Int J Multimed Info Retr
DOI 10.1007/s13735-014-0055-y

REGULAR PAPER

Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identification in TV broadcast

Hervé Bredin · Anindya Roy · Viet-Bac Le · Claude Barras

Received: 16 July 2013 / Revised: 11 March 2014 / Accepted: 10 April 2014

Abstract This work introduces a unified framework for mono-, cross- and multi-modal person recognition in multimedia data. Dubbed person instance graph models the person recognition task as a graph mining problem: i.e., finding the best mapping between person instance vertices and identity vertices. Practically, we describe how the approach can be applied to speaker identification in TV broadcast. Then, a solution to the above-mentioned mapping problem is proposed. It relies on integer linear programming to model the problem of clustering person instances based on their identity. We provide an in-depth theoretical definition of the optimization problem. Moreover, we improve two fundamental aspects of our previous related work: the problem constraints and the optimized objective function. Finally, a thorough experimental evaluation of the proposed framework is performed on a publicly available benchmark database. Depending on the graph configuration (i.e., the choice of its vertices and edges), we show that multiple tasks can be addressed interchangeably (e.g., speaker diarization, supervised or unsupervised speaker identification), significantly outperforming state-of-the-art mono-modal approaches.

H. Bredin (✉)
LIMSI/CNRS, Rue John Von Neumann, 91400 Orsay, France
e-mail: bredin@limsi.fr
URL: http://herve.niderb.fr/

A. Roy
LIMSI/CNRS, Orsay, France
e-mail: roy@limsi.fr

V.-B. Le
Vocapia Research, Orsay, France
e-mail: levb@vocapia.com

C. Barras
LIMSI/CNRS, Université Paris-Sud, Orsay, France
e-mail: barras@limsi.fr

Keywords Person recognition · Speaker identification · Multimedia · Cross-modal processing · Graph mining · Integer linear programming

1 Introduction

Multi-modal and cross-modal information processing is an essential human faculty which we depend on quite often. For example, when visiting a new place, we match lexical information in terms of place names on a map to visual scenes of the actual places we are in, or to acoustic or speech information in the form of directions provided by passersby. This is part of a constant learning process where not only do we simultaneously analyze information from multiple modalities (multi-modal processing), but also use one modality to help understand another one (cross-modal processing).

To some extent, modern information processing systems already emulate this multi- or cross-modal processing capability. This includes systems which deal with automatic content-based segmentation, annotation, summarization and retrieval of multimodal content in the form of audio, video and text [10,25,31,32,39,42].

In this paper, we study automatic annotation and retrieval of multimedia data, specifically in the context of automatic person identification in TV broadcast. Multiple sources of information can be combined to achieve automatic person identification, including the visual stream (e.g., face recognition and overlaid name detection), the audio stream (e.g., speaker identification and speech transcription) and textual metadata (e.g., electronic program guide and cast list). Automatically recognized person identities can be very useful in many higher level multimedia analysis tasks, such as semantic indexing and retrieval, interaction analysis and video summarization.

Published online: 09 May 2014

Existing studies on person identification using cross-modal analysis typically involve person names extracted from the output of automatic speech recognition (ASR) [7, 13, 24, 28, 41], from overlaid text in videos [6, 36, 37], or from subtitles and transcripts aligned with automatically detected face tracks [2, 9].

Identifying speakers using pronounced names extracted from the ASR output was first proposed in [7]. Names were manually classified based on their lexical context to indicate whether they refer to the speaker, the addressee or someone else. Tranter et al. [41] automatically learn these patterns from n-gram sequences, while Mauclair et al. [28] used a semantic classification tree to match names with speaker turns. Estève et al. [13] and Jousse et al. [24] further developed and analyzed this approach. These approaches differ from our work in that they only rely on the audio stream to identify speakers (while our work seamlessly extends from audio-only to audio-visual processing). Hence, they are very dependent on the quality of the ASR output as they cannot rely on the visual stream to address this limitations. For instance, [24] reports that error rates increase from 17 up to 75 % when switching from manual to automatic speech transcription.

More recently, three cross-modal methods were proposed by Poignant et al. [37] to automatically propagate written names (obtained from overlaid text using video optical character recognition) to speaker clusters. These cross-modal unsupervised methods achieved better performance than a mono-modal supervised speaker identification solution. However, the performance of name propagation is very dependent on the quality of the initial speaker diarization step—while our proposed framework achieves (and improves both) speaker diarization and name propagation at the same time. Hence, [37] reports that error rates increase from 23 up to 33 % when switching from manual to automatic speaker diarization.

Apart from naming speakers from spoken or written names, another approach is to align TV series transcripts with face tracks, and use this alignment to train character models in a weakly supervised learning scenario [2, 9]. Contrary to our proposed framework that can achieve fully unsupervised speaker identification, these approaches do rely on manual (and potentially ambiguous) labels to train face models later used for supervised identification.

To our knowledge, none of the existing works proposes a unified framework for mono-, cross- or multi-modal person identification in multimedia data. In this context, the main contribution of this paper is the introduction of a generic structure called person instance graph. Depending on its configuration (i.e., the choice of its vertices and edges), it can be used to model and outperform both existing mono- or cross-modal approaches, as well as supervised or unsupervised person recognition algorithms. Most of all, it has the potential to seamlessly combine all these variants into a universal multi-modal one. Section 2 details the graphical structure of the proposed framework and describes how it can be setup in practice for speaker identification in TV broadcast. For reference purposes, notations introduced in this section and used in the rest of the paper are gathered in Table 1.

Section 3 details the proposed integer linear programming solution that is used to mine person identity information from person instance graphs. Although it is based on our previous work reported in [6], the current work brings several major contributions with respect to that work. First, theoretical aspects are much more detailed than it was in [6]. Then, a weighted extension of the objective function is introduced in Sect. 3.2. Finally, Sect. 3.4 describes how (previously strict) transitivity constraints can be relaxed to achieve better person identification results.

Section 4 provides a detailed description of the experimental protocol, including the REPERE benchmark database [19] and an in-depth definition of the evaluation metrics. Finally, in Sect. 5, a thorough experimental evaluation of the proposed framework is reported on this benchmark database. Multiple modalities are combined towards speaker identification—including speech turns extracted from the acoustic signal, spoken names obtained from speech transcription, and written names given by video optical character recognition. Two learning scenarios (unsupervised and supervised) and two applications (speaker identification and diarization) are studied.

Section 6 concludes this paper and highlights how the proposed framework could be extended to various applications of interest.

2 Building person instance graphs

A person instance graph is a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$ where $\mathcal{V}$ is the set of vertices, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges, and $p \in [0, 1]^{\mathcal{E}}$ associates a weight to every edge.

Vertices $\mathcal{V}$. Each vertex $v \in \mathcal{V}$ represents either a person (identity vertex) or an instantiation of a person (instance vertex). For example, the person instance graph describing the video sequence in Fig. 1 would contain two identity vertices (one for `Nicolas_SARKOZY` and one for `Barack_OBAMA`) and four instance vertices (a face instance, a speech turn instance and a written name instance of the former, and a spoken name instance of the latter).

As illustrated in Fig. 2, instance vertices are localized in time (with start and end times). An in-depth description of instance vertices is provided in Sect. 2.1.

On the other hand, identity vertices are meta-vertices representing one person each. They are described in Sect. 2.2.

Table 1 Notations

Speaker identification		
$\mathbb{I}$	Universal set of person identities	
$\texttt{ID}$	Maps each vertex to its true identity	
λ_i	Acoustic speaker model of person *i*	
Person instance graph		
$\mathcal{G}$	Graph	
$\mathcal{V}$	Set of vertices	$v \in \mathcal{V}$
$\mathcal{T}$	Set of speech turns	$t \in \mathcal{T}$
$\mathcal{W}$	Set of written names	$w \in \mathcal{W}$
$\mathcal{S}$	Set of spoken names	$s \in \mathcal{S}$
$T(t)$	Temporal support of speech turn *t*	
$\lvert T(t) \rvert$	Duration of speech turn *t*	
$\mathcal{I}_\mathcal{W}$	Set of written names identities $\mathcal{I}_\mathcal{W} = \{\texttt{ID}(w) \mid w \in \mathcal{W}\}$	$i_w \in \mathcal{I}_\mathcal{W}$
$\mathcal{I}_\mathcal{S}$	Set of spoken names identities $\mathcal{I}_\mathcal{S} = \{\texttt{ID}(s) \mid s \in \mathcal{S}\}$	$i_s \in \mathcal{I}_\mathcal{S}$
$\mathcal{I}^*$	Set of identities for which a speaker model λ is available	$i^* \in \mathcal{I}^*$
$\mathcal{E}$	Set of edges	$(v, v') \in \mathcal{E}$
$p_{vv'}$	Probability that v and v' are the same person $p_{vv'} = p(\texttt{ID}(v) = \texttt{ID}\,(v') \mid v, v')$	
Optimization		
$\Delta_\mathcal{V}$	Set of clustering functions	$\delta \in \Delta_\mathcal{V}$
$\mathcal{L}$	Objective function	
α, β	Hyper-parameters	
Evaluation		
r, h	Reference and hypothesis	
DER	Diarization error rate	
IER	Identification error rate	

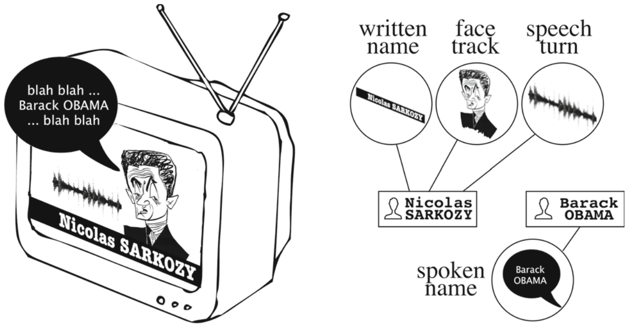

Fig. 1 A TV sequence and the corresponding person instance graph containing two identity vertices and four instance vertices (a speech turn, a face track, a written name and a spoken name)

Edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Every edge $(v, v') \in \mathcal{E}$ connects two vertices in the graph. A person instance graph does not contain any self-loop: $\forall v \in \mathcal{V}$, $(v, v) \notin \mathcal{E}$. Moreover, it may be incomplete: $\exists (v, v') \in \mathcal{V} \times \mathcal{V}$ s.t. $v \neq v'$ and $(v, v') \notin \mathcal{E}$.

Weights $p \in [0, 1]^\mathcal{E}$. Every edge $(v, v') \in \mathcal{E}$ is weighted by the probability $p_{vv'}$ that vertices v and v' correspond to the same identity. In other words, $p_{vv'} = p(\texttt{ID}(v) = \texttt{ID}(v'))$ where the function $\texttt{ID}\colon \mathcal{V} \rightarrow \mathbb{I}$ maps each vertex v to its identity among the universal set of person identities $\mathbb{I}$. Note how weights $p_{vv'}$ are therefore symmetrical (i.e., $p_{vv'} = p_{v'v}$), thus making the graph $\mathcal{G}$ undirected. Section 2.3 describes how these weights are obtained in practice.

2.1 Instance vertices

While four different types of instance vertices can be added to a person instance graph, we only integrate three of them (speech turns in Sect. 2.1.1, written names in Sect. 2.1.2 and spoken names in Sect. 2.1.3).

The proposed framework can be extended at no extra cost to the face modality and we do plan to try and integrate face tracks in the future. The only reason why we did not integrate the face modality is because we did not have access to face detection, clustering and recognition technologies at the time.

2.1.1 Speech turns $\mathcal{T}$

The first set of instance vertices added to the graph are speech turns $t \in \mathcal{T}$. They are automatically extracted from the audio stream of the TV broadcast following the initial steps of LIMSI multi-stage speaker diarization system described in [1].

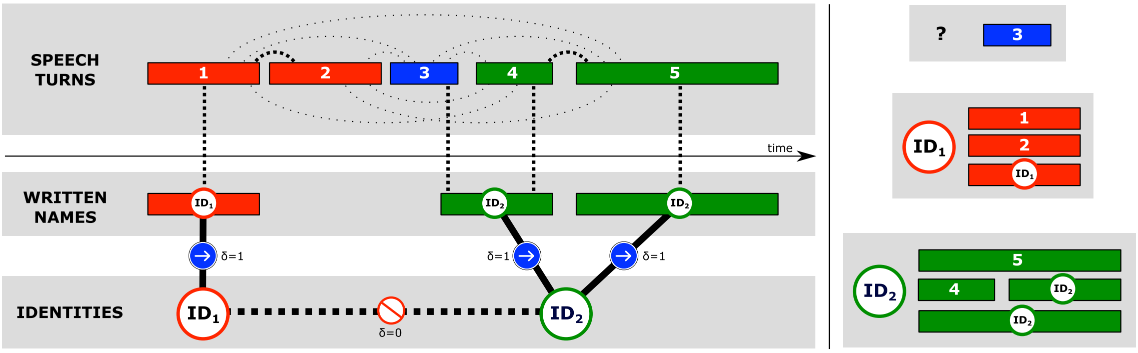

Fig. 2 A person instance graph (*left*) and the expected output of clustering (*right*)

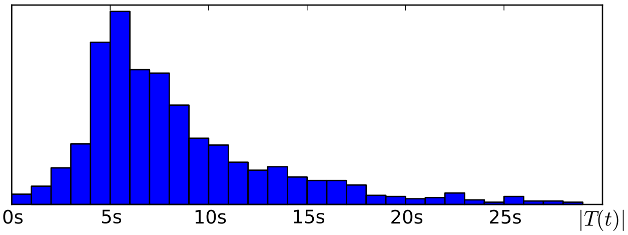

Fig. 3 Distribution of speech turn durations on the test set

First, 12 Mel frequency cepstral coefficients (MFCC) and log-energy are extracted from the audio signal every 10 ms using a 30 ms Hamming window on the 0–8 kHz bandwidth [22]. Then, speech activity detection is performed with a Viterbi decoding using one 64-Gaussians mixture model (GMM) per class: speech, noisy speech, speech over music, pure music, and silence or noise. These GMMs were trained on approximately one hour of matching data selected from radio broadcast news [16]. This approach reaches nearly perfect (96 % accuracy) speech vs. non-speech classification on the corpus used in our experiments, i.e., TV broadcast where speech is usually prepared (as opposed to spontaneous) and recorded in a controlled environment.

Speech segments are further segmented into smaller homogeneous segments by detecting speaker changes [8]. This is achieved by looking for maxima of the local Gaussian divergence $G(w_{\mathrm{L}}, w_{\mathrm{R}})$ between two adjacent sliding windows w_{L} (left) and w_{R} (right) of 5 s. The Gaussian divergence is defined as follows:

$$G(w_{\mathrm{L}}, w_{\mathrm{R}}) = (\mu_{\mathrm{R}} - \mu_{\mathrm{L}})^T \cdot \Sigma_{\mathrm{L}}^{-1/2} \cdot \Sigma_{\mathrm{R}}^{-1/2} \cdot (\mu_{\mathrm{R}} - \mu_{\mathrm{L}}) \quad (1)$$

where the MFCC coefficients extracted from each window are modeled as Gaussian with diagonal covariance matrix $\mathcal{N}(\mu, \Sigma)$. Each maximum is compared to a threshold θ to decide whether the corresponding timestamp is a speaker change. θ is optimized on a development set so that the resulting speech turns t are almost pure (i.e., contains speech from one speaker only).

Figure 3 shows the distribution of the duration $|T(t)|$ of the speech turns t on the test set described in Sect. 4.1: 90 % are shorter than 10 s. On average, speech turns t are pure at 96.4 % (a proper definition of purity will be given in Sect. 4.2).

2.1.2 Written names $\mathcal{W}$

As shown in Fig. 4, reporters in TV news (or guests in talk shows) are often introduced visually by a title block containing their names. Similarly to what we did in our previous work [6], we automatically extract every occurrence of these names and add them to the graph as written name instance vertices $w \in \mathcal{W} \subset \mathcal{V}$. Figure 2 underlines the fact that they are localized in time, and that the name of the same person can appear several times over the duration of a TV show.

In practice, we rely on the video optical character recognition (OCR) system proposed by Poignant et al. in [36] to automatically extract this information. First, overlaid text boxes are detected using a coarse-to-fine approach with temporal tracking. Then, the open-source Tesseract toolkit [40] provides one transcription every tenth frame. These transcriptions are finally merged to produce a unique transcription for each text box. Poignant et al. [36] reports precision of 97 % on a TV broadcast corpus similar to the one used in our experiments (i.e., overlaid text with clean flat background and, by design, easily readable fonts).

An additional filtering step is needed because, depending on the TV channel, not every detected box is used to introduce a person. In Fig. 4, for instance, a few text boxes are used to provide news flash (left) or the name of the TV show (right). However, TV channels tend to always use the same visual layout. Therefore, to remove those unwanted text boxes, the training set described in Sect. 4.1 is used in combination with a large list of person names from Wikipedia to automatically learn the spatial positions of text boxes the most likely to contain introductory names. Text boxes located at other spatial positions are filtered out.

One name: $p_{tw} = 0.956$ Two names: $p_{tw} = 0.996$

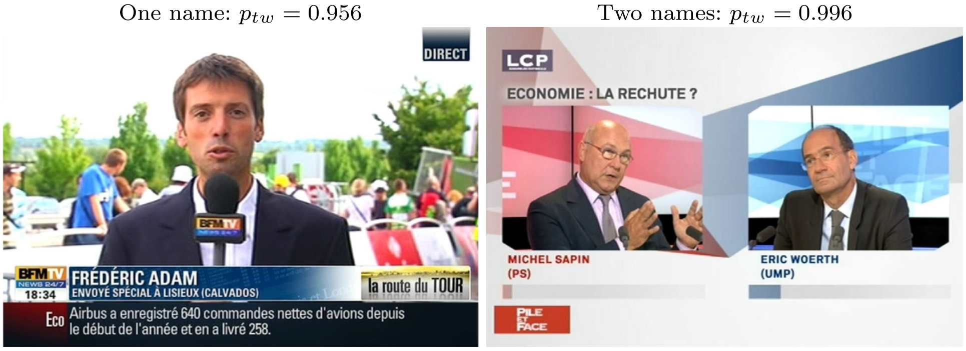
DIRECT
BFM TV
FRÉDÉRIC ADAM
ENVOYÉ SPÉCIAL À LISIEUX (CALVADOS)
18:34
la route du TOUR
Eco
Airbus a enregistré 640 commandes nettes d'avions depuis le début de l'année et en a livré 258.
LCP
ECONOMIE : LA RECHUTE ?
MICHEL SAPIN (PS)
ERIC WOERTH (UMP)
PILE & FACE

Fig. 4 Cross-modal probabilities p_{tw} depend on the number of simultaneous written names w

2.1.3 Spoken names $\mathcal{S}$

Not only are person names displayed on screen, they are also frequently pronounced, either by the anchor to introduce a guest or an interviewee, or by a reporter as part of a TV news report. Similarly, they can be added to the graph as spoken name instance vertices $s \in \mathcal{S} \subset \mathcal{V}$.

In practice, the automatic extraction of spoken names is a two-step process. First, automatic speech recognition (or speech-to-text) provides the textual transcription of the spoken words. Then, named-entity recognition aims at extracting person names from the resulting textual document.

However, though our automatic speech-to-text system [17] performs relatively well (Word error rate = 16.9 %) on the REPERE corpus, we were not able to obtain reasonably good person name detection results using the named-entity detection system described in [11]: Slot error rate = 60 %. Therefore, in the rest of the paper, all experiments involving spoken name instances $\mathcal{S}$ are based on manual speech transcription and manual person name detection.

2.2 Identity vertices $\mathcal{I}$

While there can be multiple instance vertices of the same person in a graph (e.g., one for every speech turn, one written name instance for every time it appears on screen, etc.), there cannot be more than one identity vertex $i \in \mathbb{I}$ per person. To ensure unicity, a unique standardized identifier is given to each person, using the following naming convention: `First-Name_LASTNAME`. For instance, the identifiers for the first and third authors of this paper are `Herve_BREDIN` and `Viet-Bac_LE`.

Identity vertices can be obtained in three different manners: $\mathcal{I} = \mathcal{I}_\mathcal{W} \cup \mathcal{I}_\mathcal{S} \cup \mathcal{I}^*$. First, $\mathcal{I}_\mathcal{W} = \{\texttt{ID}(w) \mid w \in W\}$ is the set of identity vertices provided by the video OCR system described in Sect. 2.1.2. Similarly, $\mathcal{I}_\mathcal{S} = \{\texttt{ID}(s) \mid s \in S\}$ is derived from the output of the spoken name detection described in Sect. 2.1.3. Simple heuristics are used to derive the standardized identifiers from the original textual output of both approaches. Finally, an additional set of identity vertices $\mathcal{I}^*$ is provided by the acoustic-based speaker identification system described in Sect. 2.3.2. $\mathcal{I}^*$ contains one identity vertex per speaker for which a voice model is available.

2.3 Weighted edges $(\mathcal{E}, p)$

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

$$\begin{aligned} p : \mathcal{E} &\rightarrow [0, 1] \\ (v, v') &\mapsto p_{vv'} = p(\texttt{ID}(v) = \texttt{ID}(v') \mid v, v') \end{aligned} \qquad (2)$$

2.3.1 Speech turn similarity p_{tt}

As illustrated in the upper timeline of Fig. 2, the first set of edges is $\mathcal{T} \times \mathcal{T} \subset \mathcal{E}$. Every pair of speech turns (t, t') is connected by an edge—hence, making the speech turn subgraph $\mathcal{G}_\mathcal{T} = (\mathcal{T}, \mathcal{T} \times \mathcal{T}, p)$ complete.

The weights $p_{tt'}$ are then estimated as follows. First, each speech turn $t \in \mathcal{T}$ is modeled with one Gaussian with full covariance matrix Σ_t trained on the $D = 12$-dimensional MFCC and energy. Then, the similarity $d_{tt'}$ between two speech turns t and t' is defined as the Bayesian Information Criterion $\Delta\text{BIC}(t, t')$ [8]:

$$\begin{aligned} d_{tt'} = {} & (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\ & - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\ & - \frac{1}{2} \cdot \lambda \cdot \left(D + \frac{1}{2} D \, (D + 1) \right) \log \, (n_t + n_{t'}) \end{aligned} \qquad (3)$$

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 $p_{tt'}$:

$$\begin{aligned} p_{tt'} &= p(\texttt{ID}(t) = \texttt{ID}(t') \mid d_{tt'}) \\ &= \frac{1}{1 + \dfrac{\pi_{\neq}}{\pi_{=}} \cdot \dfrac{p(d_{tt'} \mid \texttt{ID}(t) \neq \texttt{ID}(t'))}{p(d_{tt'} \mid \texttt{ID}(t) = \texttt{ID}(t'))}} \end{aligned} \qquad (4)$$

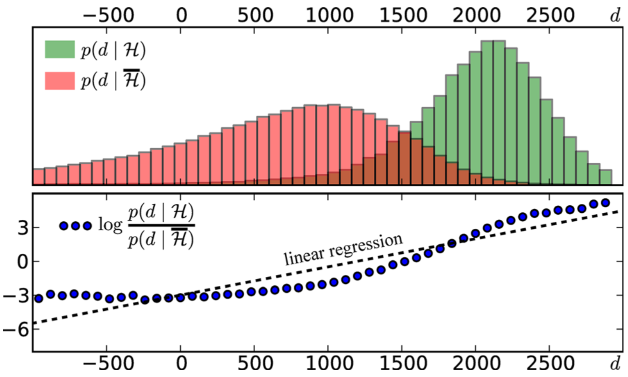

Fig. 5 Estimation of the log-likelihood ratio on the training set. *Top* likelihood under hypothesis $\mathcal{H} \equiv \texttt{ID}(t) = \texttt{ID}(t')$ (*rightmost distribution*, *green*) and $\overline{\mathcal{H}} \equiv \texttt{ID}(t) \neq \texttt{ID}(t')$ (*leftmost distribution*, *red*). *Bottom* estimated log-likelihood ratio (*bullet*) and linear regression

where the prior probabilities are assumed equal ($\pi_= = \pi_{\neq}$) and the likelihood ratio is estimated on the training set described in Sect. 4.1. This estimation process is illustrated in Fig. 5 and is achieved using linear regression by minimization of the sum of squared error in the logarithmic space.

2.3.2 Similarity to speaker model p_{ti}

Though those edges are not shown in Fig. 2, one can directly connect speech turn instance vertices $t \in \mathcal{T}$ to identity vertices $i \in \mathcal{I}^*$, by means of an acoustic-based speaker identification system.

In this work, we rely on a standard Gaussian mixture model (GMM) system based on adapted universal background model (UBM) [26]. It has proved to be very successful for text-independent speaker recognition, since it allows for robust estimation of speaker models λ_i even with a limited amount of enrollment data [38].

Acoustic features x are extracted from the speech signal on the 0–8 kHz bandwidth every 10 ms using a 30 ms Hamming window. Feature vectors x consist of 15 PLP-like cepstrum coefficients [22] plus 15 delta coefficients and delta energy, for a total of 31 features. Feature warping normalization is performed using a sliding window of 3 seconds to reduce the effect of the acoustic environment [33].

First, a gender-independent multilingual UBM with a mixture of 256 diagonal Gaussians was trained on a multilingual broadcast corpus [38]. Then, three annotated data sources were used to train one model λ_i per speaker $i \in \mathcal{I}^*$: the REPERE training [19], the ETAPE training and development data [20] and additional French politicians data extracted from French radio broadcast. Only speakers with more than 30 s training data were kept, resulting in $|\mathcal{I}^*| = 611$ speaker identity vertices. For each speaker $i \in \mathcal{I}^*$, a speaker-specific GMM λ_i is trained by MAP adaptation of the means of the UBM [18].

Given a speech turn t and a target identity $i \in \mathcal{I}^*$, the speaker identification score d_{ti} is defined as the following log-likelihood ratio:

$$d_{ti} = \frac{1}{n_t}\left[\log \prod_{x \in X_t} f\left(x|\lambda_i\right) - \log \prod_{x \in X_t} f\left(x|\lambda_\Omega\right)\right] \tag{5}$$

where X_t is the set of n_t feature vectors extracted from speech turn t, $f\,(x|\lambda_i)$ is the likelihood of feature vector x for speaker model λ_i and $f\,(x|\lambda_\Omega)$ its likelihood for the UBM.

Identification scores d_{ti} are then calibrated into probabilities p_{ti} following the open-set speaker identification paradigm:

$$\begin{aligned} p_{ti} &= p(\texttt{ID}(t) = i \mid d_{ti}) \\ &= \frac{\pi_i \cdot \dfrac{p(d_{ti} \mid \texttt{ID}(t) = i)}{p(d_{ti} \mid \texttt{ID}(t) \neq i)}}{\pi_{\textcircled{?}} + \sum_{i' \in \mathcal{I}^*} \pi_{i'} \cdot \dfrac{p(d_{ti'} \mid \texttt{ID}(t) = i')}{p(d_{ti'} \mid \texttt{ID}(t) \neq i')}} \end{aligned} \tag{6}$$

where $\pi_{\textcircled{?}}$ is the prior probability that speaker is unknown (i.e., $i \notin \mathcal{I}^*$) and $p(d_{ti} \mid \texttt{ID}(t) \neq i)$ is an approximation of $p(d_{ti} \mid \texttt{ID}(t) = \textcircled{?}\,)$. Prior probabilities π_i are assumed to be equal (i.e., $\pi_i = (1 - \pi_{\textcircled{?}}\,)/|\mathcal{I}^*|\,$). In practice, likelihood ratios are estimated like in Sect. 2.3.1.

2.3.3 Written names propagation p_{tw}

As already discussed in Sect. 2.1.2, written names $w \in \mathcal{W}$ are usually overlaid on screen to introduce the speaker of the current speech turn $t \in \mathcal{T}$. In other words, a cross-modal edge (t, w) should be added to the graph as soon as $t \sqcap w \neq \varnothing$ where the $\sqcap$ operator returns the temporal intersection. This is illustrated in Fig. 2 with thick dotted vertical edges between speech turns and written names.

Though the corresponding probability p_{tw} is very high, it is strictly smaller than 1 for various reasons. In TV news reports, for instance, the speech of a foreign speaker is usually replaced by the voice of the translator. In talk shows, the speech of the current speaker (whose name is overlaid) can be interspersed with interruption from another guest.

In practice, p_{tw} is estimated by a simple frequency count on the training set. As illustrated in the header of Fig. 4, its value depends on the number of co-occurring written names: $p_{tw} \approx 0.95$ in case there is exactly one written name, and $p_{tw} \approx 0.99$ when there are two names. In this latter case, the identity unicity constraints defined later in Sect. 3.1 will make sure at most one written name is associated to the speech turn.

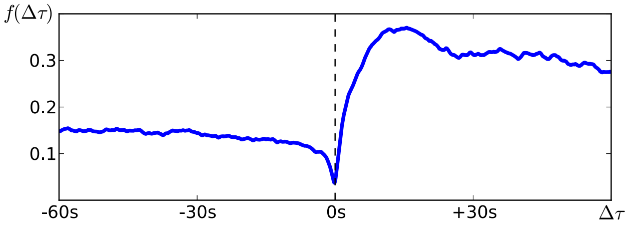

Fig. 6 Probability $f(\Delta\tau)$ that a spoken name s at time τ_s is the name of a potential speaker at time $\tau_s + \Delta\tau$

2.3.4 Spoken names propagation p_{ts}

While written names $w \in \mathcal{W}$ are often used to introduce the current speaker, 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.

Given a spoken name s pronounced a time τ_s, Fig. 6 shows the probability $f(\Delta\tau)$ that a potential speaker at time $\tau_s+\Delta\tau$ is the person whose name was pronounced at time τ_s. It was estimated using the training set described in Sect. 4.1.

The maximum at $\Delta\tau = 15$ s corresponds to the fact that a speaker is typically named just before (s)he starts speaking. $f(\Delta\tau = 0) = 0.04$ indicates that speakers rarely name themselves in TV broadcast. We also observe that values of f are lower in general for negative values of $\Delta\tau$ than positive ones. This shows that speakers are named less frequently after they spoke (e.g., for thanking them) than before they speak (e.g., for introducing them).

We rely on function f to add edges between each pair of spoken name $s \in \mathcal{S}$ and speech turn $t \in \mathcal{T}$ as long as they are <60 s apart:

$$\begin{aligned} p_{ts} &= p(\mathtt{ID}(t) = \mathtt{ID}(s) \mid t, s) \\ &= \frac{1}{|T(t)|} \int\limits_{\tau \in T(t)} f(\tau - \tau_s)\, \mathrm{d}\tau \end{aligned} \tag{7}$$

where $T(t)$ is the temporal support of speech turn t.

2.3.5 Hard edges ($p_{wi_w} = p_{si_s} = 1$)

Finally, every written name w (resp. spoken name s) is connected with probability 1 to the corresponding identity vertex $i_w = \mathtt{ID}(w) \in \mathcal{I}_\mathcal{W}$ (resp. $i_s = \mathtt{ID}(s) \in \mathcal{I}_\mathcal{S}$) introduced in Sect. 2.2.

These edges are denoted $w \Leftrightarrow i_w$ and $s \Leftrightarrow i_s$ in the rest of this article (as opposed to regular edges $t \leftrightarrow t'$ or $t \leftrightarrow i$, for instance) to highlight the fact that they are weighted with probability 1.

3 Mining person instance graphs

Figure 2 contains a simple person instance graph involving three persons (whose respective instance and identity vertices are colored in red, green and blue). It contains five speech turn vertices $\mathcal{T} = \{t_1, t_2, t_3, t_4, t_5\}$, three written name vertices $w \in \mathcal{W}$ providing only two identity vertices $\mathcal{I}_\mathcal{W} = \{\mathtt{ID}_1, \mathtt{ID}_2\}$. Mining this graph for speaker identification consists in automatically assigning the correct identity vertex to each speech turn: $t_1 \rightarrow \mathtt{ID}_1$, $t_2 \rightarrow \mathtt{ID}_1$, $t_3 \rightarrow \textcircled{?}$, $t_4 \rightarrow \mathtt{ID}_2$ and $t_5 \rightarrow \mathtt{ID}_2$. Notice how the graph does not contain the actual identity of speech turn t_3: speech turn t_3 therefore remains anonymous ($\textcircled{?}$).

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 $\mathtt{ID}$ defined as follows:

$$\begin{aligned} &\mathtt{ID} : \mathcal{V} \rightarrow \mathcal{I} \cup \{\textcircled{?}\} \\ &v \mapsto \begin{cases} v & \text{if } v \in \mathcal{I} \text{(i.e., } v \text{ is an identity vertex);} \\ i & \text{if } \exists\, i \in \mathcal{I} \text{ s.t. } v \text{ is an instance of } i; \\ \textcircled{?} & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

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). The expected output of such clustering is illustrated in the right part of Fig. 2.

Clustering has been addressed in numerous scientific fields in the past: from graph mining and community detection [30] to natural language processing and co-reference resolution [14]. Classical clustering algorithms include K-means and hierarchical (agglomerative or divisive) clustering [23]. However, they suffer from three main limitations.

First, though heuristics were proposed to estimate it automatically [29,34], K-means and its variants usually rely on the assumption that the number of clusters K is known a priori. Moreover, most approaches do not guarantee global optimality. In hierarchical agglomerative clustering approaches, two clusters are merged because they are (locally) close to each other, independently of how similar (or dissimilar) they are to other clusters. Finally, state-of-the-art approaches usually rely on complete affinity matrices and cannot deal with situations where the affinity matrices are incomplete (e.g., missing edges in person instance graphs).

Inspired by [14], we proposed in [6] to model clustering as an Integer Linear Programming (ILP) problem, addressing all three shortcomings. Sections 3.1–3.3 provide a more detailed description of this previous work. Two major improvements are proposed in Sect. 3.2 (i.e., weighted objective function) and Sect. 3.4 (i.e., transitivity constraints relaxation).

ILP has been used before by Dupuy et al. [12] in the framework of speaker diarization. However, our approach differs from [12] both in the actual formulation of the ILP problem (they rely on the assumption that a few speech turns are cluster centroids, we do not), and in the fact that their approach is purely mono-modal and is limited to speaker diarization (ours is multi-modal and can also be used for speaker identification).

3.1 Clustering function

Any output of a valid clustering algorithm can be described by a clustering function $\boldsymbol{\delta}$, as follows:

$$\begin{aligned} \boldsymbol{\delta} \colon \mathcal{V} \times \mathcal{V} &\rightarrow \{0, 1\} \\ (v, v') &\mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

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:

$$\Delta_{\mathcal{V}} = \begin{cases} \boldsymbol{\delta} \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} \quad \text{s.t.} \ \forall \left(v, v', v''\right) \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 \\ \text{(b) } \delta_{vv'} = \delta_{v'v} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 \end{cases} \tag{10}$$

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 Eq. (11):

$$\begin{aligned} \forall \left(v, v', v''\right) \in \mathcal{V}^3, \quad & \delta_{vv'} + \delta_{v'v''} - \delta_{vv''} \leq 1 \\ & \delta_{v'v''} + \delta_{v''v} - \delta_{v'v} \leq 1 \\ & \delta_{v''v} + \delta_{vv'} - \delta_{v''v'} \leq 1 \end{aligned} \tag{11}$$

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

$$\forall v \in \mathcal{V}, \quad \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1 \tag{12}$$

In particular, when combined with reflexivity constraints ($\delta_{ii} = 1$), Eq. (12) implies that two identity vertices cannot end up in the same cluster:

$$\forall \left(i, i'\right) \in \mathcal{I}^2, \quad i \neq i' \implies \delta_{ii'} = 0 \tag{13}$$

These implicit constraints are marked as red "forbidden" traffic signs in Fig. 2. Finally, we explicitly constrain written names w (resp. spoken names s) to be in the same cluster as their corresponding identity vertex i_w (resp. i_s):

$$\begin{aligned} \forall w \in \mathcal{W}, \quad \delta_{wi_w} = 1 \\ \forall s \in \mathcal{S}, \quad \delta_{si_s} = 1 \end{aligned} \tag{14}$$

These constraints are marked as blue traffic signs →⃝ in Fig. 2.

3.2 Objective 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).

In other words, 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 (12) and (14) that maximizes the intra-cluster similarity while minimizing the inter-cluster similarity:

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

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:

$$\begin{aligned} \mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1} \Bigg[\alpha \cdot & \overbrace{\sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster} \\ \text{similarity}}} \\ & + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\substack{\text{inter-cluster} \\ \text{dissimilarity}}} \Bigg] \end{aligned} \tag{16}$$

By design, a person instance graph usually contains many more $t \leftrightarrow t'$ edges (between any two speech turns) than it does $t \leftrightarrow w$ edges (only between co-occurring speech turn and written name). Therefore, Eq. (16) 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:

$$\begin{aligned} \mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) = \sum_{\substack{x \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\} \\ y \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}\left(\boldsymbol{\delta}, \mathcal{E} \cap (x \times y), p\right) \\ \boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \Delta_{\mathcal{V}}}{\text{argmax}} \ \mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) \end{aligned} \tag{17}$$

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 [21].

The resulting optimal solution $\boldsymbol{\delta}^*$ can then be used to associate a unique identity to each instance vertex:

$$\begin{aligned} \mathtt{ID}_{\boldsymbol{\delta}^*} &: \mathcal{V} \rightarrow \mathcal{I} \cup \{\textcircled{?}\} \\ v &\mapsto \begin{cases} i & \text{if } \exists\, i \in \mathcal{I} \text{ s.t. } \delta^*_{vi} = 1, \\ \textcircled{?} & \text{otherwise.} \end{cases} \end{aligned} \tag{18}$$

Note that constraints (12) 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: $\mathtt{ID}_{\boldsymbol{\delta}^*}(v) = \textcircled{?}$.

3.4 Transitivity constraints relaxation

As far as person identification is concerned, Eq. (18) 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 Eq. (10.c) can be relaxed in the following way

$$\begin{aligned} &\forall \left(v, v', i\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^2 \times \mathcal{I}, \\ &\delta_{vi} = 1 \wedge \delta_{v'i} = 1 \;\not\Rightarrow\; \delta_{vv'} = 1 \end{aligned} \tag{19}$$

Formally, this is achieved by replacing the strict transitivity constraints defined in Eq. (11) by the following loose transitivity constraints (20) and (21):

$$\forall \left(v, v', v''\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^3, \quad \begin{array}{l} \delta_{vv'} + \delta_{v'v''} - \delta_{vv''} \leq 1 \\ \delta_{v'v''} + \delta_{v''v} - \delta_{v'v} \leq 1 \\ \delta_{v''v} + \delta_{vv'} - \delta_{v''v'} \leq 1 \end{array} \tag{20}$$

$$\forall \left(v, v', i\right) \in \{\mathcal{V} \setminus \mathcal{I}\}^2 \times \mathcal{I}, \quad \begin{array}{l} \delta_{vv'} + \delta_{vi} - \delta_{v'i} \leq 1 \\ \delta_{vv'} + \delta_{v'i} - \delta_{vi} \leq 1 \end{array} \tag{21}$$

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 (as shown in Table 10 of Sect. 5 devoted to experimental results).

3.5 Applications

Depending on the targeted application, a person instance graph may contain only a subset of vertices and edges. Table 2 provides a few possible configurations.

Speaker diarization (configuration 5A in Table 2), for instance, is the task of partitioning and labeling an audio stream into homogeneous speech segments according to the identity of the speaker. One does not care about the actual identity of the speaker. This is actually a speech turn clustering problem. The corresponding graph only contains speech turn vertices $t \in \mathcal{T}$ and speech turn to speech turn edges $t \leftrightarrow t'$. It does not contain any identity vertices.

Standard supervised speaker identification can also be modeled as a person instance graph mining problem. Thus, configuration 7A simply connects every speech turn $t \in \mathcal{T}$ to a set of identity vertices $\mathcal{I}^*$ for which acoustic models were obtained using a manually annotated training set.

Configurations 8A and 8C allow cross-modal speaker identification. Basically, these configurations deal with unsupervised speaker identification in the sense that no acoustic model of the speakers is available a priori. One must uncover the identity of the speaker from other modalities: either names written on the screen (configuration 8A) or pronounced, or a combination of both written and spoken names (configuration 8C).

Table 2 Various person instance graph configurations and corresponding applications

#	Targeted application	Instance vertices			Identity vertices			Edges			
								$t \leftrightarrow t'$	$t \leftrightarrow i^*$	$t \leftrightarrow w \Leftrightarrow i_w$	$t \leftrightarrow s \Leftrightarrow i_s$
5A	Speaker diarization	$\mathcal{T}$	–	–	–	–	–	$t \leftrightarrow t'$	–	–	–
7A	Mono-modal speaker identification	$\mathcal{T}$	–	–	$\mathcal{I}^*$	–	–	–	$t \leftrightarrow i^*$	–	–
7B		$\mathcal{T}$	–	–	$\mathcal{I}^*$	–	–	$t \leftrightarrow t'$	$t \leftrightarrow i^*$	–	–
8A	Cross-modal speaker identification	$\mathcal{T}$	$\mathcal{W}$	–	–	$\mathcal{I}_{\mathcal{W}}$	–	$t \leftrightarrow t'$	–	$t \leftrightarrow w \Leftrightarrow i_w$	–
8C		$\mathcal{T}$	$\mathcal{W}$	$\mathcal{S}$	–	$\mathcal{I}_{\mathcal{W}}$	$\mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t'$	–	$t \leftrightarrow w \Leftrightarrow i_w$	$t \leftrightarrow s \Leftrightarrow i_s$
9	Multi-modal speaker identification	$\mathcal{T}$	$\mathcal{W}$	–	$\mathcal{I}^*$	$\mathcal{I}_{\mathcal{W}}$	–	$t \leftrightarrow t'$	$t \leftrightarrow i^*$	$t \leftrightarrow w \Leftrightarrow i_w$	–
		$\mathcal{T}$	$\mathcal{W}$	$\mathcal{S}$	$\mathcal{I}^*$	$\mathcal{I}_{\mathcal{W}}$	$\mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t'$	$t \leftrightarrow i^*$	$t \leftrightarrow w \Leftrightarrow i_w$	$t \leftrightarrow s \Leftrightarrow i_s$

Possible sets of vertices include speech turns $\mathcal{T}$, written names $\mathcal{W}$, spoken names $\mathcal{S}$ and identity vertices $\mathcal{I}^*$, $\mathcal{I}_{\mathcal{W}}$ or $\mathcal{I}_{\mathcal{S}}$. $x \leftrightarrow y$ stands for $\mathcal{X} \times \mathcal{Y}$ edges, and $x \Leftrightarrow y$ for hard constraints ($\delta_{xy} = 1$). Column # refers to table and experiment identifiers used in Sect. 5

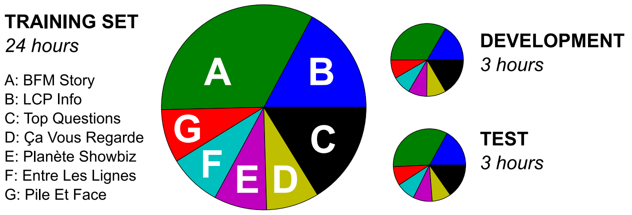

Fig. 7 Training, development and test sets each contain seven different types of shows (*A*–*G*)

4 Experimental protocol

4.1 Corpora

Figure 7 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 188 videos (30 h) recorded from 7 different shows broadcast by the French TV channels BFM TV and LCP. The audio stream is manually annotated with labeled speech turns ("who speaks when?"). In other words, a reference function r is available for each video:

$$\begin{aligned} r : T &\to \mathcal{P}(\mathbb{I} \cup \{\textcircled{?}\}) \\ \tau &\mapsto \{i_1, \ldots, i_{n_\tau}\} \end{aligned} \tag{22}$$

where T is the temporal support of the video, $\mathbb{I}$ is the universal set of person identities introduced in Sect. 2, $\mathcal{P}(\mathbb{A}) = \{A \mid A \subseteq \mathbb{A}\}$ is the set of all subsets of $\mathbb{A}$, and n_τ is the number of simultaneous speakers at time τ.

In practice, $n_\tau \in \{0, 1, 2\}$. $n_\tau = 0$ is used for non-speech regions (i.e., $r(\tau) = \varnothing$). $n_\tau = 2$ corresponds to overlapping speech regions where two persons speak simultaneously. In some rare cases (e.g., during street interviews), $\textcircled{?} \in r(\tau)$ indicates that the corresponding speaker could not be identified and therefore remains anonymous. Table 3 aggregates the duration of these special cases for the test set.

Manual speech transcription and person names entity detection are also provided: they were already mentioned in Sect. 2.1.3 for the description of spoken name vertices $s \in \mathcal{S}$.

The video stream is also manually annotated, but not as extensively as its audio counterpart. Only one frame every 10 seconds is annotated with manual detection and transcription of overlaid texts, and manual detection and identification of faces. Note that these visual annotations are not used in this work (except in Sect. 2.1.2 to automatically learn the usual spatial positions of title blocks).

Table 3 Test set statistics

Total	T	02:56:55	–
Non-speech	$\|r(\tau)\| = 0$	00:07:45	(4.4 %)
Overlapping speech	$\|r(\tau)\| = 2$	00:06:40	(3.8 %)
Anonymous speaker	$\textcircled{?} \in r(\tau)$	00:02:00	(1.1 %)

Table 4 Identification error ξ_r^h.

$r(\tau)$	$h(\tau)$	$\xi_r^h(\tau)$	Error type
$\varnothing$	$\varnothing$	0	Correct (no error)
$\{i\}$	$\{i\}$	0	
$\{\textcircled{?}\}$	$\{\textcircled{?}\}$	0	
$\varnothing$	$\{i\}$	1	False alarm
$\{i\}$	$\varnothing$	1	Missed detection
$\{i, i'\}$	$\{i\}$	1	
$\{i\}$	$\{i'\}$	1	Confusion
$\{i\}$	$\{\textcircled{?}\}$	1	
$\{i, i'\}$	$\{i''\}$	2	Confusion and missed detection

4.2 Evaluation metrics

For evaluation purposes, the manual reference r is compared to the hypothesis h obtained automatically as follows:

$$\begin{aligned} h : T &\to \mathcal{P}(\mathcal{I} \cup \{\textcircled{?}\}) \\ \tau &\mapsto \{\mathtt{ID}_{\delta^*}(t) \mid t \in \mathcal{T}, \tau \sqsubset t\} \end{aligned} \tag{23}$$

where optimal clustering function $\boldsymbol{\delta^*}$ is given by Eq. (17), $\mathtt{ID}_{\delta^*}$ is defined in Eq. (18) and $\tau \sqsubset t$ means that speech turn t overlaps time τ. Unless otherwise stated, tables report values aggregated over the 28 videos of the test set (for a total duration of 3 h).

Identification error rate (IER) The identification error rate (IER) is defined as the proportion (in duration) of the reference r incorrectly identified by the hypothesis h:

$$\mathrm{IER}(r, h) = \frac{\int_{\tau \in T} \xi_r^h(\tau)\, \mathrm{d}\tau}{\int_{\tau \in T} |r(\tau)|\, \mathrm{d}\tau} \tag{24}$$

where $\xi_r^h(\tau)$ returns the number of errors in the hypothesis h at a given time $\tau \in T$:

$$\begin{aligned} \xi_r^h : T &\to \mathbb{N} \\ \tau &\mapsto \max(|r(\tau)|\,, |h(\tau)|) - |r(\tau) \cap h(\tau)| \end{aligned} \tag{25}$$

As shown in Table 4, the IER evaluation metric takes various types of error into account. In particular, incorrect speech vs. non-speech detection (or the lack of an overlapping speech detection step) may result in false alarms or missed detections.

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

$$\text{Precision}\,(r,h) = \frac{\sum_{i' \in \mathcal{I}_h \setminus \textcircled{?}} \kappa(i', i')}{\sum_{i' \in \mathcal{I}_h \setminus \textcircled{?}} \left(\sum_{i \in \mathcal{I}_r} \kappa(i, i')\right)} \tag{26}$$

$$\text{Recall}\,(r,h) = \frac{\sum_{i \in \mathcal{I}_r \setminus \textcircled{?}} \kappa(i, i)}{\sum_{i \in \mathcal{I}_r \setminus \textcircled{?}} \left(\sum_{i' \in \mathcal{I}_h} \kappa(i, i')\right)} \tag{27}$$

where $\kappa(i, i')$ is the total duration of co-occurrence between speaker i of reference r and speaker i' of hypothesis h:

$$\begin{aligned} \kappa : \mathcal{I}_r \times \mathcal{I}_h &\rightarrow \mathbb{R}^+ \\ (i, i') &\mapsto \int_{\tau \in T} \mathbb{1}_{\substack{ci \in r(\tau) \\ i' \in h(\tau)}}(\tau)\,\mathrm{d}\tau \end{aligned} \tag{28}$$

Diarization error rate (DER) While the ultimate goal of the proposed approaches is to improve speaker identification, we also report some experiments around speaker diarization. In this framework, the actual identity of the speaker does not matter—we only aim at finding the best speech turns clustering as possible. The diarization error rate (DER) was first introduced in the framework of NIST rich transcription evaluation campaigns [15].

Let us denote $\mathcal{I}_r$ the list of speakers in the reference r and $\mathcal{I}_h$ the list of speech turns clusters in the hypothesis h. Without loss of generality, we can assume that $|\mathcal{I}_r| = |\mathcal{I}_h|$. The mapping function $m^* \in {\mathcal{I}_h}^{\mathcal{I}_r}$ is defined as a bijection between $\mathcal{I}_r$ and $\mathcal{I}_h$ that maximizes the total duration of co-occurrence:

$$m^* = \underset{m \in {\mathcal{I}_h}^{\mathcal{I}_r}}{\text{argmax}} \sum_{i \in \mathcal{I}_r} \kappa(i, m(i)) \tag{29}$$

The diarization error rate (DER) can then be defined by:

$$\text{DER}\,(r,h) = \text{IER}\,\left(r, m^*(h)\right) \tag{30}$$

where $m^*(h)$ is an (abusive) shortcut to denote the hypothesis h for which each speaker is mapped to the corresponding speaker in reference r, using the optimal mapping function m^*.

In case $|\mathcal{I}_r| < |\mathcal{I}_h|$ (resp. $|\mathcal{I}_r| > |\mathcal{I}_h|$), one can artificially add dummy silent speakers $i_\varnothing$ in the reference (resp. the hypothesis), such that $\kappa(i_\varnothing, i) = 0$ for all $i \in \mathcal{I}_h$ (resp. $\kappa(i, i_\varnothing) = 0$ for all $i \in \mathcal{I}_r$).

Purity and coverage In complement to the DER, we also report the complementary values of purity and coverage:

$$\text{Purity}\,(r,h) = \frac{\sum_{i' \in \mathcal{I}_h} \max\limits_{i \in \mathcal{I}_r} \kappa(i, i')}{\sum_{i' \in \mathcal{I}_h} \left(\sum_{i \in \mathcal{I}_r} \kappa(i, i')\right)} \tag{31}$$

$$\text{Coverage}\,(r,h) = \frac{\sum_{i \in \mathcal{I}_r} \max\limits_{i' \in \mathcal{I}_h} \kappa(i, i')}{\sum_{i \in \mathcal{I}_r} \left(\sum_{i' \in \mathcal{I}_h} \kappa(i, i')\right)} \tag{32}$$

Purity measures the ratio between the duration of the speech turns of the dominating speaker in a cluster and the total duration of all speech turns in the cluster (higher is better) [8]. Coverage is the dual measure, and accounts for the dispersion of the speech turns of a given speaker across clusters (higher is better) [16].

4.3 Setup

As illustrated in Fig. 7, the REPERE corpus is divided into three sub-corpora: training set, development set and test set. The training set is used to estimate p_{tt}, p_{ti}, p_{tw} and p_{ts} introduced in Sect. 2.3. The development set is used to select the optimal value for hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ introduced in Sect. 3.2:

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\text{argmin}}\ \mathbb{E}_{\text{dev}}\,[\text{XER}(r, h)] \tag{33}$$

where XER $\in$ {IER, DER} depending on the application (speaker identification or diarization). Hyper-parameter tuning is achieved using random search. Indeed, Bergstra and 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 [3]. Finally, the test set is used for evaluation.

5 Results and discussion

5.1 Speaker diarization

Table 5 summarizes the first set of experiments focusing on speaker diarization. The proposed approaches (5A to 5D) are compared with a state-of-the-art BIC clustering baseline (5E) based on the same input segmentation into speech turns [1]. While the audio-only approach (5A) is slightly worse than the baseline (21.1 vs. 19.8 %), it does yield much purer clusters (94.7 vs. 92.1 %).

Figure 8 illustrates how the parameter α can be used to find the right balance between cluster purity and coverage. For instance, one can increase the purity of clusters by reducing the value of α [i.e., it gives more importance to the inter-cluster dissimilarity in Eq. (16)].

However, the main strength of the proposed approach is how easily it can be extended to the multi-modal (5B) and supervised (5C) cases. For instance, adding both written name vertices and supervised identification edges to the

Table 5 Speaker diarization experiments

#	Vertices	Edges	DER (%)	Purity (%)	Coverage (%)
5A	$\mathcal{T}$	$t \leftrightarrow t'$	**21.1**	**94.7**	83.6
5B	$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	18.3	93.4	85.8
5C	$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	18.2	94.0	86.1
5D	$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**17.8**	93.9	85.7
5E	BIC clustering baseline [1]		**19.8**	**92.1**	86.8

Strict transitivity constraints
Results in bold are discussed in the text

graph results in a major performance improvement: configuration 5D is 2 % better than the baseline 5E.

5.2 Oracle performance

All experiments reported in the rest of the paper focus on speaker identification. However, 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 his/her name is never written on screen nor spoken. To determine the IER lower bound, we performed oracle experiments, reported in Table 6.

An oracle is capable of correctly identifying any detected speech turn as long as the corresponding identity vertex is available in the graph. For instance, configuration 6B shows that it is theoretically possible to correctly identify 63.8 % of the total speech duration in an unsupervised way by propagation of the detected written names. When all sources of information are combined (6G), one cannot expect to get better than IER = 8.7 %.

5.3 Mono-modal speaker identification

Table 7 summarizes mono-modal supervised speaker identification experiments: they are mono-modal because they only rely on acoustic data, and supervised because they rely on prior speaker models $\mathcal{I}^*$ to identify speech turns.

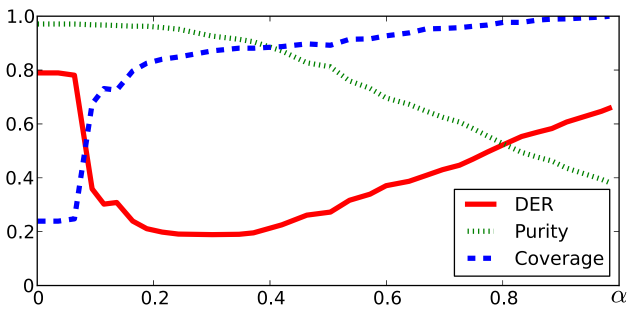

Fig. 8 Influence of parameter α on the development set. The best DER = 18.5 % is obtained for $\alpha = 0.25$

Table 6 Oracle performance

#	Vertices	IER (%)	Precision (%)	Recall (%)
6A	$\mathcal{I}^*$	35.0	100.0	67.9
6B	$\mathcal{I}_{\mathcal{W}}$	**39.3**	100.0	**63.8**
6C	$\mathcal{I}_{\mathcal{S}}$	35.8	100.0	67.3
6D	$\mathcal{I}_{\mathcal{W}} \cup \mathcal{I}_{\mathcal{S}}$	21.5	100.0	81.9
6E	$\mathcal{I}^* \cup \mathcal{I}_{\mathcal{W}}$	13.4	100.0	90.1
6F	$\mathcal{I}^* \cup \mathcal{I}_{\mathcal{S}}$	10.9	100.0	92.5
6G	$\mathcal{I}^* \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{I}_{\mathcal{S}}$	**8.7**	100.0	94.9

IER and recall do not sum to one because of non-speech, speech and overlapping speech detection errors
Results in bold are discussed in the text

It can be demonstrated that solving the optimization problem with configuration 7A leads to the following solution:

$$\mathtt{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\operatorname{argmax}}\ p_{ti} & \text{if } \ p_{ti^*} > (1 - \alpha_{\mathcal{T}\mathcal{I}^*}) \\ \textcircled{?} & \text{otherwise.} \end{cases} \tag{34}$$

This is basically 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. The only difference with the GMM–UBM baseline lies in the fact that this decision is taken at speech turn level instead of cluster level (from a preliminary speaker diarization step) for the baseline. This explains why configuration 7A leads to slightly worse results than the baseline (+0.5 % IER).

Table 7 Mono-modal speaker identification, with relaxed transitivity constraints

#	Vertices	Edges	IER (%)	P. (%)	R. (%)
7A	$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t$	49.4	54.7	54.3
7B	$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	**47.9**	55.9	**55.8**
7C	GMM–UBM baseline		**48.9**	57.5	54.3

Results in bold are discussed in the text
P. precision, *R.* recall

However, this limitation is addressed in configuration 7B by adding edges between speech turns ($t \leftrightarrow t'$) to the graph. A closer look at the hyper-parameters tuned on the development set tells us that speaker diarization ($\beta_{\mathcal{T}\mathcal{T}} = 0.55$) is given slightly more importance than supervised speaker identification ($\beta_{\mathcal{T}\mathcal{I}} = 0.45$). Moreover, $\alpha_{\mathcal{T}\mathcal{T}}$ is automatically set to 0.19, enforcing pure speech turn clusters (according to Fig. 8). Ultimately, this leads to better results than the baseline (−1 % IER).

5.4 Cross-modal speaker identification

Table 8 summarizes cross-modal unsupervised speaker identification experiments: they are unsupervised because they do not rely on any prior speaker models, and cross-modal because identities are propagated across modalities (from written or spoken names to speech turns).

Configuration 8A shows the most promising results. Indeed, although it is an unsupervised approach, it does perform better than the best mono-modal supervised approach (7B in Table 7). Moreover, its performance (IER = 46.5 %) is very close to the one of the corresponding oracle (6B in Table 6, IER = 39.3 %). It also obtains slightly better results than the baseline system 8D described in the introduction [37].

On the other side, it seems that the asynchronous nature of $t \leftrightarrow s$ edges (a speaker rarely pronounces its own name) is not well suited for the proposed framework. As a matter of fact, configuration 8B focusing on named speaker identification yields poor performances (IER = 81.8 %) even though both the speech transcription and the named-entity detection steps are done manually. However, the integration of spoken name vertices does bring a small (−0.9 % IER) improvement to configuration 8A (yet not as significant as we would have expected based on the performance of oracle 6D).

5.5 Multi-modal speaker identification

In Table 9, the last set of experiments shows how the best mono-modal supervised approach (configuration 7B, IER = 47.9 %) and the best cross-modal unsupervised one (configuration 8A, IER = 46.5 %) can be advantageously combined into a joint multi-modal speaker identification approach (configuration 9, IER = 25.3 %).

This major performance leap can be explained by the intrinsical complementarity of both approaches. Table 9 provides a detailed analysis of their behavior. Indeed, the REPERE corpus also comes with annotation of speaker roles: anchor, journalist, reporter, guest or other. The supervised approach 7B works very well for anchors (IER = 20.3 %) because a large amount of acoustic data is available in the training set to learn their models. Conversely, the unsupervised approach 8A performs very poorly (IER = 79.4 %) because the anchors names are very rarely displayed on screen. Reciprocally, it is very good (IER = 34.4 %) at recognizing journalists, reporters or guests because they are nearly systematically introduced by an overlaid title block.

Finally, Table 10 highlights the effect of transitivity constraints relaxation on the performance of the best proposed configuration. As envisioned in Sect. 3.4, strict transitivity constraints should be preferred if speaker diarization is the targeted application, while loose transitivity constraints lead to better speaker identification results. Strict constraints tend to yield purer clusters (+5.7 %) and higher precision

Table 8 Cross-modal speaker identification experiments

#	Vertices	Edges	IER (%)	Precision (%)	Recall (%)
8A	$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**46.5**	66.8	56.9
8B	$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	81.8	21.5	21.4
8C	$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$i_w \Leftrightarrow w \leftrightarrow t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	**45.6**	62.7	58.2
8D	Late name propagation baseline [37]		47.5	**90.5**	55.5

Relaxed transitivity constraints
Results in bold are discussed in the text

Table 9 Multi-modal speaker identification experiments

#	Vertices	Edges	All			Anchors			All but anchors		
			IER (%)	P. (%)	R. (%)	IER (%)	P. (%)	R. (%)	IER (%)	P. (%)	R. (%)
7B	$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9	55.9	55.8	**20.3**	**86.6**	**79.7**	51.8	50.8	48.3
8A	$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5	66.8	56.8	79.4	31.8	20.6	34.4	76.9	65.8
9	$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**	**79.4**	**78.6**	23.9	82.9	76.1	**22.4**	**82.5**	**77.9**

Relaxed transitivity constraints
Results in bold are discussed in the text

Table 10 Effect of transitivity constraints relaxation on the multi-modal configuration 9

Constraints	Strict (%)	Loose (%)
DER	**17.8**	20.0
Purity	**93.9**	88.2
Coverage	85.7	**91.1**
IER	27.6	**25.3**
Precision	**85.3**	79.4
Recall	75.9	**78.6**

Results in bold are discussed in the text

(+5.9 %), while looser ones favor higher coverage (+5.4 %) and better recall (+2.7 %).

6 Conclusion

The first contribution of this paper is the introduction of a unified framework for mono-, cross- and multi-modal person recognition in multimedia data. Dubbed person instance graph, this framework models the person recognition task as a graph mining problem: person instance or identity vertices are connected with edges weighted by the probability that they are from the same person. Practically, we described how the proposed framework can be applied to speaker identification in TV broadcast—with speech turn, written name and spoken name instance vertices.

The second contribution of this paper is related to the use of integer linear programming to solve the problem of clustering person instances based on their identity. In particular, we proposed two major extensions of our previous work [6]: a weighted version of the objective function and the relaxation of transitivity constraints.

Finally, the third contribution of this paper is a thorough experimental evaluation of the proposed framework on a publicly available benchmark database. In particular, depending on the graph configuration (i.e., the choice of its vertices and edges), we showed that multiple tasks can be addressed interchangeably (e.g., speaker diarization, supervised or unsupervised speaker identification), outperforming state-of-the-art mono-modal approaches.

While this work focused only on speaker identification, the proposed framework can be easily extended to face recognition. Indeed, state-of-the-art face detection and tracking algorithms are now robust enough to obtain reliable face tracks instance vertices f. Face similarity measures could provide weights for $f \leftrightarrow f'$ or $f \leftrightarrow i$ edges. Even when those weights are missing, the proposed framework could be used to perform speech-based face recognition (e.g., using $t \leftrightarrow f$ edges weighted by lip-sync measures [5]).

Another promising research direction is cross-show processing, i.e., building one unique person instance graph for a whole video collection—instead of one per video. This could lead to significant improvements in terms of recall: the identity of a person formally introduced in one video could be propagated automatically to other videos where he/she cannot be identified. Oracle studies on the subject tend to confirm this assumption [35]. However, scaling up the proposed approaches (based on computationally expensive integer linear programming) to such large graphs is not a trivial task. We may have to look at similar problems addressed in the graph mining community, such as community detection [4] or complex (i.e., made of heterogeneous vertices) graph clustering [27].

Acknowledgments This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency). Thanks to Johann Poignant for providing the output of video OCR.

References

1. Barras C, Zhu X, Meignier S, Gauvain JL (2006) Multi-stage speaker diarization of broadcast news. IEEE Trans Audio Speech Lang Process 14(5):1505–1512
2. Bäuml M, Tapaswi M, Stiefelhagen R (2013) Semi-supervised learning with constraints for person identification in multimedia data. In: International conference on computer vision and pattern recognition (CVPR)
3. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305
4. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008
5. Bredin H, Chollet G (2007) Audio-visual speech synchrony measure: application to biometrics. EURASIP J Adv Signal Process 2007(1):070186. doi:10.1155/2007/70186
6. Bredin H, Poignant J (2013) Integer linear programming for speaker diarization and cross-modal identification in TV broadcast. In: Interspeech 2013, 14th annual conference of the International Speech Communication Association, Lyon
7. Canseco L, Lamel L, Gauvain JL (2005) A comparative study using manual and automatic transcriptions for diarization. In: Proceedings of the IEEE automatic speech recognition and understanding, workshop, pp 415–419
8. Chen SS, Gopalakrishnan P (1998) Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: DARPA broadcast news transcription and understanding workshop. Virginia
9. Cour T, Sapp B, Nagle A, Taskar B (2010) Talking pictures: temporal grouping and dialog-supervised person recognition. In: International conference on computer vision and pattern recognition (CVPR)
10. Dimitrova N, Zhang HJ, Shahraray B, Sezan I, Huang T, Zakhor A (2002) Applications of video-content analysis and retrieval. IEEE Multimed 9(3):42–55
11. Dinarelli M, Rosset S (2011) Models cascade for tree-structured named entity detection. In: Proceedings of 5th international joint conference on natural language processing, Asian Federation of Natural Language processing, Chiang Mai, pp 1269–1278
12. Dupuy G, Rouvier M, Meignier S, Estève Y (2012) i-Vectors and ILP clustering adapted to cross-show speaker diarization. In: Inter-

speech 2012, 13th annual conference of the International Speech Communication Association

13. Estève Y, Meignier S, Deléglise, P, Mauclair J (2007) Extracting true speaker identities from transcriptions. In: Proceedings of interspeech, pp 2601–2604
14. Finkel JR, Manning CD (2008) Enforcing transitivity in coreference resolution. In: Annual meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT)
15. Fiscus JG, Garofolo, JS, Le, AN, Martin, AF, Pallett D, Przybocki MA, Sanders GA (2004) Results of the Fall 2004 STT and MDE evaluation. In: Fall 2004 rich transcription workshop (RT-04). Palisades
16. Gauvain JL, Lamel L, Adda G (1998) Partitioning and transcription of broadcast news data. In: Proceedings of international conference on spoken language processing (ICSLP 98), Sydney, pp 1335–1338
17. Gauvain JL, Lamel L, Adda G (2002) The limsi broadcast news transcription system. Speech Commun 37(1–2):89–109
18. Gauvain JL, Lee CH (1994) Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans Speech Audio Process 2(2):291–298
19. Giraudel A, Carré M, Mapelli V, Kahn J, Galibert O, Quintard L (2012) The REPERE corpus: a multimodal corpus for person recognition. In: International conference on language resources and evaluation (LREC)
20. Gravier G, Adda G, Paulson N, Carré M, Giraudel A, Galibert O (2012) 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
21. Gurobi Optimization Inc (2012) Gurobi optimizer reference manual. http://www.gurobi.com. Accessed 5 May 2014
22. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 87(4):1738–1752. doi:10.1121/1.399423
23. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
24. Jousse V, Petitrenaud S, Meignier S, Estève Y, Jacquin C (2009) Automatic named identification of speakers using diarization and ASR systems. In: ICASSP 2009, IEEE international conference on acoustics, speech, and signal processing, Taïpei
25. Lawto J, Gauvain JL, Lamel L, Grefenstette G, Gravier G, Despres J, Guinaudeau C, Sebillot P (2011) A scalable video search engine based on audio content indexing and topic segmentation. In: Networked and electronic media (NEM) summit : implementing future media internet
26. Le VB, Barras C, Ferras M (2010) On the use of GSV-SVM for speaker diarization and tracking. In: Proceedings of Odyssey 2010—the speaker and language recognition workshop, Brno, pp 146–150
27. Long B, Zhang MZ, Yu PS, Tianbing X (2008) Clustering on complex graphs. In: Proceedings of the twenty-third AAAI conference on artificial intelligence
28. Mauclair J, Meignier S, Estève Y (2006) Speaker diarization: about whom the speaker is talking? In: IEEE Odyssey
29. Mouysset S, Noailles J, Ruiz D, Guivarch R (2011) On a strategy for spectral clustering with parallel computation. High Perform Comput Comput Sci VECPAR 2010:408–420
30. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582
31. Pan JY, Yang HJ, Faloutsos C (2004) MMSS: Multi-modal story-oriented video summarization. In: Proceedings of the fourth IEEE international conference on data mining (ICDM)
32. Pan JY, Yang HJ, Faloutsos C, Duygulu P (2004) Automatic multimedia cross-modal correlation discovery. In: Proceedings of the 10th ACM SIGKDD conference
33. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: Proceedings of Odyssey 2001—the speaker recognition workshop, Crete, pp 213–218
34. Pelleg D, Moore AW (2000) X-means: extending K-means with efficient estimation of the number of clusters. Proceedings of the seventeenth international conference on machine learning, ICML '00Morgan Kaufmann Publishers Inc., San Francisco, pp 727–734
35. Poignant J, Besacier L, Le VB, Rosset S, Quénot G (2013) Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both? In: Interspeech 2013, 14th annual conference of the International Speech Communication Association, Lyon
36. Poignant J, Besacier L, Quénot G, Thollard F (2012) From text detection in videos to person identification. In: International conference on multimedia and expo (ICME)
37. Poignant J, Bredin H, Le VB, Besacier L, Barras C, Quénot G (2012) Unsupervised speaker identification using overlaid texts in TV broadcast. In: Interspeech 2012, 13th annual conference of the International Speech Communication Association, Portland
38. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digit Signal Process 10(1–3):19–41
39. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
40. Smith R (2007) An overview of the tesseract OCR engine. In: Proceedings of the ninth international conference on document analysis and recognition, vol 02, ICDAR '07IEEE Computer Society, Washington, DC, pp 629–633
41. Tranter SE (2006) Who really spoke when? Finding speaker turns and identities in broadcast news audio. In: Proceedings of the ICASSP, pp 1013–1016
42. Wang Y, Liu Z, Huang JC (2000) Multimedia content analysisusing both audio and visual clues. IEEE Signal Process Mag 17(6):12–36