Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs

Mathieu Ben, Michaël Betser, Frédéric Bimbot, Guillaume Gravier

IRISA/METISS, Campus Universitaire de Beaulieu
35142 Rennes Cedex, France
{mben,mbetser,bimbot,ggravier}@irisa.fr

Abstract

In this paper, we present an approach for speaker diarization based on segmentation followed by bottom-up clustering, where clusters are modeled using adapted Gaussian mixture models. We propose a novel inter-cluster distance in the model parameter space which is easily computable and which can both be used as the dissimilarity measure in the clustering scheme and as a stop criterion. Using adapted Gaussian mixture models enables a good description of the feature vector distribution within a cluster while adaptation prevents over-training for clusters with few data. Experiments carried out on broadcast news data in French demonstrate the potential of the proposed approach which exhibits performance similar to BIC clustering. However, our clustering method appeared to be more sensitive to segmentation errors than the BIC approach.

1. Introduction

In the context of spoken document indexing, speaker diarization is the process of determining speaker turns and of grouping together turns uttered by the same speaker. It results in a structure of the audio document according to speakers, and therefore provide information useful to the structuring and indexing of the document. Diarization is also an important step in the speech transcription process as it provides valuable information for unsupervised acoustic model adaptation.

As opposed to speaker tracking, another speaker indexing related task, for which the speakers to track are known beforehand and training data are provided for those speakers, speaker diarization assumes no prior knowledge of the speakers and therefore no training data. The actual number of speakers is not known either and has to be determined automatically.

Diarization is typically carried out as a three step process. The first step consists in segmenting the document into speech segments which hopefully contain speech from a single speaker – with the exception of segments containing overlapping speech. The second and third steps consist in determining the actual number of speakers and in grouping together segments from the same speaker. In most systems, the grouping step is achieved by a hierarchical clustering algorithm in a bottom-up or top-down manner. In the hierarchical clustering approach, determining the number of speakers is carried out using a stop criterion in the clustering algorithm in order to determine the optimal number of clusters. The Bayesian information criterion (BIC) is probably the most popular one [1].

In the clustering step, most approaches rely on a bottom-up clustering algorithm in which clusters are built iteratively by finding out the two closest clusters and by merging them if this results in a better clustering according to some criterion. Finding out the two closest clusters requires to define a distance between two clusters. The various inter-cluster distance measures proposed in the literature can be separated into two main families depending on whether the underlying model corresponds to a single Gaussian or to a mixture of Gaussians. Classical single Gaussian based distance measures are the symmetric Kullback-Leibler distance and the ΔBIC distance. In Gaussian mixture model (GMM) based approaches, the cross likelihood ratio distance is often considered [2].

This work focuses on a bottom-up clustering scheme with adapted GMMs [2] for a speaker diarization task. Based on the principles of Maximum A Posteriori (MAP) adaptation of GMMs, we define a novel inter-cluster distance which is directly computable from the model parameters. We use this distance as the dissimilarity measure for the clustering process and as a stop criterion for the algorithm. We compare the performance of the proposed approach to that of a classical BIC-based approach.

The rest of the paper is organized as follows. In section 2, the segmentation and clustering principle using the Bayesian information criterion is recalled. The proposed GMM-based approach is then detailed in section 3 and experimental evaluation is carried out in section 4.

2. Segmentation and clustering with BIC

Segmentation and clustering via the Bayesian information criterion was initially proposed in [1] and we recall here the basic principle of this approach.

In a very general way, given a segment y and a model Λ for the segment, the Bayesian information criterion is defined as

$$\mathrm{BIC}(\Lambda) = \ln \mathcal{L}_\Lambda(y) - \lambda \frac{\#\Lambda}{2} \log N_y \ , \qquad (1)$$

where $\mathcal{L}_\Lambda(y)$ denotes the segment likelihood given the model, $\#\Lambda$ the number of free parameters in the model and N_y the observation length. This criterion can be used both for segmentation and for clustering.

2.1. Segmentation

Detecting an abrupt change using the BIC is carried out by looking at the difference between the criterion under the assumption that there is a change in the sequence and the criterion under the dual assumption. Assuming an underlying Gaussian model, a change is detected between the two segments $a = \{y_1 \ldots y_{t-1}\}$ and $b = \{y_t \ldots y_T\}$ if

$$\Delta\mathrm{BIC}(t) = R(t) - \frac{\lambda_s}{2}\left(d + \frac{d(d+1)}{2}\right)\ln T \qquad (2)$$

is negative, where d is the feature vector dimension and $R(t)$ is the log-likelihood ratio given by

$$R(t) = T \ln \sqrt{|\Sigma_{ab}|} - T_a \ln \sqrt{|\Sigma_a|} - T_b \ln \sqrt{|\Sigma_b|} \ . \quad (3)$$

In the above equation, T_a denotes the length of segment a and Σ_a the maximum likelihood estimate of the covariance matrix for segment a.

2.2. Clustering

Initially, each segment is a cluster by itself and the clusters are modeled by a single Gaussian. At each step of the clustering algorithm, a similarity measure is calculated for each pair of clusters. The two closest clusters are merged if the corresponding BIC variation, given by Eq. (2), is negative. If the difference is positive, the algorithm is stopped.

Various distances between clusters based on single Gaussian modeling have been proposed, such as the arithmetic-harmonic sphericity measure or the symmetric Kullback-Leibler distance. However, in the BIC framework, the likelihood ratio measure given by Eq. (3) is well suited and has given good results. Note that the smallest inter-cluster distance can also be used as a criterion to stop the algorithm, in addition to the BIC decrease.

3. Clustering using MAP adapted GMMs

At the end of the clustering, when clusters contain a large amount of speech, the single Gaussian model might be too simple to approximate the complex distribution of acoustic features. To be able to take account for more complex distributions, Gaussian mixture cluster models can be used. However, such models have more parameters and maximum likelihood estimation can be unreliable at the beginning of the clustering, when clusters have few data. To prevent this parameter estimation problem, we adapt the cluster GMM parameters from a cluster-independent background model. Based on the very principle of the adaptation, a similarity measure between clusters is defined in the model space.

We first discuss GMM adaptation before detailing the proposed similarity measure in the model space and the corresponding stop criterion.

3.1. Cluster GMM adaptation

The probability density function (pdf) of a K-component GMM for a d-dimensional random variable y is defined as

$$p(y|\Lambda) = \sum_{k=1}^{K} w_k \mathcal{G}_k(y|m_k, S_k) \quad (4)$$

where $\mathcal{G}_k(\cdot)$ is the Gaussian density function and $\Lambda = \{w_k, m_k, S_k\}$ is the set of parameters, *i.e.* respectively the weights (with the constraint $\sum_{k=1}^{K} w_k = 1$) the d-dimensional mean vectors and the $d \times d$ covariance matrices.

In the MAP approach, the GMM for a cluster is adapted from a cluster-independent background model whose parameters are used as prior values in the adaptation process. This background model parameters can be estimated on a large amount of data as in speaker recognition. Alternately, the background model parameters can be estimated on the whole speech portions of the document on which clustering is performed. Such a model will be referred to as a Document Speech Background Model (DSBM).

In this work, cluster models are obtained by adapting the mean vectors of a DSBM with diagonal covariance matrices. Component weights and covariance matrices are not adapted and are therefore left unchanged. Adaptation of cluster model is performed via the EM algorithm [3] using all the feature vectors for the cluster considered. Re-estimation is done according to

$$\widehat{m}_k = \frac{\gamma_k}{\gamma_k + \tau_k} \overline{y_k} + \frac{\tau_k}{\gamma_k + \tau_k} \mu_k \quad (5)$$

where γ_k is the total occupation rate of Gaussian $\mathcal{G}_k$, $\overline{y}_k$ is the posterior mean calculated under $\mathcal{G}_k$ and μ_k is the prior mean of m_k (see [3] for details).

In the above equation, τ_k is a relevance factor which controls the amount of adaptation for mean vector m_k. As can be seen from equation (5), the new estimate $\widehat{m}_k$ is a weighted sum of the posterior and prior means, the balance between the two being controlled by the relevance factor τ_k and by the occupation rate γ_k. This means that if $\mathcal{G}_k$ receives few data, the estimate $\widehat{m}_k$ is closed to its prior value μ_k, while if it receives a large amount of data, $\widehat{m}_k$ is dominated by the posterior mean $\overline{y}_k$ (*i.e.* the maximum likelihood estimate). On the one hand, this mechanism prevents over-adaptation when clusters contain only a small amount of data. On the other hand, when clusters become larger, more Gaussians can be adapted leading to more representative models.

3.2. Similarity measure between mixture models

The clustering process relies on the definition of a similarity measure between cluster models obtained by adaptation. A well known dissimilarity between two general pdf's p and $\tilde{p}$ is the Kullback-Leibler symmetric divergence $KL2(p, \tilde{p})$, defined as the sum of the two dual oriented divergence $KL(p||\tilde{p})$ and $KL(\tilde{p}||p)$,

$$KL2(p, \tilde{p}) = KL(p|\tilde{p}) + KL(\tilde{p}||p) = E_p[\log \frac{p}{\tilde{p}}] + E_{\tilde{p}}[\log \frac{\tilde{p}}{p}] \ .$$

In this equation $E_p[\cdot]$ and $E_{\tilde{p}}[\cdot]$ denote the expectations calculated under the pdf's p and $\tilde{p}$ respectively. For GMMs, there is no analytic form for the divergence which can be approximated using a Monte-Carlo method as in [4]. However, the Monte-Carlo approximation is computationally expensive.

Using the properties of MAP adaptation, we explored a new similarity measure derived from the KL symmetric divergence. For two K-component GMMs p and $\tilde{p}$ defined as in (4), it can be shown [5] that the KL divergence from p to $\tilde{p}$ is upper bounded by

$$KL(p||\tilde{p}) \leq KL(\mathbf{w}||\tilde{\mathbf{w}}) + \sum_{k=1}^{K} w_k . KL(\mathcal{G}_k||\tilde{\mathcal{G}}_k) \quad (6)$$

where $KL(\mathbf{w}||\tilde{\mathbf{w}})$ is the KL divergence between the probability mass functions $\mathbf{w} = (w_1, ..., w_K)$ and $\tilde{\mathbf{w}} = (\tilde{w}_1, ..., \tilde{w}_K)$ and $KL(\mathcal{G}_k||\tilde{\mathcal{G}}_k)$ is the KL divergence between Gaussian k in p and Gaussian k in $\tilde{p}$. In the context of this paper, the set of weights and covariance matrices are common to all GMMs as only mean vectors were adapted. This implies that the first term in the right hand side of equation (6) is null. Furthermore, considering diagonal covariance matrices, this leads to the following upper bound for the KL symmetric divergence:

$$KL2(p, \tilde{p}) \leq \sum_{k=1}^{K} \sum_{d=1}^{D} w_k . \frac{(m_{k,d} - \tilde{m}_{k,d})^2}{\sigma_{k,d}^2} \quad (7)$$

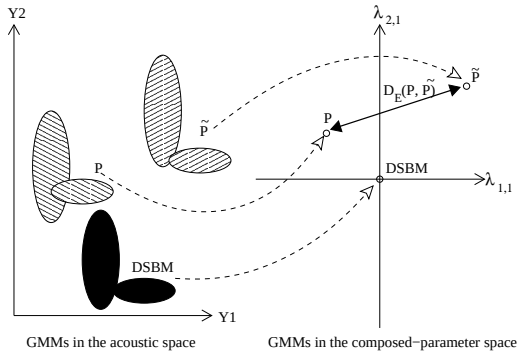

Figure 1: Representations of GMMs: from the acoustic space to the composed-parameter space

where $m_{k,d}$ and $\tilde{m}_{k,d}$ are the d components of the mean vectors of $\mathcal{G}_k$ and $\tilde{\mathcal{G}}_k$ respectively, and $\sigma^2_{k,d}$ is the $d \times d$ elements of the common covariance matrix S_k. The term on the right hand side of this inequality is homogeneous to a squared Euclidean distance between two points in the parameter space, $\{\lambda_{k,d}\} = \{\sqrt{w_k}\frac{\Delta m_{k,d}}{\sigma_{k,d}}\}$, where $\Delta m_{k,d} = m_{k,d} - \mu_{k,d}$ is the relative bias of the mean with respect to the DSBM one. In this space, a cluster GMM is represented with a single point and the DSBM corresponds to the origin (see figure 1). We will denote this Euclidean distance $D_E(p, \tilde{p})$:

$$D_E(p, \tilde{p}) = \sqrt{\sum_{k=1}^{K}\sum_{d=1}^{D} w_k . \frac{(m_{k,d} - \tilde{m}_{k,d})^2}{\sigma^2_{k,d}}} \qquad (8)$$

The advantage of this metric is that it can be computed directly from the parameters of the GMMs, with a low computational cost. In this paper, the squared distance ${D_E}^2(p, \tilde{p})$ is used as the similarity measure between two cluster GMMs. Moreover, we observed experimentally that this distance measure is highly correlated with the Kullback-Leibler divergence estimated with a Monte Carlo method, which indicates that both distance should lead to very similar clustering results.

3.3. Stopping criterion

During the clustering process, at each iteration the two clusters for which the squared Euclidean distance ${D_E}^2(p, \tilde{p})$ is minimum are selected and merged if the distance is below a threshold. If the minimal inter-cluster distance is above the threshold, then the agglomerative clustering is stopped. The threshold is optimized on development data and used on the test set to determine the expected number of clusters in each document.

4. Experiments

The proposed clustering algorithm was evaluated on the speaker diarization task of the French ESTER Broadcast News evaluation campaign [6]. We recall here briefly the data before giving results.

4.1. Evaluation corpus

The corpus used for this experiment is a corpus of radio broadcast news in the French language. The corpus is divided into a development set and a test set, according to the ESTER phase 1 specifications (see [6] for details). Each set contains a total of six broadcasts corresponding to 2h40 of France-Inter (Inter, 4 broadcasts) and 2h of Radio France International (RFI, 2 broadcasts). The diarization task is performed independently for each file and grouping of similar speakers across broadcasts is not considered.

Diarization performance is evaluated primarily in terms of classification error rate according to the performance measure defined by NIST for the Rich Transcription 2003 evaluation [7]. The classification error rate is calculated on the time basis, after determining the best mapping between speaker names and arbitrary cluster names. To help diagnose the strengths and weaknesses of the proposed method, the clustering is also evaluated in terms of average speaker and cluster purities as defined in [8]. The former is related to the number of clusters to which a speaker is associated while the latter relates to the number of speakers in the cluster. Clearly, as clusters build, speaker purity is bound to increase while cluster purity is bound to decrease.

In the experiments detailed below, performance are given for each individual broadcast as well as averaged over all the broadcasts.

4.2. Results

In all the experiments, Mel-frequency cepstral coefficients were used, along with the first derivatives for GMM-based clustering. Low energy frames are also removed for GMM-based clustering.

Results on the development set are summarized in table 1. The top table shows error rates obtained with a manual reference segmentation while error rates in the middle table were obtained with the automatic segmentation. These figures correspond to the optimal global threshold determined a posteriori and reported in column 8. For both the manual and automatic segmentations, GMM-based clustering yield significant improvements over BIC clustering with few Gaussians and best results are obtained with 16 component models. However, detailed results on each broadcasts show that this improvement mainly comes from two broadcasts (Inter 2 and RFI 2) while other broadcasts exhibit comparable performance, sometimes to the advantage of the BIC clustering. Moreover, the performance gap between manual and automatic segmentations is greater for the proposed approach than for the BIC. This is probably due to the fact that initial segments are quite short which advantages the BIC algorithm. However, it suggests that our method is more sensitive to segmentation than the BIC one.

Results on the test set show a slight advantage to BIC clustering. This result can be explained by two factors. First, as mentioned previously, the GMM approach seems to be more sensitive to automatic segmentation. As the segmentation process was optimized on the development set, more segmentation errors on the test set penalize more the GMM approach than the BIC one. This is confirmed by posterior experiments made on the test corpus reference for which the GMM approach gives better results (with 9% error rate) than the BIC approach (11% error rate). Second, a study of the posterior optimal threshold on each broadcast of the test set show that the optimal value varies greatly from one broadcast to another in the GMM approach. This was also the case on the development set but to a lesser extent. Experiments where also made using BIC as stop criterion in the GMM approach instead of the proposed distance, but it yielded approximately 2% more errors on both development and test corpora.

Figure 2 shows the evolution of cluster purity and speaker purity across clustering iterations for one file of the develop-

dev / ref	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	θ	All
MAP-8GMM	8.44	3.87	9.07	**6.97**	5.88	7.76	1.1	6.19
MAP-16GMM	**5.04**	**2.45**	12.59	9.47	5.72	**7.37**	1.4	**6.07**
MAP-32GMM	8.26	3.64	8.68	9.47	5.24	10.76	1.5	6.94
MAP-64GMM	8.65	8.73	7.07	11.04	12.25	18.45	1.5	11.84
BIC	6.90	13.94	**6.33**	7.12	**3.85**	19.25	3300	10.27
dev / bic	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	θ	All
MAP-8GMM	10.67	10.71	9.7	10.26	12.53	13.70	1.3	11.64
MAP-16GMM	9.45	**9.75**	11.46	12.99	10.18	12.44	1.7	**10.50**
MAP-32GMM	10.49	11.41	13.85	18.19	**8.07**	**10.99**	1.6	11.12
MAP-64GMM	14.07	10.82	**8.32**	11.65	14.83	20.79	1.7	13.58
BIC	**8.77**	16.29	8.76	**8.76**	8.10	26.16	2700	13.69
test / bic	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	θ	All
MAP-16GMM	14.54	**13.03**	17.76	10.65	16.80	**24.36**	1.7	16.4
BIC	**9.77**	16.29	**8.35**	**8.76**	**9.56**	26.16	2700	**15.10**

Table 1: Classification error rates for each broadcasts (columns 2 through 7), stop threshold and global classification rate across all broadcasts on the dev. set for the manual (above) and automatic (middle) segmentation and on the test set (below).

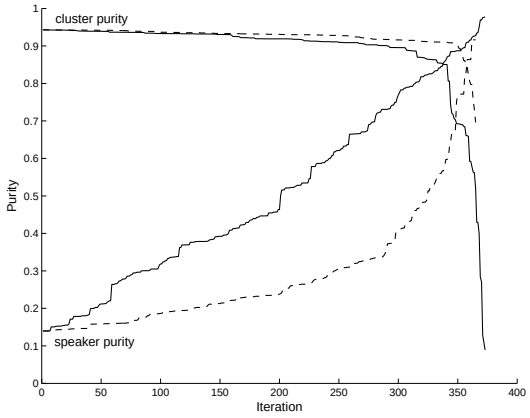

Figure 2: Speaker and cluster purity across iterations for GMM (solid) and BIC (dashed) clustering algorithms.

ment set, both for the GMM and BIC methods. Evolution of cluster purity is quite similar for the two approaches while speaker purity has a different behavior. Speaker purity regularly increases for the GMM/MAP approach while the BIC approach leads to a slow increase at the beginning of the clustering and a sudden fast increase when reaching the optimal clustering point. This results in a larger threshold range in the GMM approach for which both speaker and cluster purity are high.

5. Conclusions

We presented an approach based on adapted Gaussian mixture models for speech segment clustering using a novel inter-cluster distance in a diarization task. The proposed distance is directly derived from the model parameters and can easily be determined with a very low computational cost.

The method was evaluated on a broadcast news corpus in French. Comparison with a baseline method based on the Bayesian information criterion showed the potential advantage of the proposed approach which yielded promising results on the development set. However, this approach turned out to be sensitive to segmentation errors which resulted in a slight advantage of the BIC method on the test set.

In the future, we will focus our research efforts on improving the segmentation step in order to make the proposed approach more robust and competitive. Work should also be focused on the stop criterion which has been observed to be unstable from one file to another.

6. References

[1] S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Broadcast News Transcription and Understanding Workshop*, 1998.

[2] D. Reynolds, E. Singer, B. Carlson, J. McLaughlin, G. 0'Leary, M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics" in *Proceedings of ICSLP 98*

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, 2000.

[4] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances," in *Intl. Conf. on Acoustic Speech and Signal Proc.*, 2002.

[5] M. N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," in *IEEE Signal Processing Letters*, April 2003.

[6] G. Gravier, J. F. Bonastre, S. Galliano, E. Geoffrois, K. M. Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference*, 2004.

[7] NIST, "Rich transcription spring 03 evaluation plan," http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf.

[8] J. Ajmera, H. Bourlard, I. Lapidot and I. McCowan, "Unknown-multiple speaker clustering using HMM" in *Intl. Conf. on Spoken Language Proc.*, 2002.