Localization in Wireless Networks using Decision Trees and *K*-means Clustering

Khalid K. Almuzaini and T. Aaron Gulliver
Department of Electrical and Computer Engineering
University of Victoria, Victoria, BC, Canada
email: {kmuzaini, agullive}@ece.uvic.ca

***Abstract*—Node localization is employed in many wireless networks as it can be used to improve routing and enhance security. In this paper, we propose a new algorithm based on decision tree classification and *K*-means clustering which are well known techniques in data mining. Several performance measures are used to compare the *K*-means localization algorithm with those using linear least squares (LLS) and weighted linear least squares based on singular value decomposition (WLS-SVD). It is shown that the proposed algorithm performs better than the LLS and WLS-SVD algorithms even when the geometric anchor distribution about an unlocalized node is poor.**

***Index Terms*—localization, positioning, ad hoc networks, range-based, wireless sensor network, *K*-means, decision trees.**

I. Introduction

The process of finding the spatial location of nodes in a wireless network is commonly called localization. In a wireless network, we can classify the nodes into three categories. Nodes that know their position or coordinates are called *anchors* while those that do not know their position are called *unlocalized*. Nodes that were unlocalized but subsequently had their positions estimated using a localization algorithm are called *localized*.

The location information of nodes in a wireless network is useful for many reasons. It can help track mobile nodes that enter the coverage area of a network, monitor the coverage area of the network over time, determine the coverage area, support load and traffic management, node lifetime control, cluster formation, and enhance network routing. There are many different aspects to the localization problem, such as when localization should be performed and how frequently. Upon network start up, nodes must all be initially localized. This may have to be repeated periodically, for example if there are mobile nodes in the network.

The quality or resolution of the localization is an important consideration. Sometimes, node locations are required within meters of their actual positions, and other times within a few centimetres. Some applications only require relative localization such as node A is in region 1 and node B is in region 2, or node A is north of node B or within its range. For example, monitoring people in a building when we only need to know if an employee enters a certain room during the day.

When the number of anchor nodes is low, they cannot cover the entire wireless network. This means that some unlocalized nodes may not be within range of the signals from the anchor nodes. In this case, localized nodes can participate in the localization process by acting as anchors. This is called cooperative localization.

Localization algorithms can be divided into two categories: range-based and range-free. Range-free algorithms depend on proximity sensing or connectivity information to estimate the node locations. These include for example CPE [8], centroid [9], APIT [10], and the distributed algorithm in [4]. Range-based algorithms estimate the distance between nodes using location metrics such as time of arrival (ToA) [1], time difference of arrival (TDoA) [5], received signal strength (RSS) [6], or angle of arrival (AoA) [7].

The proposed approach differs from conventional solutions to the localization problem in wireless networks. Typically the locations of the anchor nodes within range and the estimated distances between the unlocalized node and these nodes are used to directly estimate its location. Instead, we use a multi-step process. Two approaches from data mining called decision tree classification and *K*-means clustering are used to retain the best (candidate) points, and these are averaged to get the estimated location of the unlocalized node.

We use the linear least squares (LLS) [13] and the weighted linear least squares singular value decomposition (WLS-SVD) algorithms [2] as a basis for comparison. In [2], the WLS-SVD algorithm is compared with a maximum likelihood (ML) algorithm [14], multidimensional scaling (MDS) [16], and the best linear unbiased estimator approach based on least squares calibration (BLUE-LSC) [15]. According to [2], WLS-SVD performs better than these three algorithms.

The remainder of the paper is organized as follows. Dilution of precision is explained in Section II. The decision tree classification and *K*-means clustering algorithms are explained in Section III. The proposed algorithm is presented in Section IV. Some performance results are given in Section V, and finally some conclusions are given in Section VI.

978-1-4673-1881-5/12/$31.00 ©2012 IEEE

II. Dilution of Precision

Dilution of precision is a metric which describes how good an anchor node geometry is for localization. The distance measurements used to compute the node coordinates always contain some error. These measurement errors result in errors in the computed node coordinates. The magnitude of the final error depends on both the measurement errors and the geometry of the structure induced by the nodes. The contribution due to geometry is called the geometric dilution of precision (GDOP). GDOP is used extensively in the GPS community as a measure of localization performance [12]. Another version of GDOP is the generalized geometry of dilution precision GGDOP [11] given by

$$\Gamma_m = \frac{\psi_m}{\gamma_m^2} \tag{1}$$

where

$$\gamma_m = \sum_{i=1}^{m} \frac{1}{\sigma_i^2} \tag{2}$$

and

$$\psi_m = \sum_{i=1}^{m} \sum_{j=1, j>i}^{m} \frac{\sin^2(\alpha_i - \alpha_j)}{\sigma_i^2 \sigma_j^2} \tag{3}$$

The distance error for node i has a Gaussian distribution with variance σ_i^2. The angle α_i is the orientation of the ith anchor or localized node relative to the node whose location is being estimated, as shown in Fig. 1, and m is the total number of anchor and localized nodes around this node. As the GGDOP increases, the localization error decreases.

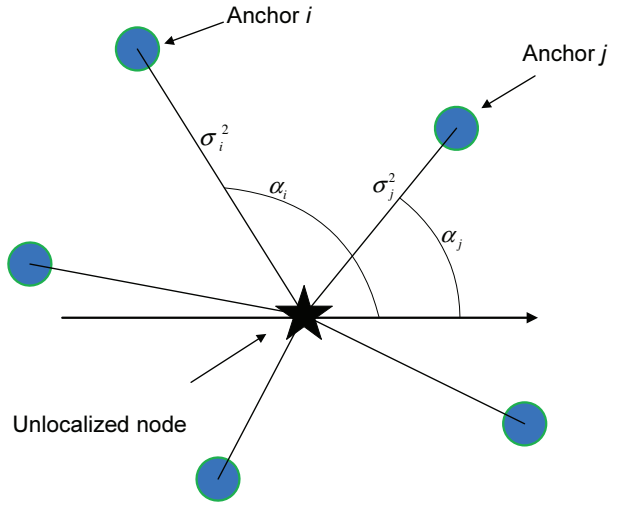

Fig. 1: An unlocalized node with multiple anchor nodes within its range.

III. Decision Tree Classification and K-means Clustering Algorithms

Classification is the task is assigning an object to one or more categories. Decision trees are a simple and widely used classification technique [3]. Moving from one level to another in a decision tree requires a test condition to decide which branch to follow. This process is continued until a leaf node is reached. In the proposed algorithm, we use the shortest distance between nodes as the test.

K-means is a clustering algorithm commonly used in data mining. It defines a cluster or class in terms of a centroid which is the mean of a group of nodes [3]. An advantage of K-means is that it is a simple algorithm. The algorithm starts by selecting K points randomly as centroids. It then tests all other nodes by calculating the distance between each point and each of the K initial centroids. Each point is assigned to the cluster which it is closest to. In the second iteration, the centroids are updated by taking the average of all cluster member locations as the new centroids, and then the nodes are reassigned to clusters. This process is repeated until the centroids do not change. The K-means algorithm is given in Algorithm 1. K-means may have difficulty detecting clusters which have non-spherical shapes or widely different numbers of points or densities [3]. However, in the scenario described in this paper, K-means works well because the number of points in a cluster is not large.

The accuracy of K-means can be improved by carefully selecting the initial centroids. In the proposed algorithm, the points with the minimum and maximum densities are selected as the initial centroids, i.e., $K = 2$. Our aim is to cluster together the points which are close to the actual position of the unlocalized node, and cluster together the points far from this position (outliers) together. The density is given by [3]

$$density(p, Q) = \left(\frac{\sum_{y \in N(p,Q)} d(p, y)}{|N(p, Q)|} \right)^{-1} \tag{4}$$

which is the inverse of the mean distance to the Q nearest neighbours of point p. $N(p, Q)$ is the set containing these nearest neighbours, and $|N(p, Q)|$ is the size of the set.

Algorithm 1 The K-means Algorithm

1: select K points as initial centroids
2: **repeat**
3: form K clusters by assigning each point to its closest centroid
4: recompute the centroid of each cluster by averaging the points in each cluster
5: **until** the centroids do not change

IV. The Proposed Algorithm

The first step in localization is to obtain the distance estimates for the unlocalized nodes from the anchor and localized nodes which are within range. These estimates provide the radii for circles around the nodes, as shown in Fig. 2 for two anchors. Next, the intersection of these

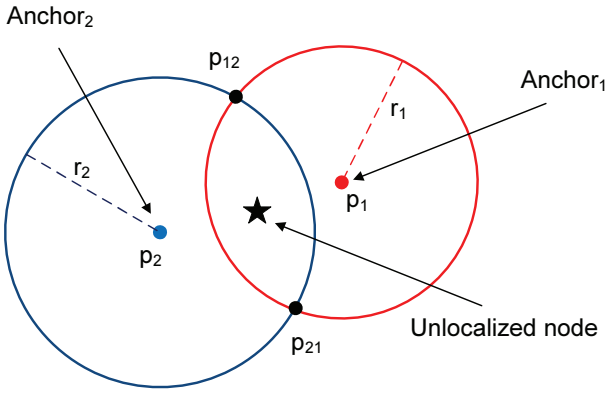

Fig. 2: Intersection of the distance estimates for two anchor nodes.

circles is determined. In the ideal case the circles intersect on the node. For example, when we have three anchor nodes, three intersection points lie on the unlocalized node, while the other three do not. However, in practical situations this event is unlikely.

In Fig. 2, two anchor nodes are located at $p_1 = (x_1, y_1)$ and $p_2 = (x_1, y_1)$ and the distance between them is

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{5}$$

The intersection points of the circles around p_1 and p_2 are denoted as p_{12} and p_{21}.

Each unlocalized node estimates the distance from the anchor or localized nodes that it can receive a signal from. A node can estimate its position only if it hears from three or more of these nodes. The intersections of the estimates around an unlocalized node produce a set of intersection points. If we have m anchor and/or localized nodes, then they form g groups

$$g = \begin{pmatrix} m \\ 2 \end{pmatrix} = \frac{m!}{2!\,(m-2)!} \tag{6}$$

Each group consists of two points as a result of the intersection between two anchor and/or localized node estimates, as shown in Fig. 2.

The third step in estimating the location of an unlocalized node is constructing two decision trees to select the inner intersection points. The root or starting point in both trees can be chosen arbitrarily, but they should belong to the same group. In Fig. 3 the root points are p_{12} and p_{21}, and they are produced by the intersection of the Anchor$_1$ and Anchor$_2$ range estimates. The subtrees in this figure correspond to three anchor and/or localized nodes around the unlocalized node. To traverse from the root points down to the leaf points, the branches with the smallest metric are chosen. If all anchor and/or localized nodes intersect, the maximum number of points is $2g$, and we need to select g inner points, one from each group. The Euclidean distance between two points is

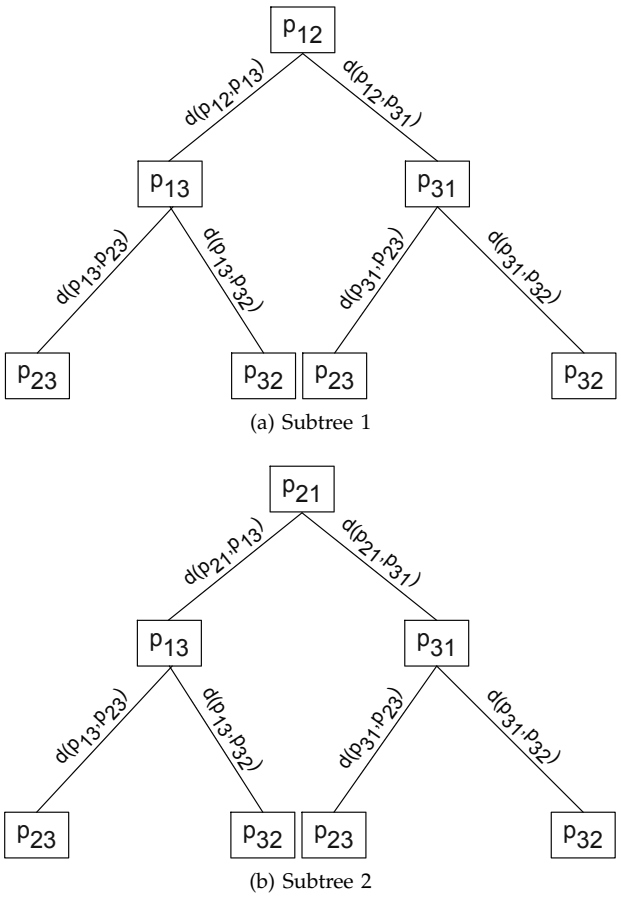

Fig. 3: Decision tree for three anchor nodes divided into two subtrees.

used as the decision tree metric. The path with the minimum distance is followed in each subtree.

Suppose that in Fig. 3a the selected path has points p_{12}, p_{13}, and p_{32}, and in Fig. 3b the selected path has points p_{21}, p_{31}, and p_{23}. Then we have two sets each consisting of g points, one from each subtree. The set with the minimum total distance is chosen as the inner intersection points.

In some cases, for example when a distance estimate is obtained by measuring the received signal strength (RSS), it may be small due to fading and other effects, so that not all circles intersect. However, the intersection points can still be calculated, but they will be complex numbers. If this occurs, we consider the real part as the intersection point in subsequent calculations.

For example, with four anchor nodes, after the distance estimates are determined, the $2g = 12$ intersection points are found, and then 6 inner intersection points are selected. Note that if some circles do not intersect, there will be fewer intersection points.

Now, we apply K-means clustering with $K = 2$ to divide the inner points into dense points and outliers as the fourth step. To calculate the density of each inner

point, we set $Q = g - 1$, i.e., each point considers its Q nearest neighbours in calculating its density. Then we select the cluster which has the maximum number of points. If more than one cluster has the maximum number of points, choose the one with the lowest sum of squared errors (SSE) [3]

$$SSE_i = \sum_{p \in C_i} d(c_i, p)^2, \quad i = 1, \ldots, K \tag{7}$$

where p is an intersection point, C_i is the ith cluster, c_i is the centroid of cluster C_i, and K is the number of clusters. A lower value of SSE means a denser cluster.

If the selected cluster is $\mathbf{v} = \{v_1, v_2, \ldots, v_q\}$ where q is the number of points in the cluster, then the estimated location of the unlocalized node is the average of these (candidate) points

$$\hat{\mathbf{u}} = \frac{\sum_{i=1}^{q} v_i}{q} \tag{8}$$

V. Performance Results

In this section, the proposed algorithm, LKmeans, is compared with the WLS-SVD [2] algorithm via simulation. We first consider distance variance to measure the accuracy of both techniques. 100 nodes are deployed with the anchor nodes chosen randomly. The deployment area is $A = 100 \times 100$ m^2, and the range is $r = 10$ m. The distance error has a Gaussian distribution with variance which is a percentage of the actual distance. The mean error is used as a performance measure and is defined as

$$\text{mean error} = \frac{\sum_{t} \sum_{i=1}^{u} \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{tu} \tag{9}$$

where u is the number of unlocalized nodes, t is the number of trials, $(\hat{x}, \hat{y})$ is the estimated unlocalized node position, and (x, y) is the actual position. The results were averaged over 10^5 trials. Localized nodes were used with the anchor nodes to localize those unlocalized nodes which were not within range of a sufficient number of anchors in the previous iterations. The localization process ends when all nodes are localized or all remaining unlocalized nodes are isolated, i.e., not in the range of three or more anchor or localized nodes. Fig. 4 shows that with 20% anchor nodes, the mean error with the proposed algorithm is lower than with LLS and WLS-SVD, and the rate of change of the error is also lower.

Next all algorithms are compared considering the transmission range of the wireless nodes. The deployment area and the number of nodes are the same as before, but the number of anchor nodes is 50% and the transmission range varies from 10 to 50 meters. The distance error variance is fixed at 10% of the actual distance between nodes. The results were again averaged over 10^5 trials. Fig. 5 shows that the proposed algorithm performs better than LLS and WLS-SVD at low

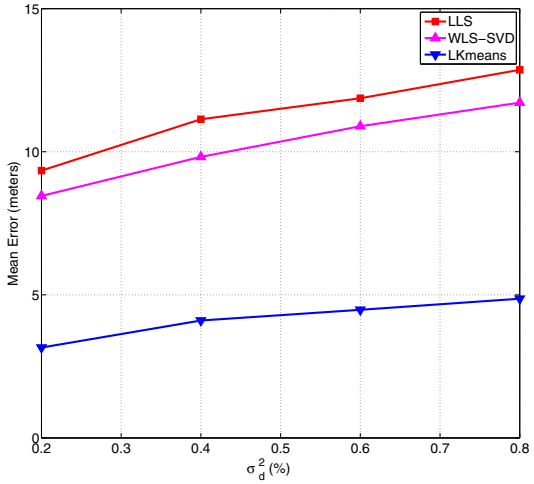

Fig. 4: Mean error versus distance variance.

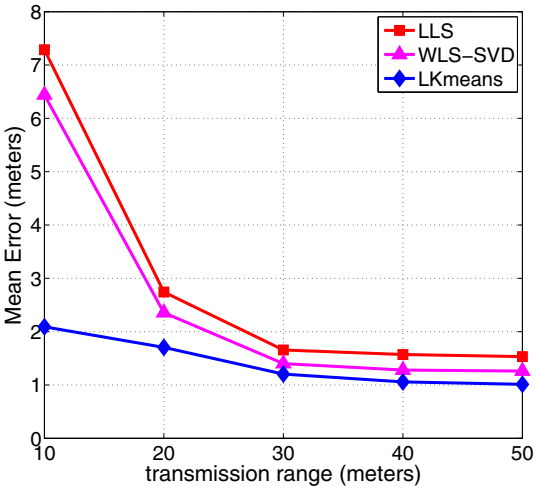

Fig. 5: Mean error versus transmission range.

transmission ranges where the unlocalized nodes can be reached by a small number of nodes, and also at high transmission ranges.

The anchor node ratio is one of the most important factors affecting localization accuracy. Thus, we next compare the algorithms with a varying percentage of anchor nodes. In this case we are more interested in the performance when the anchor node ratio is small because in a practical system the number of anchor nodes will be much less than the number of unlocalized nodes. The deployment area and the number of nodes are the same as before but the anchor node ratio varies from 20% to 80%. The transmission range is fixed at 10 meters and the distance error variance is fixed at 10% of the actual distance. The results are again averaged over 10^5 trials. Fig. 6 shows that the proposed algorithm again outperforms both the LLS and WLS-SVD algorithms,

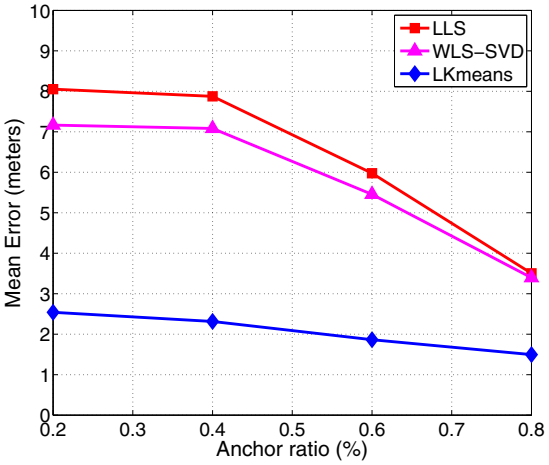

Fig. 6: Mean error versus anchor node ratio.

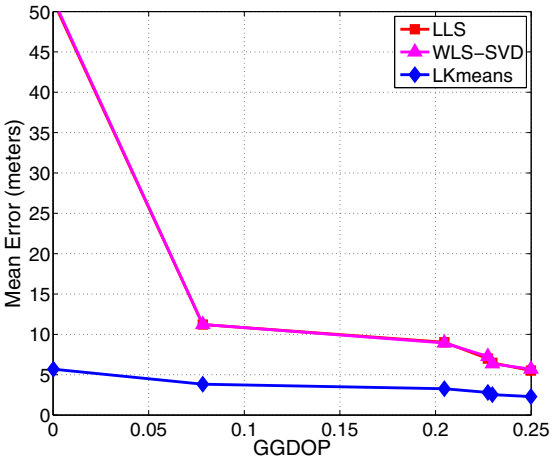

Fig. 7: Mean error versus GGDOP.

particularly at low anchor node ratios.

We now consider the effect of the node geometry on performance, with GGDOP used as the geometry measure. Three anchor nodes are deployed on a circle with a fourth unlocalized node in the center. The transmission range is set to 30 meters to ensure that the unlocalized node is within range of the three anchor nodes, and the distance error variance is set at 10%. The anchor nodes a_1, a_2, and a_3 are distributed around the unlocalized node u by changing the angles $\measuredangle a_1 u a_2$ and $\measuredangle a_2 u a_3$ from 1° to 101° (with both angles the same). The results are averaged over 10^5 trials, and are shown in Fig. 7. This shows that the proposed algorithm perform better, particularly with poor geometry, i.e., low GGDOP or small angles. Note that the LLS and WLS-SVD algorithms have similar performance at all GGDOP values.

VI. Conclusions

A new range-based localization algorithm has been presented. The idea is based on a new approach to the localization problem in wireless networks. Decision tree classification and *K*-mean clustering algorithms from data mining are employed. The proposed algorithm is based on choosing the best candidates among a set of intersection points from the anchor and localized nodes within range of an unlocalized node.

References

[1] I. Guvenc and Z. Sahinoglu, "Threshold-based TOA estimation for impulse radio UWB systems," in *Proc. IEEE Int. Conf. on Ultra-Wideband*, pp. 420–425, Sept. 2005.

[2] C.-H. Park and K.-S. Hong, "Source localization based on SVD without a priori knowledge," in *Proc. Int. Conf. on Advanced Commun. Tech.*, pp. 3–7, Apr. 2010.

[3] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Addison Wesley, Boston, 2006.

[4] K. Almuzaini and T. A. Gulliver, "A new distributed range-free localization algorithm for wireless networks," in *Proc. IEEE Vehic. Tech. Conf.*, pp. 20–23, Sept. 2009.

[5] X. Wei, L. Wang, and J. Wan, "A new localization technique based on network TDOA information," in *Proc. IEEE Int. Conf. on ITS Telecommun.*, pp. 127–130, June 2006.

[6] A. Hatami, K. Pahlavan, M. Heidari, and F. Akgul, "On RSS and TOA based indoor geolocation - A comparative performance evaluation," in *Proc. IEEE Wireless Commun. and Networking Conf.*, pp. 2267–2272, Apr. 2006.

[7] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," in *Proc. IEEE INFOCOM*, pp. 1734–1743, Mar.-Apr. 2003.

[8] L. Doherty, K. S. J. Pister, and L. El Ghaoui, "Convex position estimation in wireless sensor networks," in *Proc. IEEE INFOCOM*, pp. 1655–1663, Apr. 2001.

[9] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," *IEEE Personal Commun.*, vol. 7, no. 5, pp. 28–34, Oct. 2000.

[10] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher, "Range-free localization schemes for large scale sensor networks," in *Proc. ACM MobiCom*, pp. 81–95, Sept. 2003.

[11] S. Venkatesh and R. M. Buehrer, "Multiple-access design for ad hoc UWB position-location networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, pp. 1866–1873, Apr. 2006.

[12] D. B. Jourdan, D. Dardari, and M. Z. Win, "Position error bound for UWB localization in dense cluttered environments," in *Proc. IEEE Int. Conf. Commun.*, pp. 3705–3710, June 2006.

[13] I. Guvenc, C.-C. Chong, and F. Watanabe, "Analysis of a linear least-squares localization technique in LOS and NLOS environments," in *Proc. IEEE Vehic. Tech. Conf.*, pp. 1886–1890, May 2007.

[14] D. J. Torrieri, "Statistical theory of passive location systems," *IEEE Trans. Aerospace and Electronic Systems*, vol. 20, no. 2, pp. 183–198, Mar. 1984.

[15] F. K. W. Chan, H. C. So, J. Zheng, and K. W. K. Lui, "Best linear unbiased estimator approach for time-of-arrival based localisation," *IET Signal Process.*, vol. 2, no. 2, pp. 156–163, June 2008.

[16] H. C. So and F. K. W. Chan, "A generalized subspace approach for mobile positioning with time-of-arrival measurements," *IEEE Trans. Signal Process.*, vol. 55, no. 10, pp. 5103–5107, Oct. 2007.