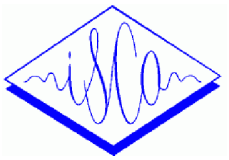
I-vectors and ILP clustering adapted to cross-show speaker diarization

Grégor Dupuy, Mickael Rouvier, Sylvain Meignier, Yannick Estève

ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

I-vectors and ILP clustering adapted to cross-show speaker diarization

Grégor Dupuy, Mickael Rouvier, Sylvain Meignier, Yannick Estève

LUNAM Université, LIUM, Le Mans, France
first.lastname@lium.univ-lemans.fr

Abstract

We propose to study speaker diarization from a collection of audio documents. The goal is to detect speakers appearing in several shows. In our approach, each show of the collection is processed separately before being processed collectively, to group speakers involved in several shows. Two clustering methods are studied for the overall processing of the collection: one uses the NCLR metric and the other is inspired by techniques based on i-vectors, mainly used in the speaker verification field. Both methods were evaluated on the whole training corpus of ESTER 2. The method based on the use of i-vectors achieves error rates similar to those obtained by the NCLR method, however, the computation time is on average 8.66 times faster. Therefore, this method is suitable for processing large volumes of data.

Index Terms: speaker diarization, cross-show diarization, i-vectors, ilp clustering.

1. Introduction

The diarization task has been defined by the NIST in the context of the *Rich Transcription* evaluation campaign as the partitioning of an input audio stream into segments, and the clustering of those segments according to the identity of the speakers. Speaker diarization is independently applied to each show that has to be processed, without *a priori* knowledge about speakers.

Until recently, most speaker diarization systems followed this definition of the task where shows are processed and evaluated individually. In this context, detected speakers are identified by anonymous labels specific to each recording: one same speaker involved in two shows is identified by two different labels.

Speaker diarization plays an important role in many applications of automatic speech processing, such as automatic transcription, named entity detection and speaker role detection. Considering the growing amount of multimedia resources available, it has become interesting and necessary to consider speaker diarization in a broader context. The major drawback of the usual approach in speaker diarization is not to take into account the interventions of some recurring speakers in several shows. This situation is very common in broadcast news programs where hosts, journalists and other guests may appear recurrently. The notion of cross-show speaker diarization on a collection has recently been introduced to deal with this kind of situation [1][2]. The authors present different overall approaches to detect and group speakers within a whole collection of recordings from a single source. Thus, a speaker involved in several shows is always identified by the same anonymous label in each of the shows.

This research was supported by ANR (French National Research Agency) under contract number ANR-2010-CORD-101-01 (SODA project)

In this paper, we compare two clustering methods adapted to the cross-show speaker diarization task when applied to French broadcast news recordings. We used a two-stage architecture system which combines single-show diarization, where shows are handled individually, and cross-show diarization, dealing with the comprehensive collection of recordings as a whole.
Subsequent sections are organized as follows: first, we describe the baseline LIUM single-show diarization system and the cross-show diarization architecture. Then, we present the data set and the evaluation metrics, the system configuration and the experimental results.

2. Single-show diarization system

Experiments were carried out using the *LIUM_SpkDiarization* system[1] [3]. This system was developed during the ESTER 2 evaluation campaign [4] and achieved the best results in the speaker diarization task on French broadcast news records.

LIUM_SpkDiarization relies on acoustic segmentation and hierarchical agglomerative clustering using BIC (Bayesian Information Criterion) both as similarity measure between speakers and as stop criterion for the merging process. Each speaker is modeled by a Gaussian distribution with a full covariance matrix. Segment boundaries are adjusted through a Viterbi decoding using GMMs (Gaussian Mixture Models) with 8 components learned on the data of each speaker via the EM algorithm (Expectation-Maximization). A segmentation into speech/non-speech regions is also carried out to remove non-speech segments. Segmentation, clustering and decoding are performed using 12 MFCC parameters (Mel-Frequency Cepstral Coefficients), completed with energy.

At this point, each speaker is not necessarily represented by a single cluster. The system then performs a hierarchical agglomerative clustering using NCLR (Normalized Cross-Likelihood Ratio) [5] both as similarity measure between speakers and as stop criterion for the merging process. Unlike the previous stages, acoustic parameters are now normalized (centered/reduced and *feature warping* calculated on each segment). The purpose of parameter normalization is to minimize channel contribution. Speaker models are obtained, for each cluster, by applying a MAP (Maximum *A Posteriori*) adaptation on a UBM (Universal Background Model). This 512 UBM components results of the concatenation of four 128 component GMMs, gender- and bandwidth-dependent.

3. Cross-show diarization architecture

While a single-show diarization system allows to detect speaker utterances within a show, a cross-show diarization system, in

[1] http://www-lium.univ-lemans.fr/en/content/liumspkdiarization

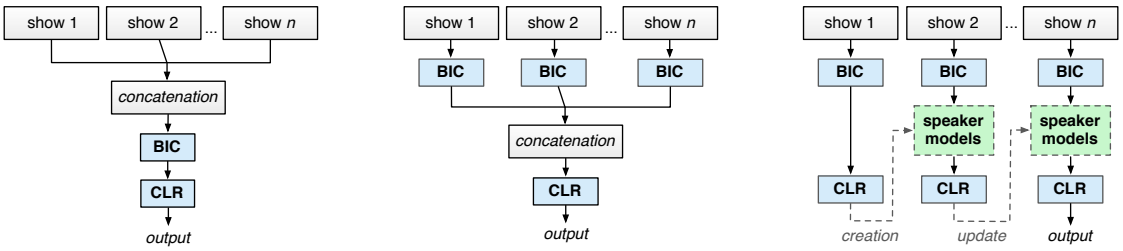


Figure 1: The three architectures for cross-show speaker diarization already studied [1]: left, the concatenation of all shows; middle, the *hybrid* system; right, the *incremental* system.

addition, deals with speakers appearing across multiple shows of a collection. Three different architectures have already been compared [1][2] (Figure 1):

1. a *concatenation* of all shows into a single one, on which a usual single-show diarization system is run (close to the one described in section 2);
2. a *hybrid* system, in which a BIC clustering is performed individually on each show, before the concatenation of the outputs is used for an overall BIC clustering [2] or CLR clustering [1];
3. an *incremental* system, which processes the shows individually one after another. Only speaker models extracted from shows already processed can help the diarization of the show currently being processed. Speaker models learned on each show are used and updated over the processing of the collection.

The *concatenation* and *hybrid* systems are comparable in terms of performance. The *incremental* system distinguishes itself through processing speed. This architecture is most suitable for the insertion of new shows in the collection, but it has two drawbacks: its diarization error rate on the whole collection is higher than that obtained by the two other systems, and the order in which shows are processed affects the results. These experiments show that the best results are achieved at the expense of processing time, and *vice-versa*.

We considered a different approach in which we chose to implement a system suitable for processing large volumes of data. Such a system has to be efficient both in terms of speaker diarization error rate and in terms of computational time and memory usage. We drew inspiration from the *hybrid* system by testing two different clustering methods for the global processing of the collection. Figure 2 shows the two clustering methods tested: the first one implements a NCLR and the second one is formulated as an Integer Linear Programming (ILP) problem based on i-vectors. In both cases, each show is processed individually, using the single-show diarization system described in section 2, before attempting to identify speakers reappearing in several shows within the collection. The *collection* is obtained by concatenating the outputs from individual processings.

3.1. NCLR cross-show diarization system

With this variant, the overall processing of the collection is performed by a NCLR clustering. The architecture of this system is really close to the *hybrid* architecture [1] previously presented, the only noticeable difference is the presence of a NCLR clustering at the individual processing stage.

3.2. i-vectors and ILP cross-show diarization system

The i-vectors, mainly used in speaker verification field [6], allow to reduce large amounts of acoustic data into vectors of smaller dimensions, only retaining relevant information about speakers. This approach was adapted to speaker diarization using the *k-means* algorithm, applied to distances between i-vectors, to find utterances of speakers within a corpus where the number of speakers is known *a priori* [7].

Here, we have to deal with an unknown number of speakers. An i-vector j is extracted from each cluster j of the individual NCLR clustering stages, using 19 MFCC parameters completed with energy, their first and second derivatives, along with a 1024 UBM-GMM. The N resulting i-vectors are then normalized in an iterative process [8]. The clustering problem consists in minimizing, on the one hand, the number K of cluster centers chosen among the N i-vectors and, on the other hand, the dispersion of i-vectors within each cluster. The value $K \in \{1, \ \ldots, \ N\}$ is to be automatically determined.

We propose to express this clustering problem as an Integer Linear Programming (ILP) problem, where the objective solving function (eq. 1) is minimized subject to constraints:

Minimize

$$\sum_{k=1}^{N} x_{k,k} + \frac{1}{D} \sum_{k=1}^{N} \sum_{j=1}^{N} d(k,j) x_{k,j} \qquad (1)$$

Subject to

$$x_{k,j} \in \{0,1\} \qquad \forall k, \forall j \qquad (1.2)$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad \forall j \qquad (1.3)$$

$$d(k,j) x_{k,j} \leq \delta \qquad \forall k, \forall j \qquad (1.4)$$

Where $x_{k,k}$ (eq. 1) is a binary variable equal to 1 when the i-vector k is a center. The number of centers K is implicitly included in the equation 1, indeed $K = \sum_{k=1}^{N} x_{k,k}$. The distance $d(k,j)$ is computed using the *Mahalanobis* distance between i-vectors k and j [8]. D is a normalization factor equal to the longest distance $d(k,j)$ for all k and j. The binary variable $x_{k,j}$ is equal to 1 when the i-vector j is assigned to the center k. Each i-vector j will be associated with a single center k (eq. 1.3). The i-vector j associated with the center k (*i.e.* $x_{k,j} = 1$) must have a distance $d(k,j)$ shorter than a threshold δ empirically determined (eq. 1.4).

Preliminary experiments show that solving this ILP problem gives a better clustering than a hierarchical agglomerative clustering, regardless of the linkage criteria. This ILP-based clustering was first adapted to the single-show diarization task, without *a priori* knowledge on speakers, in a parallel work [9].

Corpus ID	Radio	Recording year	Broadcasting time slot	# of shows	Total duration (hours)	# of speakers	# of reliable speakers	# of cross-show reliable speakers
Dev	RFI	2000	9:30 - 10:30 am	15	15	358	206	48
Test 1	RFI	2000	11:30 - 12:30 am	15	15	298	143	41
Test 2	France Inter	1999	7:00 - 7:20 pm	5	2	66	50	11
Test 3	France Inter	1999	7:00 - 8:00 am	10	10	235	139	50
Test 4	France Inter	1999	8:00 - 9:00 am	10	10	181	94	24
Test 5	RFI	2001	9:00 - 10:00 am	9	9	256	165	45
Test 6	RFI	2001	10:00 - 11:00 am	9	9	244	110	28
Test 7	France Inter	2002	7:00 - 7:00 pm	5	5	151	68	15
Test 8	RFI	2002	8:00 - 9:00 am	5	5	115	88	20
Test 9	RFI	2002	0:00 - 1:00 am	5	5	113	91	15
Test 10	RFI	2002	2:00 - 2:00 pm	5	5	141	93	21
Test 11	RFI	2002	8:00 - 9:00 pm	5	5	166	91	20
Test 12	Africa	2003	all time slots	13	5	130	91	19

Table 1: Information used to divide the ESTER 2 training corpus and, for each corpus, the duration and the number of shows that compose it, the total number of speakers, the number of reliable speakers and the number of cross-show reliable speakers.

4. Experiments

4.1. Data

The data selected to perform our experiments represent the entire training corpus from the ESTER 2 French evaluation campaign [4]. It consists of 100 hours of manually transcribed French radio broadcast news recorded between 1999 and 2003. This training corpus has been divided into thirteen smaller subsets, on which we performed independent experiments. The distribution of data within the thirteen corpora was carried out according to broadcasting year and time slots criteria of each show. We used the first corpus as a development corpus, to tune

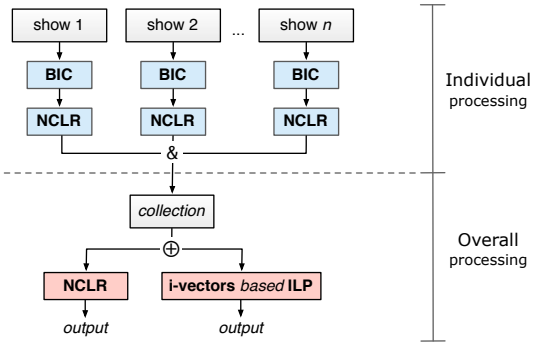


Figure 2: The cross-show diarization architecture we experimented, with its two variants: an overall NCLR clustering and an overall ILP clustering (performed by solving an ILP problem dealing with distances between i-vectors).

In order to assess the cross-show diarization task, speakers appearing in several shows must necessarily be identified by the same label in each show. The evaluation focuses only on reliable speakers, *i.e.*, speakers formally identified by their full name in the references. Other labels (Christelle, speaker#151, journaliste_rfi ...) do not provide any certainty about the true identity of speakers: the same speaker may be identified by different labels in various shows of the collection.

Information used to divide the ESTER 2 training corpus into smaller ones are presented in Table 1. The duration and the number of shows, the total number of speakers, the total number of reliable speakers and the number of reliable speakers reappearing at least in two shows, for each corpus, are also presented in this table.

4.2. Evaluation metrics

The Diarization Error Rate (DER) is the metric used to measure performance. DER was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker, using the best matching between references and hypothesis speaker labels. The scoring tool we used was developed by the LNE[2] as part of the REPERE campaign[3].

This tool allows to distinguish two different error rates: on the one hand, we use the *single-show DER* when the evaluation has to be performed by considering shows independently. The resulting value corresponds to the mean of DERs calculated individually on each show, weighted by their duration. On the other hand, we use the *cross-show DER* when evaluation has to be performed simultaneously on each show of the collection. The *cross-show DER* takes into account multiple appearances of a speaker in several shows, as if all shows were merged into a single one.

4.3. Cross-show diarization systems configuration

The UBM is learned on the test corpus provided during the ESTER 1 French evaluation campaign [4]. The ESTER 1 training corpus is used to train the i-vectors required for normalization step. Speaker/cluster models are obtained by performing a single iteration of the MAP algorithm. A single iteration of MAP allows to save time on the "update" of the cluster models in the hierarchical clustering stage: a new cluster model is obtained by merging the saved statistical accumulators. To accelerate the computation of these likelihoods, only the five top Gaussians of the cluster models are considered. The ILP problem is solved by the *Branch and Bound* algorithm implemented in the *GNU Linear Programming Toolkit* [4].

Both systems were implemented using the development corpus to determine the configuration that gives the best performance in terms of diarization error rate. The optimal NCLR threshold of the individual processing stage is 0.97, and the one with the overall processing stage is 0.82. The optimal distance threshold δ for the ILP clustering (eq. 1.4) was set to 120. Each of these thresholds has been applied similarly to process the twelve other corpora.

[2]The French National Laboratory of Metrology and Testing
[3]http://www.defi-repere.fr/
[4]http://www.gnu.org/software/glpk/

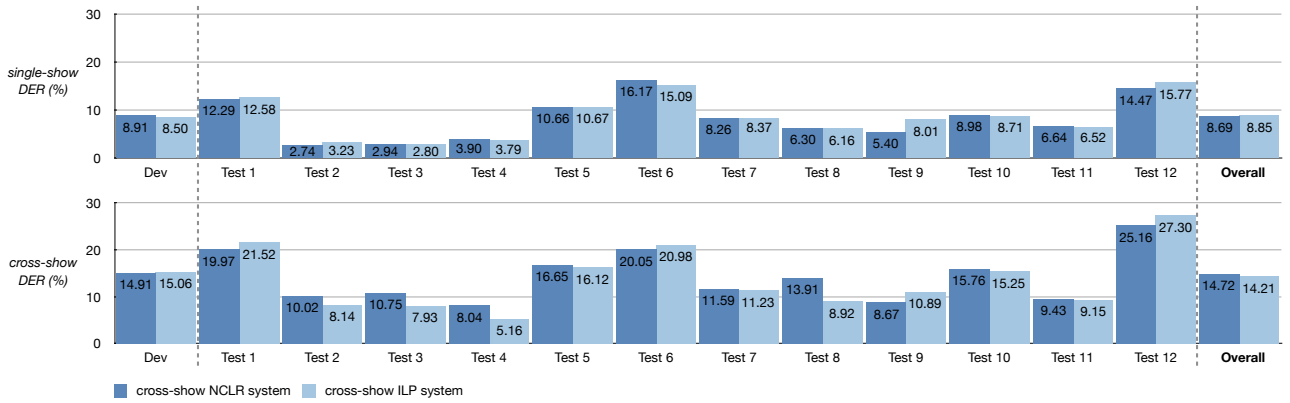


Figure 3: Evaluation results in terms of *single-show DER* (top graph) and *cross-show DER* (bottom graph), on the thirteen corpora, with the cross-show NCLR and ILP systems. The dev corpus was not included in the calculation of the overall mean.

4.4. Results and discussion

In Figure 3, we present the comparison between the results obtained by the cross-show NCLR system with those obtained by the cross-show ILP system. Experiments were performed individually on the twelve test corpora and evaluated both in terms of *single-show DER*, on the top graph, and *cross-show DER*, on the bottom graph. The overall means were calculated on the test corpora only, with the DER of each corpus weighted by their duration (as if we consider the whole collection as a single corpus, and the twelve test corpora as twelve shows).

Single- and *cross-show DER* results are similar between the two systems: the overall mean of the *cross-show DER* evaluation is slightly higher with the ILP system (14.21% for the ILP system against 14.73% for the NCLR system). The trend is reversed with the overall means of the *single-show DER* evaluation (8.69% for the NCLR system against 8.85% for the ILP system).

We measured the computational time of the two overall clustering methods on each corpus. ILP clustering processing time is on average 8.66 times faster than NCLR. The speed factor depends upon the processed data; that factor ranges from 2.90 to 17.67 in our experiments. The average processing time of ILP clustering compared to NCLR clustering on corpora of same total duration is as follows:

5-hour corpora: 18 minutes vs 1:50 hours,
10-hour corpora: 1:58 hours vs 10:57 hours,
15-hour corpora: 3:55 hours vs 60:24 hours.

Adding new shows in the collection is not penalizing since the ILP clustering is faster to process than the NCLR one. In this context, the overall clustering must obviously be performed from the beginning, however, already learned speaker models can be reused. We do not have a valid explanation for the poor results obtained on test corpora 1, 6 and 12. The only noticeable difference between these corpora and the others is the proportion of female speakers which is twice as high. Nevertheless, the separation of the overall clustering stage according to speaker genders, by using two gender-dependent UBM, should give better results and allow to further reduce the computational time. This improvement is easy to perform since gender detection is realized during the individual processing stage.

On a few corpora, we notice that the use of an overall clustering improves the *single-show DER* of the individual processing stage regardless the overall clustering method chosen. The overall *single-show DER* mean of the individual processing stage is 9.06% against 8.69%, when using an overall NCLR clustering and 8.85% when using an overall ILP clustering.

5. Conclusions

We proposed a new clustering approach adapted to the cross-show diarization task. In this approach, the speakers are modeled by i-vectors and the classification itself is expressed as an ILP problem dealing with distances between i-vectors. Performance of the system implementing overall ILP clustering is comparable, in terms of *single-* and *cross-show DER*, to that of the implementation of the overall NCLR clustering.

The overall ILP clustering is more effective than overall NCLR, in terms of processing speed, while remaining reasonable in terms of memory consumption. This method is particularly appropriate for dealing with large collections.

6. References

[1] 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, Italie, 2011.

[2] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italie, 2011.

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

[4] 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, UK, 2009.

[5] V. B. Le, O. Mella, and D. Fohr, "Speaker diarization using normalized cross-likelihood ratio," in *Proceedings of Interspeech*, Antwerp, Belgique, 2007.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," in *Proceedings of IEEE TASLP*, vol. 19, 2011, pp. 788–798.

[7] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Proceedings of Interspeech*, Florence, Italie, 2011.

[8] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Proceedings of Interspeech*, Florence, Italie, 2011.

[9] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Odyssey Workshop*, Singapore, 2012.