IMPROVED SPEAKER DIARIZATION SYSTEM FOR MEETINGS

Elie El-Khoury, Christine Sénac, Julien Pinquier
Toulouse University, IRIT Laboratory
{khoury, senac, pinquier}@irit.fr

ABSTRACT

In this paper, we investigate new approaches to improve speech activity detection, speaker segmentation and speaker clustering. The main idea behind them is to deal with the problem of speaker diarization for meetings where error rates are relatively high. In opposition to existing methods, a new iterative scheme is proposed considering those three tasks as only one problem. New bidirectional source segmentation is proposed based on the GLR/BIC method. The well-known BIC clustering is also reviewed and a new unsupervised post-processing is added to increase clusters purity. Those new proposals applied on meeting data show a relative improvement of about 40% compared to a standard speaker diarization system.

Index Terms— speaker diarization, speaker segmentation, speaker clustering, speech activity detection, meetings

1. INTRODUCTION

In the context of speech processing in meeting data, one of the most difficult and unresolved problems is "speaker diarization". It consists in segmenting and clustering an audio file into its different speakers without *a priori* knowledge about their number. An overview of automatic speaker diarization systems is presented in [1]. They are usually composed of three common and essential stages which are generally independent: speech activity detection, speaker segmentation and speaker clustering.

However, due to the high interaction between speakers in meeting data, those stages become laborious tasks. In this paper, a new scheme for speaker diarization is proposed: it considers those three stages as only one iterative problem. Moreover, for each of those components, new proposals are given to improve performance of the global system.

This paper is organized as follow: in section 2, we present briefly our state-of-the-art system for Speech/Non-Speech separation and its behavior on meeting data. In section 3, we describe our existing speaker segmentation based on GLR/BIC algorithm and we propose two hypotheses for improvements. Then, in section 4, the well-known BIC clustering is reviewed and a post-processing step is added to enhance the purity of the resulting clusters. In section 5, we describe the whole scheme that cures simultaneously and iteratively the weaknesses of the above tasks. Experiments and results on EPAC[1] corpus are presented in section 6.

2. SPEECH ACTIVITY DETECTION

Existing methods for speech activity detection are often based on Gaussian Mixtures Models [2] for both Speech and Non-Speech components. Those models need learning and depend on the training data. Unsupervised methods use robust features like the 4Hz modulation energy [3]. The fusion of these two techniques was developed in our team and gave results among the best in the last ESTER evaluation campaign [4] on radio broadcast news. For more details about those methods, please refer to [5].

However, for meeting data, particularly on regions where two people talk simultaneously, the value of the 4Hz modulation energy is not always relevant. Due to a thresholding decision, this method may introduce additional missed detections and imprecise boundaries location of speech regions. Table 1 shows the results of our speech detection algorithm on two different databases: the first contains 10 hours of radio broadcast news (the test set of ESTER corpus) and the second one contains 7 hours of radio broadcast meeting (test set of EPAC corpus).

	ESTER'05	EPAC'08
Speech detection rate	96.4%	93.7%

Table1. Speech activity detection rate for news and meeting data

Moreover, speech activity detection is used as the first step in the most of existing speaker diarization systems. That is why final results are subject to high cumulative errors. To avoid those errors, we propose to postpone the decision step to later stages (cf. section 5).

3. SPEAKER SEGMENTATION

Recent papers in speaker change detection show that methods based on the Bayesian Information Criterion are the best among all existing approaches [1]. Moreover, in our

[1] EPAC is a French ANR project that aims to explore methods for information extraction and document structuring applied on meeting data: http://epac.univ-lemans.fr

previous work [6], we proposed a method for speaker segmentation that consists of applying the GLR (Generalized Likelihood Ration) algorithm several times until convergence to the best repartition of Gaussian distributions. Then, it uses the BIC (Bayesian Information Criterion) algorithm [7] to choose points that correspond to speakers change. The ΔBIC expression is:

$$\Delta BIC = \frac{N_X}{2}\log|\Sigma_X| - \frac{N_{X1}}{2}\log|\Sigma_{X1}| - \frac{N_{X2}}{2}\log|\Sigma_{X2}| - \lambda P \quad (1)$$

where the window X is divided into two sub-windows X_1 and X_2. Σ_X, Σ_{X1} and Σ_{X2} are respectively the covariance matrices of the acoustic vectors of X, X_1 and X_2. N_X, N_{X1} and N_{X2} are the numbers of those vectors. λ is the penalty coefficient and P the penalty term which depends on N_x and the dimension of the acoustic vectors (MFCCs).

Tests done on broadcast news show the efficiency of this method compared to other methods which also use the BIC but need many tuning parameters like in [8].

However, when applying this method on meeting data, some errors occurred in regions containing multiple speakers. For example, in the scenario illustrated in the ground truth of Figure 1.a where "*Spkr1 continues speaking even when Spkr2 starts his turn*", the GLR/BIC segmentation may fail in detecting speaker change because the theoretical boundary region presents some homogeneity. In subsections 3.1 and 3.2, two hypotheses are proposed to resolve this problem.

3.1. Bidirectional GLR/BIC segmentation

Due to the shifted variable size window introduced in the GLR/BIC method (please refer to the algorithms proposed in [6] or [8] for more details), processing from "*left to right*" may detect different points of change than processing from "*right to left*", and therefore, there is a chance that a missed boundary in the first direction can be detected in the other direction and vice versa. Figure 1 illustrates the three possible corrections: *S1* (respectively *S2*) corresponds to the set of boundaries provided by the "*left to right*" segmentation (respectively "*right to left*") and $S1 \cup S2$ is the resulting union. Those corrections can be divided into two types: perfect and partial corrections. Practically, we have noticed that partial corrections outnumber the perfect ones.

Because this segmentation method is based on an acoustic homogeneity criterion, it has the ability of segmenting the audio stream into different sources: different types of music, different speakers and silence. This advantage is used to help Speech /Non-Speech separation. (cf. section 5).

3.2. Penalty coefficient decreasing technique

In Equation 1, we notice that when the penalty λ decreases, ΔBIC increases. Therefore, it is possible that ΔBIC becomes positive and an additional point of change is detected in this case. However, the decreasing of λ in an unsupervised manner can be harmful to the system performance because it may introduce many false alarms. That is why we must be sure that the region under investigation is unstable i.e. it contains an interaction zone as in the example shown in Figure 1. In section 5, a framework is proposed to handle the detection of the unstable segments.

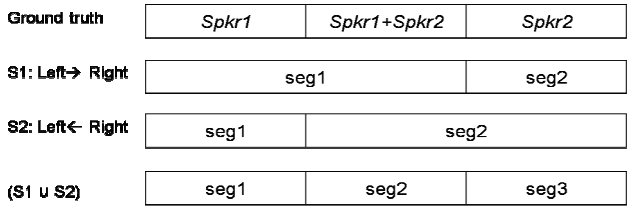

Fig. 1.a. Correction due to S1 and S2 (perfect correction)

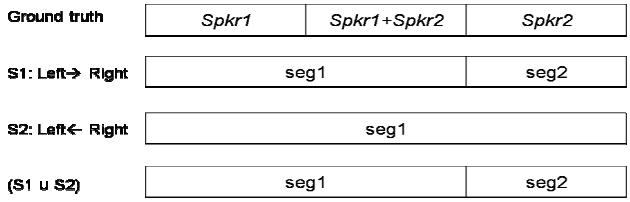

Fig. 1.b. Correction due to S1 (partial correction)

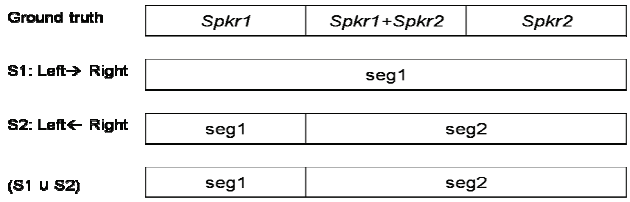

Fig. 1.c. Correction due to S2 (partial correction)

4. SPEAKER CLUSTERING

The speaker clustering consists in grouping all segments corresponding to the same speaker. In general, it is done with a hierarchical bottom-up manner. Many criteria like BIC [7] or EVSM [9] (Eigen Vector Space Model) were proposed in the literature to resolve this problem in an unsupervised manner. Moreover, speaker identification (SID) clustering that needs training process can be added as described in [10]. In this section, we choose to review the BIC clustering and we propose some modifications in order to fit characteristics of meetings.

The BIC clustering uses the same Equation 1 presented for speaker segmentation. But in this case, *X1* and *X2* denote the clusters under investigation and *X* the resulting cluster.

Due to the high interaction in meeting data, the average length of speaker turns is relatively low contrary to broadcast news, and the regions where many people talk simultaneously are more numerous. Table 2 shows the difference of those two kinds of data.

The above two factors decrease the segments purity, and so introduce a risk of cumulative errors in the clustering

process. It is obvious that homogenous segments with long duration are more confident and provide better clustering. To deal with this problem, two hypotheses were proposed in subsections 4.1 and 4.2.

	ESTER'05	EPAC'08
Average length of speaker turns	20.22s	8.33s
time ratio of multi-speakers turns	0.21%	5.26%

Table2. Main difference between broadcast news (ESTER) and meetings (EPAC) corpora

4.1. Local /global clustering

In the standard hierarchical clustering, the initial clusters correspond to segments, and as described above for meeting data, those segments have relatively small duration. Due to the iterative structure of the clustering, it is very probable that the comparison is done between clusters of very different sizes. In this case, the BIC-based inter-cluster similarity is not precise as explained in [11], and may introduce cumulative errors in the clustering process. Our solution to cure this weakness is to do a local clustering every N consecutives segments (practically *N=20*) before processing the global one. The goal is to build a first level of confident clusters with balanced sizes.

4.2. Similarity matrix and updating of clusters

At the end of a clustering process, each segment is theoretically assigned to the cluster providing the highest BIC similarity. However, due to the hierarchical bottom up manner, there are some segments that do not respect this hypothesis. To correct these errors and therefore enhance the clusters purity, we propose to compute the similarity matrix between segments S_i and clusters C_j and then reclassify segments regarding this matrix. For example, in figure 2, the similarity matrix shows that the segment S_8 will be assigned to the cluster C_3 *(-ΔBIC=0.7)* instead of C_1 *(-ΔBIC=0.1)* as in the previous clustering.

-ΔBIC	S_1	...	S_8	S_9	...	S_{14}	S_{15}	...	S_{20}
C_1	0.6	...	0.1	-0.3	...	-0.6	-0.2	...	-0.7
C_2	0.2	...	-0.4	0.8	...	0.5	0.3	...	-0.1
C_3	-0.6	...	0.7	-0.1	...	0.2	0.9	...	0.3

Fig.2. Similarity matrix between segments and clusters

5. ITERATIVE SPEAKER DIARIZATION

After reviewing the strengths and weaknesses of each essential component of a speaker diarization, we propose our iterative system (Figure 3). It can be summed up by the following algorithm:

1) **Parameters extraction** where the Mel Frequency Cepstrum Coefficients (MFCCs), the 4Hz modulation energy and the log-likelihoods of Speech and Non-Speech GMMs are computed.
2) First **Bidirectional GLR/BIC segmentation** using a penalty coefficient $\lambda = \lambda 1$ (practically *λ1=1).*
3) **Speech/Non-Speech Separation** by merging the 4Hz modulation Energy (M_E) and Speech and Non-Speech GMM scores for each segment.
4) **Local BIC clustering** each N consecutive segments.
5) **Global BIC clustering** based on clusters obtained from previous step.
6) Computation of the **Similarity matrix** between segments S_i *(i=1 to* N_s*)* and clusters C_j *(j=1 to* N_c*)* where N_s is the number of segments and N_c is the number of clusters.
7) **Updating clusters** by assigning each segment S_i to $\arg\max_{C_j}(-\Delta BIC(S_i, C_j))$ when *j* varies from *1* to N_c.
8) **Splitting unstable segments** using the bidirectional GLR/BIC segmentation with $\lambda = \lambda 2$, $\lambda 2 < \lambda 1$ *(practically λ2=0.8)* as explained in subsection 3.2. Unstable segments are segments for which $-\Delta BIC(S_i, C_j) < 0$ i.e. the similarity between segment S_i and its corresponding cluster is bad.
9) **Stop loop** if no more splitting can be done. Otherwise, do a **speech/Non-Speech separation** and go back to **step 6** and so on.
10) **Final SID clustering** in order to group clusters corresponding to the same speaker but under different backgrounds.

We notice that the number of segments N_s and the number of clusters N_c are dynamic: N_s can decrease at the end of step 3 and increase at the end of step 8. However N_c can only decrease at the end of step 7 due to the reassignment of segments.

6. EXPERIMENTS AND RESULTS

In order to evaluate the proposed speaker diarization system on meeting data, a test set is chosen from the EPAC corpus. This set contains 10 shows for a total duration of 7 hours recorded from 3 French radio stations.

In these experiments, a centisecond approach is used i.e. the soundtrack is decomposed into 10 ms frames. Speech and Non-Speech GMMs use vectors composed of 12 MFCC coefficients, the energy and their associated derivatives. The GLR/BIC segmentation uses 12-dimensional MFCC vectors and the clustering uses 15-dimensional MFCC vectors.

Figure 3 illustrates the difference between our proposed system and the standard system used in our experiments. This standard system gives similar results as the best one on ESTER evaluation test [4].

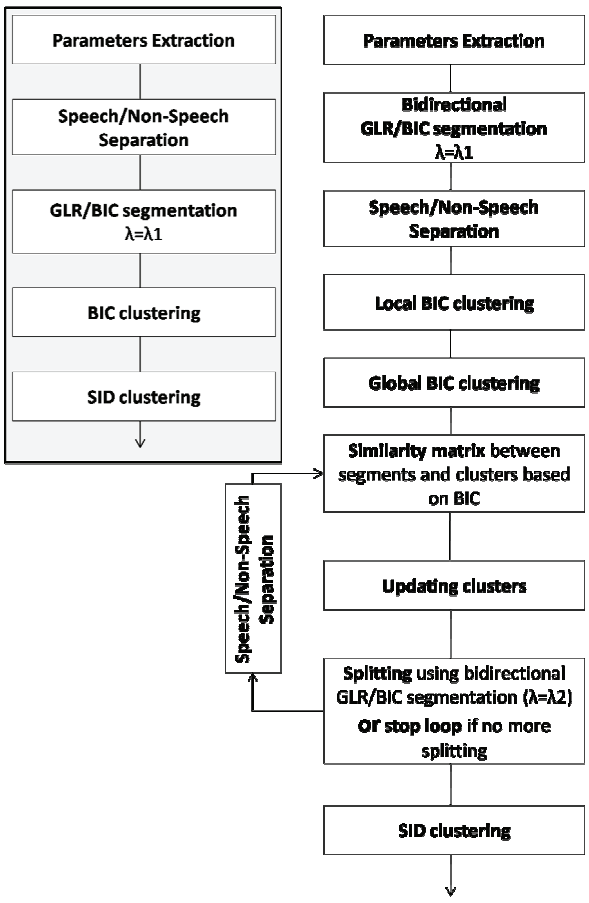

Fig.3. Standard **vs** Improved Speaker Diarization systems

In table 3, NIST diarization error rates (DER) show that results are relatively improved by 32.5% (16.72 instead of 24.77) or 40.4% (11.66 instead of 19.55) when multi-speakers regions are excluded (b). Moreover, because the bidirectional GLR/BIC segmentation provides precise boundaries, missed detection error rate decreases by 0.8 when performing Speech/Non-Speech separation in a later step contrary to the standard system. Finally, we notice that the highest impact is provided by the clustering corrections because speaker error rate significantly decreases.

	All speakers turns (a)		Exclusion of multi-speakers turns (b)	
	Standard system	**Improved system**	Standard system	**Improved System**
% Missed detection	9.7	**8.9**	3.9	**3.1**
% False alarm	0.6	**0.1**	0.7	**0.2**
% Speaker error	14.5	**7.6**	15.0	**8.4**
Diarization error rate	24.77	**16.72**	19.55	**11.66**

Table3. Diarization error rates for both standard and improved systems

7. CONCLUSIONS

In this paper, a new speaker diarization system for meeting data is presented. We propose the bidirectional GLR/BIC segmentation, the penalty coefficient decreasing technique, the local/global clustering and the BIC similarity matrix in order to build an iterative scheme that aims to enhance the purity of clusters. Our system outperforms the baseline by 40.4%. As future works, speed optimization and multi-speaker detection technique like in [12] can be done.

11. REFERENCES

[1] S.E. Tranter, D.A. Reynolds, "An overview of automatic speaker diarization systems", IEEE transactions on Audio, Speech and Language Processing, pp. 1557-1565, 2006.

[2] J.-H. Chang, N.S. Kim and S.K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models", IEEE Transaction on Signal Processing, 54, , pp. 1965-1976, 2006.

[3] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", ICASSP, pp. 1331–1334, 1997.

[4] S. Galliano, E. Geofrois, De. Mosterfa, J.F. Bonastre, and G. Gravier, "the Ester phase II evaluation campaign for the rich transcription of the French broadcast news," EUROSPEECH, pp. 1149–1152, 2005.

[5] J. Pinquier, J-L Rouas, R. André-Obrecht. "A Fusion Study in Speech/Music Classification", ICASSP, pp. 17-20, 2003.

[6] E. El-Khoury, C.Sénac and R.André-Obrecht, "Speaker Diarization: Towards a more Robust and Portable System", ICASSP, pp. 489-492, 2007.

[7] S.S. Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", DARPA Speech Rec. Workshop, 1998.

[8] M. Cettelo and M. Vescosi, "Efficient audio segmentation algorithms based on the BIC", ICASSP, pp. 537-540, 2003.

[9] W.H. Tsai, S.S. Cheng, Y.H. Chao and H.M. Wang, "Clustering speech utterances by speaker using Eigenvoice-Motivated vector space models", ICASSP, pp. 725-728, 2005.

[10] C. Barras, X. Zhu, S. Meignier and J.-L. Gauvain, "Multi-stage speaker diarization of broadcast news", IEEE Transactions on Audio, Speech and Language Processing, pp. 1505-1512, 2006.

[11] K.J. Han, S.S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system", Interspeech, pp. 1853-1856, 2007.

[12] K. Boakye, B. Trueba-Hornero, O. Vinyals and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meeting", ICASSP, pp. 4353-4356, 2008.