Open-Set Speaker Identification pipeline in live criminal investigations

Maël Fabien[1,2], *Petr Motlicek*[1]

[1]Idiap Research Institute, Martigny, Switzerland,
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

mael.fabien@idiap.ch, petr.motlicek@idiap.ch

Abstract

Speaker recognition has many applications in conversational data, including in forensic science where Law Enforcement Agencies (LEAs) aim to assess the identity of a speaker on a specific recorded telephone call. However, speaker identification (SID) systems require initial enrollment data, whereas LEAs might start a case with text or video evidence, and few to no enrollment data. In this paper, we introduce the ROXANNE simulated dataset, a multilingual corpus of acted telephone calls following a screenplay prepared by LEAs. We also present a process to build criminal networks from SID, by addressing practical constraints of these investigations. Our process reaches a speaker accuracy of 92.4% on the simulated data and a conversation accuracy of 84.9%. We finally offer some future directions for this work.

Index Terms: speaker identification, conversational data, organized crime

1. Introduction

This paper address the scenario of criminal investigations, during which LEAs intercept evidence on suspects and start wiretapping a list of telephone numbers belonging to the suspects, in order to understand connections between each suspect and external speakers and establish a criminal network which is then used for link prediction [1], node classification, community detection, central characters identification [2], or network disruption [3].

In large investigations, SID is typically used to automate the attribution of a call to existing speakers, instead of having to listen to hours of intercepted speech every day. However, in traditional closed-set SID, several seconds or minutes of speech of a speaker are selected as enrollment, and once all speakers have been enrolled, the system is deployed. Such enrollment data often do not exist for live ongoing cases. We propose a process that iteratively enrolls speakers based on decision thresholds and addresses practical issues of SID in criminal conversational data in an open-set fashion. We also introduce ROXANNE simulated data, a dataset of simulated telephone calls matching real-life conditions of criminal investigations.

The paper is organized as follows: Section 2 presents the related works. Section 3 introduces the ROXANNE simulated data. Section 4 describes our iterative SID process, and Section 5 presents experimental results. Finally, in Section 6, we discuss some of the results and future directions to take.

2. Related work

Speaker recognition is a common task in forensic science [4], and is composed of several sub-tasks, including Speaker Verification (SV), whose aim is to verify the identity of a speaker in a recording, and SID. Most LEAs focus on SV to speed up the decision of voice comparison and bring evidence in court [5, 6, 7].

SID has also been studied in the context of criminal investigations for decades now [7, 8], and especially when a significant amount of enrollment data per speaker is available. Due to the importance of decisions being made when identifying speakers in criminal investigations, semi-automatic SID techniques are still commonly used. These methods allow human intervention in the processing of audio samples or in the final decision [9].

3. The dataset

The lack of openly accessible criminal conversational data, that follow a realistic screenplay, or are published from a real case, limits the number of works in the field of SID for criminal conversations. As a result, few works and commercial solutions also tackle the construction of criminal networks from SID, especially as a tool for investigators.

We created a set of simulated data, jointly with LEAs, which matches most of the constraints of real-life investigations. Over 100 telephone calls, lasting a few seconds to several minutes, from 24 speakers, representing 155 minutes of speech, sampled at 8kHz, have been recorded on Twilio [1]. Two speakers talk in each telephone call, each of them on a separate channel. The speakers read a pre-defined scenario, either in their own language or a foreign language. The dataset is multilingual (Czech, Slovak, Russian, Vietnamese, German, and English, used as a second language by 6 speakers), which is a well-known factor for performance degradation, as discussed in [10, 11]. Transcripts in the original language are available, and the screenplay was prepared by LEAs of the consortium to match the conditions of real investigations. All recordings are encoded as PCM-16.

The screenplay involves the Prague anti-drug unit of the Czech police which investigates three cases at the same time: a first drug distribution case involving Czech and Russians students, named DDA (Drug Distribution A), a drug lab, ran by Vietnamese suspects, named DLA (Drug Lab A), and another drug distribution case which involves German speakers, named DDB. All cases are linked through a single character, who is managing the drug deal network.

4. The pipeline

Criminal investigations raise several challenges compared to classical SID systems in which we enroll all speakers on given enrollment files and run the identification on the rest of the test files. In live conversational criminal data, investigators start with some evidence (text or video) regarding suspects, and few to no speech enrollment data. The list of suspects is also expanding dynamically, and the amount of enrollment data col-

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

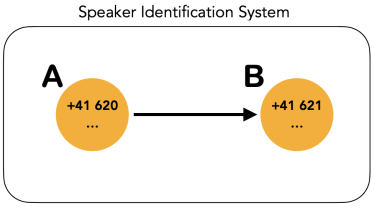

Figure 1: *Starting point of the speaker identification pipeline*

lected for each speaker might remain very limited. We propose a iterative method that is suitable for criminal conversations starting with a single phone call. In this section, we describe our iterative SID approach and the SID system used.

4.1. Open-set SID approach

Most cases start with a few suspects being wire-tapped. These initial suspects start to produce a first telephone call, between two speakers, knowing that at least one of them is a wire-tapped speaker. This first telephone call is used to enroll the two speakers as illustrated in Figure 1. The names of the speakers, A and B, are arbitrary here, but can also be adapted manually to match the knowledge of investigators, or automatically through a coreference resolution module on top of automatic speech recognition transcripts.

As the case moves on, additional telephone calls C_k are intercepted. Therefore, as presented Figure 2, if the telephone number of speaker B makes a call, we need to assess the identity over both channels. For the recording r_q belonging to telephone number of B, we have to run our SID system since several speakers might use the same telephone. In this example, our SID system correctly identifies B as being the speaker. For the speaker on the second channel, we compute s_{r_q}, the score of our SID system against every enrolled speaker. If the log-likelihood ratio is below a given threshold ($\tau_{new-speaker}$), usually set to 0 in our experiments, we consider that this sample does not correspond to any known speaker in our SID system. We, therefore, create a new speaker and enroll it based on the speech sample. Otherwise, if this speaker is recognized as being an existing speaker, we can assess that B has talked to an existing enrolled speaker, which adds an edge to the criminal network under investigation.

A constraint of SID on conversational data is that a telephone call cannot involve the same speaker on both ends. Therefore, a constraint must be set on the decision here. We select the identified speaker with the highest score, and if the second speaker has a similar identity, we select the second best speaker $s_{r_{q2}}$, as long as its score is above the threshold $\tau_{new-speaker}$ that we have set to 0.

So far, our SID system relies only on a single enrollment file. If the output score of a speaker is above a second threshold ($\tau_{new-enrollment}$), fixed to 10 in our experiments, we add the speech sample collected during the conversation to the existing enrollment files for this speaker r_q, hence generating longer enrollment files for speakers if the confidence level is met. Towards the end of the case, this typically creates enrollment data of more than 5 to 10 minutes per speaker in the ROXANNE simulated data. The iterative SID process we developed is more formally defined in Algorithm 1.

4.2. SID system

The ROXANNE simulated data is not large enough to train an entire SID system. Therefore, we used Idiap's submission to the NIST SRE 2019 Speaker Recognition Evaluation [12], a pre-trained SID system prepared for NIST Speaker Recognition Evaluation (SRE19) dataset. The system relies on a Time-Delay Neural Networks (TDNN) [13] X-vector architecture [14, 15] with a Probabilistic Linear Discriminant Analysis (PLDA) [16] back-end.

We first down-sampled speech to 8 kHz, and apply Bandpass filtering between 20 and 3700 Hz. We then extract windows of 25 ms, with a frame-shift of 10 ms. From these windows, we extract Mel Frequency Cepstral Coefficients (MFCCs) of 23 dimensions. Since conversations contain a lot of non-speech frames, we apply an energy-based Voice Activity Detection (VAD) to remove these frames.

We trained the X-vector system on Voxceleb dataset [17] and on the augmented versions of Switchboard dataset [18] and SRE 2004 to 2010 with additive noise (MUSAN dataset [19]) and reverberation (RIR dataset [20]). Finally, the PLDA classifiers were trained on augmented versions of SRE.

5. Experimental results

Speaker accuracy is the most common metric for SID systems. We jointly established with LEAs involved in the ROXANNE project that both speaker accuracy and conversation accuracy should be presented as output performances. The notion of conversation accuracy was introduced in one of our previous works [21], and represents the percentage of conversations where both (or all) speakers have been correctly identified, which guarantees that no wrong edge was added to the network and that investigators follow the correct track. On the ROXANNE simulated data, using threshold values $\tau_{new-speaker} = 0$ and $\tau_{new-enrollment} = 10$, the speaker accuracy we obtained in 92.4%. The corresponding conversation accuracy is 84.9%.

Errors mostly arise in short utterances, since no specific minimum duration for the speech was set to process a recording. Additional filters can be defined in order to avoid processing files which lead to poor performances, and therefore leave the decision on these files to human experts.

However, not all mistakes in a network have the same importance. Figure 3 depicts an example network created by our algorithm, with both wrong and correct edges. Edges in red display a wrong link added between two characters, and edges in green are for the correct ones. We notice that a wrong edge was added between speakers G01 and V01, which would suggest that these two characters know each other, whereas they belong to different groups in the screenplay. This might bias the network analysis of investigators, and especially the community detection task.

6. Discussions

No tool currently allows investigators to process a large number of conversations in an automated manner, with few to no human intervention required. This is partly due to the lack of relevant conversational criminal data available for researchers, and the lack of realistic SID solutions starting with few to no initial speech data on the suspects.

The ROXANNE simulated dataset matches a screenplay defined by LEAs. The dataset will soon be available for all researchers. This first step could open the road to additional con-

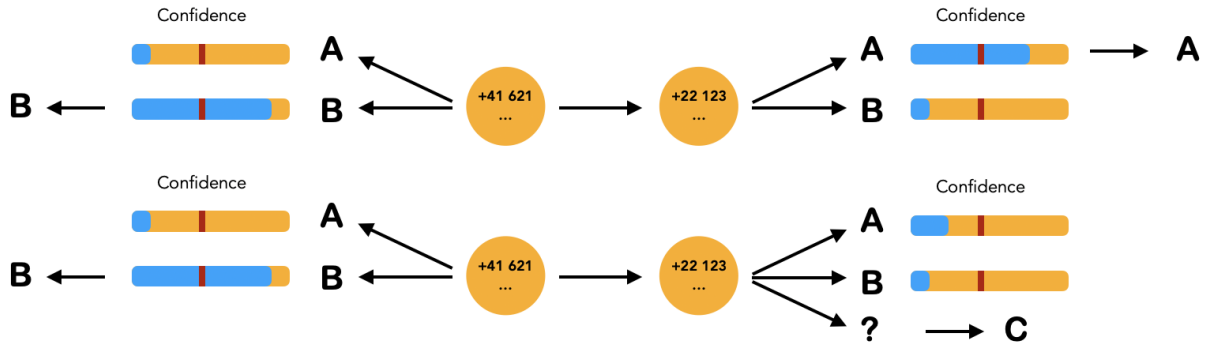

Figure 2: *New telephone call decision process*

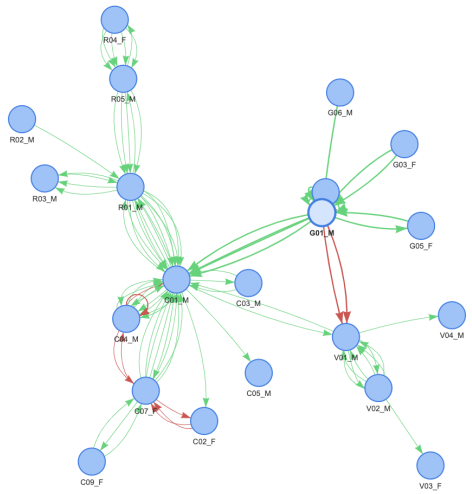

Figure 3: *Correct and wrong edges on the output network*

tributions in the area, and become a standard evaluation dataset for various tasks in the context of criminal investigations.

A fully automated approach to build criminal networks from iterative SID seems so far not realistic since missing 15.1% of the network links can drastically change the understanding of a case by LEAs. During our experiments, it appeared that a semi-automatic approach might be the correct approach, in order to build trust in the tool, and avoid important errors in the network. For example, if a recording does not meet pre-defined criteria (length, SID score...), human intervention might be needed. We will also further experiment whether human intervention in the first utterances can help the performance of SID system.

7. Conclusions

We introduced in detail the ROXANNE simulated data as a candidate for criminal conversations. We also introduced a iterative process to build a SID system in live criminal conversations, starting with no speech evidence. We obtained 92.4% speaker accuracy and 84.9% conversation accuracy. Although encouraging, these results also suggest that human intervention might be needed in such a tool given the importance of decisions being

Algorithm 1: Speaker identification process in criminal conversations

```
k = 0;
S = empty network;
τ_new−speaker = 0;
τ_new−enrollment = 10;
while New telephone calls C_k intercepted do
    Initialize empty list of speakers L;
    if k = 0 (first conversation) then
        for each recording r_q in C_k do
            Enroll speaker s_rq based on r_q;
            Store s_rq in the list of speakers L;
        end
    else
        for each recording r_q in C_k do
            Run speaker identification on r_q;
            Select speaker s_rq with highest score
              score_srq;
            if s_rq already identified in C_k then
                Select second best speaker s_rq2;
                if score_srq2 > τ_new−speaker then
                    s_rq = s_rq2;
                    score_srq = score_srq2
                end
            end
            if score_srq < τ_new−speaker then
                Enroll new speaker
            else if score_srq > τ_new−enrollment then
                Add speech to enrollment of s_rq
            Store s_rq in the list of speakers L;
        end
    end
    Create an edge between the two nodes stored in L
      and store in S;
    k += 1;
end
Result: Display output network
```

made.

8. Acknowledgment

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

9. References

[1] Francesco Calderoni, Salvatore Catanese, Pasquale De Meo, Annamaria Ficara, and Giacomo Fiumara, "Robust link prediction in criminal networks: A case study of the Sicilian Mafia," *Expert Systems with Applications*, vol. 161, pp. 113666, Dec. 2020.

[2] Sylvert Prian Tahalea and Azhari Sn, "Central Actor Identification of Crime Group using Semantic Social Network Analysis," *Indonesian Journal of Information Systems*, vol. 2, no. 1, pp. 24, Aug. 2019.

[3] Lucia Cavallaro, Annamaria Ficara, Pasquale De Meo, Giacomo Fiumara, Salvatore Catanese, Ovidiu Bagdasar, and Antonio Liotta, "Disrupting Resilient Criminal Networks through Data Analysis: The case of Sicilian Mafia," *arXiv:2003.05303 [cs, stat]*, Mar. 2020, arXiv: 2003.05303.

[4] Phil Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech & Language*, vol. 20, no. 2, pp. 159–191, Apr. 2006.

[5] Lawrence Solan and Peter Tiersma, "Hearing Voices: Speaker Identification in Court," *HASTINGS LAW JOURNAL*, vol. 54, pp. 65.

[6] J Sarwono and M I Mandasari, "Forensic speaker identification: an experience in Indonesians court," p. 3.

[7] Anders Eriksson, "Tutorial on Forensic Speech Science," p. 14.

[8] The Aerospace Corporation Law Enforcement Development Group, "Applications of semi-automatic speaker identification techniques," .

[9] R. Rodman, D. McAllister, D. Bitzer, L. Cepeda, and P. Abbitt, "Forensic speaker identification based on spectral moments," *International Journal of Speech, Language and the Law*, vol. 9, no. 1, pp. 22–43, Mar. 2002, Number: 1.

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

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

[12] Seyyed Saeed Sarfjoo, Srikanth Madikeri, Mahdi Hajibabaei, Petr Motlicek, and Sébastien Marcel, "Idiap submission to the NIST SRE 2019 Speaker Recognition Evaluation," Idiap-RR Idiap-RR-15-2019, Idiap, Rue Marconi 19, 1920 Martigny, 11 2019.

[13] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," p. 5.

[14] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Apr. 2018, pp. 5329–5333, IEEE.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker Recognition for Multi-speaker Conversations Using X-vectors," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 5796–5800, IEEE.

[16] Sergey Ioffe, "Probabilistic Linear Discriminant Analysis," in *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz, Eds., vol. 3954, pp. 531–542. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, Series Title: Lecture Notes in Computer Science.

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv:1706.08612 [cs]*, May 2018, arXiv: 1706.08612.

[18] John J. Godfrey, Edward C. Holliman, and Jane McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, San Francisco, California, Mar. 1992, ICASSP'92, pp. 517–520, IEEE Computer Society.

[19] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484 [cs]*, Oct. 2015, arXiv: 1510.08484.

[20] Igor Szoke, Miroslav Skacel, Ladislav Mosner, Jakub Paliesek, and Jan "Honza" Cernocky, "Building and Evaluation of a Real Room Impulse Response Dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, Aug. 2019, arXiv: 1811.06795.

[21] Mael Fabien, Seyyed Saeed Sarfjoo, Petr Motlicek, and Srikanth Madikeri, "Improving Speaker Identification using Network Knowledge in Criminal Conversational Data," *arXiv:2006.02093 [cs, eess]*, June 2020.