO Joao, MOŠNER Ladislav, NOVOTNÝ Ondřej, PLCHOT Oldřich, ROHDIN Johan A., SILNOVA Anna, SLAVÍČEK Josef, STAFYLAKIS Themos, WANG Shuai and ZEINALI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

  - Trained on: Telephony and/or multimedia data
  - Technology status: Lab
  - Languages: Mainly English, works reasonably for other languages and can be adapted for better performance

- Technology: Speaker diarization
  - Description: Different variants of variational Bayes speaker diarization systems[24]
  - Trained on: Telephony and/or multimedia
  - Technology status: Lab
  - Languages: Mainly English, works reasonably for other languages

- Technology: Language identification
  - Description: DNN embedding based system, with classification performed by Gaussian Linear Classifier[25].
  - Trained on: Telephony and/or multimedia, mainly data distributed by U.S. NIST and LDC.
  - Technology status: Lab
  - Languages: Amharic, Arabic_Egyptian, Arabic_Iraqi, Arabic_Levantine, Arabic_Maghrebi, Arabic_MSA, Assamese, Azerbaijani, Bangla/Bengali, Bosnian, Cantonese, Cebuano, Chinese Min, Southern Min, Chinese Wu, Shanghai Wu, Creole (Hatian Creole French), Croatian, Czech, Dari, Dholuo, EnglishIndian, EnglishAmerican+FAE+unknown dialects, Farsi_Persian, French, Georgian, Guarani, Hausa, Hindi, Igbo, Japanese, Javanese, Kazakh, Korean, Kurdish, Laotian, Lithuanian, Mandarin, Halh_Mongolian, Pashto, Punjabi, Polish, Portuguese, Russian, Spanish_latam, Slovak, Swahili, Tagalog, Tamil, Telugu, Thai, Tibetan, Tigrigna, Pisin, Turkish, Uighur, Ukranian, Urdu, Uzbek, Vietnamese, Zulu

- Technology: Speech to text
  - Description: A large collection of speech to text systems[26]
  - Trained on: Various data from LDC, and acquired in collaborative projects.
  - Technology status: Lab
  - Languages: English (incl. Non-native), Levantine Arabic, Gulf Arabic, languages from DARPA and IARPA programs: Somali, Swahili, Cantonese, Assamese, Bengali, Pashto, Turkish, Georgian, Tagalog, Vietnamese, Haitian Creole, Lao, Tamil, Zulu, Kurmanji, Tok Pisin, Cebuano, Kazakh, Telugu, Lithuanian, Guarani, Javanese, Igbo, Mongolian, Dholuo, Guarani, Amharic.

Hossein. ABC NIST SRE 2019 CTS System Description. In: Proceedings of NIST SRE 2019. Sentosa, Singapore, 2019

[24] DIEZ Sánchez Mireia, BURGET Lukáš, LANDINI Federico Nicolás and ČERNOCKÝ Jan. Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, 2019, 28- 1, pp. 355-368.

[25] LOZANO Díez Alicia, PLCHOT Oldřich, MATĚJKA Pavel, NOVOTNÝ Ondřej and GONZALEZ-RODRIGUEZ Joaquin. Analysis of DNN-based Embeddings for Language Recognition on the NIST LRE 2017. In: Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop. Les Sables d'Olonne, 2018, pp. 39-46. ISSN 2312-2846.

[26] KARAFIÁT Martin, BASKAR Murali K., SZŐKE Igor, MALENOVSKÝ Vladimír, VESELÝ Karel, GRÉZL František, BURGET Lukáš and ČERNOCKÝ Jan. BUT OpenSAT 2017 speech recognition system. In: Proceedings of Interspeech 2018. Hyderabad, 2018, pp. 2638-2642.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

4. Initial tests on ROXANNE data-sets

4.1 Motivation for data selection

Acquiring real investigation data for testing speech, text and video technologies is (despite number of LEA partners present in the consortium) for the moment not possible due to legal and ethical issues. While work is in progress on Roxanne simulated data (mimicking real data, Task 4.6 in WP4), it was necessary to bootstrap technical work and cooperation among partners. More details can also be found in D4.1 deliverable on: "Overview and analysis of lawfully intercepted and publicly available data[27]".

Therefore, work on three data-sets was launched:

1. Crime Scene Investigation (CSI) data
2. Set extracted from NIST Speaker recognition evaluations data
3. ENRON telephone calls.

We are aware that none of these sets is ideal and have carefully judged the advantages and drawbacks of individual data-sets:

Crime Scene Investigation (CSI) data:

+ audiovisual, face and scene recognition and audio processing recognition can be combined.

+ data follow a story of a criminal case.

+ variety of signal qualities for both audio and video recognition.

+ ground-truth transcriptions.

- data is acted, with professional speakers, very far from spontaneous nature of real investigation data.

- limited number of subjects, not suitable for quantitative evaluations.

- persons contain investigators, laboratory technicians, etc, their data does not occur in real investigation data.

- few telephone calls, the audio files are of better quality than the ones encountered in real investigation data.

Set extracted from NIST Speaker recognition evaluations data

+ real telephone calls.

+ sufficient number of speakers to perform statistical analysis of performance.

+ sufficiently long conversations that can eventually be chopped to test robustness.

- data do not follow a logical story line, topics were attributed to conversations by a SW platform.

- no accompanying text transcriptions.

- no related video or text data.

ENRON telephone calls

[27] https://www.roxanne-euproject.org/results

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

+ extracted from a real case that has a “story” and time-line.

+ availability of large quantity of related text materials (ENRON emails).

+ real telephone calls

- low number of speakers involved in both the telephone calls and the emails.

- personal information and illegal activities have often been deleted from audio data.

- no related video data.

4.2 CSI

Data

Crime Scene Investigation (CSI) is a popular criminal investigation television series in the United States[28]. Episodes of the series include a video of around 40 minutes, an audio file, and transcript. The audio and video are extracted from the DVD of the show. The transcripts were published by the University of Edinburgh. The transcripts also contain the role of each speaker (Suspect, Killer or Other). Each episode involves a team of investigators, journalists, suspects, and a killer.

We collected transcripts of 39 episodes, and video/audio of 6 episodes. Each episode involves on average more than 30 characters. Utterances last on average 3 to 4 seconds. This dataset is so far the most complete dataset collected, allowing potential use of all the technologies developed by the partners: automatic speech recognition, speaker identification, speaker diarization, gender and age detection, keyword and topic detection, named entity recognition, place diarization through video, network analysis, etc.

One of the challenges in the CSI dataset is the lack of precision of some timestamp annotations by the University of Edinburgh[29], which leads to having sometimes several speakers in the same utterance. The results presented below take into account this mis-alignment which we plan to take care of by running a speaker diarization on the raw audio files of the episode for example. Another limit of the CSI dataset is the lack of temporal landmarks. A cut in a scene does not indeed mean that the next scene took place only few seconds after the previous one. By default, we just consider the data as sequential, and will not conduct further analysis on time between communications.

Automatic speech recognition on CSI data

In order to determine ASR performance on CSI data, a set of segments was extracted from 6 episodes of CSI. These were transcribed automatically and the resulting transcript compared to a human generated reference. Table 1 provides an overview of these file and the respective transcription word-error-rate (WER, the accuracy simply corresponds to 1-WER).

An analysis of the errors committed yielded the following insights:

– the ASR models (AM and LM, see above) have been trained on data from broadcast-news. As such, the vocabulary and language use reflects topics and styles which are common in news. In contrast, the CSI episodes exhibit special terminology, slang and colloquialisms. The mismatch

[28] https://en.wikipedia.org/wiki/List_of_CSI:_Crime_Scene_Investigation_episodes

[29] L. Frermann, S. B. Cohen, and M. Lapata. 2017. Who-dunnit? Crime Drama as a Case for Natural Language Understanding.Transactions of the Associa-tion for Computational Linguistics, 6:1–15

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 833635.

caused by this (domain mismatch) accounts for a large share of errors. Adjusting the LM and vocabulary is expected to yield improved performance.

- Kind of speech: characters within CSI exhibit accented and (strongly) emotional language, leading to changes in language characteristics which cannot be adequately captured by the ASR models. Adjustment of pronunciations and vocabulary as well as re-training of the LM may yield improved performance.
- Speaking style: dialogues in CSI often contain very short turns resulting in short audio segments. In addition, further speech may be present in the background
- Acoustic/recording conditions: the audio produced by the particular recording conditions does not correspond to the audio used to create the AM of the ASR system, leading to an acoustic-mismatch in conditions. Far-field microphones, reverberation, background noise and a variety of sounds effects all contribute to ASR-problems .
- when measuring performance (calculating the WER), a reference transcript - created by humans - is compared to the output of the ASR system. This comparison is performed using a dynamic alignment between reference and ASR-output. It was observed that due to differences in segmentation of the audio, several instances of mismatched in this comparison took place. This leads to additional phantom-errors, increasing the WER. based on this observation it can be assumed that the calculated WERs form an upper-bound of the actual WER.

Table 1: Parameters of CSI data and ASR.

	s01e07	s01e08	s01e19	s01e20	s02e01	s02e04
Total duration	30:01	29:06	30:30	32:33	30:04	28:25
# segments	316	350	377	410	389	385
# words	3703	3950	5015	4571	5104	4276
WER	59.7%	53.8%	50.8%	56.5%	63.4%	59.6%

Gender detection on CSI data

Gender detection is made through a simple technique. Once the speaker identification has been made, we concatenate all audio recordings of the speaker until the given time-stamp of the new recording. Using this concatenated audio recording, we extract Mel-frequency cepstral coefficients (MFCC) from the speech. These coefficients are well known in speech processing and convey a lot of information. A Gaussian Mixture Model (GMM) was previously trained on Audioset, a large corpus of audio data containing information about the gender of the speakers. The GMM then predicts whether the recording is more likely to belong to a male or a female. On the experiments we ran, expect for the miss-alignments, the system identifies correctly the gender of the speakers.

Speaker Identification on CSI data

Due to the low volume of data available, training a speaker identification system on CSI data is impossible. Therefore, we leverage pre-trained systems, trained on NIST Speaker Recognition Evaluation (SRE) dataset. The pipeline of a speaker identification system is to structure the audio into enrolment and test audio. We select speakers from CSI for which we have at least 30 seconds of audio samples. We then keep 30 seconds as enrolment, and everything in test. This heads to 14 different speakers among the 31 available.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

During the enrolment, in the front end, we extract first extract audio features (MFCCs), perform Voice Activity Detection (VAD) to remove frames without speech. Then, we extract state-of-the-art features (X-vectors). We then have one "model" per speaker. In test, we compare the vector extracted for each speaker in each conversation with the speaker models using Probabilisitic Linear Discriminant Analysis (PLDA). We finally attribute the audio sample to the speaker model heading the highest score. The pipeline is described in the figure below (which corresponds to the general scheme in Fig. 1):

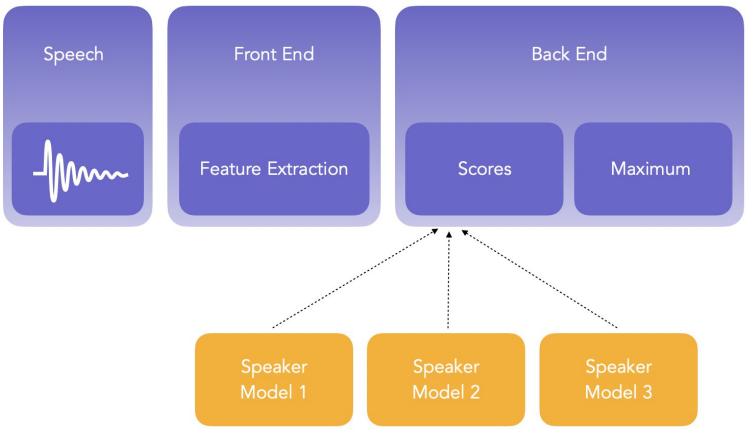

Figure 8: Speaker Identification Pipeline.

On episode 07 of season 01, the speaker identification heads a Top 1 score, i.e. how often the correct speaker model of the correct speaker reached the highest score, of 91.7%. Out of 96 audio files from CSI conversations, 88 were correct. The Top 5 score, i.e. how often the correct speaker model is in the 5 highest scores, is 100%.

Experiments are currently being ran on the improvement of these results using graph knowledge.

All the results (speaker identification, gender detection, Automatic speech recognition) were then included in a network analysis tool. We were able to display for each node in the network, the identity predicted by the speaker identification system and the gender predicted.

Video analysis on CSI data

For video analysis, the targeted capability is a face and/or place diarization service for open source videos enabling the identification of segments sharing a same face or location.

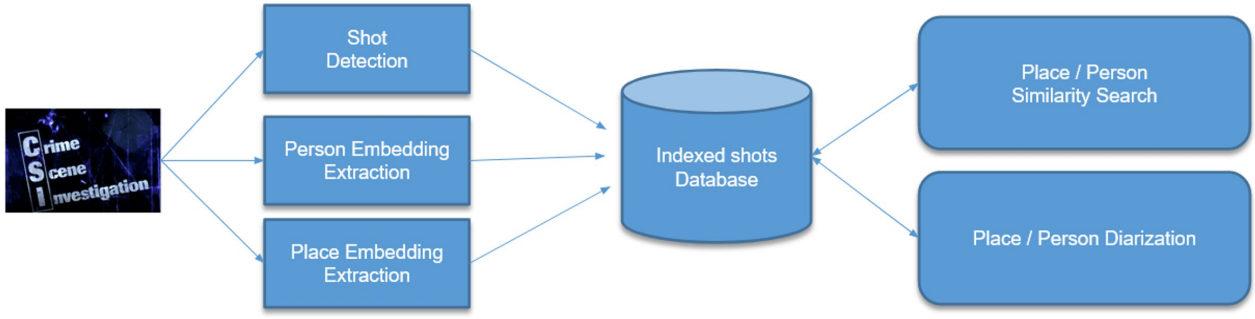

Figure 9: Place / Location diarization core components.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Airbus Defence and space provides its image signature extraction learning and exploitation pipeline described in the following figure.

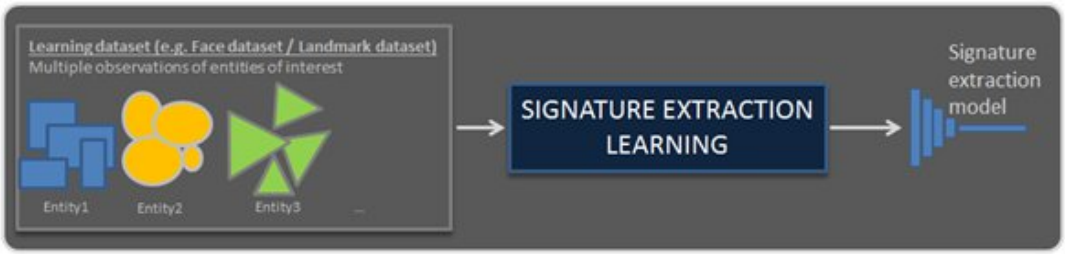

Figure 10: Signature extraction learning pipeline.

The "Signature Extraction Learning" module takes as input a dataset of entities of interest consisting of several observations of a large set of different entities belonging to a same class. In Roxanne the classes that could be considered are:

- Faces (entities are then faces of different persons, for instance from the Celebrity Face Recognition dataset[30] consisting of around 800k images of more than 1000 celebrities)
- Places (entities are then different geographic landmarks, for instance from a cleaned version of the Google Landmark Dataset[31] providing tens of observations for hundreds of thousand different landmarks)

The signature extraction learning pipeline is based on a trainable deep learning network which is iteratively optimized to extract similar signatures for different images of a same entity, and different signatures for images of different entities.

The obtained signature extractor can then be used for image retrieval or near duplicates search purposes as illustrated in the following figure.

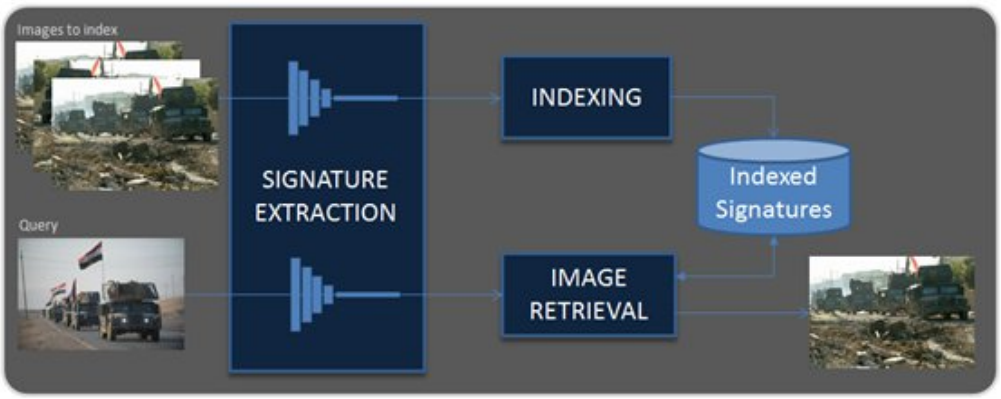

Figure 11: Image Indexing and Retrieval pipeline.

[30] Celebrity Face Recognition Dataset : http://github.com/prateekmehta59/Celebrity-Face-Recognition-Dataset

[31] Google Landmark Dataset v2 : https://github.com/cvdfoundation/google-landmark

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

In the Roxanne project, these modules could be combined to provide a place and/or face diarization service in videos aiming at structuring and summarizing videos according to the faces and places they contain. Besides, the computed video segments will be provided to network analysis modules, as complementary inputs to segments computed by speaker diarization.

The targeted video diarization services are described in the following picture.

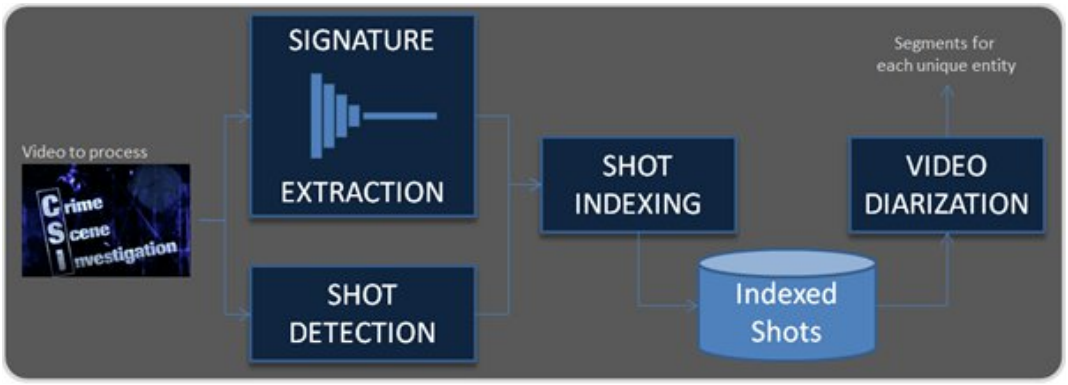

Figure 12: Targeted video diarization pipeline.

The targeted pipeline will combine the signature extraction of one or several entities of interest (e.g. faces and/or places), a shot detection module aimed at detecting shot discontinuities in the video, a shot indexing module resulting in a shot index which will be the main input to the video diarization module.

The output of the video diarization will be a list of segments of unique entities (faces or places).

The available core components of this service were tested on 6 CSI episodes. These include:

- A **shot detection module**: it takes a video file as input and extracts shot limits. In this context, a shot is a continuous footage or sequence between two edits or cuts. Shots are used as primary elements of the diarization process and are defined by a starting time-stamp and an ending time-stamp. Our shot detector is based on a pixel motion detector that tracks pixel motions between subsequent images. If the tracking of most of the points fails, a discontinuity is detected, and a shot limit is marked.
- A **place or face embedding (or signature) extraction module**: this process runs on images taken at a fixed frequency on the input video file. For a given image, a signature is computed using a convolutional neural network trained using deep learning methods. Each signature is an array of 256 floating numbers and contains information about the specific parts of the image which are representative of the observed location or faces. These signatures enable to compare / link images using a simple distance measure which will be close to zero for images taken in a same location or for faces of a same person, and far from zero in other cases.
- An **object detection module**: it runs on images taken at a fixed frequency on an input video file. For a given image, objects are detected and localized with rectangle bounding boxes using a neural network. This network was trained on Open Image Dataset, a dataset containing about 5000 object classes, and is used along with the RetinaFace face detection model, to detect faces in images.
- A **similarity search module**: The similarity search is a process taking as input the signature associated to an image and outputs indexed images that are the most similar given the selected signature. If the location signatures are used, this search returns the images that are containing the same locations. If the face signature is used, this search returns images containing faces of the same person. In our case, the similarity search has been improved to take into account video shots: when a request is done for the signature associated with one image, this signature is compared to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

the signatures of all images belonging to each shot. The distance taken into account is the minimum distance in a given shot. The final output consists of shots observing the same location than the query image in case of location signatures or shots observing the requested person in case of face signatures.

The following images illustrate results of similarity searches performed for a selection of faces or locations on the available CSI episodes. The left column images are the query images, the next columns are the more similar shots found by the similarity search. For each image, the lower left caption indicates the similarity score in percent. Finally, the lower right caption indicates the name of the video and the shot properties (start and end timestamps), showing the identification of same places or same faces across episodes.

Figure 13: Place similarities across episodes (first column = query, other columns: retrieved shots from same locations).

Similarly, face similarities results obtained across available episodes are illustrated in the following picture.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Figure 14: Face similarity results across episodes (Left: face query, other columns: examples of retrieved faces).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Topic detection on CSI data

As an example, we applied a topic detection process in one of the episodes of the CSI dataset, namely the episode S01E07. Figure 15 (left plot) exemplifies the semantically related terms found it in the transcripts, and, on the right side, we show a plot of the categorization result of the same episode. On one hand, the plot on the left depicts the most relevant terminology found for each concept identified in the dataset (words associated with the same concept, appear in the same color), in this example 5 concepts. On the other hand, the plot on the right represents the utterances organization according to the found concepts, where each dot represents one utterance of the S01E07 episode (54 in total), and the number in the circle indicates its associated cluster. For this particular example, the most relevant associated terminology for each cluster is shown in the Table 2:

Table 2: Examples of word clusters on CSI.

Number of Cluster	Associated terminology
Cluster 0	'pretty', 'quick', 'look', 'actually', 'grabbed'
Cluster 1	'picture', 'freak', 'cutting', 'kill', 'process'
Cluster 2	'print', 'sexual', 'beating', 'stab', 'marks'
Cluster 3	'brenda', 'grissom', 'nickwarrick', 'tina', 'collins'

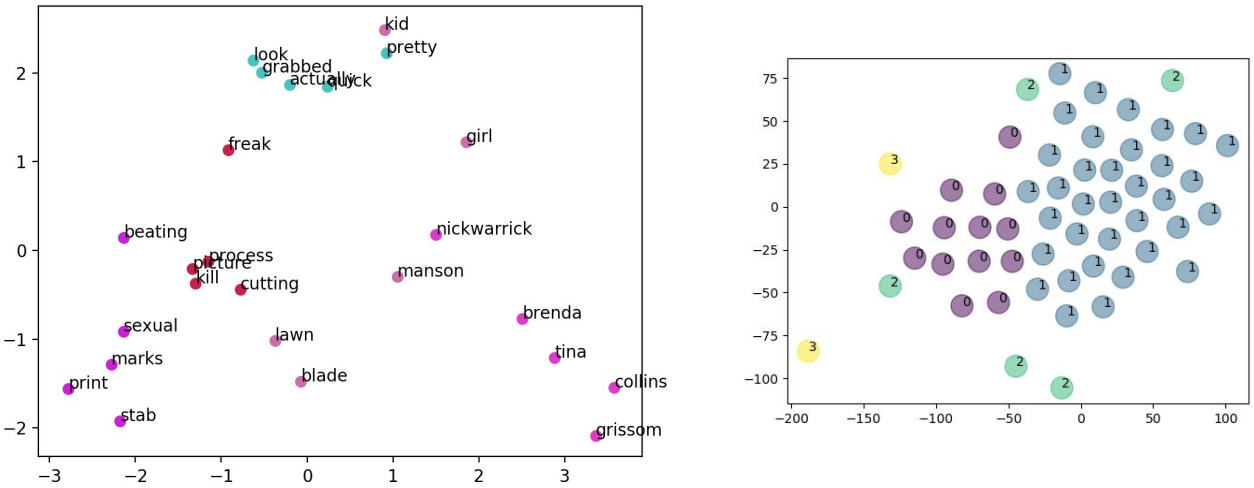

Figure 15: Concepts (left) and Clusters (right) organization of S01E07 episode from the CSI dataset.

4.3 NIST SRE

Data

National Institute of Standards and Technology (NIST) has performed speaker recognition evaluations (SRE) for more than two decades. The data from these evaluations is attractive because they consist of telephone calls and the numbers of speakers and calls are large. However, it has the drawback that it is not designed for network analysis. In order obtain a data set suitable for network analysis we used datasets from several different evaluations carried out by NIST from 2005 to 2010 (NIST Speaker Eval 2005, 2006,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

2008 and 2010) and tried to find clusters of speaker such that for any two speakers in a cluster, say A and B, there is "path of calls" between them meaning that Speaker A has talked to Speaker X, which in turn has talked to speaker Y, ..., which in turn has talked to speaker B. We found two big such clusters. The biggest cluster has 2046 speakers who make on average 6.1 calls to on average 5.6 different speakers. The second biggest cluster has 463 speakers who make on average 3.0 calls to on average 2.9 different speakers. The remaining "clusters" are small, usually having only 2 speakers[32]. The obtained data set is thus not one specific NIST dataset but rather a combination of several existing NIST datasets.

Speaker identification on NIST data

As first experiment, we considered all possible trials (utterance pairs) that could be formed from the data set. This is approximately 100 million non-target (different-speaker) trials and 60K target (same-speaker) trials. We used a time delay neural network based embedding extractor[33,34] trained on the VoxCeleb[35] data set. As backend we used cosine similarity and probabilistic linear discriminant analysis (PLDA). As evaluation metric, we used equal error rate (EER).

The results were EER 3.23% and 2.44% for cosine similarity and PLDA respectively. When comparing these results to standard (NIST) benchmarks, it should be noted that the data set in the experiment is relatively easy because it includes

- Only English data
- Only long utterances
- Cross-gender trials
- Possibly same phone number target trials

On the other hand, the models are not ideal for the task because the VoxCeleb data it was trained on is quite different from the NIST data both with respect to the domain (multimedia vs telephone) and the duration (short vs long).

Automatic speech recognition on NIST data

The audio content selected for processing in ROXANNE consist of telephone conversations mostly in American English, split into two channels (A and B) and saved as two separate files in raw format. For the ASR experiments, these mono files were converted into WAV and processed with an International English model.

Figure 16 shows an excerpt from the recognizer's XML output for an example audio file. There is only a single word error in this piece, where the utterance "hate" is mis-recognized as "haid", with a low confidence score. Since no manual transcripts are available for this data-set, we cannot provide a numerical evaluation of the results. However, from subjective evaluation, the main causes of potential errors are seen as follows:

- Audio quality mismatch: The current ASR system, which natively accepts 16 kHz-sampled audio , produces suboptimal acoustic features with 8 kHz telephony conditions (the wave files are up-sampled before processing)
- Topic mismatch: The language model of the current system is optimized for broadcast-news transcriptions and this may cause mismatches in predicting the words of a natural telephone conversation.

[32] These numbers are preliminary. In particular, issues related to same utterances being present in more than one databases need to be sorted out. The data set may be updated in due course of the project.

[33] D. Snyder, et al.,Speaker recognition for multi-speaker conversations using x-vectors, in: ICASSP, 2019

[34] P. Matějka et al. 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE, CSL 2020.

[35] A. Nagrani et al., VoxCeleb: a large-scale speaker identification dataset, in INTERSPEECH 2017

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- Accented speech: Most of the speakers in this data-set use a heavily accented English, which affects the recognition accuracy.

All of these issues will be addressed during the course of the project. The ASR system will be retrained with 8kHz data, the language models will be adapted for the relevant topics, and the accent detection system will be used as a guideline to choose the right model for decoding.

```
<segment  start="0" end="2400" nonSpeechRatio="0.02125" snr="1.15011" type="speech">
<nbest id="0" score="1.00">
<word conf="0.66" end="25">For</word>
<word conf="0.77" end="66">travel</word>
<word conf="0.73" end="344">it's</word>
<word conf="0.98" end="406">hard</word>
<word conf="1.00" end="412">I</word>
<word conf="1.00" end="433">mean</word>
<word conf="0.98" end="454">I</word>
<word conf="0.99" end="475">if</word>
<word conf="1.00" end="491">it's</word>
<word conf="1.00" end="502">up</word>
<word conf="1.00" end="518">to</word>
<word conf="1.00" end="535">me</word>
<word conf="1.00" end="551">I'd</word>
<word conf="1.00" end="578">rather</word>
<word conf="0.91" end="614">drive</word>
<word conf="1.00" end="664">everywhere</word>
<word conf="0.96" end="677">I</word>
<word conf="1.00" end="700">just</word>
<word conf="0.51" end="739">haid</word>
<word conf="1.00" end="764">going</word>
<word conf="1.00" end="778">on</word>
<word conf="0.88" end="823">planes</word>
<word conf="1.00" end="860">because</word>
<word conf="1.00" end="868">I</word>
<word conf="1.00" end="890">have</word>
<entity>
<etype origin="number" value="2">cardinal</etype>
<ewords><word conf="0.56" end="907">two</word></ewords>
</entity>
<word conf="1.00" end="935">small</word>
<word conf="1.00" end="968">children</word>
```

Figure 16: An example XML output from the ASR system.

4.4 ENRON

Data

Enron is a company that remains famous for the amount of wilful corporate fraud and corruption it was involved in. Two years after the bankruptcy of the company, emails from 150 managers were made public by the Federal Energy Regulatory Commission (available for instance here[36]). Overall, over 500'000

[36] https://www.cs.cmu.edu/~enron/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

emails were collected, with the content of each email and email addresses of each recipient, email address of the sender as well as the full name. Some of these emails highlighted integrity issues from some of the managers. Several laboratories, including SRI International, worked on removing the named of employees involved in such activities.

In 2004, telephone recordings of several managers were made public[37]. The recordings consist of 64 recordings, of 5 minutes on average, each recording consisting of several phone calls, for a total of 6 hours of recordings. All transcripts were also made public in PDF image version, with the first name of each speaker.

This dataset has the great advantage to be from a real-world scenario. It also involves corporate fraud, and the structure of the network, the frequency of the exchange between characters and the content should reflect it. All the timestamps of the phone calls and the emails are available, meaning that a temporal analysis can be conducted.

In order to prepare the dataset, transcripts first had to be extracted in text files. These transcripts will be used as a ground truth for the automatic speech recognition system evaluation. We used an optical character recognition system, and cleaned the output by hand.

Parts of the audio files were deleted when the information they conveyed was private. This results in some blanks in the conversations. Also, different phone calls can occur in the same recording. The phone calls are however identified at first by a ring tone. We were able to split the recordings on the ring tones, and identify the unique phone calls in each recording. The new recordings were then saved, and the timestamps corresponding to the phone call was derived from the timestamps of the recording to which we add the time before this phone call starts. Note that this approach comes with some limits. We do not have any guarantee that the phone calls were made sequentially, without any break. We listened to some audio files, and the order of the phone calls seems to follow a logic. If A was talking to B, and A needs to confirm an information with C, then A will call C after. But the ground truth timestamp of this second call cannot be known.

The next task to prepare this dataset is to match the names of the characters in the phone calls and in the emails. The emails carry information about the first name and the last name of each of the 150 managers. The phone calls only carry information about the first name. Therefore, we must match the recordings based on the first name of the characters, and drop characters for which we have multiple candidates. This leads us to 15 speakers, who exchanged 1638 phone calls or emails between June 2000 and March 2001. We restricted the period over which we consider the emails, in order to match the period over which the phone calls were recorded.

Planned experiments

The consortium partners started to work on ENRON data shortly before the submission of this Deliverable D5.1. Therefore, results are not yet available. We will compare our results with those obtained by other researchers[38]. The advantage of ENRON data is their placement on a clear time-line and correlation with text materials, so that experiments close to a real investigation work can be performed.

[37] https://web.archive.org/web/20070219025955/http://www.enrontapes.com/files.html

[38] Gao, Ning, Gregory Sell, Douglas W. Oard, and Mark Dredze. "Leveraging side information for speaker identification with the Enron conversational telephone speech collection." In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 577-583. IEEE, 2017.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

5. Preparation for integration

In this section, we are dealing with the preparation of speech, text and video technologies for the integration in Roxanne platform and for the first round of Field tests (scheduled originally for M9 but postponed due to the corona-virus crisis). In M4-M5 of the project, the selection of technologies was discussed and partners agreed that for the platform, one technology should be made ready (marked in green in Table 3) while simultaneously, alternative or more advanced versions are being prepared for the next Milestones (yellow). The following sections provide the important implementation details of the technologies, while the integration aspects (such as API description, notes on dockerization, license management, etc.) are left for the respective WP7 deliverables.

Table 3: Technologies to be integrated in the ROXANNE Platform = For the first field test = Later

	Technology\Partner	IDIAP	BUT	PHO	SAIL	USAAR	AIRBUS
Speech	Speaker ID						
	Diarization						
	Gender detection						
	Accent detection						
	Age detection						
	Speech-to-text						
Text	Entity detection						
	Topic detection						
Video	Face diarization						
	Place diarization						

5.1 Diarization, Gender detection, Age estimation, Language and dialect recognition and Speaker recognition - PHO

The mentioned technologies are implemented in Phonexia Speech Engine v3 (SPE3). It is a server application with REST API interface through which all available speech technologies can be accessed. Both, Linux 64bit and Windows 64bit operating systems are supported. SPE3 provides RESTfull application programming interface to access various technologies. Aside from technologies themselves, the SPE has implemented other functionalities supporting work with speech technologies, recordings and streams, and others.

The main purpose of SPE is to work as a processing unit for all Phonexia technologies. It has the following main features:

- **entity oriented** - when processing any recording or stream with any technology, SPE has information about this particular recording or stream as long as it exists. Once the recording is deleted, or stream is ended, SPE removes all information, metadata and technology results from the database.
- **file processing and stream processing** (dependent on available technologies). HTTP or RTP streams are supported.
- **user management -** there are several roles defined with various rights. This enables to let various SPE users work with their data only and prevent them to see any recordings, metadata or technology results of other users.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- **load management -** SPE can queue incoming requests and serve them one by one based on the capacity of the current installation. This means that user or partner application can request any number of queries and can just wait till all are answered.
- **audio management -** SPE can split stereo recordings, cut one audio to several files, save incoming stream and others.
- **flexibility in providing results -** results are returned in XML/JSON format. The result can be obtained using several ways - polling, WebSockets or webhooks.

SPE supports a variety of Input formats. For formats, which are not natively supported by SPE, an embedded third party format converter (FFmpeg or Sox) is used. It can identify speech on audio recordings, in video recordings with audio channels and it can also identify speech in streams. Typical loss less format for processing - WAV, FLAC or RAW (8 or 16 bits linear coding), A-law or Mu-law, PCM, 8kHz or 16kHz sampling frequency - are supported.

The processing speed varies depending on technology:

- VAD: This type of process is very fast. The speed might be approx. 150x faster than real-time processing on 1 CPU core, ie. standard 8 CPU core server processes 28,800 hours of audio in 1 day of computing time.
- Gender and Age estimation: first, a voiceprint is extracted from a speech recording that is a unique biometric identifier describing a speaker. The voiceprint extraction is the most time-consuming part of the process, but even so, it is very fast. Typically, it is possible to process 50 minutes of speech in 1 minute of real-time on one CPU core. The following age and gender estimation is extremely fast (several thousands voice-prints per second).
- Language identification: from 4 to 50x faster than real-time processing on 1 CPU core (e.g. standard 8 CPU core server can process 3,840 hours of audio in 1 day of computing time). The speed varies according to the model of technology used.
- Speaker identification: see above for the extraction of voice-prints, the same ones as for age and gender estimations are used. The following speaker recognition is extremely fast (several thousands voice-print comparisons per second).

5.2 Speech to text - SAIL

Several ASR modules will be made available as docker images, one image per language. The module will contain the ASR-engine as well as a model for a particular language. All input is assumed to be in that particular language. The module will accept an audio file as input and produce a transcript of words, associated time-tags and confidence values as output. The output format is JSON

Internally, the ASR module segments the audio into homogeneous sections for normalization purposes as well as for (internally) pipelined processing. Within the project, an external, pre-computed segmentation may be provided and this processing step will be skipped. Sections containing non-speech (silence, noise,...) will be skipped by speech-processing. A predefined level of noise and/or music content is specified in a parameter file and can also be used to eliminate processing of segments. The remaining (speech) segments are processed in a pipelined manner, each segment appearing as such in the resulting transcript. Finally, all segments are combined and returned in a single JSON structure.

The ASR model externalizes a REST interface allowing to

- submit an audio-file for transcription (parameters are the audio file and the language the audio is in)
- query the status of transcription and
- receive the result of the transcription (as JSON)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Depending on the language, models may use up to 2GB of memory. This may lead to an initial latency when first using the component. As the model stays in memory, subsequent processing will be faster and faster-than-realtime processing can be expected.

5.5 Entity detection - USAAR

The Named-Entity Recognition module is available through docker container supporting REST APIs for data transmission.

We support extracting named-entities from a input text. The input will be a JSON file including the text that needs to be processed, along with all parameters required for running the correct model (language of the text, model type etc.). The output will also be a JSON file containing entity tags found in the input text, along with the positions they appear.

Depending on the model size, the text-processing module (trained neural networks) will be either integrated in the container or stored on the host machine and can be accessed by the container.

5.10 Topic detection - IDIAP

The topic detection module is available through docker container supporting API interfaces for easy integration. It follows the standard JSON format for content specification.

The basic operations through the API interfaces includes:

- Get the configuration parameters.
  - method, number of concepts, number of topics, number of words per topic, BERT -model, stop-words file, preprocessing option, and output file.
- Set the configuration parameters.
  - Method, number of concepts, number of topics, number of words per topic, preprocessing option.
- Get the topic detection results based on configuration.
  - concepts, clusters with associated keywords, output plots, and models.
- Get inference for the input text
  - concepts, associated topic (cluster) with its corresponding keywords, and output plots.

5.11 Video analysis - AIRBUS

Video analytics modules (mainly place and / or face diarization) will not be integrated until the second field tests.

6. Activities related to ROXANNE

This section covers the achievements and activities of ROXANNE research teams in areas connected to the project, although not directly connected to ROXANNE data or integration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

6.1 2019 NIST Speaker recognition evaluation

Since 2005, members of the BUT team have been participating in speaker and language recognition evaluations organized by the U.S. National Institute of Standards and Technology (NIST). An evaluation for speaker recognition was held in the end of 2019 with the main goal of exploring, measuring and supporting the development of the latest technologies in speaker recognition in two areas:

1. Conversational telephone speech (CTS)
2. Audiovisual recognition from videos from YouTube or similar channels (VAST).

BUT and Phonexia participated in a consortium with other academic and industrial partners (Omilia Conversational Intelligence - Athens, Greece, CRIM - Montreal, Canada, Speechlab - Shanghai Jiao Tong University, China, and Audias-UAM - Universidad Autonoma de Madrid, Spain). Our consortium, in competition with 50 other groups from around the world, has once again confirmed that it is one of the world's leaders in speaker recognition.

The CTS system[39] was a fusion of four systems based on x-vectors (low-dimensional vectors from a neural network). For the VAST domain[40] we also used a system for verifying the speaker from the video, which we merged with audio systems. While the performance of our audio system was excellent, in the next evaluations, we need to be careful about a selection of robust state-of-the-art video component.

6.2 Webinar on Phonexia speech technologies

Phonexia released the first 30-minute-long Webinar with a topic: "Discover the latest version of our Speech Platform", covering the latest release of our Speech Platform for Government.

Our experts covered the most essential product updates such as new languages for Speech to Text (STT) and Keyword Spotting (KWS), improved STT accuracy, Speaker Identification (SID) improvements, and many more.

6.3 Spring 2020 speech evaluations

BUT SID/LID and ASR groups are working towards two challenges in spring 2020:

1. *CHiME-6* is about distant multi-microphone conversational speech diarization and recognition in everyday home environments[41]. The results will be announced at the (virtual) workshop on Monday 4th May, after the submission of this deliverable.
2. The goal of *Short-duration Speaker Verification (SdSV) Challenge 2020*[42] is to evaluate SdSV with varying degree of phonetic overlap between the enrolment and test utterances. The challenge is

[39] ALAM Jahangir, BOULIANNE Gilles, GLEMBEK Ondřej, LOZANO Díez Alicia, MATĚJKA Pavel, MIZERA Petr, MONTEIRO Joao, MOŠNER Ladislav, NOVOTNÝ Ondřej, PLCHOT Oldřich, ROHDIN Johan A., SILNOVA Anna, SLAVÍČEK Josef, STAFYLAKIS Themos, WANG Shuai a ZEINALI Hossein. ABC NIST SRE 2019 CTS System Description. In: Proceedings of NIST. Sentosa, Singapore: United States Department of Commerce, National Institute of Standards and Technology, 2019.

[40] ALAM Jahangir, BOULIANNE Gilles, BURGET Lukáš, GLEMBEK Ondřej, LOZANO Díez Alicia, MATĚJKA Pavel, MIZERA Petr, MOŠNER Ladislav, NOVOTNÝ Ondřej, PLCHOT Oldřich, ROHDIN Johan A., SILNOVA Anna, SLAVÍČEK Josef, STAFYLAKIS Themos, WANG Shuai, ZEINALI Hossein, DAHMANE Mohamed, ST-CHARLES Pierre-Luc, LALONDE Marc, NOISEUX Cédric a MONTEIRO Joao. ABC System Description for NIST Multimedia Speaker Recognition Evaluation 2019. In: Proceedings of NIST 2019 SRE Workshop. Sentosa, Singapore: United States Department of Commerce, National Institute of Standards and Technology, 2019.

[41] https://chimechallenge.github.io/chime6/

[42] https://sdsvc.github.io/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

organized by Hossein Zeinali (ex BUT senior researcher) and runs on Farsi DeepMine database[43], that is one of the largest corpora for text-dependent and text-independent SID. The results are available in the form of "leader-boards" - BUT team scored the 1st in the text-dependent task, and had good results in the text-independent one. All results will be thoroughly discussed during the SdSV Challenge 2020 special session at Interspeech 2020.

6.4 SAIL's new CAVA Framework

SAIL LABS introduced an update to a component of their Media Mining Indexer software, called the CAVA: Continuous Automatic Vocabulary Adaptation Framework. CAVA allows semi-automatic and periodic updates of the ASR vocabulary and language model from relevant and new data, making the ASR system to be self-aware of previously unknown words, such as "Brexit" or "coronavirus". These tasks are performed by a Web crawler searching for such words along with their contexts, some NLP components, core algorithms for vocabulary selection, and the Language Model Toolkit (LMT), SAIL's tool for updating ASR models. The framework was presented during the Show&Tell session at Interspeech 2019 in Graz, Austria.[44]

7. Future work

7.1 Within ROXANNE

In **SID**, the general trend is lowering the error rates approximately to 1/2 every two years, but challenges still remain in language dependency (especially for Asian languages), in dependency on duration of available recordings and also in inappropriate performance on low bit rate (VoIP-like) communication channels. The same holds for speaker diarization. Increasing the robustness will be tackled

- in the standard way (ie. without relation to network analysis) where work will be done on novel neural architectures, data preparation and probabilistic interpretation of results. The progress will be checked in open international evaluations.
- in connection to network analysis, by making use of prior information, for example by setting the minimum number of speakers in the conversation (a typical problem in diarization) or by exploiting the conversational nature of speech.

Concerning **ASR**, the following activities have been carried out or are still ongoing:

- transcription of several episodes of CSI data -> production of baseline results
- transcription of several sets of NIST data, including the improvement of segmentation to account for structure of audio

As a next step for ASR, the segmentation component will be extended to also accept external (pre-computed) segmentation information. This will result in a skipping of the internal one, yielding transcripts with segments which better align with those of other technologies. Furthermore, the training of ASR models for low-quality audio (telephony) is planned.

[43] http://data.deepmine.ir/en/

[44] Erinç Dikici, Gerhard Backfried, Jürgen Riedler, The SAIL LABS Media Mining Indexer and the CAVA Framework, INTERSPEECH 2019: 4630-4631.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

The component for the **detection of Named-Entities** has been extended and will be provided as a separate component. This is expected to allow to tag crime-related content and criminal jargon characteristic of different types of organized crime.

For **video analysis**, the performed similarity search tests show that both signature extraction modules (faces and places) are powerful enough to retrieve consistent sets of places and faces. These qualitative tests will be complemented with quantitative tests, giving more precise results using metrics based on ground truth data extracted from multiple videos. Based on these core modules, the diarization processing chain will be developed exploiting signature similarity and clustering techniques.

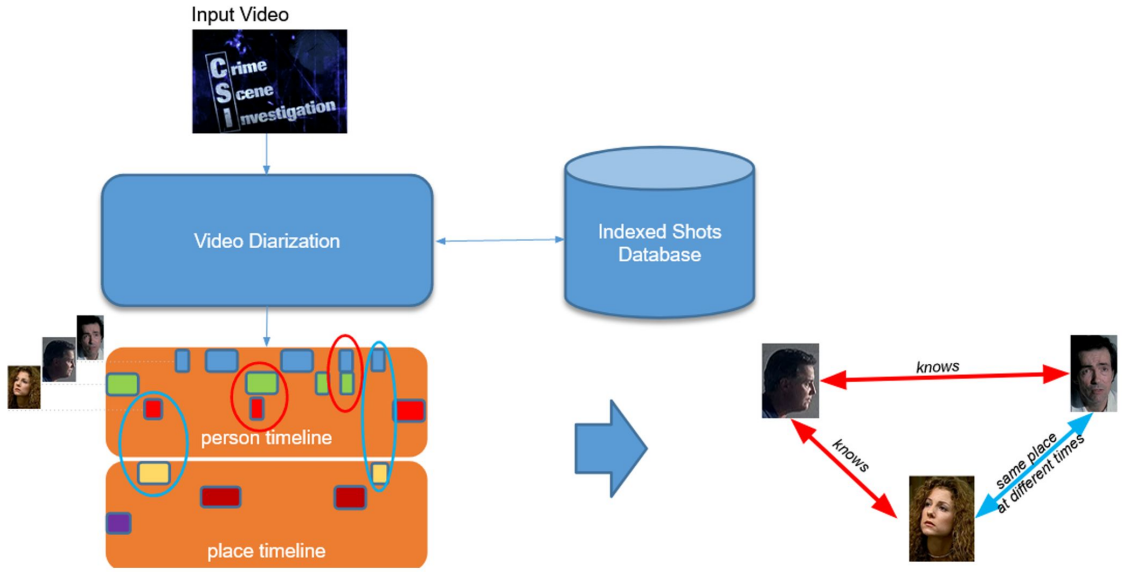

Figure 17: Exploitation of the person and place diarization outputs

Using the information from timeline and shots it will be possible to establish links between people, and links between places and people and higher level information.

7.2 Related to ROXANNE

SAIL LABS is a member of the H2020 ELG (The European Language Grid) project[45]. In line with the goals of the Multilingual Digital Single Market, ELG aims to become the primary platform for natural language processing (NLP) in Europe. It will not only provide the infrastructure for a wide portfolio of NLP technologies and resources but also function as a marketplace for innovation in NLP and AI. SAIL aims to connect the results of ROXANNE (its own as well as those of partners) with those of ELG, e.g. by communication of results, open-calls, opportunities for partnering and events and ultimately even by the integration of technologies. Furthermore, where possible SAIL aims to align technology stacks in order to enable such connections (e.g. the choice of containerization and orchestration).

BUT will analyze the results of SdSV and Chime6 evaluations and assess their relevance for Roxanne. Another edition of NIST Speaker recognition evaluation is being scheduled for the end of 2020 or beginning 2021, intensive collaborative efforts will take place at BUT for this evaluation.

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

In September 2021, the most important speech conference Interspeech 2021[46] will be organized in Brno (after Hyderabad, India 2018, Graz , Austria 2019, Shanghai China 2020). The conference regularly attracts >1500 participants from academia, industry and Government, the 2019 edition had a record participation of over 2000 speech enthusiasts. Interspeech 2021 is strongly related to ROXANNE, as it is organized by the BUT team, Dr. Petr Motlicek (IDIAP, the project PI) serves as one of the technical chairs and Phonexia (also located in Brno) provides an industrial support.

[46] https://www.interspeech2021.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.