Open access • Journal Article • DOI:10.1109/LSP.2007.914792

Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification

Ville Hautamäki, Tomi Kinnunen, Ismo Kärkkäinen, Juhani Saastamoinen ...+2 more authors

Institutions: University of Eastern Finland

Published on: 22 Jan 2008 - IEEE Signal Processing Letters (IEEE)

Topics: Maximum a posteriori estimation, Speaker recognition, Vector quantization, NIST and Mixture model

Maximum *a Posteriori* Adaptation of the Centroid Model for Speaker Verification

Ville Hautamäki, Tomi Kinnunen, Ismo Kärkkäinen, Juhani Saastamoinen, Marko Tuononen, and Pasi Fränti

***Abstract*—Maximum a posteriori adapted Gaussian mixture model (GMM-MAP) is widely used in speaker verification. GMMs have three sets of parameters to be adapted: means, covariances, and weights. However, practice has shown that it is sufficient to adapt the means only. Motivated by this, we formulate maximum a posteriori vector quantization (VQ-MAP) procedure which stores and adapts the mean vectors (centroids) only. Experiments on the NIST 2001 and NIST 2006 corpora indicate that VQ-MAP gives comparable accuracy with GMM-MAP with simpler implementation and faster adaptation.**

***Index Terms*—Bayesian methods, centroid model, maximum a posteriori (MAP) adaptation, speaker verification, vector quantization.**

I. Introduction

In *speaker verification* [1], the unknown speech utterance is introduced to the system accompanied by a claim. The task is to decide whether the claim was true or false, by matching the unknown test utterance to a previously stored model. Speaker recognition systems typically have used generative models such as *vector quantization* (VQ) [1] (aka the *centroid model*) or *Gaussian mixture models* (GMMs) [2].

The generative model is typically trained using the maximum likelihood (ML) principle. The ML approach usually does not generalize well to unseen speech data with finite amount of training material. *Maximum a posteriori* (MAP) objective training attacks this problem of by using a so-called *universal background model* (UBM) [2]. In the MAP approach, prior knowledge of the distribution of the model parameters is incorporated into the modeling process. Even if some areas of the feature space are less represented in the training data, the prior information about the parameters can help to overcome the problem. However, incorporating the prior information is not trivial because prior parameter distribution has its own parameters, known as *hyperparameters*, which can be difficult to estimate.

Maximum *a posteriori* training for Gaussian mixtures was first formulated in [3], where Gauvain and Lee solved two key issues in MAP estimation of Gaussian mixture parameters, namely, choice of the prior distribution family, and specification of the parameters of the prior densities, which led to simple *expectation maximization* (EM) re-estimation formulae. The original motivation of the authors was to enhance the performance of a HMM-based speech recognition system. The idea was later successfully applied to speaker verification as well [2]. To overcome the parametrization problem, Reynolds *et al.* [2] proposed to use the universal background model as a set of parameters of the prior distribution. Some constraints to the parameters and a relevance factor were introduced to solve the prior parametrization problem.

Gaussian mixtures have three sets of parameters to be adapted: mean vectors (*centroids*), covariance matrices, and weights. However, experiments have indicated that best results are obtained by adapting the mean vectors only [2]. Motivated by this, we formulate maximum *a posteriori* vector quantization (VQ-MAP), which is a special case of the GMM-MAP method [2]. The advantages of the proposed centroid-based model are simpler implementation and much faster adaptation. The speedup originates mostly from the replacement of the Gaussian density computations with squared distance computations, leaving out the exponentiation and additional multiplications.

II. VQ-MAP

In vector quantization, the goal is to estimate parameter vector modeling the speaker, denoted as $\Theta = (\boldsymbol{c}_1^t, \ldots, \boldsymbol{c}_K^t)^t$. Here $\boldsymbol{c}_i$ are the centroids and K is the model size, which is a trade-off between the representation accuracy and the speaker model size.

The maximum *a posteriori* modeling paradigm, irrespective of the actual model in question, is formulated as a way to find Θ that maximizes the posterior probability density function (pdf). Formally

$$\begin{aligned}\Theta_{\mathrm{MAP}} &= \arg\max_{\Theta} p(\Theta|X) \\ &= \arg\max_{\Theta} p(X|\Theta)g(\Theta) \end{aligned} \quad (1)$$

where $p(X|\Theta)$ is the likelihood of the training set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ given parameters Θ, and $g(\Theta)$ is the prior pdf of the parameters.

Three subproblems need to be solved so that a maximization algorithm can be derived:

- likelihood function $p(X|\Theta)$ needs to be defined in terms of vector quantization;
- appropriate prior distribution $g(\Theta)$ needs to be defined;
- prior distribution contains its own set of parameters, which also needs to be estimated.

In the following, we address these points and derive the maximization algorithm.

Manuscript received June 18, 2007; revised November 3, 2007. The work was supported by the National Technology Agency of Finland (TEKES) as the projects New Methods and Applications of Speech Technology (PUMS). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian Kan-Wing Mak.

The authors are with the Speech and Image Processing Unit, Department of Computer Science and Statistics, University of Joensuu, FIN-80101 Joensuu, Finland (e-mail: villeh@cs.joensuu.fi; tkinnu@cs.joensuu.fi; iak@cs.joensuu.fi; juhani@cs.joensuu.fi; mtuonon@cs.joensuu.fi; franti@cs.joensuu.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2007.914792

A. Modeling VQ as a Gaussian Mixture

For the VQ-MAP algorithm, we must formulate (1) in the vector quantization framework. Since VQ is not a parametric probabilistic model, we need to specify a likelihood pdf that corresponds to the *mean squared error* (MSE), which defines the VQ model. We can model the likelihood $p(X|\Theta)$ as a Gaussian mixture as in [4]. The density of the kth component is defined as

$$p(\boldsymbol{x}|\boldsymbol{c}_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}-\boldsymbol{c}_k\|^2\right\} \quad (2)$$

where $\Sigma_k = I\epsilon$, and ϵ is constant. It is shown in [4] that with this model, the EM algorithm reduces to k-means algorithm and that component prior weights π_k do not play any role in the algorithm. The weights just reflect the proportion of the data vectors in a given cluster.

B. Defining the Prior Density

When selecting the appropriate prior density $g(\Theta)$, a good choice would be the *conjugate prior* of the $p(\boldsymbol{x}|\boldsymbol{c}_k, \Sigma_k)$ as in [4]. Prior distribution is called a conjugate prior, if its algebraic form is the same as the resulting posterior distribution. The conjugate prior of a multivariate Gaussian with a known covariance matrix is a multivariate Gaussian. We can therefore model the prior of the component k as

$$g(\boldsymbol{c}_k|\boldsymbol{\mu}_k, \hat{\Sigma}_k) = B_k \exp\left\{-\frac{1}{2}(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^t\hat{\Sigma}_k^{-1}(\boldsymbol{c}_k-\boldsymbol{\mu}_k)\right\} \quad (3)$$

where $\hat{\Sigma}$ is the covariance matrix of the prior distribution, and

$$B_k = \frac{1}{(2\pi)^{D/2}|\hat{\Sigma}_k|^{1/2}}. \quad (4)$$

Assuming independence between the parameters of the individual mixture components, as was done in [3], we conclude the definition of the prior model by noting that

$$g(\Theta) \propto \prod_{k=1}^{K} g(\boldsymbol{c}_k|\boldsymbol{\mu}_k, \hat{\Sigma}_k). \quad (5)$$

C. MAP Estimates for Vector Quantization

In order to maximize the posterior pdf in (1), we need to jointly determine the observation posteriors and the model parameters of each component. Unfortunately, the maximization cannot be performed directly [3]. Instead, locally optimal solution can be obtained by EM algorithm [4] for the Gaussian mixture models and by k-means algorithm for the vector quantization models. Both algorithms work essentially in a similar manner:

1) find observation posteriors (E-step);
2) given the posteriors, re-estimate the parameters (M-step).

In k-means, the observation posteriors correspond to hard partitioning of the dataset. In the M-step, the parameters are maximized by calculating new centroid estimates. Now the corresponding steps need to be defined in the new framework so that MAP parameters can be optimized.

Interestingly, the term $g(\Theta)$ affects the maximization of the posterior distribution only in the M-step [4]. Optimal Θ with respect to observation posteriors can then be calculated by maximizing the following auxiliary function [3]:

$$R(\Theta, \hat{\Theta}) = Q(\Theta, \hat{\Theta}) + \log g(\Theta) \quad (6)$$

where $\hat{\Theta}$ are the parameters estimated in the previous iteration, and Θ are the parameters to be estimated. The function Q is the expectation of the complete-data log likelihood and can be expressed as [4]

$$Q(\Theta, \hat{\Theta}) = \sum_{i=1}^{N}\sum_{j=k}^{K}\gamma_{ik}\left\{\ln\pi_k + \ln p(\boldsymbol{x}|\boldsymbol{c}_k, \Sigma_k)\right\} \quad (7)$$

where π_k is the prior weight of the component k, and γ_{ik} is the posterior probability of the observation i for the component k. By letting $\epsilon \to 0$ in (2), the complete-data log likelihood function becomes the MSE [4]

$$Q(\Theta, \hat{\Theta}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2. \quad (8)$$

The values of r_{ik} form a binary matrix, where

$$r_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_j \|\boldsymbol{x}_i - \boldsymbol{c}_j\| \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

In vector quantization literature, MSE describes the distortion when observations $\boldsymbol{x}_i$ are encoded as their nearest centroids $\boldsymbol{c}_j$.

By substituting (5) and (8) into (6), we arrive at a new auxiliary function form

$$R(\Theta, \hat{\Theta}) = -\sum_{i=1}^{N}\sum_{j=1}^{K} r_{ik}\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 - \sum_{k=1}^{K}(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^t\hat{\Sigma}_k^{-1}(\boldsymbol{c}_k-\boldsymbol{\mu}_k). \quad (10)$$

We need to find such $\boldsymbol{c}_k$ for each component that minimizes the above equation. The $\boldsymbol{\mu}_k$ and $\hat{\Sigma}_k$ are our prior parameters for the component k, and they are selected from a previously trained universal background model as in [2]. However, in the VQ model, covariance matrices (variance parameters) are not recorded as a part of the UBM. Therefore, we substitute $\hat{\Sigma}_k = I$ for all components. This is motivated by our model assumptions in (2). Now, $R(\Theta, \hat{\Theta})$ can be written as

$$R(\Theta, \hat{\Theta}) = r_{11}\|\boldsymbol{x}_1 - \boldsymbol{c}_1\|^2 + \cdots + r_{NK}\|\boldsymbol{x}_N - \boldsymbol{c}_K\|^2 + \|\boldsymbol{c}_1 - \boldsymbol{\mu}_1\|^2 + \cdots + \|\boldsymbol{c}_K - \boldsymbol{\mu}_K\|^2. \quad (11)$$

Now, let us denote by $S_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|S_k|}\}$ the set of training vectors that are mapped to $\boldsymbol{c}_k$. We denote by R_k the terms of $R(\Theta, \hat{\Theta})$ that contain centroid $\boldsymbol{c}_k$

$$\begin{aligned} R_k &= \|\boldsymbol{x}_1 - \boldsymbol{c}_k\|^2 + \cdots + \|\boldsymbol{x}_{|S_k|} - \boldsymbol{c}_k\|^2 + \|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2 \\ &= 2|S_k|\langle\bar{\boldsymbol{x}}_k, \boldsymbol{c}_k\rangle + |S_k|\|\boldsymbol{c}_k\|^2 + \|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2 \\ &= 2|S_k|\langle\bar{\boldsymbol{x}}_k, \boldsymbol{c}_k\rangle + (|S_k|+1)\|\boldsymbol{c}_k\|^2 + 2\langle\boldsymbol{c}_k, \boldsymbol{\mu}_k\rangle \end{aligned} \quad (12)$$

where $|S_k|$ is the number of vectors mapped to centroid $\boldsymbol{c}_k$, and $\bar{\boldsymbol{x}}_k$ is the average of all vectors in the same cluster. Taking the gradient with respect to $\boldsymbol{c}_k$ from (12), we obtain the centroid re-estimation formula for the M-step as

$$\boldsymbol{c}_k = \frac{|S_k|}{|S_k|+1}\bar{\boldsymbol{x}}_k + \frac{1}{|S_k|+1}\boldsymbol{\mu}_k. \tag{13}$$

III. Speaker Verification Using VQ-MAP

In speaker verification, the prior distribution of the speaker model parameters is represented by a universal background model, which is created by collecting a large number of representative training utterances from a number of speakers. These utterances are then converted into sequences of spectral feature vectors and pooled into a single training set. The UBM can be trained using any clustering algorithm such as k-means, and it is presented by a set of centroid vectors denoted as $U = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$.

A. Enrollment Phase

A new speaker is enrolled by adapting the UBM with the MAP procedure. Given the training data for a new speaker, $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, the adaptation is performed using the following steps:

1) initialization: Set $\boldsymbol{c}_k = \boldsymbol{\mu}_k$ for $k = 1, 2, \ldots, K$;
2) for $i = 1, 2, \ldots, I$, do
   a) For each training vector $\boldsymbol{x}_n$, find the index of the nearest neighbor in the adapted model

$$q_n = \arg \min_{1 \le k \le K} ||\boldsymbol{x}_n - \boldsymbol{c}_k||^2. \tag{14}$$

   b) For the kth cluster, define the set of vectors mapped into that cluster as $S_k = \{\boldsymbol{x}_n \in X | q_n = k\}$ and calculate the centroid vector

$$\bar{\boldsymbol{x}}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{x}_n \in S_k} \boldsymbol{x}_n. \tag{15}$$

   c) For the kth cluster, update the adapted vector as

$$\boldsymbol{c}_k = w_k \bar{\boldsymbol{x}}_k + (1 - w_k)\boldsymbol{\mu}_k \tag{16}$$

   where

$$w_k = \frac{|S_k|}{|S_k| + r} \tag{17}$$

   and r is a fixed constant known as the *relevance factor*.

The result is the adapted model $C = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$. The adaptation process is demonstrated in Fig. 1.

The relevance factor r is a parameter that is used to control the balance of how much the UBM vector and the speaker-specific training data are weighted in adaptation. The "unit" of the parameter is vector count and can be given an interpretation as follows. By having $w_k > 1/2$ in (17), we have $|S_k| > r$. Thus, the relevance parameter can be interpreted as the number of training vectors that are needed in order the computed centroid $\bar{\boldsymbol{x}}_k$ to be equally reliable to the corresponding UBM vector.

B. Verification Phase

Given a sequence of feature vectors, $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, and the claimed speaker model, $C = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$, we compute the log likelihood ratio and compare it against a verification threshold to make the decision. In the VQ model, the log-likelihood is related to the negative square error given in (8). Thus, the match score can be defined as

$$\begin{aligned} \text{score} &= -\text{MSE}(X, C) - (-\text{MSE}(X, U)) \\ &= \text{MSE}(X, U) - \text{MSE}(X, C) \end{aligned} \tag{18}$$

where

$$\text{MSE}(X, Y) = \frac{1}{|X|} \sum_{\boldsymbol{x}_i \in X} \min_{\boldsymbol{y}_k \in Y} ||\boldsymbol{x}_i - \boldsymbol{y}_k||^2. \tag{19}$$

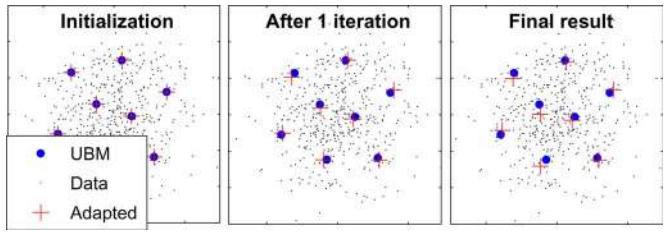

Fig. 1. Illustration of the VQ-MAP procedure with two-dimensional data. The centroid vectors of the universal background model (•) define the prior distribution of the model parameters. Given the training data (.), the adapted model (+) is derived by making local changes to the prior model. Here, the relevance factor is 10.

C. Relation to GMM-MAP

In principle, the resulting adaptation (14)–(17) are simply a special case of the more general GMM-MAP equations [2]. By assuming a diagonal covariance with all dimensions sharing the same variance and by replacing the observation posteriors with the binary 0/1 values (1 for the most probable cluster and 0 for all the others) and interpreting the proportion of data vectors in each cluster as the mixture weight, GMM-MAP equals VQ-MAP. In practice, the proposed model is much simpler to implement and results in faster adaptation.

The VQ-MAP and GMM-MAP algorithms require $O(NKI)$ squared distance and Gaussian density computations, where I is the number of iterations. By counting the number of elementary operations in the squared Euclidean distance and the diagonal covariance Gaussian, the speedup ratio can be written as

$$\text{speedup} = \frac{T_{\text{exp}} + (2D+2)T_{\text{mul}} + 2DT_{\text{add}}}{DT_{\text{mul}} + 2DT_{\text{add}}} \tag{20}$$

where T_{exp}, T_{mul}, and T_{add} denote the costs of exponentiation, multiplication, and addition, respectively, and D denotes the feature vector dimensionality. Exponentiation takes more time than multiplication and addition, which yields a significant speedup in practice, as will be demonstrated.

IV. Experiments

We have used the NIST 2001 SRE corpus for optimizing the control parameters and NIST 2006 SRE corpus for validating the results.[1] The results are presented on the one-speaker detection set of each corpus. The NIST 2001 corpus contains 2 min of training material per each of the 174 target speakers and 2038 test segments. The NIST 2006 corpus contains 816 target speakers trained using 5 min of data and 3735 test segments. We have trained the UBMs using the development set of the NIST 2001 SRE corpus.

[1]http://nist.gov/speech/tests/spk/

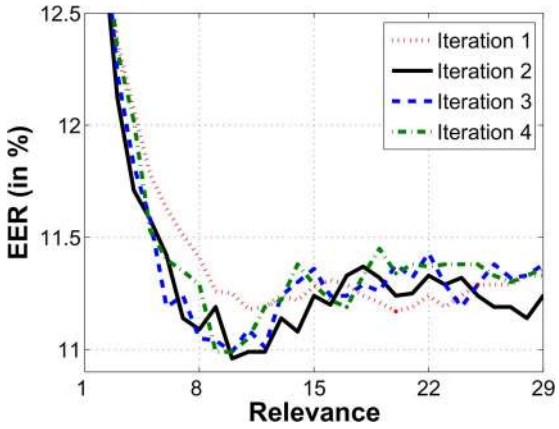

Fig. 2. Effect of control parameters of VQ-MAP on the equal error rates of the NIST 2001 corpus using a model of size $K = 64$.

We use the first 12 mel-frequency cepstral coefficients (MFCCs), their deltas, and double deltas as the features, implying 36-dimensional feature space. A periodicity-based voice activity detector was used [5], and the detected speech vectors were normalized to have zero mean and unit variance. The UBM for VQ was generated by a recursive split algorithm, followed by fine-tuning with k-means. This model was used as the initial guess for the GMM model with diagonal covariances and fine-tuned by EM.

Experiment with different settings of the relevance parameter (r) and the number of iterations (I) indicates that, independent on the number of iterations used, the relevance factor should be set to 8 or higher as shown in Fig. 2. There is no significant difference between different iterations, which is similar behavior that was observed also for the GMM-MAP method [2]. It can also be seen in Fig. 1 that the adapted model does not change much with further iterations. We fix the parameters as $r = 12$ and $I = 2$ for further experiments.

For the GMM-MAP, we use the well-established values of $r = 16$ and $I = 1$ [2]. In addition to the means-only adapted GMM, we also present the results for GMM in which the variances are also adapted using the same relevance factor.

The detection error trade-off (DET) curves for the NIST 2001 and NIST 2006 corpuses are shown in Fig. 3. The running times on the NIST 2001 corpus have been summarized in Table I. Finally, Table II summarizes the speedup factors for different model sizes relative to two VQ-MAP iterations. The running times include only the work done in adapting the models; the overhead due to feature extraction and file I/O has been excluded.

VQ-MAP and GMM-MAP provide accuracies close to each other. The variance-adapted GMM gives slightly better result at low false acceptance rates on the NIST 2001 corpus. However, the difference vanishes on the NIST 2006 corpus which includes more training data and more difficult channel conditions.

The VQ-based systems run much faster. In particular, the adaptation step is significantly faster, even though VQ-MAP performs two iterations. An advantage of the obtained speedup would be on platforms with significantly limited CPU power and fixed-point arithmetics, such as mobile phones [6].

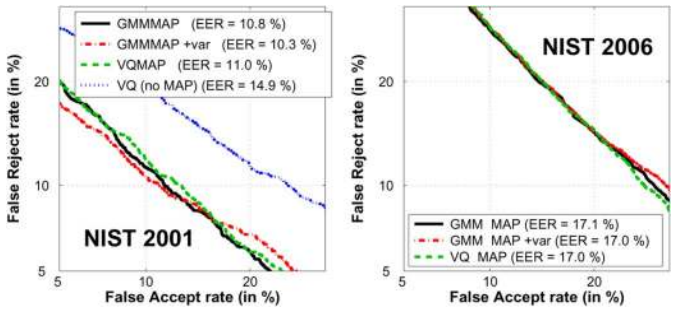

Fig. 3. Results for the NIST 2001 corpus (left) and NIST 2006 corpus (right), model size $K = 64$.

TABLE I
COMPARISON OF THE CPU TIME (IN SECONDS) ON THE NIST 2001 CORPUS (MODEL SIZE $K = 64$)

Method	UBM train	Adaptation avg.	Adaptation std	Matching avg.	Matching std
VQ-MAP ($I = 1$)	170	0.01	0.005	0.14	0.06
VQ-MAP ($I = 2$)	170	0.03	0.009	0.14	0.06
GMM-MAP	514	0.18	0.054	0.54	0.23
GMM-MAP +var	514	0.18	0.055	0.54	0.23

TABLE II
ADAPTATION SPEEDUP FACTORS ON THE NIST 2001 CORPUS RELATIVE TO TWO VQ-MAP ITERATIONS

Method	Model size (K) 64	128	256	512	1024
VQ-MAP ($I = 1$)	2:1	2:1	2:1	2:1	2:1
VQ-MAP ($I = 2$)	1:1	1:1	1:1	1:1	1:1
GMM-MAP	6:1	6:1	9:1	19:1	19:1
GMM-MAP +var	6:1	6:1	9:1	19:1	20:1

V. CONCLUSIONS

We have formulated the MAP algorithm originally developed for GMM to work with the VQ-based model. Experimental results show that the proposed method provides similar recognition accuracy than the GMM-based algorithm but with simpler and faster implementation.

REFERENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, S. Meignier, T. Merlin, J. Ortega-Garcia, I. Magrin-Chagnolleau, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, no. 4, pp. 430–451, 2004.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1, pp. 19–41, 2000.

[3] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[4] C. M. Bishop, *Pattern Recogntion and Machine Learning*. New York: Springer-Verlag, 2006.

[5] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)*, Moscow, Russia, Oct. 2007, vol. 2, pp. 645–650.

[6] J. Saastamoinen, E. Karpov, V. Hautamäki, and P. Fränti, "Accuracy of MFCC based speaker recognition in Series 60 device," *J. Appl. Signal Process.*, no. 17, pp. 2816–2827, Sep. 2005.