



# Speaker Diarization using Normalized Cross Likelihood Ratio

Viet-Bac Le, Odile Mella, Dominique Fohr

Speech Group, LORIA  
 Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France  
 {vietbac.le, odile.mella, dominique.fohr}@loria.fr

## Abstract

In this paper, we present the Normalized Cross Likelihood Ratio (NCLR) and the advantages of using it in a speaker diarization system. First, the NCLR is used as a dissimilarity measure between two Gaussian speaker models in the speaker change detection step and its contribution to the performance of speaker change detection is compared with those of BIC and Hostelling's  $T^2$ -Statistic measures. Then, the NCLR measure is modified to deal with multi-gaussian adapted models in the cluster recombination step. This step ends the step-by-step speaker diarization process after the BIC-based hierarchical clustering and the Viterbi re-segmentation steps. By comparing the NCLR measure with the CLR (Cross Likelihood Ratio) one, more than 30% of relative diarization error is reduced in ESTER evaluation data.

**Index Terms:** speaker diarization, speaker change detection, cluster recombination, NCLR.

## 1. Introduction

Many applications in speech processing domain (like automatic speech transcription, spoken document retrieval ...) need speaker diarization to partition and classify an audio document into homogeneous segments according to speakers. In speaker diarization, it has been assumed that no *a priori* information knowledge about speakers (like number of speakers, reference data or model of speakers...) is available.

Almost recent speaker diarization systems have a general architecture called multistage [1] or step-by-step speaker diarization [2]. Such a system contains the following steps:

- *Speech activity detection:* an audio stream may consist of some acoustic activities like speech, noise, music, background conversation, advertisement. Therefore, non-speech regions must be detected and removed from the audio stream.

- *Speaker change detection:* inside every speech region, a speaker change (or speaker turn) detector is used to find points in the audio stream which are candidates for speaker change points. To do this, a distance is computed between two Gaussian modeling data of two adjacent given-length windows. By sliding both windows on the whole audio stream, a distance curve is obtained. A peak in this curve is thus considered as a speaker change point if its distance value is higher than a predefined threshold.

- *Segment recombination:* too many speaker turn points detected during the previous step results in a lot of false alarms. A segment recombination using BIC is needed to recombine adjacent segments uttered by the same speaker. The BIC threshold value must be tuned on a development corpus in order to reduce the number of false alarms without increasing the number of new missed detection errors [3].

- *Speaker clustering:* in this step, speech segments of the same speaker are clustered. Top-down clustering techniques [2] or bottom-up hierarchical clustering techniques [1, 2]

using BIC can be used.

- *Viterbi re-segmentation:* the previous clustering step provides enough data for every speaker to estimate multi-gaussian speaker models. These models are used by a Viterbi algorithm to refine the boundaries between speakers.

To perform speaker diarization, several of these steps apply dissimilarity measures between Gaussian or multi-gaussian speaker models. In this paper, we present the use of Normalized Cross Likelihood Ratio distance measure in SELORIA<sup>1</sup> Speaker Diarization system. Section 2 introduces the dissimilarity measures compared in our study: first, some well-known distance metrics between two Gaussian models such as KL2, Hostelling's  $T^2$ -Statistic, BIC and then NCLR measures. The NCLR measure between two adapted speaker models obtained from Universal Background Model are also proposed in section 2. Section 3 presents the architecture of SELORIA diarization system. The experimental results are shown in section 4. Section 5 concludes the work and gives some future perspectives.

## 2. Dissimilarity measures between two speaker models

### 2.1. Some distance metrics

The symmetric Kullback-Leibler distance (KL2) was firstly used in speaker segmentation by M. Siegler [4]. If two audio segments  $X_i$  and  $X_j$  are modeled by two Gaussian models  $M_i$  and  $M_j$ , then the KL2 distance between these segments can be calculated as:

$$KL2 = \frac{1}{2}(\mu_i - \mu_j)^T (\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + \frac{1}{2}tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I) \quad (1)$$

Another distance named Hostelling  $T^2$ -statistic distance (or  $T^2$  distance) was also used in [5]. In this distance metric, we assume that the covariances of two Gaussian models are equal but unknown. Thus the difference between two models is:

$$T^2 = \frac{N_i N_j}{N_i + N_j} (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) \quad (2)$$

where  $N_i, N_j$  are the number of frames of speech data  $X_i, X_j$ .

When the audio segment is short (<5s), R. Huang has said that the KL2 distance work incorrectly because of insufficient data in the estimation of the covariance. The  $T^2$  distance can overcome this problem because, under the equal covariance assumption, it uses more data to estimate the covariance and reduce the impact of insufficient data in the estimation [5]. If audio segments are longer (>5s), other distances based on likelihood ratio such as BIC and NCLR may work better.

### 2.2. Bayesian Information Criterion

Bayesian Information Criterion (BIC) is a model selection

<sup>1</sup> SELORIA: système de SEgmentation en LOcuteurs du LORIA

criterion which maximizes the log-likelihood penalized by the complexity of the model [6]. Let  $X = (x_1, x_2, \dots, x_{N_x})$  be audio data,  $N_x$  be the size in frame of  $X$  and  $M$  be the desired parametric model, the BIC criterion is defined as:

$$BIC(M) = \log L(X|M) - \lambda \frac{1}{2} P \log(N_x) \quad (3)$$

where  $\log L(X|M)$  is the log-likelihood function of data  $X$  and its model,  $\lambda$  is the penalty weight (theoretically,  $\lambda=1$ ),  $P$  is the model perplexity or number of independent parameters of  $M$ .

To calculate the BIC-based distance between two audio segments or clusters  $X_i$  and  $X_j$ , two hypotheses are tested [7]:

- $H_0$ : both  $X_i$  and  $X_j$  are modeled by the same model  $M$ .
- $H_1$ :  $X_i$  and  $X_j$  are modeled by 2 different models  $M_i$  and  $M_j$ .

Let  $X = X_i X_j$  be the merged data of  $X_i$  and  $X_j$ . Let  $N_i, N_j, N$  be the frame sizes of  $X_i, X_j$  and  $X$ , respectively. By applying the model selection problem described above to test  $H_0$  and  $H_1$ , we have:

$$BIC(M) = \log L(X|M) - \lambda \frac{1}{2} P \log(N) \quad (4)$$

$$BIC(M_i, M_j) = \log L(X_i|M_i) + \log L(X_j|M_j) - \lambda \frac{1}{2} (2P) \log(N) \quad (5)$$

and the BIC score is:  $\Delta BIC = BIC(M_i, M_j) - BIC(M) =$

$$= \frac{1}{2} [N \log(|\Sigma|) - N_i \log(|\Sigma_i|) - N_j \log(|\Sigma_j|)] - \lambda \frac{1}{2} P \log(N) \quad (6)$$

where  $|\Sigma_i|, |\Sigma_j|, |\Sigma|$  are the determinants of covariance matrices for models  $M_i, M_j, M$ , respectively.

In fact, the BIC-based distance can be applied both during the speaker segmentation and the speaker clustering steps in a speaker diarization system.

### 2.3. Normalized Cross Likelihood Ratio (NCLR)

#### 2.3.1. NCLR between two Gaussian speaker models

Normalized Cross Likelihood Ratio was firstly presented as a distance measure between two speaker models by D. Reynolds [8]. It was used to select a background speaker in a speaker identification and verification system. Given two speaker models  $M_i$  and  $M_j$ , the NCLR distance is defined as:

$$NCLR(M_i, M_j) = \frac{1}{N_i} \log \left( \frac{L(X_i|M_i)}{L(X_i|M_j)} \right) + \frac{1}{N_j} \log \left( \frac{L(X_j|M_j)}{L(X_j|M_i)} \right) = \frac{1}{N_i} [\log L(X_i|M_i) - \log L(X_i|M_j)] + \frac{1}{N_j} [\log L(X_j|M_j) - \log L(X_j|M_i)] \quad (7)$$

We note that the cross likelihood ratio  $\frac{L(X_i|M_i)}{L(X_i|M_j)}$  measures

how well speaker model  $M_j$  scores with speaker data  $X_i$  relative to how well speaker model  $M_i$  scores with its own data  $X_i$  [8]. In equation (7), each log-likelihood value is normalized by the amount of corresponding data.

#### 2.3.2. NCLR between two adapted GMM models

As we know, speaker model adaptation techniques could be applied to speaker diarization. In this case, a Universal Background Model (UBM) is firstly trained with a huge amount of audio data according to the gender (male, female) and the channel condition (studio, telephone) [9]. Then, speaker models were derived by adapting the UBM model's parameters with speaker speech data. The adaptation method usually used is the MAP (Maximum A Posteriori) adaptation [10]. Comparing to the previous dissimilarity measures, the

log-likelihood function  $\log L(X|M)$  of data  $X$  given the model  $M$  must be replaced by the log-likelihood ratio function (called LLR) because speaker model  $M^*$  is derived or adapted from the UBM model. Log-likelihood ratio is defined as [8]:

$$LLR(X, M^*) = \log L(X|M^*) - \log L(X|UBM) \quad (8)$$

where  $L(X|M^*)$  is likelihood of data  $X$  given the adapted model  $M^*$  and  $L(X|UBM)$  is likelihood of data  $X$  given the UBM model.

By using Cross Likelihood Ratio (CLR) distance measure, C. Barras applied a speaker model adaptation in an extra step for speaker clustering (called speaker recombination) [1]. The CLR measure was firstly used in [11] to compute the distance between two adapted speaker models and it was defined as:

$$CLR(M_i^*, M_j^*) = \frac{1}{N_i} \log \left( \frac{L(X_i|M_i^*)}{L(X_i|UBM)} \right) + \frac{1}{N_j} \log \left( \frac{L(X_j|M_j^*)}{L(X_j|UBM)} \right) \quad (9)$$

where  $M_i^*, M_j^*$  are adapted speaker models for speaker  $i$  and speaker  $j$ , respectively.

In our work, we have extended the NCLR distance measure presented in equation (7) to adapted speaker model case. Indeed, by replacing the log-likelihood for Gaussian speaker model  $\log L(X|M)$  in equation (7) with the log-likelihood ratio function for adapted speaker models  $LLR(X, M^*)$ , we have:

$$NCLR(M_i^*, M_j^*) = \frac{1}{N_i} [LLR(X_i, M_i^*) - LLR(X_i, M_j^*)] + \frac{1}{N_j} [LLR(X_j, M_j^*) - LLR(X_j, M_i^*)] \quad (10)$$

By applying equation (8), we get:

$$LLR(X_i, M_i^*) - LLR(X_i, M_j^*) = \log L(X_i|M_i^*) - \log L(X_i|UBM) - \log L(X_i|M_j^*) + \log L(X_i|UBM) = \log \frac{L(X_i|M_i^*)}{L(X_i|M_j^*)} \quad (11)$$

Similarly, we have:

$$LLR(X_j, M_j^*) - LLR(X_j, M_i^*) = \log \frac{L(X_j|M_j^*)}{L(X_j|M_i^*)} \quad (12)$$

Finally, by applying (11) and (12) to (10), the NCLR distance between two adapted speaker models is calculated as follows:

$$NCLR(M_i^*, M_j^*) = \frac{1}{N_i} \log \frac{L(X_i|M_i^*)}{L(X_i|M_j^*)} + \frac{1}{N_j} \log \frac{L(X_j|M_j^*)}{L(X_j|M_i^*)} \quad (13)$$

In the cluster recombination experiments, we will compare the performance of CLR- and NCLR-based clustering method in the French ESTER data.

## 3. SELORIA Diarization System

The architecture of SELORIA, shown in Figure 1, is similar to the general architecture introduced in section 1 except the two following main modifications.

Firstly, we don't use any speech activity detection to pre-segment the audio stream into speech and non-speech (noise, music, advertisement...) regions.

Secondly, after the Viterbi re-segmentation and refinement process, we apply a second speaker clustering step (called cluster recombination) using adapted speaker models. This step is similar to the Speaker Identification Clustering step using the CLR distance measure presented in [1]. But instead of the CLR measure, we investigate the NCLR measure as presented in section 2.3.2. Both the CLR and the NCLR measure work on adapted speaker models based on

UBM models with 128 diagonal Gaussians.

Briefly, the other distinctive features of SELORIA modules are:

- *Speaker Change Detection*: our speaker change detector was based on  $T^2$ , NCLR and BIC dissimilarity measures between two Gaussian models.

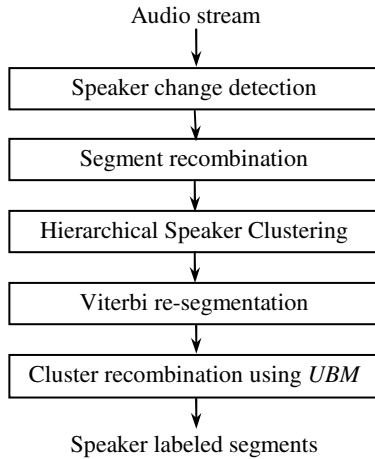


Figure 1: SELORIA System architecture

- Both *Segment recombination* and *hierarchical speaker clustering* are based on BIC measure.

- *Viterbi re-segmentation* and *cluster recombination modules*. We used and adapted the *mClust* toolkit developed by the LIUM Laboratory, France [12]. The Viterbi decoding tool works with GMM models trained by EM to refine boundaries between speaker's segments. For cluster recombination, speaker models are adapted from a UBM model with 128 diagonal Gaussians built by merging 4 gender and bandwidth dependent GMMs. We modified the CLR-based cluster recombination module of the *mClust* toolkit to work with our NCLR-based clustering method.

## 4. Experiments

### 4.1. Experimental Framework

#### 4.1.1. Database

The SELORIA diarization system was evaluated on the speaker diarization task of the Evaluation Campaign for the Rich Transcription of French Broadcast News (ESTER) [13]. This campaign covered three categories of tasks: transcription (T), segmentation (S) and information extraction (E). We focused on speaker diarization (SRL) task in category S.

We used as development corpus, called ESTER-Dev, 4 hours of French radio broadcast news extracted from ESTER training corpus.

The test corpus, called ESTER-Eval, is composed of 10 hours of radio broadcast news shows.

#### 4.1.2. Performance evaluation measures

Both speaker change detection and speaker diarization performance measures are evaluated in our work. Firstly, for speaker change detection, 3 performance measures are used: Recall (RCL), Precision (PRC) and F-measure (F). These measures are defined as:

$$RCL = \frac{\# \text{ of correct detected speaker changes}}{\# \text{ of speaker changes}} \quad (14)$$

$$PRC = \frac{\# \text{ of correct detected speaker changes}}{\# \text{ of detected speaker changes}} \quad (15)$$

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \quad (16)$$

As no speech activity detector was used in our system, we compute speaker diarization error (SER) instead of overall diarization error (DER) which is expressed in terms of miss speech, false alarm and SER. The speaker diarization error is defined as:

$$SER = \frac{\# \text{ of frames incorrectly speaker-labeled}}{\# \text{ of speech frames}} \quad (17)$$

### 4.2. Experimental Results

#### 4.2.1. Speaker change detection evaluation

These experiments were conducted in the framework of the STORECO project whose aim is to detect speaker changes in order to help captioning of TV shows and movies. In this context, we were mainly interested in well detecting speaker changes and not necessarily in determining the number of different speakers or in attributing each speech segment to the good speaker. This is the reason why we especially focused on speaker change detection evaluation. Moreover, we wanted to know if using only the two first steps of SELORIA system achieves sufficient performance. Therefore, the recall, precision and F-measure were evaluated after the segment recombination and after the Viterbi re-segmentation steps.

Before this evaluation, we tuned some parameters on the ESTER-Dev development corpus. Optimal values of minimal segment duration for speaker change detection are 2s, 2.5s and 3s for  $T^2$ , BIC and NCLR-based method, respectively. Moreover, full covariance matrices work better than diagonal ones. Finally, the threshold  $\lambda$  for the BIC-based segment recombination procedure is set at 2.5.

Table 1 shows speaker change detection (SCD) results after the segment recombination module on both ESTER-Dev and ESTER-Eval corpora. It is clear that BIC- and NCLR-based speaker change detection results are similar: NCLR works slightly better than BIC on ESTER-Dev corpus but it is overcome by BIC on ESTER-Eval corpus.

SCD Method	ESTER-Dev			ESTER-Eval		
	%RCL	%PRC	%F	%RCL	%PRC	%F
$T^2$	55.39	51.78	51.55	49.65	59.81	53.51
BIC	<b>59.32</b>	58.97	57.99	<b>53.73</b>	<b>67.82</b>	<b>59.13</b>
NCLR	58.19	<b>61.57</b>	<b>58.85</b>	51.18	66.96	57.36

Table 1: Speaker change detection measures evaluated after the segment recombination step

Table 2 shows the same evaluation after the Viterbi decoding step on the ESTER-Eval corpus. The results show that speaker clustering and Viterbi re-segmentation steps significantly improve speaker change detection. Besides, using NCLR measure in the speaker change detector gives better performance than BIC and  $T^2$ .

SCD Method	ESTER-Eval		
	%RCL	%PRC	%F
$T^2$	61.13	77.46	67.46
BIC	62.13	77.20	68.29
NCLR	<b>63.76</b>	<b>78.56</b>	<b>69.87</b>

Table 2: Speaker change detection measures evaluated after the Viterbi decoding step

#### 4.2.2. Speaker diarization error (SER)

We note that the SER evaluated after the Viterbi decoding step on the ESTER-Eval corpus are 17.0. In order to demonstrate the advantage of the cluster recombination step, we computed the SER achieved by the whole system. The influence of the CLR- and NCLR-based methods in this cluster recombination step was also evaluated.

As presented in figure 2, the optimal thresholds for CLR- and NCLR-based methods which were separately fixed on ESTER-Dev corpus are equal to 1.5 and 3.8, respectively.

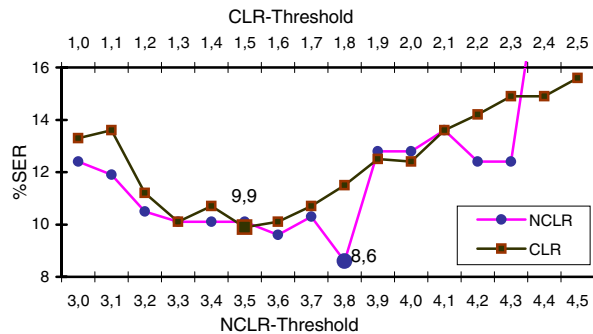


Figure 2: Speaker diarization error of CLR-based and NCLR-based cluster recombination on the ESTER-Dev corpus as a function of CLR and NCLR thresholds

We then applied these optimal thresholds to evaluate the speaker diarization error obtained by both CLR- and NCLR-based methods on the ESTER-Eval corpus. The SER is reduced from 10.8% with CLR-based method to 7.3% with NCLR-based method. Although the optimal thresholds were fixed on ESTER-Dev corpus, the results (SER) from changing CLR and NCLR thresholds on the ESTER-Eval corpus are also illustrated in figure 3.

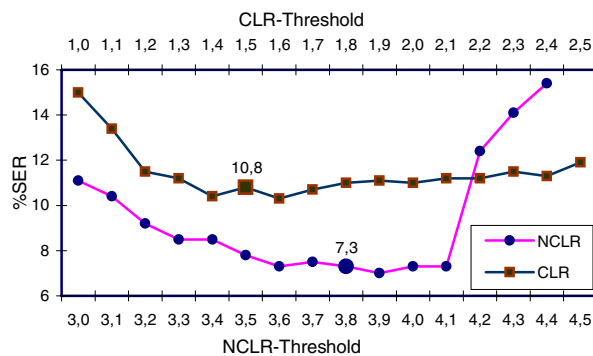


Figure 3: Performance of different methods of cluster recombination on ESTER-Eval corpus

## 5. Conclusions

In summary, we have presented in this paper the advantages of using the Normalized Cross Likelihood Ratio (NCLR) in a speaker diarization system. On the one hand, using NCLR as a dissimilarity measure between two Gaussian speaker models in speaker change detection achieves better performance than using BIC or  $T^2$ . On the other hand, NCLR-based method performs better than CLR-based method in the final cluster recombination step using multi-gaussian adapted speaker models. An improvement of 32.4% is achieved for the SER computed on the ESTER-Eval corpus of the Evaluation Campaign for the Rich Transcription of French Broadcast News. In addition, the SELORIA system achieved a better SER (7.3%) than the best SER obtained in the ESTER

Campaign (9.8%) [13].

In the future, we will investigate the use of NCLR measure in other steps of speaker diarization such as segment recombination or hierarchical cluster clustering.

## 6. Acknowledgements

The authors would like to thank the “Speech” team at LIUM for its *mClust* toolkit and the French Ministry for Research for its support of the STORECO project through the program RIAM (*Recherche et Innovation en Audiovisuel et Multimédia*).

## 7. References

- [1] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.-L., “Multistage Speaker Diarization of Broadcast News”, *IEEE Trans. on ASLP*, vol. 14, no. 5, pp. 1505-1512, September 2006.
- [2] Fredouille, C., Moraru, D., Meignier, S., Bonastre, J-F. and Besacier, L., “Step-by-step and Integrated approaches in broadcast news speaker diarization”, *Computer Speech and Language*, vol. 20, issues. 2-3, pp. 303-330, April-July 2006.
- [3] Delacourt, P. and Wellekens, C. J., “DISTBIC: A speaker-based segmentation for audio data indexing”, *Speech Communication*, vol. 32, pp. 111-126, 2000.
- [4] Sieglar, M., Jain, U., Raj, B. and Stern, R., “Automatic Segmentation, Classification and Clustering of Broadcast News Audio”, *DARPA Speech Recognition Workshop*, Chantilly, VA, February 1997.
- [5] Huang, R. and Hansen, J.H.L., “Advances in Unsupervised Audio Segmentation for the Broadcast News and NGSW Corpora”, *ICASSP'04*, vol. 1, pp. 741-744, Montreal, Canada, May 2004.
- [6] Schwarz, G., “Estimating the dimension of a model”, *The Annals of Statistics*, vol. 6, pp 461-464, 1978.
- [7] Chen, S. and Gopalakrishnan, P., “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion”, *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, Landsdowne, VA, 1998.
- [8] Reynolds, D., “Speaker identification and verification using Gaussian mixture speaker models”, *Speech Communication*, vol. 17, issue 1-2, pp. 91-108, 1995.
- [9] Reynolds, D., Quatieri, T. and Dunn, R., “Speaker verification using adapted gaussian mixture models”, *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.
- [10] Gauvain, J.-L. and Lee, C. H., “Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains”, *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, April 1994.
- [11] Reynolds, D., Singer, E., Carlson, B., O’Leary, G., McLaughlin, J. and Zissman, M., “Blind clustering of speech utterances based on speaker and language characteristics”, *ICSLP’98*, Sydney, Dec 1998.
- [12] Deleglise, P., Esteve, Y., Meignier, S. and Merlin, T., “The LIUM Speech Transcription System: a CMU Sphinx III-based System for French Broadcast News”, *Interspeech’05*, pp. 1653-1656, Lisbon, Portugal, 2005.
- [13] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J-F., and Gravier, G., “The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News”, *Interspeech’05*, pp. 1149-1152, Lisbon, Portugal, 2005.