COMPARING MAXIMUM A POSTERIORI VECTOR QUANTIZATION AND GAUSSIAN MIXTURE MODELS IN SPEAKER VERIFICATION[*]

Tomi Kinnunen, Juhani Saastamoinen, Ville Hautamäki, Mikko Vinni, Pasi Fränti

Speech and Image Processing Unit (SIPU), Dept. of Computer Science and Statistics
University of Joensuu, P.O. Box 111, FI-80101 Joensuu, FINLAND
E-mail: {tkinnu,juhani,villeh,mvinni,franti}@cs.joensuu.fi

ABSTRACT

Gaussian mixture model - universal background model (GMM-UBM) is a standard reference classifier in speaker verification. We have recently proposed a simplified model using vector quantization (VQ-UBM). In this study, we extensively compare these two classifiers on NIST 2005, 2006 and 2008 SRE corpora, while having a standard discriminative classifier (GLDS-SVM) as a reference point. We focus on parameter setting for N-top scoring, model order, and performance for different amounts of training data. The most interesting result, against a general belief, is that GMM-UBM yields better results for short segments whereas VQ-UBM is good for long utterances. The results also suggest that maximum likelihood training of the UBM is sub-optimal, and hence, alternative ways to train the UBM should be considered.

Index Terms— Speaker verification, MFCCs, Gaussian mixture model (GMM), vector quantization (VQ), MAP training

1. INTRODUCTION

Typical speaker verification systems use mel-frequency cepstral coefficients (MFCCs) to parameterize speech signal. Feature extraction is followed by speaker modeling, for which two approaches have been dominant in the 21st century: generative modeling based on *maximum a posteriori* (MAP) adaptation of a speaker-independent *universal background model* (UBM) [1, 2], and discriminative modeling based on the concept of separating hyperplane [3, 4]. Latest solutions also use a so-called *eigenchannel* transformation and *joint factor analysis* (JFA) to reduce the effects of channel and session variability in the speaker models [5].

We use MFCCs and focus on the speaker modeling by Gaussian mixture model with UBM (GMM-UBM) [1], vector quantizer with UBM (VQ-UBM) [2] and generalized linear discriminant sequence support vector machine (GLDS-SVM) [3]. We set the following limitations in order to keep the baseline simple: (1) we use only telephone data for background modeling, (2) we do not use any intersession variability compensation, (3) we do not make use of ASR component, (4) we do not make use of language information, (5) we do not use additional score normalization such as T-norm [6]. More complete systems used in recent NIST speaker recognition evaluations use such techniques in conjunction with each other. Our simplifications allow us to focus more deeply on the modeling component, but on the other hand, weaken the overall performance in comparison to more complete systems, especially for non-telephony data.

Vector quantization speaker modeling was popular in the 1980s and 1990s [7, 8], but after the introduction of the background model concept for GMMs [1], GMM has been the dominant approach. Even so, usually only the mean vectors of the GMM are adapted while using shared (co)variances and weights for all speakers. This raises a question whether the variances and weights are needed at all. To answer this question, we derived MAP adaptation algorithm for the VQ model [2] as a special case of the MAP adaptation for GMM, involving only the centroid vectors. The VQ approach achieves speed-up in training compared to GMM with comparable accuracy.

In this paper, we further explore the inherent differences of the GMM-UBM and the VQ-UBM classifiers in the speaker verification task, while having the GLDS-SVM classifier as a reference point. The results presented here are based on our submissions to NIST 2006 and NIST 2008 speaker recognition evaluations. We focus on parameter setting for fast N-top scoring, model order, performance for different amounts of training data and effects of mismatched data. In [2], our main focus was in formal derivation of the algorithm rather than in extensive testing. This paper serves for that latter purpose.

Since the VQ-model has less free parameters to be estimated, it may be hypothesized that VQ-based classifier will outperform GMM for small amounts of data; see, for instance, [9] for such an observation. This hypothesis is probably true if both models are trained using maximum likelihood (mean square error minimization). However, it is less clear how the situation changes when using MAP training. In this paper, we will show surprising experimental evidence that suggests the opposite: GMM-UBM is better for short utterances whereas VQ-UBM outperforms GMM-UBM when the length of training and test data increases. We discuss the possible reasons for this and its implications.

2. SYSTEM DESCRIPTION

2.1. Feature Extraction and Classifier Training

The MFCCs are extracted from 30 msec Hamming-windowed frames with 50 % overlap. We use 12 MFCCs computed via 27-channel mel-frequency filterbank. The MFCC trajectories are smoothed with RASTA filtering, followed by appending of the Δ and Δ^2 features. The last two steps are voice activity detection (VAD) and utterance-level mean and variance normalization in that order. For the VAD, we use an energy-based algorithm that uses file-dependent detection threshold based on maximum energy level.

GMM-UBM system follows the standard implementation with diagonal covariance matrices [1]. We use two gender-dependent UBMs trained by deterministic splitting method, followed by seven

[*] EXTENDED VERSION OF THE PAPER HAS BEEN ACCEPTED FOR PUBLICATION IN *PATTERN RECOGNITION LETTERS*.

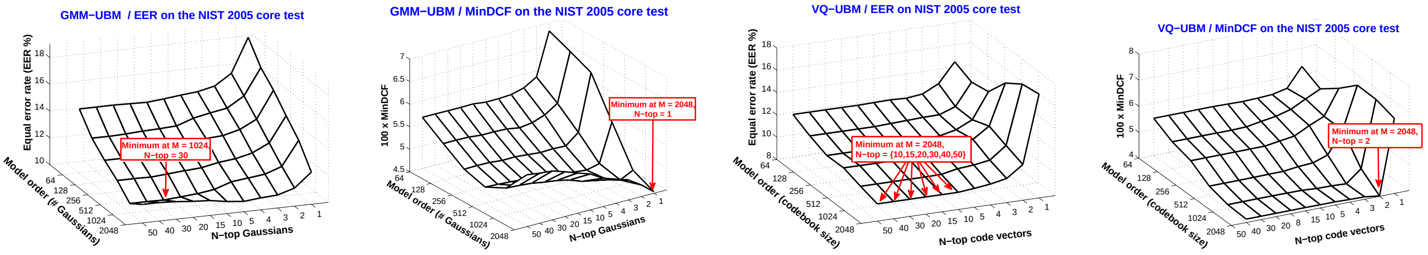

Fig. 1. Effect of N-top scoring to recognition accuracy.

K-means iterations and further two EM-iterations. Only the mean vectors are adapted using a relevance factor $r = 16$. During recognition, the N top-scoring Gaussians are found from the UBM for each feature vector, and only the corresponding adapted Gaussians in the target model are evaluated. Match score is the difference of the target and UBM log-likelihoods.

The VQ-UBM system [2] is similar to GMM-UBM but consists of only centroid vectors without any variance or weight information. Two gender-dependent UBMs are trained using the splitting method, followed by 20 K-means iterations. For adaptation, we use a relevance factor $r = 12$ and an iteration count $I = 2$ [2]. During recognition, the N nearest UBM vectors are found for each vector. In the speaker model, nearest neighbor search is limited on the corresponding adapted vectors only. Match score is the difference of the UBM and target quantization errors.

The GLDS-SVM system follows the implementation presented in [3]. The 36-dimensional MFCCs are first expanded by calculating all the monomials up to order 3, implying 9139-dimensional features. The expanded features are then averaged to form a single *supervector* for each utterance. A separating hyperplane with the target speaker on the positive side and the background speakers on the negative side is then trained using the commonly available *Statistical Pattern Recognition Toolbox*[1]. As in GMM-UBM and VQ-UBM, we use gender-dependent background sets for GLDS-SVM. The match score is computed as the inner product between the model vector and the supervector of the test utterance.

In addition to the three base classifiers, we consider their *fusion* by linear match score weighting as implemented in the FoCal toolkit[2]. In preliminary experiments, we experimented with several other solutions such as Bayes nets and neural networks but the logistic regression yielded most robust result and was therefore chosen.

2.2. Corpora and Performance Evaluation

We use NIST 2005 and NIST 2006 speaker recognition evaluation (SRE) data sets for optimizing the parameters, of which the most important is the *model order* (number of Gaussians and centroids in GMM and VQ, respectively). Furthermore, we use the latest NIST 2008 SRE corpus to investigate the effect of mismatched data - the 2008 SRE data contains, for instance, interview data that is not present in the other corpuses. Fusion accuracy is also evaluated on NIST 2008, while the fusion weights are trained on NIST 2006.

In all three corpora, we focus on two test conditions: the "core" test (1conv-1conv, short2-short3) containing 5 minutes of train and test data, and the shorter 10sec-10sec test containing 10 seconds of data. The 1conv training files of the NIST 2004 corpus (246 males and 370 females) are used as the background utterances for all three classifiers. To simplify system optimization and save processing time, we use the same background training set for all three corpora.

In evaluating our recognizer performance, we use two well-known metrics. The first one, *equal error rate* (EER), corresponds to the decision threshold that gives equal false acceptance rate (FAR) and false rejection rate (FRR). The second measure, referred to as *minimum detection cost function* (MinDCF), punishes heavily false acceptances. It is used in the NIST SRE evaluations and defined as the minimum value of the function $0.1 \times \text{FRR} + 0.99 \times \text{FAR}$.

3. OPTIMIZATION RESULTS: NIST 2005/2006

First, we study the N-top scoring algorithm for GMM-UBM and VQ-UBM, because we are not aware of a systematic study on the effect of N-top value to accuracy. In [1], it is stated that $N = 5$ top scoring components are enough. We hypothesized that for higher model orders, more Gaussians would be required for accurate recognition as the likelihood computation gets more accurate. On the other hand, VQ-UBM obtains *exactly the same result* as full search if the nearest code to the unknown vector is in the N-top list. This made us hypothesize that VQ-UBM may require a smaller value of N.

From the results displayed in Fig. 1, we make the following immediate observations. First, GMM-UBM is somewhat sensitive to the selection of N; the optimum value depends on both the objective function (EER, MinDCF) and the model size. VQ-UBM, on the other hand, is less sensitive to value of N; any value $N \geq 10$ minimizes both EER and MinDCF. Moreover, the result is fairly independent of the model order. The GMM-UBM and VQ-UBM have some similarities as well. In particular, both models achieve a small EER for "large" N and a small MinDCF for "small" N.

Does larger model require more N-top components as we hypothesized? According to the results shown here the answer is **no**. Even the opposite can happen. For instance, the MinDCF of the GMM-UBM increases with N for large model sizes. In other words, the more inaccurate the computation of the log-likelihood ratio, the better MinDCF! This is an indirect indication of sub-optimal speaker model density estimation, and possibly some other violations in the modeling assumptions. For the rest of the experiments, we fix the values $N = 10$ for the GMM-UBM and $N = 5$ for the VQ-UBM. These values were chosen to give a small EER with significant speed-up compared to full search.

Next, we present model order optimization results in Fig. 2 which displays EER against MinDCF. For the GMM-UBM and VQ-UBM classifiers, results are shown for different model orders M. The GLDS-SVM does not have similar control parameter and hence is presented by a single point.

The following observations can be made:

- Optimal model order depends on both the test condition and on the model type (GMM-UBM or VQ-UBM); for shorter

[1] http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

[2] http://niko.brummer.googlepages.com/focal

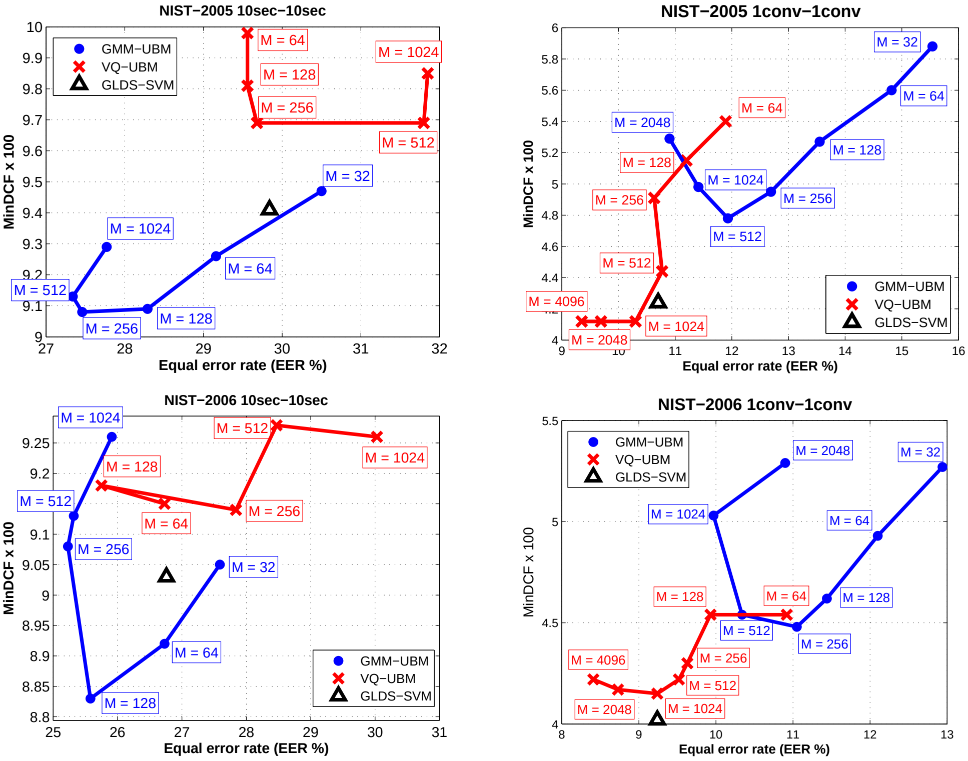

Fig. 2. Accuracy on the NIST 2005 and NIST 2006 corpora for the 10sec-10sec and 1conv-1conv tests.

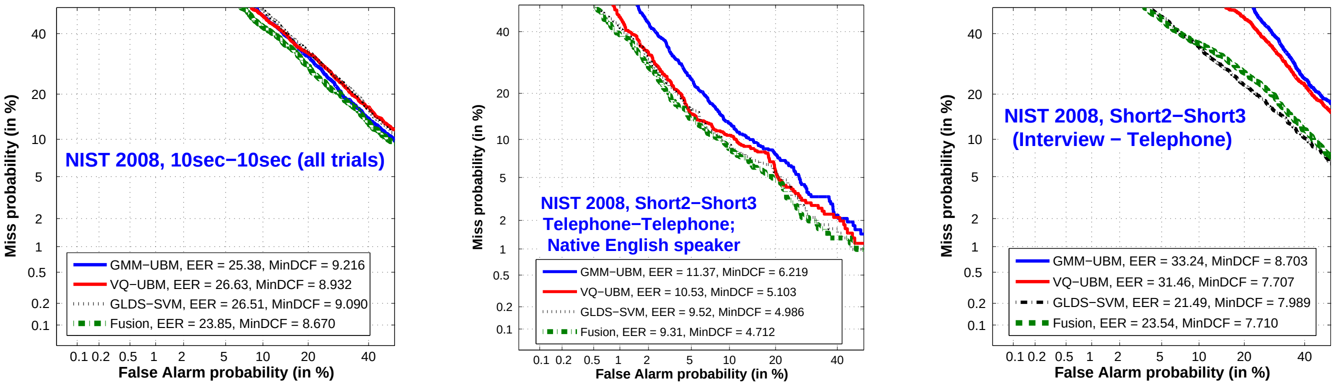

Fig. 3. Selected results on the NIST 2008 corpus.

data, the optimal model order is lower compared with longer data.

- GMM-UBM outperforms VQ-UBM for the short data condition (10sec-10sec), and vice versa, VQ-UBM outperforms GMM-UBM on the longer data condition (1conv-1conv).

- The GMM-UBM performance is consistent across the two corpora giving nicely convex error curves

- Accuracy of the GLDS-SVM lies in between the other two classifiers for the 10-second cases. It is comparative with VQ-UBM on the longer data (1conv-1conv). It also shows consistent (predictable) performance for the 1conv-1conv case across the two corpora

4. RESULTS ON THE NIST 2008 CORPUS

The optimized classifiers were then evaluated on the NIST 2008 data. The results shown here are based on our primary submission system to the NIST 2008 SRE campaign. Due to page limitations, only a few selected cases are shown.

The following model sizes were used for Short2-Short3 and 10sec-10sec cases, respectively: 512, 256 (GMM), and 2048, 128 (VQ). Selected results shown in Fig. 3 are two-fold. On one hand, for the 10sec-10sec test case, the observations made for NIST 2006 results generalize well to the 2008 corpus: GMM-UBM is the best individual classifier. Fusion of the three systems also slightly improves accuracy. The results for the short2-short3 telephone data is also consistent with those of NIST 2006 1conv-1conv case as GMM-UBM is still the worst. However, GLDS-SVM performs now slightly better than VQ-UBM.

On the other hand, the interview material does not exist in the NIST 2005 and 2006 data. The methods tuned for these corpora do therefore not apply well to the trials where interview data is present. The results with the worst channel mismatch (interview-telephone case), GLDS-SVM appears to be most robust.

Table 1. Comparing VADs for the VQ-UBM system (EER %).

	NIST VAD + Energy VAD	NIST VAD
interw. - interw.	30.42	21.52
interw. - interw., same mic	15.60	10.46
interw. - interw., diff mic	30.71	21.47
interw. - teleph.	31.45	24.72

During the NIST 2008 SRE, we also studied two alternative VADs for the interview data. The interview data contains data from the interviewee and interviewer recorded simultaneously with several microphones. NIST provided automatically generated speaking turn intervals for the interviewees (used as target speakers) to refrain participants solving the speaker segmentation problem. We call these indicators “NIST VAD”. In the second approach we apply an energy VAD on top of the NIST VAD. We hypothesized that the NIST VAD keeps too many non-speech frames that should be eliminated by an additional energy VAD. The results in Table 1 show that 5-10 % unit better recognition accuracy is achieved with the NIST VAD, which contradicts with our hypothesis. Post-evaluation analysis indicated that the two-stage VAD retains on average 43 % of speech frames whereas NIST VAD retained as many as 67 %. Moreover, the two-stage VAD generated nearly 600 files with less than 10 % of speech frames. To sum up, the two-stage VAD was too aggressive. One hypothesis is that the energy threshold optimized for the telephone data should be re-optimized for the interview data.

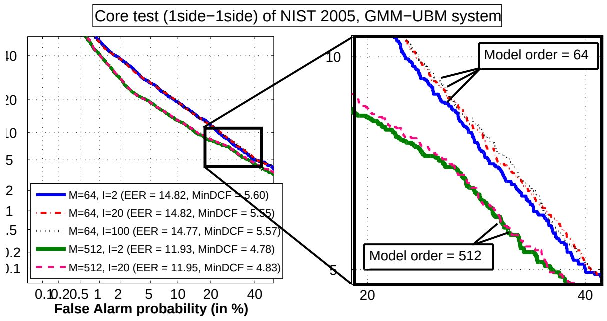

Fig. 4. Effect of EM iteration count in UBM training.

5. DISCUSSION

The observation that VQ-UBM outperformed GMM-UBM on the longer training and test data contradicts intuition and our initial hypothesis: since speaker models in the VQ-UBM approach have less parameters, one would expect it to perform better on short samples.

Are the differences between the GMM-UBM and VQ-UBM due to the inherent differences in the models themselves or just because of differences in their parameter settings? One may argue that, as we used only 2 EM iterations to train the background model for the GMM-UBM system and 20 K-means iterations for the VQ-UBM, the setting is unfair for GMM. To study this, we varied the number of EM iterations for the UBM training in the GMM-UBM system as a post-evaluation analysis. The results displayed in Fig. 4 clearly indicates that the number of EM iterations is an insignificant parameter compared to model order. In other words, maximum likelihood criterion training of the UBM is not optimal; if this was the case, further iterations would improve accuracy.

6. CONCLUSION

In this paper, we have experimentally compared VQ- and GMM-based speaker models trained using MAP criterion. The most surprising observation was that VQ-UBM gave better results for longer training and test segments whereas GMM-UBM was better for short segments. The differences seem not only due to parameter settings but in the model types themselves. It would be interesting to study the combination of the VQ-UBM and support vector machine as already done for the GMM-UBM by several authors [4, 10].

In this work, we purposely kept the recognizers simple enough and computationally efficient so that they could be implemented in a real-time platform. In future, the observations need to validated on more complete systems including JFA compensation [5] and T-norm [6].

7. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digit. Sign. Proc.*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[2] V. Hautamäki, T. Kinnunen, I. Kärkkäinen, M. Tuononen, J. Saastamoinen, and P. Fränti, “Maximum a posteriori estimation of the centroid model for speaker verification,” *Sign. Proc. Lett.*, vol. 15, pp. 162–165, 2008.

[3] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Comp. Speech & Lang.*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[4] K.A. Lee, C. You, H. Li, T. Kinnunen, and D. Zhu, “Characterizing speech utterances for speaker verification with sequence kernel SVM,” in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 1397–1400.

[5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE T. Audio, Speech & Lang. Proc.*, July 2008.

[6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, January 2000.

[7] J. He, L. Liu, and G. Palm, “A discriminative training algorithm for VQ-based speaker identification,” *IEEE T. Speech & Audio Proc.*, vol. 7, no. 3, pp. 353–356, May 1999.

[8] F.K. Soong, A.E. Rosenberg A.E., B.-H. Juang, and L.R. Rabiner, “A vector quantization approach to speaker recognition,” *AT & T Tech. J.*, vol. 66, pp. 14–26, 1987.

[9] P. David, “Experiments with speaker recognition using GMM,” in *Proc. of Radioelektronika 2002*, Bratislava, May 2002, pp. 353–357.

[10] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *Sign. Proc. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.