2021 ISCA Symposium on Security and Privacy in Speech Communication
10-12 November 2021, Online

Graph2Speak: Improving Speaker Identification using Network Knowledge in Criminal Conversational Data

Maël Fabien[1,2], *Seyyed Saeed Sarfjoo*[1], *Petr Motlicek*[1], *Srikanth Madikeri*[1]

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

mael.fabien@idiap.ch, saeed.sarfjoo@idiap.ch, petr.motlicek@idiap.ch,
srikanth.madikeri@idiap.ch

Abstract

Criminal investigations mostly rely on the collection of speech conversational data in order to identify speakers and build or enrich an existing criminal network. Social network analysis tools are then applied to identify the central characters and the different communities within the network. This paper introduces a new method, Graph2Speak, to re-rank individuals after applying a speaker identification step, by leveraging the frequency of previous interactions extracted from a graph. We deploy our method on two candidate datasets for criminal conversational data, Crime Scene Investigation (CSI), a television show, and the ROXANNE simulated data. We demonstrate that our method can reduce the error rates of the speaker identification baseline by up to 12% (relative).

Index Terms: criminal conversational data, criminal networks, network analysis, speaker identification

1. Introduction

Conversational data is characterized by interactions between a set of characters, through text messages, telephone, or video calls for example. In criminal investigations, Law Enforcement Agencies (LEAs) collect criminal conversational data and build criminal networks to assess the links between suspects. This work is part of the ROXANNE project [1], a European Union's Horizon 2020 research project that develops a tool for LEAs to extract network knowledge from intercepted phone calls.

Speaker identification is a well known tool for criminal investigations, whether used by investigators [1, 2], or in court [3]. However, real-world criminal data is hard to access (high sensitivity), and to collect, due to the variety of modalities and channels required, the privacy, the information about structures of criminal networks, etc. Criminal data is rarely released publicly, and if so, only partial and anonymized data is published (e.g. only the graph structure) [4].

Audio data from criminal investigations are analyzed with a graph-like structure, consisting of nodes representing the identity of speakers and edges reflecting links between them, altogether describing the topology of the network. Networks are used for link prediction [4], node classification, community detection, central characters identification [5], or network disruption [6].

In this paper, we test the hypothesis whether the information contained in the topology of criminal networks, built on a set of conversations, can improve speaker identification performance, by favoring strong existing relationships between speakers. We propose an extension to existing works by Gao et al. [7] by computing for a dataset the speaker identification scores of each speaker, and re-ranking the potential speakers based on how frequently they talked to each other over the past. The main contribution of our approach is that it does not rely on an external source of data, since previous discussions between speakers are used to build the criminal network, which itself will influence the re-ranking of the next conversations. It also handles more than two speakers per conversation. We introduce two datasets employed to simulate criminal data, the CSI and the ROXANNE simulated datasets, that satisfy most of the requirements of criminal data, and also introduce the metric of the conversation accuracy, which is relevant for LEAs investigations. We release the code of our approach on GitHub [2].

Section 2 presents the existing works we rely on as well as our approach, Graph2Speak. Section 3 describes the candidate datasets for criminal investigation data and the evaluation metrics used. The results of a baseline speaker identification as well as the output of our experiments are presented in Section 4. Finally, section 5 discusses obtained results, the dataset, and the metrics used, as well as the future direction of network-based improvement of speaker identification, and section 6 depicts our conclusions.

2. Re-ranking speakers

2.1. Related work

The idea of improving the speaker identification process using an external source of data is not new. It has been explored by Khelif et al. in [1], by operating an inter-task fusion, using the accent, the language, or the gender as an external source of information. In [8], a relative improvement of 8% in terms of Equal-Error-Rate (EER) was obtained using a probabilistic fusion on NIST SRE 2008.

Gao et al. [7] have shown in previous works on Enron email [9] and phone call[3] databases that we can re-rank speaker pairs using network information. The knowledge present in the email database was used to build a network, and assess how often speakers talked to each other. This information was then used to re-compute the score of a pair of speakers on the Enron phone call database, increasing the score of the pair if speakers talked to each other frequently in the past. This work has shown an improvement in classification error and on the harmonic mean of the rank of the known speaker. This approach requires an external source of data, such as emails in the case of Enron, and focuses on the sub-case of conversations between 2 characters.

Due to the lack of realistic data, the impact of network anal-

[1]https://roxanne-euproject.org/

[2]https://github.com/maelfabien/Graph2Speak

[3]https://web.archive.org/web/20060208051051/,http://www.enrontapes.com/files.html

10.21437/SPSC.2021-3

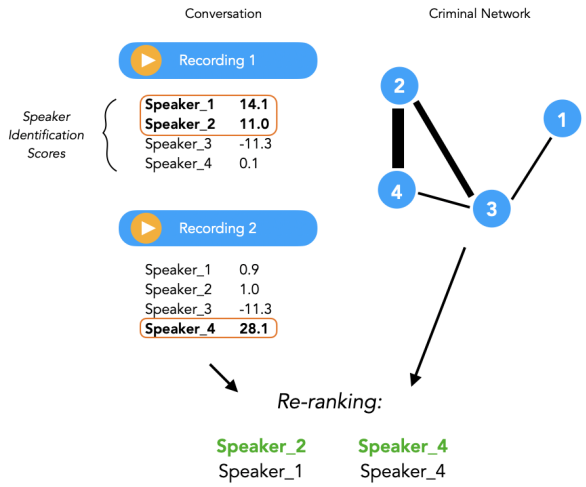

Fig. 1: *Graph2Speak re-ranking process*

ysis in speaker identification for criminal conversational data has not been extensively explored.

2.2. Graph2Speak re-ranking method

The logic behind the Graph2Speak approach is to compute a weighted average of the acoustic scores of the combination of potential speakers. Then, we multiply the resulting score by the relative frequency of the conversations between all the permutations of the speakers, hence increasing the score by a large factor if several characters have spoken a lot over the past.

More formally, we introduce s_{mc} as being a joint score of all speakers in a conversation c, considering the combination of speakers m. Our aim is to score all combinations of speakers (M_c in total) in conversation c, and choose the combination which maximizes the score. In the given conversation, there are N_{mc} different speakers. For each speaker k, the acoustic score s_k from the X-vector baseline is obtained. Relative centrality is defined as C_k as the number of interactions of speaker k divided by E_c, the total number of interactions in the network at that moment of the conversation c. The score of the combination m of speaker for conversation c can be written as:

$$s_{mc} = \frac{\sum_{k=1}^{N_{mc}} s_k(1+C_k)}{N_{mc}} \prod_{i<j}^{N_{mc}} (1+\lambda\frac{e_{i,j}}{E_c}). \qquad (1)$$

The second part of the score takes into account all the permutations of speakers, two-by-two, denoted i and j, within the list of candidates m. $e_{i,j}$ is the number of times speakers i and j talked to each other over the past. The factor λ denotes a weighting factor, which we set to 1 by default, but can be increased or lowered based on the expected importance of the previous interactions. For a given conversation c, we will select the optimal combination of speakers m such that $s^*_{mc} = \arg\max_{m \in M_c} s_{mc}$.

The overall process of our method is presented in Figure 1. In the first recording, speakers 1 and 2 have similar scores, whereas, in the second recording, speaker 4 has by far the highest score. From the topology of the network, we see that speakers 2 and 4 have been talking a lot over the past, and speakers 1 and 4 never met. Through the re-ranking process, the score of the pair 2 and 4 is higher than the pair 1 and 4, and therefore, the re-ranking favors speakers with a high frequency of interactions in the past.

The novelty of our approach compared to [7] is to focus on a single data source, and not external ones while handling conversations of more than 2 speakers. The re-rankings that we operate have impacts on the network we build, which will itself influence the next re-ranking. Since some of the computations imply evaluating all combinations between all speakers involved in a conversation, it can create a large combinatorial factor. Therefore, we apply a threshold on the scores under which we decide not to consider a speaker as a potential candidate for a given recording (i.e. -15 in our experiments).

3. Datasets

To the best of our knowledge, apart from the Enron e-mail database augmented with the Enron phone call database, as described by Gao et al. [7], no real-condition criminal conversational database exists. In the case of the Enron dataset, most fraudulent conversations were even removed, and the topology of the network we can build does not anymore reflect the fraudulent activities of Enron.

3.1. CSI dataset

We propose to use CSI television show[4], as a potential candidate for criminal investigation data. Each episode of the series includes a video of around 40 minutes, an audio file, and a manual transcript. The audio and video are extracted from the DVD of the show. The transcripts are publicly available [5]. We collected transcripts of 39 episodes and video/audio of 6 episodes, resulting in 4 hours of speech. Each episode involves more than 30 speakers and 56 distinct scenes/conversations per episode on average.

In the episodes of CSI, the main characters are the investigators. Suspects often remain secondary characters in the screenplay. A real investigation would obviously collect data on suspects only. However, we suppose that the structure of the networks that we can extract from this dataset is relevant for criminal investigation, by the number of speakers, the frequency of the interactions between the speakers, the number of sub-groups in each episode or the role of central characters who act as information bottlenecks.

3.2. ROXANNE simulated dataset

In order to test our approach on more realistic data too, the ROXANNE consortium has also collected simulated data. A screenplay, based on a real drug-trafficking use-case, was carefully prepared by one of the LEA partners, and the phone calls were made by members of the consortium, and recorded using Twilio. The simulated dataset contains 24 speakers, with a total of 100 phone calls, resulting in 155 minutes of speech, on topics related to drug dealing, in Czech, Russian, Vietnamese, German, and English. Although acted, this dataset provides close-to-real-life conditions. The topology of the network is meant to reflect the structure of a real-life criminal network, as illustrated in Figure 2. The multilingual framework, the background noises, and the short duration utterances present in the dataset are also challenging elements matching real-life conditions.

[4] As suggested by LEA partners from the consortium

[5] https://github.com/EdinburghNLP/csi-corpus

Fig. 2: *Network representation of the simulated data.*

3.3. Evaluation metrics

Speaker accuracy is a natural candidate metric for a speaker identification task in criminal investigations, but adding a wrong edge to the network of known connections could lead investigators on the wrong track, especially in cases where conversations imply many characters and a single misidentification can lead to several wrong edges. We introduce the conversation accuracy metric, which is computed as the percentage of conversations for which we could identify all speakers. In a criminal case with C conversations, each of the conversations involves the list of speakers s_c. Using a speaker identification system, we predict the list of speakers as being s_{pc}. The conversation accuracy then becomes $acc_C = \frac{1}{C}\sum_{c=1}^{C}\delta(s_{pc}, s_c)$ where $\delta(s_{pc}, s_c)$ is an indicator function equal to 1 if s_{pc} is equal to s_c.

This metric is easy to interpret for non-experts, and close to their objective of building accurate and reliable criminal networks at the conversation level. To compute it, we split the raw audio file of each episode into a sequence of conversations. These sequences of conversations were manually annotated for each of the 6 episodes of CSI, but could also be inferred automatically [10].

4. Experiments

4.1. Speaker identification baseline

Due to the relatively low volume of data available in both datasets, we used a pre-trained speaker identification system prepared for the NIST Speaker Recognition Evaluation (SRE19) [11]. The pre-trained system is described in Idiap's submission to the NIST SRE 2019 Speaker Recognition Evaluation [12]. The submission relies on Time Delay Neural Network (TDNN) [13] X-vector front-end [14, 15], followed by a Probabilistic Linear Discriminant Analysis (PLDA) [16] backend.

The speech data was first downsampled to 8 kHz (with an application of band-pass filtering between 20 and 3700 Hz). Then, 23-dimensional mel frequency cepstral coefficients (MFCCs) were extracted on 25 ms speech windows, with a frame-shift of 10 ms. To remove non-speech frames, energy-based Voice Activity Detection (VAD) was applied.

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]). The PLDA classifiers were trained on augmented versions of SRE. For CSI dataset, we only selected speakers for which we were able to collect at least 20 seconds of audio samples, with an average of 13 to 15 speakers per episode. For the ROXANNE simulated data, each of the 24 speakers was enrolled using its first utterance.

For the CSI dataset, the X-vector baseline correctly identified all speakers in 77.5% of the conversations on average and reached an average speaker accuracy of 89.6%. For the ROXANNE simulated data, the speaker accuracy is 88.1%, while the conversation accuracy is 79.6%.

4.2. Re-ranking Results

Using the Graph2Speak approach on CSI, the error rates on both the conversation and the speaker accuracy dropped by 12% (relative). For the ROXANNE simulated data, the conversation accuracy was improved by 2% (absolute), at 81.6%, and the speaker accuracy by 1.1% (absolute). A summary of the results is presented in Table 1.

Table 1: *Graph2Speak performance summary, in percentage.*

Source	Speaker Accuracy (%)		Conversation Accuracy (%)	
	Baseline	Graph2Speak	Baseline	Graph2Speak
CSI S01E07	91.6	**92.7**	84.4	**88.8**
CSI S01E08	91.9	**95.3**	80.6	**88.8**
CSI S02E01	**88.0**	**88.0**	71.4	**73.5**
CSI S02E04	88.1	**89.0**	**76.1**	**76.1**
CSI S02E06	86.2	**88.6**	70.9	**74.5**
CSI S02E09	**92.3**	91.3	**81.4**	79.1
CSI Average	89.6	**90.8**	77.5	**80.1**
Simulated data	88.1	**89.2**	79.6	**81.6**

In most of CSI episodes, the members of the investigation police are divided into groups. In this case, the network reflects the groups working on a task, and our approach significantly improves the speaker accuracy and the conversation accuracy. This is typically the type of network topology that we expect to encounter in real investigations. In other cases, the whole team works on the investigation without any distinct group being made, e.g. in S02E01 or S02E09, and the frequency of the conversations between speakers is constantly modified. In that case, our method does generally not worsen the result, but it illustrates some limits of the approach in networks which quickly change the structure. On the simulated data, where the scenario and the network topology strictly follow the ones of a real criminal investigation, Graph2Speak improves both the speaker and the conversation accuracy.

5. Discussion

In this paper, we first presented the CSI dataset as a potential candidate for criminal investigation data. This dataset shares some commons characteristics with real cases (number of characters, information bottlenecks, sub-groups, etc.). CSI is however not a perfect dataset, since it captures conversations between investigators. We also introduced the ROXANNE simulated dataset, a collection of phone calls between members of the ROXANNE consortium, with a screenplay prepared by LEAs. This dataset respects the main characteristics of a criminal investigation (recorded phone calls, network structure, discussed topic, length of conversations, etc.), and will be further used in ROXANNE project as a baseline to evaluate various systems.

Regarding the method, the Graph2Speak approach explores the various permutations of speakers for each conversation. Our approach, as opposed to [7], does not rely on an external source of data to estimate the number of links between speakers, but exclusively on previous interactions in the same data source. Therefore, re-ranking decisions affect the criminal network built, which itself influences the next re-ranking. We also offer an extension by applying our approach to more than 2 speakers in a conversation. The re-ranking approach that we propose remains simple, and several types of fusion could be also explored in future works. Network attributes, other than relative degree centrality and number of edges between two characters, could also be leveraged. Other types of conversational data could also be explored.

6. Conclusions

We introduced both CSI dataset and the ROXANNE simulated data as potential candidates for criminal data. We described the metric of conversation accuracy, and have shown that our re-ranking method based on previous interactions can improve both the conversation and the speaker accuracy on the CSI dataset and the ROXANNE simulated data.

We also discussed the limits of these datasets and of our re-ranking method, and offer some future directions to take by including social network analysis tools in speaker identification.

7. 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).

8. References

[1] Khaled Khelif, Yann Mombrun, Gideon Hazzani, Petr Motlicek, Srikanth Madikeri, Farhan Sahito, Damien Kelly, Luca Scarpato, and Emmanouil Chatzigavriil, "SIIP: An Innovative Speaker Identification Approach for Law Enforcement Agencies," p. 14, 2018.

[2] Geoffrey Stewart Morrison, Farhan Hyder Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, and Caroline Goemans Dorny, "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic Science International*, vol. 263, pp. 92–100, June 2016.

[3] Lawrence Solan and Peter Tiersma, "Hearing Voices: Speaker Identification in Court," *HASTINGS LAW JOURNAL*, vol. 54, pp. 65, 2003.

[4] 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.

[5] 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.

[6] 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.

[7] Ning Gao, Gregory Sell, Douglas W. Oard, and Mark Dredze, "Leveraging side information for speaker identification with the enron conversational telephone speech collection," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, , no. 2, pp. 577–583, 2017.

[8] Srikanth Madikeri, Petr Motlicek, and Subhadeep Dey, "A Bayesian Approach to Inter-task Fusion for Speaker Recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 5786–5790, IEEE.

[9] Bryan Klimt and Yiming Yang, "The Enron Corpus: A New Dataset for Email Classification Research," in *Machine Learning: ECML 2004*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, Eds., vol. 3201, pp. 217–226. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, Series Title: Lecture Notes in Computer Science.

[10] Philippe Ercolessi, Christine Senac, Philippe Joly, and Herve Bredin, "Segmenting TV series into scenes using speaker diarization," p. 4, 2011.

[11] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2019 NIST Audio-Visual Speaker Recognition Evaluation," p. 7, 2019.

[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, 2015.

[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, number 3, 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.