HAL
open science

Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast

Hervé Bredin, Johann Poignant

To cite this version:

Hervé Bredin, Johann Poignant. Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH, 2013, Lyon, France. hal-00953095

HAL Id: hal-00953095

https://inria.hal.science/hal-00953095v1

Submitted on 3 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast

Hervé Bredin

CNRS / LIMSI UPR 3251
Université Paris-Sud, Orsay, France
bredin@limsi.fr

Johann Poignant

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP
CNRS / LIG UMR 5217, Grenoble, France
johann.poignant@imag.fr

Abstract

Most state-of-the-art approaches address speaker diarization as a hierarchical agglomerative clustering problem in the audio domain. In this paper, we propose to revisit one of them: speech turns clustering based on the Bayesian Information Criterion (a.k.a. BIC clustering). First, we show how to model it as an integer linear programming (ILP) problem. Its resolution leads to the same overall diarization error rate as standard BIC clustering but generates significantly purer speaker clusters. Then, we describe how this approach can easily be extended to the audiovisual domain and TV broadcast in particular. The straightforward integration of detected overlaid names (used to introduce guests or journalists, and obtained via video OCR) into a multimodal ILP problem yields significantly better speaker diarization results. Finally, we explain how this novel paradigm can incidentally be used for unsupervised speaker identification (i.e. not relying on any prior acoustic speaker models). Experiments on the REPERE TV broadcast corpus show that it achieves performance close to that of an oracle capable of identifying any speaker as long as their name appears on screen at least once in the video.

Index Terms: speaker diarization, integer linear programming, speaker identification, multimodal fusion, optical character recognition

1. Introduction

Speaker diarization is the task of partitioning and labeling an audio stream into homogeneous speech segments according to the identity of the speaker. Most state-of-the-art approaches address it as a hierarchical (either agglomerative, divisive or a combination of both [1]) clustering problem in the audio domain [2, 3, 4]. In this paper, we propose to revisit one of them: speech turns clustering based on the Bayesian Information Criterion (a.k.a. BIC clustering) [5].

One of the main limitations of this type of approach is that each clustering iteration is performed locally and does not guarantee global optimality. Two clusters are merged because they are close to each other, independently of how similar (or dissimilar) they are to other clusters. New approaches based on spectral clustering [6] were recently introduced to try and cope with this limitation [7, 8, 9]. Indeed, spectral clustering techniques rely on the complete speech turns similarity matrix to project each of them into a low-dimensional manifold more suited for clustering. However, they still rely on a subsequent agglomerative clustering step with the same limitations as above.

Inspired by *Dupuy et al.* [10] who recently proposed a global optimization framework based on Integer Linear Programming (ILP), we investigate the use of ILP as a replacement for BIC clustering. Our approach differs from [10] both in the actual formulation of the ILP clustering problem (introduced in Section 2), and in the fact that we use it in place of BIC clustering (Section 3) while they still rely on BIC clustering as a preliminary step.

Section 4 describes how the monomodal (audio-only) ILP problem is extended to the audiovisual domain and TV broadcast in particular [11]. Building on our previous related work [12], we integrate overlaid person names (automatically detected by video OCR) into the ILP problem in order to jointly improve speaker diarization performance and achieve unsupervised speaker identification (i.e. not relying on any prior acoustic speaker models).

We present the results of our experiments, both for speaker diarization and cross-modal speaker identification, in Section 5. Section 6 concludes the paper.

2. Clustering as an ILP Problem

Clustering has been addressed in numerous scientific fields in the past: from graph mining and community detection [13] to natural language processing and co-reference resolution [14]. Classical clustering algorithms include K-means and hierarchical (agglomerative or divisive) clustering [15].

We describe the clustering of N items by a function $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} \colon [\![1, N]\!]^2 \to \{0, 1\}$$

$$(i, j) \mapsto \delta_{ij} = \begin{cases} 1 & \text{if items } i \text{ and } j \text{ are in the same cluster } (\mathcal{H} \text{ hypothesis}) \\ 0 & \text{otherwise } (\overline{\mathcal{H}} \text{ hypothesis}) \end{cases}$$

A few constraints ensure that $\boldsymbol{\delta}$ codes for a valid clustering:

Reflexivity. Every element must belong to its own cluster:

$$\forall i \in [\![1, N]\!], \quad \delta_{ii} = 1 \tag{1}$$

Symmetry. If i is in the same cluster as j, then j should be in the same cluster as i:

$$\forall (i, j) \in [\![1, N]\!]^2 \quad \delta_{ij} = \delta_{ji} \tag{2}$$

Transitivity. If i is in the same cluster as j ($\delta_{ij} = 1$) and j is in the same cluster as k ($\delta_{jk} = 1$), then i must be in the same cluster as k ($\delta_{ik} = 1$):

$$\forall (i, j, k) \in [\![1, N]\!]^3 \quad \delta_{ij} + \delta_{jk} - \delta_{ik} \leq 1 \tag{3}$$

Let $p_{ij} = p(\mathcal{H} \mid d_{ij})$ be the posterior probability (conditionally to their similarity d_{ij}) that elements i and j are in the same cluster. *Finkel & Manning* [14] propose to use Integer

Linear Programming to find the optimal clustering function $\boldsymbol{\delta}^*$ that maximize the intra-cluster similarity while simultaneously minimizing the inter-cluster one:

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\, f(\boldsymbol{\delta}) \tag{4}$$

$$f(\boldsymbol{\delta}) = \alpha \cdot \underbrace{\sum_{i,j} \delta_{ij} \cdot p_{ij}}_{\substack{\text{intra-cluster}\\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{i,j} (1-\delta_{ij}) \cdot (1-p_{ij})}_{\substack{\text{inter-cluster}\\ \text{dissimilarity}}}$$

The influence and choice of parameter $\alpha \in [0, 1]$ will be discussed later in Section 3.3. The details towards the solution of this ILP problem is out of the scope of this paper. For this work, we use the Gurobi solver [16] implementation of the *branch-and-bound* algorithm [17].

3. Application to Speaker Diarization

Our first contribution is to revisit the standard BIC clustering approach [5] by replacing its agglomerative clustering step with the ILP formulation described in previous section.

3.1. Baseline BIC clustering

Our baseline BIC clustering approach is derived from the multi-stage speaker diarization system introduced in [18].

3.1.1. Segmentation

Feature extraction yields one 38-dimensional vectors every 30ms, made of the concatenation of 12 LPC-based cepstral coefficients [19], their first- and second-order derivatives, and those of the log-energy. Speech activity detection is achieved based on Viterbi decoding with one 64-Gaussians Mixture Model (GMM) for speech, noisy speech, speech over music, pure music, and silence or noise. Speech segments are chopped into short pure segments (with a minimum duration of 2.5 seconds) using divergence-based segmentation. After training one GMM for each segment, Viterbi segmentation is used to refine segment boundaries.

3.1.2. Clustering

Agglomerative clustering is initialized with one cluster per segment, modeled with one Gaussian with full covariance matrix Σ trained on the $D = 12$-dimensional Mel Frequency Cepstral Coefficients (MFCC) and energy. The BIC criterion ΔBIC_{ij} [5] defines the similarity between clusters i and j:

$$\begin{aligned}\Delta\text{BIC}_{ij} = &(n_i + n_j) \log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| \\ &- \frac{1}{2} \cdot \lambda \cdot \left(D + \frac{1}{2} D\,(D+1)\right) \log(n_i + n_j)\end{aligned}$$

where n_k is the number of samples in cluster k and λ the penalty weighting coefficient. Each iteration merges the two most similar clusters i and j until the stopping criterion $\Delta\text{BIC} < 0$.

3.2. ILP BIC clustering

We propose to replace the clustering step of paragraph 3.1.2 by the ILP clustering introduced in Section 2. For fair comparison, we also use the BIC criterion as similarity measure ($d_{ij} = \Delta\text{BIC}_{ij}$). We apply Bayes' theorem to obtain the posterior probability $p_{ij} = p(\mathcal{H} \mid d_{ij})$ used in Equation 4:

$$p(\mathcal{H} \mid d) = \frac{1}{1 + \dfrac{p(d \mid \overline{\mathcal{H}})}{p(d \mid \mathcal{H})} \dfrac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}} \tag{5}$$

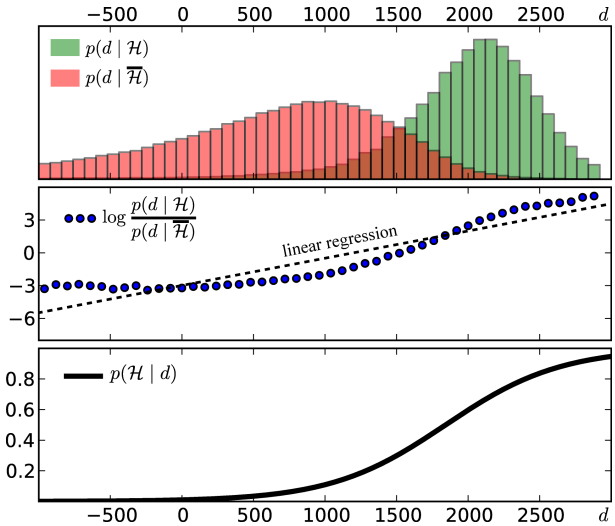

Figure 1: Estimation of posterior probability $p(\mathcal{H} \mid d)$ on the training set. Top: likelihood under hypothesis $\mathcal{H}$ (rightmost distribution, green) and $\overline{\mathcal{H}}$ (leftmost distribution, red). Middle: estimated log-likelihood ratio (•) and linear regression. Bottom: estimated posterior probability.

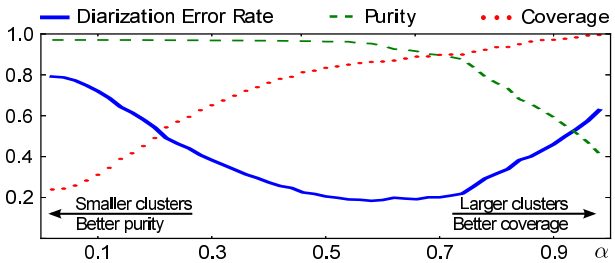

Figure 2: Influence of parameter α on the development set.

The likelihood ratio $p(d \mid \mathcal{H})/p(d \mid \overline{\mathcal{H}})$ is estimated using the training set described in Section 5. As shown in Figure 1, linear regression by minimization of the sum of squared error is used to fit an affine function to the log-likelihood ratio. Finally, the prior probability ratio is estimated using the same training set. We obtain $p(\overline{\mathcal{H}}) = 4.97 \times p(\mathcal{H})$: this leads to a shift to the right of the abscissa of $p(\mathcal{H} \mid d) = p(\overline{\mathcal{H}} \mid d) = 0.5$ in Figure 1.

3.3. Role of parameter α

Though detailed experimental results will be presented in Section 5, we here provide a general discussion about the behavior of the proposed ILP approach. In particular, Figure 2 illustrates the influence of parameter α on both cluster purity, cluster coverage [5] and resulting diarization error rate [20]. It shows that smaller values of α tend to generate small pure clusters while larger ones will result in large covering clusters at the expense of purity. Rewriting Equation 4 in both extreme regions ($\alpha \approx 0$ and $\alpha \approx 1$) explains this expected behavior:

$$[\alpha \approx \mathbf{0}]\ \boldsymbol{\delta}^* \approx \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \sum_{ij} \delta_{ij} \cdot (1 - p_{ij}) = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \sum_{ij} \delta_{ij}$$

$$[\alpha \approx \mathbf{1}]\ \boldsymbol{\delta}^* \approx \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{ij} \delta_{ij} \cdot p_{ij} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{ij} \delta_{ij}$$

The optimal value for diarization error rate is around $\alpha = 0.6$ though we notice a plateau for $\alpha \in [0.5, 0.7]$ common to both our development and test sets.

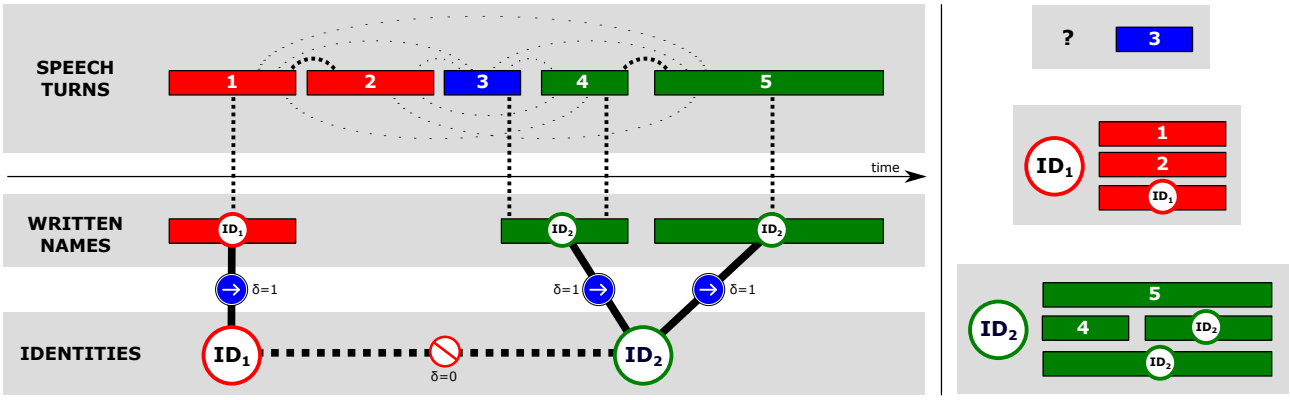

Figure 3: Cross-modal probability graph & expected clusters.

One name: $p = 0.956$ Two names: $p = 0.996$

Figure 4: Cross-modal probabilities depend on the number of simultaneous written names.

4. Cross-Modal Speaker Diarization

Most state-of-the-art speaker diarization approaches address this problem in the audio domain – the baseline and ILP BIC clustering are no exception. This is illustrated in the upper part of the graph in Figure 3 where each pair of speech turn vertices i and j is connected with an edge weighted by the posterior probability p_{ij}. However, in the case of TV broadcast such as TV news or talk-shows, other sources of information can be used to jointly improve speaker diarization and perform speaker identification. As a matter of fact, guests or reporters are often introduced to the viewer using overlaid title blocks containing their name, such as the ones shown in Figure 4. In this section, we propose to augment the speech turn graph with written name vertices (middle part of Figure 3) and identity vertices (bottom).

4.1. Written names

We use the video Optical Character Recognition (OCR) system proposed by *Poignant et al.* in [21] to automatically extract this information. Overlaid text boxes are first detected using a coarse-to-fine approach with temporal tracking. We then use the open-source `Tesseract` OCR system [22] to provide one transcription every ten frames. The transcriptions are finally merged to produce one transcription for each text box. As seen in Figure 4, not all detected text boxes are used to introduce a person: some provide other kind of information (name of the show, news flash, *etc.*). In order to only extract title blocks with person names, we use a list of names extracted from `Wikipedia` and the annotated training set introduced in Section 5 to automatically learn the spatial positions the most likely to be used for title blocks.

4.2. Cross-Modal Probability Graph

One vertex is added per written name occurrence. As shown by ⊖ traffic signs in Figure 3, each of them is constrained ($\delta_{ij} = 1$) to be in the same cluster as the corresponding identity vertex. Moreover, any two identity vertices are prevented from being in the same cluster by an additional constraint (⃠ sign, $\delta_{ij} = 0$).

To benefit from the influence of written names for speech turns clustering, cross-modal edges are added between cooccurring speech turns and written names. As a matter of fact, when a speech turn i and one (or more) written name(s) j occur simultaneously, the probability p_{ij} that the latter corresponds to the former is very high. This probability is learned using the annotated training set and depends on the number of cooccurring written names, as illustrated in Figure 4. For instance, in case two names are written while one person is speaking, the corresponding speech turn vertex is connected to both written name vertices with probability $p = 0.996$.

Since the resulting cross-modal probability graph is no longer complete, the objective function of Equation 4 is updated to take missing edges into account:

$$f(\boldsymbol{\delta}) = \alpha \cdot \sum_{\substack{i,j \\ \exists\, i \leftrightarrow j}} \delta_{ij} \cdot p_{ij} + (1-\alpha) \cdot \sum_{\substack{i,j \\ \exists\, i \leftrightarrow j}} (1-\delta_{ij}) \cdot (1-p_{ij})$$

where $\exists\, i \leftrightarrow j$ means that vertices i and j are connected.

4.3. Speaker Identification

The right part of Figure 3 describes the expected output of the resulting multimodal ILP clustering process: one cluster per person containing every corresponding vertex (speech turn, written name and identity when available). Therefore, the proposed framework can also be used for speaker identification.

After optimization, each speech turn is simply given the identity of the unique identity vertex belonging to the same cluster. Since the ILP problem is designed to prevent identity vertices from ending up in the same cluster (using ⃠ constraints), there can be at most one identity vertex in every cluster. In Figure 3, speech turns #1 and #2 are given the identity ID_1 and speech turns #4 and #5 the identity ID_2. However, it may happen that a cluster does not contain any identity vertex: speech turns simply remain with unknown identity (as does speech turn #3).

5. Experiments

5.1. Corpora

Figure 5 provides a graphical overview of the REPERE video corpus used in our experiments [23]. It contains 188 videos (30 hours) recorded from 7 different shows broadcast by the French TV channels *BFM TV* and *LCP*. The audio stream is fully annotated with labeled speech turns (*"who speaks when?"*). [23] provides a comprehensive description of the corpus and the associated annotation process.

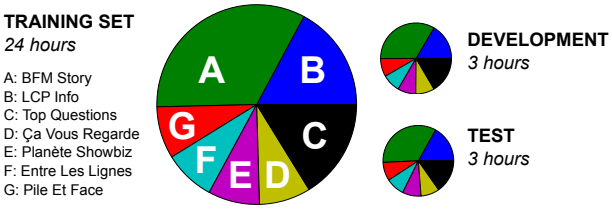

Figure 5: Training, development and test sets each contain 7 different types of shows (A to G).

5.2. Experimental Protocol

The whole corpus is divided into three corpora. The training set is used to estimate the log-likelihood and prior probability ratios introduced in Equation 5. While speaker audio annotations are complete, visual annotations are only available for one frame every 10 seconds. Therefore, we ran our written name extraction system on the training set and used its output to estimate the cross-modal probabilities of Figure 4. The development set is used to select the optimal value for parameters λ (for standard BIC clustering) and α (for ILP BIC clustering). The test set is used for evaluation.

Depending on the evaluated task (speaker diarization or speaker identification), parameters are optimized with respect to two evaluation metrics: diarization error rate (DER) for the former and identification error rate (IER) for the latter [20].

5.3. Late, Intermediate or Early Fusion

For comparison purposes, we also evaluated two other multimodal speaker identification approaches (dubbed "*Poignant et al.* [12]" and "*Constrained HAC*" in Tables 1 and 2). They differ from the proposed ILP approach in how early in the fusion pipeline the audio and visual modalities are combined.

The late fusion approach by *Poignant et al.* [12] consists in first applying standard BIC clustering and then propagating written names onto each speaker clusters and/or speech turns. The early fusion approach "*Constrained HAC*" relies on standard Hierarchical Agglomerative Clustering (HAC) with average linkage [15] based on ΔBIC speech turn similarity defined in Section 3.1. Starting with one cluster per speech turn, each iteration merges the two closest clusters for which there is no conflict in their cooccurring written names. Our proposed audio-visual ILP approach falls into the intermediate fusion category.

We also define two oracles ($\mathcal{O}_{\text{SD}}$ for speaker diarization and $\mathcal{O}_{\text{SI}}$ for identification) that provide error rate lower bounds.

5.4. Speaker Diarization

Table 1 summarizes the performance in terms of speaker diarization. The oracle $\mathcal{O}_{\text{SD}}$ relies on the same automatic pre-processing as the proposed approaches but differs in the clustering process which is perfect. The non-zero DER obtained by $\mathcal{O}_{\text{SD}}$ in Table 1 is mostly due to the lack of an overlapping speech detection module in pre-processing. Indeed, the REPERE corpus contains TV news and talk-shows where guests tend to speak simultaneously. Overall, it leads to a high missed detection error rate (only one speaker detected instead of two).

Audio-only ILP clustering achieves performance as good as standard BIC clustering. Yet, it yields much purer clusters (92% *vs.* 94%). Therefore, we expect that integrating this ILP approach into a multi-stage (BIC+CLR/*i-vector* [24]) one will lead to a sensible improvement in performance. Moreover, the integration of the visual modality leads to a significant improvement (-1% absolute DER) mostly due to an increase in coverage. It was expected as two small clusters cooccurring with the same written name tend to be grouped together into a larger one.

Approach	DER	Purity	Coverage	F-measure
Oracle $\mathcal{O}_{\text{SD}}$	7.4%	100.0%	96.5%	98.2%
BIC clustering	19.8%	92.1%	86.8%	89.4%
ILP (audio)	19.9%	94.1%	85.0%	89.3%
ILP (audio-visual)	**19.0%**	**94.2%**	85.7%	**89.7%**
Constrained HAC	21.6%	88.3%	**88.8%**	88.5%

Table 1: Speaker diarization.

Approach	IER	Precision	Recall	F-Measure
Oracle $\mathcal{O}_{\text{SI}}$	39.3%	100.0%	62.1%	76.6%
Poignant et al. [12]	44.4%	79.8%	59.4%	68.1%
ILP (audio-visual)	44.9%	**90.6%**	58.2%	**70.9%**
Constrained HAC	**42.2%**	83.9%	**61.2%**	70.7%

Table 2: Speaker identification.

5.5. Speaker Identification

Table 2 summarizes the performance in terms of speaker identification. The oracle $\mathcal{O}_{\text{SI}}$ is able to correctly identify any speaker as long as its name appears at least once in the same video, according to our written name extraction system. It does not have to appear simultaneously with any of the speaker speech turns: such an oracle is therefore very optimistic. The high IER obtained by $\mathcal{O}_{\text{SI}}$ in Table 2 is mostly due to the fact that anchors are very rarely introduced by overlaid names and therefore cannot be recognized in any way.

Looking at both IER and F-Measure, all three speaker identification approaches have very similar performance. The room for improvement to reach the oracle performance is also very limited. In the future, we will therefore try to lower the oracle performance bound by integrating supervised speaker identification approaches to better recognize anchors (for which training data can easily be collected). This can be achieved seamlessly by adding edges connecting speech turn and identity vertices, weighted by a probability derived from the identification score.

Though they have similar performance, the various approaches behave very differently from each other. *Audio-visual ILP* has much higher precision than *constrained HAC* for instance. One could try to combine them by first applying audio-visual ILP identification and then benefiting from the nearly perfect recall of the *constrained HAC* approach to identify the remaining unknown speakers.

6. Conclusion and Future Works

In this paper, we revisited BIC clustering as an Integer Linear Programming problem. We showed that standard and ILP-based BIC clusterings perform equally well with the exception that the latter yields much purer clusters. We therefore envision significant improvement over state-of-the-art approaches once it is integrated into a full-fledged multi-stage diarization system [18].

Then, we showed how easily the proposed framework can be extended to the audiovisual domain and lead to significant performance improvement. Overlaid written names are not the only source of information available in TV broadcast. For instance, we plan to extend this approach to multimodal person recognition [25] by adding face track vertices to the graph, or to named speaker identification [26] using addressee name vertices extracted from the automatic speech transcription with named entity detection.

Acknowledgments. This work was realized as part of the OSEO Quaero Program and the ANR QCompere project.

7. References

[1] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A Comparative Study of Bottom-Up and Top-Down Approaches to Speaker Diarization," *Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 2, pp. 382–392, Feb 2012.

[2] S. E. Tranter and D. A. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, Sept.

[3] D. A. Reynolds, P. Kenny, and F. Castaldo, "A Study of New Approaches to Speaker Diarization," in *Interspeech 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, September 2009.

[4] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 2, pp. 356–370, February 2012.

[5] S. S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[6] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2001, pp. 849–856.

[7] H. Ning, M. Liu, H. Tang, and T. Huang, "A Spectral Clustering Approach to Speaker Diarization," in *Interspeech 2006, 7th Annual Conference of the International Speech Communication Association*, 2006.

[8] J. Luque and J. Hernando, "On the use of Agglomerative and Spectral Clustering in Speaker Diarization of Meetings," in *Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 130–137.

[9] S. Shum, N. Dehak, and J. Glass, "On the Use of Spectral and Iterative Methods for Speaker Diarization," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, September 2012.

[10] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "i-Vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, 2012.

[11] A. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal Speaker Diarization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 79–93, 2012.

[12] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot, "Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, September 2012.

[13] M. E. J. Newman, "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, June 2006.

[14] J. R. Finkel and C. D. Manning, "Enforcing Transitivity in Coreference Resolution," in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT)*, 2008.

[15] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[16] Gurobi Optimization, Inc., "Gurobi Optimizer Reference Manual," http://www.gurobi.com, 2012.

[17] A. H. Land and A. G. Doig, "An Automatic Method of Solving Discrete Programming Problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, July 1960.

[18] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[19] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[20] J. G. Fiscus, J. S. Garofolo, A. N. Le, A. F. Martin, D. S. Pallett, M. A. Przybocki, and G. A. Sanders, "Results of the Fall 2004 STT and MDE Evaluation," in *Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004.

[21] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From Text Detection in Videos to Person Identification," in *International Conference on Multimedia & Expo (ICME)*, 2012.

[22] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ser. ICDAR '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 629–633.

[23] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus: a Multimodal Corpus for Person Recognition," in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

[24] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[25] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, ""Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[26] V. Jousse, S. Petitrenaud, S. Meignier, Y. Estève, and C. Jacquin, "Automatic Named Identification of Speakers using Diarization and ASR Systems," in *ICASSP 2009, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taïpei, Taïwan, April 2009.