First investigations on self trained speaker diarization

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

First investigations on self trained speaker diarization

Gaël Le Lan[1,2]*, Sylvain Meignier*[1]*, Delphine Charlet*[2]*, Anthony Larcher*[1]

[1]LIUM, University of Le Mans, France
first.lastname@lium.univ-lemans.fr
[2]Orange Labs, France
first.lastname@orange.com

Abstract

This paper investigates self trained cross-show speaker diarization applied to collections of French TV archives, based on an i-vector/PLDA framework. The parameters used for i-vectors extraction and PLDA scoring are trained in a unsupervised way, using the data of the collection itself. Performances are compared, using combinations of target data and external data for training. The experimental results on two distinct target corpora show that using data from the corpora themselves to perform unsupervised iterative training and domain adaptation of PLDA parameters can improve an existing system, trained on external annotated data. Such results indicate that performing speaker indexation on small collections of unlabeled audio archives should only rely on the availability of a sufficient external corpus, which can be specifically adapted to every target collection. We show that a minimum collection size is required to exclude the use of such an external bootstrap.

1. Introduction

Speaker diarization aims at uniquely label speakers across one or more audio recordings, without a priori knowledge about the speakers. The increasing amount of multimedia data, which needs to be indexed, requires an effective diarization framework. Such data can be processed as collections, in a task called cross-show speaker diarization a.k.a. speaker attribution. Cross-show diarization is a global task which consists in processing a collection of raw recordings (unsegmented, containing multiple speakers) to extract a representation of "who speaks when". A same speaker should always be labeled in the same way. This task is usually decomposed in two steps : the within-recording diarization, which is about segmenting and clustering speaker occurences within a same recording, and across-recording speaker linking, which aims at regrouping the within-recording clusters across the whole collection (with or without allowing further within-recording merges). Other implementations are possible, where all the recordings can be concatenated in an artificial within-recording diarization problem [1], or where a cluster boundaries correction step can be added. The two-pass approach is the mostly used, since it is the easiest and that it allows to process large collections of data.

This cross-show diarization task can be applied to collections of radio and TV archives [2, 3, 4, 1, 5], phone recordings [6, 7, 8, 9] or meeting recordings [10].

The state-of-the-art speaker modeling is based on i-vector/PLDA framework, requiring speaker labeled data, indicating the speakers identity and time stamps. The key step of such frameworks is the between-speaker variability modeling, thus training corpora must include multiple occurrences of a same speaker in various acoustic conditions. Manually speaker labeled data are expensive to produce and not always available for target data, leading to two different strategies : unsupervised training on target data, or supervised training on external annotated training data, resulting in a conditions mismatch between the train and target data. Solutions to the mismatch problem have already been proposed in the context of speaker verification, based on unsupervised domain adaptation [7]. In the context of unsupervised PLDA training, the idea of iterative training [11] has also been investigated to improve the modeling quality. In both papers, training was performed on mono-speaker unlabeled data. In [3], it was shown that an unsupervised diarization system, trained on multi-speaker unsegmented data, could perform as good as a supervised one, indicating that annotations or labels are not mandatory for training. In this paper we want to address speaker diarization when no training corpus is available, and the only available data for training is the unannotated test data a.k.a. target data itself.

We investigate the task of self trained speaker diarization, which only uses the data from the target collection itself, without external training nor adaptation data. The idea is to propose a system which can perform multimedia clustering, requiring the minimal amount of a priori knowledge. Thus, we investigate different variations of the well known i-vector/PLDA framework through the perspective of training with or without external data. Combinations of labeled and unlabeled data for modeling are compared for the task of cross-show speaker diarization.

Subsequent sections are organized as follows: first, we describe the diarization framework and define the perimeter of the self trained diarization task. Then we describe the data used for the experiments and conclude with a discussion about the performances of the self trained system and the possible improvements.

2. Diarization Framework

Figure 1 describes the two-pass diarization process, which is detailed below. This process, which is widely described in [3], includes a i-vector/PLDA system trained in a supervised or unsupervised way. It is decomposed in a within-recording diarization followed by a cross-recording speaker linking step.

2.1. Within-recording speaker diarization

After the MFCC extraction and Viterbi-based speech detection, a standard BIC segmentation and clustering is applied. Each BIC cluster is considered pure and representing a single speaker. However, since several clusters can be related to the same speaker, another clustering step is required.

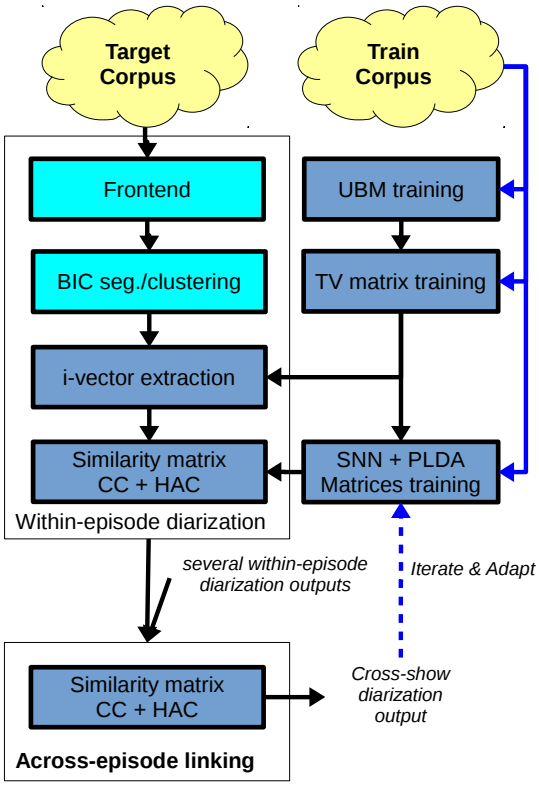

Figure 1: Overview of the diarization framework for supervised (plain blue lines) and adapted (both plain and dashed lines) training.

The channel contribution is removed with a zero mean and unit variance normalization computed for each cluster. Each cluster is modeled by an i-vector, normalized by Spherical Nuisance Normalization [12]. PLDA is used to calculate log likelihood ratio for pairs of i-vectors [13], which we call PLDA score in the following of the paper. The term PLDA matrix indicates the matrix of the PLDA scores.

For each file, a PLDA matrix is computed, which can be seen as a connected graph. A complete-linkage Hierarchical Agglomerative Clustering (HAC) on subsets of this graph (connected components (CC) acquired by keeping the links above a defined threshold) is applied. This approach is similar to the one presented in [14], where the CC+HAC combination replaces the ILP clustering[1].

The dimension of the feature vectors is 39: 13 MFCCs including c_0 coefficient supplemented with the first and second order derivatives. The GMM is composed of 256 Gaussians with diagonal covariance matrix, the dimension of the i-vectors is 200 and PLDA eigenvoice matrix has a dimension of 100 with no eigenchannel matrix. I-vector and PLDA parameters estimation are computed using the SIDEKIT toolkit [15].

[1]ILP clustering is a clustering method formulated as an Integer Linear Programming (ILP) problem

2.2. Cross-Show Speaker linking

Once speaker diarization has been applied to each show separately, the collection of shows is considered as a global speaker linking problem and the within-recording clustering process is reapplied on newly formed clusters. Each within-recording cluster is modeled with an i-vector, a PLDA matrix is computed and CC+HAC is applied in a global way.

3. Self trained diarization

In [3], we showed that unsupervised model training (e.g. without annotations) for speaker diarization can perform as good as supervised model training for large training corpora.

Figure 1 represents the overview of the diarization framework, including the possible strategies for the training process. Three strategies are presented : one is represented with the blue plain lines only, another with the blue dashed lines only and the last one with both lines.

As presented in figure 1 we propose to train an unsupervised i-vector/PLDA framework with the target data (blue dashed lines only), in an approach similar to [3], but where the target corpus itself takes the role of training data. We compare this framework with a supervised one, trained on annotated external data (blue plain lines only), which is our baseline, and with another supervised system, trained on the test data but using annotations, which is our oracle. Finally, we propose to adapt the baseline system with the target data itself, and to apply an iterative PLDA training framework (both blue dashed and plain lines).

According to the diarization framework described in section 2.1, required models for diarization are a Universal Background Model (UBM), a Total Variability matrix (TV) and a PLDA model. When data are labeled by speaker, supervised modeling can be performed. However, when we want to train a system on target data itself, the information about the speakers need to be extracted in an unsupervised way, as detailed hereafter.

3.1. Universal Background Model

The Universal Background Model (UBM) is required for the Baum-Welch statistics needed during i-vector extraction. It needs to be trained on clean speech segments with few background noise that are representative of the test set in terms of duration, gender and channel. Thus, the output of a GMM based speech / non-speech detector on the target data can be used to train the UBM, which contains 256 Gaussians with diagonal covariance matrices.

3.2. Total Variability

The Total Variability (TV) matrix is required to extract i-vectors. It must be trained over segments containing a unique speaker. From our experience, we choose the BIC penalty coefficient so that resulting clusters are pure (they contain only one speaker). According to this observation, we consider each class produced by the BIC diarization as a segment representing a unique speaker and the total variability matrix is trained on those with a rank of 200.

3.3. PLDA

Spherical Nuisance Normalization requires the mean vector and the within-class co-variance matrix of the i-vectors, whereas PLDA parameters are estimated using normalized speaker la-

beled i-vectors. Both systems are designed to compensate for the inter-session speaker variability. In both cases, the data used for training must be labeled by speaker and session. Several sessions are needed per speaker to obtain robust models. To train our PLDA parameters, we only consider the speakers with occurrences in at least three different recordings, with a minimum amount of speech of 10s for each recordings.

Hence, performing an i-vector/cosine based (no i-vector normalization, cosine distance instead of PLDA likelihood) cross-show speaker diarization on the target corpus allows to automatically obtain speaker clusters, containing data from the same speaker and from different recordings. PLDA matrices can then be trained over the i-vectors extracted from the clusters obtained using an unsupervised cosine distance.

4. Experimental context

Contrastive models for diarization systems were trained on manually annotated corpus. In this corpus the speakers are identified by their first and last names, providing several sessions for a large set of speakers. About 200 hours of French broadcast news drawn from REPERE [16], ETAPE [17] and ESTER[18] 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. Speakers appearing in more than one episode of a corpus are called recurring (R.) speakers, as opposed to one-time (O.T.) speakers, who only speak in one episode. 88

4.1. Target corpora

In this paper, we define two target corpora built from the REPERE train and test corpora. The first one, named LCP_{target}, is the collection of all available episodes of the show $LCPInfo$, a French news broadcast show. The second target corpus, named BFM_{target}, is the collection of all available episodes of the TV news talk-show $BFMStory$. Those two corpora have been selected 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 1. Since both corpora are partially annotated, only details for annotated speakers are presented.

Corpus	LCP_{target}	BFM_{target}
Episodes	45	42
Episode duration	25m	60m
Evaluated (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 1: Composition of target corpora. Annotated speakers numbers are presented.

4.2. Complementary train corpus

The $train$ corpus, used for complementary experiments, is composed of 344 audio files from train and development corpora of the three previously cited campaigns, for a total of 200 hours of speech duration. For each show, all available episodes are taken, meaning many speakers appear in more than one episode. Some speakers also appear in different shows (politicians, for example). The shows selected as test corpora are not present in the $train$ corpus. The corpus contains 3888 unique speakers. Among those speakers, 391 meet our requirement for PLDA parameters estimation : they appear in at least three recordings, with a minimum speech time per recordings of 10s. Thus, this corpus is well suited for a i-vector PLDA system training.

5. Experiments

The metric used to measure performance in speaker diarization is the Diarization Error Rate (DER). DER was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker, using the best match between references and hypothesis speaker labels. The scoring tool [19] is employed for within-recording and across-recording speaker diarization. In this last case, a recurring speaker should always be labeled in the same way, in every recording that composes a collection. For DER computation, a boundaries error tolerance of 250ms is allowed, and the overlapping speech segments are included for evaluation.

5.1. Baseline and Oracle

For our $baseline$, a supervised i-vector/PLDA system was trained on the complementary train corpus. This system was used to perform speaker diarization on the two target corpora. It represents what is usually done when a new corpus without annotations needs to be processed, and a speaker annotated train corpus is available. We also trained another supervised system on each target corpora, using the annotations. This represents our $oracle$, what diarization performances could be achieved if the between-speaker variability of the target corpus was perfectly estimated.

Finally, an unsupervised system, using only the ressources from the target corpora, without any annotations, was trained. This system is i-vector/cosine based, since no information about recurring speakers is available. It is our target-centric baseline ($baseline_{target}$), when no use of external data is allowed. Results are presented in table 2.

Results show that without any a priori information, using only unsupervised training on the target corpora, the diarization performances are way worse than using only supervised training on external data (29.68% vs 17.72% and 27.62% vs 13.22% respectively), as no across-speaker variability is modeled.

Concerning Oracle experiments, training a supervised PLDA on the target corpus is not always possible: for the BFM_{target} corpus, PLDA training does not succeed, the EM algorithm does not converge, probably due to the low number of speakers available (35). On the contrary, when the target corpus contains enough recurring speaker, oracle results in table 2 show that using the a priori information about the target corpus for PLDA generation gives the best results with an across-recording DER of 10.87%).

5.2. Minimum requirements for PLDA training

Previous results showed that for some target corpora, there might not be enough data available in the target corpus to build an $oracle$ system. Figure 2 shows that a suitable PLDA-based system can be built for at least 37 episodes of the LCP_{target} corpus, containing 40 different recurring speakers, each appearing in 7.31 different episodes, in average. While BFM_{target} corpus contains 42 episodes, only 35 different speakers meet the requirements for PLDA training (at least sessions from three episodes per speaker), each appearing in 5.45 different episodes, in average.

These results show that a minimum amount of recurring speakers is required to try to build a self trained PLDA-based system, but defining exactly the minimal conditions required to estimate PLDA parameters is still under investigations. When the minimum amount condition is not met, the EM algorithm for parameters estimation does not converge. It is interesting to note that in the figure, for episodes 32 to 37, the number of recurring speakers stays at 40, while the DER decreases greatly. It means that for 32 episodes, the amount of speech per speaker is too low, and with the increasing amount of episodes, new sessions of those speakers appear and improve the PLDA parameters estimation.

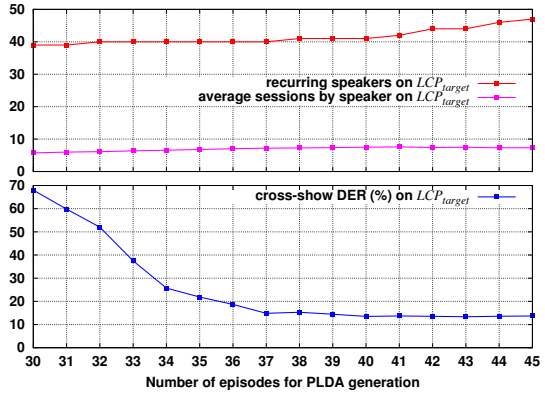

Figure 2: Metrics on the self trained supervised i-vector/PLDA clustering of LCP_{target} data, as a function of the number of episodes used for training. Episodes are selected in chronological order.

5.3. Unsupervised training

In the following of this paper, the use of provided annotations is forbidden for the two target corpora. In [3], we showed that training an unsupervised i-vector/PLDA based system on BIC diarization sessions could work as good as a supervised system based on manual annotations.

We trained an unsupervised UBM and TV matrix, based on BIC diarization classes instead of manual annotations. On top of that background, we trained a supervised PLDA system with data from the complementary train corpus ($iter_{0-a}$), using the provided annotations. Results are presented in table 2 and indicate that using $target_{unsup}$ UBM and TV matrix, with PLDA trained on $train_{sup}$ data, performs better than using an external TV matrix : for both target corpora, we observe a decrease of the across-recording DER (from 17.72% to 16.58% and from 13.22% to 12.60%).

On top of that system, we decided to use the output speaker classes of experiment $iter_{0-a}$ to train a $target_{unsup}$ PLDA based system, trained on target-data only ($iter_{0-b}$). Unfortunately the PLDA parameters estimation failed, the EM algorithm not converging, probably due to the fact that these classes are unpure (e.g. can refer to different speakers), not allowing to gather sufficient statistics for the EM algorithm to converge.

The data composition of the supervised and unsupervised UBM, TV, and PLDA for both corpora is presented in table 3. Due to the fact that corpora are only partially annotated and that supervised training is based on annotations, the data quantity for UBM and TV training if twice less for supervised training than for unsupervised training. The same difference is to be noted regarding the composition of supervised and unsupervised PLDA. Since both corpora are partially annotated, unsupervised training is based on much more speaker classes. Those speaker classes are extracted from the whole corpora, while supervised training is based on annotated speaker only.

5.4. Domain adaptation

Since the output speaker classes of experiment $iter_{0-a}$ are not sufficient to train a PLDA-based system and iterate, they are added to the manually annotated speaker classes of the train corpus ($iter_1$). This way, we see that the use of output speaker classes of $iter_{0-a}$ enhance the overall performance : from 16.58% to 15.60% and from 12.60% to 11.38% respectively. The recurring speakers found after experiment $iter_{0-a}$ allow to enhance the across-speaker variability modeling.

In [7], the domain adaptation problem was addressed with a combination of labeled external data and unlabeled data similar to the target data, but not the actual target data. While the cited work consists in introducing a weighting variable between the external and internal data for PLDA parameters estimation, our approach consists in concatenating data from both domains, which is equivalent to setting the weighting parameter to the relative number of recurring speaker classes between both domains. Our results show similar conclusions with the paper, with an improvement of performances when introducing adaptation, despite the fact our system is based on multi-speaker unsegmented data instead of mono-speaker i-vectors.

5.5. Iteration

Our experiments show that if we keep iterating after experiment $iter_1$, the system does not improve itself, meaning the impurity of some output speaker classes pollutes the parameters estimation.

The idea of unsupervised iterative PLDA training was addressed in [11], but it was focused on the external train corpus and bootstrapped with a first i-vector/cosine based clustering. In our work, PLDA parameters are partially estimated on the target data, and bootstrapped with an i-vector/PLDA based clustering. While the cited paper showed that iterative training could improve the performances of speaker verification, our approach did not show the same results.

Since our data are raw unsegmented multi-speaker recordings, the i-vectors used for iterative training might not be very accurate in terms of speakers. Better results might be achieved by modifying the across-recording speaker linking threshold. A stronger threshold would improve the purity of speaker clusters, allowing a better PLDA modeling at each iteration.

Experiment	UBM + TV training corpus	SNN + PLDA training corpus	LCP_{target} WR DER (%)	LCP_{target} AR DER (%)	BFM_{target} WR DER (%)	BFM_{target} AR DER (%)
$baseline$	$train_{sup}$	$train_{sup}$	7.77	17.72	9.92	13.22
$oracle$	$target_{sup}$	$target_{sup}$	6.68	10.87	X	X
$baseline_{target}$	$target_{unsup}$	$none$ (cosine-based)	7.04	29.68	12.46	27.62
$iter_{0-a}$	$target_{unsup}$	$train_{sup}$	7.80	16.58	8.30	12.60
$iter_{0-b}$	$target_{unsup}$	$target_{unsup}$	X	X	X	X
$iter_1$	$target_{unsup}$	$train_{sup} + target_{unsup}$	7.67	15.60	8.58	11.38
$iter_2$	$target_{unsup}$	$train_{sup} + target_{unsup}$	7.51	15.52	8.79	11.56
$iter_3$	$target_{unsup}$	$train_{sup} + target_{unsup}$	7.55	15.98	8.79	11.56

Table 2: within-recording (WR) and across-recording (AR) DER on target corpora for all experiments

	LCP_{target} sup	LCP_{target} $unsup$	BFM_{target} sup	BFM_{target} $unsup$
Data used for UBM/TV training	9h56m	19h17m	19h50m	39h09m
Average session duration for UBM/TV training	1m08s	4m23s	1m58s	4m10s
Number of speaker classes for PLDA training	47	130	35	190
Average number of sessions by speaker class	7.31	5.25	5.45	4.34
Average session duration	1m10s	1m25s	2m50s	2m07

Table 3: Composition of data used for UBM, TV and PLDA training, for both corpora

6. Conclusion

In this paper, we proposed a cross-show diarization framework based on a self training i-vector/PLDA strategy, in order to process relatively small collections of multi-speaker unsegmented TV archives. While previous work showed that unsupervised training on such data could be achieved, the small size of the target corpora is not sufficient for self training. The use of an external train corpus is required for bootstrapping.

We successfully applied a domain adaptation technique, which proved to be effective for speaker verification on mono-speaker data, to improve the baseline diarization system, using unlabeled data from the target collection itself. After a first diarization iteration, results showed a decrease in terms of DER for both target corpora.

Further work will be dedicated to the improvement of the training framework, introducing weighting variability in the domain adaptation parameters between the external train data and target data. We also plan to focus on the iterative aspect of the training procedure, since the literature showed that some improvements are achievable.

7. References

[1] Viet-Anh Tran, Viet Bac Le, Claude Barras, and Lori Lamel, "Comparing multi-stage approaches for cross-show speaker diarization.," in *INTERSPEECH*, 2011, number 1, pp. 1053–1056.

[2] Grégor Dupuy, , Sylvain Meignier, and Yannick Estève, "Is incremental cross-show speaker diarization efficient to process large volumes of data?," in *Proceedings of Interspeech*, Singapore, Sept 2014.

[3] Gael Le Lan, Sylvain Meignier, Delphine Charlet, and Paul Deléglise, "Speaker diarization with unsupervised training framework," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[4] Qian Yang, Qin Jin, and Tanja Schultz, "Investigation of cross-show speaker diarization.," in *INTERSPEECH*, 2011, pp. 2925–2928.

[5] David A Van Leeuwen, "Speaker linking in large data sets," Proc. Odyssey 2010.

[6] Stephen H Shum, William M Campbell, Douglas Reynolds, et al., "Large-scale community detection on speaker content graphs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7716–7720.

[7] Stephen H. Shum, Douglas A. Reynolds, Daniel Garcia-romero, and Alan Mccree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," Proc. Odyssey 2014.

[8] Zahi N Karam and William M Campbell, "Graph embedding for speaker recognition," in *Graph Embedding for Pattern Analysis*, pp. 229–260. Springer, 2013.

[9] Houman Ghaemmaghami, David Dean, Robbie Vogt, and Sridha Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4185–4188.

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

[11] Wenbo Liu, Zhiding Yu, and Ming Li, "An iterative framework for unsupervised learning in the plda based speaker verification," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 78–82.

[12] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldrich Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Speaker Odyssey Workshop*, 2012, pp. 157–164.

[13] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker Odyssey Workshop*, 2010.

[14] Grégor Dupuy, Sylvain Meignier, Paul Deléglise, and Yannick Estève, "Recent improvements towards ILP-

based clustering for broadcast news speaker diarization," in *Speaker Odyssey Workshop*, 2014.

[15] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier, "An extensible speaker identification sidekit in python," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[16] Olivier Galibert and Juliette Kahn, "The first official repere evaluation," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2013.

[17] Olivier Galibert, Jeremy Leixa, Adda Gilles, Khalid Choukri, and Guillaume Gravier, "The ETAPE Speech Processing Evaluation," in *Conference on Language Resources and Evaluation*, Reykyavik, Iceland, May 2014.

[18] 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.

[19] Olivier Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech.," in *INTERSPEECH*, 2013, pp. 1131–1134.