An Open-source State-of-the-art Toolbox for Broadcast News Diarization

Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, Sylvain Meignier

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

Abstract
This paper presents the LIUM open-source speaker diarization toolbox, mostly dedicated to broadcast news. This tool includes both Hierarchical Agglomerative Clustering using well-known measures such as BIC and CLR, and the new ILP clustering algorithm using i-vectors. Diarization systems are tested on the French evaluation data from ESTER, ETAPE and REPERE campaigns.

Index Terms: speaker diarization, broadcast news, open-source

1. Introduction
Speaker diarization, the “who spoke when” task, consists in annotating recordings with labels representing the speakers. This task is performed without any prior information: neither the number of speakers, nor their identities, nor samples of their voices are available.

Since 2004, the state-of-the-art system for broadcast news speaker diarization is composed of 5 steps. First, music and jingle regions are removed using a Viterbi decoding. Next, an acoustic segmentation followed by a Hierarchical Agglomerative Clustering (HAC) splits and then groups the signal into homogeneous parts according to speakers and background. In this step, each segment or cluster is modeled by a Gaussian distribution with a full covariance matrix and the Bayesian Information Criterion (BIC) is employed both as similarity measure and as stop criterion. Then, a Gaussian Mixture Model (GMM) is trained for each cluster via the Expectation-Maximization (EM) algorithm. The signal is then re-segmented through a Viterbi decoding. The system finally performs another HAC, using the Cross-Likelihood Ratio (CLR) measure and GMMs trained with the Maximum A Posteriori algorithm (MAP). This kind of system obtained the best results at NIST RT’04 fall [1], ESTER [2,3], ETAPE [4] and, more recently, REPERE [5] evaluation campaigns.

LIUM_SpkDiarization is a simple program for those who only need to perform speaker diarization for their own applications (speech recognition, speaker recognition, multimedia indexing, etc.). The execution of the JAR archive calls upon the

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

made available at the 2010 Sphinx Workshop [6]. This paper presents a review of the tools included in the toolkit, and describes the latest improvements.

In the next section, we present other diarization tools. Then, we describe the choices made for this tool before describing the diarization system for broadcast news, in section 3. The single-show speaker diarization system is described in section 4 and the cross-show speaker diarization is reported in section 5. In section 6 we describe experiments performed on various corpora and give results. Finally, section 7 describes a video shot boundary detector to illustrate how this toolbox could be employed in other fields.

2. Other diarization tools
Several toolkits distributed under open-source licenses are available on the web. One of the oldest is the CMU Segmentation tool which was released in 1997 [7]. It was developed during the former NIST broadcast news evaluation campaign to address specifically the task of diarization for automatic speech recognition.

AudioSeg [8], under the GPL license, is a toolbox developed by IRISA during the ESTER campaign in 2005. It includes an audio activity detector, BIC/GLR or KL2 segmentation and clustering tools, as well as a Viterbi decoder. Note that CLR-based clustering is not available.

The speech recognition library ALIZEÉ [9] also includes speaker diarization tools. Diarization is based on the E-HMM method [10] in which segmentation and clustering are done iteratively and jointly. Performance is better when dealing with meetings and phone conversations rather than broadcast news.

The speech recognition toolkit SHoUT includes a speech/non-speech detector and a diarization tool. This tool seems to be well adapted for recordings of meetings as shown by the results reported in [11].

Recently, IDIAP published DiarTK [12] where clustering and segmentation are based on the information bottleneck principle. It was developed specifically for meetings recorded using multiple distant microphones or microphone arrays.

3. Guidelines of the toolkit
LIUM_SpkDiarization is developed in Java to minimize dependency problems with the various operating systems and libraries. This tool is distributed as one self-contained JAR archive, which can be run directly, with no need of additional third-party packages. Indeed, all the required packages are included in the precompiled version, as well as models (UBM, gender or speech/non-speech models).

LIUM_SpkDiarization is a simple program for those who only need to perform speaker diarization for their own applications (speech recognition, speaker recognition, multimedia indexing, etc.). The execution of the JAR archive calls upon the
speaker diarization method dedicated to broadcast news recordings. The acoustic parameters (MFCC) are computed directly by the program using Sphinx4 (one of the included third-party packages).

LIUM SpkDiarization is also designed to carry out research. The toolkit is composed of elementary programs, such as segment and cluster generator, decoder, and model trainers. Fitting those elementary tools together in a shell script is an easy way to develop a specific diarization system.

This toolbox can employ alien features. By default, the tools read Sphinx MFCC files; however, it also reads HTK, SPPro files, as well as text files where each line corresponds to a frame. Regarding speaker models, the toolkit has its own binary file format but is also able to read ALIZE files.

LIUM SpkDiarization allows the management of a speaker diarization which works with clusters of segments. A segment is an object defined by the name of the recording it belongs to, its start time in the recording, and its duration. These three values identify a unique region of signal in a set of audio recordings. The elementary programs that compose the toolkit are able to work with a set of audio recordings. This way, models can be learnt from parts of recordings without having to cut them, making cross-show diarization systems easy to develop (see section 5).

4. Single-show diarization system

The proposed diarization system was inspired by the system [6] which won the RT’04 fall evaluation as well as the ESTER evaluation in 2005. It was developed for transcription and diarization tasks, and with the goal of minimizing both Word Error Rate and speaker error rate (i.e. Diarization Error Rate – DER).

Automatic transcription requires accurate segment boundaries. Segment boundaries have to be set within non-informative zones such as filler words. Indeed, having a word cut by a boundary disturbs the language model and increases the Word Error Rate; whereas the main objective of speaker diarization is to produce the minimum number of pure clusters, each of which corresponds to only one single speaker.

Non-speech segments must be rejected to save computation time in both tasks: in transcription, non-speech segments generate insertion of words and, in diarization, non-speech segments make the speaker models less accurate.

4.1. Diarization for speech transcription

This system is widely described in [6]. Only an outline is given below:

1. The acoustic frames are composed of 13 MFCCs with coefficients $C_0$, and feature normalization is not applied. In the two last steps, the acoustic frames are composed of 12 MFCCs ($C_0$ is removed) as well as the first-order derivatives of those coefficients.

2. The speaker segmentation is composed of two passes. The first detects the instantaneous change points using Generalized Likelihood Ratio (GLR) distance. The second pass over the signal uses BIC distances between speakers in order to fuse consecutive segments that are found to correspond to the same speaker.

3. The Hierarchical Agglomerative Clustering merges the two closest clusters at each iteration until the best BIC distance is positive. In this step and in the previous one, speakers are modeled by a Gaussian distribution with a full covariance matrix.

4. A Viterbi decoding is performed to generate a new segmentation using GMMs as speaker models. $C_0$ is removed from the features and first-order derivatives of the coefficients are added.

5. A segmentation into speech/non-speech is obtained using a Viterbi decoding in order to remove music and jingle regions. The initial segmentation is less accurate if this detection step is done before the speaker segmentation, like most of the diarization systems do. The main reason is that speaker segmentation cannot make correct decisions on the 2.5 seconds at the beginning and at the end of a segment. This problem is exacerbated when the speech/non-speech segmentation is done first because it generates many segments to cut.

6. The gender and bandwidth of clusters are detected using a GMM over normalized frames as in CLR-like clustering described below. The segment boundaries produced are not perfect: for example, some of them fall within words. In order to avoid this situation, boundaries are moved to low energy regions. Long segments are also cut in order to yield segments shorter than 20 seconds.

4.2. Diarization optimized for speaker

Step 6 described above was not required in order to perform an accurate speaker diarization, because only gender and bandwidth information, as well as boundary location, are needed for speech transcription. In the current case, another clustering step is performed instead.

4.2.1. CLR-like Clustering

The second Hierarchical Agglomerative Clustering merges the two closest clusters at each iteration until the similarity between the two candidate clusters is positive. Most systems use the CLR measure to estimate similarity between clusters [14], but it has been shown that normalized CLR measure [15] gives better results in most cases. The difference between CLR and normalized CLR resides only in the denominators, which are the UBM log-likelihood scores and the log-likelihood scores of the clusters respectively. In this step, the first order derivative is added to the 12 MFCCs ($C_0$ is removed) and the features are normalized: short-term windowed mean and variance are computed to normalize the frame, and a feature warping normalization is
applied. Only the means of the Universal Background Model (UBM) are adapted for each cluster in order to obtain the model for its speaker.

4.2.2. ILP/i-vectors clustering

A new clustering algorithm in which the problem is solved as an Integer Linear Programming (ILP) problem has been proposed [18]. Experiments were carried out using i-vectors to model and measure the similarity between clusters. This method is now available in the LIUM_SpkDiarization toolbox.

The i-vectors is a state-of-the-art method in the field of Speaker Verification [17]. The acoustic data of a speaker are compacted into a low-dimension vector, which only retains the relevant information about the speaker. This approach was first adapted to 2-speaker diarization using the $k$-means algorithm to find utterances of the two speakers within a phone call recording [19].

In broadcast news, the number of speakers is unknown a priori. The proposed method replaces the standard CLR-like clustering step. According to an initial speaker segmentation, an i-vector is extracted from each cluster using 19 MFCCs parameters completed with energy, their first and second order derivatives, and a 1024 GMM-UBM. The $N$ resulting i-vectors are then normalized in an iterative process [19]. The clustering problem consists in jointly minimizing the number $K$ of cluster centers as well as the dispersion of i-vectors within each cluster. The value $K \in \{1, \ldots, N\}$ is to be automatically determined.

This clustering problem is expressed as an ILP problem, where the objective solving function (eq. 1) is minimized subject to constraints:

\[
\text{Minimize} \quad \sum_{k=1}^{N} x_{k,k} + \frac{1}{D} \sum_{k=1}^{N} \sum_{j=1}^{N} d(k,j) x_{k,j} \quad (1)
\]

Subject to

\[
x_{k,j} \in \{0, 1\} \quad \forall k, \forall j \quad (1.2)
\]

\[
\sum_{k=1}^{N} x_{k,j} = 1 \quad \forall j \quad (1.3)
\]

\[
d(k,j) x_{k,j} \leq \delta \quad \forall k, \forall j \quad (1.4)
\]

\[
x_{k,j} - x_{k,k} \leq 0 \quad \forall j \quad (1.5)
\]

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 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 $j$. $D$ 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 $\delta$ empirically determined (eq. 1.4). Equation 1.5 ensures that the cluster $k$ is selected if an i-vector is assigned to cluster $k$.

5. Cross-show diarization system

The recently introduced cross-show speaker diarization [3, 4, 5] aims to expand the diarization task to a broader context, where speakers appearing in different recordings of the same show (the cross-show speakers) will always be identified in the same way in every recording. Each show from a collection is first individually processed with the single-show speaker diarization system, as described in section 4. Then it is processed collectively using a CLR or ILP clustering. Experiments showed that ILP clustering provides a better speed/accuracy trade-off [7].

6. Single- and cross-show diarization evaluation

6.1. Data

Evaluation has been performed on three different corpora:

- The test corpus of the ESTER evaluation campaign is composed of 26 broadcast news recordings coming from seven French radio stations. Most of the corpus contains prepared speech. 15 hours of the shows are fully annotated.
- The test corpus of the ETAPE evaluation campaign is composed of 15 TV shows recorded from two French TV channels and one French radio station. Most of the corpus contains spontaneous speech or very spontaneous speech. 7 hours of the shows are fully annotated.
- The test corpus of the REPERE 2014 evaluation campaign is composed of 3 hours of data drawn from 28 TV shows recorded from French TV channels BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, from 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

6.2. 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 match between references and hypothesis speaker labels. The scoring tool we used was developed by the LNE as part of the ETAPE and the REPERE campaigns.

The LNE evaluation tool computes two different error rates. The single-show DER is computed considering each show independently. In this context, the overall DER corresponds to the mean of individual DERs (one per show), weighted by the duration of each show. The cross-show DER is computed over all the shows, taking into account multiple appearances of the same speaker in several shows. In order to assess the cross-show diarization task, the same label must necessarily identify speakers appearing in several shows.

6.3. Results and comments

Single- and cross-show speaker DER on the three test corpora are reported in Table 2. Both CLR and ILP clustering methods were tested. The cross-show CLR clustering was performed using the single-show CLR segmentations, and the cross-show ILP clustering with the single-show ILP segmentations. Because the segmentation is unchanged during the clustering steps, False Alarms (FA) and Missed Detections (MISS)

---

References:

1. http://catalog.elra.info

*The French National Laboratory of Metrology and Testing*
obtained on a corpus remains the same, regardless of the clustering method or the evaluation metric. The values are as follows: 1.58% FA and 0.98% MISS on the ESTER 2 test corpus; 4.10% FA and 4.24% MISS on the ETAPE test corpus; 1.10% FA and 3.83% MISS on the REPERE 2012 test corpus.

It is important to note that the systems were not specifically tuned according to the test corpora, and better DER could be obtained by doing so. Single- and cross-show CLR thresholds were set to 0.35 and 0.80 respectively. Single- and cross-show ILP distance thresholds were set to 100 and 60 respectively.

<table>
<thead>
<tr>
<th>Type</th>
<th>Corpus</th>
<th>Single-show DER</th>
<th>Cross-show DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR</td>
<td>ESTER 2</td>
<td>11.27 %</td>
<td>20.43 %</td>
</tr>
<tr>
<td></td>
<td>ETAPE</td>
<td>21.57 %</td>
<td>27.79 %</td>
</tr>
<tr>
<td></td>
<td>REPERE</td>
<td>17.19 %</td>
<td>23.95 %</td>
</tr>
<tr>
<td>ILP</td>
<td>ESTER 2</td>
<td>8.35 %</td>
<td>17.51 %</td>
</tr>
<tr>
<td></td>
<td>ETAPE</td>
<td>24.49 %</td>
<td>26.31 %</td>
</tr>
<tr>
<td></td>
<td>REPERE</td>
<td>15.46 %</td>
<td>19.59 %</td>
</tr>
</tbody>
</table>

Table 2: Single- and cross-show DER on the ESTER 2, ETAPE and REPERE 2012 test corpora.

7. Extension to computer vision tasks

Recent studies show that there is quite a large similarity between audio and video tasks, especially for solving several identification and segmentation problems. In [7], the authors show that successful speaker identification techniques are also very good for face identification. Furthermore, in [7], the authors show that speaker segmentation techniques using the generalized likelihood ratio (GLR) and the Bayesian information criterion (BIC) can also be used for other segmentation tasks such as video shot boundary detection or TV program boundary detection. In this work, we give the example of shot boundary detection to show how easily LIUM_SpkDiarization can be extended.

Shot boundary detection (SBD) is a well-known segmentation process. It aims at breaking down the massive volume of video into smaller chunks. Quite a lot of approaches have been proposed in the literature in the last two decades [9]. The main ideas behind their assumption is that 1) shots, similarly to audio turns, are homogeneous segments, and 2) color features similarly to MFCC features, can be modeled by Gaussian distributions.

Because of this similarity, the implementation of SBD in LIUM_SpkDiarization is straightforward. Although the visual feature extractor module is missing, LIUM_SpkDiarization has a good option that allows the user to use external features, and in various file formats. This option makes the integration of new tasks very simple. Consequently, results obtained by our SBD implementation are equivalent to the best state-of-the-art systems (F-measure > 95% on TV Broadcast news and debates).

8. Conclusions

This paper presents a diarization toolkit mostly dedicated to broadcast news recordings. Developed by the LIUM, this toolkit is published under the GPL license. When used for broadcast news diarization, it represents the state-of-the-art in terms of performance; in addition, only minor work is required in order to reuse its components for other tasks. The LIUM_SpkDiarization toolbox, as well as its documentation and some examples, is available at: [http://www-lium.univ-lemans.fr/en/content/liumspkdiarization](http://www-lium.univ-lemans.fr/en/content/liumspkdiarization)
9. References


