# QUANTIZATION FOR ADAPTED GMM-BASED SPEAKER VERIFICATION

*Ivy H. Tseng*[*†]*, Olivier Verscheure*[‡]*, Deepak S. Turaga*[‡]*, and Upendra V. Chaudhari*[‡]

[†]Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089
[‡]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

**ABSTRACT**

State-of-the-art speaker verification systems are built around the likelihood ratio test, using Gaussian Mixture Models (GMM) for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM. This work tackles optimal quantizer design of the speech cepstral features (MFCCs) for such systems. The problem is posed as the minimization of loss of log-likelihood ratio between the quantized and unquantized speech features. First we show that the conventional mean squared error (MSE) quantizer for the top-scoring UBM Gaussian is optimal under practical assumptions. Then we derive the optimal bit allocation strategy across the dimensions of the feature vectors. Finally we demonstrate the validity of the approach against various quantization and bit allocation schemes by running experiments on the appropriately modified IBM Speaker Verification system. Experimental results on the HUB4 corpora show negligible impact on verification performance for bit rates as low as less than 1 bit per dimension on average in contrast to 32 bits per dimension in the original system.

## 1. INTRODUCTION

With the growing prevalence of mobile devices, users are starting to expect a full range of computational as well as communication services from these devices. Furthermore, given that these devices are used in a wide variety of environments, by users with differing access and operational requirements, the need for non-keyboard based (hands-free) interfaces is becoming increasingly apparent. Recent advances have significantly improved the robustness and accuracy of speech recognition technology, making speech-based interfaces for human-computer (mobile device) interaction viable. Speech may also be used in order to enable secure access to the mobile device, through the use of voice authentication and speaker verification. Unfortunately, advanced speech recognition and speaker verification algorithms, especially for large vocabulary systems, under noisy environments, are computationally expensive, and cannot be easily implemented on these mobile handsets. There is hence a need for distributed processing, where the computation is shared between the device and the network infrastructure to provide these capabilities. Such a distributed processing approach can also allow for added levels of security, by limiting amount of valuable information (such as speaker models) stored on the actual mobile device. Among current distributed speech recognition (DSR) applications, the ETSI Aurora [1] is widely referenced as the standard over the mobile cellular network. The mobile device performs the relevant feature parameter extraction and compression, as well as bit-stream framing, formatting and decoding, and additional error protection and mitigation. It then transmits these processed features over a data channel to a remote back-end system, which further processes the received features and performs the speech recognition. The features extracted by the ETSI DSR application include 13 Mel-frequency cepstral coefficients (MFCC) as well as the logarithmic energy, extracted from each frame of the speech signal. The feature vector is compressed using Split Vector Quantization (SVQ). During this process the features are grouped into pairs, and each pair is quantized using its own codebook. It has been shown that accurate speech recognition performance can be achieved with a fairly low data rate of 4,800 bps for the quantized features. In this paper we focus on compression (using quantization) of speech data for a Distributed Speaker Verification (DSV) application that is based on adapted Gaussian Mixture Models (GMMs) for speakers. In particular, we design optimal quantization schemes, for the MFCC feature vectors, to minimize the loss in speaker verification accuracy. We rely on generative models and tailor our quantization towards minimizing a distortion metric (based on the log-likelihood ratio) specific to the speaker verification algorithm used. Using our analysis we show that the quantizer that minimizes the squared loss in log-likelihood ratio, may be mapped to a conventional weighted Mean Squared Error (MSE) quantizer for both a single-speaker as well as a multi-speaker verification task. Furthermore, we also investigate variable bit allocation across the different dimensions of the feature vector, and derive analytically the number of bits per feature dimension. Finally, we evaluate the designed quantization and bit allocation schemes on the HUB4 corpora, and show significant improvements in the achievable compression with negligible impact on the accuracy of the speaker verification. This paper is organized as follows. We describe the state of the art in speaker verification using GMMs in Section 2. We then describe the design of our Minimum Log-Likelihood Ratio Difference (MLLDR) quantizer in Section 3. We investigate bit allocation schemes in Section 4. We describe the system design in Section 5 and present experimental results in Section 6. Finally, we present our conclusions in Section 7.

## 2. SPEAKER VERIFICATION USING ADAPTED GAUSSIAN MIXTURE MODELS

Top-performing speaker verification systems are built around the likelihood ratio test, using diagonal-covariance Gaussian Mixture Models (GMM) for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM [2]. The UBM model represents the background population and is trained with data from a large number of speakers so as to create a model without idiosyncratic characteristics. The mean vectors of the speaker models are derived via MAP adaptation from the UBM parameters based on speaker-specific training data, whereas the mixture weights

*This work was done while at the IBM T.J. Watson Research Center. Contact author: hsinyits@sipi.usc.edu

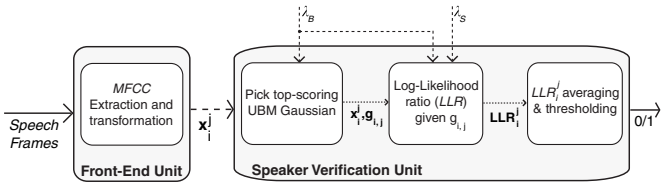


**Fig. 1**. Original speaker verification system [3]: Block diagram.

and diagonal covariance matrices are identical to the UBM's.

We briefly review the key features of a state-of-the-art speaker verification system [3]. The core components are illustrated by Figure 1. Mel-frequency cepstral coefficients (MFCC) are extracted and pre-processed from input speech frames. Let $X = \{x_i^j\}$ with $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, D$ denote a sequence of such $N$ mutually independent $D$-dimensional feature vectors. Those feature vectors feed the speaker verification unit. This unit is composed of three main parts. First the top-scoring D-dimensional Gaussian from the UBM Gaussian mixture $\lambda_B$ is identified. That is, the most likely UBM Gaussian given a feature element $x_i^j$ is picked. Let $g_{i,j}$ denote the index of this Gaussian. Then, given the feature element, the log-likelihood ratio is computed between the most likely UBM Gaussian and the corresponding Gaussian in the claimant speaker model $\lambda_S$. Please refer to Section 3 for a detailed analysis. Finally those log-likelihood ratio values are added over the entire test data of $N$ feature vectors. The result of this averaging operation $\Lambda(X)$ is compared to a decision threshold $\theta$ for accepting $(\Lambda(X) \geq \theta)$ or rejecting $(\Lambda(X) < \theta)$ the hypothesis that $X$ is indeed from the claimant speaker.

## 3. MINIMUM LOG-LIKELIHOOD RATIO DIFFERENCE (MLLDR) QUANTIZER

The goal of standard scalar quantization is to encode the data from a source, characterized by its probability density function (*pdf*), with the lowest possible rate and the smallest average distortion. The most common distortion measure is the squared error. The expected value of the distortion over the source distribution is the mean squared error between quantized and unquantized data. Standard quantizer design algorithms for this distortion measure iteratively compute encoder partitions based on the nearest neighbor condition and decoder reconstruction levels based on the centroid condition.However, given the verification task and the log-likelihood ratio test, we argue that the squared loss in log-likelihood ratio is a more appropriate distortion measure. That is, our quantizer must minimize $\left(\Lambda(X) - \Lambda(\widehat{X})\right)^2$ to minimize the impact on the verification task, where $\widehat{X}$ denotes the quantized version of $X$.

### 3.1. Simplifying the Log-Likelihood Ratio

Let $g_{i,j}$ denote the index of the top-scoring UBM Gaussian given a feature element $x_i^j$. Let $\mu_{S_{g_{i,j}}} = \mu_{B_{g_{i,j}}} + \delta_{g_{i,j}}$ denote the MAP-adapted mean of the $g_{i,j}th$ Gaussian from the speaker model. Recall that the weight coefficients $w_{g_{i,j}}$ and covariance matrices $\Sigma_{g_{i,j}}$ of the UBM mixture model and corresponding adapted speaker model are identical. Also, recall that the $D$ feature elements are mutually independent. Therefore the log-likelihood ratio $\Lambda(x_i^j)$ simplifies as

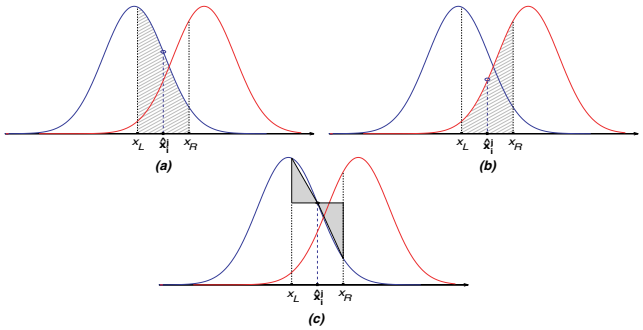


**Fig. 2**. Approximation of the log-likelihood ratio at the decoder.

follows:

$$\begin{aligned}
\Lambda(x_i^j) &= \log p(x_i^j|\lambda_S) - \log p(x_i^j|\lambda_B) \\
&= \log w_{g_{i,j}} - \frac{1}{2}\log 2\pi - \log \sigma_{g_{i,j}} - \frac{(x_i^j - \mu_{S_{g_{i,j}}})^2}{2\sigma_{g_{i,j}}^2} \\
&\quad - \log w_{g_{i,j}} + \frac{1}{2}\log 2\pi + \log \sigma_{g_{i,j}} + \frac{(x_i^j - \mu_{B_{g_{i,j}}})^2}{2\sigma_{g_{i,j}}^2} \\
&= \frac{1}{2\sigma_{g_{i,j}}^2}\left[(x_i^j - \mu_{B_{g_{i,j}}})^2 - (x_i^j - \mu_{S_{g_{i,j}}})^2\right] \\
&= \frac{\delta_{g_{i,j}}}{\sigma_{g_{i,j}}^2}\left[x_i^j - \frac{(\mu_{S_{g_{i,j}}} + \mu_{B_{g_{i,j}}})}{2}\right]
\end{aligned} \tag{1}$$

The decision is taken after $N$ feature vectors. Thus, the average log-likelihood ratio $\Lambda(X)$ can be computed as:

$$\Lambda(X) = \sum_{i=1}^{N}\sum_{j=1}^{D}\frac{\delta_{g_{i,j}}}{\sigma_{g_{i,j}}^2}\left[x_i^j - \frac{(\mu_{S_{g_{i,j}}} + \mu_{B_{g_{i,j}}})}{2}\right] \tag{2}$$

### 3.2. Computing the Log-Likelihood Ratio at the Decoder

Let $\widehat{x}_i^j$ denote the quantized value of $x_i^j$ received by the decoder. The statistics of $\widehat{x}_i^j$ are not continuous. Instead $\widehat{x}_i^j$ follows a discrete probability mass function. However we show that it is valid to use the original *pdf* at the decoder to compute $\Lambda(\widehat{x}_i^j)$. Consider indeed Figure 2. The quantization interval $[x_L, x_R]$ is fixed, the centroid of which is the reconstruction value $\widehat{x}_i^j$. Let $P(\widehat{x}_i^j|\lambda_S)$ and $P(\widehat{x}_i^j|\lambda_B)$ represent the diagonally striped areas on Figures 2a and 2b, respectively. Now consider Figure 2c. Through analysis, one can show that the two gray-shaded triangles have approximately the same area. Thus, $P(\widehat{x}_i^j|\lambda_S)$ approximately equals $p(\widehat{x}_i^j|\lambda_S)(x_R - x_L)$. Similarly, $P(\widehat{x}_i^j|\lambda_B) \approx p(\widehat{x}_i^j|\lambda_B)(x_R - x_L)$. Hence, we have:

$$\Lambda(\widehat{x}_i^j) = \log\frac{P(\widehat{x}_i^j|\lambda_S)}{P(\widehat{x}_i^j|\lambda_B)} \approx \log\frac{p(\widehat{x}_i^j|\lambda_S)}{p(\widehat{x}_i^j|\lambda_B)} \tag{3}$$

Equation 3 shows that using the original *pdf* to compute $\Lambda(\widehat{x}_i^j)$ at the decoder is a good first approximation. This approximation improves as the quantization bin size decreases (i.e. $|x_R - x_L| \mapsto 0$).

### 3.3. MLLDR Quantizer Design

From Equations 1, 2 and 3, the loss in log-likelihood ratio between an unquantized feature vector $x_i$ and its quantized version $\widehat{x}_i$ can be

written as:

$$\mathcal{L}_i = \sum_{j=1}^{D} \left( \Lambda(x_i^j) - \Lambda(\widehat{x}_i^j) \right) = \sum_{j=1}^{D} \frac{\delta_{g_{i,j}}}{\sigma^2_{g_{i,j}}} \left( x_i^j - \widehat{x}_i^j \right) \quad (4)$$

We wish to minimize the expected value of the squared loss $E\left[\mathcal{L}_i^2\right]$:

$$\begin{aligned} E\left[\mathcal{L}_i^2\right] &= E\left[ \left( \sum_{j=1}^{D} \frac{\delta_{g_{i,j}}}{\sigma^2_{g_{i,j}}} (x_i^j - \widehat{x}_i^j) \right)^2 \right] \\ &= E\left[ \sum_{j=1}^{D} \frac{\delta^2_{g_{i,j}}}{\sigma^4_{g_{i,j}}} (x_i^j - \widehat{x}_i^j)^2 \right] \\ &+ E\left[ \sum_{j=1}^{D} \sum_{k \neq j} \frac{\delta_{g_{i,j}} \delta_{g_{i,k}}}{\sigma^2_{g_{i,j}} \sigma^2_{g_{i,k}}} (x_i^j - \widehat{x}_i^j)(x_i^k - \widehat{x}_i^k) \right] \\ &= \sum_{j=1}^{D} \sum_{i=1}^{N} w_{g_{i,j}} \frac{\delta^2_{g_{i,j}}}{\sigma^4_{g_{i,j}}} E\left[ (x_i^j - \widehat{x}_i^j)^2 \right] \\ &+ \sum_{j=1}^{D} \sum_{k \neq j} \sum_{i=1}^{N} w_{g_{i,j}} \frac{\delta_{g_{i,j}} \delta_{g_{i,k}}}{\sigma^2_{g_{i,j}} \sigma^2_{g_{i,k}}} E\left[ x_i^j - \widehat{x}_i^j \right] E\left[ x_i^k - \widehat{x}_i^k \right] \end{aligned}$$

Note that the mean is usually not affected by quantization. The centroid of each interval is indeed chosen as the reconstruction value. Therefore, the second term goes to zero and we have:

$$E\left[\mathcal{L}_i^2\right] = \sum_{j=1}^{D} \sum_{i=1}^{N} w_{g_{i,j}} \frac{\delta^2_{g_{i,j}}}{\sigma^4_{g_{i,j}}} E\left[ (x_i^j - \widehat{x}_i^j)^2 \right] \quad (5)$$

Equation 5 shows that the quantizer that minimizes the squared loss in log-likelihood ratio for single-speaker verification simply is a conventional weighted *MSE* quantizer.

### 3.4. Quantization for Multiple Speakers

Speaker verification systems are usually set up so as to verify that the sequence of feature vectors $X$ is from a set of $M > 1$ claimant speakers. Thus, $M$ independent log-likelihood ratios are computed and averaged over time. Let $\mathcal{L}^m = \Lambda^m(X) - \Lambda^m(\widehat{X})$ denote the log-likelihood ratio difference for the $mth$ claimant speaker. Clearly, producing $M$ quantized versions of $X$ is not an admissible solution. Instead, we wish to find the quantizer that jointly minimizes the expected value of the log-likelihood ratio difference over all $M$ claimant speakers. Thus,

$$\begin{aligned} E\left[ \sum_{m=1}^{M} (\mathcal{L}_i^m)^2 \right] &= \sum_{m=1}^{M} E\left[ (\mathcal{L}_i^m)^2 \right] \\ &= \sum_{m=1}^{M} \sum_{j=1}^{D} \sum_{i=1}^{N} w_{g_{i,j}} \frac{\delta^2_{m,g_{i,j}}}{\sigma^4_{g_{i,j}}} E\left[ (x_i^j - \widehat{x}_i^j)^2 \right] \\ &= \sum_{j=1}^{D} \sum_{i=1}^{N} \frac{w_{g_{i,j}}}{\sigma^4_{g_{i,j}}} E\left[ (x_i^j - \widehat{x}_i^j)^2 \right] \sum_{m=1}^{M} \delta^2_{m,g_{i,j}} \\ &= M \sum_{j=1}^{D} \sum_{i=1}^{N} w_{g_{i,j}} \frac{\Delta_{g_{i,j}}}{\sigma^4_{g_{i,j}}} E\left[ (x_i^j - \widehat{x}_i^j)^2 \right], \end{aligned}$$

where $\Delta_{g_{i,j}} = \frac{1}{M} \sum_{m=1}^{M} \delta^2_{m,g_{i,j}}$. Thus, the quantizer that minimizes the squared loss in log-likelihood ratio for multi-speaker verification is also a weighted *MSE* quantizer.

## 4. VARIABLE BIT ALLOCATION

The objective is to determine the number of bits to devote to each element of a feature vector so as to minimize a cost function under a rate constraint. The rate constraint is expressed as the total number of bits $b_T$ per D-dimensional feature vector. Let $b_j$ denote the number of bits per feature dimension, and $b_c$ represent the number of bits required to code the index of the top-scoring D-dimensional Gaussian given a feature vector $x_i$. Therefore, the rate constraint is expressed as $b_q = b_T - b_c \geq \sum_{j=1}^{D} b_j$.

Let $D_j(b_j) = E\left[ \left( x_i^j - \widehat{x}_i^j \right)^2 \right]$ denote the *MSE* when $b_j$ bits are used to quantize $x_i^j$. Recall that $x_i^j$ is a random variable following a Normal distribution. Thus, for high-resolution, $D_j(b_j)$ can be expressed as:

$$D_j(b_j) = \frac{\sqrt{3}\pi}{2} \sigma_j^2 2^{-2b_j}$$

### 4.1. Single-Speaker Verification

We derive the number of bits to use by feature element $x_i^j$ given the top-scoring Gaussian $g_{i,j}$ by minimizing $\mathcal{L}_i^2$ under the rate constraint $\sum_{j=1}^{D} b_j \leq b_q$. We cast the optimization problem as the following Lagrange cost function:

$$\begin{aligned} C &= \sum_{j=1}^{D} \frac{\delta^2_{g_{i,j}}}{\sigma^4_{g_{i,j}}} D_j(b_j) + \lambda \left( \sum_{j=1}^{D} b_j - b_q \right) \\ &= \sum_{j=1}^{D} \frac{\delta^2_{g_{i,j}}}{\sigma^4_{g_{i,j}}} \frac{\sqrt{3}\pi}{2} \sigma^2_{g_{i,j}} 2^{-2b_j} + \lambda \left( \sum_{j=1}^{D} b_j - b_q \right) \end{aligned}$$

Equating its first derivative with respect to the bit assignment variable to zero and following the derivations in [4], we obtain the number of bits per dimension given the index of the top-scoring Gaussian $g_{i,j}$:

$$b^j_{g_{i,j}} = \frac{1}{D} b_q + \frac{1}{2} \left[ \frac{\delta^2_{g_{i,j}}}{\sigma^2_{g_{i,j}}} - \frac{1}{D} \sum_{k=1}^{D} \log_2 \left( \frac{\delta^2_{g_{i,k}}}{\sigma^2_{g_{i,k}}} \right) \right] \quad (6)$$

Implementation details are provided in Section 4.3.

### 4.2. Multi-Speaker Verification

Similarly but minimizing $\sum_{m=1}^{M} (\mathcal{L}_i^m)^2$ instead, we obtain:

$$b^j_{g_{i,j}} = \frac{1}{D} b_q + \frac{1}{2} \left[ \frac{\Delta_{g_{i,j}}}{\sigma^2_{g_{i,j}}} - \frac{1}{D} \sum_{k=1}^{D} \log_2 \left( \frac{\Delta_{g_{i,k}}}{\sigma^2_{g_{i,k}}} \right) \right] \quad (7)$$

where $\Delta_{g_{i,j}} = \frac{1}{M} \sum_{m=1}^{M} \delta^2_{m,g_{i,j}}$.

### 4.3. Bit Allocation: Implementation Details

Equations 6 and 7 do not guarantee that $b^j_{g_{i,j}}$ is a non-negative integer. Let $\beta = \left\{ b^1_{g_{i,1}}, \cdots, b^D_{g_{i,D}} \right\}$ denote the bit allocation vector given the input feature vector $x_i$. Enforcing the non-negative condition is known in the quantization literature. We enforce the integer condition by the following algorithm:

1. Calculate $\lfloor \beta \rfloor$ by rounding down every element of $\beta$.
2. Pick the top $\left( b_q - \sum_{j=1}^{D} \lfloor b^j_{g_{i,j}} \rfloor \right)$ elements from the vector $(\beta - \lfloor \beta \rfloor)$ and add one bit to those elements.

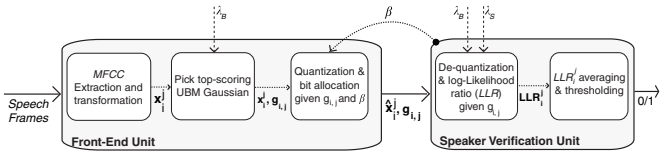


**Fig. 3**. Distributed Speaker Verification: Block diagram.

## 5. DISTRIBUTED SPEAKER VERIFICATION

Our Distributed Speaker Verification system is illustrated by Figure 3. Mel-frequency cepstral coefficients (MFCC) are extracted and pre-processed from input speech frames. The feature vectors are used to identity the top-scoring D-dimensional UBM Gaussian. The feature vector is then quantized using an MSE quantizer designed for the top-scoring Gaussian. Each feature element is coded on a number of bits given by $\beta$. The bit allocation vector $\beta$ is computed in the server given the set of claimant speakers and communicated to the front-end. Note that $\beta$ must be updated whenever the set of claimant speakers changes. The quantized feature vector together with the coded index of the top-scoring Gaussian is sent to the speaker verification unit, which de-quantizes the incoming vector and computes the average log-likelihood ratio to perform speaker verification.

## 6. EXPERIMENTS

We validate the efficacy of our approach on the appropriately modified state-of-the-art IBM Speaker Verification system [3].

### 6.1. Experimental Setup

The data consisted of the audio portion of the HUB4 Broadcast News Database (mono 16kHz PCM). A subset of 64 speakers was selected as the target speaker set. A feature vector consists of 19-dimensional MFCC and their first derivatives[1] with feature warping(i.e., $D = 38$). A rate of 100 frames per second, with $50\%$ overlap was used and the MFCC were computed over a 20 millisecond window. For each speaker, two minutes of data were set aside and used for training the final models. The UBM, trained on independent broadcast news data, contained 256 38-dimensional Gaussian components. The speaker models, being MAP-adapted from the UBM, also had 256 components. For each speaker, 30 seconds (i.e., $N = 3000$ feature vectors) were used for testing performance.

### 6.2. Experimental Results

We compare the verification performance of the original system (unquantized features; 32 bits per feature element) against two quantization methods: (i) Our quantizer with variable bit allocation, and (ii) Conventional MSE quantizer designed from the UBM model with uniform bit allocation. Note that the latter approach does not require to pick the top-scoring Gaussian before quantization. Thus, we assign the extra $b_c = 8$ bits to the best feature elements. That is, those feature elements with the highest $\frac{\delta_{g_{i,j}}}{\sigma^2_{g_{i,j}}}$ ratio.

Experiments were run for various rate constraints. Figure 4 shows the single-speaker verification performance for a *1:32* compression ratio (i.e., 1 bit per dimension, on average). Our quantization method performs extremely well considering that 1 bit per feature element translates into being left/right of the corresponding Gaussian's mean. One may suspect that most dimensions are useless such that most feature elements are simply skipped, allowing non-skipped feature elements to use a relatively high number of bits. Actually, most feature elements are indeed coded and only a few feature elements use a number of bits greater than one (maximum of 3 bits). Finally, note that the performance of the conventional MSE quantizer is bad (increases the Equal Error Rate, or EER, by more than $18\%$). Similar conclusions were drawn for loser rate constraints and multi-speaker verification performance.

[1]Quantizing the 19 MFCC elements only and computing their derivatives in the server is beyond the scope of this work but will soon be investigated.

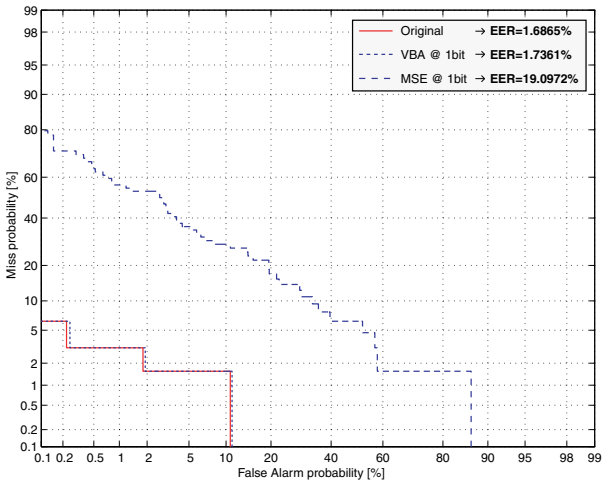


**Fig. 4**. Single-Speaker Verification on HUB4 Corpora: Performance of unquantized (original; 32 bits per dimension) and quantized feature vectors (our method & conventional MSE; 1 bit per dimension).

## 7. CONCLUSIONS

We designed an optimized quantization and variable bit allocation strategy for speaker verification systems based on adapted Gaussian mixture models. Our quantization strategy minimizes the squared loss in log-likelihood ratio, and was shown to trivialize to a conventional weighted *MSE* quantizer. A bit allocation vector was also derived analytically. Finally, we validated the efficacy of our approach on a modified version of the IBM Speaker Verification system. Results showed significant improvements in the achievable compression with negligible impact on verification accuracy.

## 8. REFERENCES

[1] ETSI ES 201 108 V1.2.2, "Speech Processing, Transmission and Quality Aspects (stq); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms," Tech. Rep., ETSI, 2000.

[2] D. A. Reynolds and T. F. Quatieri and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, 2000.

[3] G. N. Ramaswamy, J. Navratil, U. V. Chaudhari and R. D. Zilca, "The IBM System for the NIST 2002 Cellular Speaker Verification Evaluation," in *IEEE ICASSP*, Hong-Kong, April 2003.

[4] A. D. Subramaniam and B. D. Rao, "PDF Optimized Parametric Vector Quantization of Speech Line Spectral Frequencies," *IEEE Trans. on Speech & Audio Processing*, vol. 11, no. 2, 2003.