An Adaptive Method for Cross-Recording Speaker Diarization

Gaël Le Lan, Delphine Charlet, Anthony Larcher, Sylvain Meignier

An Adaptive Method for Cross-Recording Speaker Diarization

Gaël Le Lan, Delphine Charlet, Anthony Larcher, and Sylvain Meignier,

***Abstract*—Nowadays, state-of-the-art speaker diarization and linking systems heavily rely on between-recording variability compensation methods to accurately process large collections of recordings. Variability estimation is performed on consequent training datasets, which must be labeled by speaker. One major problem of such systems is the acoustic mismatch between training and target data that degrades performances. Most of the collections contain lots of speakers speaking in various acoustic conditions. In this paper, we investigate how unlabeled speakers can help improve between-recording variability estimation, to overcome the mismatch issue. We propose a scalable unsupervised adaptation framework for two types of variability compensation. The proposed framework consists in adapting a state-of-the-art diarization and linking system, trained on *out-domain* data, using the data of the collection itself. Experiments in mismatch condition are run on two French Television shows, while the initial training dataset is composed of Radio recordings. Results indicate that the proposed adaptation framework reduces the cross-recording DER of 13% in average for variable collection sizes.**

***Index Terms*—speaker diarization, speaker linking, domain adaptation.**

I. Introduction

The increasing volume of audio and video data produced every day by social or traditional media, conferences, meetings or MOOCs requires powerful tools to automatically index topics, languages or speakers. The task of speaker diarization aims at answering the question "who speaks when ?" within a recording, and can be extended to the task of cross-recording diarization when it comes to consider a collection of recordings. Cross-recording diarization is the task of detecting and labeling speaker segments in a way that each unique speaker will be identified by a single label across the entire collection.

In the literature, cross-recording speaker diarization is often considered a complimentary step to the within-recording diarization task. Even if the terminology varies (*Speaker Linking* in [1][2][3][4], *Cross-Show Speaker Diarization* for [5][6]), it tends to normalize to *Speaker Diarization and Linking* in [7][8]. Each recording is usually processed separately (within-recording diarization) before estimated speaker segments are linked across the collection (cross-recording speaker linking). In this paper, we use the terms *diarization* for within-recording diarization, and *linking* for cross-recording linking. We will also use the term *collection*, which is a more generic than *show*. A show is made of multiple episodes having similar characteristics, while a collection can be made of multiple episodes of a show or of episodes of different shows.

Manuscript received April 19, 2005; revised August 26, 2015.

G. Le Lan is with Orange Labs (email: gael.lelan@orange.com).

D. Charlet is with Orange Labs, France (email: delphine.charlet@orange.com).

A. Larcher and S. Meignier are with University of Le Mans, France (email: anthony.larcher@univ-lemans.fr; sylvain.meignier@univ-lemans.fr).

Speaker Diarization and Linking is about differentiating speakers. State-of-the-art approaches combine the *i-vector* paradigm [9] to represent speaker segments, and within- and between-speaker variability compensation to discriminate them in terms of speaker. *I-vectors* can be compared using similarity scores of likelihood ratios. Within- and between-speaker variabilities are estimated over a speaker labeled dataset, which must include multiple examples of a same speaker in various acoustic conditions. Such accurate labels are expensive to create as they require human operators. Due to the important acoustic mismatch, state-of-the-art models dedicated to a type of data (e.g. radio) do not guaranty comparable performances on other domains (television, meetings...). In speaker verification, domain adaptation has been studied for the past years to overcome this acoustic mismatch issue [10][11][12]. Domain adaptation consists in adapting statistical models dedicated to a type of data (or domain) to prevent performance degradation when applying the models on a new domain.

In our prior work [13], we investigated domain adaptation in the context of speaker diarization and linking. We proposed an adaptive diarization framework which iteratively extracts information from the collections it processes to update speaker variabilities using the Weighted Likelihood domain adaptation method [10] in mismatch conditions. Speaker variabilities are modeled through Probabilistic Discriminant Linear Analysis (PLDA) [14]. In this paper, we extend our previous work with a study on the scalability of the method, i.e., the adaptation framework is designed to process collections of variable sizes. We evaluate the proposed approach for two types of speaker variability compensations: Within Class Covariance Normalization (WCCN [**?**]) and PLDA.

The paper is organized as follows. Section II describes the general architecture of a diarization and linking system. Section III focuses on the *i-vector* paradigm and modeling of speaker variabilities. Section IV addresses domain adaptation techniques based on the *i-vector* paradigm. In Section V, the proposed diarization and linking framework is described. The data used for the experiments are presented in Section VI while experimental results are presented and analyzed in Section VII. Conclusions and perspectives are discussed in Section VIII.

II. Speaker Diarization and Linking

Diarization has to answer the question "who speaks when?". It aims at partitioning audio recordings into homogeneous segments according to the speaker identity, at different scales. When applied on a collection of recordings, the task is usually performed sequentially: 1) at a local scale, diarization is performed within each recording to produce homogeneous segments in terms of speaker, 2) at a wider scale, the resulting segments are linked per speaker across all recordings. Speakers appearing in more than one recording of a collection are called recurring (R.) speakers, as opposed to one-time (O.T.) speakers, who only speak in one recording.

A. Diarization

Diarization consists in processing a single audio recording to produce a list of segments identified by labels. Each label corresponds to a supposed speaker. As the number of speakers and channel variability is usually limited within a single recording, low complexity models can be used. Diarization is usually done in 3 steps.

1) Frontend: The frontend consists in extracting acoustic features from the audio before applying a Speech Activity Detector to remove noise and silence. The most used acoustic features are the Mel Frequency Cepstral Coefficients (MFCC).

2) Segmentation: Segmentation is about detecting speaker changes by comparing consecutive segments of speech to assess whether they are pronounced by the same speaker or not. Producing pure segments is essential for the accuracy of the following linking. This can be achieved with bottom up approach, which consists in successively applying several techniques with an increasing modeling complexity, to merge initial fixed duration segments into longer ones. The lowest level approach compares two consecutive overlapping windows of a fixed short duration, using metrics like Generalized Likelihood Ratio (GLR) [15], Kullback-Leibler divergence [16][17] or Gaussian Divergence [18]. It produces a first set of short segments.

The highest level segmentation removes the false segment boundaries detected by the lowest level one. Low-level segments of sufficient duration enable the application of a more refined method to relieve false segment boundaries. The Bayesian Information Criterion (BIC) [19] is used to fuse segments belonging to a same speaker. Contrary to the low-level methods, BIC can be used to compare consecutive segments of variable duration, but requires a decision threshold which is usually empirically chosen to favor purity of the produced segments. In order to allow higher level speaker representations, such as GMM-based or *i-vector* modeling, segments must contain only one speaker. In some cases the multi-level approach is not used. For example, in [20], since the data contain only two speakers, the authors decide to directly cluster a set of fixed duration small segments in two classes, using the *i-vector* representation and a Principal Component Analysis to separate the set into homogeneous segments.

3) Clustering and refinement: This last step at the scale of the recording consists in clustering all segments by speaker. Due to the longer duration of the segments, higher complexity representations can be used to compare any pair of segments within the recording. Three main models are described in the literature : Gaussian, using BIC as similarity measure [21], Gaussian Mixture Model (GMM), using Cross Likelihood Ratio (CLR) [22][3], and *i-vector*, using cosine distance or PLDA likelihood ratio for scoring [21]. For all segments, similarities between associated models are estimated to perform clustering. The most used clustering method is the bottom-up hierarchical agglomerative clustering (HAC) [23][15][24][16][25][26], but other methods were proposed, like k-means [27], graph-based [28], or ILP clustering [29].

Segment boundaries are sometimes refined via the Viterbi algorithm [21][30]: small GMMs are iteratively trained for each speaker cluster and a Viterbi decoding adjusts the boundaries, until some convergence criterion is met.

B. Linking

Once diarization has been applied to each recording, the segments are to be linked across the collection. The aim is to connect all segments from a same speaker across recordings. The process is similar with the clustering step of diarization II-A3, but differs on two points. First, the magnitude of the number of segments to connect is higher. Second, the variability between them is higher. Some recurring speakers may appear in various acoustic conditions across the collection. Thus, estimating the within-speaker/between-recordings variability is much more important than when working on a single recording. Depending on the chronology of the collection aging of the speakers can strongly increase the within-speaker variability [31][32].

Since the linking step is similar to the within-recording clustering, the models and scoring methods are very similar: Gaussian/BIC [6], GMM/CLR [5][18] and *i-vector*/cosine or *i-vector*/PLDA [28][33]. However, due to the important within-speaker/between-recording variability, *i-vector*-based methods involving compensation techniques like WCCN or PLDA are preferred in recent literature.

III. The *i-vector* paradigm

Introduced in [9], *i-vectors* provide compact representation of acoustic segments that were previously modeled using GMMs [34]. According to this paradigm, a GMM mean supervector $m_{i,j}$, modeling session j of speaker i, is an observation produced by a generative model described by:

$$\boldsymbol{m_{ij}} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{\phi_{ij}} \tag{1}$$

In this equation, $\boldsymbol{\phi}$ is a random variable for which the Maximum a Posteriori point estimate is the *i-vector* $\boldsymbol{\phi_{ij}}$. $\boldsymbol{\mu}$ is the speaker- and channel-independent supervector, while $\boldsymbol{T}$ is the Total Variability matrix.

A. Cosine Scoring and WCCN compensation

Cosine similarity is a simple and well established method to compare *i-vectors* [9].

$$s(\boldsymbol{\phi_1}, \boldsymbol{\phi_2}) = \frac{\boldsymbol{\phi_1}\boldsymbol{\phi_2}}{\|\boldsymbol{\phi_1}\|\|\boldsymbol{\phi_2}\|} \in [-1; 1] \tag{2}$$

It can be combined with Within Class Covariance Normalization (WCCN) where the within class covariance (WCC) matrix is the average within-speaker covariance, weighted by the number of sessions n_i of each speaker i [9]. $\boldsymbol{h_i}$ is the average *i-vector* of speaker i.

$$\boldsymbol{W} = \frac{1}{S}\sum_{i=1}^{S}\frac{1}{n_i}\sum_{j=1}^{n_i}(\boldsymbol{\phi_{ij}} - \boldsymbol{h_i})(\boldsymbol{\phi_{ij}} - \boldsymbol{h_i})^T \tag{3}$$

Once estimated the WCC matrix, the variability can be compensated by rotating the *i-vectors* according to:

$$\hat{\phi}_{ij} = \boldsymbol{L}\phi_{ij} \tag{4}$$

$\boldsymbol{L}$ is the Cholesky decomposition of $\boldsymbol{W^{-1}}$, $\boldsymbol{W^{-1}} = \boldsymbol{LL^T}$.

B. PLDA modeling

Introduced for face recognition applications [35], PLDA has been adapted to speaker identification [14]. The *i-vectors* are considered as observations of a probabilistic generative model. Each *i-vector* $\boldsymbol{\phi_{ij}}$, of dimension p, can be decomposed as:

$$\boldsymbol{\phi_{ij}} = \boldsymbol{\mu} + \boldsymbol{\Phi h_i} + \boldsymbol{\epsilon} \tag{5}$$

In this equation, the hidden variable $\boldsymbol{h_i}$, called speaker factor, is speaker-dependent and channel-independent, while $\boldsymbol{\epsilon}$ is the residual. $\boldsymbol{h_i}$ follows a standard normal distribution, while $\boldsymbol{\epsilon}$ is distributed along $\mathcal{N}(0, \boldsymbol{\Lambda})$. $\boldsymbol{\Phi}$ is of size $p \times r (r < p)$. The *i-vectors* distribution is $\mathcal{N}(0, \boldsymbol{\Phi\Phi^T} + \boldsymbol{\Lambda})$. $\boldsymbol{\Phi\Phi^T}$ represents the between-speaker variability matrix and $\boldsymbol{\Lambda}$ the within-speaker variability matrix. Estimation of $\boldsymbol{\mu}$ $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$ is performed with a speaker labeled training dataset, using the EM algorithm.

In the remaining of this paper, we consider that *i-vectors* are centered (i.e., $\boldsymbol{\mu} = 0$) and length-normalized [36]. Training the PLDA model $\Theta = (\boldsymbol{\Phi}, \boldsymbol{\Lambda})$ consists in estimating the values maximizing the likelihood:

$$L_k(\boldsymbol{\Phi\Phi^T}, \boldsymbol{\Lambda}) = \frac{1}{N}\sum_{i=1}^{S}\sum_{j=1}^{n_i} log(p(\boldsymbol{\phi_{ij}}|\boldsymbol{\Phi\Phi^T}, \boldsymbol{\Lambda})) \tag{6}$$

with

$$p((\boldsymbol{\phi_{ij}})|\boldsymbol{\Phi\Phi^T}, \boldsymbol{\Lambda}) = \mathcal{N}((\boldsymbol{\phi_{ij}}); 0, \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T + \tilde{\boldsymbol{\Lambda}}) \tag{7}$$

Where $\tilde{\boldsymbol{\Phi}}$ is a column bloc matrix containing N times $\boldsymbol{\Phi}$ and $\tilde{\boldsymbol{\Lambda}}$ is a bloc diagonal matrix containing N times $\boldsymbol{\Lambda}$, N being the total number of *i-vectors*.

The EM algorithm steps are:

- E-step: estimate the a posteriori probability of the hidden speaker variables $\boldsymbol{h_i}$, using all the observations $\{\boldsymbol{\phi_{ij}}\}_{j=1}^{n_i}$ of those speakers.

$$E[\boldsymbol{h_i}] = (N_i\boldsymbol{\Phi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi} + \boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Lambda}^{-1}\sum_{j=1}^{n_i}\boldsymbol{\phi_{ij}} \tag{8}$$

$$E[\boldsymbol{h_i}\boldsymbol{h_i^T}] = (N_i\boldsymbol{\Phi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi} + \boldsymbol{I})^{-1} + E[\boldsymbol{h_i}]E[\boldsymbol{h_i}]^T \tag{9}$$

- M-step : using the previous estimates, update the model.

$$\boldsymbol{\Phi_{new}} = \left(\sum_{i=1}^{S}\sum_{j=1}^{n_i}\boldsymbol{\phi_{ij}}E[\boldsymbol{h_i}]^T\right)\left(\sum_{i=1}^{S}N_iE[\boldsymbol{h_i}\boldsymbol{h_i^T}]\right)^{-1} \tag{10}$$

$$\boldsymbol{\Lambda_{new}} = \frac{1}{N}\sum_{i=1}^{S}\sum_{j=1}^{n_i}\left[\boldsymbol{\phi_{ij}}\boldsymbol{\phi_{ij}^T} - \boldsymbol{\Phi_{new}}E[\boldsymbol{h_i}]\boldsymbol{\phi_{ij}^T}\right] \tag{11}$$

IV. Domain adaptation

In real life, acoustic models used to process test data, called in-domain, are learnt on a different train set, called out-domain. It often happens that the mismatch between in-domain test data and out-domain train data strongly affects the performance of an automatic system processing a type of data it has not been trained for (the in-domain test data). Domain adaptation aims at compensating for the acoustic mismatch between in-domain and out-domain. Usually, a third dataset, called *development*, from the in-domain, is used to adapt the out-domain models. In our work, we do not have such a dataset and decide to directly use the test set for adaptation. Training of a speaker diarization system is especially demanding as the *i-vector*/PLDA paradigm requires speaker annotated data to estimate within- and between- speaker variability. When *in-domain* data are not sufficient to estimate a PLDA model, it can be used to adapt an *out-domain* model, e.g. estimated on *out-domain* data. In [12], it was found that the most important component in a classical *i-vector*/PLDA based system for domain adaptation is PLDA. Thus, we will only present the adaptation methods focusing on within class covariance [37] and PLDA adaptation [10][12], but other approaches were proposed, working on the Total Variability space [38][39][40][**?**].

A. Within-Speaker Covariance adaptation

In [37], the author focus on the adaptation of Within Speaker Covariance. The speaker comparison framework is based on *i-vector* modeling, combined with Linear Discriminant Analysis (LDA) for dimensionality reduction and PLDA for scoring. Since LDA relies on the computation of within- and between-class covariances, the idea is to adapt the within-speaker variability by adapting the between-domain variability.

$$\boldsymbol{W_{new}} = \boldsymbol{W} + \alpha\boldsymbol{W_{BD}} \tag{12}$$

$\boldsymbol{W_{BD}}$ represents the between-domain covariance, while $\boldsymbol{W}$ stands for the within-class covariance. Boosting the between-domain variability allows to reduce the Fisher ratio in LDA for

the between-domain variability directions, which are supposed to be speaker independent. Consequently, the LDA projection will not include these directions.

B. PLDA adaptation

Two methods of PLDA adaptation are presented. The choice of the method depends on the quantity of *test* data available for adaptation.

1) Weighted Likelihood: The main idea of Weighted Likelihood Domain Adaptation is to express the maximum of likelihood for PLDA estimation in two terms depending on each domain, introducing a weighting parameter. It was introduced in [10]. This method can be used even with a limited quantity of *in-domain* data.

$$L(\mathbf{\Phi\Phi^T}, \mathbf{\Lambda}) = \alpha L_{in}(\mathbf{\Phi\Phi^T}, \mathbf{\Lambda}) + (1-\alpha) L_{out}(\mathbf{\Phi\Phi^T}, \mathbf{\Lambda}) \quad (13)$$

Where

$$L_k(\mathbf{\Phi\Phi^T}, \mathbf{\Lambda}) = \frac{1}{N_k} \sum_{i=1}^{S_k} \sum_{j=1}^{n_{ik}} log(p((\boldsymbol{\phi_{ij}})|\mathbf{\Phi\Phi^T}, \mathbf{\Lambda})) \quad (14)$$

N_k is the number of k-*domain i-vectors* and S_k is the number of speakers. This method allows to chose how influent a dataset is in regard to the other, for the parameters estimation. Estimation of the adapted PLDA parameters is similar to classical PLDA, with the introduction of the weighting parameter.

In the literature [10], results show that the adapted PLDA gives better results than the out-domain PLDA parameters, and that the Equal Error Rate decreases when the number of speakers of the in-domain collection increases.

2) A posteriori interpolation: When the *in-domain* collection contains enough data (number of sessions superior to the *i-vectors* dimension), a faster approximation of the previous method consists in interpolating the pre-estimated PLDA parameters. Two PLDA models are separately trained on each dataset, before being interpolated.

$$\mathbf{\Phi\Phi}^{\boldsymbol{T}}_{\boldsymbol{final}} = \alpha_1 \mathbf{\Phi\Phi}^{\boldsymbol{T}}_{\boldsymbol{in}} + (1-\alpha_1) \mathbf{\Phi\Phi}^{\boldsymbol{T}}_{\boldsymbol{out}} \quad (15)$$

$$\mathbf{\Lambda}_{\boldsymbol{final}} = \alpha_2 \mathbf{\Lambda}_{\boldsymbol{in}} + (1-\alpha_2) \mathbf{\Lambda}_{\boldsymbol{out}} \quad (16)$$

C. Unsupervised adaptation

Sometimes, the *in-domain* collection is unlabeled [12][41]. *In-domain* PLDA parameters need to be estimated in an unsupervised way, using clustering to label the data. For example, in [12], *out-domain* PLDA parameters are used to compute similarities between the *in-domain i-vectors*. The resulting similarity matrix is then used to cluster the *i-vectors*. The obtained clusters allow to estimate *in-domain* PLDA parameters, which can then be interpolated with the *out-domain* parameters, using equations 15 and 16. The results show that the interpolation works best when the *in-domain* number of speakers is low. When this number increases, interpolated PLDA tends to give similar results to the *in-domain* unsupervised PLDA.

D. Domain adaptation for speaker diarization

Until recently, domain adaptation was mainly studied for the task of speaker identification/clustering, where the compared recordings only contain one speaker voice (mostly phone recordings). Common domain adaptation techniques aim to estimate a better modeling of a collection variabilities (Total Variability subspace, between- and within-speaker variability), starting from an *out-domain* model, using *in-domain* data. Speaker diarization is a more demanding task, where recordings must be segmented before linking the resulting segments.

V. Proposed Adaptative Framework for Diarization

The goal of our work is to focus on how the inaccurate knowledge of the *in-domain* collections can help improve an *out-domain*-based diarization system, using domain adaptation techniques. As seen in the previous section IV, modeling the within- and between- speaker variabilities is a key step in the adaptation process. The *in-domain* collections consist of TV shows and their size increases over time. Some speakers appear in multiple recordings: we propose to use them to adapt the system, in an unsupervised way. As we want to be able to process and adapt over any kind of collection, the adaptation method must be scalable.

The proposed framework, described in the following sections, consists in adapting a state-of-the-art diarization system ($baseline$), trained on *out-domain* data ($train$ data), using the data of the collection itself ($target$ data). The first section is dedicated to the description of the $baseline$ system, the next presents the proposed adaptation strategy and the third one focuses on the issue of the $target$ collection size variability.

A. $Baseline$ Diarization Framework

Figure 1 describes the diarization framework, detailed thereafter. It was developed using the SIDEKIT toolkit [42]. The supervised $baseline$ framework, described in this section, is presented with the plain lines, while the proposed unsupervised adaptation, described in the next one, is presented with the dashed lines.

The *i-vector* representation used in the following is estimated over a GMM/UBM of 256 Gaussians with diagonal covariance, computed on the $train$ corpus. *Out-domain* WCCN and PLDA matrices are also estimated on this speaker labeled corpus. The dimension of the *i-vectors* is 200 and PLDA eigenvoice matrix has a dimension of 100 with no eigenchannel matrix. Those parameters correspond to the best configuration we found after performing an exhaustive search.

The within-recording diarization is applied to each recording independently. The front-end consists in the extraction of 39 features (13 MFCCs with Δ and $\Delta\Delta$), followed by a 2-Gaussian Viterbi-based speech detector. Speaker change detection is performed with a standard Gaussian divergence segmentation, using a 20ms sliding window, supplemented by a Gaussian/BIC segmentation, which merges the initial segments. Finally, two complete-linkage HAC are successively used to

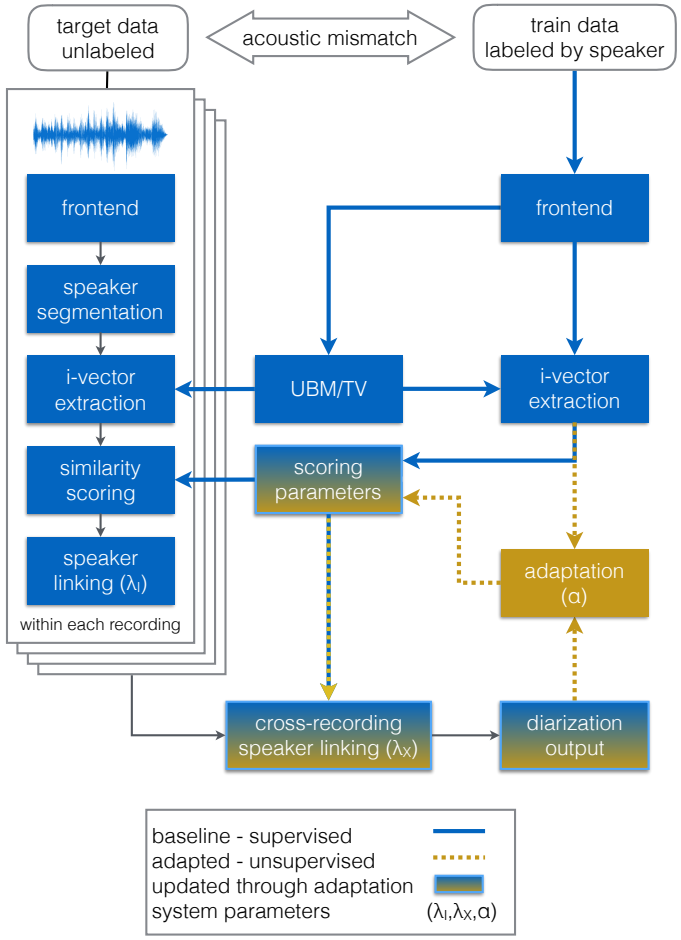

Figure 1. Overview of the diarization framework for *baseline* (plain blue lines) and *adapted* (dashed lines) training.

cluster the segments within each recording: Gaussian/BIC, followed by *i-vector*/(WCCN/cosine) or *i-vector*/PLDA as model/similarity. Before the *i-vector* extraction, MFCC features are centered with unit variance, for each BIC segment. The HAC threshold for the *i-vector*-based clustering is noted λ_I. At the end of the within-diarization step, each within-recording cluster is represented by the average of its *i-vectors*.

For the linking step, the previous averaged *i-vectors* are reused, either combined with WCCN/cosine or PLDA for similarity computation. Complete-linkage Hierarchical Agglomerative Clustering is used to cluster the *i-vectors*, with a clustering threshold λ_X.

B. Unsupervised Adaptation Framework

The proposed adaptation framework relies on an iterative adaptation approach to perform the linking. The output of the *baseline* cross-recording diarization system consists in speaker clusters. Among those clusters, some contain segments from different recordings and can be used to update speaker variabilities (modeled through WCCN and PLDA), using interpolation methods.

Using updated WCCN or PLDA parameters, the similarities between *i-vectors* from all recordings can be refreshed, which leads to updating the linking, hence the cross-recording diarization output.

Even if the *baseline* produces diarization errors, we suppose the information brought by the *target* speaker clusters should refine the within-speaker/between-recording variability estimation. Better parameters should give more accurate clusters, which could be used for another adaptation loop: we expect the accuracy of clusters to improve over iterations.

The adaptation methods used for WCCN and PLDA are detailed in the two following sections.

1) WCCN adaptation: For WCCN adaptation, we propose to compute a new WCCN matrix, as a weighted sum of $\boldsymbol{W_{train}}$, the initial WCCN matrix of the *baseline* system, and $\boldsymbol{W_{target}}$ computed over the clusters produced by the *baseline* system.

$$\boldsymbol{W_{adapt}} = \alpha \boldsymbol{W_{target}} + (1-\alpha)\boldsymbol{W_{train}} \tag{17}$$

2) PLDA adaptation: For scalability reasons, we expect not to always be able to estimate PLDA using the *target* clusters only. To adapt PLDA parameters, the Weighted Likelihood Domain Adaptation method, presented in section IV-B1 is chosen.

C. Scalability

The diarization system we work with depends on a triplet of parameters $(\lambda_I, \lambda_X, \alpha)$. λ_I and λ_X are the within-recording and cross-recording clustering thresholds, and α the adaptation parameter. The collections to process are usually stored in chronological order and their size increases over time. Depending on the collections size, it might be useful to give the *target* more or less weight in the adaptation process. If we adapt the parameters with only a few episodes, the number or speaker clusters used for adaptation is going to be limited, and the confidence in the *in-domain* contribution might need to be limited to avoid an inaccurate estimation.

Instead of setting α empirically, we propose to make it depend on some variables representative of the *target* collection, with a formula inspired by MAP adaptation [43]. When it comes to WCCN computation or PLDA estimation, the number of speakers (or clusters) and the number of sessions are key factors. If one of those criteria is too low, estimation can fail (mainly due to matrix inversion issues). In [10], the authors showed that the optimal value of α depends on the number of speakers used for adaptation, using a two-covariance model and a parameter interpolation method. We want α to be close to 0 (respectively 1) when the number of speakers is rather low (respectively high), which leads us to propose the following formula.

$$\alpha = \frac{{S_{target}}^p}{{S_{target}}^p + r^p} \tag{18}$$

S_{target} is the number of recurring speakers (ie. speaker clusters containing 3 or more sessions) from the *target* collection.

We call r the *virtual* size of the $train$ collection and p is a strength factor. r corresponds to the $target$ collection size for which the influence of both collections is identical. Some curves describing the formula are presented in figure 2. The figure represents the evolution of α, depending of the number of recurring speakers (or clusters), for $r = 20$ or $r = 60$ and p from 1 to 3.

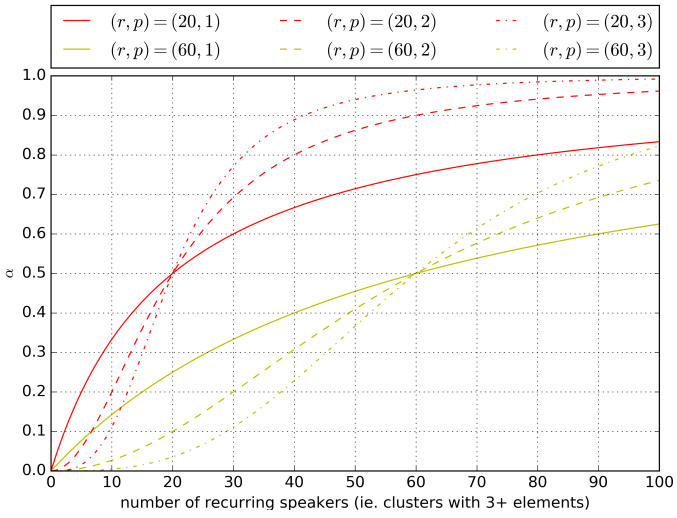

Figure 2. Examples for the proposed formula, for $r = 20$ or $r = 60$ and p from 1 to 3.

VI. Experimental collections

$Baseline$ models for diarization systems were trained on manually annotated corpus. In this corpus speakers are identified by their first and last names, providing several sessions for a large set of speakers. About 220 hours of French broadcast news drawn from REPERE [44], ETAPE [45] and ESTER[46] evaluation campaigns were used to build three corpora. The shows were broadcast between 1998 and 2007, duration of shows ranges from ten minutes to one hour. The corpora also contain some broadcasts of Moroccan radio, in French language. For each show in the corpus, multiple episodes are available. Only radio shows were used to build the train corpus, while both target corpora contain TV shows only. This was chosen to maximize the acoustic mismatch between the train and target data.

A. Train corpus

The $train$ corpus, used to train the $baseline$ system, is composed of 317 audio files from ESTER campaign corpora, taken from radio broadcasts, for a total of 190 hours of speech duration. For each show, all available episodes are taken. Many speakers appear in more than one episode, but some also appear in different shows (politicians, for example). The corpus contains 3212 unique speakers. Among those speakers, 372 meet our requirements for PLDA training: they appear in at least three recordings, with a minimum speech time per recordings of 10s. Thus, this corpus is well suited for an *i-vector*/PLDA system training.

B. Target corpora

We define two $target$ collections built from REPERE and ETAPE corpora. The first one, named LCP_{target}, is the collection of all available episodes of the show *LCP Info*, a French TV news broadcast show. The second target corpus, named BFM_{target}, is the collection of all available episodes of the TV news talk-show *BFM Story*. Those two corpora have been chosen because they both contain a decent number of episodes (more than 40), and there is a large amount of recurring speakers, who speak for more than 50% of the total speech duration of the collection. Numerical details about the two corpora are presented in table I.

Corpus	LCP_{target}	BFM_{target}
Episodes	45	42
Labeled speech duration	10h08m	19h57m
One-Time speakers	127	345
Recurring speakers (2+ occurrences)	93	77
R. speakers (3+ occurrences)	48	35
Total speakers	220	422
O.T. speakers speech proportion	20.12%	44,84%
R. speakers (2+ occurrences) s.p.	79.88%	55,16%
R. speakers (3+ occurrences) s.p.	67.06%	45.94%
Average speaker time per episode	1m08s	1m58s

Table I: Composition of target corpora. Annotated speakers numbers are presented.

VII. Experiments

Experiments were evaluated using the Diarization Error Rate (DER). DER was introduced by the NIST as the fraction of speech time which is not attributed to the correct speaker, using the best match between references and hypothesis speaker labels. The scoring tool [47] is employed for within-recording and cross-recording speaker diarization. Cross-recording speaker diarization aims at labeling a recurring speaker the same way, in every recording that composes a collection. For DER computation, a collar of 250ms is used, and the overlapping speech is included. The 250ms collar removes between 2 and 3% of the total speech time of each collection.

In this paper, we mainly focus on the evaluation of the cross-recording DER, which we will call X-DER in the following sections, as opposed to I-DER, for within-recording DER.

A. Baseline System

Results of the $baseline$ diarization system are presented in table II. Three configurations are compared, with different scoring methods: cosine without normalization, cosine with WCCN and PLDA. The $baseline$ system is trained on the $train$ dataset only: WCCN and PLDA matrices are estimated on the *out-domain i-vectors*. The UBM dimension is of 256, TV rank is 200 and PLDA rank is 100, and those dimension will be fixed for all experiments. For each scoring method and corpora, I-DER and X-DER are shown for the optimal clustering configuration (λ_I, λ_X). As seen in the table II, when looking at the cross-recording DER, optimal error rates vary from 19.5%

to 18.1% for the LCP collection and from 22.2% to 15.7% for BFM. We also note that there is no common configuration that optimizes X-DER for both collections. The best performing system uses PLDA as a scoring method, and results show that using scoring or normalization methods exploiting speaker variability outperforms the cosine-only system, as expected. However, compensation methods have more impact on BFM than LCP.

B. Oracle adaptation

The goal of our work is to focus on how the inaccurate knowledge of the *in-domain* collections can help improve an *out-domain*-based diarization system, using domain adaptation techniques. To estimate how much it can improve, we first perform an experiment where we use *oracle in-domain i-vectors* to adapt WCCN or PLDA parameters. This means, if we have a perfect knowledge about the *in-domain* speaker variabilities, how good can domain adaptation methods improve the results ?

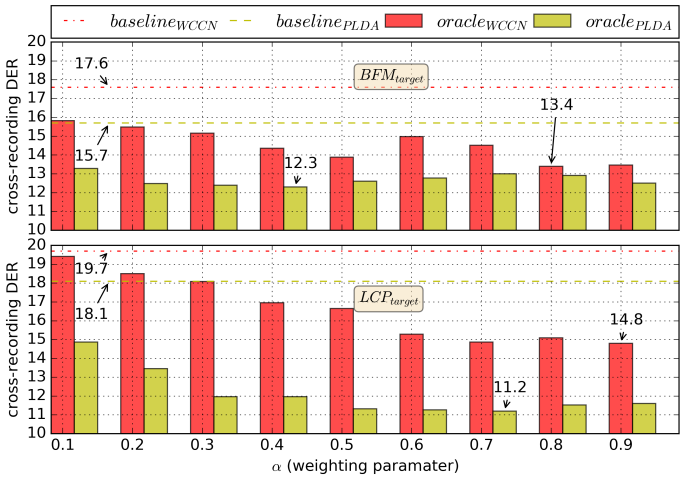

Figure 3. Effect of *oracle* domain adaptation on the cross-recording DER for both target corpora, with WCCN or PLDA parameters adapted using *oracle i-vectors*, as a function of α.

Results are presented in figure 3. Two scoring methods are tested, where the *baseline* WCCN and PLDA parameters are adapted with the *oracle in-domain i-vectors*, the weight of adaptation data α varying from 0.1 to 0.9. Experiments $\alpha = 0$ and $\alpha = 1$ are not presented, since $\alpha = 0$ corresponds to the *baseline* experiment, while with $\alpha = 1$, WCCN or PLDA estimation is not possible due to the limited quantity of *target* data (matrix inversion issues). Only the cross-recording DER is presented.

For both adaptation strategies, results show that a significant gain is achievable through adaptation (relative reduction of X-DER between 22% and 38%). We also note that if optimal values of α are not the same for both corpora, we still observe a X-DER reduction whatever α.

C. Iterative adaptation

For this experiment, we stop using *oracle i-vectors* for adaptation, and work with the set of *target i-vectors*, extracted from the segments produced by the *baseline* within-recording diarization pass. After the *baseline* linking, those *i-vectors* are grouped into speaker clusters and can be used to adapt WCCN or PLDA parameters. Similarity scores are updated and give a new clustering, which can again be used for adaptation: the process can be iterated.

Each adaptation experiment depends on a triplet $(\lambda_I, \lambda_X, \alpha)$, λ_I being the within-recording HAC threshold, λ_X the cross-recording one, and α the adaptation parameter. Results are presented in figures 4 and 5, for the PLDA-adapted experiment, with 4 successive iterations of adaptation. They show that for various triplets $(\lambda_I, \lambda_X, \alpha)$, iterative adaptation can gradually improve the baseline X-DER.

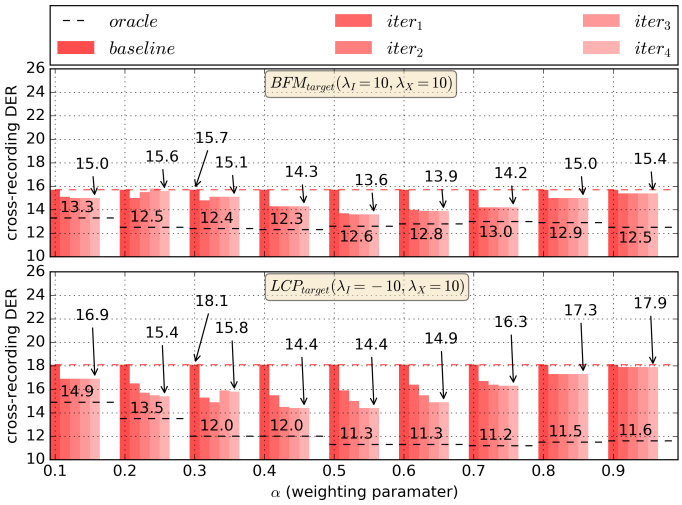

Figure 4. Cross-recording DER for both target corpora, for iterations 0 (baseline) to 4, as a function of α, using PLDA for scoring and adaptation.

Figure 4 presents the results as a function of α, λ_I and λ_X being set, while figure 5 presents them as a function of λ_X, α and λ_I being set. They show the neighborhood of the best clustering configuration for LCP ($\lambda_I = -10; \lambda_X = 10; \alpha = 0.5$) and for BFM ($\lambda_I = 10; \lambda_X = 10; \alpha = 0.5$).

For LCP, the best X-DER is obtained after 4 iterations of adaptation (14.4%), starting from a baseline X-DER of (18.1%), the *oracle* being of 11.3%. For the BFM collection, the best X-DER is of 13.6%, with a baseline of (15.7%) and an oracle of (12.6%). Figures show that the main DER reduction is obtained at the first iteration, but smaller improvements can be observed with further iterations, especially for the LCP corpus, where 2 or 3 iterations are necessary for the process to converge. In our experiments, we saw that for both collections, the optimal value for α is near 0.5. When looking at figure 5, we notice that even at an inaccurate cross-recording clustering threshold, the process can improve the *baseline*. This is true when being not too far from the best configuration, otherwise the process can slightly degrade the baseline DER (as seen for the BFM collection and $\lambda_X = -20$).

scoring	λ_I	λ_X	LCP_{tgt}		BFM_{tgt}	
			I-DER	X-DER	I-DER	X-DER
cosine	-10	0	8.6	**19.5**	13.6	23.8
	-40	-20	10.1	24.4	14.4	**22.2**
$WCCN + cos$	-30	-30	9.6	**19.7**	11.9	24.2
	-50	-40	9.8	23.0	13.3	**17.6**
	-50	-30	9.9	20.6	13.0	19.0
$PLDA$	-10	10	8.3	**18.1**	11.3	24.5
	10	10	10.0	19.1	10.6	**15.7**

Table II: Diarization results of the *baseline i-vector* based system, for cosine, WCCN and PLDA scoring.

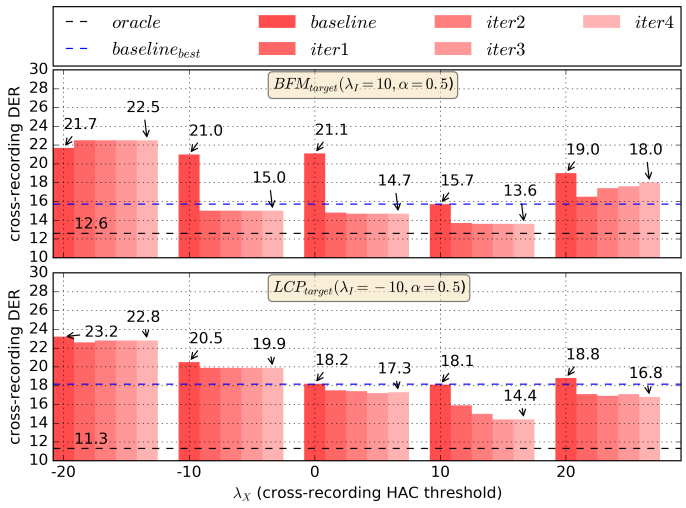

Figure 5. Cross-recording DER for both target corpora, for iterations 0 (baseline) to 4, as a function of λ_x, using PLDA for scoring and adaptation.

Table III shows the results for various configurations, including WCCN adaptation. The *dedicated* (*dedi*) configuration corresponds to a set of parameters $(\lambda_I, \lambda_X, \alpha)$ which are optimal for each couple $(show, scoring)$, while the *common* configuration corresponds to a set of suboptimal parameters, but works best for both shows, in average.

scoring	WCCN		PLDA	
show	LCP	BFM	LCP	BFM
$baseline_{dedi}$	19.7	17.6	18.1	15.7
$baseline_{common}$	20.6	19.0	19.1	15.7
$oracle_{best}$	14.8	13.4	11.2	12.5
$adapted_{dedi_{LCP}}$	**18.0**	15.4	**14.4**	15.1
$adapted_{dedi_{BFM}}$	19.8	**15.0**	14.9	**13.4**
$adapted_{common}$	**18.0**	15.4	14.9	**13.4**

Table III: Summary of the iterative adaptation results, on the complete collections, in terms of X-DER.

Table shows that whatever the chosen scoring method, iterative adaptation allows to improve the *baseline*, whatever the chosen *baseline* configuration. The adaptation configuration can be either the *dedicated* or the *common* one. However, when the adaptation configuration is *dedicated* to the BFM collection, the same configuration does not improve the best *baseline* DER of the LCP collection, using WCCN normalization and cosine scoring. We note that the best adaptation results are obtained using a configuration that does not give the best *baseline* DER. For example, we saw in figure 4 that a X-DER of 13.6% could be obtained for the BFM corpus after adapting on top of its best *baseline* configuration, but we found out that another configuration could give even better results after adapting, to achieve a DER of 13.4%.

D. Scalability

In previous section, we showed that iterative Weighted Likelihood Domain Adaptation could optimize diarization performances on two different collections. The experiments were performed on collections of 42 and 45 episodes, using an arbitrary adaptation parameter α. In this section, we want to focus on scalability. If a value of 0.5 for α seems to work for around 40 episodes, we need to verify what happens when the number of episodes is lower. A low number of episodes means a low number of recurring speakers, whose contribution to WCCN or PLDA parameters estimation might be bad. It might be better to give more weight to a high number of *out-domain* speakers than to a limited number of in-domain speakers for modeling.

We decide to repeat the previous experiment, but on variable size collections. For each of the two $target$ collections, with episodes sorted in chronological order, we define $N-1$ subsets, each subset k containing the first k episodes of the collection, for $k \in [2, N]$. For each subset, the cross-recording diarization is evaluated, after the *baseline* and 2 iterations of adaptation, for both scoring and normalization methods. The data used for adaptation is from the subset only. Experiments are run independently on each subset (*results obtained for an episode in a subset can change for the same episode in another subset*), each clustering is performed on the bag of *i-vectors* obtained after the within-recording diarization. The results obtained on a k^{th} subset have no influence on the results of the $(k+1)^{th}$.

The experiments depend on the same three parameters: λ_I, λ_X, the HAC thresholds, and α, the adaptation parameter. An exhaustive search is performed for $\alpha \in [0, 1[$. From the previous section, when working on the full $target$ collections, we found out that the optimal value for α was close to 0.5 for both collections.

The best performing configurations for PLDA adaptation are presented in figures 6 and 7. Each figure is split into two graphs.

The top one represents the weighted average X-DER for subsets 5 to N, for iterations 0 (*baseline*) to 2. Under 5 episodes, there are no recurring speakers to adapt with. The weighted average X-DER equals to the average X-DER r_k of each subset k, weighted by its total duration d_k, the formula is presented in equation 19. This allows to summarize the performances over all subsets. On the bottom graph, the bar chart represents the *baseline* I-DER for each episode.

$$waDER = \frac{\sum_{k=5}^{N} d_k r_k}{\sum_{k=5}^{N} d_k} \tag{19}$$

Figures 6 and 7 show that both iterations improve the weighted average X-DER, α being set to 0.5 and 0.3, while the waDER decreases from 15.2% to 13.3% and from 16.3% to 13.9%, respectively. As seen in section VII-C, the second iteration does not give as much improvement for the BFM collection as for the LCP one. When looking at the smallest subsets, we can see that the process improves the *baseline* starting from the 8-th subset. When the number of episodes is that small, the corresponding number of clusters used for adaptation is 4 for BFM and 4 for LCP (not presented in the figures). Even with such a low number of clusters, the use of a relatively high α seems not to be a problem.

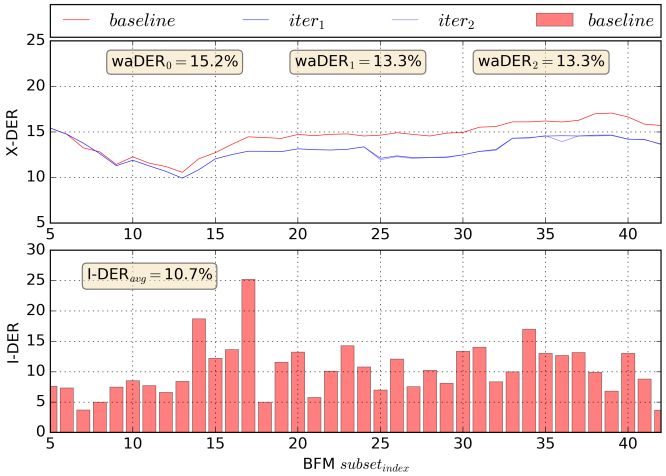

Figure 6. Cross-recording DER of the BFM corpus subsets, for iteration 0 to 2, using PLDA for scoring and adaptation. Experiment parameters are $(\lambda_I = 10, \lambda_X = 10, \alpha = 0.5)$.

As for WCCN scoring (not presented in the figures), the experiment is successful for BFM, with a weighted average *baseline* X-DER of 16.2%, decreasing to 14.0% through adaptation. However, the best result obtained for LCP was a quasi constant from 18.5% to 18.4%. When looking precisely at the experiment, we noticed that adaptation was effective starting from the 20-th subset, but below that index, adaptation could degrade the *baseline* X-DER of some subsets up to 50% in relative.

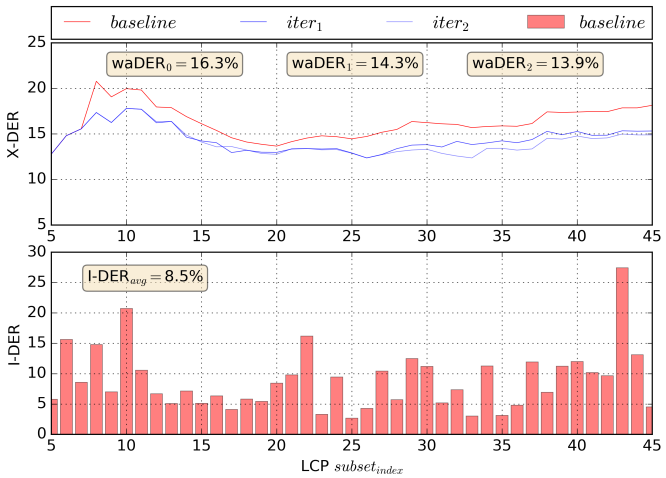

Figure 7. Cross-recording DER of the LCP corpus subsets, for iteration 0 to 2, using PLDA for scoring and adaptation. Experiment parameters are $(\lambda_I = -10, \lambda_X = 10, \alpha = 0.3)$.

E. Optimality of adaptation

Experiments of the previous section showed that the choice of a fixed value for α was not effective for the LCP collection, using WCCN/cosine scoring. Depending on the subset size, the optimal α value might vary. For each set of $(\lambda_I, \lambda_X, scoring)$, we repeat the previous experiment 10 times, but on randomized collections: the episodes order is shuffled in a different way for each experiment. With α ranging from 0 to 0.9 with a 0.1 step, we are able to observe what are the optimal values for α, for each subset, in average. Results are presented in figures 8 and 9, for PLDA and WCCN scoring, respectively.

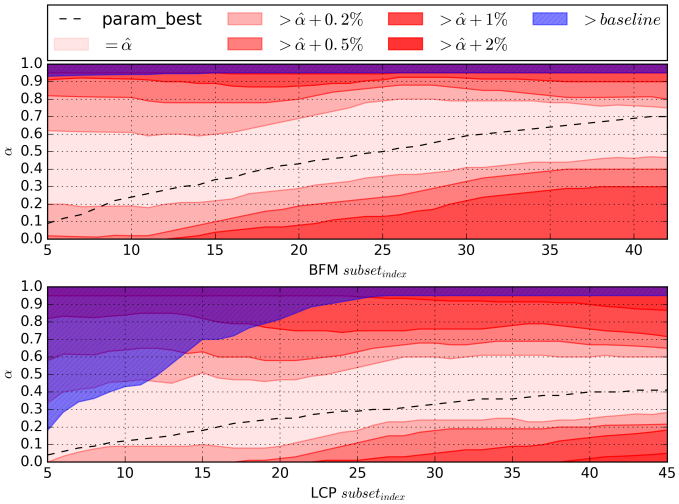

Figure 8. Isomap of X-DER, relative to the optimal α value, for the PLDA-based adaptation experiment. The map depends on the subset indexes and the values of α. BFM configuration is $(\lambda_I = 10, \lambda_X = 10)$, while for LCP it is $(\lambda_I = -10, \lambda_X = 10)$. Boundaries are smoothed using a sliding averaging window of 5 consecutive episodes (ie subset indexes).

Within each figure, the top graph is for the BFM subsets,

while the bottom one is for LCP. On each graph, various areas are presented, depending on the subset indexes and the values of α. The clearer area corresponds to the α values which give the best adapted X-DER, we note it $\hat{\alpha}$. Each boundary corresponds to the α value where the adapted X-DER reaches a limit, relative the best X-DER. For example, the second clearer area corresponds to the range of α values where the adapted X-DER is between $\hat{\alpha} + 0.2\%$ and $\hat{\alpha} + 0.5\%$, in absolute. We also display a light blue hashed area, which corresponds to an area where the adapted X-DER is above the $baseline$. We note that this *forbidden* area only appears on the upper part of the graphs, since the bottom limit *is* the $baseline$ ($\alpha = 0$ is equivalent to not adapting). When α increases from zero, it usually improves the $baseline$ X-DER up to a point where the adaptation degrades it. In conclusion, every area except the blue hashed one is an area where adaptation improves or equals the $baseline$.

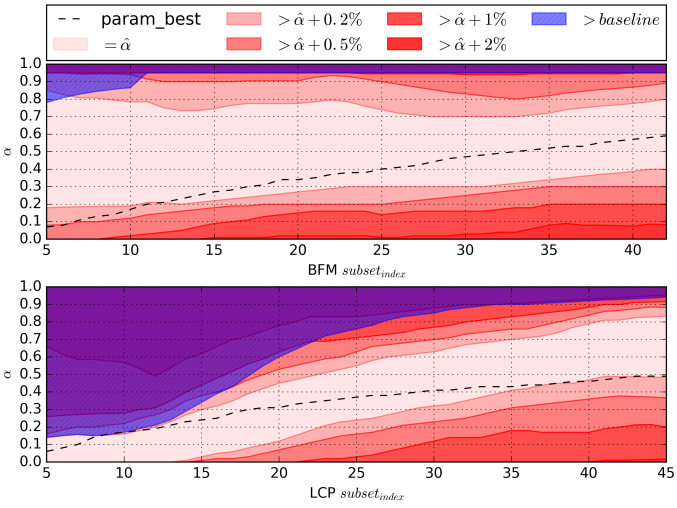

Figure 9. Isomap of X-DER, relative to the optimal α value, for the WCCN/cosine-based adaptation experiment. The map depends on the subset index and the value of α. BFM configuration is $(\lambda_I = -30, \lambda_X = -40)$, while for LCP it is $(\lambda_I = -40, \lambda_X = -30)$. Boundaries are smoothed using a sliding averaging window of 5 consecutive episodes (ie subset indexes).

The first thing we note on all graphs is that the optimal $\hat{\alpha}$ value tends to increase as the collection grows, but with various trends, depending on the collection and the scoring method. For PLDA scoring, figures look similar, with an optimality range of 0.2 to 0.4 for the $[\hat{\alpha}, \hat{\alpha} + 0.2\%]$ interval.

As for WCCN scoring, for the BFM collection, we see that the optimality range is very large at the beginning and tends to narrow down to 0.4 for the full collection, while for the LCP collection, the optimality range is very narrow for small subsets, the optimal average value being very close to 0 (little to no adaptation), and spreads to 0.3 for the full collection. As seen in the previous section, for the LCP-WCCN/cosine experiment, even for small values of α, adaptation degrades the $baseline$ DER if the size of the subset is too small.

F. Parameterization

In this section, the fixed adaptation weight α is replaced by a parametric version, in order to account for the size of the collection the system is adapted on. The parametric formula is presented in section V-C and depends on the recurring speakers found by the previous linking step. The same randomized experiments of the previous section VII-E are conducted, but using the parametric formula instead of a fixed adaptation parameter. An exhaustive search is run for $r \in \{2, 4, 8, 16, 32, 64, 128\}$ and $p \in \{1, 2, 3\}$. For each $(scoring, show)$ couple, the best waDERs, averaged over the 10 randomized collections, are compared for the fixed and parametric approach. Results are presented in table IV, and parametric α values corresponding to the best results are displayed with a dashed black line on figures 8 and 9.

scoring	WCCN		PLDA	
show	LCP	BFM	LCP	BFM
$baseline$	19.6	18.1	17.2	16.3
α	0.5	0.4	0.4	0.5
$adapted_{fix}$	18.6	14.2	15.1	13.7
r	32	16	64	16
p	1	1	1	1
$adapted_{param}$	18.3	14.3	15.0	13.8

Table IV: Summary of the iterative adaptation results, comparing a fixed adaptation parameter and the proposed parametric adaptation formula, average in terms of waDER over all randomized subsets of each collection. $adapted$ results are obtained after two iterations of adaptation.

When looking at figures 8 and 9, we see that the parametric curves follow the optimality area, which corresponds to our expectations. For the BFM collection, due to the chosen formula, the curves have to cross sub-optimality areas when the subsets contain a low number of episodes. Thus, the parametric approach performs slightly worse than the fixed one: when looking at the graphs, one can easily draw an horizontal line which stays in the optimality area for all subsets. As for the LCP collection, no horizontal curve can match the optimality area and the parametric approach outperforms the fixed one, as confirmed in table IV.

G. Analysis & Discussions

This section is dedicated to the analysis of the adaptation results. As we know the speaker composition of our collections, we want to study what really happens in terms of cluster evolution and speaker accuracy through iterative adaptation: is the DER reduction due to a better accuracy on the little speaking speakers or to the ability to better cluster recurring speakers ?

We selected one experiment from section VII-C: iterative adaptation on the full LCP collection, using PLDA scoring and parameters $(\lambda_I = -10, \lambda_X = 10; \alpha = 0.5)$, and represented it in terms of correct speech attribution from one iteration to another: it aims to visualize the contributions to Diarization

Error Rate variations. The first column displays the speakers for which speech time attribution varies from one iteration to another, the second columns shows the actual number of recordings the speakers speak in, the third is their role in the show (guest or journalist), and the four right columns describe the speech attribution variations after each iteration of adaptation. Each cell's color indicates how much speaker time was retrieved (blue) or lost (red) at each iteration (the amount of speech gained or lost being written in the cell), while the height of each line is proportional to the logarithm of the total speech time of the corresponding speaker in the collection. A way to read the figure is for example: the second line is about Germain Andrieux, which is a journalist and appears in 11 episodes of the show. After the first iteration of adaptation, 198 seconds of true speech were retrieved, while 151 additional seconds were retrieved after the second iteration, for a total of 349 seconds. The sum of a column corresponds to the total variation of Speaker Error for an iteration.

The first observation we make is that there are three kind of speakers for which changes happen. Those who gain correct speech time in multiple iterations, those who gain or lose in only one iteration and those who gain then lose (or the other way) the same amount of speech time between two iterations.

When looking at the recurrence of the speakers (the amount of recordings they speak in), we note that even some one-time speakers are affected by the adaptation process, mainly during the first adaptation. We also notice that most of the speakers who keep retrieving speech through iterations are recurring speakers. This means adaptation actually provides better modeling of the recurring speakers variabilities, which was our initial motivation for iterating.

Finally we observe that the main contribution to the DER decrease is due to the recurring speakers. In the LCP collection, we know that around 80% of the total speech time is spoken by recurring speakers.

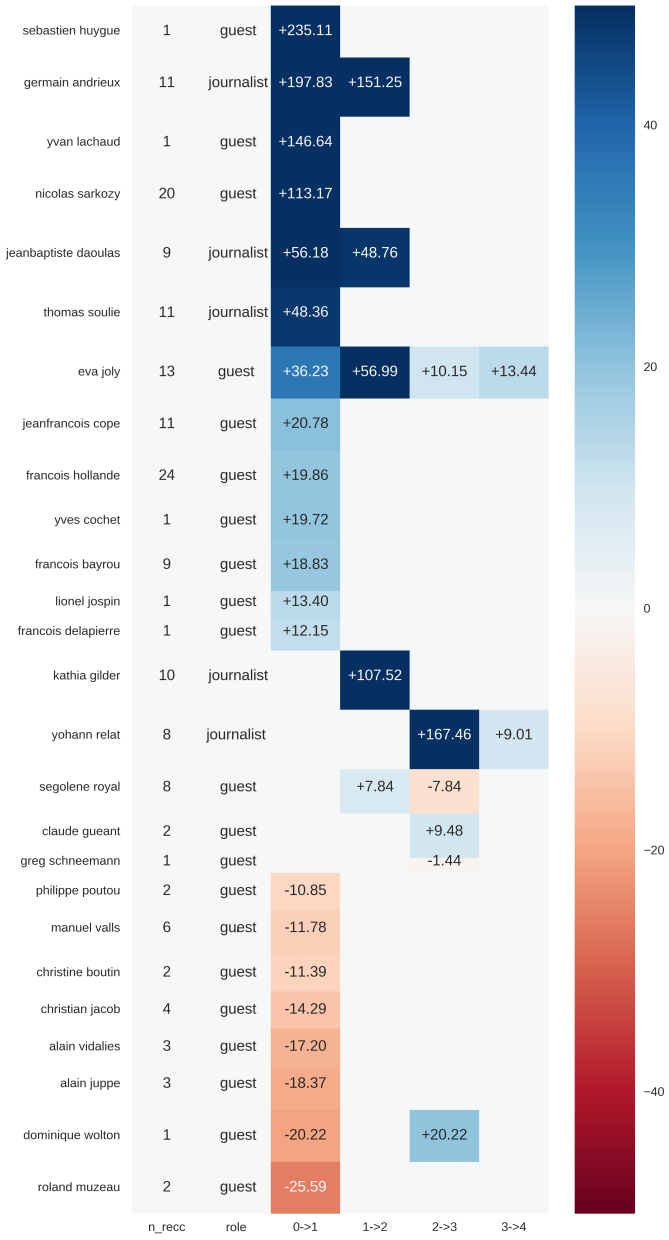

Figure 10. Analysis of the evolution of correct speech attribution during the iterative adaptation process. Experiment is on the full LCP collection, with $(\lambda_I = -10, \lambda_X = 10, \alpha = 0.5)$ and PLDA scoring. From $iter_0$ $(baseline)$ to $iter_4$ experiment, the X-DER varies from 18.1% down to 14.4%.

VIII. Conclusion

In this paper, we proposed an iterative adaptation framework for speaker diarization and linking of multimedia collections. It proved to be effective for two types of scoring: cosine with WCCN and PLDA scoring, for variable collection sizes. We observed a convergence in terms of cross-recording Diarization Error Rate after 2 or 3 iterations. Due to the observed optimality ranges of the adaptation parameter, on variable collection sizes, we proposed a parametric method, depending on the estimated number of recurring speakers of each collection, to compute the adaptation parameter.

Results analysis showed that the main contribution to the Diarization Error Rate decrease through the adaptation iterations is due to the gains obtained on the recurring speakers. This proves how good adaptation improves the modeling of within- and between-speaker variabilities, for normalization (WCCN) or scoring (PLDA). The proposed method is well suited for relatively big collections, which include some recurring speakers. The main advantage of the process is that it only requires the audio one time per recording, since the whole linking and adaptation process is based on *i-vectors*, thus computation requirements are rather low.

Our experiments were run on collections of around 40 recordings, which corresponds to the annual number of episodes of a weekly show in France. Should the method be applied to much bigger collections, such as daily shows or videos sharing platforms, a particular attention should be paid at the linking process, which is $O(n^2)$. One way to overcome the issue could be incremental linking, which forbids to change the linking of past recordings [48].

REFERENCES

[1] H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, and M. Guillemot, "Processing and Linking Audio Events in Large Multimedia Archives: The EU inEvent Project," in Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 2013.

[2] M. Ferràs and H. Bourlard, "Speaker Diarization and Linking of Large Corpora," in Proceedings of IEEE Workshop on Spoken Language Technology, Miami, Florida (USA), December 2012.

[3] H. Ghaemmaghami, D. Dean, and S. Sridha, "Speaker Attribution of Australian Broadcast News Data," in Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 2013.

[4] D. A. Leeuwen, "Speaker Linking in Large Data Sets," in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 2010.

[5] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization," in Proceedings of Interspeech, Florence, Italy, August 2011.

[6] Q. Yang, Q. Jin, and T. Schultz, "Investigation of Cross-show Speaker Diarization," in Proceedings of Interspeech, Florence, Italy, August 2011.

[7] M. Ferras, S. Madikeri, and H. Bourlard, "Speaker diarization and linking of meeting data," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, no. 99, pp. 1–1, 2016.

[8] H. Ghaemmaghami, D. Dean, and S. Sridharan, "A cluster-voting approach for speaker diarization and linking of australian broadcast news recordings," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 4829–4833.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. 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.

[10] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4047–4051.

[11] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in Proceedings of Odyssey: The Speaker and Language Recognition Workshop, 2014.

[12] S. H. Shum, D. A. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," 2014.

[13] G. L. Lan, D. Charlet, A. Larcher, and S. Meignier, "Iterative plda adaptation for speaker diarization," in INTERSPEECH, 2016.

[14] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in Speaker Odyssey Workshop, 2010.

[15] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," in IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1991, pp. 873–876.

[16] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," in Proceedings of the DARPA Broadcast News Workshop, 1997, p. 11.

[17] P. Delacourt, D. Kryze, and C. J. Wellekens, "Detection of Speaker Changes in an Audio Document," in Sixth European Conference on Speech Communication and Technology, 1999.

[18] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multi-stage speaker diarization of broadcast news," IEEE Transactions on Speech and Audio Processing, vol. 14, no. 5, pp. 1505–1512, Feb. 2006.

[19] G. Schwarz et al., "Estimating the Dimension of a Model," The annals of statistics, vol. 6, no. 2, pp. 461–464, 1978.

[20] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality assessment for speaker diarization and its application in speaker characterization," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 4, pp. 816–827, 2013.

[21] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "I-vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization," in Proceedings of Interspeech, Portland, Oregon, USA, September 2012.

[22] D. Reynolds, E. Singer, B. Carlson, G. O'Leary, J. Mc Laughlin, and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in Proceedings of International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[23] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," in Proc. DARPA Broadcast News Transcription and Understanding Workshop. Virginia, USA, 1998, p. 8.

[24] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. McLaughlin, and M. A. Zissman, "Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics," in International Conference on Spoken Language Processing (ICSLP), 1998.

[25] M.-H. Siu, G. Yu, and H. Gish, "An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers," in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2. IEEE, 1992, pp. 189–192.

[26] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering Speakers by Their Voices," in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2. IEEE, 1998, pp. 757–760.

[27] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in Proceedings of Interspeech, Florence, Italy, 2011.

[28] S. H. Shum, W. M. Campbell, and D. A. Reynolds, "Large-scale Community Detection on Speaker Content Graphs," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013, pp. 7716–7720.

[29] M. Rouvier and S. Meignier, "A Global Optimization Framework for Speaker Diarization," in Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop, Singapore, 2012.

[30] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2015–2028, 2013.

[31] G. Doddington, "The effect of target/non-target age difference on speaker recognition performance," in Odyssey 2012-The Speaker and Language Recognition Workshop, 2012.

[32] Y. Matveev, "The problem of voice template aging in speaker recognition systems," in International Conference on Speech and Computer. Springer, 2013, pp. 345–353.

[33] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "Segmentation et Regroupement en Locuteurs d'une collection de documents audio," in Proceedings of 29e Journées d'Études sur la Parole (JEP'12), Grenoble, France, June 2012.

[34] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification," in Thèse de doctorat. Georgia Institute of Technology, 1992.

[35] S. J. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1–8.

[36] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in Interspeech, vol. 2011, 2011, pp. 249–252.

[37] O. Glembek, J. Ma, P. Matějka, B. Zhang, O. Plchot, L. Bürget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014, pp. 4032–4036.

[38] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving out-domain plda speaker verification using unsupervised inter-dataset variability compensation approach," in Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015). IEEE, 2015, pp. 4654–4658.

[39] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Channel adaptation of plda for text-independent speaker verification," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 5251–5255.

[40] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4002–4006.

[41] E. Khoury, L. El Shaffey, M. Ferras, and S. Marcel, "Hierarchical speaker clustering methods for the nist i-vector challenge," in Speaker Odyssey Workshop, 2014.

[42] A. Larcher, K. Aik Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[43] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital signal processing, vol. 10, no. 1, pp. 19–41, 2000.
[44] O. Galibert and J. Kahn, "The first official repere evaluation," in Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM), 2013.
[45] O. Galibert, J. Leixa, A. Gilles, K. Choukri, and G. Gravier, "The ETAPE Speech Processing Evaluation," in Conference on Language Resources and Evaluation, Reykyavik, Iceland, May 2014.
[46] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in Proceedings of Interspeech, Brighton, Royaume Uni, Sept 2009.
[47] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech." in INTERSPEECH, 2013, pp. 1131–1134.
[48] G. Dupuy, , S. Meignier, and Y. Estève, "Is incremental cross-show speaker diarization efficient to process large volumes of data?" in Proceedings of Interspeech, Singapore, Sept 2014.

PLACE PHOTO HERE

Gaël Le Lan Biography text here.

PLACE PHOTO HERE

Delphine Charlet Biography text here.

PLACE PHOTO HERE

Anthony Larcher Biography text here.

PLACE PHOTO HERE

Sylvain Meignier Biography text here.