Appl Intell (2010) 33: 107–116
DOI 10.1007/s10489-008-0152-9

Eigenvalues Driven Gaussian Selection in continuous speech recognition using HMMs with full covariance matrices

Marko Janev · Darko Pekar · Niksa Jakovljevic · Vlado Delic

Published online: 3 December 2008

Abstract In this paper a novel algorithm for Gaussian Selection (GS) of mixtures used in a continuous speech recognition system is presented. The system is based on hidden Markov models (HMM), using Gaussian mixtures with full covariance matrices as output distributions. The purpose of Gaussian selection is to increase the speed of a speech recognition system, without degrading the recognition accuracy. The basic idea is to form hyper-mixtures by clustering close mixtures into a single group by means of Vector Quantization (VQ) and assigning it unique Gaussian parameters for estimation. In the decoding process only those hyper-mixtures which are above a designated threshold are selected, and only mixtures belonging to them are evaluated, improving computational efficiency. There is no problem with the clustering and evaluation if overlaps between the mixtures are small, and their variances are of the same range. However, in real case, there are numerous models which do not fit this profile. A Gaussian selection scheme proposed in this paper addresses this problem. For that purpose, beside the clustering algorithm, it also incorporates an algorithm for mixture grouping. The particular mixture is assigned to a group from the predefined set of groups, based on a value aggregated from eigenvalues of the covariance matrix of that mixture using Ordered Weighted Averaging operators (OWA). After the grouping of mixtures is carried out, Gaussian mixture clustering is performed on each group separately.

Keywords Gaussian mixtures · Vector quantization · Gaussian selection · Eigenvalues · Ordered weighted averaging operators · Constrained optimization

M. Janev (✉) · N. Jakovljevic · V. Delic
Faculty of Technical Sciences Novi Sad, University of Novi Sad, Novi Sad, Serbia
e-mail: markojan@uns.ns.ac.yu

N. Jakovljevic
e-mail: jakovnik@uns.ns.ac.yu

V. Delic
e-mail: vdelic@uns.ns.ac.yu

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

1 Introduction

Hidden Markov model-based (HMM-based) continuous speech recognition (CSR) systems tend to operate several times slower than real-time, which is not practical for most applications. Techniques are therefore required that would reduce the decoding time to at least real-time, while retaining or staying close to the same level of accuracy. To obtain a high level of accuracy, HMM-based CSR systems typically use continuous densities. In such systems, calculation of state likelihoods makes up a significant proportion (between 30–70%) of the computational load [8]. This is a result of the need to use multiple mixture Gaussian output distributions in a state, and each Gaussian component must be separately evaluated in order to determine the overall likelihood. A wide variety of techniques may be used to reduce the amount of computation required. Some of them involve alteration of acoustic feature vector from the full system (Linear Discriminant Analysis) [7, 16], and others involve "tying" of acoustical states (semi-continuous HMM systems) [6]. An alternative approach is to use Gaussian Selection (GS) methods [3, 10] that reduce likelihood computation time by only computing the likelihood of a selected subset of mixtures used for a particular input vector.

Most of the GS methods in CSR and Speaker Verification tasks were designed for selection of diagonal covariance matrices [1, 3, 10, 13–15, 21]. Nevertheless, the generalization to the full covariance case is straightforward. Bocchieri in [3] firstly introduced Vector Quantized Gaussian Selection. The idea was to generate a set of codewords (clusters). Each codeword is then assigned a shortlist of Gaussians. For clustering of Gaussians, weighted Euclidean distance was used. Knill and Gales in [10] modified the proposed algorithm in order for the distance to take into account the variance of a particular mixture to be clustered. In [15] Watanabe and Shinoda proposed the use of a tree structured probability density function in modeling of the acoustical space applied on Gaussian mixtures. In [14] Simonin and Delphin did something similar. The Gaussian tree was built from successive Gaussian mixture merging. Each node of the tree was associated with a Gaussian hyper-mixture, and the actual HMM densities are associated to the leaves. In general, they proposed building of a tree structure for partitioning of the acoustical space, but the algorithms were applied to the diagonal Gaussian mixtures. In application to the tied state full covariances in the system used in this work, only one level of hierarchy was acceptable from the aspect of recognition accuracy. In the one level case, it amounts to a VQ technique, i.e. generalization of a method proposed by Bocchieri. That scheme is referred to as Vector Quantization Gaussian Selection (VQGS) and it is used as a baseline approach. The basic idea is to form hyper-mixtures by putting close mixtures into a single cluster and assigning unique Gaussian parameters to it. In the decoding process, after the evaluation of all hyper-mixtures for a particular input acoustic vector, only the hyper-mixtures which are above a predefined threshold are selected. The mixtures belonging to them are evaluated and the rest of the mixtures are approximated with some sufficiently general value.

A problem emerges if there are significant overlaps between the mixtures to be clustered. In CSR systems such situations are inevitable, and considerable overlapping is always present. In this paper, a GS scheme is proposed that takes into account the most significant eigenvalues of covariance matrices in the clustering process. In this way the previously mentioned problem is addressed, as will be explained later. The main idea is to place mixtures in several groups based on the eigenvalues of their covariance matrices. The goal is for every group to contain only those mixtures with the most of their significantly high eigenvalues in some predefined range. After the grouping of the mixtures is applied, baseline VQ based Gaussian mixture clustering is performed on every group of mixtures separately. Owing to this, in any particular cluster it is likely that a large majority of mixtures will have either dominantly larger or dominantly smaller eigenvalues of their covariance matrices. The frequency of clusters that contain both kinds of mixtures is thus severely reduced. For a particular input vector, if such a grouping of mixtures is performed, it is not probable that a hyper-mixture would have high likelihood and at the same time that the corresponding cluster would contain a significant number of mixtures that have low likelihoods. Such a situation would degrade system performance regarding computational time required for decoding and recognition accuracy.

For aggregation of values on the base on which the grouping of mixtures prior to VQ is done, use of Ordered Weighted Average Operators (OWA) is proposed. The idea is to develop an aggregation method that takes into account the most significant (the largest) eigenvalues when deciding to which of the groups to assign a particular mixture.

In Sect. 2, VQGS selection scheme is presented. It is used to provide a performance baseline.

Section 3 gives the description of the proposed novel GS scheme, with emphasis on the way the eigenvalues are combined, in order to decide to which group to assign a particular mixture. Mixture grouping is performed before VQ Gaussian clustering, which is applied on each group of mixtures separately.

Section 4 presents experimental results that favor the proposed method in comparison with VQ Gaussian selection algorithm that does not take into account mixture covariance eigenvalues.

2 VQ-based Gaussian Selection

In the original implementation of the GS by Bocchieri [3], during the training phase the acoustical space is divided up into a set of vector quantized (VQ) regions. Each Gaussian component (mixture) is then assigned to one or more VQ codewords (VQ Gaussian mixture clustering). During the recognition phase, the input feature vector is vector quantized, i.e. the vector is mapped to a single VQ codeword. The likelihood of each Gaussian component in this codeword shortlist is computed exactly, whereas for the remaining Gaussian components the likelihood is approximated. In the works of Shinoda [15] and Simonin [14], acoustical space was modeled using a tree structured probability density function, with application on Gaussian mixtures. The similarity measure (distance) used in mixture clustering was KL divergence, and the parameters of hyper-mixtures were estimated using ML estimates. The method was applied to diagonal Gaussian mixtures. In the full covariance case, multilevel tree structure causes degradation in recognition accuracy. It is especially the case when state tying is applied to the HMM system, as it was done in this paper. In that case the number of mixtures is already reduced, and with rotation capability included, the coverage of the acoustical space is much better than in the case of diagonal mixtures without state tying. Actually, with two or more hierarchy levels used,

Word Error Rate (WER [%]) was larger than 8% in comparison with 5.66% of the full system. For those reasons, for the purpose of this work as a baseline method, the previous idea is applied with just one level of hyper-mixtures, i.e. mixtures are vector quantized into codewords (clusters). Every cluster corresponds to a unique hyper-mixture. This will be referred to in the following text as VQGS selection scheme, and it will be used to provide a performance baseline.

The distance between the m-th Gaussian mixture $\vartheta_m = \vartheta(\mu_m, \Sigma_m)$ and the current f-th Gaussian hyper-mixture $\vartheta_f = \vartheta(c_f, \Sigma_f)$ (in a particular iteration of the algorithm) used in the baseline iterative clustering algorithm is defined as the symmetric KL distance [13] from ϑ_m to ϑ_f:

$$\begin{aligned} d(\vartheta_m, \vartheta_f) = {} & (c_f - \mu_m)^T \Sigma_f^{-1} (c_f - \mu_m) \\ & + (c_f - \mu_m)^T \Sigma_m^{-1} (c_f - \mu_m) + \mathrm{Tr}(\Sigma_f^{-1} \Sigma_m) \\ & + \mathrm{Tr}(\Sigma_m^{-1} \Sigma_f) \end{aligned} \tag{1}$$

Terms μ_m and c_f represent the centroids of the m-th Gaussian mixture and the f-th hyper-mixture respectively, while Σ_m and Σ_f represent corresponding covariance matrices.

2.1 Evaluation of hipper-mixture parameters

The centroid and the covariance matrix of the particular hyper-mixture are used in the distance measure (1) and in the selection process. In the baseline VQGS scheme, they are estimated under the assumption that the hyper-mixture is Gaussian, as proposed in [5, 9, 13, 14]. In that case, ML estimation of the corresponding parameters of a hyper-mixture (centroid c_f and covariance matrix Σ_f) can be expressed as a function of ML estimated parameters of the belonging mixtures only. ML estimates $\hat{c}_f$, $\hat{\Sigma}_f$ of the parameters μ_f, Σ_f can be expressed as:

$$\hat{c}_f = \sum_{m=1}^{M_f} w_m \hat{\mu}_m \tag{2}$$

$$\hat{\Sigma}_f = W_f + \sum_{\substack{m,p=1 \\ m \neq p}}^{M_f} w_m w_p (\hat{\mu}_m - \hat{\mu}_p)(\hat{\mu}_m - \hat{\mu}_p)^T \tag{3}$$

$$W_f = \sum_{m=1}^{M_f} w_m \hat{\Sigma}_m \tag{4}$$

The term W_f is the pooled covariance matrix of the f-th cluster, and w_m is mixture cluster occupancy (different from the state mixture occupancy obtained in the training process) of the m-th mixture. Terms $\hat{\mu}_m$, $\hat{\Sigma}_m$, $m \in \{1, \ldots, M_f\}$ represent ML estimates of centroids and covariance matrices of mixtures belonging to the f-th cluster, evaluated in the training process.

Mixture cluster occupancy w_m for some m-th mixture for $m \in \{1, \ldots, M_f\}$ can be calculated exactly as $w_m = n_m/n$, where n_m is the actual number of observations belonging to the m-th mixture obtained at the end of the training process, while n is the overall number of observations that corresponds to the f-th cluster, i.e. $n = n_1 + \cdots + n_{M_f}$. In this paper, as in [13], the assumption is made that mixture cluster occupancies are equal for all mixtures belonging to the f-th cluster. In the VQ process the accent is on covering of acoustical space with Voronoi regions, so that approximation does not degrade GS system performance. All occupancies w_m are thus set to $w_m = 1/M_f$ for all $m \in \{1, \ldots, M_f\}$.

2.2 Gaussian mixture clustering procedure

Gaussian mixture clustering in the terms of Vector Quantization is basically a process of making the codebook. Codevectors corresponding to hyper-mixtures are obtained for every cluster.

It is an iterative process. In each iteration, particular mixture is associated to a cluster iff the distance between the mixture and the hyper-mixture corresponding to that cluster defined by (1) is minimal. It is based on a Linde-Buzo-Gray algorithm [10, 11], which performs clustering in order to minimize the average (per Gaussian component) distortion $D_{average}$, defined as:

$$D_{average} = \frac{1}{M} \sum_{m=1}^{M} \left\{ \min_{f=1}^{|X|} d(\vartheta_m, \vartheta_f) \right\} \tag{5}$$

The term M is the overall number of mixtures to be clustered, X is the set of all hyper-mixtures (clusters), $|X|$ is a predefined number of clusters, and $d(\cdot, \cdot)$ is the mixture distance as defined in (1). The number of clusters is heuristically obtained as $|X| = M/n_{average}$, where $n_{average}$ is a predefined average number of mixtures per cluster. Initially, $|X|$ hyper-mixtures are initialized as Gaussian mixtures with centroids randomly picked from the set of centroids of all mixtures to be clustered, and with covariance matrices set to the identity matrix I.

2.3 Hyper-mixture selection as the second part of GS procedure

In the recognition process, for each hyper-mixture $\chi_f \in X$, log-likelihood for a particular acoustic input vector $x \in R^p$ is evaluated using the following equation:

$$\begin{aligned} \ln f(x|\chi_f) = {} & -\frac{1}{2} \ln(\det(\hat{\Sigma}_f)) \\ & - \frac{1}{2} (x - \hat{c}_f)^T \Sigma_f^{-1} (x - \hat{c}_f) \end{aligned} \tag{6}$$

The term $\ln f(\cdot|\chi_f)$ is the log-likelihood of the hyper-mixture, where $\hat{c}_f$ and $\hat{\Sigma}_f$ are expressed by (2) and (3). The actual occupancies of the hyper-mixtures are set to be equal as in [13], and for that reason they do not appear in (6). The constant $-(p/2)\ln(2\pi)$ is ignored because it does not play any role in the selection.

If, for a particular hyper-mixture $\chi \in X$ with parameter estimates $\hat{c}_\chi$ and $\hat{\Sigma}_\chi$, log-likelihood $\ln f(x|\chi)$ evaluated on input vector x is above the predefined threshold $\theta \in R$, all mixtures belonging to χ are evaluated for that input vector. Mixture occupancies obtained from the training process are this time taken into account, thus full mixture likelihoods are evaluated. If, on the other hand, the log-likelihood $\ln f(x|\chi)$ is below the threshold, mixtures belonging to χ are not evaluated for x, and the log-likelihood for all of them is set to a constant value within the given cluster. Since in [1, 3, 10, 13–15, 21] it is not clearly explained how to "floor" these likelihoods, for this work the value is set to the log-likelihood of a Gaussian mixture with centroid $\hat{c}_\chi$ and the covariance matrix equal to the pooled matrix W_χ evaluated on x.

2.4 Gaussian selection performance

The performance of a Gaussian selection procedure is assessed in terms of both recognition performance and reduction in the number of Gaussian components calculated. Reduction is described by the Computation Fraction CF, given as:

$$CF = \frac{G_{new} + R_{comp}}{G_{full}} \tag{7}$$

Terms G_{new} and G_{full} are the average number of Gaussians calculated per frame in the VQGS and the full system respectively, and R_{comp} is the number of computations required for the system to calculate log-likelihoods of hyper-mixtures in order to decide whether the mixtures belonging to that cluster will be evaluated or not. The fact that the term R_{comp} can be approximated as $R_{comp} \approx G_{full}/n_{average}$ implies that CF can be approximated as:

$$CF \approx \frac{G_{new}}{G_{full}} + \frac{1}{n_{average}} \tag{8}$$

3 Eigenvalues Driven Gaussian Selection

The main idea presented in this paper, is for the Gaussian selection process proposed in Sect. 2, to be driven by the eigenvalues of covariance matrices of the Gaussian mixtures to be selected. The method is proposed to deal with situations when there is a significant overlapping between mixtures, which is a common situation in the CSR systems. It will be referred to as Eigenvalues Driven Gaussian Selection (EDGS). Actually, the Gaussian mixtures used in an HMM system are to be grouped on the basis of their eigenvalues into several groups, before the actual baseline VQ clustering described in the Sect. 2 is performed on each group separately.

The need for grouping of mixtures before their actual clustering rises from the following consideration. If the baseline VQ Gaussian mixture clustering described in Sect. 2 is performed on the whole set of mixtures used, then at the end of the procedure, in some cluster, there could be both mixtures for which the eigenvalues of covariance matrices are predominantly large, and those for which the eigenvalues of covariance matrices are predominantly small. This is especially the case if the degree of mixture overlapping is high, because many low-variance mixtures could be masked by high-variance ones and thus assigned to the same cluster. This comes as a consequence of the use of clustering distance (1), more precisely, its $(c_f - \mu_m)^T \Sigma_f^{-1}(c_f - \mu_m)$ component.

As a result, the covariance matrix of the Gaussian hyper-mixture that corresponds to a cluster can have predominantly large eigenvalues, although there are many mixtures belonging to that cluster with predominantly small eigenvalues of covariance matrices. Let us suppose that such a situation has actually occurred. The hyper-mixture covariance matrix can be decomposed as: $\Sigma_f = V_f \Lambda_f V_f^T$, and the belonging mixtures covariance matrices as: $\Sigma_k = U_k \Psi_k U_k^T$, for $k \in Cluster_f$. Matrices V_f and U_k are unitary, while Λ_f and Λ_k are diagonal matrices of eigenvalues. In the recognition process, for some input acoustic vector $x \in R^p$, let $x^{(f)} = V_f^T x$ and $x^{(k)} = V_k^T x$. Log-likelihood evaluated for x on ϑ_k, and used in recognition is:

$$\begin{aligned}
&\ln f(x|\vartheta_k) \\
&= -\frac{1}{2}\ln(\det(\Psi_k)) - (x^{(k)} - \mu_k^{(k)})^T \Psi_k^{-1}(x^{(k)} - \mu_k^{(k)}) \\
&\quad + \ln(w_k) \\
&= -\frac{1}{2}\sum_{j=1}^{p} \ln \psi_j^{(k)} - \sum_{j=1}^{p} \frac{(x_j^{(k)} - \mu_{k,j}^{(k)})^2}{\psi_j^{(k)}} + \ln(w_k)
\end{aligned} \tag{9}$$

In recognition i.e. decoding process, if the mixture ϑ_k has predominantly small eigenvalues, $\ln f(x|\vartheta_k)$ is often much smaller then the log-likelihood:

$$\begin{aligned}
&\ln f(x|\chi_f) \\
&= \frac{1}{2}\ln(\det(\Lambda_f)) + (x^{(f)} - c_f^{(f)})^T \Lambda_f^{-1}(x^{(f)} - c_f^{(f)}) \\
&= -\frac{1}{2}\sum_{j=1}^{p} \ln \lambda_j^{(f)} - \sum_{j=1}^{p} \frac{(x_j^{(f)} - c_{f,j}^{(f)})^2}{\lambda_j^{(f)}}
\end{aligned} \tag{10}$$

evaluated for the hyper-mixture with predominantly large eigenvalues. Actually, for small eigenvalues $\psi_j^{(k)}$in (9), the term $1/\psi_j^{(k)}$ takes advantage over $\ln\psi_j^{(k)}$, and becomes dominant when $\psi_j^{(k)} \to 0_+$. In that case, the negative components $-(x_j^{(k)} - \mu_{k,j}^{(k)})^2/\psi_j^{(k)}$ in (9) take advantage over the positive $-0.5\ln\psi_j^{(k)}$, and the likelihood is low, unless there is a significant number of projections $\mathrm{pr}_{v_j}(x-\mu_k) = x_j^{(k)} - \mu_{k,j}^{(k)}$ that are much smaller then the corresponding $\psi_j^{(k)}$. That exception is not very likely because that region is the union of at most p very narrow hyper-stripes around principal axes (corresponding to v_j), and it makes a small proportion of R^p. At the other hand, for the large eigenvalues $\lambda_j^{(f)}$ in (10), the influence of negative components $-(x_j^{(f)} - c_{f,j}^{(f)})^2/\lambda_j^{(f)}$ in (10) is small, unless there is a significant number of projections $\mathrm{pr}_{u_j}(x-c_f) = x_j^{(f)} - c_{f,j}^{(f)}$ that are much larger than the corresponding $\lambda_j^{(f)}$. One can, therefore, conclude that the distance $\|x-c_f\|_{R^p}$ is large. On the other hand, since $\|x-\mu_k\|_{R^p} - \|x-c_f\|_{R^p} \le \|\mu_k - c_f\|_{R^p}$ holds (triangle inequality), it follows that if $\|\mu_k - c_f\|_{R^p}$ is sufficiently small, the distance $\|x-\mu_k\|_{R^p}$ is also large. That particular situation is not of interest, because it means that the input vector x is far away from both mixture ϑ_k and hyper-mixture χ_f. Actually, critical situations are those where x is close enough to the hyper-mixture. The constraint that $\|\mu_k - c_f\|_{R^p}$ should be sufficiently small can be met by keeping the average number of mixtures per cluster n_{avr} reasonably small, so that a single cluster does not cover a very significant part of the acoustical space.

Keeping n_{avr} reasonably small is the condition for the EDGS scheme to retain its advantage over the VQGS scheme. Nevertheless, the similar constraint must also be met in order to obtain satisfactory recognition accuracy of any GS system. In Sect. 4, these conclusions are confirmed by simulations on real data.

As a result of situations when low-variance ("narrow") mixtures are masked by high-variance ("wide") ones, in the decoding process the following can happen. If the likelihood of a hyper-mixture evaluated on the input vector is above the predefined threshold, all mixtures in the cluster will be evaluated for that particular input vector. The evaluation will include even those mixtures with low likelihood values, that should have been excluded from the evaluation in order to obtain a sufficiently low *CF* and at the same time not to change WER [%] significantly. The result is the increase in both *CF* and WER [%].

Due to previous considerations, the most significant eigenvalues of mixture covariance matrices are combined in order to group the mixtures used in the HMM system into several groups, prior to the execution of the baseline clustering algorithm (Sect. 2) on each group. The largest eigenvalues are the most important for mixture grouping and their relative importance decreases with their value. For that reason, the eigenvalues could, for instance, be combined as follows. The simple arithmetical mean of the largest L eigenvalues of mixture covariance matrices, for $L \le p$, is taken, and the operator that extracts the value used for grouping is defined as:

$$A(\lambda) = A(\lambda_1,\ldots,\lambda_p) = \frac{1}{L}\sum_{j=1}^{L}\lambda_{\sigma(j)},$$
$$\text{for } \lambda_{\sigma(1)} \ge \cdots \ge \lambda_{\sigma(p)} > 0 \tag{11}$$

The mixture with eigenvalues $\lambda = (\lambda_1,\ldots,\lambda_p)$ is assigned to a particular g-th group where $g \in \{1,\ldots,G\}$, iff $A(\lambda)$ is in a corresponding predefined interval $[\tau_{\min}^{(g)}, \tau_{\max}^{(g)})$. The intervals $[\tau_{\min}^{(g)}, \tau_{\max}^{(g)})$ form the partition of $R \cup \{0\}$ and satisfy the conditions: $\tau_{\max}^{(g)} = \tau_{\min}^{(g+1)}$ for all $g \in \{1,\ldots,G\}$ and $\tau_{\max}^{G} = \infty$. The border values are set heuristically, and the principle adopted was to set: $\tau_{\min}^{(g+1)} = 2\tau_{\min}^{(g)}$.

Instead of (11), more general Ordered Weighted Average (OWA) aggregation operators [17–20] are used. The idea is to give more weight to more significant (larger) eigenvalues in the aggregation process. In that manner the OWA weights are optimized. They are to be applied to $\lambda = (\lambda_1,\ldots,\lambda_p)$ in the following way:

$$\mathrm{OWA}_\omega(\lambda_1,\ldots,\lambda_p) = \sum_{j=1}^{p}\omega_j\lambda_{\sigma(j)},$$
$$\text{for } 0 < \lambda_{\sigma(1)} \le \cdots \le \lambda_{\sigma(p)}, \tag{12}$$

in order to aggregate the value on the basis of which the mixture should be assigned to a particular group, if the value is in the corresponding interval.

The coefficients $\omega \in R^p$ satisfy the following constraints:

$$0 \le \omega_j \le 1, \quad \sum_{j=1}^{p}\omega_j = 1$$

The OWA operators provide a parameterized family of aggregation operators which include many of the well-known operators such as the maximum, the minimum, k-order statistics, median and the arithmetic mean. They can be seen as a parameterized way to interpolate between the minimum and the maximum value in an aggregation process. It is clear that in this particular application, the applied operator should be somewhat closer to $\max(\cdot)$ in order to favor more significant eigenvalues in the grouping process. For that reason, the method to optimally obtain OWA coefficients introduced by Yager [17] and used by O'Hagan in [12] is used in this paper. The idea is for predefined maxness $M(\omega) = \alpha \in [0,1]$ of the OWA operator [17]

defined as:

$$M(\omega) = \sum_{j=1}^{p} \omega_{p-j+1} \frac{p-j}{p-1} = \sum_{j=1}^{p} \omega_j \frac{j-1}{p-1}$$

to maximize the dispersion of weights $D(\omega)$ defined as [12, 20]:

$$D(\omega) = -\sum_{j=1}^{p} \omega_j \ln(\omega_j)$$

Thus, a Constrained Nonlinear Programming (CNP) problem [12] is obtained. For finding the optimal weights ω_{opt}, any standard CNP method can be used (for example SQP as in this work) [2, 4].

In the experiments presented in Sect. 4, it is shown that EDGS selection scheme with usage of aggregation operators (12) gives better results from the point of view of both recognition accuracy and the *CF* factor obtained. In the application of OWA operators with weights obtained by the previously described method, a high degree of predefined maxness α was required.

Block diagrams of the baseline VQGS and the proposed EDGS scheme can be described by their pseudo-code as follows:

VQGS:

– Initialization:
  - For predefined n_{avr} and the overall number of mixtures M, calculate the number of clusters as: $N_{hypc} = |X| = \lfloor M/n_{avr} \rfloor$.
  - Pick up at random (uniform distribution) N_{hypc} different centroids c_f, $f \in \{1, \ldots, N_{hypc}\}$ from the set of overall M mixture centroids used. Assign to every centroid the identity covariance matrix $\Sigma_f = I$. Let Gaussian densities $X^{(0)} = \{\chi_f(c_f, \Sigma_f) : f = 1, \ldots, N_{hypc}\}$ be initial hyper-mixtures.
– Clustering:

  Do the following, for predefined $\varepsilon > 0$:
  - To all mixtures ϑ_j, $j = 1, \ldots, M$ assign a corresponding hyper-mixture $\chi^{(j)}$ in the current k-th iteration as: $\chi^{(j)} = \arg\min_{\chi \in X^{(k-1)}} d(\vartheta_j, \chi)$, where $d(\cdot, \cdot)$ is defined with (1).
  - Evaluate hyper-mixture parameters c_f and Σ_f using ML estimates (2) and (3), to obtain $X^{(k)}$.
  - If any cluster "runs out" of mixtures, set $N_{hypc} = N_{hypc} - C$ for the next iteration, where C is the number of such clusters.
  - Until $D_{average} < \varepsilon$, for $D_{average}$ defined by (5).

EDGS:

– Initialization:
  - Specify the number of groups G.
  - Using any CNP method, obtain optimal OWE weights for predefined maxness $\alpha \in [0, 1]$ as: $\omega_{opt} = \arg\max D(\omega)$, satisfying constraints $M(\omega) = \alpha, 0 \leq \omega_j \leq 1, \sum_{j=1}^{p} \omega_j = 1$.
  - For ω_{opt}, determine the group threshold vector (elements are group borders) $\tau = [\tau_{\max}^{(1)}, \ldots, \tau_{\max}^{(G-1)}]$, and set $\tau_{\min}^{g+1} = \tau_{\max}^{g}$, $g = 1, \ldots, G-1$, and $\tau_{\min}^{1} = 0$, $\tau_{\max}^{G} = \infty$. The group borders should satisfy the constraint: $\tau_{\max}^{g+1} = 2\tau_{\max}^{g}$, for $g = 1, \ldots, G-2$, where $\tau_{\max}^{1}$ is obtained heuristically.
– Mixture Grouping:

  For every $i = 1, \ldots, M$, for mixture ϑ_i do:
  - Obtain eigenvalues $\lambda^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_p^{(i)})$.
  - Assign ϑ_i to the group g iff: $\mathrm{OWE}_{\omega_{opt}}(\lambda^{(i)}) \in [\tau_{\min}^{g}, \tau_{\max}^{g})$.
– Perform baseline VQGS method on every group separately to obtain clusters with mixtures and corresponding hyper-mixtures.

The decoding process is the same for both schemes:

Decoding:

For all observations $x_t, t = 1, \ldots, N$, where N is the number of observations in the testing process do:

For every cluster $C_k, k = 1, \ldots, N_{hypc}$ do:

- Evaluate log-likelihood $\ln f(x_t|\chi^{(k)})$, where $\chi^{(k)}$ is the hyper-mixture that corresponds to cluster C_k.
- If $\ln f(x_t|\chi^{(k)}) > \theta$, where θ is a predefined likelihood threshold, evaluate the exact likelihood for all mixtures that belong to the cluster C_k. Else, set all belonging mixture log-likelihoods to $\ln f(x_t|\Theta^{(k)})$, where $\Theta^{(k)}$ is the Gaussian mixture with centroid c_k and covariance matrix W_k defined by (4).

4 Experimental results

In this chapter results are presented, confirming considerations from previous sections and showing that Eigenvalues Driven Gaussian Selection gives better results than the baseline VQ Based Gaussian Selection that does not incorporate grouping of mixtures. It has smaller Word Error Rate (WER [%]) and achieves more favorable Computation Factor *CF* then the baseline approach.

4.1 System description

For the experiments, Continuous Speech Recognition HMM based system with full Gaussian mixtures was used, where training was done using Tree Based Clustering (TBC) algorithm [1, 8, 21], making the parameters shared by the phonetic models. System used 26 features, where 24 of them

describe spectral envelope (12 static and 12 dynamic Mel Frequency Cepstral Coefficients i.e. the first time derivates) and 2 of them describe normalized energy and its first time derivate. For purpose of the experiments, Speech base recorded on *Faculty of Technical Sciences, Novi Sad* was used. The base contains about 1000 different speakers (half of them male, half female), recorded over the public telephone network. Files were recorded in an A-low format with the sampling frequency of 8 kHz.

The recognition system used 2682 acoustical states with altogether 6984 Gaussian mixtures. Due to the fact that system uses tied states (so the number of parameters in the system is reduced) end full covariance matrices (optimal number of mixtures per state is significantly lower than with the diagonal), possible further reduction in the means of number of mixtures that has to be computed was not expected to be so significant. Nevertheless, using proposed EDGS method, reduction of nearly 50% of mixtures was obtained, with minimal degradation of the performances (WER from 5.66% in the full system, to 5.81%, which is the best result obtained in the system using EDGS selection scheme in comparison to 6.33% obtained by VQGS scheme).

4.2 Results

In Table 1 comparative performances of recognition systems with applied baseline VQGS and EDGS method for the Gaussian selection are presented. In column 1 of the table, term n_{avr} represents average number of mixtures per cluster. Predefined number of clusters was obtained as explained in Sect. 2 by dividing number of mixtures M used in system, by n_{avr}. Group threshold vector in column 2 represents vector with values of border thresholds $\tau^g_{\max}, g \in \{0, \ldots, G-1\}$ that was used for determination of the group where particular mixture should be assigned, on the base of the aggregated value obtained by OWA operator applied on the vector of its eigenvalues. The number of groups G was predefined, and it was held $G = 4$ for all experiments. Term α in column 3 represent the maxness of OWA operator used, while the term θ in column 4 represents log-likelihood threshold used for hyper-mixture evaluation on the acoustic input vectors in the recognition process. Log-likelihood threshold is held on $\theta = 10.0$ in all experiments represented in table, because it gave the best trade-off between recognition accuracy and computing efficiency. With significantly larger value of log-likelihood threshold (for example $\theta = 30.0$) it was not possible to hold WER of CSR systems (with applied VQGS and applied EDGS schemes) close enough to the WER of the full system.

It can be seen that EDGS Gaussian selection scheme obtains better results in comparison with VQGS scheme. The EDGS scheme obtained lower WER [%], and *CF* factor at the same time than VQGS scheme. It results with the better

Table 1 Competitive performances for VQGS and EDGS selection scheme

Selection scheme	n_{avr}	τ_{vec}	α	θ	WER [%]	*CF*
FULL	–	–	–	–	5.68	1.0
VQGS	140	–	–	10	6.54	0.64
	70	–	–	10	6.33	0.60
	30	–	–	10	6.61	0.61
EDGS	100	[0.05 0.1 0.2]	0.5	10	5.94	0.57
	50	[0.05 0.1 0.2]	0.5	10	6.20	0.55
	100	[0.2 0.4 0.8]	0.9	10	5.88	0.56
	50	[0.2 0.4 0.8]	0.9	10	6.14	0.54
	100	[0.4 0.8 1.6]	1.0	10	5.81	0.53
	50	[0.4 0.8 1.6]	1.0	10	5.96	0.54

trade-off between recognition accuracy and computing efficiency of a system with applied EDGS. The maxness α of the OWA operator that gave the best result is $\alpha = 1$, which due to the procedure of maximization of the dispersion of the coefficients gives that only largest 4 or 5 eigenvalues have to be taken into account in the grouping process. It confirms that the largest eigenvalues are the most significant in the grouping process.

Considerations from the previous section are confirmed on simulations and the real data respectively. Corresponding numerical results and diagrams are presented in Tables 2 and 3, and on Figs. 1 and 2. The average difference Δ_{avr} between the likelihoods evaluated on hyper-mixtures and belonging mixtures averaged over all observations and clusters is defined as:

$$\Delta_{avr} = \frac{1}{N}\frac{1}{M}\sum_{t=1}^{N}\sum_{j=1}^{M} |\ln f(x_t|\vartheta_j) - \ln f(x_t|\chi^{(j)})| \qquad (13)$$

In (13), term N represents the overall number of observations and M is the overall number of mixtures used in HMM system. Term $\ln f(x_t|\vartheta_j)$ is the log-likelihood evaluated for some t-th observation on mixture ϑ_j, $j \in \{1, \ldots, M\}$ and the term $\ln f(x_t|\chi^{(j)})$ is the log-likelihood evaluated on the hyper-mixture corresponding to the cluster that ϑ_j belongs. Both, the simulation and the real data experiments show that Δ_{avr} defined with (13) is smaller in the case of EDGS scheme than in the case of baseline VQGS scheme.

For the simulations with results presented in Table 2 and on a corresponding diagram at Fig. 1, setup was delivered as follows: A number of $M = 80$ different mixtures were generated with centroids uniformly distributed on a region $\Omega = [0, 80] \times [0, 80] \subset R^2$. Mixture covariance matrices are obtained so that half of generated centroids have covariance $R_{small} = 5.0I$, and half of them $R_{large} = 40.0I$, where I is identity matrix. Choice of those with smaller covariance was also random with uniform distribution. The intention was to make significant degree of overlapping between

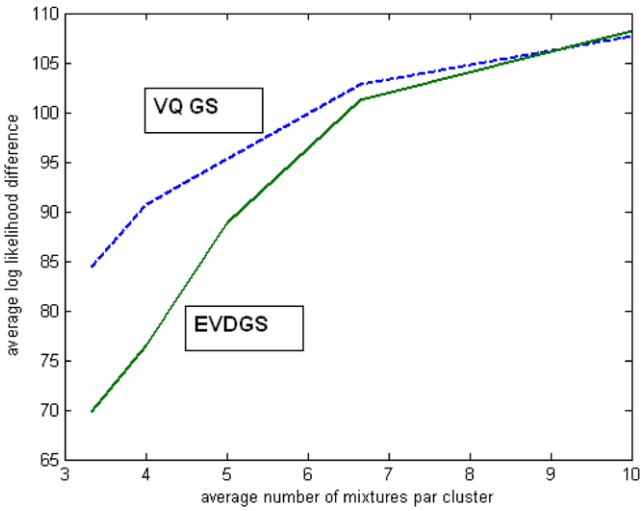

Fig. 1 Diagram Δ_{avr} in function of n_{avr}, that corresponds to Table 2. Competitive results for VQGS and EDGS scheme, obtained in simulations

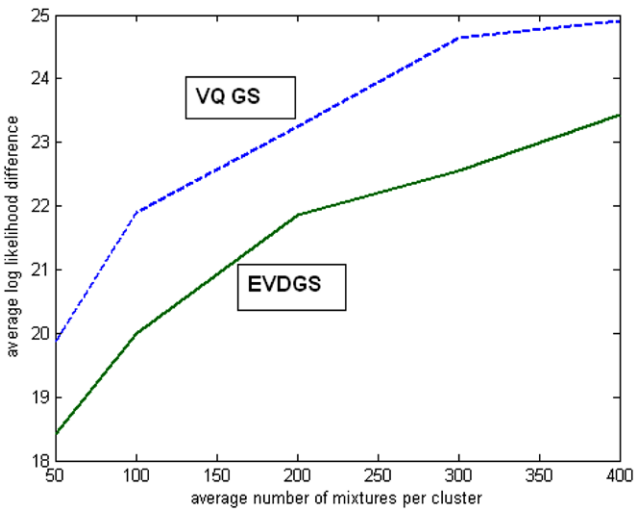

Fig. 2 Diagram Δ_{avr} in function of n_{avr}, that corresponds to Table 3. Competitive results for VQGS and EDGS scheme, obtained on real data

Table 2 Average log likelihood differences Δ_{avr} between hyper-mixtures and belonging mixtures obtained in simulation, for different values of the average number of mixtures per cluster. Competitive results for VQGS and EDGS scheme, obtained on the simulation on 80 uniformly distributed mixtures and 10000 uniformly distributed observations

N_{hypc}	n_{avr}	$n_{avr}^{percent}$ [%]	Δ_{avr}^{VQGS}	Δ_{avr}^{EDGS}
8	10	12.5	107.71	108.23
12	6.66	8.3	102.77	101.24
16	5	6.25	95.32	88.83
20	4	5	90.72	76.47
24	3.33	4.16	84.47	69.76

Table 3 Average log likelihood differences between hyper-mixtures and belonging mixtures Δ_{avr}, obtained in real system on all test labels, for different values of the average number of mixtures per cluster. Competitive results for VQGS and EDGS scheme

n_{avr}	$n_{avr}^{percent}$ [%]	Δ_{avr}^{VQGS}	Δ_{avr}^{EDGS}
400	5.7	24.91	23.43
300	4.3	24.64	22.56
200	2.9	23.26	21.85
100	1.4	21.89	20.00
50	0.7	19.88	18.42

mixtures, as it is the case in real CSR problems. A number of $N = 10000$ observations were generated uniformly on a region $\Omega' = [-20,100] \times [-20,100] \subset R^2$. Borders of region Ω' are more stretched then borders of Ω, in order to take into account variances of mixtures. Average log-likelihood difference Δ_{avr} was evaluated for two different scenarios. For Scenario 1 that simulates EDGS scheme, mixtures were placed in two disjoint groups: one containing mixtures with covariances equal to R_{small} ("narrow" mixtures), and other containing those with covariances equal to R_{large} ("wide" mixtures). The baseline clustering of mixtures (described in Sect. 2) is then performed on every group separately with a predefined number of clusters (i.e. average number of mixtures per cluster). For Scenario 2 that simulates VQGS scheme, the same clustering was performed, without any grouping of mixtures prior to that. First column N_{hypc} of Table 2 represents the predefined number of hyper-clusters. Second column is n_{avr}, obtained by dividing the overall number of mixtures M with N_{hypc}. Third column $n_{avr}^{percent}$, presents n_{avr} in percents of overall number of mixtures M. In the Table 2 and on the corresponding diagram at Fig. 1, it can be seen that if the number of mixtures per cluster n_{avr} is not too large (not larger than 10% of overall number of mixtures), average log-likelihood difference Δ_{avr} defined with (22), in the case of Scenario 1, represented with the full line, is smaller then in the case of Scenario 2, represented with the dash line. Only for $n_{avr}^{percent} = 10\%$, the bigger centroid displacement in a single cluster in the case of EDGS in comparison to VQGS scheme became significant. Nevertheless, $n_{avr}^{percent} = 10\%$ is too large for real time data tests, as it can severally reduce recognition accuracy. Actually in all real time data experiments (Table 3), Δ_{avr}^{EDGS} came out smaller then Δ_{avr}^{VQGS}, for all $n_{avr} \leq 400$.

In Table 3, and on the corresponding diagram on Fig. 2, the similar results are presented for the real data. In the recognition process, on the system trained as it is described at the beginning of a section, overall Δ_{avr} obtained for all labels from the testing base is observed for two different configurations. One configuration (Δ_{avr}^{VQGS}) with applied baseline VQGS selection scheme and the other (Δ_{avr}^{EDGS}) with applied proposed EDGS scheme. There was no pruning in

mixture evaluation (θ set on wary small value) i.e. all the mixtures were evaluated. This was done in order to explore overall Δ_{avr} averaged on all mixtures and observations, and obtained for sufficiently large numbers of labels. All experiments involving EDGS scheme were done for fixed OWE operator used. Fixed $\alpha = 1$ and ω_{opt} obtained by minimizing dispersion, as explained in the previous section. Group threshold vector was heuristically obtained and set to be $\tau = [0.4, 0.8, 1.6]$, while the number of groups was set on $G = 4$ (same as for $\alpha = 1$ in Table 1). Testing was conducted for different values of n_{avr}. It can be seen that if n_{avr} is reasonably small (so that the recognition error is not significantly higher in the case of GS system in comparison to the full system), overall Δ_{avr} averaged over all labels is smaller in the case of EDGS system then in the case of VQGS system. Actually, this is satisfied for all n_{avr} from Table 3.

Experimental results shows that EDGS gains better results in the means of Word Error Rate and computational efficiency (i.e. *CF* factor) than the baseline VQGS method, as presented in Table 1. This is consistent with the fact that EDGS scheme obtained smaller overall Δ_{avr} on testing labels than the baseline VQGS scheme, as presented in Table 3.

5 Conclusions

In this paper, a novel GS approach named "Eigenvalues Driven Gaussian Selection" is proposed. The main idea is to take into account the most significant eigenvalues in the GS process. The mixtures to be selected are grouped in to several groups on the base on eigenvalues of their covariance matrices, before the actual baseline VQ clustering described in the Sect. 2 is performed on each group separately. The combination is done in the way that gives more weight to the larger (more significant) eigenvalues in the grouping process. For that manner, the usage of OWE aggregation operators is proposed, with weights optimal in the way that for the fix maxness $M(\omega) = \alpha \in [0, 1]$, dispersion $D(\omega)$ is maximized.

Experimental results on simulations and real data (in the recognition process) favor EDGS selection scheme in comparison with VQGS scheme in the means of WER [%] and *CF* factor.

Acknowledgements This research work has been supported by the "Serbian Ministry of Science" and it has been realized as a part of "Human-Computer Speech Communication" research project.

References

1. Bahl LR, de Souza PV, Gopalakrishnan PS, Nahamoo D, Picheny MA (1991) Context dependent modeling of phones in continuous speech using decision trees. In: Proc DARPA speech and natural language processing workshop, Pacific Grove, pp 264–270
2. Biggs MC (1975) Constrained minimization using recursive quadratic programming. In: Dixon LCW, Szergo GP (eds) Towards global optimization. North-Holland, Amsterdam, pp 341–349
3. Bocchieri E (1993) Vector quantization for efficient computation of continuous density likelihoods. In: Proc ICASSP, Minneapolis, MN, vol 2, pp II-692–II-695
4. Coleman TF, Li Y (1996) An interior, trust region approach for nonlinear minimization subject to bounds. SIAM J Optim 6:418–445
5. Gales M (1999) Semi-tied covariance matrices for hidden Markov models. IEEE Trans Speech Audio Process 7(3):272–281
6. Huang XD, Lee KF, Hon HW (1990) On semi-continuous hidden Markov modelling. In: Proc ICASSP, pp 689–692
7. Hunt M, Lefebre C (1989) A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In: Proc ICASSP, pp 262–265
8. Kannan A, Ostendorf M, Rohlicek JR (1994) Maximum likelihood clustering of Gaussians for speech recognition. IEEE Trans Speech Audio Process 2(3):453–455
9. Kay SM (1993) Fundamentals of statistical signal processing: estimation theory. Prentice Hall, New York
10. Knill KM, Gales MJF, Young SJ (1996) Use of Gaussian selection in large vocabulary continuous speech recognition using HMMs. In: Proc int conf spoken language processing
11. Lindo Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun COMM 28:84–95
12. O'Hagan M (1988) Aggregating template or rule antecedents in real time expert systems with fuzzy set logic. In: Proc of the 22-th annual IEEE Asilomar conferences on signals, systems and computers, Pacific Grove, pp 681–689
13. Shinoda K, Lee C-H (2001) A structural Bayes approach to speaker adaptation. IEEE Trans Speech Audio Process 9(3):276–287
14. Simonin J, Delphin L, Damnati G (1998) Gaussian density tree structure in a multi-Gaussian HMM based speech recognition system. In: 5th int conf on spoken language processing, Sidney, Australia, 4 December 1998
15. Watanabe T, Shinoda K, Takagi K, Iso K (1995) High speed speech recognition using tree-structured probability density function. In: Proc int conf acoust speech signal process, vol 1, pp 556–559
16. Webb A (1999) Statistical pattern recognition. Oxford University Press, London. Arnold a member of the Hodder Headline Group, 338 Euston Road, London NW1 3BH, Great Britain
17. Yager RR (1988) On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans Syst Man Cybern 18:183–190
18. Yager RR, Kacprzyk J (1997) The ordered weighted averaging operators, theory and applications. Kluwer Academic, Dordrecht
19. Yager RR, Rybalov A (1996) Uniform aggregation operators. Fuzzy Sets Syst 80:111–120
20. Yager RR, Rybalov A (1998) Full reinforcement operators in aggregation techniques. IEEE Trans Syst Man Cybern 28:757–769
21. Young SJ, Odell JJ, Woodland PC (1994) Tree-based state tying for high accuracy acoustic modeling. In: Proc of the workshop on human on human language technology, pp 307–312

Marko Janev received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 2003.
Currently he is a research assistant at the "Human Computer Interaction" research project at the FTN Novi Sad, University of Novi Sad, Serbia.
His research interests include statistical pattern recognition, speech recognition and speech processing.

Darko Pekar received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia in 1998.
Currently he is a head of research at AlfaNum Speech Technologies Ltd. located in Novi Sad, Serbia.
His research interests include statistical pattern recognition, speech recognition, speech processing and speech synthesis.

Niksa Jakovljevic received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia in 2001.
Currently he is a consultant at the Faculty of Technical Sciences (FTN) in Novi Sad, University of Novi Sad, Serbia.
His research interests include statistical pattern recognition, speech recognition and speech processing.

Vlado Delic 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.