Channel Robust Speaker Verification via Bayesian Blind Stochastic Feature Transformation

Kwok-Kwong Yiu, Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA

Abstract

In telephone-based speaker verification, the channel conditions can be varied significantly from sessions to sessions. Therefore, it is desirable to estimate the channel conditions online and compensate the acoustic distortion without prior knowledge of the channel characteristics. Because no a priori knowledge is used, the estimation accuracy depends greatly on the length of the verification utterances. This paper extends the Blind Stochastic Feature Transformation (BSFT) algorithm that we recently proposed to handle the short-utterance scenario. The idea is to estimate a set of prior transformation parameters from a development set in which a wide variety of channel conditions exists in the verification utterances. The prior transformations are then incorporated into the online estimation of the BSFT parameters in a Bayesian (maximum a posteriori) fashion. The resulting transformation parameters are therefore dependent on both the prior transformations and the verification utterances. For short (long) utterances, the prior transformations play a more (less) important role. We referred the extended algorithm to as Bayesian BSFT (BBSFT) and applied it to the 2001 NIST SRE task. Results show that Bayesian BSFT outperforms BSFT for utterances shorter than or equal to 4 seconds.

1. Introduction

The acoustic mismatch between the training and recognition conditions can significantly reduce the accuracy of speaker recognition systems, and transducer variability has been shown to be the main cause of acoustic mismatch. Transducer variability occurs when a system is trained with speech data obtained from one type of transducer and is subsequently tested on speech data recorded from other types of transducers. Channel compensation is one of the possible approaches to minimizing the effect of transducer variability.

Channel compensation can be applied in feature space [1], [2], model space [3], [4], or score space [5]. Channel compensation can also be supervised or unsupervised. Supervised compensation assumes that the channel or handset characteristics are known a priori [1]. On the other hand, unsupervised compensation does not assume any knowledge of the channel characteristics.

We have recently proposed an unsupervised feature-based transformation approach, namely blind stochastic feature transformation (BSFT) [6], to address the acoustic mismatch problem. Specifically, feature-based transformations are estimated based on the statistical difference between test utterances and a composite GMM formed by combining the speaker and background GMMs. The transformations are then used to transform the test utterances before verification.

Because the BSFT algorithm estimates the transformation parameters based on the verification utterances and the clean acoustic models only, the accuracy of the transformation parameters depends on the length of the verification utterances. For short utterances, the algorithm may transform the distorted features to an undesirable region, resulting in incorrect verification decisions.

This paper extends the BSFT algorithm to handle the short-utterance scenario. The idea is to estimate a set of prior transformation parameters from a development set in which a wide variety of channel conditions exists in the verification utterances. The prior transformations are then used to derive the hyperparameters that govern the prior distribution of the transformation parameters. The prior distribution is incorporated into the online estiamtion of the BSFT parameters via the maximum a posteriori (MAP) criterion. We refer the extended algorithm to as Bayesian blind stochastic feature transformation (BBSFT). Because the prior distribution reflects the acoustic mismatches among different channels (e.g., handsets) in the development set, BBSFT is able to take the common mismatches into account during the estimation of transformation parameters. This ability is especially important for short verification utterances because short utterances may not contain sufficient channel information for an accurate estimation of the transformation parameters.

We compared the performance of BBSFT against BSFT by extracting short segments of verification utterances from the 2001 NIST SRE corpus. Experimental results suggest that BBSFT performs better than BSFT for short utterances.

2. Bayesian Blind Stochastic Feature Transformation

The blind stochastic feature transformation (BSFT) proposed in [6] is an unsupervised (blind) approach to channel mismatch compensation. The transformation is blind in that the transformation parameters are determined from a clean acoustic model and the distorted speech features derived from a claimant without any prior knowledge about the channel. Specifically, given a D-dimensional distorted vector $\mathbf{y}$, the transformed feature vector is

$$\hat{\mathbf{x}} = f_\nu(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \quad (1)$$

where $A = \text{diag}\,\{a_1, \ldots, a_D\}$ is a transformation matrix, $\mathbf{b} = [b_1, \ldots, b_D]^T$ represents a bias vector, $\nu = \{a_i, b_i\}_{i=1}^D$ is the set of transformation parameters, and $f_\nu(\cdot)$ denotes the transformation function. Intuitively, the bias $\mathbf{b}$ compensates the convolutive distortion and the matrix A compensates the effects of

This work was supported by the Research Grants Council of Hong Kong SAR (Project No. PolyU 5214/04E and CUHK1/02C).

noise.

Given a compact speech model $\Lambda = \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ (typically $M = 64$) derived from the clean speech of several speakers and distorted features $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ extracted from a verification utterance, the maximum a posteriori estimates of $\nu = \{A, \mathbf{b}\}$ can be obtained by

$$\begin{aligned} \nu_{\text{MAP}} &= \arg\max_{\nu} p(\nu|\Lambda, Y) \\ &= \arg\max_{\nu} p(Y|\nu, \Lambda)p(\nu). \end{aligned}$$

This is equivalent to maximizing the auxiliary function

$$\begin{aligned} Q(\nu'|\nu) &= Q(A', \mathbf{b}'|A, \mathbf{b}) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) \log\left\{\frac{p(f_{\nu'}(\mathbf{y}_t)|\mu_j, \Sigma_j)}{|J_{\nu'}(\mathbf{y}_t)|}\right\} \\ &\quad + \log p(A')p(\mathbf{b}') \qquad (2) \end{aligned}$$

with respect to ν'. In Eq. 2, ν' and ν represent the new and current estimates of the transformation parameters, respectively; T is the number of distorted vectors; $\nu' = \{a_i', b_i'\}_{i=1}^{D}$ denotes the transformation; $|J_{\nu'}(\mathbf{y}_t)|$ is the determinant of the Jacobian matrix, the (r, s)-th entry of which is given by $J_{\nu'}(\mathbf{y}_t)_{rs} = \partial f_{\nu'}(\mathbf{y}_t)_r / \partial y_{t,s}$; and $h_j(f_\nu(\mathbf{y}_t))$ is the posterior probability

$$h_j(f_\nu(\mathbf{y}_t)) = P(j|\mathbf{y}_t, \Lambda, \nu) = \frac{\pi_j p(f_\nu(\mathbf{y}_t)|\mu_j, \Sigma_j)}{\sum_{l=1}^{M} \pi_l p(f_\nu(\mathbf{y}_t)|\mu_l, \Sigma_l)}, \qquad (3)$$

where

$$\begin{aligned} p(f_\nu(\mathbf{y}_t)|\mu_j, \Sigma_j) &= (2\pi)^{-\frac{D}{2}} |\Sigma_j|^{-\frac{1}{2}} \cdot \\ \exp\Big\{ -\frac{1}{2}(f_\nu(\mathbf{y}_t) - \mu_j)^T \Sigma_j^{-1} (f_\nu(\mathbf{y}_t) - \mu_j)\Big\}. \qquad (4) \end{aligned}$$

In Bayesian BSFT, the transformation parameters are modeled by probability denisty functions, which are characterized by some hyperparameters. Assume that $\mathbf{b}$ follows a Gaussian distribution with mean vector $\beta = [\beta_1, \ldots, \beta_D]^T$ and precision matrix $\Gamma = \text{diag}\,\{\gamma_1^2, \ldots, \gamma_D^2\}$:

$$p(\mathbf{b}) = (2\pi)^{-\frac{D}{2}} \cdot \left\{\prod_{i=1}^{D} \gamma_i\right\} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{D}(b_i - \beta_i)^2 \gamma_i^2\right\}. \qquad (5)$$

Assume also that A follows a matrix variate normal distribution [7]

$$p(A) = \frac{|\Omega|^{-\frac{(D+1)}{2}}}{|\Phi|^{\frac{D}{2}}} \cdot \exp\Big\{ -\frac{1}{2}\text{tr}(A - \Upsilon)^T \Omega^{-1} (A - \Upsilon)\Phi^{-1}\Big\}, \qquad (6)$$

where $\Omega = \text{diag}\,\{\omega_1, \ldots, \omega_D\}$, $\Phi = \text{diag}\,\{1, \ldots, 1\}$ and $\Upsilon = \text{diag}\,\{\upsilon_1, \ldots, \upsilon_D\}$. The hyperparameters Υ and Ω can be obtained by

$$\Upsilon = \frac{1}{K}\sum_{k=1}^{K} A_k \qquad (7)$$

and

$$\Omega = \frac{1}{K}\sum_{k=1}^{K}(A_k - \Upsilon)(A_k - \Upsilon)^T, \qquad (8)$$

where A_k is the k-th transformation matrix estimated from training data and K is the number of transformation matrices (see Section 3 for details).

Ignoring the terms independent of ν' and assuming diagonal covariance (i.e., $\Sigma_j = \text{diag}\,\{\sigma_{j1}^2, \ldots, \sigma_{jD}^2\}$, Eq. 2 can be written as

$$\begin{aligned} Q(\nu'|\nu) &= \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t))\Big\{ -\frac{1}{2}\sum_{i=1}^{D}\frac{(a_i' y_{ti} + b_i' - \mu_{ji})^2}{\sigma_{ji}^2} \\ &\quad + \sum_{i=1}^{D}\log(a_i')\Big\} + \log p(A')p(\mathbf{b}'). \qquad (9) \end{aligned}$$

The maximum-a-posteriori estimates of ν can be found by the EM algorithm as follows. In the E-step, Eqs. 3 and 4 are used to compute $h_j(f_\nu(\mathbf{y}_t))$; then in the M-step, ν' is obtained by solving $\partial Q(\nu'|\nu)\partial\nu' = 0$. More specifically, we solve

$$\begin{aligned} \frac{\partial Q(\nu'|\nu)}{\partial a_i'} &= \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t))\Big\{ -\frac{y_{ti}(a_i' y_{ti} + b_i' - \mu_{ji})}{\sigma_{ji}^2} \\ &\quad + \frac{1}{a_i'}\Big\} - \frac{a_i' - \upsilon_i}{\omega_i} \qquad (10) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \frac{\partial Q(\nu'|\nu)}{\partial b_i'} &= \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t))\left\{-\frac{(a_i' y_{ti} + b_i' - \mu_{ji})}{\sigma_{ji}^2}\right\} \\ &\quad -(b_i' - \beta_i)\gamma_i^2 \qquad (11) \\ &= 0, \end{aligned}$$

which lead to

$$b_i' = \frac{p_i - q_i a_i' + \beta_i \gamma_i^2}{r_i + \gamma_i^2}, \qquad (12)$$

and

$$\begin{aligned} &\Big[1 + \omega_i s_i - \frac{\omega_i q_i^2}{r_i + \gamma_i^2}\Big] a_i'^2 + \\ &\Big[\frac{\omega_i q_i (p_i + \beta_i \gamma_i^2)}{r_i + \gamma_i^2} - \omega_i u_i - \upsilon_i\Big] a_i' - \omega_i T = 0, \qquad (13) \end{aligned}$$

where

$$p_i = \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) \mu_{ji} \sigma_{ji}^{-2}, \qquad (14)$$

$$q_i = \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) y_{ti} \sigma_{ji}^{-2}, \qquad (15)$$

$$r_i = \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) \sigma_{ji}^{-2}, \qquad (16)$$

$$s_i = \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) y_{ti}^2 \sigma_{ji}^{-2}, \text{and} \qquad (17)$$

$$u_i = \sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t)) \mu_{ji} y_{ti} \sigma_{ji}^{-2}. \qquad (18)$$

These E- and M-steps are repeated until $Q(\nu'|\nu)$ ceases to increase. The most significant improvement occurs during the first 5 iterations. Note that when there is a large variation in A and $\mathbf{b}$, then $\omega_i \rightarrow \infty$ and $\gamma_i \rightarrow 0$. This means that the terms without ω_i in Eq. 13 can be obmitted and Eqs. 12 and 13 reduce to

$$b'_i \quad = \quad \frac{p_i - q_i a'_i}{r_i}, \qquad (19)$$

and

$$\left[s_i - \frac{q_i^2}{r_i}\right] a_i'^2 + \left[\frac{q_i p_i}{r_i} - u_i\right] a'_i - T \quad = \quad 0. \qquad (20)$$

Eqs. 19 and 20 are the maximum-likelihood estimate of $\mathbf{b}$ and A, respectively (see [8], pp. 309).

3. Experiments

The Bayesian BSFT was applied to the one-speaker detection task specified in the 2001 NIST speaker recognition evaluation set [9]. During training, a 1024-component GMM-UBM was trained using the training utterances of all 60 speakers in the development set of the corpus. Then, for each target speaker in the evaluation set, a speaker-dependent GMM was created by adapting the UBM using MAP adaptation [10].

The development set of the 2001 NIST corpus was also used to estimate the hyperparameters of BBSFT and Znorm parameters. Specifically, for each speaker in the development set, a 1024-center speaker model is created by adapting the UBM using MAP adaptation. Then, for the k-th verification trial in the development set, transformation parameters $\nu_k = \{A_k, \mathbf{b}_k\}$ were estimated using BSFT (Eqs. 19 and 20) and the corresponding target-speaker model. The K sets of transformation parameters $\{\nu_k\}_{k=1}^{K}$ were then used to determine the hyperparameters (Eqs. 7 and 8), where K is the number of verification trials in the development set. For each speaker model in the evaluation set of the corpus, a set of Znorm parameters (mean and standard derivation of impostor scores) [11] was determined by presenting all of the impostor utterances in the development set to the speaker model and the UBM. The impostor utterances were transformed by either CMS, BSFT, or BBSFT. Therefore, the Z-norm parameters depend on the transformation method being used. Specifically, each target speaker has three sets of Z-norm parameters, one for CMS, one for BSFT and the another one for BBSFT.

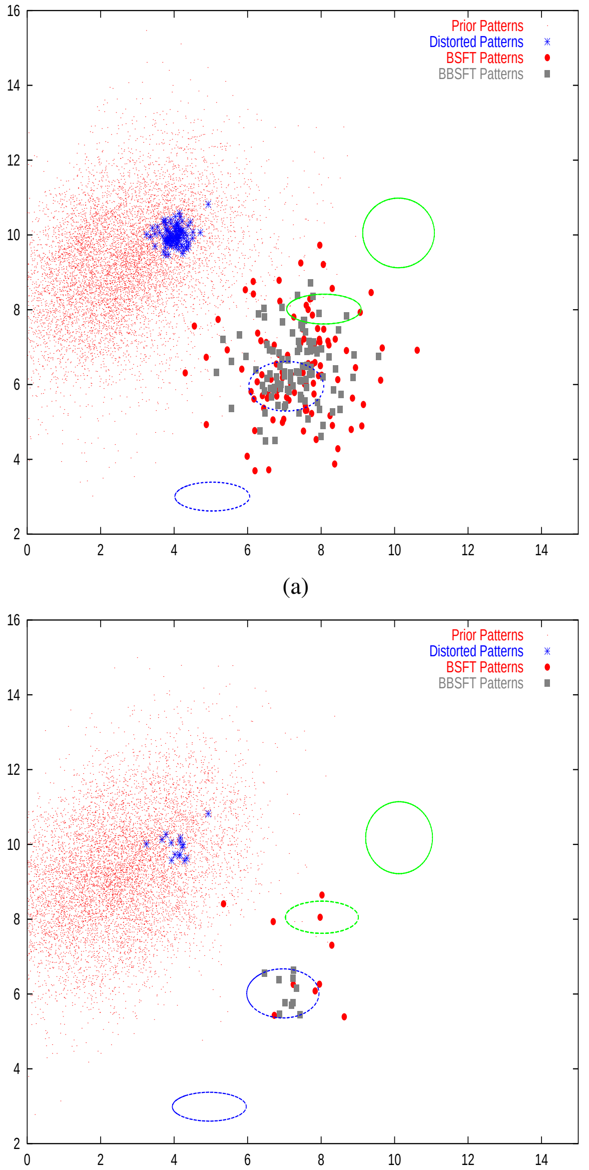
During verification, the feature sequence Y obtained from a claimant was transformed by the feature transformation parameters $f_\nu(\cdot)$ to form a sequence of transformed vectors Y_ν. The transformed vectors were then fed to a 1024-center GMM speaker model (Λ_s^N) and the 1024-center UBM (Λ_b^N) to obtain the score

$$S(Y_\nu) = \log p(Y_\nu|\Lambda_s^N) - \log p(Y_\nu|\Lambda_b^N),$$

which was further normalized by the Znorm parameters. The resulting score was compared with a global, speaker-independent threshold for decision making. In this work, the threshold was adjusted to determine an equal error rate (EER).

To compare the performance of BSFT and BBSFT, we randomly extracted short speech segments from the original verification utterances to form 5 sets of testing utterances. The 5 sets consist of speech data of duration 1, 2, 4, 8, and 16 seconds.

Mel-frequency cepstral coefficients (MFCCs) and their first-order derivatives were computed every 14ms using a Hamming window of 28ms. Cepstral mean subtraction (CMS) [12] was applied to the MFCCs to remove linear channel effects. The MFCCs and delta MFCCs were concatenated to form 24-dimensional feature vectors.



(a)

(b)

Figure 1: *A two-dimensional hypothetical problem illustrating the idea of BSFT and BBSFT for (a) 100 distorted patterns and (b) 10 distorted patterns. The red dots and blue asterisk represent the prior patterns and the distorted patterns, respectively. The blue and green ellipses represent the Gaussian mixtures in the clean speaker model and the clean impostor model, respectively. Both models have two mixtures, i.e $M = 2$. The figures show that patterns transformed by BSFT (red circles) were scattered over a large region of the feature space, whereas those transformed by BBSFT (grey squares) were confined to the regions occupied by the speaker and background models. The prior information becomes important when the number of distorted patterns is small.*

Verification Utterance Length (sec.)	**EER (%)**			**Min. DCF**		
	CMS	**BSFT**	**Bayesian BSFT**	**CMS**	**BSFT**	**Bayesian BSFT**
1	24.87	23.19	22.96	0.0875	0.0855	0.0849
2	19.83	18.08	17.73	0.0765	0.0707	0.0689
4	14.96	13.18	12.98	0.0611	0.0555	0.0549
8	12.05	10.43	10.65	0.0518	0.0448	0.0445
16	11.33	9.21	9.34	0.0454	0.0388	0.0383
Whole Utterance	10.79	8.91	9.05	0.0435	0.0413	0.0409

Table 1: *Equal error rates (in %) and minimum decision cost achieved by cepstral mean subtraction (CMS), blind stochastic feature transformation (BSFT) and Bayesian blind stochastic feature transformation (BBSFT). The scores in the three methods were normalized by their respective Z-norm parameters.*

4. Results and Discussions

Table 1 shows the equal error rates and minimum decision cost for blind stochastic feature transformation (BSFT) and Bayesian blind stochastic feature transformation (BBSFT), where the number of components M in Eq. 2 was set to 64.[1] Evidently, when testing data is limited (i.e., between 1 sec. and 8 sec.), all cases of BBSFT show reduction in error rates when compared to BSFT. On the other hand, when testing data is sufficient (e.g., 16 sec.), BSFT achieves better performance than BBSFT because the hyperparameters impose excessive constraint on the transformation parameters when sufficient testing data are available.

Fig. 2 shows the DET curves for CMS, BSFT and BBSFT. Testing data were limited to 2 second. The DET curves show that BBSFT performs better than BSFT at most of the operating points.

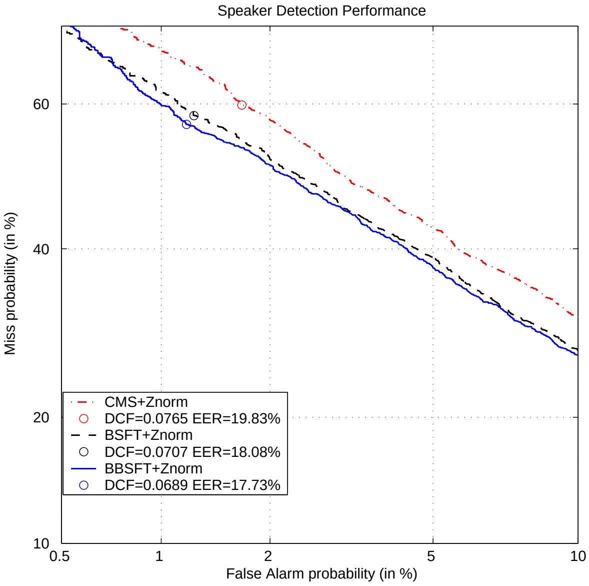


Figure 2: *DET curves comparing speaker verification performance using cepstral mean subtraction (CMS, dashdot curve), blind stochastic feature transformation (BSFT, dashed curve) and Bayesian blind stochastic feature transformation (BBSFT, solid curve). The circles represent the errors at which minimum DCF occurs. Each verification utterance contains two second of speech.*

5. Conclusions

We have extended the Blind Stochastic Feature Transformation (BSFT) algorithm to handle the short-utterance scenario during speaker verification. The proposed algorithm, Bayesian BSFT, estimates a set of prior transformation parameters from a development set containing a wide variety of channel conditions. The prior transformations are incorporated into the online estimation of the BSFT parameters by maximizing the a posteriori probability of the transformation parameters. Experimental results based on the 2001 NIST SRE task show that Bayesian BSFT outperforms BSFT for utterances shorter than or equal to 4 seconds.

6. References

[1] M. W. Mak, C. L. Tsang, and S. Y. Kung, "Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification," *EURASIP J. on Applied Signal Processing*, vol. 4, pp. 452–465, 2004.

[2] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *ICASSP'03*, 2003, vol. 2, pp. 53–56.

[3] K. K. Yiu, M. W. Mak, and S. Y. Kung, "Environment adaptation for robust speaker verification," in *Eurospeech'03*, 2003, pp. 2973–2976.

[4] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *ICSLP'00*, 2000, vol. 2, pp. 495–498.

[5] D. A. Reynolds, "Comparison of background normalization methods for text independent speaker verification," in *Eurospeech'97*, 1997, pp. 963–966.

[6] K.K. Yiu, M.W. Mak, and S.Y. Kung, "Blind stochastic feature transformation for channel robust speaker verification," *J. of VLSI Signal Processing*, to appear.

[7] A. K. Gupta and T. Varga, *Elliptically Contoured Models in Statistics*, Kluver Academic Publishers, 1993.

[8] S. Y. Kung, M. W. Mak, and S. H. Lin, *Biometric Authentication: A Machine Learning Approach*, Prentice Hall, 2005.

[9] "The NIST year 2001 speaker recognition evaluation plan," in *http://www.nist.gov/speech/tests/spk/2001/doc*.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[11] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[12] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, 1981.

[1] Theoretically, the larger the value of M, the better the results. However, setting M larger than 64 will result in unacceptably long verification time.