FAST INCREMENTAL CLUSTERING OF GAUSSIAN MIXTURE SPEAKER MODELS FOR SCALING UP RETRIEVAL IN ON-LINE BROADCAST

J. E. Rougui, M. Rziza, D. Aboutajdine [*]

GSCM, Faculté des Sciences Rabat
4, Av Ibn Battouta B.P. 1014 RP
-Rabat- Morocco
rougui@lina.univ-nantes.fr
{rziza, aboutaj}@fsr.ac.ma

M. Gelgon, J. Martinez [†]

Polytechnic school of Nantes university
LINA FRE CNRS 2729
BP 50609 - 44306 Nantes Cedex 03 - France
{*lastname*}*@polytech.univ-nantes.fr*

ABSTRACT

In this paper, we introduce a hierarchical classification approach in the incremental framework of speaker indexing. The technique of incremental generation of speaker-homogeneous segments is applied in the first phase. Then, we propose a hierarchical classification approach that applied in the speaker indexing framework. This approach benefits from the efficiency of Gaussian mixture model (GMM) merge algorithm to the high accuracy of update Gaussian mixture models which referenced by speakers tree index. The adaptive threshold algorithm reduces the cost of exploring the speakers GMM into the balanced binary tree of speaker's index, whose complexity becomes logarithmic curve.

1. CONTEXT AND GOAL

The present paper is a contribution to the field of audio document indexing and retrieval. The growing amount of multimedia documents in digital form calls for techniques that, in content organization and retrieval, offer performance both in computational cost and relevance of the answers provided to the user. In many cases, a trade-off is sought between these two qualities.

Our work considers the case of spoken radio archives, in which a continuous flow of speech is introduced into the indexing system. News and discussion programs are quite typical of our scope, since they involve several speakers. The goal of our proposal is to provide technology that scales up to long spoken documents involving very many speakers. User applications built on top of this would enable fast speaker identity-based querying or browsing. As a side remark, it is quite easy to discard automatically occasional short jingles, thanks to their acoustic properties, leaving quite clean speech sections.

The present work is set in the framework of probabilistic modeling and statistical decision criteria, which is common and effective for speaker recognition. A classical solution to the above mentioned task would consist in partitioning the flow in speaker-homogeneous segments, then grouping segments originating from a single speaker. The latter operation enables valuable forms of browsing for the user ('go directly to all occurrences of such or such speaker in all radio archives') and also benefits to the accuracy of the system, in the sense that gathering more acoustic data from a single speaker enables refining his acoustic representation. We choose to represent speaker acoustic characteristics by a probability density over a mel-cepstral multi-dimensional space. In a usual implementation of this task, the cost of comparing the acoustic representation from a speaker segment (notably, the one that has just flown into the system) to all registered speaker models has a computational cost that rises linearly with the number of speakers. The focus of this paper is put forward a new technique to organize the set of speakers to obtain sub-linear cost, by avoiding exhaustive evaluation over the set of registered speakers. In this process, the trade-off is of course to gain significant computational cost while loosing only little reliability. The task relates tightly to the very classical issue of indexing structures for multi-dimensional data. The database community has put forward a considerable amount of contributions based on a variety of tree structures. The particularity of the current problem arises from the nature of the entities to index, namely probability distributions, for which classical indexing structures are inappropriate.

2. HIERARCHICAL CLUSTERING OF *GMMS* SPEAKERS

2.1. Principles and existing work

As the incoming audio stream produces new speakers, speaker-homogeneous segments have to be match to the corresponding speaker model (if the speaker is already enrolled). The cost of an exhaustive comparison the acoustic representation from a speaker segment (notably, the one that has just own into the system) to all registered speaker models has a computational cost that rises at least linearly with the number of speakers (it can be more expensive if the matching resorts to the history of feature vectors). The focus of this section is to disclose a new technique to organize the set of speakers to

[*] This work has been conducted under the Franco-Moroccan program on NTIC directed by INRIA.

[†] LINA-GRIM, Management and summarization of multimedia data.

obtain sub-linear cost, by avoiding exhaustive evaluation over the set of registered speakers see Fig. 1. In this process, the trade-off is to gain significant computational cost while loosing only little reliability. In comparison to the closest existing proposals based on anchor models [1] [2], the present proposal saves computation in building the hierarchical. It can be seen as complementary to multi-grained modeling, such as presented in [3].

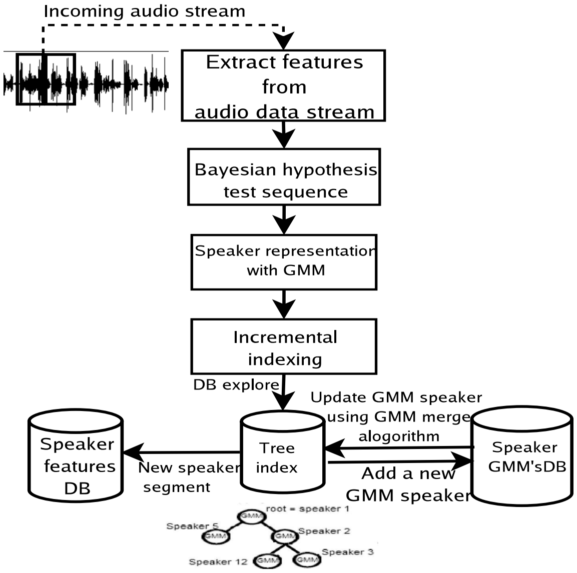


Fig. 1. Synoptic of speaker indexing system

2.2. Binary tree of incremental speaker indexing

For a large number of models detected, the binary tree representation of incremental speaker indexing remains one of a great solution, which is allows scaling up, updating and retrieval in audio document with a low cost. Then, we propose to create a tree of GMMs speaker indexing in the aim to organize the *GMMs* speaker under a binary tree form. Moreover, to increase the accuracy of speaker model retrieval, the *GMMs* merge algorithm is applied on the *GMMs* speaker already exist in *GMM DB*. The process of creating a binary tree of GMM speaker indexing system is shown in Fig. 2.

Let denote M_{GMM} a set of GMMs speaker already identified, and denote G the vector of a kl-divergence values between each model of M_{GMM} Then, the thershold η is defined:

$$\eta = \max(hist(G))$$

such as :

$$G = d(M_{GMM_i}, M_{GMM_j}), \forall i \neq j$$

The variation of threshold implies a change of framework, the level of the balanced tree, which minimizes the value depth of the tree, knowing that the threshold can selected with histogram of divergence.

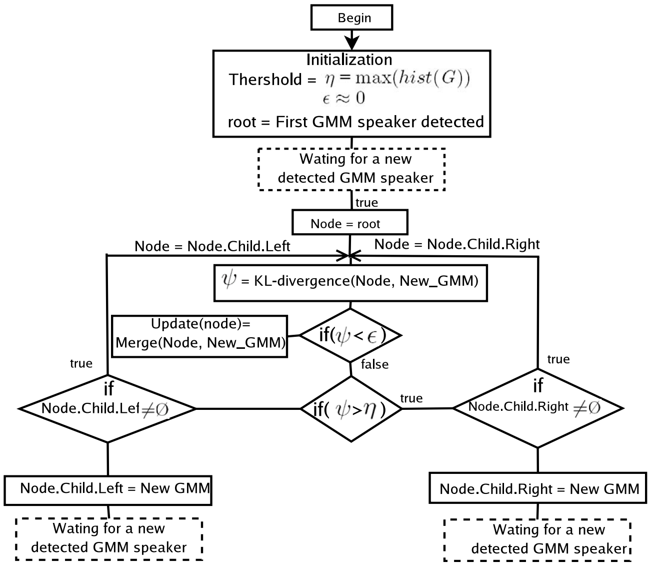

Begin
Initialization
Thershold = $\eta = \max(hist(G))$
$\epsilon \approx 0$
root = First GMM speaker detected
Wating for a new detected GMM speaker
true
Node = root
Node = Node.Child.Left
Node = Node.Child.Right
ψ = KL-divergence(Node, New_GMM)
Update(node)= Merge(Node, New_GMM)
if($\psi < \epsilon$)
false
true
if Node.Child.Lef $\neq \emptyset$
if($\psi > \eta$)
true
if Node.Child.Right $\neq \emptyset$
true
Node.Child.Left = New GMM
Node.Child.Right = New GMM
Wating for a new detected GMM speaker
Wating for a new detected GMM speaker

Fig. 2. Process of structuring a speaker *GMMs* databases

2.3. A modified Kullback-Leibler divergence of Gaussian mixtures

Here are a few similarity measures classically proposed in the literature for speaker clustering:

- Methods which require models to be trained at each utterances grouping as the generalized likelihood ratio [4] or the Bayesian information criterion (BIC) [5]. These two methods have a heavy computational cost.
- Methods which do not require models to be trained at each utterances grouping as the cross likelihood ratio or the symmetric Kullback-Leibler divergence [6].

In our work, the similarity measures between a speaker *GMMs* is given by a KL divergence measurement described in [7], which is performed without require the training data of each speaker segment. However, The *KL* divergence between two Gaussian mixtures is a closed-form expression. Thus, we use the expression Eq. 1, that have a similar properties and can be expressed in analytical form.

Lets f denote the Gaussian mixture is composed of k d-dimensional Gaussian components:

$$f(y) = \sum_{i=1}^{k} \alpha_i N(y_i; \mu_i; \Sigma_i) = \sum_{i=1}^{k} \alpha_i f_i(y_i) \qquad (1)$$

The divergence measurement between two Gaussian mixtures, $f = \sum_{i=1}^{k} \alpha_i f_i$ and $g = \sum_{i=1}^{k} \beta_i g_i$ is defined:

$$d(f, g) = \sum_{i=1}^{k} \alpha_i \min_{j=1}^{k} KL(f_i || g_j) \qquad (2)$$

Each term represents the Kullback-Leibler divergence between two Gaussian density ($N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$) is defined as:

$$KL(N_1, N_2) = \frac{1}{2}(log\frac{|\Sigma_2|}{|\Sigma_1|} + Tr(\Sigma_2^{-1}\Sigma_1) +$$

$$(\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) - d \qquad (3)$$

2.4. Merge of Gaussian mixtures

The algorithm of *GMM-merge* is proposed in order to increase the indexing performance of the audio document according to speaker occurrences. Thus, for a faithful representation and resolve the intra-variability of the speech statistical parameters, the merge technique is used to update each speaker *GMM* while a new segments of the acoustic data arrive, which already exist in a tree index. In this section, we adapt and extend a recent work in order to propose a merge algorithm of Gaussian mixture models [7] [8]. This algorithm is used to achieve very efficient speaker indexing in terms of both search time and index size. In the next, we define the component mixture operations for merge :

1. **<u>"COLLAPSE"</u>**:
   The aim of the collapse operation is to simplify the *GMM* by reducing the numbers of the model component. First, the matching function π is deduced by the minimum of the *KL* divergence between each component of M_{init} according to g (where $M_{init} = f$, is the initial model) [7]. Second, the components are collapsed and represented into a new single Gaussian density $M_{c_i}^{'\pi}$. Denote *GoM(m)* the set of Gaussian mixture with *m* components, and denote the set of the m^k mappings from $\{1, ..., k\}$ to $\{1, ..., m\}$ by S.

   For a given model $M_{init} \in GoM(m)$ The matching function which gives the minimum of distance from KL is defines:

   $$\pi^g = \arg\min_{j=1}^{m} KL(g_i || M_{init_i}) i = 1, ..., k \qquad (4)$$

   For each $\pi \in S$ and $M_{init} \in$ *GoM(m)*, the collapse *GMM* is define as:

   $$M_{c_i}^{'\pi} = \frac{\sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}} M_{init_j}}{\sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}}} \qquad (5)$$

   Where the M_{init_j} prior likelihood verify ($\sum_1^k \alpha_{M_{init_i}} = 1$) and the expression of the mean and the covariance ($\mu_j^{'}$ Σ_j') is written as:

   $$\mu_j^{'} = \frac{1}{\beta_j} \sum_i \in \pi^{-1}(j) \mu_i \qquad (6)$$

   and

   $$\Sigma_j^{'} = \frac{1}{\beta_j} \sum_{i \in \pi(j)^{-1}} \alpha_i (\Sigma_i + (\mu_i - \mu_i^{'})(\mu_i - \mu_i)^{'T}) \qquad (7)$$

   with
   $$\beta_i = \sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}} \qquad (8)$$

   From where $M_{ci}^{'\pi} = N(\mu_i^{'}, \Sigma_i^{'})$ is a Gaussian density, which obtained by collapsing the $M_{init_j}^{'\pi}$ components in only one Gaussian component such as:

   $$M_{c_i}^{'\pi} = N(\mu_i^{'}, \Sigma_i^{'})$$

   $$= \arg\min_M KL(M_c^{'\pi} || g) = \arg\min_M d(M_c^{'\pi}, g) \qquad (9)$$

2. **<u>"ADD"</u>**
   For each *k* such as $\pi^{g^{-1}}(k) = \phi$, we denote $M_{Add_k}^{'} = g_k$, that consist the component of g add to $g^{'}$, in order to keep the same components number of the *GMM* of merge.

3. **<u>"PERMUTE"</u>**:
   For each *s* such as $Card(\pi(s)) = 1$, we a change the position of components as follow: $M_p^{'}(\pi(s)) = M_{init}(s)$, in order to keep a significative components, which will represents a better original model.

The obtained Gaussian mixture $g^{'}$, present the result of the first step of merging $M_{init} = f$ with g:

$$g^{'} = \sum_{(i, \pi^{g-1}(i) \neq \phi)} (\frac{\beta_{M'_{c_i}}}{\Omega}) M_{c_i}^{'} + \sum_{(j, \pi^g(j) = \phi)} (\frac{\alpha_{Add_j}}{\Omega}) M_{Add_j}^{'}$$
$$+ \sum_{(k, Card(\pi(k)) = 1)} (\frac{\alpha_{p_k'}}{\Omega}) M_{p_k}^{'} \qquad (10)$$

with:
$$\Omega = \sum_n \beta_{M_{c_n}} + \sum_m \alpha_{Add_m} + \sum_l \alpha_{p_l}^{'}$$

Finally to obtain the merge model M_{Merge}, we repeat the previous steps by replacing M_{init} by $g^{'}$ and g by f.

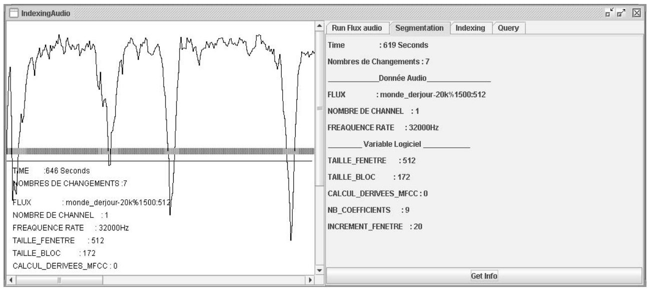
3. EXPERIMENTAL RESULTS

The proposed approaches were experimented on a archives of a *Europe1* radio operator news program, which allows to extract a flow continuously from data. Initially, The test of bayesian hypothesis is carried out with a *(5 sec)* of training data for each speaker change detected, in order to increase the accuracy of speaker recognition see Fig. 3. Thus, we use *EM* algorithm to each block of speaker acoustic data to generate a Gaussian mixture corresponds, which is composed of 16 26-dimensional Gaussian density.

The example of incremental generation of speaker homogeneous shown in as seen in Fig. 3. Four hours audio data of *"Europe1 radio operator"* is used. The segmentation presents 46 speakers segments change detected, which the acoustics data size exceeds *(5 sec)*, recall that represent the condition for a better estimate of the mixture models.

The tree of speakers model index that appear in a audio data stream carried out with incremental way, among 46 speakers segments detected 16 model speakers here been identified and inserted in Model databases. Table of index and audio meta-data is updated continuously see as seen in Fig. 5.

The hierarchical structuring in a balanced binary tree is depends directly by the initial system parameters see Fig. 4.



TIME :646 Seconds
NOMBRES DE CHANGEMENTS :7
FLUX : monde_derjour-20k%1500:512
NOMBRE DE CHANNEL : 1
FREAQUENCE RATE : 32000Hz
TAILLE_FENETRE : 512
TAILLE_BLOC : 172
CALCUL_DERIVEES_MFCC : 0

Time : 619 Seconds
Nombres de Changements : 7
____________Donnée Audio______________
FLUX : monde_derjour-20k%1500:512
NOMBRE DE CHANNEL : 1
FREAQUENCE RATE : 32000Hz
________Variable Logiciel __________
TAILLE_FENETRE : 512
TAILLE_BLOC : 172
CALCUL_DERIVEES_MFCC : 0
NB_COEFFICIENTS : 9
INCREMENT_FENETRE : 20

Fig. 3. Incremental segmentation of audio speech data

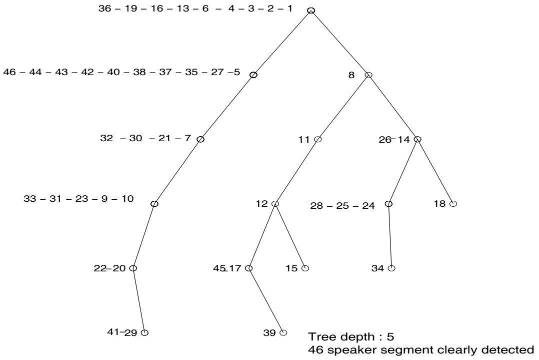
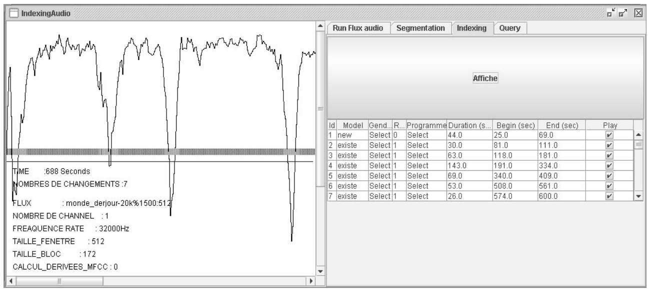


Fig. 4. The hierarchical speaker *GMMs* binary tree organization using KL-divergence measurement criterion, we add a new *GMM* speaker detected or if already exist we update his corresponding *GMM* with the *GMM-merge* alogorithm

Thus, we choose an adaptive algorithm in order to modify the threshold in order to reorganize the speakers tree index to a balanced tree form and independent of databases the speakers model what makes it possible to optimize the cost of calculation.



TIME :688 Seconds
NOMBRES DE CHANGEMENTS :7
FLUX : monde_derjour-20k%1500:512
NOMBRE DE CHANNEL : 1
FREAQUENCE RATE : 32000Hz
TAILLE_FENETRE : 512
TAILLE_BLOC : 172
CALCUL_DERIVEES_MFCC : 0

Id	Model	Gend...	R...	Programme	Duration (s...	Begin (sec)	End (sec)	Play
1	new	Select	0	Select	44.0	25.0	69.0	☑
2	existe	Select	1	Select	30.0	81.0	111.0	☑
3	existe	Select	1	Select	63.0	118.0	181.0	☑
4	existe	Select	1	Select	143.0	191.0	334.0	☑
5	existe	Select	1	Select	69.0	340.0	409.0	☑
6	existe	Select	1	Select	53.0	508.0	561.0	☑
7	existe	Select	1	Select	26.0	574.0	600.0	☑

Fig. 5. Table of index and meta data registration

4. SUMMARY

In this paper, we presented some techniques of incremental generation of speaker-homogeneous segments. We proposed a hierarchical classification approaches that applied in speaker incremental indexing framework. We applied divergence measurement method between *GMMs*. First, to discriminate between models without using of acoustics data for recognition step. Second, to create a hierarchical structure under binary tree index form. We also proposed an algorithm of merge of Gaussian mixture to update a framework and increase the accuracy of the speakers model representation.
Future work will focus on studies methods to evaluate the potentiality of hierarchical classification approaches. Experimentation of an incremental indexing system, associating an hierarchical clustering phase and concept a audio DBMS services for querying on multi-flow of continuous audio speech is also planned.

5. REFERENCES

[1] Y. Mami and D. Charlet, “Speaker identification by location in an optimal space of anchor models,” in *International Conferences on Spoken Language Processing (ICSLP ’02)*, Denver, Colorado, USA, 2002, pp. 1333–1336.

[2] D.E. Sturim, D.A. Reynolds, D.A. Singer, and E. Campbell, “Speaker indexing in large audio databases using anchor models,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’01)*, Salt Lake City, Utah, 2001, pp. 429–432.

[3] V. Upendra, J. Navratil, G. N. Ramaswamy, and S.H. Maes, “Very large population text-independant speaker identification using transformation enhanced multi-grained models,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’01)*, Salt Lake City, Utah, May 2001.

[4] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering speakers by their voices,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’98)*, Seattle, USA, May 1998, pp. 429–432.

[5] S. Chen J.F. Gales, P. Gopalakrishnan, R. Gopinath, H. Printz, D. Kanevsky P. Olsen, and L. Polymenakos, “system for transcription of broadcast news in the 1997 hub4 english evaluation,” in *Speech recognition workshop (DARPA ’98)*, 1998, pp. 389–404.

[6] R. Lutfi, M. Gelgon, and J. Martinez, “Structuring and querying documents in an audio database management system.,” *Multimedia Tools and Applications*, vol. 24, no. 2, pp. 105–123, 2004.

[7] J. Goldberger and S. Roweis, “Hierarchical clustering of mixture model,” in *Neural Information Processing Systems 17 (NIPS ’04)*, 2004, pp. 505–512.

[8] J.E. Rougui, M. Gelgon, M. Rziza, J. Martinez, and D. Aboutajdine, “Hierarchical organization of a set of gaussian mixture speaker models for scaling up indexing and retrieval in audio documents,” in *ACM Symposium on Applied Computing (SAC’06), Multimedia and Visualization (MV) track*, Dijon, France, 2006.