SPEAKER DIARIZATION: TOWARDS A MORE ROBUST AND PORTABLE SYSTEM

Elie EL KHOURY, Christine SÉNAC, Régine ANDRÉ-OBRECHT
SAMoVA team – IRIT – CNRS UMR 5505
Toulouse - France
{khoury, senac, obrecht}@irit.fr

ABSTRACT

In this paper, we describe a new method for speaker segmentation and clustering of an audio document. For the segmentation phase, we combine the Generalized Likelihood Ratio (GLR) and the Bayesian Information Criterion (BIC) in a way that avoids most of the parameters tuning. For the clustering phase, we use an existing approach that utilizes the Eigen Vector Space Model (EVSM) with a bottom-up hierarchical grouping but we make some improvements by introducing prosodic information. Evaluation is done on the audio database of the ESTER evaluation campaign for the rich transcription of French Broadcast news. Results show that our method which operates without any a priori knowledge about speakers is suitable for speaker diarization as it outperforms the traditional ones with an overall Diarization error rate (DER) of 16.72%.

Index Terms— Speaker Diarization, Generalized Likelihood Ratio, Bayesian Information Criterion, Eigen Vector Space Model, F0 Feature.

1. INTRODUCTION

In the context of audio document indexing and retrieval, speaker diarization is the process which detects speakers turns and regroups those uttered by the same speaker.
So, it's generally based on a first step of segmentation, often preceded by a speech detection phase, that consists in partitioning the regions of speech into segments where each segment must be as long as possible and must contain ideally the speech of only one speaker, followed by a clustering step that consists in giving the same label to segments uttered by the same speaker. Ideally, each cluster corresponds to only one speaker and vice versa.

Most of the systems operate without specific a priori knowledge of speakers or their number in the document. But as mentioned in [1], in spite of tremendous progress, they generally need specific tuning and parameters training.

On the contrary, we present a new approach we tried to make the more robust and portable. This paper is organized as follows: Section 2 presents the segmentation step processed without preliminary speech detection. Section 3 explains the hierarchical clustering we used. We present our experiments on broadcast news in section 4.

2. SEGMENTATION

The audio documents we process differ in the quality of recordings such as bandwidth, microphones, noise, the number of speakers and the structure of the speech: duration and sequencing of speaker turns. But they particularly contain multiple audio sources, possibly overlapped, such as music segments, jingles, commercials, noises and different speakers.

In this context, a preliminary speech detection step, that separates the regions of speech and the regions containing music, silence and noise, will necessarily miss some speech regions overlapped with music: these missed detections are then irretrievable.

It's why we chose to process the segmentation directly on the raw audio file without any preliminary speech detection.

2.1. Existing methods

Segmentation uses generally: 1) metric approaches as symmetric Kullback-Leibler [2], Hotteling's T2-Statistic [3], 2) or approaches based on model selection like the GLR [4] and the BIC criterion [5] which lead to the best systems [6], [7].

The metric approaches didn't give us sufficient results, so we turned towards the GLR and BIC criterion of which a concise presentation follows.

Two hypotheses are considered: H_0 supposes that there is one speaker in the window X, and H_i supposes that there are 2 consecutive speakers separated by a point of change *i*.
The GLR is given by:

$$GLR = \frac{P(H_i)}{P(H_0)} \tag{1}$$

Supposing that the probability density functions (PDFs) are Gaussians, the logGLR expression becomes:

$$R(i) = \frac{N_X}{2}\log|\Sigma_X| - \frac{N_{X1}}{2}\log|\Sigma_{X1}| - \frac{N_{X2}}{2}\log|\Sigma_{X2}| \tag{2}$$

where the window X is divided into two sub-windows X_1 and X_2, separated by the point i. $\sum_X$, $\sum_{X1}$ and $\sum_{X2}$ are the covariance matrices of the acoustic vectors respectively of X, X_1 and X_2. N_X, N_{X1} and N_{X2} are the corresponding number of acoustic vectors.

The expression of ΔBIC is given by:

$$\Delta BIC = R(i) - \lambda \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N_X \quad (3)$$

λ is the penalty coefficient and d the dimension of the feature vectors.

That criterion is applied within a shifted variable size window. In [6], the authors utilize many parameters that should be tuned experimentally to optimize the detection of points of change. Also, the detection of acoustic changes is sequential which is source of cumulative errors because a detected point depends from previous one. Finally, the DER is relatively high.

The proposed method presents some improvements by trying to cure this weakness.

2.2. The proposed method

Figure (1.a) illustrates a raw audio stream, containing both speech and non speech regions, where the theoretic segmentation is represented by the points R_1, R_2,…,R_n. The proposed method for the segmentation of that stream processed without preliminary speech detection follows four main steps.

2.2.1. Splitting step

It consists in splitting arbitrarily the audio stream into windows of two seconds. Then, we detect the point of change the most probable in every window. This step is shown in Figure (1.b). Mathematically, this point corresponds to the maximum of the GLR expression or to the maximum of ΔBIC [7].

The advantage of this step is that we don't need to fix a threshold for comparing the expression of GLR.

2.2.2. Most probable point detection step

In the first step, we have obtained points of change P_1...P_m which separate the best way the two mono-Gaussian models existing in every window. However, those models are not very representative because they are affected by a window with a fixed size and fixed boundaries. So, we repeat the first step using windows that are chosen as following: to detect a change point P_i', we use the window $[P_{i-1}, P_{i+1}]$. Thus, the new models will be quite close to Gaussian distributions.

If two consecutive windows vote for the same point, or for two close points (difference < 0.2 seconds), we decide to confuse those two points by considering their mean. Thus, the number of points will decrease.

This step is illustrated in Figure (1.c).

2.2.3. Re-Adjustment step

This step (Figure (1.d)) consists in repeating the second step several times until the repartition of change points is stabilized i.e. the distributions approach Gaussian distributions. Experimentally, this stabilization is mainly reached after three iterations. Points obtained are annotated q_1...q_t where t < m.

2.2.4. Definitive change detection step

At this stage, the points q_i represent the most probable positions of change. Thus the BIC criterion is applied to select only the points that are effectively points of acoustic changes (Figure (1.e)).

The final algorithm is shown below:

```
Let m = number of points q_j,
j = 1,
initialize W =[q_0, q_2]
while ( j ≤ m)
    in W, search  ΔBIC_max ,
    if ΔBIC_max ≥ 0 then
              q_j =argmax ΔBIC_max,
              increment j, S = q_j-1
    else
              increment j,
    End if
    E= q_j+1 , W=[S, E]
End while
```

3. CLUSTERING

The clustering consists in collecting all segments corresponding ideally to the same speaker. In our case, segments may contain pure speech or speech overlapped with music or music segments: we should notice that jingles are accurately separated from adjacent music regions.

In most systems, the clustering step is achieved by a hierarchical grouping algorithm in a bottom-up manner in which closest clusters – in the sense of a pair wise distance or similarity measure between each cluster – are merged iteratively. Methods differ in the selection of merging distance (mainly ΔBIC distance and Kullback-Leibler distance) and stopping criterion (mainly with a threshold).

The results we obtained with ΔBIC and Kullback-Leibler distances were unsatisfactory, so we use another method based on the Eigen Vector Space Model (EVSM) and we moreover exploited prosodic information.

3.1. Clustering with the Eigen Vector Space Model

Our clustering method is based on the work of Tsai and al. [8] which utilizes EVSM with a hierarchical bottom-up clustering.

Figure 2 presents the different steps: from all the segments S_i a universal Gaussian Mixture Model (GMM) Λ is created. This GMM is then adapted on each segment S_i to

obtain the GMM Λ_i. From each Λ_i, a super-vector V_i is created by concatenating the mean vectors of each gaussian distribution of that Λ_i.

Then, PCA (Principal Component Analysis) is applied to obtain from each vector V_i, a vector W_i with a lower dimension. After that, cosine formula calculates similarity between each two vectors (W_i, W_j).

The stopping criterion is based on a threshold comparison: if the cosine is higher than this threshold *th1*, the two segments are grouped.

3.2. A stronger merging criterion

Our contribution consists in choosing a stronger merging criterion based both on the previous similarity measure and on prosodic information which is not yet exploited by diarization systems contrary to some Speaker Recognition systems such as [9,10].

The F0 feature is estimated every 10ms on voiced regions with ESPS signal processing software which utilizes the normalized cross correlation function and dynamic. Then, a difference (called ΔF_0) between the averages of the F_0 values of each couple of segments is calculated.

We have to notice that, whatever the software, some pure music segments will be considered erroneously as voiced regions of the signal; but they will never be grouped with speakers segments because of the cosine similarity which separates them.

As illustrated in Figure 3, the new merging criterion becomes: the two segments correspond to the same speaker if 1) the similarity (cosine formula) is higher than a threshold *th1* and 2) ΔF_0 is lower than a threshold *th2*.

4. EXPERIMENTS AND RESULTS

Experiments were done on 24 hours of Broadcast News taken from ESTER campaign's data [11]: we took 4 hours from phase1 for tuning the parameters and we took 20 hours for testing : 10 hours from phase1 (different files from tuning files) and the 10 hours of test files of phase2.

To evaluate our method, we used the NIST scoring software (http://nist.gov/speech/tests/rt/rt2005/spring/). We chose 12 MFCC calculated each 10 ms for the segmentation step, and 15 MFCC + Energy for the clustering step. In addition, we considered that each segment is modeled by a GMM with 128 Gaussian distributions. Also, we fixed λ to 1, the thresholds *th1* to 0.7 and *th2* to 40Hz.

Table1 shows that the proposed segmentation "GLR+BIC" gives an absolute improvement of 12.74% compared to the "BIC + shifted variable size window" segmentation using the same "EVSM + hierarchical clustering" step. Also, the proposed clustering with prosodic information (EVSM + F0 + Hierarchical clustering) method gives additional improvement of 4.81%.

We verified that short speakers turns (>0.6second), different music styles and phone speech are well detected and the corresponding segments are well grouped. Main errors may occur for simultaneous speakers, speakers with music background, and consecutive speakers with the same gender and who are very difficult to distinguish by human ear.

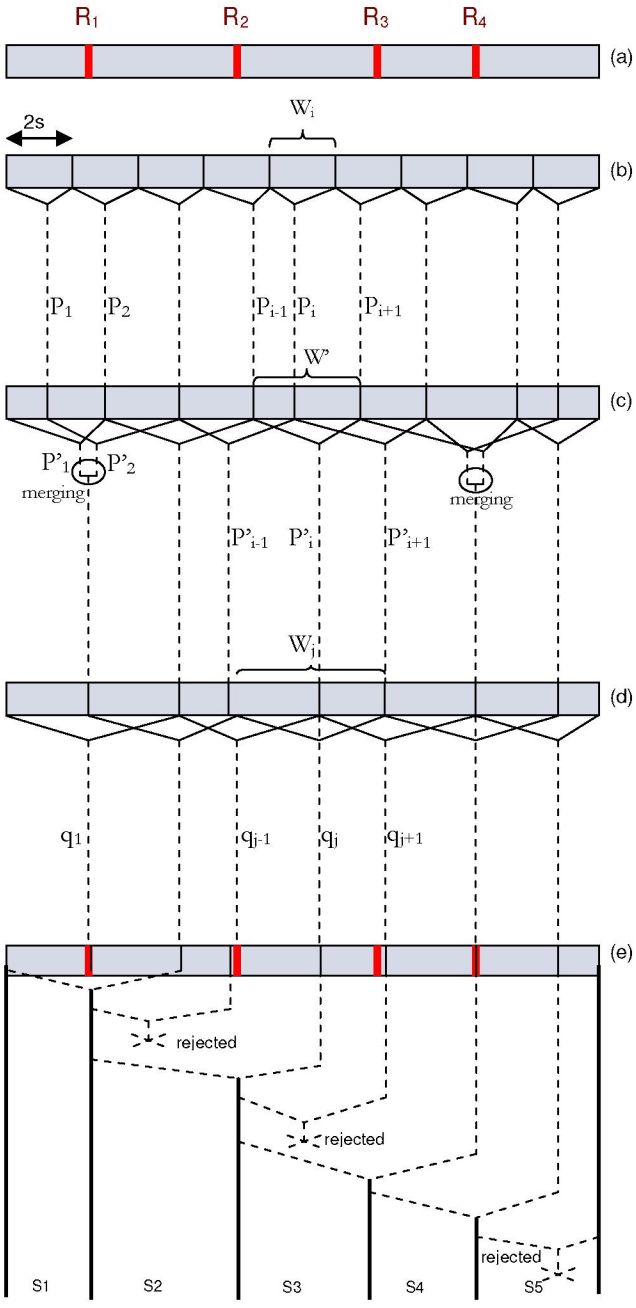

Fig.1. Segmentation steps

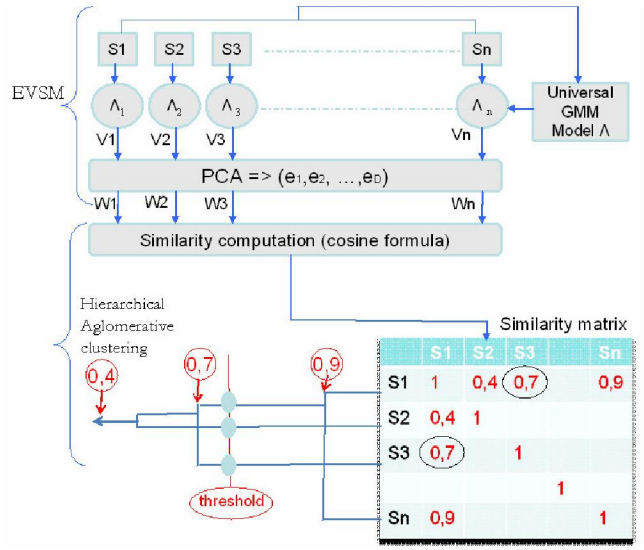

Fig.2. Clustering step: EVSM + Hierarchical agglomerative clustering

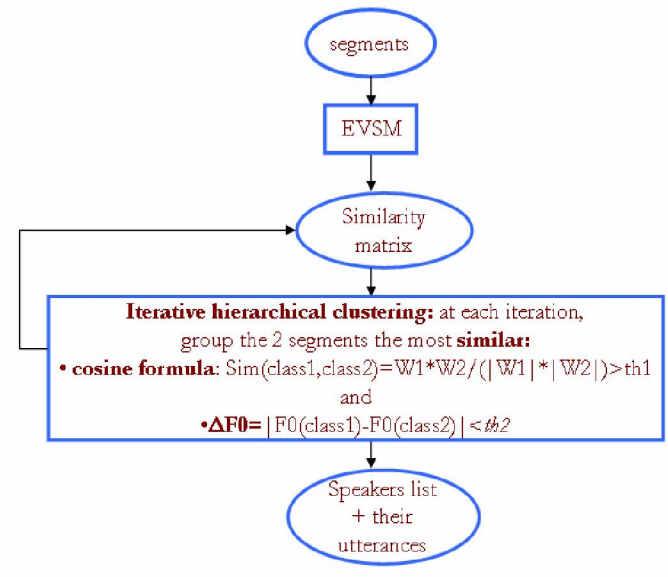

Fig.3. The proposed clustering

Segmentation step	ΔBIC + shifted variable size window	GLR+ΔBIC	GLR+ΔBIC
Clustering step	EVSM + Hierarchical clustering	EVSM + Hierarchical clustering	EVSM +ΔF_0 +Hierarchical clustering
Missed detection	3.42%	0.48%	0.14%
False alarm	0.98%	2.46%	2.4%
Speaker error rate	29.87%	18.59%	14.18%
Overall DER	34.27%	21.53%	16.72%

Table1. Comparison between the proposed methods

5. CONCLUSIONS

We have proposed, implemented and evaluated new approaches for speaker diarization of broadcast news. This system segments the audio stream into utterances using a combined GLR-BIC approach, and then, groups all the utterances corresponding to the same speaker using the EVSM and prosodic information (ΔF_0) with a hierarchical bottom-up clustering. Results show that our method made improvements: no pre-trained models, lower number of parameters to be tuned, more precision (small segments are well detected) and lower overall DER (16.72%).

6. REFERENCES

[1] S.E. Tranter, D.A. Reynolds, "An Overview of Automatic Speaker Diarization Systems", IEEE Transactions on Audio, Speech and Language Processing, vol.14, p. 1557-1565, Sept 2006

[2] M.A. Seigler, U. Jain, B. Raj and R.M Stern, "Automatic Segmentation, Classification and Clustering of Broadcast news audio", DARPA Speech Recognition Workshop, 1997

[3] B. Zhou, J.H.L Hansen, "Efficient Audio Stream Segmentation via the Combined T²-Statistic and Bayesian Information Criterion", IEEE Transactions on Speech and Audio Proc., Vol. 13, p. 467-474, 2005

[4] M. Siu, H. Gish, and R. Rohlicek, "Segregation of speaker for speech recognition and speaker identification", p. 873-876, ICASSP 1991

[5] 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

[6] P. Sivakumaran, J. Fortuna and A.M. Ariyaeeinia, "On the use of the Bayesian Information Criterion in multiple speaker detection", p. 795-798, Eurospeech 2001

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

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

[9] D.A.Reynolds, P.Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: applications to broadcast audio and telephone conversations", NIST Rich transcription workshop, November 2004

[10]J.P.Campbell, D.A.Reynolds, R.B.Dunn, "Fusing high and low-level features for speaker recognition", p. 2665-2668, EUROSPEECH 2003

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