INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Iterative PLDA Adaptation for Speaker Diarization

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

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

Abstract

This paper investigates iterative PLDA adaptation for cross-show speaker diarization applied to small collections of French TV archives based on an i-vector framework. Using the target collection itself for unsupervised adaptation, PLDA parameters are iteratively tuned while score normalization is applied for convergence. Performances are compared, using combinations of target and external data for training and adaptation. The experiments on two distinct target corpora show that the proposed framework can gradually improve an existing system trained on external annotated data. Such results indicate that performing speaker diarization on small collections of unlabeled audio archives should only rely on the availability of a sufficient bootstrap system, which can be incrementally adapted to every target collection. The proposed framework also widens the range of acceptable speaker clustering thresholds for a given performance objective.

Index Terms: speaker diarization, PLDA, unsupervised training, domain adaptation, iterative training

1. Introduction

The goal of speaker diarization is to segment and label speaker utterances across one or more audio recordings without a priori knowledge of the speakers. The amount of multimedia data produced every day, requiring automatic indexation, created a need for an effective diarization framework.

Cross-show diarization consists in processing a dataset of raw audio archives to extract the information about the speaker occurences: ”who speaks when?”. In such a task, speakers are to be identified by a same label across the dataset. Speaker diarization applied to collections is usually decomposed in two steps: within-recording diarization, aiming at segmenting and clustering speaker occurences within each recording, and cross-recording speaker linking, which aims at regrouping the within-recording clusters of a same speaker across the whole dataset.

Application domains of cross-show diarization include radio and TV [1, 2, 3, 4, 5], phone [6, 7, 8, 9] or meeting recordings [10]. The state-of-the-art systems based on i-vector/PLDA framework require speaker annotated datasets including speaker segments and identities. The training of those systems relies on between-speaker variability estimation which requires several utterances of a same speaker in various acoustic environments.

When a new collection is to be processed, the collection itself would be the best corpus for this variability estimation. Unfortunately, manual speaker labels are not available for every targeted collection, leading to two different strategies: the state-of-the-art, for which supervised training on external annotated training data results in an acoustic conditions mismatch between training and target data or our approach which uses unsupervised training on the target dataset itself.

In our previous work [2, 11], it was shown that an unsupervised diarization system, trained on a large multi-speaker unsegmented dataset, performs as well as a supervised one, indicating that annotations are not mandatory for training. However we noted that if a target dataset, used as training material, is too small, an effective unsupervised PLDA system cannot be trained. Only the adaptation of an existing system, trained on a large amount of external annotated data, is possible.

In case of acoustic mismatch between train and target data, different strategies of domain adaptation have been proposed in the context of speaker verification [7][12][13]. In the context of unsupervised PLDA training, the concept of iterative training [14] has also been investigated and proved to be effective.

In this paper, we propose an iterative adaptation process to overcome the target collections small size issue. We investigate the use of data from the target collection itself to perform iterative adaptation of PLDA parameters for speaker diarization, using manually annotated external data for bootstrap. The main difference with the system proposed in [11] is the refined adaptation step, the focus on the iterative process and the fact that only the cross-recording speaker linking is rerun through iterations, instead of the whole process.

Subsequent sections are organized as follows: first, we describe the diarization framework and the perimeter of the iterative adaptation process. Then we present the data used for the experiments and conclude with the performances of the proposed system and the possible improvements.

2. Diarization Framework

Figure 1 describes the two-pass diarization process, detailed below. It is composed of a within-recording diarization, performed on each file of the collection, followed by a cross-recording speaker linking step. This process, widely described in [2], is based on an i-vector/PLDA system trained in a supervised way, then adapted in an unsupervised way.

2.1. Within-recording speaker diarization

The front-end is composed of a MFCC extraction and Viterbi-based speech activity detection, followed by a standard BIC segmentation and clustering and i-vector extraction. The BIC penalty coefficient is chosen so that resulting clusters are pure and represent a unique speaker. Each cluster is normalized with zero mean and unit variance and an i-vector is extracted. Spherical Nuisance Normalization (SNN) [15] is applied on the whole i-vector dataset.

Afterwards, PLDA is used to calculate log likelihood ratios (LLR) for all pairs of i-vectors [16]. The opposite of the resulting LLR matrix is called PLDA score matrix in the following of the paper. For each recording, a PLDA score matrix is computed

http://dx.doi.org/10.21437/Interspeech.2016-572

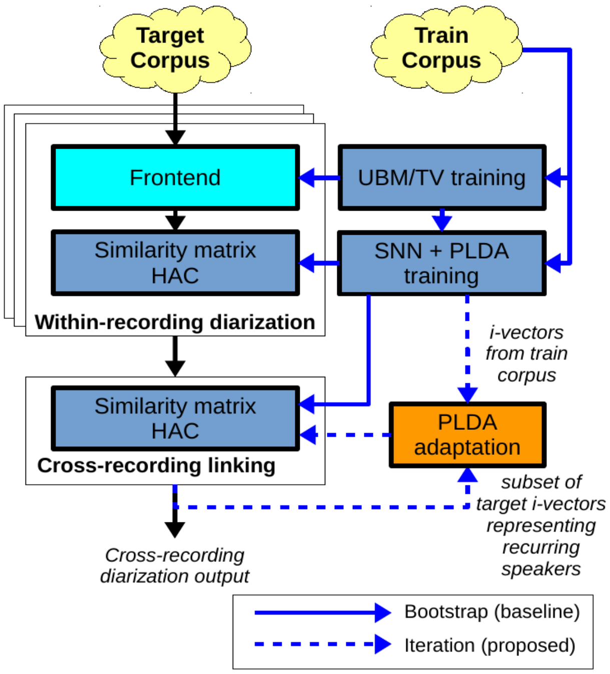

Figure 1: Overview of the diarization framework for bootstrap (plain blue lines) and adapted (dashed lines) training.

and a complete-linkage Hierarchical Agglomerative Clustering (HAC) is performed to cluster the i-vectors.

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 [17].

2.2. Cross-Show Speaker linking

Once each recording has been separately processed, the whole collection is considered as a global speaker linking problem and the previous clustering framework is reapplied on newly formed clusters. An i-vector is computed for each within-recording output cluster (by computing the mean of the initial i-vectors merged in that cluster), a new PLDA score matrix is computed and HAC is used to form global clusters.

3. Iterative adaptation of PLDA parameters

In this paper, the target collections are too small to be solely used for unsupervised estimation of PLDA parameters [11], the use of a sufficient external labeled train corpus is mandatory to estimate those parameters.

Figure 1 represents the overview of the diarization framework, including the two possible strategies for the training process: the bootstrap system (baseline) is represented with blue plain lines while the blue dashed lines correspond to the proposed adapted system. With the bootstrap system, a first step of within- and cross recording diarization is performed. We propose then to adapt the system with the target data itself and rerun the cross-recording speaker linking. We define an iterative PLDA adaptation framework (blue dashed lines). Our aim is to use the recurring speakers information, retrieved after an iteration of cross-recording diarization, to enhance the system. The within-recording process could also be rerun but for computational reasons, we decided to focus on the cross-recording part.

3.1. Unsupervised adaptation

Due to the mismatch between the train and target corpora and the lack of target data for PLDA training, we propose to apply the weighted likelihood domain adaptation technique, which is described in [12]. Γ, the inter-speaker variability matrix, and Λ, the intra-speaker variability matrix, are the PLDA parameters, trained with the EM-algorithm to maximize the weighted log-likelihood objective :

$$L(\Gamma,\Lambda) = \alpha L_{train}(\Gamma,\Lambda) + (1-\alpha)L_{target}(\Gamma,\Lambda) \qquad (1)$$

With:

$$L_k(\Gamma,\Lambda) = \frac{1}{N_k}\sum_{s=1}^{S_k} log(p(D_s|\Gamma,\Lambda)) \qquad (2)$$

Where N_k is the number of sessions and S_k the number of speakers of corpus k, D_s being the collection of i-vectors representing a speaker s. Speaker sessions from the target corpus are extracted from the output clusters of an iteration of cross-recording linking. We only consider the clusters (i.e. speakers) with occurrences in at least three different recordings, with a minimum amount of speech of 10s per recording.

The main parameter of the adaptation step is the weight α, which balances the influence of the two corpora. With $\alpha = 0$, only the target data is used and the EM algorithm does not converge, due to the dataset size. When $\alpha = 1$, the PLDA parameters obtained after adaptation are identical to the baseline.

3.2. Iterative training

Due to the improvement of diarization performance expected with the adaptation procedure, we propose to iterate the process. The main parameter of the cross-recording speaker linking process is the clustering threshold λ. It is the key parameter of the HAC applied on the score matrix. The optimal value usually depends on the PLDA scores distribution. Through the adaptation procedure, we observe that the distribution changes from one iteration to another. Since we do not want to have to tune the threshold after each iteration and keep the same as the baseline, we propose to normalize PLDA scores according to the baseline distribution, which is the distribution of PLDA scores at the initial iteration (3).

$$\hat{x}_{ijn} = \frac{x_{ijn} - \mu_n}{\sigma_n}\sigma_{baseline} + \mu_{baseline} \qquad (3)$$

with x_{ijn} being the PLDA likelihood ratio between i-vector i and j at iteration n. $\mu_{baseline}$, $\sigma_{baseline}$, μ_n and σ_n are the score distributions parameters.

Since the set of i-vectors to be clustered is the same whichever the iteration (only the cross-recording speaker linking step is rerun), distribution parameters can consequently be computed over different versions of PLDA scores between the same i-vector subsets. Scores between i-vectors used for PLDA adaptation at the previous iteration can be biased: some are considered as belonging to the same speaker, but it might be inaccurate. We decide to normalize according to the distribution of scores between i-vectors which are not used for PLDA adaptation at the previous iteration, i.e. i-vectors representing non-recurring speakers only.

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 220 hours of French broadcast news drawn from REPERE [18], ETAPE [19] and ESTER[20] 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 recording of a corpus are called recurring (R.) speakers, as opposed to one-time (O.T.) speakers, who only speak in one episode.

4.1. Train corpus

The train corpus, used for bootstrapping, is composed of 317 audio files from train and development corpora of the ESTER campaign, taken from radio broadcasts, for a total of 190 hours of speech duration. To maximize the acoustic mismatch between the train and target data, only radio shows were selected to build the train corpus, while both target corpora contain TV shows only. The train 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 to train an i-vector/PLDA system.

4.2. Target corpora

We define two target corpora built from the REPERE and ETAPE train and test 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 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. As opposed to [11], where the same shows are selected to build the target corpora, we decided to restrict our experiments to the annotated segments, in order to accurately evaluate the unsupervised part of the framework.

Target Corpus	LCP	BFM
Episodes	45	42
Labeled speech duration	10h08m	19h57m
One-Time (O.T.) speakers	127	345
Recurring (R.) speakers (2+ occurrences)	93	77
R. speakers (3+ occurrences)	48	35
Total speakers	220	422
O.T. speakers speech proportion (s.p.)	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.

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 not attributed to the correct speaker, using the best match between references and hypothesis speaker labels. The scoring tool [21] is employed for within-recording and cross-recording speaker diarization, with a collar of 250ms. The focus of this paper being the cross-show speaker linking part of the diarization process, we only present our results in terms of cross-recording DER.

5.1. Baseline and oracle

5.1.1. Definition

For both target corpora, the same *baseline* system is used, where PLDA parameters, UBM and TV matrix are trained on the external corpus. We keep the same UBM and TV in all experiments in order to focus on the variability induced by PLDA adaptation. This *baseline* is the bootstrap of our iterative adaptation framework (see 3).

Using the annotations available with the two target corpora, we decided to train an *oracle* system for each. The *oracle* system is obtained with an iteration of adaptation between the *baseline* and the recurring speakers of the target corpora, according to the corpus labels. This is the best achievable system, since it is using the ground truth about the target corpora for PLDA adaptation. The target corpora being too small, the *oracle* cannot be trained without the external bootstrap.

5.1.2. Comments

Results are presented as a function of three parameters: the adaptation parameter, α (see eq. 1), the clustering threshold, λ and the iteration number, i, ranging from 1 to 6. For a single experiment, α and λ are fixed and 6 iterations are performed. Figures 2 and 3 show results for both corpora, as a function of α and λ respectively, the other parameter not varying. The DER at each iteration is shown with a histogram bar. After the last iteration, we note the best achievable DER, which could be obtained with optimal threshold (*iter6-best*, last histogram bar).

According to figure 3, we see that the best *baseline* DER for BFM_{target} is 19.5, while it is 18.0 for LCP_{target}. In figure 2, the best *oracle* DER on the BFM corpus is of 12.2, while for the LCP corpus, it is of 11.2. This shows that when the recurring speakers of the target corpus are perfectly known, the adaptation process can greatly improve the baseline DER.

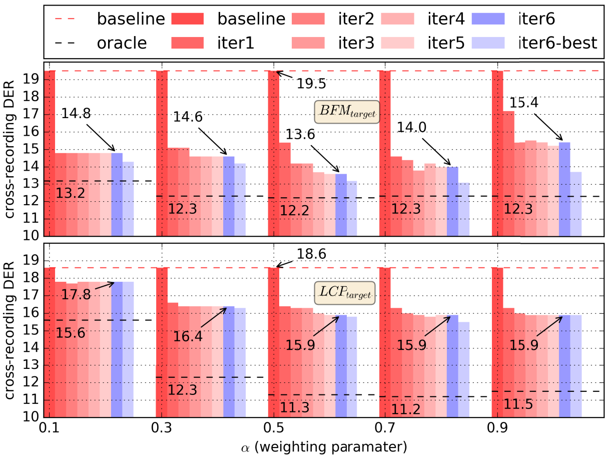

Figure 2: Cross-recording DER for both target corpora, for iterations 0 (*baseline*) to 6, as a function of α ($\lambda = -4$).

5.2. Iterative PLDA Adaptation

On figure 2, with the proposed framework, we see that a single iteration of PLDA adaptation (*iter1*) greatly improves the *baseline* DER for all values of α (for example, decrease from 19.5 to 14.0 and from 18.6 to 15.9 respectively, for $\alpha = 0.7$). This is coherent with the results presented in [12], where the adaptation process improved the baseline results too. If the major DER decrease is obtained after a single iteration, we note that in some configurations, two or three iterations are necessary to reach some kind of plateau. Sometimes, extra iterations can even give an small extra improvement.

We can also see that optimal values for α are found in a range between 0.5 and 0.7, this is coherent with the *oracle* results. When α is too small, the DER quickly reaches a plateau and the system stops improving after one iteration. This was observed in [11], where adaptation was performed with concatenation of train and target data. This is equivalent to setting α as the ratio between the two dataset, in this case, close to 0.1. Whatever the value of α, the cross-recording DER after 6 iterations *iter6* is always better than the *baseline*.

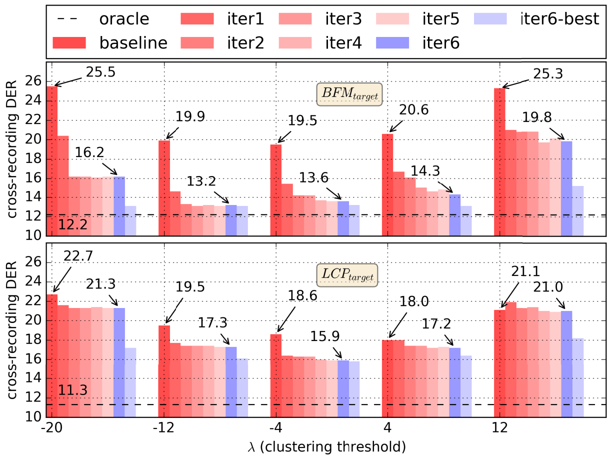

Figure 3: Cross-recording DER for both target corpora, for iterations 0 (*baseline*) to 6, as a function of λ ($\alpha = 0.5$).

5.3. Clustering threshold

In figure 3, α is set to 0.5 and the results depend on λ. Whatever the choice of λ, the DER after 6 iterations (*iter6*) is lower than the *baseline*, with a decrease after each iteration. This means that the range of acceptable thresholds, if we want to keep the DER below a certain value, is wider when iterative adaptation is performed. Iterative adaptation shows that after each iteration, the gathered information about recurring speakers of the target corpus enhances the PLDA parameters estimation.

For BFM_{target}, the best final DER is obtained for $\lambda = -16$, with a final DER value of 13.2, very close to the *oracle* (12.2), the *baseline* being 19.9. For LCP_{target}, the best final DER value of 15.9 is obtained for $\lambda = -4$, the *oracle* being 11.3 and the *baseline* being 18.6. We see that for both corpora, the threshold value corresponding to the best final DER (*iter6*) is not necessarily the optimal value for the *baseline*, but is very close. We also note that for optimal values of λ, the final DER (*iter6*) is really close to the best achievable DER (*iter6-best*).

From our observations, we note that for high values of λ (merges of clusters from different speakers), the iterative adaptation process can stagnate or even diverge. If stagnation is noticeable for the last experiment of figure 3, divergence only appears for λ values far from the baseline optimal, and a decently calibrated threshold should be sufficient to obtain DER gain with the proposed framework.

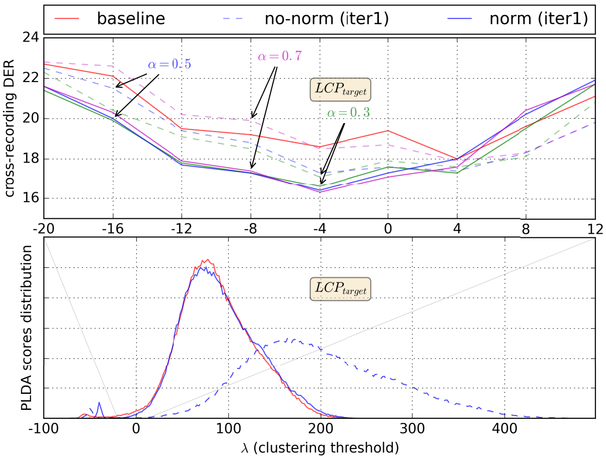

Figure 4: Effect of score normalization on distributions and DER for a single iteration on LCP_{target}

5.4. Score normalization

If the optimal α parameter seems to be easy to calibrate, the main difficulty in the speaker linking step is the clustering threshold setting (λ). In our framework, for a single experiment, it set once and applied for the baseline and adapted diarization. Early experiments with the framework showed that without any score normalization technique, the distribution of PLDA scores changes a lot from one iteration to another and the system under-performs, due to the inability to modify λ accordingly. This phenomenon is presented in figure 4, for a single adaptation iteration. When studying closely the distribution of scores, we also noticed that the combination of adaptation and score normalization tends to spread the scores in the region of the optimal threshold, allowing to improve the resolution around (not visible in the figure).

6. Conclusion

In this paper we proposed a cross-show diarization framework based on an auto-improving i-vector/PLDA system, in order to process 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 target corpora is a problem. The use of an external bootstrap is required.

Using unlabeled and unsegmented data from the target collection itself, we successfully applied the weighted likelihood domain adaptation technique, which proved to be effective for supervised speaker verification on mono-speaker data, to improve the baseline diarization system. After multiple iterations and using score normalization for better convergence, results showed a decrease in terms of cross-recording DER for both target corpora, for a wide range of speaker linking thresholds.

Further work will be dedicated to the study of the minimal requirements concerning target corpora and bootstrap for the adaptation process, in the context of incrementally growing archive collections. We will also consider training a fully unsupervised auto-improving diarization system, the main problem being the question of bootstrapping without any labeled data.

7. References

[1] 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.

[2] G. Le Lan, S. Meignier, D. Charlet, and P. Deléglise, “Speaker diarization with unsupervised training framework,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[3] Q. Yang, Q. Jin, and T. Schultz, “Investigation of cross-show speaker diarization.” in *INTERSPEECH*, 2011, pp. 2925–2928.

[4] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, “Comparing multi-stage approaches for cross-show speaker diarization.” in *INTERSPEECH*, 2011, pp. 1053–1056.

[5] D. A. Van Leeuwen, “Speaker linking in large data sets,” Proc. Odyssey 2010.

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

[7] S. H. Shum, D. A. Reynolds, D. Garcia-romero, and A. Mccree, “Unsupervised clustering approaches for domain adaptation in speaker recognition systems,” Proc. Odyssey 2014.

[8] Z. N. Karam and W. M. Campbell, “Graph embedding for speaker recognition,” in *Graph Embedding for Pattern Analysis*. Springer, 2013, pp. 229–260.

[9] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, “Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4185–4188.

[10] 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.

[11] G. Le Lan, S. Meignier, D. Charlet, and A. Larcher, “First investigations on self trained speaker diarization,” in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2016.

[12] 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.

[13] 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.

[14] W. Liu, Z. Yu, and M. 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.

[15] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, “Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis,” in *Speaker Odyssey Workshop*, 2012, pp. 157–164.

[16] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” in *Speaker Odyssey Workshop*, 2010.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.