IDIAP RESEARCH REPORT

TOWARDS A BREAKTHROUGH SPEAKER IDENTIFICATION APPROACH FOR LAW ENFORCEMENT AGENCIES

Khaled Khelif yann Mombrun Petr Motlicek
Gerhard Backfried Damien Kelly Farhan Sahito
Gideon Hazzani Luca Scarpatto
Emmanouil Chatzigavriil

Idiap-RR-29-2017

OCTOBER 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Towards a breakthrough speaker identification approach for law enforcement agencies

SIIP – Speaker Identification Integrated Project

Khaled Khelif, Yann Mombrun
Airbus
Elancourt, France
Firstname.lastname@airbus.com

Petr Motlicek
IDIAP
Martigny, Switzerland
Petr.motlicek@idiap.ch

Gerhard Backfried
SAIL LABS
Vienna, Austria
Gerhard.Backfried@sail-labs.com

Damien Kelly
Data Fusion International
Dublin, Ireland
Damien.kelly@datafusion.ie

Farhan Sahito
INTERPOL
Lyon, France
F.sahito@interpol.int

Gideon Hazzani
VERINT
Herzliya, Isreal
Gideon.Hazzani@verint.com

Luca Scarpato
NUANCE
Torino, Italy
Luca.Scarpato@nuance.com

Emmanouil Chatzigavriil
SingularLogic
Athens, Greece
echatzigavriil@ep.singularlogic.eu

***Abstract*—This paper describes a high performance innovative and sustainable Speaker Identification (SID) solution, running over large voice samples database. The solution is based on development, integration and fusion of a series of speech analytic algorithms which includes speaker model recognition, gender identification, age identification, language and accent identification, keyword and taxonomy spotting. A full integrated system is proposed ensuring multisource data management, advanced voice analysis, information sharing and efficient and consistent man-machine interactions.**

Keywords: speaker identification, audio and voice analysis, OSINT, Forensics, LEA

I. INTRODUCTION

To date, one of the prominent challenges encountered by LEAs and security agencies (SA) in fighting crime and terrorism is the use of multiple and arbitrary identities by terrorists and criminals. Being tracked by LEAs, they use increasingly sophisticated means to hide their real identity and real activities in the telecommunication domain (PSTN, Cellular, SATCOM) and in the Internet domain (peer to peer VOIP apps and social media) in order to mislead the LEAs and to make their tracking or monitoring very difficult or almost impossible. For example, criminal and terrorists can use randomly multiple prepaid cell-phones, replacing and switching between them frequently, knowing that linking prepaid cell-phone identity (MSISDN/IMSI/IMEI) with the real subscriber identity is very difficult. Moreover when using post-paid cell-phones, the criminals/terrorists change the SIM cards occasionally creating a real difficulty to link between all these SIM cards identities (‘IMSIs’). They may even use any public phone in the street or in a nearby coffee shop, a roamer phone or even a passer-by cell phone. In the Internet, the criminals and terrorists use easily, many different identities and nick names through various Voice Over IP applications.

Another challenge that LEAs/SAs face is the ‘Unknown 2nd side’ (or unknown participant) in a conversation with a suspect which is being lawful intercepted. This problem is another side of the first challenge above and is derived from it. It is important for LEAs to know who both participants are in a lawfully intercepted call, as unknown 2nd side conversations are estimated to be 30% of all transcript products in lawful interception.

The third challenge for LEA's is the possibility to use performing and efficient Voice Recognition (‘VR’) biometric technologies while preserving the public's privacy and conducting ethically in a way that respects societal norms. For example, innocent callers who use suspect's phone routinely and therefore should not be eavesdropped upon (unless they are forced by the suspect to communicate with another suspect/criminal). Or another example, suspect family's members who use the suspect phone at their home routinely for personal business, for their personal matter, although the phones under a court warrant permitting lawful interception.

These "innocent" calls must be filtered out from the Lawful Interception process. (Nevertheless, where innocent people are forced by the suspect to communicate with other suspects or criminals, these calls should be identified and intercepted).

Few more challenges that LEA face are in the context of speaker identification reliability:

- Judicial admissibility of speaker identification results depends on national legislation which is strongly influenced by the reliability of the automated speech analysis.
- A challenge to have speaker identification results presented in a standardized format before the court to enhance such reliability. It would indeed avoid subjective interpretation in the final written account.
- Voice spoofing (or voice cloning) methods used by criminal to mislead LEAs (as if they were another person who made the call) and to deal with the limited size of speaker models databases in use by state of the art speaker identification systems.

SIIP, FP7 funded European Project[1] aims to overcome the above challenges in order to enable LEAs to have better intelligence and incrimination capabilities while responding to the privacy preserving, legal and ethics considerations.

In the following, we present the analytics developed in the project, the data management mechanisms, the generic approach of the integration of the final system and finally the evaluation methodology implemented in the project.

II. Speaker Identification Analytics

A. Speaker Identification

Speaker Identification (SID) system is built around the i-vector (identity vector) approach [7], modeling a speech recording by projecting its acoustic features onto a low-dimensional representation. As such, i-vectors contain many of the variabilities observed in the original recording, e.g. speaker, channel and language, with these components lying on the i-vector low-dimensional space as well. Since i-vectors originate from a multivariate Gaussian distribution and have fixed dimensionality, compared to a variable and potentially large number of acoustic observations in the original utterance, i-vectors can be conveniently processed using statistical and machine learning techniques. In SID engine, the inter-speaker variability of i-vectors is retained and other variabilities are removed using techniques such as Linear Discriminant Analysis (LDA), within class covariance normalization (as in [7]) and Probabilistic LDA that provide better discriminability amongst speakers [8]. After applying such techniques, i-vectors are assumed to represent the speaker information in the original recording [13]

In SIIP, the performance of SID engine was further enhanced by estimating posteriors from Deep Neural Network (DNN) instead of Gaussian Mixture Model (GMM). While both DNNs and GMMs aim at incorporating phonetic information of the phrase with these posteriors, model-based SID approaches ignore the sequence information of the phonetic units of the phrase. SIIP overcomes this problem by applying a dynamic time warping architecture using speaker-informative features [9]. Further, also a combination of SID with other modalities such as with automatic speech recognition or keyword spotting engine allows the use of content information in speaker identification.

SIIP speaker identification systems have been consistently shown, through peer-reviewed publications and international challenges, to be among the best systems in the world. In the latest NIST 2016 Speaker Recognition Evaluation, SIIP systems featured among the top solutions, especially in terms of the Equal Error Rates (EERs). Overall, SIIP systems achieved EERs as low as 0.5% on previous benchmark NIST datasets in which focus more on evaluating systems in low false alarm-regions.

B. Gender and Age Identification

Both, the Gender- (GID) as well as the Age-Identification (AID) modules within SIIP are based on a GMM/UBM framework. GID aims to determine the gender of a given speaker; AID aims at identifying whether the speaker is an adult or a child (translating to a binary classification problem).

Models for both classifiers were trained in SIIP project using a combination of different corpora in English and German (WSJ [2], aGender [3], PF Star [4], CMU Kids [5], Vorleser. The total amount of acoustic data amounts to 138.25h of audio. This set was used for cross-evaluation experiments.

Monolingual experiments as well as cross-lingual experiments were carried out. A series of models of different complexities was trained and evaluated in a cross-evaluation manner to arrive at the best performing set of models which were eventually deployed in the SIIP demonstrator. The best performance for SID and AID were 98% accuracy and 89.8% accuracy respectively.

C. Language and Accent Identification

The language and accent identification engines are based on the I-vector PLDA architecture, similar to SID engines, described in Section II.A. I-vectors are extracted using a GMM/UBM and a DNN respectively. These i-vectors are length-normalized as in SID systems.

For the Accent ID (AID) task, only a PLDA module is trained to discriminate accents rather than speakers. In our SIIP implementation, we considered only English speech with several native and non-native accents to be used for training (English (Native), Chinese, Russian, Hindi and Korean). The developed AID systems tested on NIST datasets provide ~80% detection accuracy.

For Language ID (LID), it is common to distinguish between acoustic and phonotactic engines. Acoustic LID modeling attempts to find the discriminative information in acoustic data (similar to SID or AID). Successful examples of acoustic

1 http://www.siip.eu

engines exploit GMMs, SVMs, and the more recent I-vector and DNN approaches. Phonotactic LID exploits the co-occurrences of phone sequences in speech. Text-dependent phone recognizers are usually employed to tokenize speech into phonemes even if the target language is unknown. Recently, phone log likelihood ratio based features as extracted from phonetic recognizers have received particular attention in the LID field. Experimental results have shown that acoustic and phonotactic engines are orthogonal and meaningful improvements are obtained using combined engines. The SIIP language identification allows discrimination among 22 languages. Achieved results show equal-error-rates of about 3% and 0.8% for fused LID systems (combining both acoustic and phonetic approaches) when tested on 10s and 30s long utterances.

D. Keyword and Taxonomy Spotting

Keyword Spotting (KWS) within SIIP is performed by first producing a full transcript of the input audio and subsequently detecting keywords in the output structures.
The KWS components provided by SAIL LABS and Idiap in SIIP project are based on the state-of-the-art open source Kaldi toolkit [6] and follows a three-step process: The first step constitutes the pre-processing and segmentation. Here the input audio is normalized and converted into acoustic features. Based on this information, it is segmented into utterances and passed to the second step, the actual Automatic Speech Recognition (ASR)-module. In this module, the segments are converted into a network of words with associated time-tags and scores. In the third step, this network is searched for the keywords. The KWS service finally returns the corresponding file as a match if at least one of the keywords appears in the transcription. The actual scores are determined via a combination of the scores and timings of the individual keywords.
Taxonomy spotting, developed over output of ASR and KWS engines, then attempts to semantically structure the concepts and relations between different lexical outputs provided by ASR and KWS. Taxonomy spotting allows to extract the meaning of text provided by automatic transcription.

E. Results' fusion approaches

Fusion is a common approach to improving the performance of SID systems. Most of recent contributions however focused on intra-task fusion, combining different SID engines (e.g. trained on different data, applying different modeling technologies, etc.). SIIP project rather explores inter-task fusion approaches, to incorporate side information from other engines (such as accent, age, gender or language identification engines) to eventually improve SID, since these characteristics are related to speaker identity as well.
In our recent work [12], we explored two approaches, namely based on score-level and model-level techniques, to combine speaker information together with accent and language information. Experimental results on NIST speaker evaluation 2008 dataset reveal that both techniques are able to bring improvements over the baseline (i.e. no fusion, or filtering out inadequate SID scores according to side information). SIIP project further explores other ways to incorporate not only accent or language characteristics, but also other based on gender or age, to eventually improve SID.

III. MULTISOURCES DATA MANGEMENT

A. Data gathering

The SIIP system includes two distinct data gathering capabilities (voice call data and open source intelligence), for the purpose of providing a collection system for audio samples, with associated metadata, from (simulated) interception systems (voice calls) and multiple open sources.
For reasons of data protection, a synthetic interception content generation engine was developed that allows realistic voice call content construction, querying and capture. In conjunction with audio capture, associated (simulated) Call Detail Records (CDR) and Internet Protocol Detail Records (IPDR) are also available via the simulator.
The SIIP developed Lawful Interception (LI) Simulator was designed to emulate typical interception systems commonly utilized by LEAs and works as an autonomous interception voice call system, offering interception on demand and can generate interception of voice calls spontaneously.

The LI Simulator supports an array of communication channels including; SATCOM, PSTN, cellular, telecom VOIP and Internet VOIP apps.
The LI Simulator includes resampling, equalization, several compression formats and noise addition features allowing for thousands of unique voice samples to be generated. This provides a very rich repository of data to query and interrogate. SIIP project also developed a large open-source acoustic simulator to be used for the development phases, allowing for compensating for the effect of a wide variety of speech degradation processes in SID tasks [14].
In addition to the LI Simulator for voice call data, the SIIP system includes an Open Source Intelligence (OSINT) data gathering capability allowing for broad and targeted searches to be conducted against an array of specified online sources.
Through the utilization of SIIP's OSINT capabilities, the Social Media platforms of Twitter, Google+, LinkedIn, YouTube and Facebook are brought into the fold of available sources from which intelligence can be gathered.

SIIP's OSINT features expand on basic keyword search and retrieval functionality to allow Investigators query OSINT sources through an array of advanced options and search criteria including language relevance, regions, geo-location, entity associations etc.
The SIIP system allows for the filtering and funneling of OSINT results, to efficiently and accurately arrive at the targeted information required.
The SIIP system also provides OSINT capture capabilities beyond standard basic information commonly derived from such sources including metadata associated with a search result and all linked multimedia files, all of which may be captured and stored within the SIIP system.

Captured multimedia in the form of graphics, such as photographs or images, are available for inclusion and association with the entity under investigation, via the SIIP portal.
Captured video content is processed through SIIP's Video Processing Engine, extracting the audio content, splitting it to individual mono files (if not originally mono) and formatting to uncompressed, PCM, 16KHz, 16 bits, mono wave files.
Captured audio content (as distinct from captured video content) is also processed, as above, and in similar fashion to audio extracted from video content, is made available to the Speaker Identification Analytics engine through SIIP's Information Sharing Mechanisms. Original (non-processed) files are maintained for possible evidential purposes also.

B. Information sharing mechanisms

LEAs equipped with an operational SIIP system will be able to share between and compare speaker models of identified suspects. Rich metadata associated to the suspect are recorded in a separated file/database but capable of creating automated links with the voice sample/print database (e.g. Personal details, Social connections and Fake Identities), in a secure way, in order to preserve the right to privacy.
For this purpose, SIIP implements a SIIP-Info-Sharing-Center (SISC), located at Interpol and includes an Info-sharing Management-Module and very large (>1,000,000 records), secured, centralized database infrastructure of hi-quality suspect speaker models and metadata. Prior to populating the database, LEAs should provide guarantees attesting of the high quality of the data as well as their authenticity. Each LEA that is inserting new input data to this centralized database is labelled a 'Donor'.
On the other hand, all the LEAs will be able to pull Voiceprints and Metadata (in a separate file/database) about a given suspect by providing one of his known identities. The LEA that is retrieving data from this database is labelled a 'Recipient'. Each LEA can play both as Donor and Recipient.

A baseline-programming interface that enables the implementation of the secure Info Sharing Center Mediation Module is already integrated in the actual SIIP system.

IV. SYSTEM INTEGRATION AND IMPLEMENTATION

A. Generic and flexible integration

A SIIP incoming voice content is an unstructured data sometimes combined with descriptive metadata (e.g., suspect name, nationality, age, gender and many others). A SIIP module aims at analyzing the content in order to enrich the existing metadata by adding new specific properties, as mentioned above, by using Gender, Age, Accent, Language Identification engines, automatic speech recognition and keyword spotting engines.
We can consider that all these various modules have a common purpose: they analyze the unstructured audio content to extract one or several specific features that they formalize with descriptive metadata. Some of them need to reuse the results from others. For example, Speech Recognition (transcription) will be easier if the language is already identified. This means that the modules have to be performed according to a relevant sequence. In other words, a processing chain must be defined to decide what available software services must be requested and when. The processing chain defines the order and the conditions in which the service is called. According to this definition, each module will fulfill their mission (i.e., deliver the expected service) one after another. Each will receive the description of the audio content to be processed with some input metadata and it will enrich this description with new metadata by using its own outcomes.
In addition, one of the main objectives of SIIP is to provide a generic architecture that should be modular, introducing an easy and straight forward way to integrate new identification engines as well as supporting new languages and dialects.

For all these reasons, we decided to use and adapt the WebLab platform [10] as an integration facility to manage orchestration and information exchange between the SIIP modules.

The WebLab platform relies on a Service Oriented Architecture as the core paradigm for the design and integration of components. The high level functions offered to users through applications, is achieved by putting together services and calling them in the right sequence (orchestration). As a consequence, the service definition and conception is a key feature in the platform. WebLab Core is an open source technical baseline acting as a runtime environment for unstructured information processing services.
Every component to be integrated in the platform shall implement one (or several) service generic interface(s) described as service level agreements in WSDL. They offer the platform their processing capabilities that could be called by the orchestrator in order to run the business processes, or workflows. These business processes delivers the high level function offered to users.
The components are fully autonomous and do not have any knowledge of the other services deployed and consumed by the platform. However, as services need to collaborate through the WebLab workflow, a common data exchange model is used among the services. These services could then be easily chained: a "producer" service (providing a processing capability) encodes its results following the model and provides them with a "consumer" service (requesting a process) which decodes the received results and then process them.

The diagram below shows a high-level interaction diagram through the integration platform.

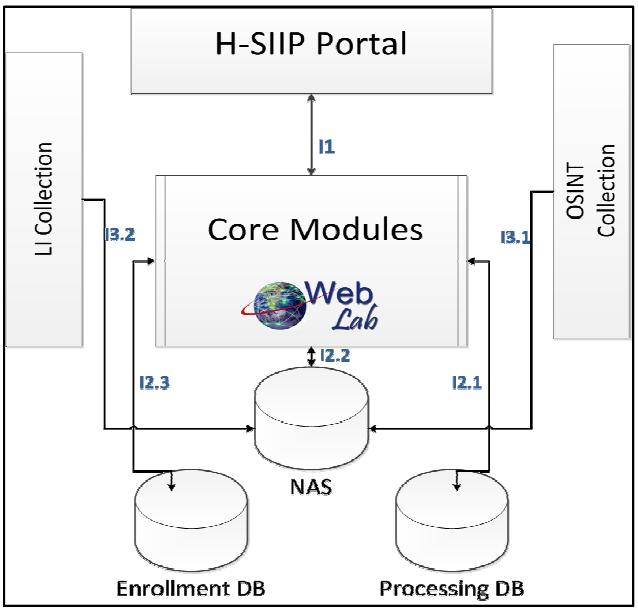

Fig. 1. SIIP sub-architecture diagram with layers interactions

- H-SIIP portal: provides access to the developed functionalities via graphical interfaces (web pages)
- LI collection: modules simulating lawful interception.
- OSINT collection: modules collecting information for open sources, especially, from social networks
- Core modules: includes the different audio processing modules, and, technical components needed for pre-processing and for the orchestration. All these modules are integrated as WebLab services.
- NAS: is shared storage, accessible under SMB (Server Message Block- a network file sharing protocol) and NFS (Network File System), which stores the audio files coming from LI and OSINT Collection and stores the audio files that are the results of segmentation process.
- Processing DB: stores information about these processes, initiated by the H-SIIP Portal through the WebLab API, and their results. Results are then returned on demand.
- Enrollment DB: used for permanent storing of Speaker information, including Speaker Models and metadata associated with the speakers.

The communication and the interaction between these components are orchestrated by the WebLab platform and via the following interfaces:
I1: communication between the portal and WebLab in order to handle the different flows and answer users' queries. This interface is ensured through a REST API.

- I2.1: storage of the results of the different processing flows (WebLab processing chains). These results are communicated to the portal through I1.
- I2.2: use of shared files candidate for processing. These files can come from OSINT through I3.1, from LI through I3.2 or directly from users (file system).
- I2.3: interaction between core modules (integrated as WebLab services) and speaker information stored in the enrollment DB. This interface is ensured through a REST API.
- I3.1: storage of audio files coming from OSINT. These files are used by core modules through I2.2.
- I3.2: storage of audio files coming from LI. These files are used by core modules through I2.2.

B. Portal design and implementation

The SIIP portal constitutes the main interaction point between end-users and the SIIP components. The primary objective of the SIIP portal is to accommodate end-users' functional requirements and to provide an intuitive interface that could enable them to easily grasp the benefits of the provided middleware and tools. Building on contemporary design and development practices SIIP portal is Web 2.0 based application supporting the seamless and efficient interaction with end-users.
Taking into account the variety of information that is made available by the project, the provision of an intuitive interface is of paramount importance. Moreover, the complexity of the provided functionality as well as the diversity between the expected end-user roles renders the design and implementation of the portal a considerable challenge.

The design and prototype implementation of the SIIP portal, apart from the specified functional user requirements, has been guided by a set of non-functional constraints and generic design principles such as the ones mentioned below:

- User-centered design: The structure of the functionality offered by the portal, the page design and whole layout have been devised in such a way so as to support the interaction with the end-user. User requirements have been considered since the onset of the portal design phase, whereas a continuous prototype- user evaluation-update process is applied for the portal development.
- Asynchronous interaction: The variety, complexity and computational cost of the provided functionality render the synchronous integration of the portal with the back-end services a rather inefficient and obtrusive approach. To accommodate the unobstructed interaction of the end-user with the system as well as to facilitate the independence of the portal with regards to the rest of the provided middleware and tools, the use of the asynchronous interaction pattern is imperative. In addition to fostering the end-user experience, the asynchronous interaction pattern enables the modular and independent development and update of the provided functionality.
- Modularity: To facilitate the development of a complex system as this portal, and to enable traceability

between requirements and implementation components the use of a modular design is of high importance. The partitioning of the provided implementation into distinct and concise logical fragments enables us to speed up the development of the portal in a multi-developer environment and to better trace between requirements or problems and implementation code.

- Security: Considering the sensitive nature of the exchanged information and of the performed actions security is a paramount requirement for the whole system. Authorization, authentication, non-repudiation, integrity and privacy are key features of the system that will have to be ensured across the whole range of tools and middleware that will be offered by the system. These aspects are also considered during the design and prototyping of the portal.
- Availability: Ensuring a high availability rate, i.e. 24/7, is considered as critical factor for achieving user satisfaction and acceptance in most of the contemporary portals. In this frame the SIIP portal will be designed and implemented in manner that will ensure high availability, but the achievement of 24/7 availability is not a critical factor for the SIIP portal.
- Resilience to failures: The complexity of the whole system renders it prone to failures, which may be raised from several sources, e.g. the underlying middleware, services offered by the project, etc. Along with the use of the asynchronous interaction pattern and the modular design, the portal is supported by proper exception handling mechanisms and tools that will enable it to report and accommodate exceptions and failures that may be raised during its operation.

Below an example of graphical interfaces – Alerting page and speaker diarization page - provided by the portal.

The alerting pages provide a listing of the alerts that are connected with the cases a user is related to. The alerting page provides an overview of the alerts listed in descending arrival time order. Hence, the most recent alerts will be presented at the top of the list. The user will be able to filter the presented list using additional criteria.

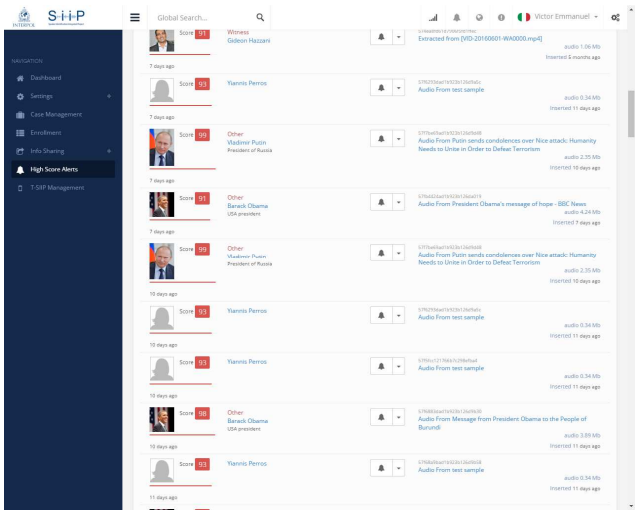

Fig. 2. SIIP Portal : Alert listing

The speaker diarization page allows to:

- Manually choose which specific segments will be sent for speaker identification
- Request for the audio file to be automatically segmented through the auto diarization function

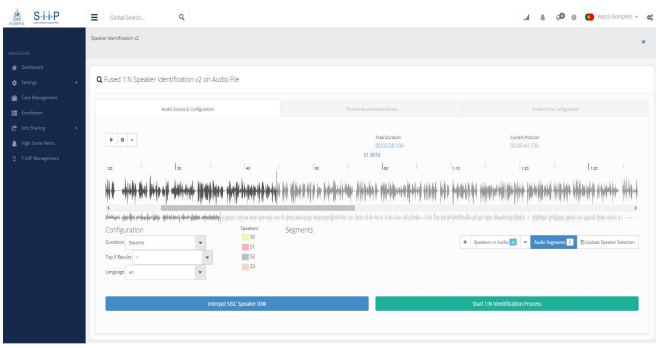

Fig. 3. SIIP Portal : Speaker diarization

To enable a transparent use of the SIIP system capabilities, we defined and exposed a REST API allowing the communication with the core system. Advantage of this architecture is that the portal does not care about integration details such as how many voice identification components are presents, where they got deployed, which data storage solution the system relies on.
A core process triggered by a portal request is illustrated by the next Figure showing the sequence of interactions involved in process.

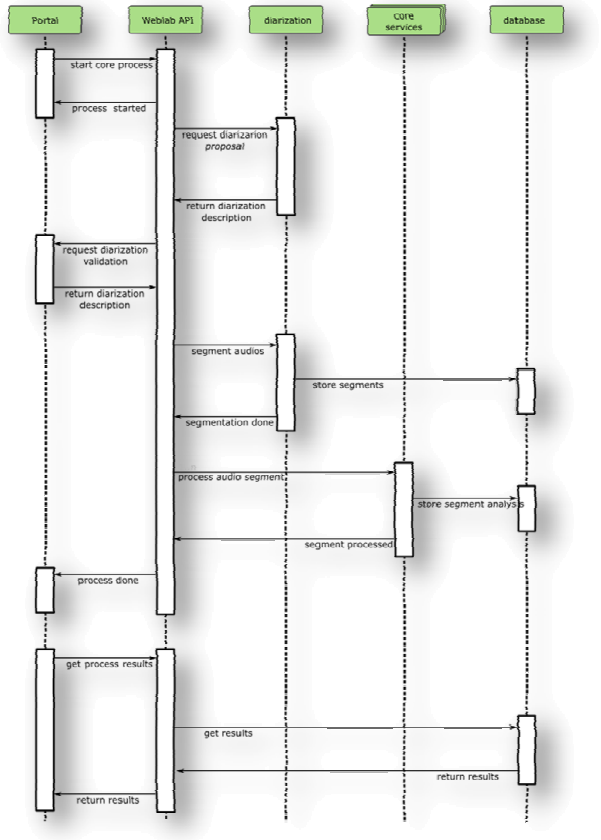

Fig. 4. Core process sequences diagram

Finally, the SIIP portal implements a communication hub supporting the interaction between the collaborating LEAs (Info Sharing mechanisms). It also provides functionalities that enable the INTERPOL operators to monitor and manage the information exchange process.

V. FIELD TESTING AND EVALUATION

A. SIIP survey questionnaires:

In the framework of SIIP project, a questionnaire was drafted by INTERPOL on end-user requirements (legal, technical and operational aspects), based on feedback provided by experts and SIIP consortium partners. The questionnaire was then circulated among the 190 INTERPOL's member countries (Translation provided in INTERPOL's four official working languages, namely Arabic, English, French and Spanish). A strong interest for the SIIP project was shown as INTERPOL received 91 responses from LEAs' cybercrime, counter-terrorism units and forensic laboratories. Subsequent telephone interviews were held with 40 survey respondents. A paper was also submitted and published in the special issue of "Forensic Science International"[9].

B. Pool of experts:

An expert Working Group, composed of law enforcement officers as well as forensic, technical and legal experts was set up to comment upon the results of the questionnaire. INTERPOL identified law enforcement, legal and technical experts from around the world in speaker identification field in order to create a pool of Experts to provide feedback and share expert information. Several field visits were also conducted with law enforcement agencies worldwide to gather end-user requirements.

C. Expert group meeting

An expert Working Group, composed of law enforcement officers as well as forensic, technical and legal experts was set up to comment upon the results of the questionnaire. A workshop with 41 participants (LEA Investigators, police officers, forensic experts, prosecutors and representatives of the academia and the private sector) was held at INTERPOL. This event was dedicated to presentations of the experts in their respective fields, followed up by the analysis of the needs of LEA in the field of Speaker Identification and sharing of expertise and good practices on the subject matter.

D. End-user meeting

An end-user meeting organized by INTERPOL held in London with police officers, forensic experts and consortium partners to collect the end-user requirements.

E. Proof of concept

The concept of the SIIP system and its contribution to speaker identification in the context of police investigations was demonstrated during the Proof of Concept event held at the Carabinieri School in Rome in June 2016. Attended by police officers from more than 20 law enforcement agencies, forensic experts and representatives from academia and the private sector, the systems capabilities were shown in a variety of scenarios.

F. Field test:

More than 130 speaker identification researchers and experts, forensic experts and police investigators from some 40 law enforcement agencies from around the world took part in the field test in March 2017 in Lisbon, Portugal. This event was held to promote an open discussion among the key stakeholders on the challenges and relevant issues to be considered for the development of a privacy-enhanced speaker identification system with a global reach.

G. Qualitative evlauation methodology

While the end-user centered assessment and evaluation provides the primary perspective on SIIP performance, addition controlled testing will be applied to complement operational findings. This methodology, in which controlled scenario testing is used to clarify or support findings from operational tests and trials, has been proven effective in numerous biometric evaluations. In the SIIP application scenario, controlled testing will be used to explore interesting or potentially anomalous findings (e.g. devices that generate unusually high failure rates or subjects who cannot reliably match against their enrolled data). From a validation perspective, many aspects of end-to-end functionality can be assessed in a controlled lab environment.

Evaluation corpuses have been set up in order to evaluate both each component and the complete chain implemented in SIIP. They are based on the corpus provided during international evaluation campaigns (more than 100 000 annotated audio files) and on data provided by the police services involved in the project (data that allowed the resolution of real cases).
The results of this evaluation campaign will be the subject of a specific publication.

VI. Conclusion and further work

The identification approach and the implemented system proposed in this paper have been presented to the international community of end-users animated by Interpol. End users were satisfied and have expressed different exploitation needs that we will try to take into account in a further work.

The qualitative evaluation of the components and the end-to-end chain has started and the first results are very encouraging.

Finally, we are working to go deeper in the standardisation and communication between the agencies using this common infrastructure. However the approach will be adaptable to provide accurate speaker identification outside the EU.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7 under REA grant agreement n° 607784.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] Douglas B. Paul , Janet M. Baker, The design for the wall street journal-based CSR corpus, Proceedings of the workshop on Speech and Natural Language, February 23-26, 1992, Harriman, New York

[3] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann: A Database of Age and Gender Annotated Telephone Speech, Proceedings of the Language and Resources Conference (LREC), 2010.

[4] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, M. Wong. *The PF STAR Children’s Speech Corpus*. In Proc. of Interspeech, 2005.

[5] M. Eskenazi, J. Mostow, and D. Graff. The CMU Kids Corpus LDC97S63. Web Download. Philadelphia: Linguistic Data Consortium, 1997.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, P. Motlicek, N. Goel, M. Hannemann et al. "The Kaldi speech recognition toolkit." In IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Tran. on Audio, Speech and Language Processing, pp. 788–798, 2011.

[8] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in IEEE 11th International Conference on Computer Vision (ICCV). IEEE, 2007, pp. 1–8.

[9] G.Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, C. G. Dorny. “INTERPOL survey of the use of speaker identification by law enforcement agencies”. Forensic Science International, Volume 263, June 2016

[10] S. Brunessaux and P. Giroux, “WebLab: 10 years, the age of maturity for the WebLab platform”, 1st edition of the Practical Applications of Artificial Intelligence (APIA), Rennes (France), July 2015.

[11] S. Dey, P. Motlicek, S. Madikeri, M. Ferras, “Exploiting sequence information for text-dependent speaker verification, in proceeedings of Icassp, New Orleans, USA, March 2017.

[12] M. Ferras, S. Madikeri, S. Dey, P. Motlicek, H. Bourlard, “Inter-task Fusion for Speaker Recognition, in proceedings of Interspeech 2016,

[13] P. Motlicek, S. Dey, S. Madikeri, L. Burget, “Employment of Subspace Gaussian Mixture Models in Speaker Recognition, in proceedings of Icassp, 2015.

[14] M. Ferras, S. Madikeri, P. Motlicek, S. Dey, H. Bourlard, A Large-Scale Open-Source Acoustic Simulator for Speaker Recognition, IEEE Signal Processing Letters, 2016.

.