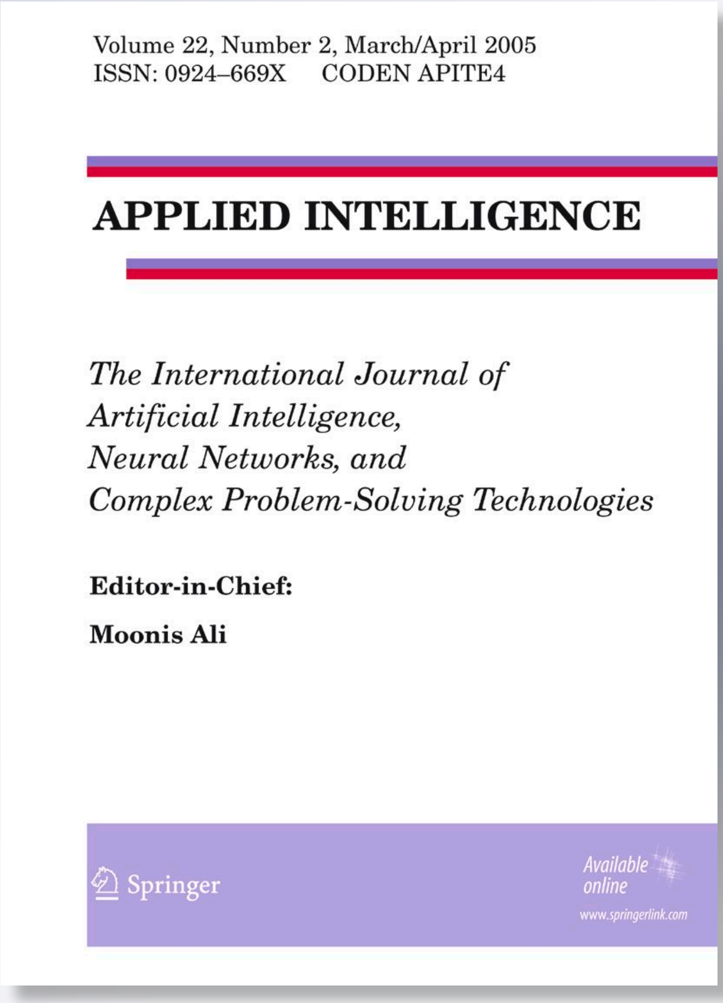
A novel split-and-merge algorithm for hierarchical clustering of Gaussian mixture models

Branislav Popović, Marko Janev, Darko Pekar, Nikša Jakovljević, Milan Gnjatović, Milan Sečujski & Vlado Delić

Applied Intelligence
The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies

ISSN 0924-669X
Volume 37
Number 3

Appl Intell (2012) 37:377-389
DOI 10.1007/s10489-011-0333-9



Volume 22, Number 2, March/April 2005
ISSN: 0924–669X CODEN APITE4

APPLIED INTELLIGENCE

The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies

Editor-in-Chief:

Moonis Ali

Springer

Available online
www.springerlink.com

Springer

Appl Intell (2012) 37:377–389
DOI 10.1007/s10489-011-0333-9

A novel split-and-merge algorithm for hierarchical clustering of Gaussian mixture models

Branislav Popović · Marko Janev · Darko Pekar · Nikša Jakovljević · Milan Gnjatović · Milan Sečujski · Vlado Delić

Published online: 12 January 2012

Abstract The paper presents a novel split-and-merge algorithm for hierarchical clustering of Gaussian mixture models, which tends to improve on the local optimal solution determined by the initial constellation. It is initialized by local optimal parameters obtained by using a baseline approach similar to k-means, and it tends to approach more closely to the global optimum of the target clustering function, by iteratively splitting and merging the clusters of Gaussian components obtained as the output of the baseline algorithm. The algorithm is further improved by introducing model selection in order to obtain the best possible trade-off between recognition accuracy and computational load in a Gaussian selection task applied within an actual recognition system. The proposed method is tested both on artificial data and in the framework of Gaussian selection performed within a real continuous speech recognition system, and in both cases an improvement over the baseline method has been observed.

Keywords Gaussian mixtures · Split-and-merge operation · Hierarchical clustering · Continuous speech recognition

1 Introduction

Gaussian Mixture Models (GMMs) are extensively used for density estimation and data clustering in the fields of image and speech processing [1, 2], as well speech and emotion recognition and speaker identification [3]. Clustering procedure is one of the most important components in any recognition task [4–6]. Thus, clustering of Gaussian mixture components, i.e., Hierarchical Gaussian Mixture Model Clustering (HGMMC) (see [7, 8]) is a key component of Gaussian Selection (GS) methods [9–12]. These are used for increasing the speed of recognition systems based on GMM models, especially in the areas of Continuous Speech Recognition (CSR), speaker verification [3] and speaker adaptation [13]. The purpose of HGMMC is to group and re-represent the Gaussian components from the original Gaussian mixture, and also to create a compact, i.e., simplified representation of the underlying mixture, with the restriction that no original component could be split in the reduced representation. The aim is to simplify the underlying GMM distribution that has already been learned from the observed data set; by forming a simplified hyper-distribution, i.e. hyper-GMM. The Gaussian selection techniques are designed to reduce the likelihood computation load by computing only the likelihood of a selected subset of mixtures for a particular input vector. The basic idea is to cluster all

B. Popović (✉) · N. Jakovljević · M. Gnjatović · M. Sečujski · V. Delić
Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
e-mail: bpopovic@uns.ac.rs

N. Jakovljević
e-mail: jakovnik@uns.ac.rs

M. Gnjatović
e-mail: gnjatovi@uns.ac.rs

M. Sečujski
e-mail: secujski@uns.ac.rs

V. Delić
e-mail: vdelic@uns.ac.rs

M. Janev
Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
e-mail: markojan@uns.ac.rs

D. Pekar
Alfanum Speech Technologies, Novi Sad, Serbia
e-mail: darko.pekar@alfanum.co.rs

Gaussian components that are used in the recognition system, and to assign unique hyper-Gaussians with appropriately estimated parameters to these clusters. In the decoding process, for a given observation, likelihoods of only those Gaussian components which belong to a predefined percent $\theta[\%]$ of clusters whose corresponding hyper-Gaussians have the highest likelihoods, are evaluated directly, while the likelihoods of all other Gaussian components are approximated (floored). The approximate value for a particular Gaussian component can be e.g. the likelihood of the corresponding hyper-Gaussian evaluated for a particular observation.

The idea of GS was first introduced in [8], and it was refined and efficiently applied to CSR problems in the work of Gales [9], but the method still did not include the full Gaussian covariance in the actual clustering and selection procedure. In [10] and [11], the method was generalized so it could be hierarchical, and to use Gaussian component covariances in clustering. The mathematical aspects of a similar, but more formal approach were presented in [7], and according to this approach, efficient Gaussian component clustering algorithm, suitable for application in any of the aforementioned recognition tasks, has been developed. The HGMMC algorithm presented in [7] is used in this paper as a baseline algorithm, as well as for comparison of the experimental results. However, since it is based on vector quantization, i.e., the Lindo-Buzo-Gray [14, 15] approach, which is similar to k-means, after a finite number of iterations it converges only to a local optimal solution. To our knowledge, the similar property holds for all GS methods currently present in the literature [8–11]. We also note that there are other approaches beside GS, related to the acceleration of GMMs, subspace tying techniques (see [16–18]) and kd-tree based techniques (see [19, 20] and [21]).

A novel Split-and-Merge algorithm for Hierarchical Clustering of Gaussian Mixture Models (S&M HGMMC) proposed in this paper tends to improve on the local optimal solution determined by its initial constellation. The algorithm is initialized with locally optimal values obtained by using the baseline HGMMC method presented in [7], and it tends to approach to the global optimum of the clustering target function more closely by iteratively splitting, and merging the clusters of Gaussian components obtained as the output of the baseline method. The novel clustering algorithm is also applied to GS, in order to achieve a better trade-off between speed and accuracy in a recognition task (particularly CSR) than in case of the baseline method presented in [7], as well as in case of other similar methods.

Our motivation stems from a study by Ueda et al. [22], who proposed the Split-and-Merge operation in order to address the problem of local convergence of the standard EM algorithm for estimating the parameters of GMM models. There are also several other researchers who have studied the problem of finding the criteria for efficient selection of Split-and-Merge candidates [23, 24] in the task of EM learning of GMM models from observed data. These ideas have been widely used in computer vision, pattern recognition, and signal processing. In all these works, relative to the task of GMM learning, the merge operation is well posed, as it is based on ML estimation of Gaussian parameters in case there is only one Gaussian component. On the other hand, as a split criterion for a particular cluster of observations, local Kullback-Leibler (KL) divergence has been proposed (see [24]), i.e. the Gaussian component with the greatest local KL divergence is the one to be split. It is an ill-posed problem, as there is no unique solution to it.

In our work, the concept of Split-and-Merge operations initially proposed in [24] is incorporated in the framework of HGMMC proposed in [7], which corresponds to a GS process. As a merge criterion, we selected the minimal KL divergence between two arbitrary Gaussian components, obtained by collapsing the underlying GMMs attached to the corresponding clusters. The merge operation is a well-posed problem, as there is a closed form solution to it. The optimal way to estimate the parameters of the hyper-Gaussian that corresponds to the resulting cluster is also proposed however, the generalization of the split operation is problematic. On one hand, the natural choice of the candidate cluster for the split operation is the cluster with the largest KL divergence between the hyper-Gaussian that represents it, and the whole underlying GMM that corresponds to it. On the other hand, there is no closed form expression for the KL divergence between two Gaussian mixtures (except in the trivial case when each mixture contains only one Gaussian component) thus, two different approximations for the mentioned KL divergence are used, both are proposed by Hershey and Olsen in [22], in order to implement the S&M method in a feasible way. The first, which is more precise, but much more computationally demanding is the Monte Carlo Sampling Approximation (MCSA), and the second one, much less computationally demanding, but still sufficiently precise, is the Lower Bound Variational Approximation (LBVA). The split operation is an ill-posed problem, a possible solution to it is proposed within this research, also a possible way to incorporate the model selection mechanism into the proposed S&M HGMMC approach is also suggested, aiming at a better trade-off between computational efficiency and accuracy, when the clustering is applied within a GS task, in a real recognition system.

Experiments have been conducted on artificial data as well as on a real GS system, used within a CSR task. In the experiments conducted on artificial data, for the proposed S&M HGMMC, an improvement in terms of the values of the cost function in comparison to the baseline HGMMC was obtained in a large percent of cases, for all the tested dimensions, and for both approximations of mixture KL divergence that were used. Furthermore, in the experiments on a real GS system, a better trade-off between recognition accuracy and computational efficiency was achieved when using the proposed S&M HGMMC, in comparison to the GS system using the baseline HGMMC. Even better results were obtained when the model selection mechanism mentioned above was incorporated.

The paper is organized as follows: Sect. 2 defines the problem and gives the description of the baseline algorithm used as a starting point for Split-and-Merge iteration, as well as a reference point for evaluation of experimental results. In Sect. 3, a novel S&M HGMMC is presented, along with the appropriate Split-and-Merge criteria and operations. In Sect. 4 it is explained how the model selection mechanism can be incorporated into the proposed method. Section 5 presents the experimental results, which favor the proposed approach over the baseline algorithm, confirming the considerations from previous sections. In Sect. 6, several conclusions have been drawn.

2 Problem formulation and the baseline algorithm

Let us consider the original (ground-truth) Gaussian Mixture Model (GMM) f with $k > 1$ d-dimensional Gaussian components:

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

where α_i is the occupancy of the i-th component in the mixture and $f_i = N(y; \mu_i, \Sigma_i)$ is the actual i-th Gaussian component.

The task of clustering GMMs is to cluster the components of the original Gaussian mixture f and to create a new, more compact, i.e. simplified representation g of the original mixture f, with $m \ll k$ components, following the restriction that no original component could be split in the simplified representation. The distribution g is called the hyper-mixture or hyper-GMM, and its components are called hyper-Gaussians. Goldberger and Roweis [7] proposed an extended model-based approach that performs hierarchical clustering of a GMM, while still preserving the component structure from the original model. Since the most natural criterion, based on the KL divergence between two GMMs, leads to an intractable optimization problem, they introduced a new, analytically more tractable distance measure between the models f and g:

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

Let us denote the set of all d-dimensional GMMs with at most m components by $MoG(m)$, and the set of all partitions of length m (i.e. of m elements in the partition) of the set $\Omega_k = \{1, \ldots, k\}$, which is the set that indexes all Gaussian components in the original mixture f, by P_m. It means that any $\pi \in P_m$ can be represented by $\pi = \{\pi_1, \ldots, \pi_m\}$, where: $\Omega_k = \bigcup_{j=1}^{m} \pi_j$, $\pi_j \neq \varnothing$, $i \neq j \Rightarrow \pi_i \cap \pi_j = \varnothing$ holds for all $i, j \in \{1, \ldots, m\}$. Each partition $\pi \in P_m$ is considered as one particular clustering into m clusters, of the set of Gaussian components $\{f_i\}_1^k$, where π_j, $j \in \{1, \ldots, m\}$ are particular clusters of Gaussians, and π is referred to as the matching partition. Let $\pi \in P_m$ be fixed. Then for any $i \in \Omega_k$, let us denote $\pi(i) = \pi_j$, if $i \in \pi_j$. The term $\pi(i)$ is the class of equivalence that $i \in \Omega_k$ belongs to, given the unique equivalence relation that the partition $\pi \in P_m$ generates. If an auxiliary function $d(f, g, \pi)$ is defined as

$$d(f, g, \pi) = \sum_{i=1}^{k} \alpha_i KL(f_i \parallel g_{\pi(i)}) \tag{3}$$

for any $\pi \in P_m$, $g \in MoG(m)$, for a given $g \in MoG(m)$, the matching partition $\pi^g \in P_m$ is

$$\pi^g(i) = \arg\min_{j=1}^{m} KL(f_i \parallel g_j), \quad i = 1, \ldots, k \tag{4}$$

and it can be easily shown (see [7]) that

$$d(f, g) = d(f, g, \pi^g) = \min_{\pi \in S} d(f, g, \pi) \tag{5}$$

Given a matching partition $\pi \in P_m$, the function $g^\pi \in MoG(m)$ is defined as the mixture in which the Gaussian component g_j^π is obtained by collapsing the distribution

$$f_j^\pi = \frac{\sum_{i \in \pi_j} \alpha_i f_i}{\sum_{i \in \pi_j} \alpha_i} \tag{6}$$

into a single Gaussian. The mean and the covariance of f_j^π are obtained as

$$\begin{aligned} \tilde{\mu}_j &= \frac{1}{\beta_j} \sum_{i \in \pi_j} \alpha_i \mu_i, \\ \tilde{\Sigma}_j &= \frac{1}{\beta_j} \sum_{i \in \pi_j} \alpha_i (\Sigma_i + (\mu_i - \tilde{\mu}_j)(\mu_i - \tilde{\mu}_j)^T), \\ \beta_j &= \sum_{i \in \pi_j} \alpha_j \end{aligned} \tag{7}$$

where $f_i = N(\mu_i, \Sigma_i)$ are the ground-truth Gaussian components. It has been shown (see [7], Lemma 1), that for the single Gaussian obtained as $g_j^\pi = N(\tilde{\mu}_j, \tilde{\Sigma}_j)$ with occupancies β_j, the following holds:

$$g_j^\pi = \arg\min_{g \in MoG(1)} KL(f_j^\pi \parallel g) = \arg\min_{g \in MoG(1)} d(f_j^\pi, g) \tag{8}$$

where minimization is performed on the set of d-dimensional Gaussian densities, implying that g_j^π is in fact obtained by collapsing f_j^π into a single Gaussian. If one denotes the collapsed version of f according to some fixed $\pi \in P_m$ as g^π, i.e.

$$g^\pi = \sum_{j=1}^{m} \beta_j g_j^\pi \tag{9}$$

then it has been shown in [7] that, given a Mixture of Gaussians (MoG) f and a matching partition π, the MoG g^π is a unique minimum point for $d(f, g, \pi)$. As the main double minimization problem $(\hat{g}, \hat{\pi}) = \arg\min_{g \in MoG(m)} \times \min_{\pi \in P_m} d(f, g, \pi)$ cannot be solved analytically for arbitrary $(g, \pi) \in MoG(m) \times P_m$, an alternating minimization procedure is proposed (see [7]) in order to reach a local minimum of $d(f, g)$. It is formulated as the alternating and iterative application of regroup and refit operations

$$\pi^g = \arg\min_{\pi} d(f, g, \pi) \quad \text{Regroup} \tag{10}$$

$$g^\pi = \arg\min_{g} d(f, g, \pi) \quad \text{Refit} \tag{11}$$

It is clear that performing these operations decreases the cost function $d(f, g)$ defined by (2) monotonically (see [7]), and since P_m is finite, and thus also $MoG(m)$, the cost function converges to a local minimum $\hat{g}$ of $d(f, g)$ on $MoG(m)$.

3 A novel Split-and-Merge algorithm for Gaussian mixture clustering

A study by Ueda et al. [24] has been used as motivation to enforce Split-and-Merge iterations in order to overcome the problem of a local optimal solution of the EM algorithm applied to learning of GMM from the observed data points. For the split criterion, the local KL divergence, given by

$$J_{split}(i; \psi) = \int f_i(x; \psi) \log \frac{f_i(x; \psi)}{p(x; \psi_i)} dx \tag{12}$$

has been adopted. It represents the distance between the two distributions, the local data density around the i-th Gaussian component and the i-th Gaussian density specified by the current parameter estimate (see [23] or [24] for more details). The Gaussian component with the greatest distance $J_{split}(i; \psi)$ has the least precise estimate, and therefore it should be split into two components whose parameters are then estimated using EM in the ML manner. As a merge criterion, the maximal correlation coefficient of two Gaussian components has been used. This particular concept is further developed in [23].

In this section, a novel Split-and-Merge algorithm for Hierarchical Clustering of Gaussian Mixture Models (S&M HGMMC) is proposed. It shares the basic idea with the S&M algorithms presented for learning GMMs [22–24] insofar as it tends to improve on a local optimal solution for cost (2) by iteratively applying Split-and-Merge operations. We proceed further with the idea by rewriting (3) in the following equivalent form

$$d(f, g, \pi) = \sum_{j=1}^{m} A_j, \quad A_j = \sum_{i \in \pi_j} \alpha_i KL(f_i \parallel g_j) \tag{13}$$

for a given ground truth distribution $f \in MoG(k)$, arbitrary $g \in MoG(m)$ and $\pi \in P_m$, for a fixed $m \in N$.

Let us suppose that $(\hat{g}, \hat{\pi}) \in MoG(m) \times P_m$ is obtained as the local minimum of $d(f, g, \pi)$, given by (3) and (13), and let $m > 3$. The cost (13) for obtained $(\hat{g}, \hat{\pi})$ can be represented as

$$d(f, \hat{g}, \hat{\pi}) = \hat{A}_{\hat{k}_1} + \hat{A}_{\hat{k}_2} + \hat{A}_{\hat{j}} + \hat{A}, \quad \hat{A} = \sum_{\substack{j \in \{1,\ldots,m\} \\ j \notin \{\hat{k}_1, \hat{k}_2, \hat{j}\}}} \hat{A}_j \tag{14}$$

for some fixed $\hat{k}_1, \hat{k}_2, \hat{j} \in \{1, \ldots, m\}$, where $\hat{A}_j = \sum_{i \in \hat{\pi}_j} \alpha_i KL(f_i \parallel \hat{g}_j)$. Let us suppose, without loss of generality, that clusters $\hat{\pi}_{\hat{k}_1}, \hat{\pi}_{\hat{k}_2} \in \hat{\pi}$ have been chosen to be merged, according to some adopted merge criterion, thus forming a new cluster $\hat{\pi}_{\hat{k}} = \hat{\pi}_{\hat{k}_1} \cup \hat{\pi}_{\hat{k}_2}$ and a new matching partition $\pi' = \{\hat{\pi}_{\hat{k}}\} \cup \{\hat{\pi}_j | j \in \{1, \ldots, m\} \backslash \{\hat{k}_1, \hat{k}_2\}\}$. The Gaussian component $\hat{g}_{\hat{k}}$ and occupancy $\hat{\beta}_{\hat{k}}$ that correspond to $\hat{\pi}_{\hat{k}}$ are estimated using (7). Note that $\pi' \in P_{m-1}$, and that it is obtained from $\hat{\pi}$ by simply merging $\hat{\pi}_{\hat{k}_1}$ and $\hat{\pi}_{\hat{k}_2}$ into a single cluster. Also note that $g' \in MoG(m-1)$, which corresponds to π', is given by its components: $\hat{g}_j$, $j \notin \{\hat{k}_1, \hat{k}_2\}$, $\hat{g}_{\hat{k}}$ and their corresponding occupancies. Moreover, for both π' and components of g', re-indexing is performed using the index set $\{1, \ldots, m-1\}$. Let us also suppose, without loss of generality, that the cluster $\hat{\pi}_{\hat{j}} \in \hat{\pi}$ is chosen to be split, based on some adopted split criterion, and that two new clusters $\hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_2}$, where $\hat{\pi}_{\hat{j}} = \hat{\pi}_{\hat{j}_1} \cup \hat{\pi}_{\hat{j}_2}$, are obtained by using an appropriate split method, where the estimates for parameters of corresponding Gaussian components $\hat{g}_{\hat{j}_1}, \hat{g}_{\hat{j}_2}$ and occupancies $\hat{\beta}_{\hat{j}_1}, \hat{\beta}_{\hat{j}_2}$ are obtained by using (7). In that way, the new partition $\tilde{\pi} \in P_m$ is obtained, defined as $\tilde{\pi} = (\pi' \backslash \{\hat{\pi}_{\hat{j}}\}) \cup \{\hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_2}\}$, and the corresponding hyper-GMM $\tilde{g} \in MoG(m)$, containing $m - 3$ components $\hat{g}_j$ for $j \notin \{\hat{k}_1, \hat{k}_2, \hat{j}\}$ and three additional Gaussian components $\hat{g}_{\hat{k}}$, $\hat{g}_{\hat{j}_1}$ and $\hat{g}_{\hat{j}_2}$, that are obtained by collapsing the GMMs that correspond to clusters $\hat{\pi}_{\hat{k}}$, $\hat{\pi}_{\hat{j}_1}$ and $\hat{\pi}_{\hat{j}_2}$ (see (6) and (7)), together with the corresponding occupancies. Re-indexing of components of $\tilde{\pi}$ and $\tilde{g}$ using the index set $\{1, \ldots, m\}$ is also performed. Note that our actual proposal for the mentioned Split-and-Merge criteria and the split method is given in Sects 3.1 and 3.2. Now, let us define $\tilde{A}_j = \sum_{i \in \tilde{\pi}_j} \alpha_i KL(f_i \parallel \tilde{g}_j)$, $\tilde{A} = \sum_{\substack{j \in \{1,\dots,m\} \\ j \notin \{\hat{k}, \hat{j}_1, \hat{j}_2\}}} \tilde{A}_j$, and consider the condition

$$\hat{A}_{\hat{k}_1} + \hat{A}_{\hat{k}_2} + \hat{A}_{\hat{j}} \geq \tilde{A}_{\hat{k}} + \tilde{A}_{\hat{j}_1} + \tilde{A}_{\hat{j}_2}. \tag{15}$$

Since the relation $\tilde{A} = \hat{A}$ holds, this condition implies that $d(f, \tilde{g}, \tilde{\pi}) \leq d(f, \hat{g}, \hat{\pi})$, and because of (5) and the fact that $\hat{\pi} = \pi^{\hat{g}}$ and $\tilde{\pi} = \pi^{\tilde{g}}$ (for definition of π^g see Sect. 2) it eventually implies that $d(f, \tilde{g}) \leq d(f, \hat{g})$.

Therefore, with such Split-and-Merge criteria and an adequate method of splitting cluster $\hat{\pi}_{\hat{j}}$ into $\hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_2}$, which would lead to (15) with sufficient possibility, the introduction of S&M operations in the regroup-refit algorithm given by (10) and (11) could, in a majority of cases, lead to lower local minima of $d(f, g)$ on $MoG(m)$. In view of the previous fact, the actual Split-and-Merge criteria, together with the method of performing the split operation, are crucial for making the S&M operations performed within the regroup-refit algorithm (given by (10) and (11)) efficient in lowering the cost function (2). The split and the merge criterion are chosen in a natural way, using the KL divergence as the well established informational pseudo-distance. Also, the unique solution for obtaining the optimal parameters for $\hat{g}_{\hat{k}}$ is provided, which makes the proposed merge operation well-posed. Nevertheless, as there is no expression for the KL divergence between two arbitrary GMMs, approximations are introduced in order to make the proposed split criterion functional. Moreover, as the actual split operation is ill-posed, one simple and effective solution is proposed in this paper.

The considerations concerning (14) and (15) are related to the case when the merge operation is performed prior to the split operation. It should be noted that similar conclusions could also be drawn in the opposite case. This fact is used in Sect. 4, where the model selection paradigm is incorporated into the proposed S&M Gaussian model clustering framework in order to obtain a better trade-off between computational load and accuracy in the GS task applied within a real CSR system, while the experiments are presented in Sect. 5.2.

3.1 Merge operation

The merge operation is proposed as follows: For the currently obtained partitioning $\hat{\pi} \in P_m$ and the hyper-mixture $\hat{g} \in MoG(m)$, the informational criterion based on informational KL distance is chosen. This criterion is considered to be the most natural and potentially efficient, as it could lead to the condition (15) being satisfied most frequently. Two clusters $\hat{\pi}_{\hat{k}_1}, \hat{\pi}_{\hat{k}_2} \in \hat{\pi}$, with their corresponding Gaussian components, are chosen to be merged if

$$(\hat{k}_1, \hat{k}_2) \in \operatorname*{arg\,min}_{(k_1,k_2) \in (\{1,\dots,m\})^2} KL(\hat{g}_{k_1} \parallel \hat{g}_{k_2}) \tag{16}$$

The expression for $KL(h_1 \parallel h_2)$ for any d-dimensional Gaussians h_1 and h_2 exists in a closed form given by

$$KL(h_1 \parallel h_2) = \frac{1}{2}\Bigg(\log \frac{|\Sigma_2|}{|\Sigma_1|} + \operatorname{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) - d \Bigg) \tag{17}$$

where (μ_1, Σ_1) and (μ_2, Σ_2) are parameters of Gaussians h_1 and h_2 respectively. If multiple solutions exist, any one of them is chosen arbitrarily. Next, a well-posed merge operation is performed: the cluster $\hat{\pi}_{\hat{k}} = \hat{\pi}_{\hat{k}_1} \cup \hat{\pi}_{\hat{k}_2}$ and the corresponding hyper-Gaussian $\hat{g}_{\hat{k}}$ (see Sect. 3) are obtained, using the estimates given by (7), for the newly obtained mixture

$$f_{\hat{k}}^{\pi} = \frac{\sum_{i \in \hat{\pi}_{\hat{k}_1}} \alpha_i f_i + \sum_{i \in \hat{\pi}_{\hat{k}_2}} \alpha_i f_i}{\sum_{i \in \hat{\pi}_{\hat{k}_1} \cup \hat{\pi}_{\hat{k}_2}} \alpha_i} \tag{18}$$

Note that the merge operation used in the S&M GMM learning framework [23, 24], contains the explicit EM optimization step for a GMM consisting of only one Gaussian component that corresponds to the newly obtained cluster (see [24]). In the proposed framework of S&M in a Gaussian mixture clustering task, when the merge operation is performed, the actual "optimization" is done implicitly, since the hyper-Gaussian $\hat{g}_{\hat{k}}$ that corresponds to the cluster $\hat{\pi}_{\hat{k}}$ yields the minimal KL divergence to the GMM (18), as it is obtained by using estimates (7).

3.2 Split operation

For the same reasons as in the case of the merge operation, we use the KL divergence as the criterion for the split operation. Let $\hat{\pi} \in P_m$ be the partition obtained in the current step of the S&M procedure and let and $\hat{g} \in MoG(m)$ be its corresponding hyper-mixture. For the split operation, the cluster $\hat{j} \in \hat{\pi}$, satisfying the condition

$$\hat{j} \in \operatorname*{arg\,max}_{j \in \{1,\dots,m\} \backslash \{\hat{k}_1, \hat{k}_2\}} KL(f_j^{\hat{\pi}} \parallel \hat{g}_j) \tag{19}$$

is chosen, where $f_j^{\hat{\pi}}$ is the Gaussian mixture that corresponds to the cluster $\hat{\pi}_j$, $\hat{g}_j$ is the component of $\hat{g}$ that corresponds to the j-th cluster, and the clusters $\hat{k}_1, \hat{k}_2$ are selected for the merge operation that is to be conducted prior to the split operation. As in the case of merge operation, if there are multiple solutions, one of them is chosen arbitrarily.

As it is well known (see e.g. [22]), contrary to the KL divergence between two Gaussians, for which there is a closed form expression, as given by (17), there is no closed form expression for the KL divergence between two arbitrary Gaussian mixtures. Thus, in order to make the criterion (19) feasible, we used two different approximations for $KL(f_j^{\hat{\pi}} \parallel \hat{g}_j)$ which exist in the literature (see [22]), and tested both of them experimentally. The approximations that were used are presented in [25], and are briefly explained in the sequel. The first is Monte Carlo Sampling Approximation (MCSA), given by:

$$D_{MCSA}(f_j^{\hat{\pi}} \parallel \hat{g}_j) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f_j^{\hat{\pi}}(x_i)}{\hat{g}_j(x_i)} \tag{20}$$

where $\{x_i\}_{i=1}^n$ are i.i.d. samples drawn from the distribution $f_j^{\hat{\pi}}$. In the experiments, $\{x_i\}_{i=1}^n$ are actually drawn from one of the Gaussian components f_i, $i \in \hat{\pi}_j$ independently. For the actual number of observations n_i generated from a particular f_i, we take $n_i = \lceil \alpha_i / \sum_{i\in\pi_j} \alpha_i \rceil$, having in mind that $n = \sum_{i\in\pi_j} n_i$. As its variance of estimation converges to 0 when $n \to \infty$ (see [22]), for a large number of generated samples n, an arbitrary good estimate of the true $KL(f_j^{\hat{\pi}} \parallel \hat{g}_j)$ can be obtained. On the other hand, regardless of the fact that actual clustering is performed off-line, evaluation of such an estimation could be too computationally demanding for real world applications.

In order to obtain a significantly lower computational load in the S&M HGMMC task, the lower bound variational approximation (LBVA) is used (see [22]). It is given by

$$D_{LBVA}(f_j^{\hat{\pi}} \parallel \hat{g}_j) = \sum_{i\in\hat{\pi}_j} \alpha_i \ln \frac{\sum_{k\in\hat{\pi}_j} \alpha_k e^{-KL(f_i \| f_k)}}{e^{-KL(f_i \| \hat{g}_j)}} \tag{21}$$

which is a closed form, since $KL(f_i \parallel f_k)$ and $KL(f_i \parallel \hat{g}_j)$, the KL divergences between particular Gaussian components, also exist in the closed form given by (17). It is a much less computationally demanding approximation than MCSA, but sufficiently efficient nevertheless, and has been confirmed in the experiments presented in Sect. 5.

Let us assume that the cluster $\hat{\pi}_{\hat{j}}$ is chosen to be split into clusters $\hat{\pi}_{\hat{j}_1}$ and $\hat{\pi}_{\hat{j}_2}$. Firstly, the parameters of the hyper-Gaussians $\hat{g}_{\hat{j}_1}^{(0)}, \hat{g}_{\hat{j}_2}^{(0)}$ and the corresponding occupancies are initialized as:

$$\hat{\beta}_{\hat{j}_1}^{(0)} = \hat{\beta}_{\hat{j}_2}^{(0)} = \frac{\hat{\beta}_{\hat{j}}}{2}, \qquad \hat{\Sigma}_{\hat{j}_1}^{(0)} = \hat{\Sigma}_{\hat{j}_2}^{(0)} = \frac{\hat{\Sigma}_{\hat{j}}}{2} \tag{22}$$

$$\hat{\mu}_{\hat{j}_1}^{(0)} = \hat{\mu}_{\hat{j}} + \frac{\sqrt{\lambda_{max}}}{2} v_{max}, \qquad \hat{\mu}_{\hat{j}_2}^{(0)} = \hat{\mu}_{\hat{j}} - \frac{\sqrt{\lambda_{max}}}{2} v_{max} \tag{23}$$

where $(\hat{\mu}_{\hat{j}}, \hat{\Sigma}_{\hat{j}})$ are the parameters of the Gaussian $\hat{g}_{\hat{j}}$ that corresponds to the cluster $\hat{\pi}_{\hat{j}}$, and $\lambda_{\max}$, $v_{\max}$ are the maximum eigenvalue and the corresponding eigenvector obtained for $\hat{\Sigma}_{\hat{j}}$. For such $\hat{g}_{\hat{j}_1}^{(0)}, \hat{g}_{\hat{j}_2}^{(0)}$, a regroup operation given by (10) is performed in order to obtain $\hat{\pi}_{\hat{j}_1}^{(0)}, \hat{\pi}_{\hat{j}_2}^{(0)}$. Next, the regroup-refit algorithm given by (10) and (11), described in Sect. 2, is conducted on the cluster pair $(\hat{\pi}_{\hat{j}_1}^{(0)}, \hat{\pi}_{\hat{j}_2}^{(0)})$, until it converges in a number of steps S, thus obtaining $(\hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_1}) = (\hat{\pi}_{\hat{j}_1}^{(S)}, \hat{\pi}_{\hat{j}_2}^{(S)})$, the corresponding hyper-Gaussians $(\hat{g}_{\hat{j}_1}, \hat{g}_{\hat{j}_2}) = (\hat{g}_{\hat{j}_1}^{(S)}, \hat{g}_{\hat{j}_2}^{(S)})$ and occupancies $(\hat{\beta}_{\hat{j}_1}, \hat{\beta}_{\hat{j}_2}) = (\hat{\beta}_{\hat{j}_1}^{(S)}, \hat{\beta}_{\hat{j}_2}^{(S)})$. This is actually the optimization step. Note that, as the split operation is ill-posed, one possible solution is given, in which the parameters are initialized in a simple way. Also, note that since the initialization could affect even the two clusters problem, especially in the large dimensional case (such as the CSR task), there is some space to improve the performance of the proposed algorithm by using "better" initialization, but it is left for some future work.

As the last step of each iteration of the S&M algorithm, after the split and the merge steps, the baseline algorithm initialized with their output $(\tilde{\pi}, \tilde{g})$ is performed until it converges and (π^*, g^*) is obtained. If $d(f, g^*, \pi^*) < d(f, \hat{g}, \hat{\pi})$, where $(\hat{\pi}, \hat{g})$ is the result of the previous iteration of the S&M algorithm, then $(\hat{\pi}, \hat{g}) \leftarrow (\pi^*, g^*)$ is performed, and the algorithm proceeds with the next iteration. If not, the process is finished, and $(\hat{\pi}, \hat{g})$ is adopted as the final solution.

The block diagrams of the baseline HGMMC and the proposed S&M HGMMC, are given below:

HGMMC

- Initialization:
  - For a predefined n_{avr} and the overall number of Gaussian components k, obtain the predefined number of clusters as: $m = \lfloor k/n_{avr} \rfloor$.
  - Select at random (uniform distribution) m different centroids μ_j from the set of k mixture centroids used. Assign to every centroid the identity covariance matrix $\Sigma_j^{(0)} = I$. Perform one iteration of the regroup-refit algorithm given by (10) and (11), and obtain $(\hat{\pi}^{(0)}, \hat{g}^{(0)})$.
- Clustering:

  For a predefined $\varepsilon > 0$:
  - Start with $(\hat{\pi}^{(0)}, \hat{g}^{(0)})$ and repeat the regroup-refit algorithm given by (10) and (11) until in a p-th iteration $|d(f, \hat{g}^{(p)}, \hat{\pi}^{(p)}) - d(f, \hat{g}^{(p-1)}, \hat{\pi}^{(p-1)})| < \varepsilon$ holds. The final $(\pi^*, g^*) = (\pi^{(p)}, g^{(p)})$ is adopted.

S&M HGMMC

- Initialization:
  - Perform the regroup-refit algorithm described in Sect. 2 until it converges to $(\hat{\pi}, \hat{g})$ for the predefined n_{avr}, i.e., the predefined number of clusters $m = \lfloor k/n_{avr} \rfloor$ and a predefined $\varepsilon > 0$, and obtain initial $(\hat{\pi}^{(0)}, \hat{g}^{(0)}) = (\hat{\pi}, \hat{g})$.

– Clustering:
Start with $(\hat{\pi}^{(0)}, \hat{g}^{(0)})$ and repeat the following until in a p-th iteration $d(f, \hat{g}^{(p)}, \hat{\pi}^{(p)}) \geq d(f, \hat{g}^{(p-1)}, \hat{\pi}^{(p-1)})$:

- Let $(\hat{\pi}, \hat{g}) \leftarrow (\hat{\pi}^{(p)}, \hat{g}^{(p)})$. Select candidates $\hat{\pi}_{\hat{k}_1}$, $\hat{\pi}_{\hat{k}_2} \in \hat{\pi}$ for merge operation based on the criterion (16). Perform the merge operation described in Sect. 3.1, by merging $\hat{k}_1$ and $\hat{k}_2$ into $\hat{k}$, i.e. $\hat{\pi}_{\hat{k}} = \hat{\pi}_{\hat{k}_1} \cup \hat{\pi}_{\hat{k}_2}$.
- Perform the optimization step: obtain $\hat{g}_{\hat{k}}$ that corresponds to the mixture (18), using estimates (7).
- Select the candidate $\hat{\pi}_{\hat{j}} \in \hat{\pi} \backslash \{\hat{\pi}_{\hat{k}_1}, \hat{\pi}_{\hat{k}_2}\}$ for performing the split operation by using the criterion (19), together with some chosen approximation ((20) or (21)).
- Perform the optimization step: obtain initial $\beta^{(0)}_{\hat{j}_1}, \beta^{(0)}_{\hat{j}_2}$, $(\mu^{(0)}_{\hat{j}_1}, \Sigma^{(0)}_{\hat{j}_1})$ and $(\mu^{(0)}_{\hat{j}_2}, \Sigma^{(0)}_{\hat{j}_2})$ using (22) and (23). Perform one regroup operation in order to obtain the corresponding cluster pair $(\hat{\pi}^{(0)}_{\hat{j}_1}, \hat{\pi}^{(0)}_{\hat{j}_2})$. Perform the regroup-refit algorithm described in Sect. 2 on $(\hat{\pi}^{(0)}_{\hat{j}_1}, \hat{\pi}^{(0)}_{\hat{j}_2})$, until it converges in an S-th iteration for the predefined $\varepsilon > 0$, thus obtaining $(\hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_2}) = (\hat{\pi}^{(S)}_{\hat{j}_1}, \hat{\pi}^{(S)}_{\hat{j}_2})$, $(\hat{g}_{\hat{j}_1}, \hat{g}_{\hat{j}_2}) = (\hat{g}^{(S)}_{\hat{j}_1}, \hat{g}^{(S)}_{\hat{j}_2})$, $(\hat{\beta}_{\hat{j}_1}, \hat{\beta}_{\hat{j}_2}) = (\hat{\beta}^{(S)}_{\hat{j}_1}, \hat{\beta}^{(S)}_{\hat{j}_2})$ as described in Sect. 3.2.
- Perform one iteration of the regroup-refit algorithm given by (10) and (11), on $(\hat{\pi}\backslash\{\hat{\pi}_{\hat{k}_1}, \hat{\pi}_{\hat{k}_2}, \hat{\pi}_{\hat{j}}\}) \cup \{\hat{\pi}_{\hat{k}}, \hat{\pi}_{\hat{j}_1}, \hat{\pi}_{\hat{j}_2}\}$ (see Sect. 3), until it converges with the predefined $\varepsilon > 0$, thus obtaining $(\hat{\pi}^{(p)}, \hat{g}^{(p)})$.

Adopt final $(\pi^*, g^*) = (\pi^{(p-1)}, g^{(p-1)})$.

4 Incorporating model selection into the proposed Split-and-Merge framework

One of the important things in any clustering task is the model selection, i.e. the question of the "optimal" number of clusters m^* that is to be imposed on the model in order to avoid overfitting. In the task of GMM learning, there are various model selection techniques, such as e.g. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or the more efficient cross-validation technique [26]. While AIC and BIC are not sufficiently reliable for many applications, cross-validation is extremely computationally demanding. Nevertheless, being based on informational criteria, all three techniques strive to get close to the optimal (or a sub-optimal) size of GMM that describes the underlying distribution (which is not exactly known, but the set of observations drawn from it is available), so that some informational distance between the underlying distribution and that particular GMM is as low as possible. Nevertheless, in the case of HGMMC, if we consider the underlying distribution to be equal to f, and if we search for the minimal KL divergence and disregard the fact that the regroup-refit algorithm obtains only a sub-optimal solution, it is clearly $m^* = k$ and $g^* = f$, implying that a view that is solely informational leads to ambiguity. Our goal is to incorporate the model selection, i.e. to obtain a number of clusters m that yields the best trade-off between computational load and accuracy, when performing the GS task within a real recognition system.

Thus, the actual Gaussian components $\{f_i\}_{i=1}^{k}$ are considered as objects in the k-means based algorithm, where KL is used as the pseudo-distance, as defined in Sect. 2. One of the simplest model selection techniques, which can be applied to any k-means based clustering, but yet efficient in a majority of cases, is the following (see [15, 26]): Start with a smaller number of clusters (for example $m = 2$), perform a k-means based algorithm until it converges, and continue by performing the same algorithm on an increased number of clusters. The values of the cost function will decrease rapidly until the "optimal" number of clusters is approached, and it will significantly decelerate its decrease as the true number of clusters is overreached, so that one can choose the "optimal" number of clusters to be the one that corresponds to the knee-point of the previously described characteristic. The problem is that one must always perform the complete clustering task from the beginning, for all different predefined numbers of clusters, which results in a combinatorial explosion and makes the method practically unfeasible. We propose a version of the S&M Gaussian mixture models clustering algorithm that uses the split and the merge operations given in Sect. 3, but performs iterations of S&M on an increasing number of clusters. It performs model selection "on the fly", i.e., in just one pass of the algorithm.

The proposed method starts with the baseline method given in Sect. 2, where the initial number of clusters is set to $m^{(0)} = 2$, and performs the regroup-refit algorithm given by (10) and (11) until it converges to some $(\hat{\pi}_0^{(0)}, \hat{g}_0^{(0)}) \in P_2 \times MoG(2)$. Next, the split and the merge operations, described in Sects. 3.1 and 3.2, are performed, starting with $(\hat{\pi}_0^{(0)}, \hat{g}_0^{(0)})$. If in a k-th Split-and-Merge step we obtain $(\hat{\pi}^{(0)}_{s,k}, \hat{g}^{(0)}_{s,k})$ and $(\hat{\pi}^{(0)}_{m,k}, \hat{g}^{(0)}_{m,k})$ respectively, where $(\hat{\pi}^{(0)}_{k-1}, \hat{g}^{(0)}_{k-1})$ are the partition and the hyper-Gaussian obtained at the end of the previous, $k-1$-th attempt, then if $d(f, \hat{g}^{(0)}_{m,k}, \hat{\pi}^{(0)}_{m,k}) < d(f, \hat{g}^{(0)}_{k-1}, \hat{\pi}^{(0)}_{k-1})$ holds, $(\hat{\pi}^{(0)}_{m,k}, \hat{g}^{(0)}_{m,k})$ is adopted as the current setting and it is continued with the Split-and-Merge operations. On the contrary, if $d(f, \hat{g}^{(0)}_{m,k}, \hat{\pi}^{(0)}_{m,k}) \geq d(f, \hat{g}^{(0)}_{k-1}, \hat{\pi}^{(0)}_{k-1}), m^{(1)} = 3$ is adopted, and $(\hat{\pi}_0^{(1)}, \hat{g}_0^{(1)}) \leftarrow (\hat{\pi}^{(0)}_{s,k}, \hat{g}^{(0)}_{s,k})$ as well, thus $(\hat{\pi}_0^{(1)}, \hat{g}_0^{(1)}) \in P_3 \times MoG(3)$. The number of clusters $m^{(p)} \leftarrow m^{(p-1)} + 1$ is further increased using the described procedure, until, in a Kth iteration, $d(f, \hat{g}_0^{(K)}, \hat{\pi}_0^{(K)}) \geq d(f, \hat{g}_0^{(K-1)}, \hat{\pi}_0^{(K-1)}) - T$ holds, where $T > 0$ is a predefined threshold. The "optimal" number of clusters is then $m^* = m^{(K)}$, and the final setting $(\pi^*, g^*) = (\hat{\pi}_0^{(K)}, \hat{g}_0^{(K)}) \in P_{m^*} \times MoG(m^*)$ is obtained. It means that the number of clusters will increase until the lowering of the cost function becomes sufficiently insignificant, i.e., below a predefined threshold T. The threshold T controls the trade-off obtained in the actual GS system that uses the proposed method, in the sense that a lower T produces more clusters and thus larger accuracy, but results in a larger computational load.

In the experiments obtained in the GS task on a real CSR system, as it can be seen in Table 4, Sect. 5, using an appropriate threshold T, a better trade-off between recognition accuracy and computational efficiency is obtained, for a different predefined $\theta[\%]$ (percent of hyper-Gaussians for which the underlying Gaussian components are directly evaluated in the GS process, see [9] for basics concerning GS). Moreover, performing S&M operations on an increasing number of clusters, as proposed in the previous considerations, also shows better results in obtaining lower average values of cost function (2), its lower variance, as well as a larger number of improvements observed on almost all artificial test examples, as can be seen in Tables 1, 2 and 3, respectively.

The block diagram of the proposed S&M MS HGMMC method is given below:

S&M MS HGMMC

- Initialization:
  - Start with the initial number of clusters $m^{(0)} = 2$ and perform regroup-refit described in Sect. 2, thus obtaining $(\hat{\pi}_0^{(0)}, \hat{g}_0^{(0)})$.
- Clustering:
  Repeat the following until $d(f, \hat{\pi}_0^{(K)}, \hat{g}_0^{(K)}) \geq d(f, \hat{\pi}_0^{(K-1)}, \hat{g}_0^{(K-1)}) - T$ in a K-th iteration, for a predefined threshold $T > 0$:
  - Perform Split-and-Merge operations: In a k-th Split-and-Merge step, $(\hat{\pi}_{s,k}^{(p)}, \hat{g}_{s,k}^{(p)})$ and $(\hat{\pi}_{m,k}^{(p)}, \hat{g}_{m,k}^{(p)})$ are obtained (p denotes the index of the iteration). If $d(f, \hat{g}_{m,k}^{(p)}, \hat{\pi}_{m,k}^{(p)}) < d(f, \hat{\pi}_{k-1}^{(p)}, \hat{g}_{k-1}^{(p)})$ holds, where $(\hat{\pi}_{k-1}^{(p)}, \hat{g}_{k-1}^{(p)})$ are obtained in the previous, $k-1$-th step, $(\hat{\pi}_k^{(p)}, \hat{g}_k^{(p)}) \leftarrow (\hat{\pi}_{m,k}^{(p)}, \hat{g}_{m,k}^{(p)})$ is set, and if not, a new iteration is performed, and $p \leftarrow p+1, m^{(p+1)} \leftarrow m^{(p)} + 1$ and $(\hat{\pi}_0^{(p+1)}, \hat{g}_0^{(p+1)}) \leftarrow (\hat{\pi}_{s,k}^{(p)}, \hat{g}_{s,k}^{(p)})$ are set.

  Adopt final $(\pi^*, g^*) = (\hat{\pi}_0^{(K)}, g_0^{(K)})$.

5 Experimental results

In this section, experimental results that favor the proposed approach in comparison to the baseline algorithm are presented, supporting our previous considerations. Experiments have been conducted on artificial data (see Sect. 5.1), as well as on a real GS system used within the CSR task (see Sect. 5.2).

5.1 Experiments on artificial data

In the experiments conducted on artificial data, the results of which are given in Tables 1 to 3, for the proposed S&M MS HGMMC and S&M HGMMC, improvements were obtained in terms of lower average cost function and its lower variance in a large percentage of cases, in comparison to the baseline HGMMC, for all tested dimensions, and for both approximations of mixture KL divergence used. Simulations were conducted using two different approximations for $KL(f_j^{\hat{\pi}} \parallel \hat{g}_j)$: the MCSA and the LBVA, defined in Sect. 3.2. A total number of 2800 simulations on artificial data were carried out, i.e. 100 testing examples for each particular combination of the number of dimensions and the number of Gaussian components, for a given algorithm and KL approximation. The field “dimension” in Tables 1–3 refers to the dimension of the feature space of data, i.e. the dimension of centroids of Gaussian components, while the field “No. of Gaussians” refers to the number of Gaussian components in a particular experiment. The results were obtained as follows: S&M MS HGMMC algorithm is executed first, finding the “optimal” number of clusters $m^* = m^{(K)}$ in some K-th iteration (see Sect. 4). The iteration K is obtained as the one in which the decrease in the cost function is lower than 1% in comparison to the value obtained in the previous, $K-1$-th iteration. After the convergence of the model selection algorithm, S&M HGMMC starts with m^* clusters.

Table 1 shows the average improvement in terms of the average cost function (2), i.e., the average distance (KL divergence) between the clusters and their associated Gaussian components, obtained by using the proposed S&M MS HGMMC and S&M HGMMC algorithms, in comparison to the baseline HGMMC algorithm initialized with the same number of clusters. It can be seen in Table 1 that both S&M HGMMC and S&M MS HGMMC obtain better results in terms of lower average cost function, in comparison to the baseline HGMMC. Moreover, as it was mentioned in Sect. 4, performing the S&M operations on an increasing number of clusters reaching m^*, as it is done by S&M MS HGMMC, additionally lowers the average cost in comparison to the S&M HGMMC initialized with the same number of clusters m^*. It can be noted that the lower average of the cost function (2) directly corresponds to the actual recognition accuracy in GS task, on a real recognition system.

Table 2 shows the average reduction in standard deviation of the cost function (2), for S&M MS HGMMC and S&M HGMMC algorithms, in comparison to the baseline HGMMC algorithm; initialized with the same number of clusters. It can be seen that both S&M HGMMC and S&M MS HGMMC give better results in comparison to the baseline HGMMC in terms of the reduction in standard deviation of the cost function (2). Moreover, as it was mentioned in Sect. 4, performing the S&M operations on the increasing number of clusters, reaching some predefined m^*, as it

Table 1 Improvement in terms of the average reduction of the cost function, obtained by using the proposed S&M MS HGMMC and S&M HGMMC algorithms, in comparison to the baseline HGMMC algorithm. Experiments are conducted on artificial data

No. of Gaussians	Dimension	Average improvement [%]			
		LBVA MS	LBVA	MCSA MS	MCSA
100	2	29.84	12.73	32.39	15.58
100	10	21.35	13.25	22.12	14.06
200	2	18.94	13.44	25.83	15.76
200	10	19.67	20.48	24.23	20.62
500	2	17.11	13.12	18.31	14.59
500	10	20.68	18.57	22.73	19.55
500	20	19.73	17.44	23.33	19.05
Total average		21.05	15.58	24.13	17.03

Table 2 Average reduction of the standard deviation of the cost function, for the proposed S&M MS HGMMC and S&M HGMMC algorithms, in comparison to the baseline HGMMC algorithm. Experiments are conducted on artificial data

No. of Gaussians	Dimension	Average reduction in standard deviation [%]			
		LBVA MS	LBVA	MCSA MS	MCSA
100	2	37.71	30.11	63.42	32.35
100	10	12.31	8.03	15.63	9.08
200	2	26.05	21.14	40.04	28.91
200	10	7.13	6.05	21.98	17.36
500	2	39.51	29.92	51.23	45.31
500	10	14.62	10.44	17.71	11.56
500	20	6.12	5.08	9.07	8.32
Total average		20.49	15.82	31.30	21.84

Table 3 The number of simulations in percents (field "percent of improvements") in which a reduction of the cost function is observed, when comparing the proposed S&M MS HGMMC and S&M HGMMC algorithms to the baseline HGMMC algorithm. Experiments are conducted on artificial data

No. of Gaussians	Dimension	Percent of improvements			
		LBVA MS	LBVA	MCSA MS	MCSA
100	2	92	71	98	65
100	10	91	62	96	65
200	2	75	55	84	71
200	10	80	67	89	68
500	2	82	71	86	81
500	10	78	63	78	67
500	20	88	75	95	86
Total average		83.71	66.29	89.43	71.86

is done by S&M MS HGMMC, additionally lowers the average standard deviation of the cost, in comparison to the S&M HGMMC initialized with the same number of clusters m^*. It can be noted that lower standard deviation of the cost function is relevant for the clustering algorithm. This implies that it is less likely for some particular instance of the clustering to produce clusters that lead to the recognition error in GS task being higher than the average recognition error for that particular clustering method.

Table 3 shows the percentage of simulations (field "percent of improvements") in which a reduction of the cost function is observed in the experiments, for S&M MS HGMMC and S&M HGMMC algorithms. The experimental settings for all three experiments presented in Tables 1–3 were identical. Although the proposed S&M approach does not explicitly guarantee lowering of the cost function, it provides information on how frequently the cost function decreases. It can be seen from Table 3 that there is a large percentage of cases where the cost function decreases as a result of the proposed S&M approach, for all experimental settings used.

The average value and standard deviation of the cost function (2) obtained within the experiments on artificial data with a sufficiently wide range of data parameters, such as the dimension of the centroid vectors or the number and diversity of parameters of Gaussian components, give a good prediction of the performance of the proposed methods in recognition and GS tasks within real recognition systems. From the results presented in Tables 1 and 2, it is clear that the proposed S&M HGMMC and S&M MS HGMMC result in lower average values and standard deviation of the cost function (2) in a large majority of cases (see Table 3). A total reduction greater than 15% is achieved for both the average value and the standard deviation of the cost function in comparison to the baseline HGMMC. The following subsection gives an overview of the performance of the proposed S&M concept within a real CSR system.

5.2 Application in the GS task on real CSR system

S&M HGMMC and S&M MS HGMMC, proposed in Sects. 3 and 4 respectively, were applied in the GS task on a real CSR system, and compared to the baseline HGMMC presented in Sect. 2, as well as to the system that does not use a GS approach, which will be referred to as the full system.

For all experiments a CSR HMM based system that incorporates GMMs with full covariance matrices was used, while training was performed using the Tree Based Clustering algorithm (TBC) [27, 28], allowing the parameters to be shared between phonetic models. The system used 26 features, 24 of which describing the spectral envelope (12 static and 12 dynamic mel-frequency cepstral coefficients, i.e., the first time derivatives), and the remaining two of them describing normalized energy and its first time derivative. The speech database used in the experiments was recorded at the Faculty of Technical Sciences, Novi Sad. The database contains utterances from about 1000 different speakers (with approximately equal gender distribution), recorded over the public telephone network. The files were recorded in the A-law format, with the sampling frequency of 8 kHz. The recognition system used 5826 acoustical states with $K = 29558$ Gaussian mixture components altogether.

The performance of the GS method was assessed in terms of both recognition performance and reduction in the number of directly calculated Gaussian components. The recognition performance was measured in terms of standard Word Error Rate (WER). Reduction is expressed in terms of the computational fraction factor (CF) [9], given as

$$CF = \frac{G_{new} + R_{comp}}{G_{full}} \qquad (24)$$

Terms G_{new} and G_{full} are the average numbers of Gaussians calculated per frame in the system that uses GS and the full system respectively, and the term R_{comp} is the number of computations required for the system to calculate log-likelihoods of hyper-mixtures in order to decide whether the Gaussian mixture component that belongs to a particular cluster will be evaluated exactly or not. The term R_{comp} is approximated with $R_{comp} \approx G_{full}/n_{avr}$, where n_{avr} is the average number of baseline Gaussians per cluster, simplifying the previous expression to:

$$CF \approx \frac{G_{new}}{G_{full}} + \frac{1}{n_{avr}} \qquad (25)$$

In the decoding stage (i.e., recognition stage) of GS, for every particular observation vector, the log-likelihoods of all hyper-Gaussians are evaluated. For a predefined percentage $\theta[\%]$ of hyper-Gaussians with the largest log-likelihoods, the underlying Gaussian components that belong to the corresponding clusters are evaluated exactly on a particular observation, while for the remaining components log-likelihoods are approximated with log-likelihoods of corresponding hyper-Gaussians evaluated for that particular observation.

In Table 4, a comparison of the performance of the following methods is presented: GS system that uses the baseline, GS system that uses the S&M HGMMC, GS system that uses the S&M MS HGMMC, as well as the full system. In all experiments, only the LBVA approximation for KL divergence, given by (21), is used, as it is computationally too expensive to use MCSA in a recognition system that uses such a large number of Gaussian components as a CSR system, regardless of the fact that the actual clustering is performed off-line. It can be seen that the S&M HGMMC gives better results in terms of the trade-off between the recognition accuracy and the computational efficiency in comparison to the baseline HGMMC for all values of n_{avr}, and thus for all the predefined numbers of clusters $m = \lfloor K/n_{avr} \rfloor$ that are tested. It can also be seen that using the S&M MS HGMMC, with the appropriate choice of the threshold T (obtained empirically as $T = 2.3$), a better trade-off between the recognition accuracy (WER) and the computational load (CF) in comparison to the other methods was obtained, for all values of θ used in the experiments. The experimental results presented in Table 4 show that the proposed S&M HGMMC and S&M MS HGMMC perform better than the baseline HGMMC in a GS task within a real CSR system. Moreover, the efficiency of model selection procedure, introduced in the S&M framework in Sect. 4, is also experimentally confirmed.

It should be pointed out that the execution times for the proposed S&M HGMMC and S&M MS HGMMC are significantly larger than for the baseline HGMMC. However, it should be noted that this does not affect the application of the proposed algorithms, as both these algorithms are used for off-line learning and the speed of actual recognition does not depend on their computational load in any way.

Table 4 Comparison of the proposed S&M HGMMC and S&M MS HGMMC with the full and the baseline HGMMC systems, in terms of WER and *CF* factor, applied to the GS task within a real CSR system

Selection scheme	n_{avr}	T	m	$\theta[\%]$	WER[%]	*CF*
Full	–	–	–	–	2.12	1.0
HGMMC	50	–	591	40	2.31	0.42
	100	–	295	40	2.34	0.43
	150	–	197	40	2.39	0.43
	50	–	591	30	2.77	0.32
	100	–	295	30	2.81	0.33
	150	–	197	30	2.93	0.34
	50	–	591	20	3.37	0.22
	100	–	295	20	3.42	0.23
	150	–	197	20	3.63	0.23
S&M HGMMC	50	–	591	40	2.25	0.41
	100	–	295	40	2.28	0.41
	150	–	197	40	2.32	0.42
	50	–	591	30	2.57	0.32
	100	–	295	30	2.72	0.32
	150	–	197	30	2.79	0.33
	50	–	591	20	3.15	0.21
	100	–	295	20	3.19	0.22
	150	–	197	20	3.23	0.23
S&M MS HGMMC	–	2.3	583	40	2.18	0.43
	–	2.3	583	30	2.43	0.32
	–	2.3	583	20	3.08	0.22

6 Conclusions

The research presented in this paper concerns a novel Split-and-Merge algorithm for Hierarchical Gaussian Mixture Models Clustering, which tends to improve on the local optimal solution determined by the initial constellation. The algorithm is initialized by local optimal parameters obtained by using a baseline approach similar to k-means. It tends to approach the global optimum of the target clustering function more closely, by iteratively splitting and merging the clusters of Gaussian components obtained as the output of the baseline algorithm. In order to obtain a better trade-off between recognition accuracy and computational load in a GS task applied within an actual recognition process, model selection is also incorporated. The proposed method was tested on artificial data, as well as in the framework of a GS task performed within an actual CSR system. Better results in comparison to the baseline approach presented in [7] were obtained.

Acknowledgements This research work has been supported by the Serbian Ministry of Education and Science, and it has been realized as a part of "Development of Dialogue Systems for Serbian and Other South Slavic Languages" research project (id TR 32035).

References

1. Wang J (2007) Discriminative Gaussian mixtures for interactive image segmentation. In: Proc ICASSP, Honolulu, HI, vol 1, pp I-601–I-604. doi:10.1109/ICASSP.2007.365979
2. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286. doi:10.1109/5.18626
3. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 3(1):72–83. doi:10.1109/89.365379
4. Shin KS, Jeong Y-S, Jeong MK (2011) A two-leveled symbiotic evolutionary algorithm for clustering problems. Appl Intel (published online 08 July 2011), doi:10.1007/s10489-011-0295-y
5. Bahrampour S, Moshiri B, Salahshoor K (2011) Weighted and constrained possibilistic C-means clustering for online fault detection and isolation. Appl Intell 35(2):269–284. doi:10.1007/s10489-010-0219-2
6. Korkmaz EE (2010) Multi-objective genetic algorithms for grouping problems. Appl Intell 33(2):179–192. doi:10.1007/s10489-008-0158-3
7. Goldberger J, Roweis S (2005) Hierarchical clustering of a mixture model. Adv Neural Inf Process Syst 17:505–512
8. Bocchieri E (1993) Vector quantization for efficient computation of continuous density likelihoods. In: Proc ICASSP, Minneapolis, MN, vol 2, pp II-692–II-695. doi:10.1109/ICASSP.1993.319405
9. Knill KM, Gales MJF, Young SJ (1996) Use of Gaussian selection in large vocabulary continuous speech recognition us-

ing HMMs. In: Proc ICSLP, vol 1, pp 470–473. doi:10.1109/ICSLP.1996.607156

10. Simonin J, Delphin L, Damnati G (1998) Gaussian density tree structure in a multi-Gaussian HMM based speech recognition system. In: 5-th Int Conf Spok Lang Process, Sidney, Australia
11. Watanabe T, Shinoda K, Takagi K, Iso K-I (1995) High speed speech recognition using tree-structured probability density function. In: Proc ICASSP, vol 1, pp 556–559. doi:10.1109/ICASSP.1995.479658
12. Marko J, Pekar D, Jakovljevic N, Delic V (2010) Eigenvalues driven Gaussian selection in continuous speech recognition using HMM's with full covariance matrices. Appl Intell 33(2):107–116. doi:10.1007/s10489-008-0152-9
13. Shinoda K, Lee C-H (2001) A structural Bayes approach to speaker adaptation. IEEE Trans Speech Audio Process 9(3):276–287. doi:10.1109/89.906001
14. Linde Y, Buzo A, Gray R (1980) An algorithm for vector quantizer design. IEEE Trans Commun 26(1):84–95. doi:10.1109/TCOM.1980.1094577
15. McCrosky J (2008) A new measure for clustering model selection. Master thesis, University of Waterloo, Waterloo, Ontario, Canada
16. Axelrod S, Goel V, Gopinaht RA, Olsen PA, Visweswariah K (2005) Subspace constrained Gaussian mixture models for speech recognition. IEEE Trans Speech Audio Process 13(6):1144–1160. doi:10.1109/TSA.2005.851965
17. Dharanipragada S, Visweswariah K (2006) Gaussian mixture models with covariances or precisions in shared multiple subspaces. IEEE Trans Audio Speech Lang Process 14(4):1255–1266. doi:10.1109/TSA.2005.860835
18. Olsen PA, Gopinaht RA (2004) Modeling inverse covariance matrices by basis expansion. IEEE Trans Speech Audio Process 12(1):37–46. doi:10.1109/TSA.2003.819943
19. Sun J, Kaban A (2008) A fast algorithm for robust mixtures in the presence of measurements errors. IEEE Trans Neural Netw 21(8):1206–1220. doi:10.1109/TNN.2010.2048219
20. Verbeek JJ, Nunnink JRJ, Vlassis N (2006) Accelerated EM-based clustering of large data sets. Data Min Knowl Disc 13:291–307. doi:10.1007/s10618-005-0033-3
21. Moore AW (1999) A very fast EM-based mixture model clustering using multiresolution kd-trees. In: Adv Neural Inf Process Syst, vol 11. MIT Press, Cambridge, pp 543–549. ISBN: 0-262-11245-0
22. Hershey JR, Olsen PA (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. In: Proc ICASSP, Honolulu, HI, vol 4, pp IV-317–IV-320. doi:10.1109/ICASSP.2007.366913
23. Zhang Z, Chen C, Sun J, Chan KL (2003) EM algorithms for Gaussian mixtures with split-and-merge operation. Pattern Recognit 36(9):1973–1983. doi:10.1016/S0031-3203(03)00059-1
24. Ueda N, Nakano R, Ghahramani Z, Hinton GE (2000) Split and merge EM algorithm for improving Gaussian mixture density estimates. J VLSI Signal Process Syst Signal Image Video Technol 26(1/2):133–140. doi:10.1023/A:1008155703044
25. Delic V (2007) A review of R&D of speech technologies in Serbian and their applications in western Balkan countries. Keynote lecture at 12th SPECOM (Speech and Computer), Moscow, Russia, pp 64–83
26. Webb AR (1999) Statistical Pattern Recognition. Defence Evaluation and Research Agency, Arnold, UK
27. Kannan A, Ostendorf N, Rohlicek JR (1994) Maximum likelihood clustering of Gaussian mixtures for speech recognition. IEEE Trans Speech Audio Process 2(3):453–455. doi:10.1109/89.294362
28. Young SJ, Odell JJ, Woodland PC (1994) Tree-based state tying for high accuracy state modeling. In: Proc ARPA Workshop Hum Lang Technol, pp 307–312. doi:10.3115/1075812.1075885

Branislav Popović received his M.Sc. degree in electrical and computer engineering from the Faculty of technical sciences, Novi Sad, Serbia. He is currently pursuing the Ph.D. degree in electrical and computer engineering. His current research interests include multimodal human-computer interaction, speech and image processing, speech recognition and speech synthesis.

Marko Janev received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 2003. He received his M.Sc. and Ph.D. degree in Applied Mathematics from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 2009 and 2011, respectively. Currently he is a researcher at the Mathematical institute of Serbian academy of sciences and arts. His research interests include: statistical pattern recognition, speech recognition, speech processing, image processing and fractional calculus.

Darko Pekar received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 1998. He has been leading speech technologies R&D team since 2003. as CEO of AlfaNum company, and associate researcher at FTN. Currently he is still CEO of the AlfaNum company and a researcher at the FTN. His research interests include speech recognition, speech processing, speech synthesis and adaptive dialogue management in human-machine interaction.

Nikša Jakovljević received the B.Sc. and M.Sc. degree in electrical engineering from the Faculty of technical sciences, Novi Sad, Serbia. He is currently pursuing the Ph.D. degree in electrical engineering. His current research interests include speech and audio processing, speech recognition, machine learning.

Milan Gnjatović received the Ph.D. degree in computer science from Otto-von-Guericke University Magdeburg, Germany, in 2009. Currently, he works as a postdoctoral researcher at the Department of Power, Electronics, and Communications Engineering at the University of Novi Sad, Serbia. He has also worked as a research assistant at the Department of Knowledge Processing and Language Engineering at Otto-von-Guericke University Magdeburg. His research interests include adaptive dialogue management in human-machine interaction, cognitive technical systems, affective computing, and natural language processing.

Milan Sečujski received his M.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN) in Novi Sad, Serbia, in 2002. He received his Ph.D. degree in electrical engineering from FTN in 2009. Currently, he is assistant professor at FTN and one of the key researchers at the "Human Computer Interaction" research team. His research interests include natural language processing, speech recognition, speech synthesis and acoustics.

Vlado Delić received his M.Sc. degree in electrical engineering from the School of Electrical Engineering in Belgrade, Serbia, in 1993. He received his Ph.D. degree in electrical engineering from the Faculty of Technical Sciences (FTN) Novi Sad in 1997. Currently, he is associate professor at the Faculty of Technical Sciences (FTN) Novi Sad and the leader of the "Human Computer Interaction" research team. His research interests include statistical pattern recognition, speech recognition, speech processing and speech synthesis and acoustics.