Biometric security based on ECG

Biometric Security Based on ECG

Lingni Ma, J.A. de Groot and Jean-Paul Linnartz
Eindhoven University of Technology
Signal Processing Systems, Electrical Engineering
l.ma.1@student.tue.nl j.a.d.groot@tue.nl j.p.m.g.linnartz@tue.nl

***Abstract*—Recently the electrocardiogram (ECG) has been proposed as a novel biometric. This paper aims to construct a reliable ECG verification system, in terms of privacy protection. To this end, an improved expression to estimate the capacity in the autocorrelation (AC) of the ECG is derived, which not only gives an upper bound for the number of identifiable subjects using ECG, but also helps to determine the number of independent features needed for verification. This capacity derivation, to our best knowledge, is new and can be easily extended to other biometrics. In addition, we apply the Quantization Index Modulation (QIM) scheme to the verification system for template protection. Based on our database, we found that, the capacity in the AC of ECG is approximately 20 bits information contained in the 23 independent features. We have also shown that compared to unprotected systems, the QIM scheme greatly enhances privacy protection of ECG templates without significant sacrificing of the identification performance.**

I. Introduction

The electrocardiogram (ECG) is a measure of electrical activities of the heart, which provides wealthy information of cardiac features. Despite the resemblance of ECG wave patterns among people, it has been shown that the ECG signal demonstrates distinctiveness and long-term stability. Since the heart activities are difficult to disguise, in recent years, the ECG has been proposed as a promising biometric for human identification and authentication.

In literature, the validity to use ECG for human identification has been explored in [1] [2]. The method used in these papers relies the biometric feature extraction heavily on fiducial detectors, which measure the geometrical properties of ECG waveforms. However, since there is no universally acknowledged rule to uniquely define the measurement, this method is less attractive for identification systems. A alternative approach to extract biometric features has been proposed in [3] [4], which performs the Discrete Cosine Transform (DCT) on the autocorrelation (AC) coefficients of ECG segments. A considerably high human identification rate has been reported using this method.

As with any other biometric, template protection is an important task in practical ECG identification systems. The templates stored in the public database require proper protection, in order to prevent the misuse of biometric privacy. Since individual ECG recordings are not exactly reproducible, it increases the difficulty for cryptography. To solve this problem, Linnartz *et al.* [5] introduced a shielding function with δ-contracting and ϵ-revealing properties. In addition, they also proposed the Quantization Index Modulation (QIM) scheme to center biometric measurements on the quantization interval with helper data. Based on the same concept of shielding functions, Tuyls *et al.* [6] suggested a Reliable Components Scheme (RCS), in which two types of helper data are used for noise reduction and error correction. In the above protection schemes, the helper data are always somehow correlated to the secrets they conceal. As a result, they leak certain amount of information. Zero leakage can be achieved with the method proposed in [7] [8], which uses a pre-distortion function to yield uniformly distributed secrets.

In previous researches, protection schemes are tested merely with fingerprints, face and iris. In contrast to these biometrics, the ECG has a much smaller number of independent features. Furthermore, the protection issue of the ECG templates has not been explored. Another interesting question is how much capacity is contained in a biometric and how many features are required for a verification system. However, there is no simple method to estimate ECG capacity from a practical set of measurements yet.

In this paper, the ECG template protection for authentication systems is discussed. In general, we have made two contributions. First, we have proposed an improved expression to estimate the capacity in the autocorrelation of ECG data. This derivation of capacity, to our best knowledge is new and can be easily extended to other biometrics. Second, we have applied the QIM protection

scheme to construct an ECG authentication system, which is not only more reliable in terms of privacy protection, but it also maintains a good identification performance.

II. Methods

In this paper, we consider a biometric verification system with three stages: signal preprocessing, feature extraction and subject classification. The ECG capacity is calculated both upon the extracted features and based on a verification system. In our research, autocorrelation and DCT are performed to extract biometric features. Thereby, we refer the capacity in the autocorrelation of ECG, shortly as ECG capacity in the rest of text. Note that the DCT does not lead to information loss. In addition, we also refer the capacity measured in an authentication system as the achieved ECG capacity.

A. Data Collection and Preprocessing

The preprocessing of the ECG signal is essential for noise reduction and artifacts detection. The energy of ECG mostly concentrates within a frequency band of 2Hz to 30Hz. The raw ECG recording is always contaminated by baseline wander and power line interference, which can be considered as irrelevant signals. Therefore we apply a 4th order Butterworth bandpass filter with cutoff frequencies of 1Hz and 40Hz for noise reduction. The automatic ECG artifacts detection requires sophisticated techniques that is not only outside the scope of this paper, but also is no more reliable than visual inspection [9]. Therefore, to detect artifacts, we segment the filtered ECG signal into short epochs and exclude the contaminated ones based on visual inspection. The remaining segments are assumed to be 'clean' and passed on to further analysis. This preprocessing procedure is common practice in ECG the analysis.

B. Feature Extraction

For feature extraction, the method using the Autocorrelation (AC) and the Discrete Cosine Transform (DCT) is proposed in [3] [4]. This approach first performs AC on a windowed ECG signal, which measures the similarity within ECG patterns as a function of time and blends all samples into a sequence of the averaged products of shifted version with itself. By doing so, the requirement for ECG pulse synchronization and fiducial vectors measurement are released. However, the AC coefficients are highly correlated. Therefore, DCT is applied to decorrelate these coefficients, which reduces the m AC coefficients to k most significant DCT coefficients. These k independent coefficients contain most ECG information and thereby are selected as ECG features. The calculation of AC and DCT coefficients, denoted as $\tilde{R}_{xx}(m)$ and $C(n)$, respectively, is specified by

$$\tilde{R}_{xx}(m) = \frac{R_{xx}(m)}{R_{xx}(0)} = \frac{\sum_{i=0}^{N-|m|-1} x(i)x(i+m)}{R_{xx}(0)}, \quad (1)$$

$$C(n) = \alpha(n) \sum_{m=0}^{M-1} \tilde{R}_{xx}(m) \cos\left(\frac{(2m+1)n\pi}{2M}\right), \quad (2)$$

where variable N is the length of the windowed ECG epoch, variable M is the length of $\tilde{R}_{xx}(m)$. The variable $\alpha(n)$ is the normalizing factor defined by

$$\alpha(n) = \begin{cases} \sqrt{\frac{1}{N}} & \text{if } n = 0 \\ \sqrt{\frac{2}{N}} & \text{otherwise} \end{cases}.$$

C. ECG Capacity Analysis

Biometric systems generally consist of two phases, enrollment and verification. To calculate the capacity of a biometric, we modeled the system as a communication channel where one tries to send information from the enrollment to the verification [10]. Such a model is depicted in Fig. 1. In the model, $\boldsymbol{\xi}_i$ is the biometric data of the ith subject, which is assumed to be independent and identically distributed (*i.i.d*) Gaussian variables. The biometric vector is transformed by matrix W to form the feature vector $\mathbf{X}_i \in \mathbb{R}^k$. The noise $\mathbf{N}_e$, $\mathbf{N}_v$ added to the enrollment and verification, are assumed to be uncorrelated to $\mathbf{X}_i$ and are both k-dimensional multivariate Gaussian variables. It has been shown in [10], the biometric capacity is the mutual information between the enrollment and the verification, specified by

$$C = I(\mathbf{Y}_i; \mathbf{Z}_i) = h(\mathbf{Y}_i) + h(\mathbf{Z}_i) - h(\mathbf{Y}_i, \mathbf{Z}_i), \quad (3)$$

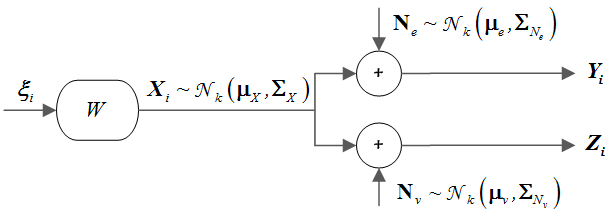


Fig. 1: Capacity analysis model: the biometric data $\boldsymbol{\xi}_i$ is processed by W to generate feature vector $\mathbf{X}_i$, which leads to the enrollment template $\mathbf{Y}_i$ and the verification sample $\mathbf{Z}_i$, respectively.

where $\mathbf{Y}_i$ is the enrolled template and $\mathbf{Z}_i$ is the new sample of the ith subject to be verified. Since $\mathbf{Y}_i = (\mathbf{X}_i + \mathbf{N}_e)$ and $\mathbf{Z}_i = (\mathbf{X}_i + \mathbf{N}_v)$, they are also multivariate Gaussian variables in $\mathbb{R}^k$. Based on the result in [11], the entropy terms in Equation (3) can be calculated as

$$\begin{aligned} h(\mathbf{Y}_i) &= \frac{1}{2}\log_2\left((2\pi e)^k|\Sigma_Y|\right) \\ h(\mathbf{Z}_i) &= \frac{1}{2}\log_2\left((2\pi e)^k|\Sigma_Z|\right) \\ h(\mathbf{Y}_i, \mathbf{Z}_i) &= \frac{1}{2}\log_2\left((2\pi e)^{2k}|\Sigma_{YZ}|\right) \end{aligned} . \quad (4)$$

Substituting Equation (4) into Equation (3) gives,

$$I(\mathbf{Y}_i; \mathbf{Z}_i) = \frac{1}{2}\log_2 \frac{|\Sigma_Y| \times |\Sigma_Z|}{|\Sigma_{YZ}|} , \quad (5)$$

where Σ_{YZ} can be expressed as a $2k \times 2k$ block matrix by

$$\Sigma_{YZ} = \begin{bmatrix} \Sigma_X + \Sigma_{N_e} & \Sigma_X \\ \Sigma_X & \Sigma_X + \Sigma_{N_v} \end{bmatrix} .$$

As known from linear algebra, the determinant of a matrix equals to the product of its eigenvalues. Thus, $|\Sigma_Y|$ and $|\Sigma_Z|$ can both be expressed as the product of k eigenvalues. However, since Σ_{YZ} is double size of the previous two, $|\Sigma_{YZ}|$ corresponds to the product of $2k$ eigenvalues. This difference in number of eigenvalues prevents us from computing the logarithm as the sum of k terms in Equation (5) and therefore prevents the direct examination of how capacity changes with respect to the number of features of the biometric. To solve this problem, we use an important property of covariance matrices which is they are always positive semi-definite. In our case, since Σ_X is the covariance matrix of the first k most significant DCT coefficients, Σ_X is even positive definite, and so is the matrix $\Sigma_X + \Sigma_{N_v}$. As a result, they are both invertible. Therefore, $|\Sigma_{YZ}|$ is equivalent to

$$\begin{aligned} |\Sigma_{YZ}| = &|(\Sigma_X + \Sigma_{N_e})(\Sigma_X + \Sigma_{N_v}) \\ &- \Sigma_X(\Sigma_X + \Sigma_{N_v})^{-1}\Sigma_X(\Sigma_X + \Sigma_{N_v})| \end{aligned} . \quad (6)$$

The term inside the determinant on the right side of Equation (6) is a matrix in $\mathbb{R}^{k\times k}$ rather than $\mathbb{R}^{2k\times 2k}$. Because of this dimension reduction, we are now able to simplify the computation of $|\Sigma_{YZ}|$ as a product of k eigenvalues of its equivalent matrix. As a result, the logarithm in Equation (5) is also transformed into the sum of k terms as

$$\begin{aligned} I(\mathbf{Y}_i; \mathbf{Z}_i) &= \frac{1}{2}\log_2 \frac{\prod_{i=1}^k \lambda_{i(Y)}\lambda_{i(Z)}}{\prod_{i=1}^k \lambda_{i(YZ)}} \\ &= \frac{1}{2}\sum_{i=1}^k \log_2 \frac{\lambda_{i(Y)}\lambda_{i(Z)}}{\lambda_{i(YZ)}} \end{aligned} , \quad (7)$$

where $\lambda_{i(Y)}$, $\lambda_{i(Y)}$ and $\lambda_{i(YZ)}$ are the ith eigenvalue of the covariance matrix Σ_Y, Σ_Z and equivalent covariance matrix of Σ_{YZ}, respectively.

At this stage, we have derived an expression for capacity estimation. In practice, the covariance matrices Σ_X, Σ_{N_e} and Σ_{N_v} cannot be observed directly due to noise. We approximate these terms as follows. Given N subjects, M measurements are collected for each individual, which are further split into K disconnected segments. We assume that, the noise is reduced by averaging. Therefore, Σ_{N_e} and Σ_{N_v} are approximated with the averaged within class covariance Σ_w while Σ_X is approximated with the between class covariance matrix Σ_b, specified by

$$\Sigma_\omega = \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M (\mathbf{Y}_{i,j} - \boldsymbol{\mu}_{i,j})(\mathbf{Y}_{i,j} - \boldsymbol{\mu}_{i,j})^T ,$$

$$\Sigma_b = \frac{1}{MN-1}\sum_{i=1}^N\sum_{j=1}^M (\boldsymbol{\mu}_{i,j} - \boldsymbol{\mu})(\boldsymbol{\mu}_{i,j} - \boldsymbol{\mu})^T ,$$

where $\boldsymbol{\mu}_{i,j}$ denotes the average of the jth measurement from the ith subject and $\boldsymbol{\mu}$ denotes the average of all measurements, respectively.

D. Template Protection

To protect a template stored in the public database, we adopt the Quantization Index Modulation (QIM) scheme to construct the authentication system [5]. Two stages are considered, *i.e.* the enrollment and the verification. The block diagram of our proposed system is shown in Fig. 2.

During the enrollment, DCT coefficients with a value range $[0, L]$ is evenly divided to set the quantization interval of length q. The extracted biometric feature, is quantified as $s_i = \lfloor u_i/q \rfloor$. Meanwhile, the helper

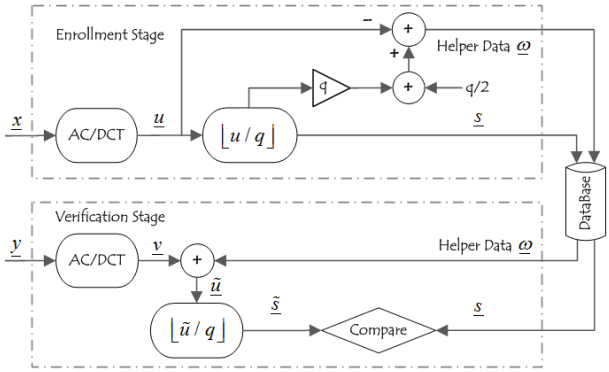


Fig. 2: Block diagram for QIM protection scheme.

data is also constructed by $\omega_i = \lfloor u_i/q \rfloor q - u_i + q/2$. Since biometrics are always contaminated with noise and thereby not exactly reproducible, the helper data is used to center each measurement with the purpose that, all the measurements from the same subject can be mapped into one value. Both the quantized secret and the helper data are stored in the public database. At the verification stage, the biometric features are again measured and the secret is reconstructed by

$$\tilde{s}_i = \left\lfloor \frac{v_i + \omega_i}{q} \right\rfloor = \left\lfloor \frac{v_i - u_i}{q} + \frac{1}{2} \right\rfloor + s_i \, .$$

A sample is admitted if it stays within the threshold, which is set to the maximum Hamming distance between the enrolled and verified samples.

In the above section, the ECG capacity is calculated directly upon the extracted features. After implementing the protection scheme, the capacity achieved in practice can again be estimated. To this end, each feature is modeled as a Binary Symmetric Channel (BSC) since the secret is transferred from the enrollment to the verification. With the transition probability P_i collected for each channel, the overall capacity is estimated by

$$C = \sum_{i=1}^{k} \big(1 - h(P_i)\big) \, . \tag{8}$$

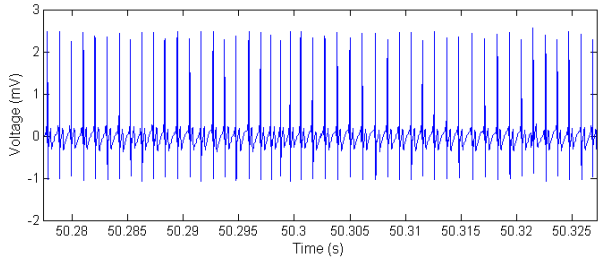
III. Results and Discussion

To examine the performance of the proposed method and system, experiments are implemented. The database used in this paper is collected from 21 subjects (7 females and 14 males, aged between 19 to 26). For every subject three to five measurements are recorded and each is approximately of one-hour duration. To simulate the real life situation, the recordings of every individual are taken with several weeks separation in between. The entire database consists of 74 records, sampled with lead I at 1024Hz by the TMSI Mobi system.

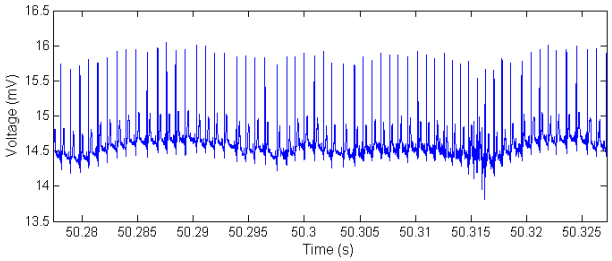
A. Preprocessing and Feature Extraction

Using the preprocessing method explained above, the ECG signal before and after processing are presented in Fig. 3. It can be observed that, the filtered signal contains no obvious noise or serious artifacts and thereby is suitable for further further analysis.

To extract ECG features, we segment the above "clean" recording into 10-second epochs and apply autocorrelation to each with a maximum lag of 0.3 second. Given 1024 sampling rate, this computation results in 308 AC coefficients for nonnegative lags, based on which DCT is performed. The AC and DCT coefficients of



(a) The raw ECG recording



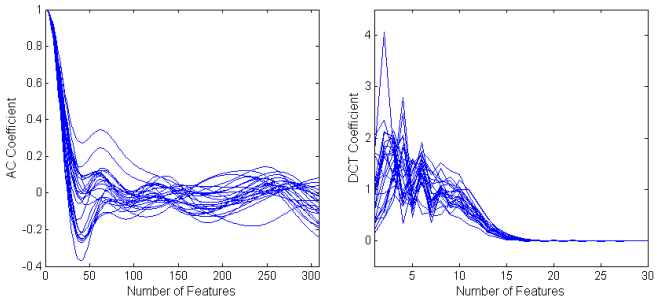
(b) The ECG recording after noise and artifacts removal

Fig. 3: ECG signal before and after preprocessing.

21 subjects are shown in Fig. 4. It can be seen that both AC and DCT coefficients present quite significant variation among different subjects. In addition, DCT greatly reduces the number of independent features to approximately 30.

B. Verification Performance

To construct the verification system, we select the first 30 most significant DCT coefficients as extracted ECG features. For each subject, we use the average of two recordings for enrollment and take the remaining recordings for verification. The False Acceptance Rate (FAR) and False Rejection Rate (FRR) are estimated and



(a) The autocorrelation coefficients of 21 subjects

(b) The corresponding DCT coefficients

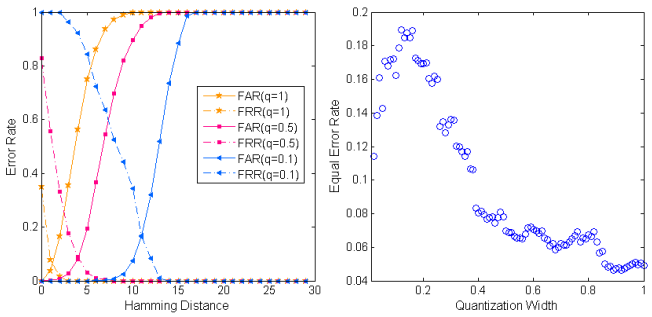
Fig. 4: ECG feature extraction of 21 subjects, where both AC and DCT coefficients demonstrate quiet significant variations among different subjects.

used as the measure for verification performance. Since the quantization width has a significant influence on the QIM scheme, the point that approximate the Equal Error Rate (EER) is selected for comparison. The results are depicted in Fig. 5. It is shown that as the quantization step gets wider, FRR decreases and FAR increases, which together contribute to a lower EER. Furthermore, with a considerably large quantization width, the EER can drop to as low as 4.2%, which almost approaches to the EER of a system without protection. Noted that a better system can be devised if the quantization step is optimized for each dimension. As can be seen in Fig. 4b, the DCT coefficients behave rather different for each feature. In general the first 15 dimensions vary in a much winder range than the last 15 ones.

C. ECG Capacity

The result of ECG capacity based on the autocorrelation coefficients and the the verification system with the QIM scheme is presented in Fig. 6. For comparison, we also include the capacity estimated with a system without any protection. The accumulative capacity with respect to the number of features shows that, the ECG based on autocorrelation provides approximately 20 bits information contained in its 23 independent features, using which about 1,000 individuals can be identified. Moreover, this capacity can also be considered as the upper bound for capacity of a practical verification system. For the QIM scheme, the quantization is one major reason for the loss of information.

In addition, it should be noted that the capacity of a practical verification system is closely related to the choice of the threshold and the quantization level. For instance, the capacity for systems with QIM scheme



(a) FAR and FRR (b) EER against different quantization width

Fig. 5: Verification performance of the authentication system with the QIM scheme.

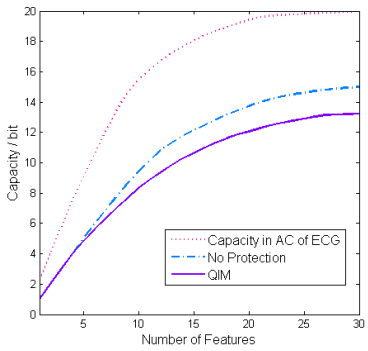


Fig. 6: Capacity analysis result: the comparison is made among capacity based on the AC coefficients of ECG, systems without protection and with the QIM protection.

can be made extremely large if a sufficiently wide quantization step is used. However, a wide quantization leads to serious information leakage [5], thereby it is undesirable for privacy protection. For the QIM scheme, the quantization step is always a tradeoff between the verification performance, achievable capacity and information leakage.

IV. Conclusions

In this paper, we have explored the template protection for ECG data in the verification system. We have derived an improved expression for the biometric capacity estimation, which allows us to exam the number of information a biometric contains and the number of independent features we need to extract this information. Applied to the ECG, it has been found that ECG contains approximately 20 bits of information in its 23 independent features, which allows the identification of about 1,000 individuals. In addition, the QIM protection scheme has been applied to construct a more reliable verification system, in terms of privacy protection. We have shown that compared to unprotected schemes, the QIM scheme greatly enhances the protection of templates privacy without sacrificing significant identification performance. That is, QIM is able to reach the Equal Error Rate (EER) of about 4.2%, which is similar to an unprotected system.

References

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: a new approach in human identification," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, pp. 808–812, June 2001.

[2] S. A. Israel, J. M. Irvine, A. cheng, M. D. Wiederhold, and B. K. Wiederhold, "ECG to identify individuals," *Pattern Recognition*, vol. 38, pp. 133–142, Jan. 2005.

[3] N. P. Plataniotis, D. Hatzinakos, and J. K. M. Lee, "ECG biometric recognition without fiducial detection," in *Proc. Biometrics Symp.: Special Session Research at the Biometric Consortium Conf*, pp. 1–6, 2006.

[4] F. Agrafioti and D. Hatzinakos, "ECG based recognition using second order statistics," in *Proc. 6th Annual Communication Networks and Services Research Conf. CNSR 2008*, pp. 82–87, May 2008.

[5] J.-P. Linnartz and P. Tuyls, "New shielding functions to enhance privacy and prevent misuse of biometric template," in *Audio- and Video-Based Biometric Person Authentication*, Springer, 2003.

[6] P. Tuyls, A. H. M. Akkermans, T. A. M. Kevenaar, G.-J. Schrijen, A. M. Bazen, and B. N. J. Veldhuis, "Practical biometric authentication with template protection," in *Audio- and Video-Based Biometric Person Authentication*, pp. 436–446, Springer, 2005.

[7] J. A. de Groot and J.-P. M. G. Linnartz, "Zero leakage quantization scheme for biometric verification," in *Proc. IEEE Int Acoustics, Speech and Signal Processing (ICASSP) Conf*, pp. 1920–1923, May 2011.

[8] J. A. de Groot and J.-P. M. G. Linnartz, "Improved privacy protection in authentication by fingerprints," in *Proc. WIC symposium on Information Theory in the Benelu*, May 2011.

[9] G. D. Clifford, F. Azuaje, and P. E. Mesharry, *Advanced Methods and Tools for ECG Data Analysis*. Artech House, Inc., 2006.

[10] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in *Proc. IEEE Int Information Theory Symp*, June 2003.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Willey & Sons, Inc., second ed., 2005.