Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Person Instance Graphs for Named Speaker Identification in TV Broadcast

Hervé Bredin[1], *Antoine Laurent*[1], *Achintya Sarkar*[1],
Viet-Bac Le[2], *Sophie Rosset*[1], *Claude Barras*[31]

[1] LIMSI / CNRS, Orsay, France
[2] Vocapia Research, Orsay, France
[3] Université Paris-Sud, Orsay, France

bredin@limsi.fr — http://herve.niderb.fr/

Abstract

We address the problem of named speaker identification in TV broadcast which consists in answering the question "*who speaks when?*" with the *real* identity of speakers, using person names automatically obtained from speech transcripts. While existing approaches rely on a first speaker diarization step followed by a local name propagation step to speaker clusters, we propose a unified framework called *person instance graph* where both steps are jointly modeled as a global optimization problem, then solved using integer linear programming. Moreover, when available, acoustic speaker models can be added seamlessly to the graph structure for joint named and acoustic speaker identification – leading to a 10% error decrease (from 45% down to 35%) over a state-of-the-art *i-vector* speaker identification system on the REPERE TV broadcast corpus.

1. Introduction

Named speaker identification is the task aiming at answering the question "*who speaks when?*" in an audio document using the sole knowledge of person names pronounced in this audio document. As such, it can be seen as a fully unsupervised speaker identification problem, where prior acoustic speaker models are not available.

It was first proposed in 2004 by *Canseco et al.* [1]. Names were manually classified based on their lexical context to indicate whether they refer to the speaker themselves, the addressee or someone else. *Tranter et al.* extended this approach in 2006 via automatic learning of these patterns from n-gram sequences centered on person names [2]. However, though they used automatic speech transcription, person name detection was still done manually. Also in 2006, *Mauclair et al.* used another approach based on semantic classification trees (SCT) to match names with speaker turns [3]. *Estève et al.* compared the two approaches in the same experimental conditions and concluded that, while *Tranter*'s approach performs best on manual transcription, *Mauclair*'s SCTs give significantly better results when applied on automatic transcription [4]. Finally, *Jousse et al.* further developed the SCT approach, and performed a detailed analysis of the influence of transcription or diarization errors on the overall speaker identification performance [5]. For instance, they report that identification error rates increase from 17% up to 75% when switching from manual to fully automatic name detection.

All existing approaches [1, 2, 3, 5] have in common that they rely on a preliminary and potentially erroneous speaker diarization step (where speech turns of the same speaker are automatically tagged with the same anonymous label) followed by a *local* name propagation step.

In contrast, in this paper, we adapt our previous work addressing unsupervised audiovisual speaker identification (using names written in overlaid text on TV) [6] to audio-only named speaker identification. Our proposed approach relies on a graphical representation (called person instance graph, introduced in Section 2) of the whole audio document and on a *global* optimization technique based on integer linear programming (described in Section 3) to jointly achieve speaker diarization and name propagation *at the same time*.

Moreover, another advantage of the proposed approach is that, when available, acoustic speaker models can be added seamlessly to the graph structure for joint named and acoustic speaker identification. It does not necessitate any additional (and potentially fallible) late fusion step, as in *El Khoury*'s work [7] where belief functions are used to combine a standard acoustic-based and a transcript-based (SCT) system.

Person instance graphs are introduced in Section 2. The resulting global optimization problem is described in Section 3. Section 4 introduces the experimental setup and results are discussed in Section 5. Section 6 concludes the paper.

2. Person Instance Graph

A person instance graph is a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$ where $\mathcal{V}$ is its set of vertices, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is its set of edges, and $p \in [0,1]^{\mathcal{E}}$ is a function associating a weight to every edge. Each vertex $v \in \mathcal{V}$ represents either a person (*identity* vertex) or an instantiation of a person (*instance* vertex). Every edge $(v, v') \in \mathcal{E}$ connects two vertices in the graph and is weighted by the probability $p_{vv'}$ that vertices v and v' correspond to the same person.

As illustrated in Figure 1 and depending on the targeted application, the same audio recording can lead to person instance graphs with different configurations. *Instance* vertices are represented as rectangles (large ones for speech turns, smaller ones for spoken person names) and circles stand for *identity* vertices.

For instance, the graph from configuration 1 only contains speech turns vertices $t \in \mathcal{V}$. It is a complete graph where every pair of speech turn vertices t and t' is connected by an edge weighted by the probability $p_{tt'}$ that they are the same person. However, a person instance graph is not necessarily complete. Hence, the graph from configuration 4 only contains edges between speech turns t and identities i, weighted by the probability p_{ti} that speech turn t was uttered by person i. It does not contain any edge between speech turns, for instance.

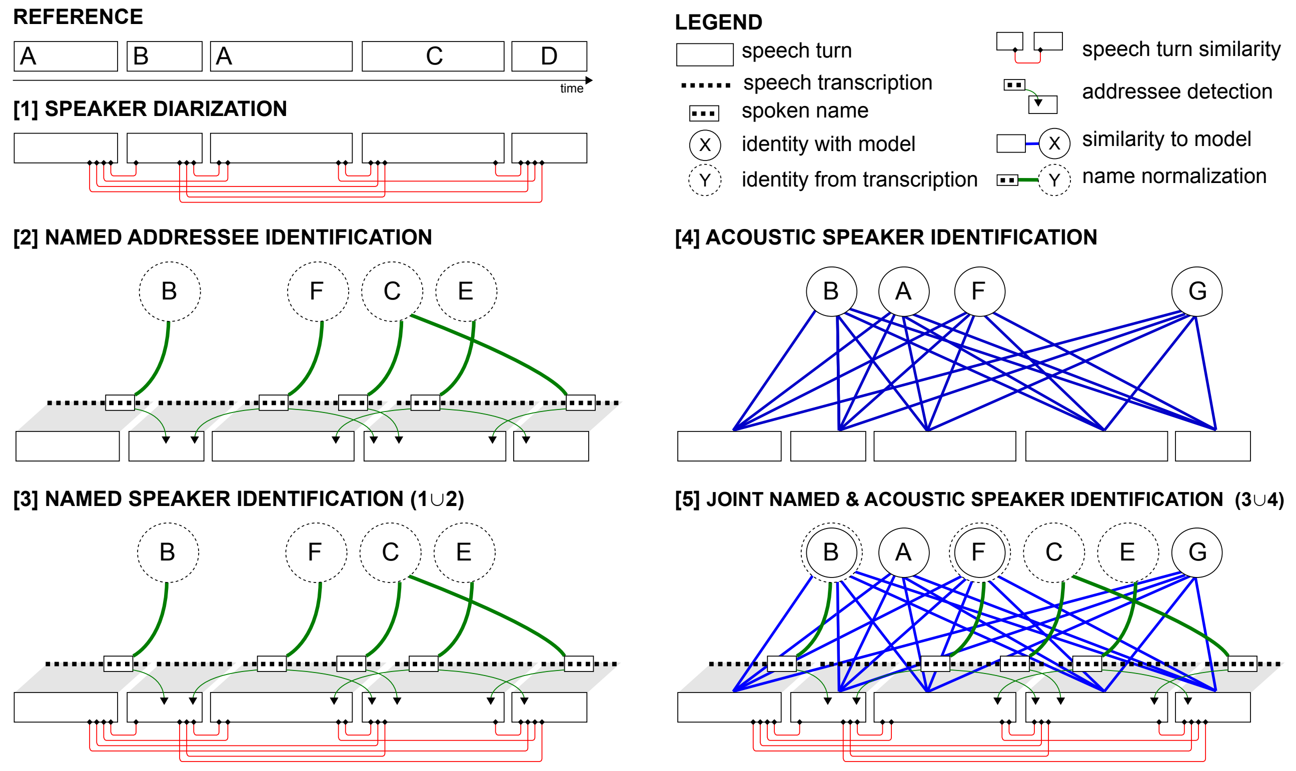

Figure 1: Person instance graph – different configuration for different applications.

2.1. Vertices

Figure 2 summarizes how vertices are automatically generated from an audio recording.

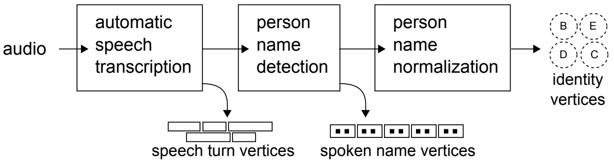

Figure 2: Speech processing pipeline.

2.1.1. Automatic speech recognition

First, a state-of-the art off-the-shelf speech-to-text system for French is used to transcribe the audio stream [8]. No task-specific adaptation was made other than adding person first and last names from the training set into the system vocabulary. One vertex $t \in \mathcal{T}$ is added to the graph for each resulting speech turn – obtained following a segmentation pipeline similar to the one described in [9].

2.1.2. Person name detection

Then, starting from the automatic speech transcription, the objective of the person name detection module is to detect person mentions whatever their forms are. A person mention can contain only part of the names (first name, middle name or last name) or any possible combination of them.

Two different classes of models were trained, using the Wapiti [10] implementation of conditional random fields (CRF). The first set of models is specialized in detecting parts of person mentions. The second one focuses on detecting complete person name mentions. Generally speaking, the models use standard features as in [11].

In order to take advantage of the complementarity of these models, a voting system (also trained with CRF) was learned using their outputs as features. As shown in Figure 2, one spoken name vertex $s \in \mathcal{S}$ can be added for each person mention detected by this final system.

2.1.3. Person name normalization

While there can be multiple *instance* vertices of the same person in a graph (one for every of their speech turns, one spoken name instance for every time their name is pronounced), there cannot be more than one *identity* vertex $i \in \mathcal{I}$ per person. To ensure unicity, a unique standardized identifier is given to each person, using the following naming convention: `First-Name_LASTNAME`. Simple heuristics are used to derive the standardized identifiers from the original speech transcript.

First, consecutive first and last names (detected by the previous person name detection system) are concatenated and capitalized as required. Then, normalized names sharing the same beginning `First-Name_LASTNAME` are grouped together. A majority vote allows to select one unique normalized version: every person name in the class are renamed accordingly. For instance, this allows to rename `Jean-Remi_BAUDOT_JEAN-REMI` and `Jean-Remi_BAUDOT_MERCI` into `Jean-Remi_BAUDOT` (French TV anchor). Finally, using the annotated training set, forced alignment of the automatic transcription onto the manual transcription allows to learn common speech recognition errors and build a mapping table as showed in Table 1.

Automatic transcription		Corrected person name
`Valerie_ROSSO_DEBORD`	⇒	`Valerie_ROSSO-DEBORD`
`Emmanuel_VALLS`	⇒	`Manuel_VALLS`
`Gaetane_MELIN`	⇒	`Gaetane_MESLIN`
`Rosine_BACHELOT`	⇒	`Roselyine_BACHELOT`

Table 1: Learned automatic transcription errors

These simple steps allows to improve the Slot Error Rate for normalized person name detection from 45.0% down to 35.7% on the test set. As shown in configuration 3 of Figure 1, each spoken name vertex $s \in \mathcal{S}$ can be connected with probability $p_{si_s} = 1$ to the corresponding normalized identity vertex $i_s \in \mathcal{I}_\mathcal{S}$, (where $\mathcal{I}_\mathcal{S}$ is the set of detected names after normalization).

2.2. Edges

Once vertices $\mathcal{V}$ are added to the person instance graph, edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ can be between selected pairs of vertices. The objective of this section is two-fold: describe which edges are added, and how the weighting function p is practically estimated:

$$\begin{aligned} p\colon \mathcal{E} &\rightarrow [0,1] \\ (v,v') &\mapsto p_{vv'} = p(\texttt{ID}(v) = \texttt{ID}(v') \mid v, v') \end{aligned} \tag{1}$$

2.2.1. *Speech turn similarity*

The first set of edges is $\mathcal{T} \times \mathcal{T} \subset \mathcal{E}$. Hence, every speech turns pair (t, t') can be connected to obtain configuration 1 in Figure 1.

In order to estimate $p_{tt'}$ weighting probabilities, each speech turn $t \in \mathcal{T}$ is first modeled with one Gaussian with full covariance matrix Σ_t trained on the $D = 12$-dimensional MFCC and energy. The similarity $d_{tt'}$ between two speech turns t and t' is then defined as the Bayesian Information Criterion $\Delta\text{BIC}(t, t')$ [12].

$$\begin{aligned} d_{tt'} = &(n_t + n_{t'}) \log |\Sigma_{t+t'}| \\ &- n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\ &- \frac{1}{2} \cdot \lambda \cdot \left(D + \frac{1}{2} D \left(D + 1 \right) \right) \log \left(n_t + n_{t'} \right) \end{aligned} \tag{2}$$

where n_t is the number of MFCC samples in speech turn t and λ a penalty weighting coefficient. Finally, we apply Bayes' theorem to obtain the posterior probability $p_{tt'}$:

$$\begin{aligned} p_{tt'} &= p(\texttt{ID}(t) = \texttt{ID}(t') \mid d_{tt'}) \\ &= \frac{1}{1 + \dfrac{\pi_{\neq}}{\pi_{=}} \cdot \dfrac{p(d_{tt'} \mid \texttt{ID}(t) \neq \texttt{ID}(t'))}{p(d_{tt'} \mid \texttt{ID}(t) = \texttt{ID}(t'))}} \end{aligned} \tag{3}$$

where the prior probabilities are assumed equal ($\pi_= = \pi_{\neq}$) and the likelihood ratio is estimated using isotonic regression in the logarithmic space, on the training set described in Section 4.

This is similar to our previous work [6]. It only differs in the fact the we use isotonic regression instead of linear regression. Indeed, we found that the former is more robust in low (and high) similarity regions with very few samples.

2.2.2. *Addressee detection*

In the REPERE corpus introduced in Section 4, speakers seldom pronounce their own name. Instead, spoken names $s \in \mathcal{S}$ are used either to address another particular speaker or to talk about someone else.

In this paragraph, we aim at deciding whether a name s pronounced during a speech turn t refers to the speaker of the previous speech turn t^-, the next speech turn t^+ or to someone else. This allows to add $s \leftrightarrow \{t^- \mid t^+\}$ edges to the graph, as illustrated in configuration 2 of Figure 1.

In order to estimate p_{st^-} and p_{st^+} probabilities, we rely on the n-gram approach proposed by *Tranter* [2]. As shown in Table 2, we first build the list of all contextual patterns (containing up to 4 words to the left of the detected name s, and up to 4 words to the right) from the training set introduced in Section 4. The precision of each pattern for (previous or next) addressee detection is then estimated on the same training set, and used as weights p_{st^-} and p_{st^+} of $s \leftrightarrow t^-$ and $s \leftrightarrow t^+$ edges when the same pattern is detected around a given s in the test document.

2.2.3. *Similarity to model*

Up to this point, no biometric supervision whatsoever was introduced in the person instance graph. Hence, mining the graph with configuration 3 from Figure 1 would result in completely unsupervised speaker identification. However, as shown in configuration 4 of Figure 1, it happens that voice models can be obtained from an annotated training set for a limited set of target speakers $\mathcal{I}^*$, allowing supervised open-set speaker identification using *i-vector* and probabilistic linear discriminant analysis (PLDA).

Following the *i-vector* approach [13], target speaker i and test speech turn t data is projected onto a lower dimensional subspace retaining the speaker and channel information, the *total variability space*. Speaker and channel factors are further decomposed through PLDA [14, 15] during scoring. The score d_{ti} between the *i-vector* w_t of speech turn t and the *i-vector* w_i of target speaker i is calculated as follows

$$d_{ti} = \log \frac{p(w_t, w_i \mid \texttt{ID}(t) = i)}{p(w_t, w_i \mid \texttt{ID}(t) \neq i)} \tag{4}$$

where hypothesis $\texttt{ID}(t) = i$ indicates that w_t and w_i are from the same speaker and hypothesis $\texttt{ID}(t) \neq i$ states that they are two different speakers. Identification scores d_{ti} are then calibrated into probabilities p_{ti} following the open-set speaker identification paradigm:

$$p_{ti} = \frac{\pi_i \cdot \exp\left(d_{ti}\right)}{\pi_{\textcircled{?}} + \displaystyle\sum_{i' \in \mathcal{I}^*} \pi_{i'} \cdot \exp\left(d_{ti'}\right)} \tag{5}$$

where $\pi_{\textcircled{?}}$ is the prior probability that speaker is unknown (*i.e.* $i \notin \mathcal{I}^*$) and prior probabilities π_i are assumed to be equal.

As illustrated in configuration 4 of Figure 1, one *identity* vertex per target speaker $i \in \mathcal{I}^*$ can be added to the person instance graph, connected to every speech turn vertex t with probability p_{ti}, leading to the overall set of *identity* vertices $\mathcal{I} = \mathcal{I}_\mathcal{S} \cup \mathcal{I}^*$. Note that the intersection of $\mathcal{I}^*$ and $\mathcal{I}_\mathcal{S}$ (spoken name *identity* vertices) is not necessarily empty – as illustrated by *identity* vertices B and F in configuration 5.

3. Mining Person Instance Graph

All person instance graphs of Figure 1 describe the same audio recording involving four persons (`A`, `B`, `C` and `D`). They contain five speech turn vertices $\mathcal{T} = \{t_1, t_2, t_3, t_4, t_5\}$ and five spoken name vertices $s \in \mathcal{S}$. Mining those graphs for speaker identification consists in automatically assigning the correct *identity*

Left context				Right context			Counts	Precision		English translation
w_{-3}	w_{-2}	w_{-1}		w_1	w_2	w_3		p_{st^-}	p_{st^+}	
–	merci	,	[s]	–	–	–	9	88%	0%	thank you, [s]
–	–	merci	[s]	pour	ces	–	5	80%	0%	thank you [s] for those
avec	nous	,	[s]	on	–	–	6	66%	0%	with us, [s] we
parole	est	a	[s]	–	–	–	29	3%	72%	let's listen to [s]
–	l'actualité	avec	[s]	–	–	–	17	0%	76%	latest news with [s]
–	–	–	[s]	revient	sur	–	6	0%	66%	[s] discusses

Table 2: Sample addressee detection patterns, with their counts and precisions on the training set

vertex to each speech turn: $t_1 \rightarrow \texttt{A}$, $t_2 \rightarrow \texttt{B}$, $t_3 \rightarrow \texttt{A}$, $t_4 \rightarrow \texttt{C}$ and $t_5 \rightarrow \textcircled{?}$. Notice how even the most complete graph (configuration 5) does not contain the actual identity of speech turn t_5, which would therefore remain unknown ($\textcircled{?}$).

More generally, given a person instance graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$ with *identity* vertices $\mathcal{I} \subset \mathcal{V}$, we aim at finding the optimal identification function ID defined as follows:

$$\begin{aligned} \texttt{ID}\colon \mathcal{V} &\rightarrow \mathcal{I} \cup \{\textcircled{?}\} \\ v &\mapsto \begin{cases} v & \text{if } v \in \mathcal{I} \text{ (\textit{i.e.} } v \text{ is an \textit{identity} vertex);} \\ i & \text{if } \exists\, i \in \mathcal{I} \text{ s.t. } v \text{ is an instance of } i; \\ \textcircled{?} & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

This can also be seen as a clustering problem where all instances of a given identity must be grouped together (alongside the actual identity itself).

Inspired by [16], we proposed in [6] to model clustering as an Integer Linear Programming (ILP) problem. While we relied on written names (obtained from overlaid text in video) to achieve unsupervised speaker identification in [6], the present work addresses a much more difficult task by only relying on the audio stream and more ambiguous spoken names.

3.1. Clustering function

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other groups (clusters). Any output of a valid clustering algorithm can be described by a *clustering function* $\boldsymbol{\delta}$, as follows:

$$\begin{aligned} \boldsymbol{\delta}\colon \mathcal{V} \times \mathcal{V} &\rightarrow \{0, 1\} \\ (v, v') &\mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

However, reciprocally, a function $\boldsymbol{\delta} \in \{0,1\}^{\mathcal{V}\times\mathcal{V}}$ does not always correspond to a clustering output. Additional constraints are needed in order to guarantee a valid clustering: (a) reflexivity, (b) symmetry and (c) transitivity. We define $\Delta_{\mathcal{V}} \subset \{0,1\}^{\mathcal{V}\times\mathcal{V}}$ the subset of functions verifying these constraints:

$$\Delta_{\mathcal{V}} = \begin{cases} \boldsymbol{\delta} \in \{0,1\}^{\mathcal{V}\times\mathcal{V}} \text{ s.t. } \forall\,(v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 \\ \text{(b) } \delta_{vv'} = \delta_{v'v} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 \end{cases} \tag{8}$$

While it is trivial to integrate reflexivity (a) and symmetry (b) constraints in the ILP framework, the transitivity constraints (c) need a little bit of work, summarized in Equations (9):

$$\begin{aligned} \forall \left(v, v', v''\right) \in \mathcal{V}^3,\ \ & \delta_{vv'} + \delta_{v'v''} - \delta_{vv''} \leq 1 \\ & \delta_{v'v''} + \delta_{v''v} - \delta_{v'v} \leq 1 \\ & \delta_{v''v} + \delta_{vv'} - \delta_{v''v'} \leq 1 \end{aligned} \tag{9}$$

Additionnally, each *instance* vertex can correspond to at most one *identity*. Therefore, the following constraints are added to the ILP problem:

$$\forall v \in \mathcal{V},\ \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1 \tag{10}$$

In particular, when combined with reflexivity contraints ($\delta_{ii} = 1$), Equation (10) implies that two *identity* vertices cannot end up in the same cluster:

$$\forall \left(i, i'\right) \in \mathcal{I}^2,\ i \neq i' \implies \delta_{ii'} = 0 \tag{11}$$

Finally, we explicitely constrain spoken names s to be in the same cluster as their corresponding *identity* vertex i_s:

$$\forall s \in \mathcal{S},\ \delta_{si_s} = 1 \tag{12}$$

3.2. Objective function

When clustering a person instance graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$, we aim at finding the clustering function $\boldsymbol{\delta} \in \Delta_{\mathcal{V}}$ with constraints (10) and (12) that maximizes the intra-cluster similarity while minimizing the inter-cluster similarity:

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \Delta_{\mathcal{V}}}{\operatorname{argmax}}\ \mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) \tag{13}$$

where $\alpha \in [0, 1]$ is an hyper-parameter controlling the size of the clusters, and the objective function $\mathcal{L}^{\alpha}$ is defined as follows:

$$\begin{aligned} \mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1} \Big[& \alpha \cdot \overbrace{\sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\text{intra-cluster similarity}} \\ & + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \Big] \end{aligned} \tag{14}$$

By design, a person instance graph usually contains many more $t \leftrightarrow t'$ edges (between any two speech turns) than it does $t \leftrightarrow s$ edges (only between a spoken name and the previous or following speech turns). Therefore, Equation (14) implicitly gives more importance to the former, at the expense of the latter. To compensate for this behavior, we extend the objective function in the following way:

$$\mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) = \sum_{\substack{x \in \{\mathcal{T}, \mathcal{S}, \mathcal{I}\} \\ y \in \{\mathcal{T}, \mathcal{S}, \mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}\left(\boldsymbol{\delta}, \mathcal{E} \cap (x \times y), p\right)$$

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \Delta_{\mathcal{V}}}{\operatorname{argmax}}\ \mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) \tag{15}$$

with $\alpha_{xy} \in [0,1]$, $\beta_{xy} \in [0,1]$ and $\sum_{x,y} \beta_{xy} = 1$. In other words, depending on the value of hyper-parameter $\boldsymbol{\beta}$, edges may be weighted differently depending on the type of vertices they connect.

3.3. Solution

This optimization problem falls into the Mixed-Integer Linear Programming (MILP) category. As such it can be solved by the *Gurobi Optimizer*, available freely for academic research purposes [17]. The resulting optimal solution $\boldsymbol{\delta}^*$ can then be used to associate a unique identity to each *instance* vertex:

$$\begin{aligned} \texttt{ID}_{\boldsymbol{\delta}^*} : \mathcal{V} &\to \mathcal{I} \cup \{\textcircled{?}\} \\ v &\mapsto \begin{cases} i & \text{if } \exists\, i \in \mathcal{I} \text{ s.t. } \delta^*_{vi} = 1, \\ \textcircled{?} & \text{otherwise.} \end{cases} \end{aligned} \tag{16}$$

Note that constraints (10) make sure that each *instance* vertex is connected to at most one *identity* vertex. Moreover, it might happen that an *instance* vertex v is not connected to any *identity* vertex. Hence, it remains anonymous: $\texttt{ID}_{\boldsymbol{\delta}^*}(v) = \textcircled{?}$.

3.4. Transitivity constraints relaxation

As far as person identification is concerned, Equation (16) shows that the only important objective is that every *instance* vertex v is associated to its correct *identity* vertex $i \in \mathcal{I}$. In particular, there is no need for two *instance* vertices v and v' of the same person i to be connected to each other ($\delta_{vv'} = 1$), as long as they are correctly connected to the correct *identity* vertex i ($\delta_{vi} = 1$ and $\delta_{v'i} = 1$). Therefore, strict transitivity constraints defined in Equation (8.c) can be relaxed in the following way:

$$\begin{aligned} &\forall \left(v, v', i\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^2 \times \mathcal{I}, \\ &\qquad \delta_{vi} = 1 \wedge \delta_{v'i} = 1 \;\not\Rightarrow\; \delta_{vv'} = 1 \end{aligned} \tag{17}$$

Formally, this is achieved by replacing the strict transitivity constraints defined in Equation (9) by the following loose transitivity constraints (18) and (19):

$$\forall \left(v, v', v''\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^3, \quad \begin{array}{l} \delta_{vv'} + \delta_{v'v''} - \delta_{vv''} \leq 1 \\ \delta_{v'v''} + \delta_{v''v} - \delta_{v'v} \leq 1 \\ \delta_{v''v} + \delta_{vv'} - \delta_{v''v'} \leq 1 \end{array} \tag{18}$$

$$\forall \left(v, v', i\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^2 \times \mathcal{I}, \quad \begin{array}{l} \delta_{vv'} + \delta_{vi} - \delta_{v'i} \leq 1 \\ \delta_{vv'} + \delta_{v'i} - \delta_{vi} \leq 1 \end{array} \tag{19}$$

Relaxing transitivity constraints has two main practical implications. The first one is that the size of the optimization problem is reduced and can therefore be solved more quickly. But, most of all, the second benefit of relaxing constraints is that it leads to better speaker identification performance [18].

4. Experiments

4.1. REPERE corpus

Figure 3 provides a graphical overview of the REPERE video corpus used in our experiments [19] and to be released publicly by ELDA in 2014.

It contains 267 videos (45 hours) recorded from 7 different shows broadcast by the French TV channels *BFM TV* and *LCP*. Selected shows mostly consist of talk shows, celebrity shows or news. The audio stream is manually annotated with labeled speech turns (*"who speaks when?"*). Manual speech transcription and person name mentions are also provided.

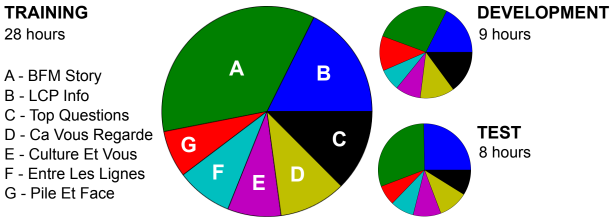

Figure 3: Training, development and test sets each contain 7 different types of shows (A to G).

4.2. Evaluation metrics

Given the reference annotation r and the automatic hypothesis h, we define the Identification Error Rate (IER) as the main evaluation metric:

$$\text{IER}(r, h) = \frac{\text{confusion} + \text{miss} + \text{fa}}{\text{total}} \tag{20}$$

where total is the total speech duration in the reference r, confusion is the duration of speech incorrectly identified in hypothesis, and miss and fa measure the duration of speech activity detection errors (missed detection and false alarms, respectively). An hypothesis is considered correct if the person name is correctly normalized (*e.g.* using `BREDIN` in place of `Herve_BREDIN` is incorrect).

Though the IER conveniently provides a unique value to compare different approaches, we also report the complementary values of precision and recall to help analyse their behavior.

4.3. Experimental protocol

The training set is used to estimate parameters for computation of p_{tt}, p_{ti} and p_{ts} introduced in Section 2. The development set is used to select the optimal values for hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ introduced in Section 3:

$$\left(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\right) = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\text{argmin}} \; \mathbb{E}_{\text{dev}} \left[\text{IER}(r, h(\alpha, \beta))\right] \tag{21}$$

Hyper-parameter tuning is achieved using random search. Indeed, *Bergstra & Bengio* showed that random search is usually able to find models that are as good or better than deterministic grid search within a small fraction of the computation time [20]. Finally, the test set is used for evaluation.

4.4. *i-vector* implementation details

Acoustic feature vectors are extracted from the speech signal on the 0-8kHz bandwidth every 10ms using a 30ms Hamming window. They consist of 15 PLP-like cepstrum coefficients [21] with 15 delta coefficients and delta energy, for a total of 31 features. Feature warping normalization is performed using a sliding window of 3 seconds in order to reduce the effect of the acoustic environment [22]. The Universal Background Model is a mixture of 256 diagonal Gaussians trained on a multilingual broadcast corpus. Then, three annotated data sources were used to train one *i-vector* w_i per speaker $i \in \mathcal{I}^*$: the REPERE training [19], the ETAPE training and development data [23] and additional French politicians data extracted from French radio broadcast. Only speakers with more than 30 seconds training data were kept, resulting in $|\mathcal{I}^*| = 706$ speaker *identity* vertices. The *total variability space* is trained using 39356 speech segments of variable length (few seconds to several minutes) collected over the target speakers (around 15 segments/speaker). 400-dimensional *i-vector* is considered for characterization of the speech segment. In test phase, only speaker factor (channel factor is kept fixed equal to the dimension of *i-vector* i.e. 400) of PLDA is varied to find the optimal performance of the speaker identification system on development data set. Before PLDA, *i-vectors* are length-normalized by two iterations of Eigen Factor Radial algorithm [24].

5. Results and Discussion

Table 3 summarizes the performance of the proposed approaches depending on the configuration of the person instance graphs.

5.1. Oracles

Depending on the configuration, not all speech turns can be identified. For instance, it might happen that no acoustic model is available for a given speaker and that their name is never mentioned. In order to determine the IER lower bound, we performed oracle experiments, also reported in Table 3. An oracle is capable of correctly identifying any speech turn as long as the corresponding *identity* vertex is available in the graph.

For instance, line C in Table 3 shows that it is theoretically possible to correctly identify 56.1% (recall) of the total speech duration in an unsupervised way by propagation of the spoken names (our approach only does half of it, with 29.4% recall). When all sources of information are combined for joint named and acoustic speaker identification, one cannot expect to get better than IER $= 14.7\%$. In comparison, our best system (using both named and acoustic speaker identification) reaches IER $= 35.4\%$, which is still 10% better than the state-of-the art *i-vector* acoustic speaker identification alone.

5.2. Unsupervised speaker identification

Lines A to C provide performance comparison for fully unsupervised speaker identification. In particular, line A allows to evaluate the precision (around 30%) of spoken name propagation (because the corresponding configuration 2 in Figure 1 only contains $t \leftrightarrow s \leftrightarrow i$ edges).

As expected, adding speech turn similarity edges $t \leftrightarrow t'$ for named speaker identification significantly improves recall (it nearly doubles from 16% to 29%) because spoken name can then be propagated to other speech turns of the same speaker. However, the major improvement in terms of precision (from 30% to 54%) is less trivial to explain. Let us look at the example proposed in Figure 1. While spoken name F (pronounced during speech turn t_3) could be incorrectly propagated to both speech turns t_2 and t_4 with configuration 2, adding edge $t_2 \leftrightarrow t_4$ with a very small probability may have the cascading effect to prevent at least one of them from being tagged as F, and thus improve precision.

On a similar broadcast news corpus (ESTER 1), the approach by *Jousse et al.* based on semantic classification trees [5] led to a precision of 42% (*vs.* 53% for ours) and recall of 18% (*vs.* 29%). We acknowledge that those results cannot be fairly compared because of different experimental conditions. However, our proposed approach seems to be more robust because it does not rely on a preliminary (and potentially error-prone) speaker diarization step: speaker diarization and identification are actually achieved at the same time.

5.3. Supervised speaker identification

Lines D and E actually provides an evaluation of a state-of-the-art *i-vector* acoustic speaker identification. Indeed, it can be demonstrated that solving the optimization problem with configuration 4 (line D) leads to the following solution (with $\theta = 1 - \alpha_{\mathcal{T}\mathcal{I}^*}$):

$$\mathtt{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\operatorname{argmax}}\ p_{ti} & \text{if } p_{ti^*} > \theta \\ \textcircled{?} \text{ otherwise.} \end{cases} \tag{22}$$

This is strictly equivalent to the standard open-set speaker identification paradigm: for each speech turn, select the most probable speaker model as long as its probability is higher than a predefined threshold θ.

In practice, hyper-parameter tuning on the development set leads to the automatic selection of $\theta \approx 0.20$ (or $\alpha_{\mathcal{T}\mathcal{I}^*} \approx 0.8$), almost perfectly matching the actual unknown prior probability $\pi_{\textcircled{?}} = 0.21$ of the development set.

5.4. *"the best of both worlds..."*

Line F of Table 3 shows how the state-of-the-art acoustic speaker identification approach (IER $= 45.3\%$) and the proposed unsupervised named speaker identification approach (IER $= 72.3\%$) can be advantageously combined into a joint named & acoustic approach (IER $= 35.4\%$, a 10% absolute – 22% relative – improvement over the state-of-the-art baseline).

As illustrated in configuration 5 of Figure 1, this combination does not rely on any (potentially error-prone) additional fusion step. It suffices to merge configurations 3 and 4 into a joint graph and apply the same hyper-parameter tuning and ILP optimization as before.

While late fusion approaches usually lead to increase in precision with similar recall, we note that our proposed scheme benefits from the intrinsic complementarity of the two subsystems – with improvement for both precision $(+2.5\%)$ and recall $(+10.3\%)$.

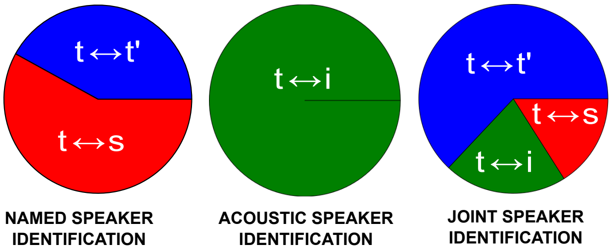

Figure 4: Hyper-parameters $\beta = [\beta_{\mathcal{T}\mathcal{T}}, \beta_{\mathcal{T}\mathcal{S}}, \beta_{\mathcal{T}\mathcal{I}}]$ after optimization on the development set.

Finally, Figure 4 provides a useful insight at the weights β given to each type of edges ($t \leftrightarrow t'$, $t \leftrightarrow s$ and $t \leftrightarrow i$) by the hyper-parameter tuning step, for the three proposed approaches. It is noticeable that, even if the supervised acoustic approach performs much better than the unsupervised one (IER $= 45.3\%$ *vs.* 72.3%), $t \leftrightarrow s$ and $t \leftrightarrow i$ edges are still given approximately the same weights in the joint approach. This confirms the observation that they are very complementary.

5.5. Name detection errors

Table 3 also contains experimental results with manual speech processing pipeline (*i.e.* manual speech transcript, manual person name detection and manual person name normalization),

		Fully automatic name detection			Fully manual name detection		
		IER	Precision	Recall	IER	Precision	Recall
A	Named addressee identification	85.6%	30.3%	16.0%	80.5%	29.8%	21.4%
B	Named speaker identification	72.3%	53.7%	29.4%	63.1%	53.3%	39.2%
C	… *vs.* oracle	45.9%	100%	56.1%	31.1%	100%	71.6%
D	Acoustic speaker identification	45.3%	68.9%	57.8%	45.3%	68.9%	57.8%
E	… *vs.* oracle	33.4%	100%	68.6%	33.4%	100%	68.6%
F	Joint named & acoustic speaker identification	35.4%	71.4%	68.1%	35.8%	71.5%	68.1%
G	… *vs.* oracle	14.7%	100%	88.0%	8.8%	100%	94.4%

Table 3: Evaluation on test set for various graph configurations (to be compared with performance of matching oracles)

allowing to estimate the influence of errors made by the initial speech processing steps on the overall speaker identification. Automatic speech recognition obtains a word error rate of WER $= 18.3\%$, spoken name detection a slot error rate of SER $= 31.9\%$ and finally, normalized spoken name detection reaches SER $= 35.7\%$.

It is remarkable that switching from manual to automatic name detection never degrades precision (while it does recall). This can be partially explained by the high precision (79%) and lower recall (67%) of the final normalized spoken name detection.

However, we also notice that the joint named & acoustic speaker identification does not seem to be impacted by person name detection errors. This leads us to conclude that most addressee name mentions are correctly identified, whereas most errors must actually correspond to mentions of person name not taking part into the conversation. We plan to carefully analyze those errors in the future.

6. Conclusion

We proposed to address the problem of named speaker identification in TV broadcast using the *person instance graph* paradigm. Each audio document is modeled as an undirected graph where speech turns, spoken names and person identities are vertices connected to each other by edges weighted by the automatically estimated probability that they represent the same person.

This unified framework allows to straightforwardly combine multiple modules (*e.g.* speech transcription, named entity detection, addressee detection or even acoustic speaker identification) into a common optimization problem solved by integer linear programming. Experimental results on the REPERE corpus show that the proposed approach leads to a 10% error decrease over a state-of-the-art *i-vector* speaker identification system (from 45% down to 35%).

While, in this paper, we only rely on the audio stream to identify speakers, the REPERE corpus actually is a TV broadcast corpus where speakers are also introduced using screen overlays containing their name. We used this source of information in a previous work to achieve completely unsupervised speaker identification [6]. While a speaker seldom pronounces their own name, a name w written on screen usually (with probability $p_{tw} = 95\%$ on the REPERE corpus) corresponds to the current speech turn t.

In order to take this reliable source of information into account, after a first step of video optical character recognition, one can simply connect each written name vertex w to the cooccurring speech turn t with probability p_{tw}. Experiments on the very same corpus show that it leads to a significant 15% error decrease (down to IER $= 20.0\%$) over the joint named & acoustic speaker identification system. This results show how easily the proposed person instance graph can integrate additional sources of information when they become available.

However, the proposed approach does have some limitations related to scalability. The cardinality of the search space $\Delta_{\mathcal{V}}$ is $\mathcal{O}(2^{|\mathcal{V}|\times|\mathcal{V}|})$ and the number of constraints is $\mathcal{O}(|\mathcal{V}|^3)$. The resulting integer linear programming problem quickly becomes intractable for $|\mathcal{V}| \gg 100$. While *Dupuy et al.* can massively prune their ILP problem into simpler problems, their proposed simplification cannot be directly applied to our own ILP formulation [25]. Therefore, alternative graph-mining techniques should be investigated when the number of vertices increases. For instance, graph embedding approaches described in [26] for acoustic-only speaker recognition could be extended to the proposed multimodal case.

7. Acknowledgment

This work was realized in the framework of the QCompere project (2011-2014) funded by the Agence National de la Recherche (ANR).

8. References

[1] Leonardo Canseco, Lori Lamel, and Jean-Luc Gauvain, "A Comparative Study Using Manual and Automatic Transcriptions for Diarization," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2005, pp. 415–419.

[2] Sue E. Tranter, "Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 1013–1016.

[3] Julie Mauclair, Sylvain Meignier, and Yannick Estève, "Speaker Diarization : about whom the Speaker is Talking?," in *IEEE Odyssey*, 2006.

[4] Yannick Estève, Sylvain Meignier, Paul Deléglise, and Julie Mauclair, "Extracting true speaker identities from transcriptions," in *Proceedings of the International Speech Communication Association*, 2007, pp. 2601–2604.

[5] Vincent Jousse, Simon Petitrenaud, Sylvain Meignier, Yannick Estève, and Christine Jacquin, "Automatic Named Identification of Speakers using Diarization and ASR Systems," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Taïpei, Taïwan, April 2009.

[6] Hervé Bredin and Johann Poignant, "Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast," in *Proceedings of the 14th*

Annual Conference of the International Speech Communication Association, Lyon, France, August 2013.

[7] Elie El-Khoury, Antoine Laurent, Sylvain Meignier, and Simon Petitrenaud, "Combining Transcription-based and Acoustic-based Speaker Identifications for Broadcast News," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4377–4380.

[8] Lori Lamel, Sandrine Courcinous, Julien Despres, Jean-Luc Gauvain, Yvan Josse, Kevin Kilgour, Florian Kraft, Viet Bac Le, Hermann Ney, Markus Nussbaum-Thom, Ilya Oparin, Tim Schlippe, Ralf Schlüter, Tanja Schultz, Thiago Fraga da Silva, Sebastian Stüker, Martin Sundermeyer, Bianca Vieru, Ngoc Thang Vu, Alexander Waibel, and Cécile Woehrling, "Speech Recognition for Machine Translation in Quaero," in *International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, CA, December 2011.

[9] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "Partitioning and Transcription of Broadcast News Data," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Dec. 1998, pp. 1335–1338.

[10] Thomas Lavergne, Olivier Cappé, and François Yvon, "Practical Very Large Scale CRFs," in *Proceedings the 48^{th} Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 504–513.

[11] Marco Dinarelli and Sophie Rosset, "Models Cascade for Tree-Structured Named Entity Detection," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, November 2011, pp. 1269–1278, Asian Federation of Natural Language Processing.

[12] Scott S. Chen and Ponani 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.

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

[14] Simon J.D. Prince, "Computer Vision: Models Learning and Inference," in *Cambridge University Press, 2012, In press*.

[15] Mohammed Senoussaoui, Patrick Kenny, Niko Brümmer, Edward de Villiers, and Pierre Dumouchel, "Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011.

[16] Jenny Rose Finkel and Christopher D. Manning, "Enforcing Transitivity in Coreference Resolution," in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.

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

[18] Hervé Bredin, Anindya Roy, Viet-Bac Le, and Claude Barras, "Person Instance Graphs for Mono-, Cross- and Multi-Modal Person Recognition in Multimedia Data. Application to Speaker Identification in TV Broadcast," *International Journal of Multimedia Information Retrieval*, 2014.

[19] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The REPERE Corpus: a Multimodal Corpus for Person Recognition," in *International Conference on Language Resources and Evaluation*, 2012.

[20] James Bergstra and Yoshua Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machin Learning Research*, vol. 13, pp. 281–305, Mar. 2012.

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

[22] Jason Pelecanos and Sridha Sridharan, "Feature Warping for Robust Speaker Verification," in *Proceedings of Odyssey 2001 - The Speaker Recognition Workshop*, Crete, Greece, June 2001, pp. 213–218.

[23] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert, "The ETAPE Corpus for the Evaluation of Speech-based TV Content processing in the French language," in *International Conference on Language Resources, Evaluation and Corpora*, Turkey, 2012.

[24] Pierre-Michel Bousquet, Driss Matrouf, and Jean-François Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011.

[25] Grégor Dupuy, Sylvain Meignier, Paul Deléglise, and Yannick Estève, "Recent Improvements towards ILP-based Clustering for Broadcast News Speaker Diarization," in *Proceedings of Odyssey 2014 - The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014.

[26] Zahi N. Karam and William M. Campbell, "Graph Embedding for Speaker Recognition," in *Graph Embedding for Pattern Analysis*, Yun Fu and Yunqian Ma, Eds., pp. 229–260. Springer New York, 2013.