Clustering of Wireless Sensors based on Ultra-Wideband Geo-Regioning

Christoph Steiner and Armin Wittneben
ETH Zurich, Communication Technology Laboratory, CH-8092 Zurich, Switzerland
Email: {steinech,wittneben}@nari.ee.ethz.ch

***Abstract*— This paper introduces a novel approach for the clustering of Ultra-Wideband nodes based on their channel impulse responses (CIRs) to a central unit. We propose to model the CIRs as realizations of a mixture density, where the individual component probability density functions represent the clusters. It is shown that the clustering can be simply achieved by the application of the expectation maximization algorithm. Furthermore, this work studies the impact of a-priori knowledge and the number of available CIR observations on the clustering performance, which is evaluated based on measured CIRs.**

I. Introduction

Ultra-Wideband (UWB) technology is a very promising candidate for energy efficient transceiver structures and localization. In wireless networks, position location information and node clustering are essential for energy-saving techniques like data fusion, location aware routing protocols, or distributed source coding [1].

The most promising UWB localization approaches exploiting the wide bandwidth are based on time of arrival estimation [2]. However, a general problem of localization and tracking systems using time of arrival estimates is the performance degradation under non-LOS conditions [3]. Another localization paradigm is based on comparing a fingerprint/signature extracted from the received signal to entries in a database at the receiver. A-priori information is required to generate this database. Possible types of fingerprint information are, e.g., received signal strength, angular power profile, and power delay profile. The accuracy of such localization methods can be increased by using more distributed receivers, which in turn increases the complexity and the amount of data exchange. On the other hand, it is possible to increase the accuracy by using more signal parameters as fingerprint information like the power delay profile rather than the received signal strength.

It has been shown in [4], [5] that communication systems with sufficiently high signaling bandwidth can directly use the channel impulse response (CIR) as fingerprint information. In these references, only one receiver, which takes over the whole computational complexity, is required. Consequently, the transmitters can be low cost devices and need no additional signal processing or hardware. Moreover, no time synchronization between transmitter and receiver is required. Therefore, this is a very appealing technique for sensor networks, where the sensors are energy limited and the receiver (central unit) is connected to a power supply allowing for more complexity. One additional advantage of this method is its successful operation even in non-LOS situations, since it does not rely on the direct path only, but uses the whole CIR for positioning. This localization method introduced in [4] is called UWB Geo-Regioning. In this reference, it is assumed that the absolute positions of the regions are known, which implies that the positioning errors are determined by the dimensions of the regions. However, if these absolute positions are unknown, UWB Geo-Regioning can still pool close transmitters into clusters and perform a clustering of all nodes within the network. Therefore, the terms *region* and *cluster* are used interchangeably in this work.

One drawback of this method is the required a-priori knowledge to generate the database at the central unit (CU). In [5], it is assumed that about 400 CIRs per region are measured beforehand by an extensive site survey, which is not feasible for a practical application. In this paper, a novel UWB Geo-Regioning approach is proposed, which requires significantly less a-priori known CIRs for the clustering of nodes. We investigate, how many a-priori CIRs (reference measurements) are necessary, such that the nodes can be correctly assigned to their clusters. The performance of the clustering approach is evaluated based on measured CIRs and compared to the UWB Geo-Regioning approach used in [5].

Notation: All vectors are column vectors, $(\cdot)^T$ denotes transposition, $(\cdot)^H$ denotes complex conjugate transposition. The operator E(·) denotes expectation. An estimate for parameter θ is denoted as $\hat{\theta}$. $\mathcal{CN}$ denotes the multivariate proper complex Gaussian probability distribution.

II. Probabilistic CIR Modeling

For this work, the wireless propagation channel is interpreted as a linear time-invariant system, which is fully described by its impulse response. A discrete time version of the CIR is considered in equivalent baseband, where the sampling frequency is chosen such that the sampling theorem is fulfilled, when the CIR is bandlimited to the system bandwidth. The channel taps are assumed to be proper complex Gaussian distributed, which can be justified by the central limit theorem. Many reflected and scattered partial waves from different directions superimpose at the receive antenna and contribute to one channel tap with varying amplitude and phase. However, as the tap duration becomes smaller (approaching UWB) less partial waves contribute to one channel tap. This fact questions the applicability of the central limit theorem. In literature there exist various studies on the tap statistics for UWB

channels. For the UWB channel tap amplitudes the Nakagami [6], Lognormal [7], and Weibull [8] distributions are proposed. However, also Rayleigh and Ricean amplitude distributions arising from the complex Gaussian channel tap distribution are supported by some channel measurement campaigns [9].

We consider a wireless communication scenario, where several low complexity UWB nodes (or nodes signaling with sufficiently high bandwidth) are spatially distributed within communication range of a CU. The CU has the ability to estimate CIRs, whereas the nodes can have noncoherent receivers. It is assumed that the coverage area of the CU can be divided into M regions. The regions or clusters are characterized by a probability model for the discrete time CIRs of nodes located within this region to the CU.

A sampled CIR with K taps from node n is denoted as $\vec{x}_n = [x_n[1], \ldots, x_n[K]]^T$. If the node is located in region m the CIR is modeled as jointly proper complex Gaussian random vector with zero mean vector and covariance matrix Σ_m. The corresponding probability density function (PDF) is given by

$$p\left(\vec{X}_n|m\right) = \mathcal{CN}\left(\vec{0}, \Sigma_m\right). \tag{1}$$

If the region of node n is not known, the observed CIR could be distributed according to any of the M distributions $\mathcal{CN}\left(\vec{0}, \Sigma_m\right)$ for $m = 1, 2, \ldots, M$. Therefore, we propose to describe the CIR from such a node n with a jointly proper complex Gaussian mixture density according to

$$p\left(\vec{X}_n\right) = \sum_{m=1}^{M} \alpha_m p\left(\vec{X}_n|m\right), \tag{2}$$

where the component PDF weights α_m sum up to one $\left(\sum_{m=1}^{M} \alpha_m = 1\right)$. The set of parameters describing this density is denoted by $\Theta = \{\alpha_1, \alpha_2, \ldots, \alpha_M, \Sigma_1, \Sigma_2, \ldots, \Sigma_M\}$.

The covariance matrix Σ_m acts as fingerprint information and is estimated from N_{ref} a-priori known CIRs from nodes located within region m. The accuracy of the estimate depends on the available a-priori knowledge. The maximum likelihood (ML) estimator using a set of N_{ref} i.i.d. CIR observations $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{N_{\text{ref}}}\}$ is given according to [10] by

$$\hat{\Sigma}_m = \frac{1}{N_{\text{ref}}} \sum_{i=1}^{N_{\text{ref}}} \vec{x}_i \vec{x}_i^H. \tag{3}$$

The estimated covariance matrix $\hat{\Sigma}_m$ is distributed according to a Wishart distribution [10], where the variance of the estimate decreases linearly with the number of observations N_{ref}.

III. Node Clustering

This section describes three approaches for node clustering, when the CU has already built up its database consisting of M estimated covariance matrices $\left\{\hat{\Sigma}_1, \ldots, \hat{\Sigma}_M\right\}$. The goal is to find the region/cluster, where a node is located in, based on its CIR to the CU. The database at the CU can be obtained either by using CIRs of reference nodes with known region, by a site survey, or by a-priori knowledge about the geometry of the environment using ray-tracing methods.

A. Maximum Likelihood Clustering

The CU estimates the CIR $\vec{x}_n$ from a node n with unknown position. Equal a-priori probabilities $(\mathcal{P}(\text{ Node } n \in \text{ Region } m_1) = \mathcal{P}(\text{ Node } n \in \text{ Region } m_2))$ for $m_1, m_2 = 1, 2, \ldots, M$ are assumed. The optimal algorithm for node classification based on the modeling assumptions is given by the following ML rule:

$$\hat{m} = \max_{m \in \{1,2,\ldots,M\}} f\left(\vec{x}_n|\hat{\Sigma}_m\right), \tag{4}$$

where the expression $f\left(\vec{x}_n|\hat{\Sigma}_m\right)$ denotes the likelihood of $\vec{x}_n$ being distributed according to $\mathcal{CN}\left(\vec{0}, \hat{\Sigma}_m\right)$. The corresponding probabilities of misclassification conditioned on region m are defined by

$$P_m = \mathcal{P}\left(\hat{m} \neq m|\text{ Node } n \in \text{ Region } m\right). \tag{5}$$

The performance of the ML classification depends among other things on the accuracy of the estimation of the covariance matrices, and therefore on the number of reference CIRs available for parameter estimation (N_{ref}). A thorough performance analysis of the ML classification algorithm can be found in [11].

In practical scenarios, there can exist very few reference nodes per region implying inaccurate parameter estimates. This necessitates iterative clustering algorithms, which use already classified CIRs to increase the accuracy of the parameter estimates and, consequently, the performance. The next section presents such a clustering algorithm based on the expectation maximization (EM) principle.

B. Expectation Maximization Clustering

It is assumed that the CU can collect $M \cdot N_{\text{batch}}$ CIR estimates from nodes with unknown cluster information. This number is limited by the number of static nodes in communication range to the CU. However, if mobile nodes are assumed, the CU can use previously recorded CIRs from the same node. The idea is that these CIRs can be thought of as i.i.d. realizations of the complex Gaussian mixture density given in (2), whereas CIRs originating from region m are distributed according to the corresponding component PDF given in (1).

The EM algorithm (cf. [12]) is an iterative approach to find ML estimates of parameters of a distribution from a given data set, which is incomplete or has missing values. It is also possible to use this method to optimize involved likelihood functions by assuming the existence of missing parameters, which renders the likelihood function mathematically tractable. Consequently, the EM algorithm can be used to find ML parameter estimates for the mixture density in (2).

The EM parameter estimates for iteration step i, $\left(\hat{\Theta}[i] = \left\{\hat{\alpha}_1[i], \hat{\alpha}_2[i], \ldots, \hat{\alpha}_M[i], \hat{\Sigma}_1[i], \hat{\Sigma}_2[i], \ldots, \hat{\Sigma}_M[i]\right\}\right)$,

based on the old parameter estimates of step $i-1$ and the set of CIR observations $\{\vec{x}_1, \ldots, \vec{x}_{M \cdot N_{\text{batch}}}\}$ are given by

$$\hat{\alpha}_m[i] = \frac{1}{M \cdot N_{\text{batch}}} \sum_{n=1}^{M \cdot N_{\text{batch}}} \mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right),$$

$$\hat{\Sigma}_m[i] = \frac{\sum_{n=1}^{M \cdot N_{\text{batch}}} \vec{x}_n \vec{x}_n^H \mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right)}{\sum_{n=1}^{M \cdot N_{\text{batch}}} \mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right)},$$

where $\mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right)$ is the probability mass of the event that the CIR $\vec{x}_n$ originates from region m conditioned on the parameter estimates $\hat{\Theta}[i-1]$ and is given by

$$\mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right) = \frac{\hat{\alpha}_m[i-1] f\left(\vec{x}_n|\hat{\Sigma}_m[i-1]\right)}{\sum_{m'=1}^{M} \alpha_{m'}[i-1] f\left(\vec{x}_n|\hat{\Sigma}_{m'}[i-1]\right)}.$$

The expression $f\left(\vec{x}_n|\hat{\Sigma}_m[i-1]\right)$ denotes the likelihood of $\vec{x}_n$ being distributed according to $\mathcal{CN}\left(\vec{0}, \hat{\Sigma}_m[i-1]\right)$. The cost function to optimize is given by

$$\mathcal{Q}\left(\hat{\Theta}[i], \hat{\Theta}[i-1]\right) = \\ = \sum_{m=1}^{M} \sum_{n=1}^{M \cdot N_{\text{batch}}} \log\left(\hat{\alpha}_m[i] f\left(\vec{x}_n|\hat{\Sigma}_m[i]\right)\right) \mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i-1]\right).$$

This cost function increases monotonically in i and converges only to a local optimum for i^*, since it is in general non-convex. The reached local optimum depends on the initial parameter set $\hat{\Theta}[0]$. The initial covariance matrices are obtained from the database entries. The initial parameters $\{\hat{\alpha}_1[0], \hat{\alpha}_2[0], \ldots, \hat{\alpha}_M[0]\}$ are set to $1/M$ each, because it is assumed that the same number of reference CIRs per region were used for initializing the database. After convergence of the EM algorithm, the new parameter estimates $\hat{\Theta}[i^*]$ are written to the database and the nodes are classified according to the following rule:

$$\hat{m} = \max_{m \in \{1,2,\ldots,M\}} \mathcal{P}\left(m|\vec{x}_n, \hat{\Theta}[i^*]\right), \qquad (6)$$

where the conditional probabilities of misclassification are defined the same way as in (5).

The advantage of this iterative clustering method over ML clustering is the inherent usage of already classified CIRs for parameter estimation. Furthermore, the contribution of a classified CIR for parameter estimation is weighted by the respective likelihood value. This means that rather unsure classification results do not contribute significantly to the new parameter estimates. However, the drawback of this method is that lots of CIRs ($M \cdot N_{\text{batch}}$) are required at once for the EM algorithm to perform well. The following section presents an adaptive EM approach, where this number is reduced significantly.

C. Adaptive Expectation Maximization Clustering

In this scenario, the CU has only a very limited number of CIR estimates N_{adapt} per time slot available. This would correspond to a realistic scenario, where only very few nodes can transmit to the CU at the same time due to limited resources. The CU could wait for some time to collect enough CIRs and apply the standard EM clustering as described above. However, this approach imposes a long delay and requires the CU to store a lot of data. The adaptive method proposed here uses the available CIRs in one time slot for EM clustering and afterwards updates the database with the obtained parameter estimates. In the next time slot, the updated database entries are used to initialize a new EM run, which clusters the newly estimated CIRs. This approach would be able to adapt to changes in the network configuration (mobile nodes) by introducing a forgetting factor.

The performance of these three proposed clustering approaches is evaluated based on measured CIRs. The next section shortly describes the measurement campaign. Section V presents the performance results.

IV. CIR Measurements

Fig. 1 shows the room of size 7.4 m $\times$ 15 m and height 6 m, where CIRs from different transmitter positions within 22 different regions to a receiver have been measured. The maximum distance between two regions is approximately 16 m, whereas the minimum separation of two transmitter positions in two different regions is approximately 10 cm.

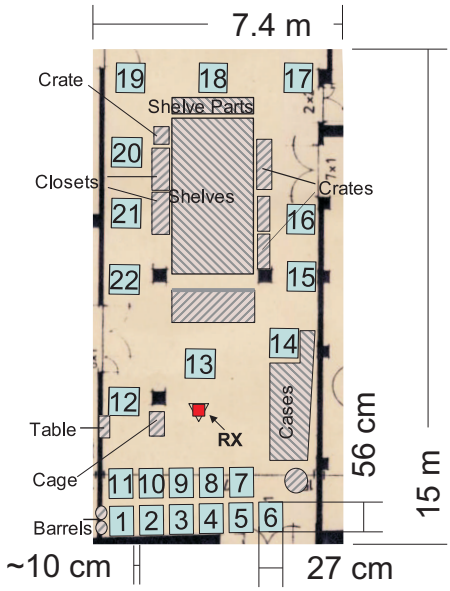

Fig. 1. Schematics of Measurement Site

In total $N_{\text{total}} = 600$ CIRs per region for different transmitter positions have been measured. The measurement frequency range is roughly limited from 3 GHz to 6 GHz by the transfer function of the UWB antenna and the cut-off frequency of the amplifier. A more detailed description of the measurements can be found in [11].

V. Performance Results

We consider CIRs within an observation window of 10 ns sampled at Nyquist rate. This implies that a CIR vector in equivalent baseband consists of $K = 30$ complex channel taps. Two sets of $M = 4$ regions are selected out of the measured 22 regions. Set 1 consists of regions $\{3, 4, 8, 9\}$ and constitutes a worst case scenario, because all regions are adjacent and separated only by approximately 10 cm. Since the largest wavelength in the measured bandwidth is $\lambda_{\text{max}} = 10$ cm, there exist CIRs in different regions, which are correlated. On the other hand, set 2 consists of regions $\{1, 15, 18, 21\}$ and constitutes a best case scenario, since the regions are separated by several meters. The conditional probabilities of misclassification defined in (5) are used as performance measure. In order to calculate expected error probabilities a Cross Validation Method is applied, where the error probabilities are averaged over 100 trials with randomized CIRs sets for a-priori parameter estimation (N_{ref}) and for algorithm testing (ML) and EM parameter estimation $(M \cdot N_{\text{batch}})$.

Fig. 2 and Fig. 3 depict the expected probabilities of classification errors for each region depending on the number of reference CIRs N_{ref} per region, when all remaining CIRs $(M \cdot N_{\text{batch}} = M \cdot (N_{\text{total}} - N_{\text{ref}}))$ are used for the EM clustering. Furthermore, they show the expected error probabilities for the ML clustering algorithm, when N_{ref} CIRs per region are used for parameter estimation. It can be seen that there is a significant performance improvement especially for small N_{ref}, when the iterative EM clustering method is used. As N_{ref} increases, this performance gap becomes smaller, since the error probabilities are not dominated by the parameter estimation errors anymore. The worst case set exhibits rather high probabilities of misclassification, because there exist node positions in two different regions, which are only 10 cm apart. This implies that the corresponding measured impulse responses still show a significant correlation, which leads to wrong classifications. If the distance between the regions is big enough, as for set 2, it can be seen that even for only $N_{\text{ref}} = 20$ reference CIRs 90% of all nodes located in region 18 can be classified correctly. Nodes from other regions are classified with error probabilities smaller than 0.02. However, the ML clustering algorithm fails, because the number of i.i.d. samples for parameter estimation is smaller than the length of the random vector, which implies inaccurate covariance matrix estimates.

The influence of the available number of CIRs (N_{batch}) on the performance of the EM clustering algorithm for set 2 is depicted in Fig. 4. Here the number of reference CIRs is set to $N_{\text{ref}} = 30$ per region. It can be seen that by decreasing N_{batch}, the classification errors increase due to the reduced number of i.i.d. samples for EM parameter estimation. However, there is still a significant performance improvement noticeable compared to the ML clustering algorithm even for $N_{\text{batch}} = 200$.

Fig. 5 and Fig. 6 show the performance results for the adaptive EM clustering approach. Here, only $N_{\text{adapt}} = 10$ CIRs

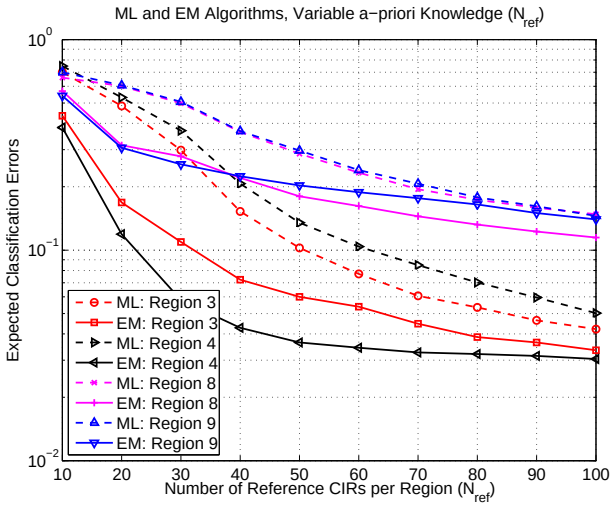

Fig. 2. Expected probabilities of misclassification for regions {3,4,8,9} as a function of N_{ref} for $N_{\text{batch}} = 600$.

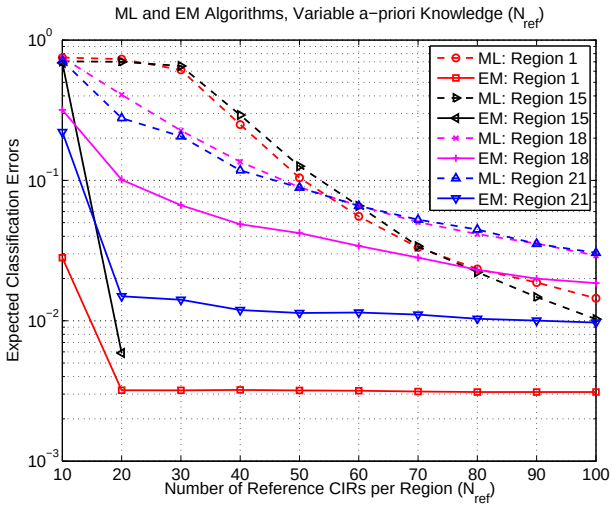

Fig. 3. Expected probabilities of misclassification for regions {1,15,18,21} as a function of N_{ref} for $N_{\text{batch}} = 600$.

per run are used for EM parameter estimation. It is reasonable that the CU can estimate the CIRs to 10 nodes within a very short time. The parameter outputs after each EM run are used to initialize a new EM run. Therefore, a continuous reduction of the error probabilities is achieved. Fig. 5 and Fig. 6 depict the steady state error probabilities as a function of N_{ref}, after all remaining CIRs $(M \cdot N_{\text{batch}} = M \cdot (N_{\text{total}} - N_{\text{ref}}))$ are used in portions of $N_{\text{adapt}} = 10$ for EM parameter estimation. It can be seen that more reference CIRs are required for this approach to achieve similar results as in Fig. 2 and Fig. 3. This is due to the rather high error probabilities in the beginning, when the parameter estimates are still inaccurate. However, the results show that also this approach converges to reasonable low error probabilities.

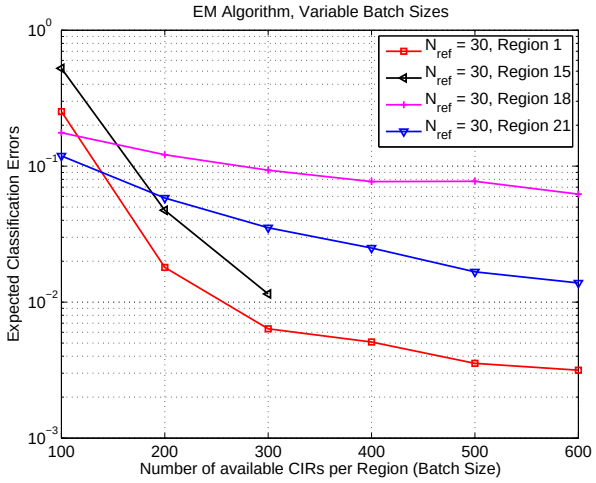

Fig. 4. Expected probabilities of misclassification for regions {1,15,18,21} as a function of N_{batch} for $N_{\text{ref}} = 30$.

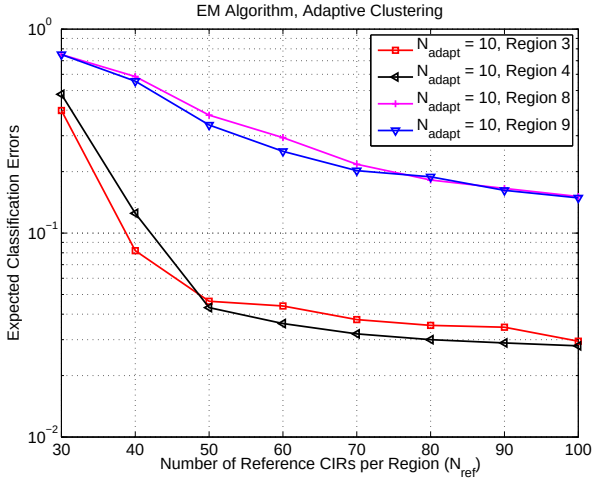

Fig. 5. Expected probabilities of misclassification for regions {3,4,8,9} as a function of N_{ref} for $N_{\text{adapt}} = 10$.

VI. Conclusions

We presented a statistical model (complex Gaussian mixture density) for CIRs from wireless UWB devices to a CU, which accounts for the fact that the nodes can be located in different clusters/regions. We propose to apply the EM algorithm in order to estimate the parameters of this mixture density and to perform node clustering. We show by means of performance results based on measured CIRs that the proposed clustering approach performs with reasonable low error probabilities even for a very small amount of a-priori knowledge. Furthermore, we presented a slight modification of the EM clustering, which allows for a faster response to mobility within the network.

VII. Acknowledgements

The work presented in this paper was partially supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

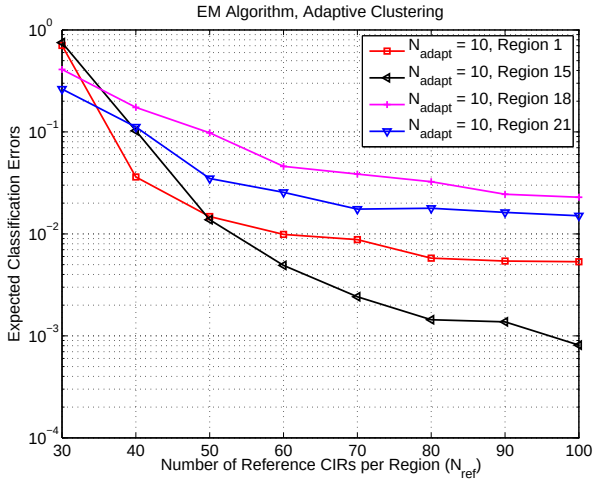

Fig. 6. Expected probabilities of misclassification for regions {1,15,18,21} as a function of N_{ref} for $N_{\text{adapt}} = 10$.

References

[1] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, Sept. 2004.

[2] S. Gezici, Z. Tian, G. Giannakis, H. Kobayashi, A. Molisch, H. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 70–84, July 2005.

[3] B. Denis, J. Keignart, and N. Daniele, "Impact of NLOS propagation upon ranging precision in UWB systems," in *IEEE Conference on Ultra Wideband Systems and Technologies*, Reston, USA, Nov. 2003, pp. 379 – 383.

[4] F. Althaus, F. Troesch, and A. Wittneben, "UWB geo-regioning in rich multipath environment," in *IEEE Vehicular Technology Conference*, Dallas, USA, Sept. 2005, pp. 1001–1005.

[5] C. Steiner and A. Wittneben, "UWB Geo-Regioning using multivariate channel statistics," in *IEEE International Conference on Communications 2007*, Glasgow, Scottland, June 2007, pp. 5763–5768.

[6] D. Cassioli, M. Z. Win, and A. F. Molisch, "The ultra-widebandwidth indoor channel: From statistical model to simulations," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 6, pp. 1247–1257, Aug. 2002.

[7] J. Keignart and N. Daniele, "Channel sounding and modelling for indoor UWB communications," in *Proc. Int. Workshop on Ultra Wideband Systems, IWUWBS*, Oulu, Finland, June 2003.

[8] P. Pagani, P. Pajusco, and S. Voinot, "A study of the ultra-wide band indoor channel: Propagation experiment and measurement results," in *Proc. Int. Workshop on Ultra Wideband Systems, IWUWBS*, Oulu, Finland, June 2003.

[9] U. G. Schuster and H. Bölcskei, "Ultrawideband channel modeling on the basis of information-theoretic criteria," *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2464–2475, July 2007, originally submitted to JSAC, accepted for publication, but withdrawn because of a severe constraint on the overall length of the final paper. Resubmitted to TWireless with many modifications. [Online]. Available: http://www.nari.ee.ethz.ch/commth/pubs/p/schuster07-07a

[10] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. John Wiley and Sons, Inc., 2003.

[11] C. Steiner, F. Althaus, F. Troesch, and A. Wittneben, "Ultra-Wideband Geo-Regioning: A novel clustering and localization technique," *EURASIP Journal on Advances in Signal Processing, Special Issue on Signal Processing for Location Estimation and Tracking in Wireless Environments*, Nov. 2007.

[12] J. A. Blimes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *Technical Report, University of Berkeley, ICSI-TR-97-021*, 1997.