Speaker Diarization With Unsupervised Training Framework

Gaël Le Lan, Sylvain Meignier, Delphine Charlet, Paul Deléglise

SPEAKER DIARIZATION WITH UNSUPERVISED TRAINING FRAMEWORK

Gaël Le Lan[1,2]*, Sylvain Meignier*[1]*, Delphine Charlet*[2]*, Paul Deléglise*[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 single and cross-show diarization based on an unsupervised i-vector framework, on French TV and Radio corpora. This framework uses speaker clustering as a way to automatically select data from unlabeled corpora to train i-vector PLDA models. Performances between supervised and unsupervised models are compared. The experimental results on two distinct test corpora (one TV, one Radio) show that unsupervised models perform as good as supervised models for both tasks. Such results indicate that performing an effective cross-show diarization on new language or new domain data in the future should not depend on the availability of manually annotated data.

Index Terms— Speaker diarization, speaker linking, unsupervised training.

1. INTRODUCTION

Speaker diarization, as defined in the NIST evaluations on Rich Transcription [1], aims to uniquely identify speakers across an audio stream, without a priori knowledge about the speakers. Until recently, speaker diarization was applied on individual shows.

With the increasing amount of data and the need for indexation of multimedia documents, performing diarization has become mandatory, not only on individual shows but on collections of shows. This task is known as cross-show speaker diarization [2], speaker linking [3] or speaker attribution [4]. Thus, some implementations have been proposed [2, 3, 5, 6, 7]. The usual approach consists in applying multistage clusterings, considering shows one by one and then merging the speaker clusters in the cross-show step. The ETAPE and REPERE campaigns [8, 9] addressed the problem of speaker diarization on French TV broadcast news data and an implementation has been proposed, where the system was trained on annotated data [7]. The MGB Challenge [10] addressed the same problem on English TV broadcast data, but no annotations were provided for training data.

The state of the art speaker modeling is based on i-vector/PLDA frameworks, which require data labeled by speakers, indicating who speaks when. Such frameworks are based on inter session speaker variability modeling, thus training corpora require multiple occurrences of a same speaker across different shows. Such labeled data are expensive to produce and are not always available for target data, resulting in a mismatch of language and/or noise conditions. Solutions to this problem have already been proposed in the context of speaker verification, based on domain adaptation [11] or unsupervised iterative PLDA training [12]. For both systems, training was done on mono-speaker unlabeled data.

In this paper we address the problem of cross-show diarization, using a priori models trained on unlabeled data. We propose an unsupervised framework for data selection, in order to train models needed to extract i-vectors (Universal Background Model (UBM) and Total Variability (TV) matrix) and to compute PLDA (eigenvoice and noise covariance matrices). Those models are trained offline and their performances on single and cross-show speaker diarization are compared with supervised models, trained on the same audio data, but using speakers labels.

Subsequent sections are organized as follows: first, we describe the baseline single-show and cross-show diarization systems. Then, we present a solution for the data selection problem for model training and conclude with the experimental results using unsupervised and supervised models.

2. DIARIZATION FRAMEWORK

Figure 1 describes the diarization process, which is detailed below. This process can be based on a i-vector/PLDA system trained in a supervised or unsupervised way.

2.1. BIC segmentation and clustering

As presented in the first part of Figure 1, after a Viterbi-based speech/non speech detection, a two pass acoustic segmentation, followed by a hierarchical agglomerative clustering (HAC), is applied over the speech segments. The segmentations and the clustering are described in [13] where the standard BIC is replaced by the segmental square-root BIC described in [14]. From previous experience, the segmental square-root BIC outperforms the standard BIC on target data.

2.2. Single Show Speaker clustering

At this point, each cluster is pure and supposed to represent a single speaker; however, several clusters can be related to the same speaker. In the previous steps, features were not normalized because the channel contribution was useful to differentiate the speakers. In this step, the channel contribution is removed with a zero mean and unit variance normalization computed for each cluster. A final clustering stage is then performed in order to obtain one cluster by speaker.

2.2.1. Speaker Modeling: i-vector/PLDA

Speaker clusters are modeled with i-vectors, normalized by Spherical Nuisance Normalization [15]. PLDA is used to calculate log likelihood ratio for pairs of i-vectors [16], which we call PLDA scores

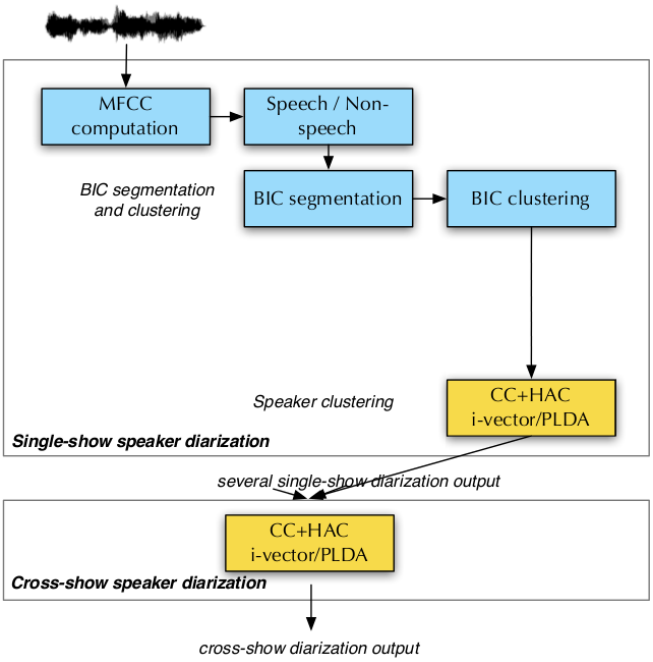

Fig. 1. Overview of speaker diarization systems.

or PLDA distances in the following of the paper. The term PLDA matrix indicates the matrix of the log likelihood ratios. I-vector and PLDA are computed using the Alize toolkit [17]. The dimension of the feature vectors is 60: 20 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 ivectors is 200 and PLDA eigenvoice matrix has a dimension of 100.

For experimental purposes, we also present a system which does not use PLDA nor SNN, where the scores between i-vectors are the cosine distance between them.

2.2.2. Clustering Process: CC+HAC

For the last step, a graph based clustering followed by a HAC is performed on the PLDA matrix. The negative of the PLDA distance matrix [1] can be interpreted as a connected graph, illustrated in Figure 2, where the clusters are represented by the nodes, and scores between clusters are represented by the links. This approach is similar to the ILP clustering presented in [18].

This graph can be simplified by removing all the unnecessary links corresponding to scores above a threshold δ. After this simplification, the graph now contains a set of subgraphs corresponding to the connected components (CC) of the graph.

A complete-linkage HAC is then performed on each complex subgraph (ie. not a star-graph). In this case, the HAC tries to find classes within a diameter less than a threshold β.

2.3. Cross-show Speaker Diarization

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

[1] we change the sign of PLDA score to minimize modifications in the classification algorithm that usually works with distances.

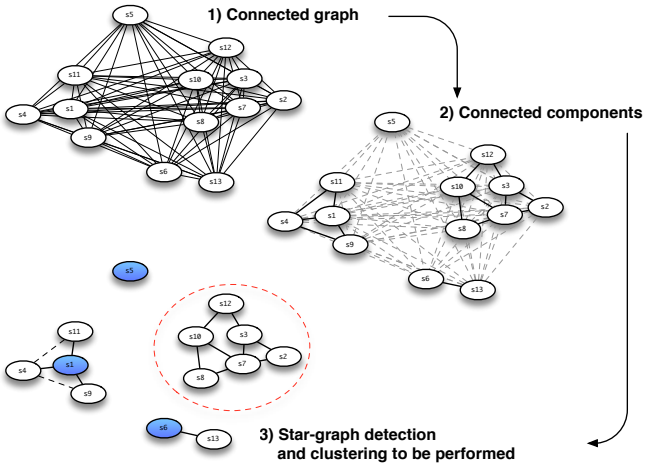

Fig. 2. Graph clustering using sub-components. The dashed circle indicates a sub-graph, for which a HAC needs to be performed; the colored clusters are identified as star-graph centers.

3. 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 [9], ETAPE [8] and ESTER[19] evaluation campaigns were used to build three corpora. The shows were broadcast between 1998 and 2007, each show during between 10 minutes and an hour. The corpora also contain some broadcasts of Moroccan radio, in French language. For each show in the corpus, multiple episodes are available.

3.1. Train corpus

The $train$ corpus is composed of 344 audio files from train and development corpora of the three previously cited campaigns, for a total of 1772 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 corpus contains 3888 unique speakers. Among those speakers, 929 appear in at least two different episodes (from the same or from different shows), we call them Cross-Show (CS) speakers, as opposed to Single-Show speakers (SS), who only speak in one audio. Thus, this corpus is well suited for a i-vector PLDA system training.

3.2. Test corpora

In this paper, we define two test corpora built from REPERE and ESTER test corpora. The first one, named $ESTER_{test}$, is the test corpus from ESTER campaigns, it contains Radio broadcasts. The

second test corpus, named BFM_{test}, is the collection of all available episodes of the TV news talk-show $BFMStory$. It has been selected because it is the one with the highest number of episodes (42), and there is a large amount of CS speakers, who speak for more than 55% of the total speech duration of the collection. Numerical details about the two corpora are presented in table 1.

Corpus	$ESTER_{test}$	BFM_{test}
Episodes	24	42
SS speakers	454	358
CS speakers	42	79
Total speakers	496	437
Total duration	14h43m00s	43h12m17s
SS speakers speech proportion	74,60%	44,62%
CS speakers speech proportion	25,40%	55,38%

Table 1. Composition of test corpora.

4. DATA SELECTION FOR UNSUPERVISED MODEL TRAINING

The proposed selection method has been inspired by the IDIAP method[20] presented during the NIST i-vector Machine Learning Challenge 2014 for speaker recognition. In this challenge, unlabeled i-vectors are directly provided, without audio data. The training corpora need to be automatically constructed, using clustering methods to train session compensation, UBM, TV matrix and PLDA. Our task is more demanding than the i-vector challenge: the i-vectors need to be extracted from unsegmented audio data.

According to the diarization system described in section 2, required models for diarization are a Universal Background Model (UBM), a Total Variability matrix (TV) and a PLDA matrix. For the data selection task, only the audio of the training corpus is usable, annotations provided in the corpus are not authorized.

4.1. 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 the GMM based speech / non-speech detector on the training corpus is used to train the UBM, which contains 256 Gaussians with diagonal covariance matrices.

4.2. Total Variability

The Total Variability (TV) matrix is required to extract i-vectors. It must be trained over segments containing a unique speaker. On the test corpus, most of the classes produced by the BIC diarization system are pure (they contain only one speaker). According to this observation, we consider each class produced by the BIC diarization applied to the $train$ corpus as a segment representing a unique speaker and the total variability matrix is trained on those with a dimension of 200.

4.3. PLDA

Spherical Nuisance Normalization requires the mean vector and the within-class co-variance matrix of the i-vectors, whereas PLDA is trained using normalized speaker labeled i-vectors. Both systems are designed to compensate the inter-session speaker variability. In both cases, the training data must be labeled by speaker and session. Several sessions are needed per speaker to obtain robust models.

Hence, we need to perform a cross-show speaker diarization on the training corpus to automatically obtain speaker clusters, containing data from the same speaker and from different audios. PLDA will then be trained over the i-vectors extracted from the automatically obtained speaker clusters.

An i-vector for each class produced by the BIC diarization is extracted using the 256-UBM and TV matrix trained over the training corpus. A cosine distance matrix between all the i-vectors is computed, without normalization, and a CC+HAC clustering is applied, thresholds δ and β being set at the same value (see 2.2.2). In the resulting clusters, we only keep the ones containing at least 3 initial classes. The kept clusters are considered as containing multiple occurrences of a single speaker.

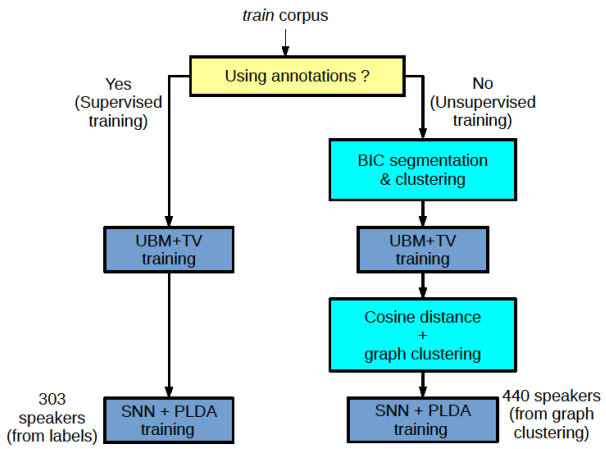

Fig. 3. Unsupervised selection methods for $train$ corpus.

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 is employed for single-show and cross-show speaker diarization. In this last case, a cross-show speaker would always be labeled in the same way, in every recording that composes a collection.

5.1. Data selection

Supervised PLDA was trained using labeled data from the $train$ corpus, over speakers appearing in at least 3 episodes. From previous experience, keeping speakers with a least 2 occurrences gives slightly lower performances. This process allowed us to retrieve 303 different speakers. Those speakers represent $61,34\%$ of the total speech time of the $train$ corpus. We call that proportion the data coverage (proportion of the $train$ data used for PLDA modeling).

For the unsupervised PLDA, CC+HAC clustering is performed on the $train$ corpus. Figure 4 shows the clustering results, depending on the clustering threshold applied on cosine distances between i-vectors representing the initial BIC classes. Resulting clusters are

used to generate the unsupervised $PLDA_{unsup}$. The results are plot in terms of number of automatically extracted clusters, their average purity and the data coverage (proportion of initial data kept) they represent. The figure also presents single-show DER results on the $ESTER_{test}$ corpus, for each generated PLDA.

The DER is lower than 8.5% for a wide amount of threshold, varying from -0.7 to 0.1, for a purity varying from 93% to 40% and for a data coverage varying from 40% to 100%. At the lowest threshold, the model is still performing well although the data coverage is lower than for the supervised PLDA (40%), generating 395 different clusters, with high purity (93%). On the other hand, at the highest threshold, the number or clusters is around 200, purity is relatively low, all data from the corpus are kept but the added data do not affect too much the performances. What seems to be important for the PLDA to be effective is generating enough clusters, which a relative high purity. When the number of clusters falls under 200, too many different speakers are agglomerated and the modeling is too poor.

The unsupervised PLDA which gives the lowest DER on the $ESTER_{test}$ corpus is achieved for a threshold of -0.4. For this threshold, the single-show DER is 7.17%, the purity is 85.98% and the data coverage is 73.96%. The clustering process at this threshold generates 440 clusters (i.e. supposed recurring speakers) from the $train$ corpus. Analyzing the composition of those clusters, we note that the average recall of all speakers is 90%, meaning that all classes from a same speaker are in majority regrouped in a same cluster. But we also note that 17% of the speakers have a recall less than 90%, representing 27% of the total speech time, some of the main speakers of the corpus being split in different clusters.

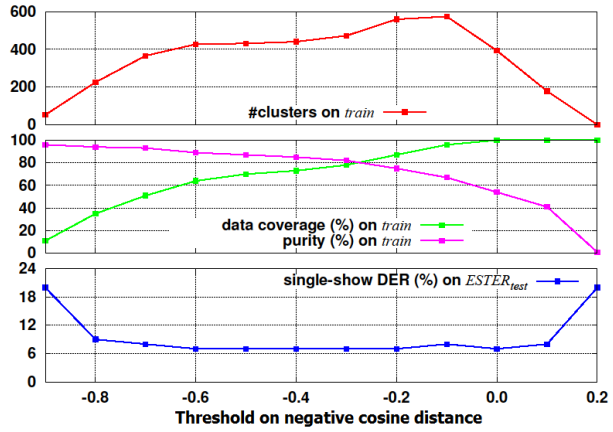

Fig. 4. Metrics on the unsupervised i-vector cosine clustering of training data for $PLDA_{unsup}$ generation

5.2. Single and cross-show diarization

Table 2 shows single and cross-show DER on both test corpora, using cosine and PLDA distance metrics (see 2).

5.2.1. TV matrix influence

Cosine distance is only based on i-vectors, without normalization. Thus, performances with cosine distance show the difference between supervised and unsupervised modeling of the UBM and TV matrix. Performances between both modelings are similar, except for the cross-show DER on $ESTER_{test}$.

5.2.2. PLDA modeling

PLDA-based scoring results allow to compare the full supervised and unsupervised systems. Single-show DER is quite similar on both configurations on $ESTER_{test}$, while unsupervised modeling provides better results on BFM_{test}. The unsupervised system performs slightly better than the baseline on the cross-show task, despite the imperfect purity (85.98%) of the generated clusters for the unsupervised PLDA modeling.

We suppose the difference of performance between the supervised and unsupervised PLDA-based system is explained by the fact that some speakers with a high speech time and many occurrences (see 5.1, having a high inter-session variability, are considered as separate speakers in the unsupervised system, and give a better modeling : the amount of clusters for PLDA training is higher when generated automatically (440 versus 303 speakers generated from labels).

While previous work [12] shows that an unsupervised PLDA-based performs less than a supervised PLDA-based system on the speaker verification task on telephone data, our experiments show that both supervised and unsupervised PLDA-based systems give similar results on TV/Radio data. We suppose it is due to the fact that our unsupervised system is trained on data which is processed in the same way than the test data (clustered BIC diarization classes), contrary to the supervised system, where training data is obtained on manual annotations. In the NIST i-vector challenge, there is no segmentation problem, i-vectors used for unsupervised PLDA training are already provided, while in our system, they need to be produced from unsegmented audio data.

	$ESTER_{test}$		BFM_{test}	
	SS DER	CS DER	SS DER	CS DER
$Cosine - TV_{ref}$	8,67	14,15	11.42	14,99
$Cosine - TV_{unsup}$	8.46	13.09	11.14	15,01
$PLDA_{ref}$	7.14	12.25	10.76	14.22
$PLDA_{unsup}$	7.17	11.02	9.87	13.20

Table 2. single (SS) and cross-show (CS) DER on test corpora.

6. CONCLUSION

While ESTER and REPERE campaigns allowed to train a single and cross-show diarization system based on i-vector/PLDA frameworks trained on labeled data, we proposed to train the i-vector/PLDA framework models over automatically selected data from the training corpus, using data from those campaigns. We compared it with supervised models (e.g. trained on the same corpus, using provided manual labels) on the cross-show speaker diarization task.

We show that having manual labels to build a performing system is not mandatory, and that automatically select training data allows to build a sufficient i-vector/PLDA framework for cross-show speaker diarization on TV and radio broadcast news data.

Further work should be dedicated to better understanding of the composition of the unsupervised clusters, and to iterative unsupervised domain adaptation for processing new types of unlabeled collections of data.

7. REFERENCES

[1] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/docs/rt04f-eval-plan-v14.pdf, Aug 2004.

[2] 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.

[3] 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.

[4] 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.

[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] 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.

[8] 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.

[9] 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.

[10] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. Woodland, "The mgb challenge: Evaluating multi-genre broadcast media transcription," in *IEEE ASRU*, 2015.

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

[12] 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.

[13] S. Meignier and T. Merlin, "LIUM SpkDiarization: An open-source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, Texas (USA), 2009.

[14] T. Stafylakis, V. Katsouros, and G. Carayannis, "The segmental bayesian information criterion and its applications to speaker diarization," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 857–866, Oct 2010.

[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, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. S. Mason, J.-Y. Parfait, and U. ValidSoft Ltd, "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition," in *Proceedings of Interspeech*, 2013, pp. 1–5.

[18] G. Dupuy, S. Meignier, P. Deléglise, and Y. Estève, "Recent improvements towards ILP-based clustering for broadcast news speaker diarization," in *Speaker Odyssey Workshop*, 2014.

[19] 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.

[20] 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.