

Improving Speaker Identification using Network Knowledge in Criminal Conversational Data

Maël Fabien^{1,2}, Seyyed Saeed Sarfjoo¹, Petr Motlicek¹, Srikanth Madikeri¹

¹Idiap Research Institute, Martigny, Switzerland,

²Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

{mfabien, ssarfjoo, petr.motlicek, msrikanth}@idiap.ch

Abstract

Criminal investigations rely on the collection of conversational data. The identity of speakers must be assessed in order to build or improve the accuracy of an existing criminal network. Investigators use social network analysis tools to identify the most central character and the different communities within the network. We introduce Crime Scene Investigation (CSI) television show as a potential candidate for criminal conversational data. We also introduce the metric of conversation accuracy in the context of criminal investigations. In this paper, a speaker identification baseline is improved by re-ranking candidate speakers based on the frequency of previous interactions between speakers and the topology of the criminal network. The proposed method can be applied to conversations involving two or more speakers. We show that our approach outperforms the baseline speaker accuracy by 1.3% absolute (1.5% relative), and the conversation accuracy by 3.7% absolute (4.7% relative) on CSI data.

Index Terms: speaker identification, network analysis, criminal networks

1. Introduction

Conversational data defines data created 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. Hereby, we introduce ROXANNE [1], a European Union’s Horizon 2020 research project leveraging real-time network, text, and speaker analytics for combating organized crime.

In this paper, we test the assumption that past interactions and the topology of criminal networks can be used to improve speaker identification systems on conversational data, hence favoring strong existing relationships between speakers.

We propose an extension to existing works by Gao et al. [2] by computing, for a conversation, 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 and the topology of the network. The main difference of our approach is to offer an extension for more than two speakers per conversation, in the context of criminal investigations, without having to rely on any external source of data.

Section 2 presents the Crime Scene Investigation (CSI) data as a potential candidate for criminal conversational data. The evaluation metrics used are described in Section 3. The results of a baseline speaker identification will be presented in Section 4 Section 5 describe our re-ranking method and experimental results, while Section 6 discusses obtained results, the dataset,

and the metrics used, as well as the future direction of network-based improvement of speaker identification.

2. Criminal Investigation Data

Criminal networks are made of nodes, representing the identity of characters, and edges, which reflect links between characters, all together describing the topology of the network. When audio files are collected, the identity of the characters involved is assessed by a speaker identification system, given the enrolled models from the speakers. Based on the detected identities, a link between two characters in the network is added. Edges can then be weighted to reflect the number of previous interactions between two given characters. Edges with larger weights reflect a high frequency of interactions in the past, which can be crucial information in investigations.

Real-condition criminal conversational data are hard to collect due to the variety of modalities and channels required. Criminal data also require timestamps of all the interactions, names, and roles of each character. The topology of criminal networks is also specific, and this type of data is by nature private. For these reasons, criminal conversational data are specific and to our knowledge, apart from the Enron e-mail database [3] augmented with the Enron phone call database [4], as described by Gao et al. [2], no such real-condition database exists. However, most fraudulent conversations were removed from Enron database, and the topology of the network we can build does not entirely reflect the fraudulent activities of Enron. Although no ideal candidate database has been identified, we propose to use CSI television show as a potential candidate for criminal investigation data.

CSI is a popular criminal investigation television series in the United-States. Each episode of the series includes a video of around 40 minutes, an audio file, and a transcript. The audio and video are extracted from the DVD of the show. The transcripts were published by the natural language processing group of the University of Edinburgh for previous work on LSTM-based killer identification in CSI episodes [5]. The transcripts also describe the role of each speaker (suspect, killer, or other) and can be downloaded publicly on GitHub [6]. Each episode involves a team of investigators, journalists, victims, the family of the victims, suspects, and killers.

We collected transcripts of 39 episodes and video/audio of 4 episodes. Each episode involves on average more than 30 speakers. Utterances last on average 3 to 4 seconds. There are around 45 to 50 distinct scenes/conversations per episode. Figure 1 presents the distribution of the speech duration per speaker in season 1 episode 7.

We suppose that the structure of the networks that we can extract from this information is relevant for criminal investigation. One major limitation is that the episodes of CSI focus on

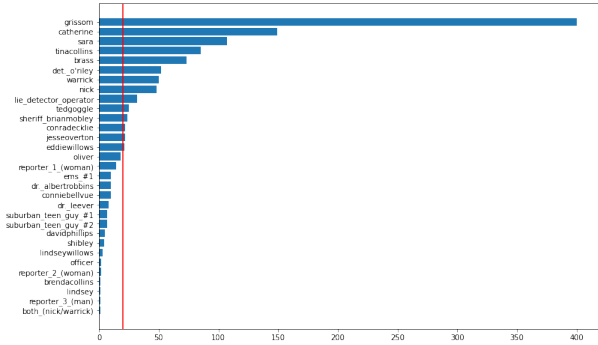


Figure 1: Distribution of aggregated speech time, in seconds, per speaker in season 1 episode 7.

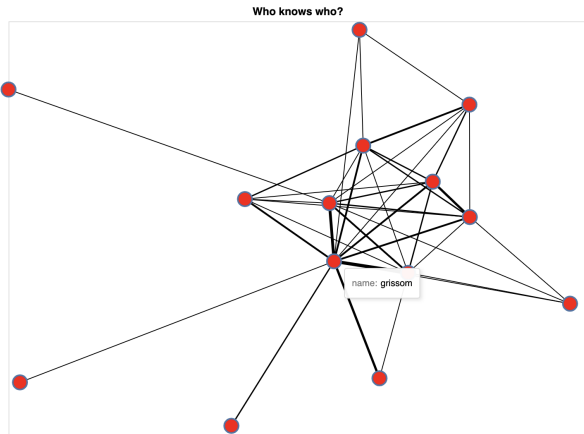


Figure 2: Network representation of Season 1 Episode 7.

the investigators as well as suspects, whereas a real investigation would only collect data on suspects. The episodes sequentially display a murder, the body/bodies are then discovered, police start the investigation, gather evidence, interrogate suspects, and identify the killer. We do not have information on the exact time at which each scene took place. Therefore, we consider the data as sequential by default and do not conduct any analysis on the time between conversations.

We build the "ground-truth" network using the transcripts provided, as illustrated in Figure 2. In the interactive tool developed, the thickness of the edges reflects the number of interactions between the speakers, and the tool-tip displayed on the node shows the name of the character. Note that the network we built is only a sub-network from the whole Season 1 Episode 7, focusing on the 14 characters with more than 20 seconds of speech, as illustrated in Figure 1. Speakers without sufficient data were mostly journalists or local police officers that did not have a central role in the conversations.

3. Evaluation metrics

Speaker accuracy is a natural candidate metric in criminal investigations. Another metric which is relevant for LEAs is the percentage of conversations for which we could identify all speakers, which we define as the conversation accuracy. In a criminal case with C conversations, each of the conversations involves the list of speakers s_c . Using our speaker identification system,

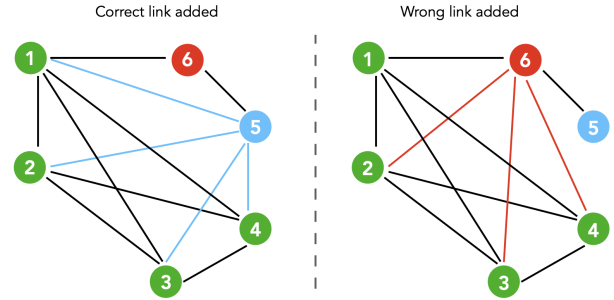


Figure 3: Effect of adding a wrong edge to an existing network.

we predict the list of speakers as being s_{pc} . The conversation accuracy then becomes:

$$acc_C = \frac{1}{C} \sum_{c=1}^C \delta(s_{pc}, s_c), \quad (1)$$

where $\delta(s_{pc}, s_c)$ is an indicator function equal to 1 if s_{pc} is equal to s_c .

The motivation behind the conversation accuracy is that adding a wrong edge to the network of known connections could lead investigators on the wrong track. Let us illustrate the process of misclassifying one of the speakers in a conversation in Figure 3, based on an existing network. The existing links between speakers are presented in black. In a new conversation, we suppose that speakers 1,2,3,4 and 5 were talking. On the left part of Figure 3, we added the correct edges in blue. On the contrary, if we misclassify speaker 5 as being speaker 6, three wrong edges are added at once, and the topology of the network is completely modified, which motivates the use of this metric in criminal investigations.

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 4 episodes, but could also be inferred automatically using speaker diarization techniques [7]. The identity of each speaker in a conversation is then assessed, and after processing all the conversations, the conversation accuracy is computed.

4. Speaker identification baseline

Due to the relatively low volume of data available, we used a pre-trained speaker identification system prepared for the NIST Speaker Recognition Evaluation (SRE19) dataset [8]. The pre-trained system is described in Idiap's submission to the NIST SRE 2019 Speaker Recognition Evaluation [9]. The submission relies on Time Delay Neural Network (TDNN) [10] X-vector systems [11, 12] with a Probabilistic Linear Discriminant Analysis (PLDA) [13] back-end.

We first downsampled speech data 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 [14] and on the augmented versions of Switchboard dataset [15] and SRE 2004 to 2010 with additive noise (MUSAN dataset [16]) and reverberation (RIR dataset [17]). The PLDA classifiers were trained on augmented versions of SRE.

We only selected speakers from CSI for which we were able to collect at least 20 seconds of audio samples. We then keep 20 seconds as enrolment and everything in test. Depending on the episode, we have 13 to 15 speakers among a total of 28 to 33 speakers. The X-vector/PLDA baseline heads an average speaker accuracy of 89.9% on the 4 episodes and an average conversation accuracy of 78.1%.

5. Re-ranking algorithm

Gao and al. [2] have shown in previous works on Enron email and phone call databases that we can re-rank speaker pairs using network information. The knowledge present in the email database was used to assess how often speakers talked to each other. This information was then used to re-compute the score of a pair of speakers, improving 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. However, these conclusions are linked to the topology of the network inherent to Enron phone call and email collection, and the approach is focused on conversations between two speakers only. Moreover, this approach requires an external source of data, such as emails in the case of Enron. In CSI dataset, several conversations involve more than 2 characters, and we don't have any external data source.

5.1. Method description

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 a conversation, and choose the combination which maximizes the score. In the given conversation, there are N_{mc} different speakers. For each speaker k , we obtain the acoustic score s_k from the X-vector baseline. We define the relative degree C_k as the number of interactions of speaker k divided by the total number of interactions. The score of the combination m for conversation c can be written as:

$$s_{mc} = \frac{1}{N_{mc}} \left(\sum_{i=1}^{N_{mc}} s_k (1 + C_k) \right) S_m \left(1 + \lambda \frac{e_{k_1, k_2}}{E_c} \right), \quad (2)$$

where S_m denotes all the permutations of speakers, two-by-two, denoted k_1 and k_2 , within the list of candidates m , e_{k_1, k_2} is the number of times speakers k_1 and k_2 talked to each other over the past, and E_c is the total number of conversations in the graph at the moment of the conversation c . The factor λ denotes a weighting factor, which we set to 1 by default, but has been adjusted to 0.2 in some of our experiments.

The logic behind this scoring approach is to weight the acoustic scores of each speaker by their degree centrality to favor speakers who have talked much over the past. Then, we multiply the resulting score by the frequency of the conversations between all the permutations of the speakers. For example, if characters A, B, and C talked frequently over the past, then the two-by-two permutations between A and B, B and C, A and C will lead to a large increase of the score of this combination of candidates. For a given conversation c , we will select the optimal combination of speakers m such that:

$$s_{mc}^* = \arg \max_{m \in M_c} s_{mc}. \quad (3)$$

The process of our method is presented in Figure 4. We can notice that for the first recording in the conversation, Speaker 1

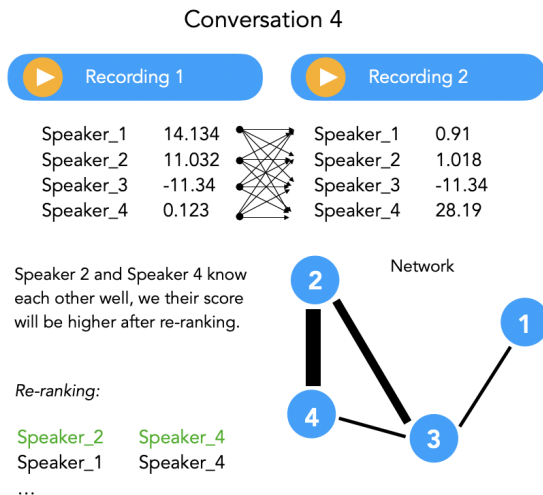


Figure 4: Re-ranking process.

and Speaker 2 have acoustic scores from the speaker identification system which are relatively high, and close. For the second recording, Speaker 4 is by far the candidate with the highest score. However, from the topology of the network, we see that Speakers 2 and 4 have been talking a lot over the past and Speaker 1 and 4 never spoke together. Through the re-ranking process, we multiply the score of Speaker 2 by its relative degree of centrality and the score of Speaker 4 by its degree of centrality. We then average the two scores and multiply the result by the relative number of interactions between Speaker 2 and 4. The score we obtain for the pair of Speaker 2 and 4 is higher than the pair of Speaker 1 and 4. Therefore, the re-ranking favors speakers with a high frequency of interactions in the past.

The novelty of our approach compared to [2] is to focus on a single data source, and not external ones. The re-rankings that we operate have impacts on the ways we built the network. We then estimate the number of interactions and the centrality of characters on this network, 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 number of combinations to compute. In order to limit the number of combinations tested, we apply a threshold on the scores under which we decide not to consider a speaker as a potential candidate for a given recording as part of a conversation.

5.2. Experimental results

We compute the scores on 4 episodes of CSI, resulting in 3 hours of conversations. For each episode, instead of re-using the same network, we created a new one. The X-vector baseline correctly identified all speakers in 78.1% of the conversations on average. Using the re-ranking approach, conversation accuracy reaches 81.8%. Speaker accuracy has also been improved from 89.9% to 91.25%. A summary of the results is presented in Table 1. The baseline is represented by the X-vector speaker identification system and the "network" is the proposed approach.

Table 1: *Re-ranking performance summary on CSI episodes.*

Approach	Speaker acc.	Conv. acc.
S01E07 baseline	91.6%	84.4%
S01E07 network	92.7%	88.8%
S01E08 baseline	91.9%	80.6%
S01E08 network	95.3%	88.8%
S02E01 baseline	88.0%	71.4%
S02E01 network	88.0%	73.5%
S02E04 baseline	88.1%	76.1%
S02E04 network	89.0%	76.1%
Average baseline	89.9%	78.1%
Average network	91.25%	81.8%

We reached a relative improvement of 4.7% in terms of conversation accuracy and 1.5% in speaker accuracy. For conversation and speaker accuracy, we obtained absolute improvements of 3.7% and 1.3%, respectively. In CSI, the way teams work on investigations is usually structured. The members of the investigation police are split into groups, and each group works on specific tasks. Most conversations hence take place within rather small groups. In this case, the network topology reflects the structure of groups working on a case, and our approach improves the speaker accuracy and the conversation accuracy by a significant factor. In other cases, the whole team works on the investigation without any distinct group being made, e.g. in S02E01. In that case, speaker accuracy is not improved, and we observe a slight improvement in conversation accuracy. Note that in this episode, the baseline X-vector approach performs quite poorly since the episode takes place in a casino, with a lot of background noises.

6. Discussions

In this paper, we first present the CSI dataset as a potential candidate for criminal investigation data. Although the main characters of the episodes are the members of the investigation police, CSI remains a good potential candidate for criminal conversational data. The network is time-varying, meaning that characters are discovered gradually and new interactions take place sequentially. The topology of the network reflects the creation of various groups (investigation, suspects, ...). The identification of the central speakers is also coherent since the main investigators also appear as the most central characters in our analysis.

However, CSI dataset has several limitations. The scenes are acted, and the audio quality, apart from some background music, is better than telephone data we would collect in real criminal investigations. The number of speakers we consider, due to the volume of data we can collect on each speaker, remains pretty low compared to real investigations. We do not have timestamps of each conversation, although having this kind of information would allow us to make recent conversations account more in the re-ranking. Finally, depending on the structure of the episode, i.e. whether investigation groups were built at the beginning of the investigation or not, the performance of our approach is impacted.

The proposed re-ranking approach explores the different permutations of speakers for each conversation. There are three novel elements in our approach compared to [2]. We do 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 made previously impact how we estimate the number of links between 2 candidate speakers in a conversation later on. We then offer an extension by applying our approach to more than 2 speakers in a conversation, thus creating large combinatorial factors controlled by thresholds. Finally, we apply our method in the context of a criminal investigation and show interesting improvements in speaker and conversation accuracy.

We have shown an interesting marginal gain in the context of criminal investigations, on CSI data. We conducted experiments in the context of the ROXANNE project and focused on criminal conversational data. Conversational data is, however, broader than criminal investigations, and we do expect that one can improve speaker identification systems in other contexts, on larger volumes of data and wider networks.

Network attributes, other than relative degree centrality and number of edges between two characters, could also be leveraged. We could indeed include notions from community detection or hierarchical embedding in the re-ranking algorithm.

7. Conclusions

We introduced CSI dataset as a potential candidate from criminal investigation data. We have introduced the metric of conversation accuracy in the context of a criminal investigation. We have shown that our re-ranking method based on previous interactions can improve speaker identification on CSI dataset by a relative 1.5%, and conversation accuracy by 4.7%.

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

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] [Online]. Available: <https://roxanne-euproject.org/>
- [2] N. Gao, G. Sell, D. W. Oard, and M. 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.
- [3] B. Klimt and Y. Yang, “The Enron Corpus: A New Dataset for Email Classification Research,” in *Machine Learning: ECML 2004*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3201, pp. 217–226, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-540-30115-8_22
- [4] [Online]. Available: <https://web.archive.org/web/20060208051051/http://www.enrontapes.com/files.html>
- [5] L. Frermann, S. B. Cohen, and M. Lapata, “Whodunnit? Crime Drama as a Case for Natural Language Understanding,” *arXiv:1710.11601 [cs]*, Oct. 2017, arXiv: 1710.11601. [Online]. Available: <http://arxiv.org/abs/1710.11601>

- [6] [Online]. Available: <https://github.com/EdinburghNLP/csi-corpus>
- [7] P. Ercolessi, C. Senac, P. Joly, and H. Bredin, "Segmenting TV series into scenes using speaker diarization," p. 4.
- [8] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST Audio-Visual Speaker Recognition Evaluation," p. 7.
- [9] S. S. Sarfjoo, S. Madikeri, M. Hajibabaei, P. Motlicek, and S. Marcel, "Idiap submission to the NIST SRE 2019 Speaker Recognition Evaluation," Idiap, Rue Marconi 19, 1920 Martigny, Idiap-RR Idiap-RR-15-2019, 11 2019.
- [10] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," p. 5.
- [11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, no. 3. Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333. [Online]. Available: <https://ieeexplore.ieee.org/document/8461375/>
- [12] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. 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: IEEE, May 2019, pp. 5796–5800. [Online]. Available: <https://ieeexplore.ieee.org/document/8683760/>
- [13] S. Ioffe, "Probabilistic Linear Discriminant Analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 3954, pp. 531–542, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/11744085_41
- [14] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv:1706.08612 [cs]*, May 2018, arXiv: 1706.08612. [Online]. Available: <http://arxiv.org/abs/1706.08612>
- [15] J. J. Godfrey, E. C. Holliman, and J. 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*, ser. ICASSP'92. San Francisco, California: IEEE Computer Society, Mar. 1992, pp. 517–520.
- [16] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484 [cs]*, Oct. 2015, arXiv: 1510.08484. [Online]. Available: <http://arxiv.org/abs/1510.08484>
- [17] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. H. 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. [Online]. Available: <http://arxiv.org/abs/1811.06795>