Adaptive Signal Processing Laboratory (ASPL)
Electrical and Computer Engineering Department
University of Florida

Clustering Methods for Extremely Low Frequency Subsurface Signals

ASPL Report No. Rep_2004-12-003

Suju Rajan and K. Clint Slatton

December 12, 2004

Point of Contact:
Prof. K. Clint Slatton
University of Florida; PO Box 116130; Gainesville, FL 32611
Tel: 352.392.0634, Fax: 352.392.0044, E-mail: slatton@ece.ufl.edu

Problem statement:

During fall 2003 and spring 2004, I and Suju Rajan wrote this paper. It was declined immediates when I submitted it to IEEE TGARS because it did not pertain enough to "remote sensing". I will have to edit it heavily to make it suitable for submission somewhere else. In it's current form, it is not state-of-the-art in pattern recognitions so it can't be submitted as a full-length paper to such a journal (too much discussion of things that are already well known in that community). However, there is much good text in it, so I don't wish to lose that. That is why it is being "saved" as an ASPL report.

Clustering Methods for Extremely Low Frequency Subsurface Signals

Suju Rajan[1] *Student Member, IEEE,* K. Clint Slatton[2] *Member, IEEE*

1: University of Texas at Austin, Electrical and Computer Engineering Dept.
Austin, TX 78705
2: University of Florida, Electrical and Computer Engineering Dept.
Gainesville, FL 32611

Abstract – Extremely Low Frequency (ELF) electromagnetic signals are of great interest in a variety of applications, including the transmission of communication signals over long distances in seawater and below the Earth's surface, deep geophysical sounding, and the study of electrical atmospheric phenomena known as sferics. In subsurface signaling applications, wireless electromagnetic methods generally require far less infrastructure than wireline approaches and can provide effective signaling over much longer distances than acoustic methods. However, the ELF band is highly susceptible to interfering signals emanating from electronic and electro-mechanical equipment that is often associated with geophysical exploration. This interference can significantly degrade the demodulation and subsequent analysis of the received signals. To improve the demodulation of ELF communication signals, data can be segmented via clustering methods into sequences representing different signal states. In this work, several clustering approaches are compared on their ability to segment ELF signals in the presence of severe interference. K-Means and 'k-Means like' clustering methods are found to be superior overall.

Keywords – data segmentation, cluster analysis, pattern recognition, subsurface propagation, low-frequency interference

Introduction

Extremely low frequency (ELF) electromagnetic signals (<3 kHz) are used for deep geophysical sounding, underground and undersea signaling, and the study of lightning induced electric field perturbations known as sferics [1] [2] [3]. While these long wavelength signals have the ability to propagate through the Earth's subsurface media and through seawater, this frequency regime is highly susceptible to interference from a wide range of natural and anthropogenic sources [4]. Various approaches have been investigated to mitigate the effects of low-frequency interference, such as protective shielding or redundant sensors placed at different locations, but quite often the interference cannot be satisfactorily removed. In applications where the received subsurface signals are analog geophysical measurements, which often exhibit significant temporal correlation, data corrupted by short bursts of relatively high-power interference can sometimes be removed from the signal with acceptable loss of information. In some cases, the signal source may even be modeled well enough that estimated values can be inserted in place of the removed signal segments [5]. In the case of digitally modulated signals, however, simply removing signal segments corrupted by interference leads to an unacceptable loss of information.

In general, increasing transmitter power or adding redundancy via repeated codes (transmitting longer messages) does not solve the problem because many field-deployed systems have severe constraints on transmitter power consumption. Furthermore, it is not usually possible to fully characterize or even identify all possible sources of interference associated with geophysical exploration infrastructure or other equipment in the vicinity of geophysical signal receivers. The amplitude, frequencies, and duration of the interference sources vary in unpredictable ways, resulting in highly nonstationary interfering signals, thus complicating subsequent demodulation and analysis of the received data. Mitigating approaches such as spread spectrum that have been used to successfully overcome interference in high-frequency (narrow band) applications are not viable for the large bandwidth-to-carrier ratios encountered with ELF communications [6]. However, pattern recognition methods can be used to segment the received data into different signal states, which in turn can be used as *a priori* information for subsequent signal processing. In this work, we investigate and compare different clustering approaches to determine which methods appear best suited for automated interference identification for phase modulated digital signals in the ELF regime. This analysis could also be applied to broadband analog (electromagnetic or acoustic) geophysical signals.

Clustering is the process of finding natural groupings in data [7], where the groupings are defined by some similarity measure such as the distance between two samples in data record. Much work has been done in the clustering of a broad spectrum of signals. Clustering using Kohonen Networks has helped identify crack related interferences in Acoustic Emission Signals [8]. Hidden Markov Models (HMMs) while being used for speech recognition for years have also found application in the clustering of vector time series for monitoring manufacturing machines [9]. Plicker, *et al*. [10] and Guedalia, *et al*. [11] used clustering to characterize non-stationary time series data. EEG signals have long been subjected to clustering techniques in an effort to automate the process of discovering certain biophysical events [12]. [13] is an application of clustering pseudo-HMMs for indexing video signals.[14] and [15] used neural network based approaches for clustering neuron signals and employed simpler methods like template matching that use a set of pre-defined signal segments as cluster prototypes.

Much of the research on interference suppression in digital communications to date has focused on high-frequency signals where interference can be greatly reduced by partitioning the useable bandwidth [16]. In this work, we focus on the automatic detection of interfering signals in feature space. A clustering approach was motivated by the poor performance obtained with correlation methods due to the severe in-band interference. We evaluated the performance of several well-known clustering algorithms on low frequency signals to see if the process of knowledge extraction could be automated. The clusters obtained can be used to form a codebook, which can then be used to classify incoming signal segments as ones containing ambient noise, signal with noise, signal with noise and interference, etc. Since the results of the different clustering algorithms are to be used only as a look-up table, this work focuses only on those clustering algorithms that differ in their inherent bias and assumptions about the data. Speed and scalability are of secondary importance and are not studied in this work. The ultimate goal being the identification of the clustering algorithm that can best discover inherent structure in real-world ELF data that is subjected to interference. We focus on the well known methods of K-Means, Agglomerative Clustering, Self Organizing Maps, Graph Partitioning, and Cluster Ensembles. Cluster validity indices like the Dunn indices, Davies-Bouldin index and the average Silhouette Width are used to determine the “best” algorithm. While these measures are not guaranteed to determine the overall optimal clustering method, a consensus about which algorithms are best suited to the present application is arrived at by comparing the performance of the different methods using these indices.

Section 2 describes a typical received data record obtained by transmitting a phase-modulated signal through the ground. Section 3 deals with the issue of preprocessing of data and identification of a suitable set of features. Sections 4 and 5 explain the different clustering algorithms that are studied and various cluster validation indices that are used to evaluate the performance of each algorithm. Results and conclusions are presented in sections 6 and 7.

Data Description

The communication signals used in our experiments propagated through several hundred meters of subsurface material. The data were acquired in a region of North-central Texas that consists primarily of Cretaceous age marine and near shore deposits [17]. Manual analysis of a subset of the test data was performed to identify signal segments that contain useful information (transmitted signal). This group of useful signal segments can be split into two main categories: signals that are transmitted from the surface down into the subsurface, called the downlinks, and the signals that are sent to the surface from a transmitter deep within the Earth’s surface, called the uplinks. The uplinks are more susceptible to interference because of their low relative power. The surface receiver also recorded broadband noise due to the electromagnetic background environment, periodic noise induced by mechanical rotation of metal parts, and impulsive noise due to the intermittent mechanical or electrical events. Some of the interference signals are capable of causing saturation of the signal receiving hardware resulting in a partial or complete loss of the carrier signal. An example of a test data record is shown in Figure 1, while normalized views of various signal segments of interest are shown in Figure 2.

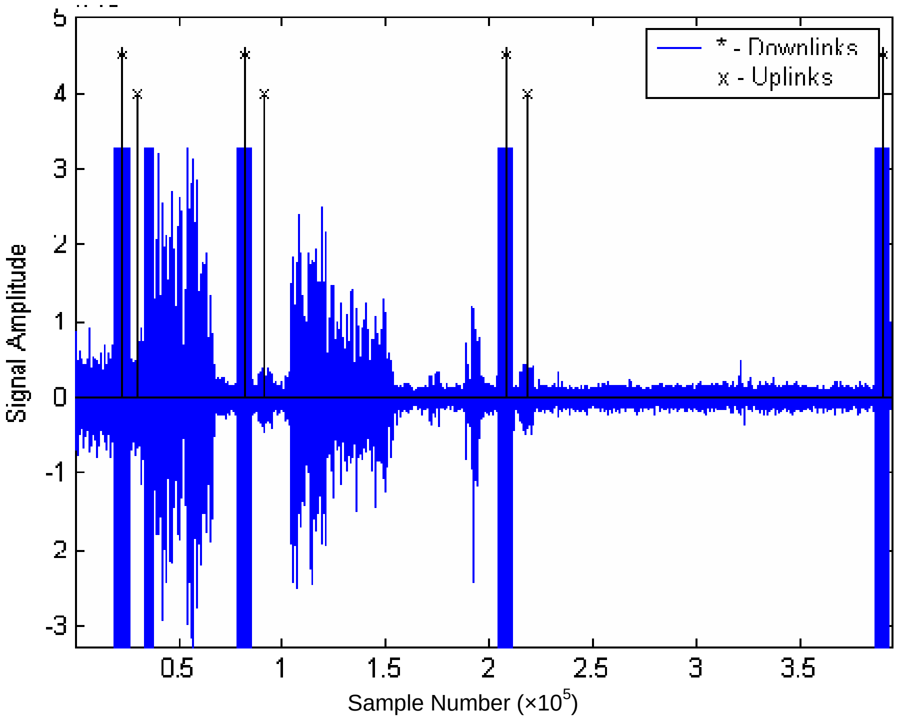


Figure 1: Typical recorded ELF signal (relative to the maximum received amplitude) showing manually identified uplinks (x) and downlinks (*).

The examples in Figure 2 represent the baseline cases of the respective signal segments. In practice the uplink signal segment might be highly corrupted by interference. Since the classification needs to accommodate non-stationary signals, clustering algorithms were used to partition long data records into short segments corresponding to the uplinks, downlinks, corrupted uplinks, and ambient noise.

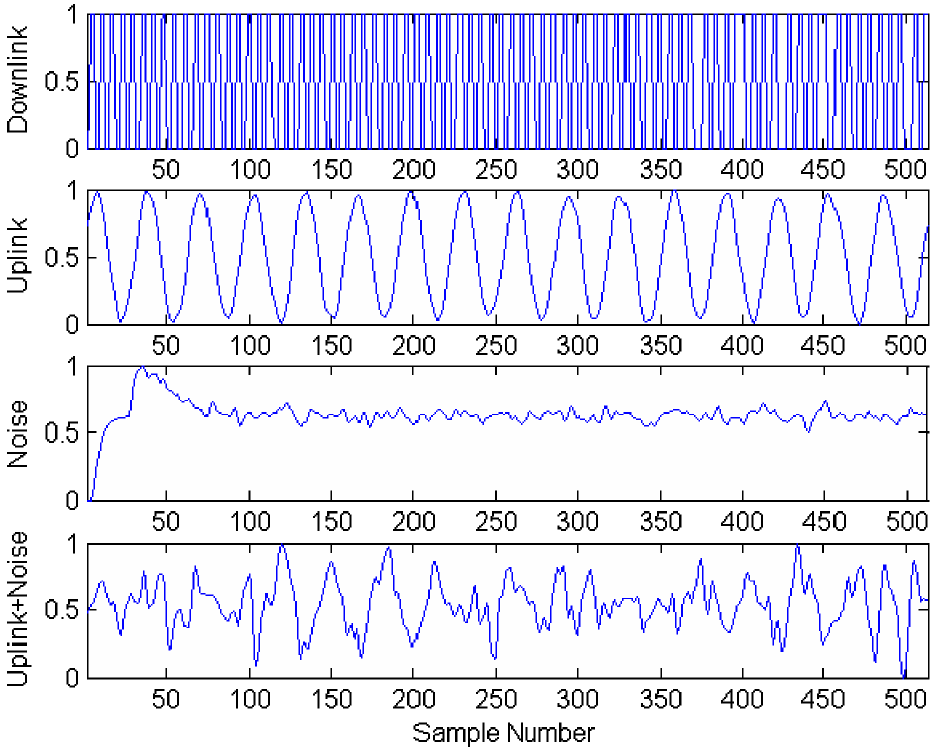


Figure 2: The different signal segments of interest (normalized to unity amplitude).

Feature Space Selection

The success of any clustering algorithm depends to a large extent on the feature space used. Preliminary analysis of the available uplinks and downlinks helped identify a feature space based on the Short Time Fourier Transform (STFT) which proved to be the most efficient among all the feature spaces considered for this specific problem.

The time-frequency localization ability of the Short Time Fourier Transform (STFT) [18] enables STFT-based features to capture the nonstationary frequency information inherent in the signals as a vector. Each signal segment was encoded as a vector of discrete frequency indices by using a 256-point STFT on a Hamming-windowed segment of length 512 and an overlap of 64. Each such STFT transformation yields a matrix of the estimate of the short-term (time-localized) frequency content of the signal. The columns represent the increase in time while the rows of the matrix correspond to increasing frequencies. It was found that the discrete frequency index at which each signal segment achieved its maximum served as a good discriminator for the different signal types of interest. As can be seen in Figure 3 the uplinks achieved their maximum amplitude at the same frequency across all time localizations. Uplinks corrupted by the noise showed some variation about the uplink frequency (about 10%), while noise signal segments exhibited much larger variation. The segments corresponding to downlinks map into a totally different frequency range over the different time localizations. The variance of these indices and the mean amplitude of the signal in each of the segments were also identified as good attributes for encoding the signal segments. Most of the parameters for the STFT computation were determined empirically by applying the transformation to signal segments about which we had prior class information in order to obtain a discriminative feature space.

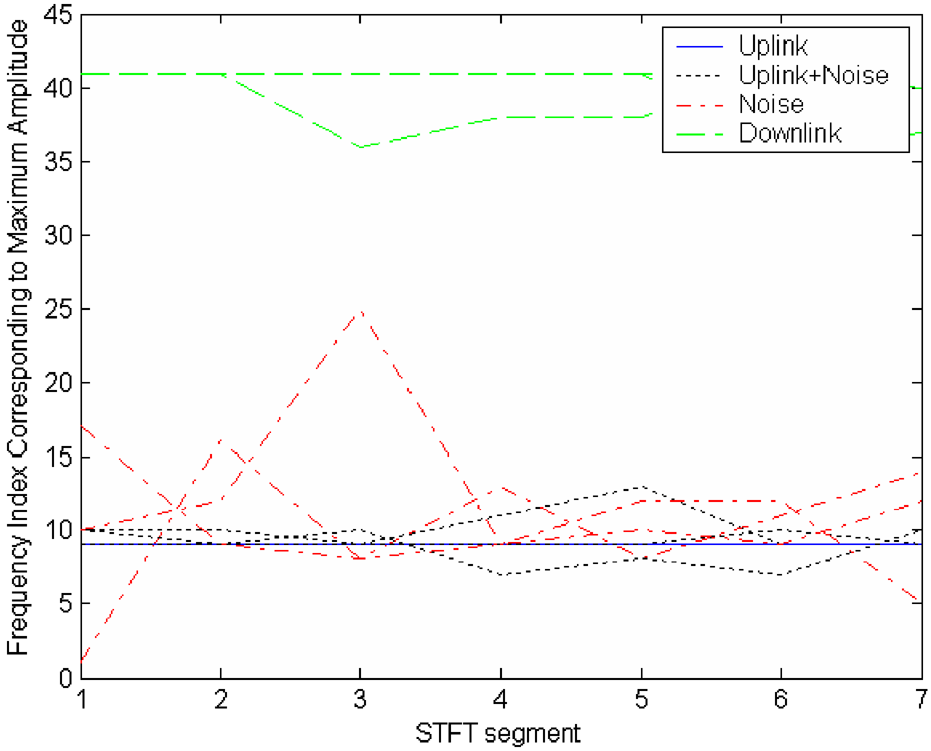


Figure 3: STFT representation of the four different nominal signal segments of interest.

Clustering Methods

No single clustering algorithm will perform uniformly well on all datasets [19] [20] [21]. The huge variety of methods available for representing the data, computing similarity/distance measures and for the aggregation of the data points has resulted in a plethora of clustering algorithms. As mentioned in [19] each of these different methods will produce clusters when presented with data, even if the data itself has no inherent structure. All such clusters produced by these different methods are equally valid as long as the clusters do not occur by chance or by some quirk of the clustering algorithm itself. More often than not, the choice of a particular algorithm and its corresponding parameters is a matter of trial and error in which the practitioner experiments with the available design choices until clusters that meet certain preset criteria are obtained.

While it is impossible to experiment with all the available clustering algorithms and their respective parameter spaces, choosing a few methods that differ significantly from each other in their inherent bias and assumptions about the data can help to focus attention on those algorithms that are similar to the most "successful" one thereby narrowing down the search space considerably. This being the ultimate goal of our work we chose for comparison the following clustering algorithms (1) k-Means (2) Hierarchical Agglomerative Clustering (3) Self Organizing Maps (4) Graph Partitioning and (5) Cluster Ensembles. A short description of each algorithm follows.

Let $X = \{x_1,...,x_N\}$ represent the set of data points. A partitioning (clustering) of these N objects into k clusters results in sets of objects denoted by C_m, $m \in (1,\ldots,k)$ or a label vector λ in $\Re^{N\times 1}$. A super-script on the data point denotes the cluster membership while a subscript refers to a specific data sample. In all cases, the distance measure used between any pair of data points or their variants (such as the mean) is the Euclidean distance denoted by $d(.,.)$.

4.1 k-Means:

One of the simplest and most popular clustering techniques, the k-Means algorithm [22] belongs to a class of algorithms called the partitional methods in which a single partition (clustering) of the data $\boldsymbol{X}$ into k clusters is obtained by optimizing a local or global cost function. The sum of squares criterion, which is widely used, has the following cost function [23]:

$$Cost(C_m) = \sum_{j=1}^{|C_m|} d\left(x_j^m, \overline{x^m}\right)$$

(1)

where $\overline{x}$ denotes the centroid of each cluster. The k-Means algorithm works by choosing k random points as the cluster centers and assigning each x_i to the cluster center that is closest to it based on the similarity measure $d(x_i, \overline{x})$. The new cluster means are then computed and the data points are reassigned until the criterion function converges. It has been shown [19] [20] [24] that the algorithm works well when the clusters are compact and well separated. Besides having to specify the number of clusters, the algorithm has been known to converge to a local optimum that is very far from the global one. Sensitivity to outliers and to the initial choice of cluster centers are some of the other drawbacks.

4.2 Hierarchical Agglomerative Clustering:

As the name implies this method belongs to a group of hierarchical clustering algorithms which yield a dendrogram or a tree-like structure in which the leaves represent the data objects $\boldsymbol{X}$ and the nodes represent subsets of $\boldsymbol{X}$. In the agglomerative method each data point is considered to be its own cluster and pairs of clusters are then successively merged until all objects belong to one cluster. The various agglomerative algorithms differ in the way the subsets of $\boldsymbol{X}$ are merged, i.e. in the way the cost function of merging two sets of objects is defined. Most algorithms use single-link, complete-link or average-link metrics or their variants. In our study we use the average link cost function, which is defined as follows:

$$Cost(C_r, C_s) = \frac{1}{\left(|C_r||C_s|\right)} \sum_{\substack{x_i^r \in C_r \\ x_j^s \in C_s}} d\left(x_i^r, x_j^s\right)$$

(2)

It has been shown [25] that single-link metrics produce straggly, chain-like clusters, complete-link metrics produce spherical clusters and average-link metrics, which attempt to minimize the maximum variance, produces clusters that are between the two. These methods are computationally expensive as the linkage metrics are computed over a proximity matrix of order $N \times N$. Unlike the partitional methods where the data points can be relocated, the hierarchical methods do not revisit clusters that are already formed, which might result in sub-optimal clusterings as a best merge at one stage need not be so in the later stages. A number of variants, which integrate hierarchical clustering methods with other techniques, have been shown to alleviate these drawbacks to a certain extent [24].

4.3 Self Organizing Maps:

Self Organizing Maps (SOMs) [26] are a class of artificial neural networks that are based on the idea of competitive learning. The SOMs normally consist of a set of source nodes that map into a one or two-dimensional lattice of output neurons that learn by competing with one another to selectively respond to certain classes of input patterns resulting in a topographic map. The spatial location of these output neurons in the topography corresponds to features of the input data. The neurons learn by adjusting their weight vectors so as to minimize the Euclidean distance between the input vectors x_i and the weight vectors w_m. The update rule for the weight vectors is given by [27]

$$w_m(n+1) = w_m(n) + \eta(n) h_{i,m(x)}(n)[x - w_m(n)]$$

(3)

where η and h are the user-defined learning rate and neighborhood functions respectively. The variable n represents the iteration over time. The output map is an approximation of the input space and is ordered by the features in the input data. The output map also reflects the distribution of the input data as larger domains of the output space correspond to those vectors that have a high probability of occurrence. The final partitioning produced by the SOMs depends on the choice of the initial weights and potentially other user-specified parameters. Like the k-Means methods, SOMs favor spherical clusters.

4.4 Graph Partitioning:

The problem of clustering has been related to the problem of graph partitioning in which the goal is to partition a vertex-weighted graph G into k unconnected components of approximately equal sizes. In a clustering framework each of the data points are considered as the vertices of the graph, and any two vertices are connected together by an undirected edge weighted by the similarity (inverse of the distance) between those two data-points. The idea is to split the graph G by a set of edges ξ, (x_i, x_j) pairs, such that we obtain k-unconnected components. The set of edges ξ are chosen so as to satisfy the objective function

$$\min_{\xi} \sum_{(x_i, x_j) \in \xi} d^{-1}(x_i, x_j)$$

(4)

It is also assumed that the normalized vertex weights satisfy a balancing constraint, which ensures an equal number of data points in each cluster. Finding an optimal edge separator is a NP-hard problem [28]. However several approximation algorithms and heuristics are available such as METIS [29] and Spectral Bisection [30]. In our study we use METIS, which handles the constrained optimization problem in three phases which involve (i) coarsening, (ii) initial partitioning, and (iii) refinement. The graph partitioning algorithms are better suited to data of large dimensions. The balancing constraint presupposes an equal distribution of data points across all clusters which may not always be the case.

4.5 Cluster Ensembles:

Cluster Ensembles is the method of combining the outputs of different clustering algorithms in order to improve the quality and robustness of the resulting clusters. In this method, a set of labels $\lambda^{(1,\ldots r)}$ from the different clustering algorithms are combined to form a single consensus labeling, where λ^{i} is the label vector produced by a particular algorithm [31]. As long as each of the individual algorithms have different biases and generalize in distinct ways we can say that the cluster ensemble will perform as good as or better than the individual clustering algorithms, when each algorithm is provided with a subset of features or a subset of the data. However there is no guarantee for the better performance of the ensemble technique when the individual algorithms have an unrestricted access to the data. Cluster ensemble methods are more relevant in knowledge reuse [32] and distributed computing frameworks than in our present application, but for completeness we evaluate the performance of the "average clustering" in comparison with that of the individual methods.

5 Cluster Validation

As mentioned previously, any clustering algorithm when presented with data will produce clusters regardless of whether or not the data have any inherent structure. The partitioning of the data thus depends on the inherent characteristics of the data and the clustering algorithm as well as the parameters of the clustering algorithm, such as the number of clusters in the case of the k-Means algorithm. Improper choices of these parameter values will result in a partitioning that is not optimal. Hence some measure of the performance of the algorithms for different choices of parameters is necessary. Such measures, called cluster validity indices, help to determine the presence of structure in the data, the number of clusters, and their validity [33].

The validity indices maybe grouped into three broad categories [34]

1. Indices based on external criteria that make use of some external information that is not available to the clustering algorithm such as the category labels of the data.
2. Indices based on internal criteria which evaluate the clustering result of an algorithm by making use of the quantities and features that are inherent to the dataset.
3. Relative criteria which choose a clustering scheme among different schemes based on relative merit, which is measured according to some pre-specified criterion.

All these indices attempt to measure the quality of the cluster output based on the compactness and the separability of the resulting clusters.

In our framework, since we do not have access to any external category labels and the intent is to compare the performance of the different clustering algorithms, we use the relative criteria as a measure of cluster validity. While most validity indices are used to obtain an optimal number of clusters, in this work these indices serve the purpose of identifying the algorithm that performs best over a range of cluster numbers. To this end we consider three measures of cluster validity, the Dunn Index, the Davies-Bouldin index and Silhouette Width.

5.1 Dunn Index:

The Dunn index proposed in [35] is a measure that attempts to identify compact and well-separated clusters based on the distance between two clusters C_i and C_j and the diameter of each cluster C_i.

$$Dist(C_i, C_j) = \min\{d(x,y)\} \text{ where } x \in C_i \text{ and } y \in C_j$$

$$Diam(C_i) = \max\{d(x,y)\} \text{ where } x, y \in C_i$$

Here, $d(x,y)$ can be any distance measure. In our formulation d is the Euclidean distance.

The Dunn Index can then be defined as

$$DunnIndex = \min_{1 \le i \le C} \left\{ \min_{\substack{1 \le j \le C \\ j \ne i}} \left\{ \frac{Dist(C_i, C_j)}{\max\limits_{1 \le k \le C} \{Diam(C_k)\}} \right\} \right\}$$

(5)

For compact and well-separated clusters the distance between clusters based on the data points in the cluster has to be large while the diameter of each cluster should be small. Given that good clustering algorithms attempt to maximize the inter-cluster distance and minimize the intra-cluster scatter, large values of the Dunn-index correspond to good clusters. However the Dunn index is extremely sensitive to the presence of outliers in the clusters as it is based on the individual data points in each cluster. Hence it has been argued in [36] that more generalized indices based on the Dunn index will yield better information about the cluster quality. These measures however do not perform well for chain or shell-type clusters.

5.2 Davies-Bouldin Index:

The Davies-Bouldin index [37], unlike the Dunn index, measures cluster quality based on both the data-points in each cluster as well as their means. However like the Dunn index the cluster quality is measured as a ratio of the inter-cluster distance and the intra-cluster scatter, which are defined below.

$$Scatter(C_i) = \left(\frac{1}{|C_i|} \sum_{x \in C_i} \|x - v_i\|_2^{\,q} \right)^{1/q}$$

where v_i is the cluster mean and $q > 1$. Then

$$Dist(C_i, C_j) = \left\| v_i - v_j \right\|_w$$

where $w > 1$. Then define

$$R(C_i) = \max_{\substack{j \\ j \neq i}} \left[\frac{Scatter(C_i) + Scatter(C_j)}{Dist(C_i, C_j)} \right]$$

The Davies-Bouldin index is given by

$$\frac{1}{C} \sum_{i=1}^{|C|} R(C_i)$$

(6)

Unlike the Dunn index, a lower value of the Davies-Bouldin index indicates good clustering as we would like to minimize the within-cluster scatter and maximize the between-cluster separation.

5.3 Silhouette Width:

The Silhouette Width [38] measure evaluates the silhouette width for each and every data sample. The algorithm that produces the lowest overall average silhouette width for the entire dataset obtained by averaging over each cluster is regarded as the best clustering.

The Silhouette width is given by

$$S(i) = \frac{(a(x_i) - b(x_i))}{\max\{a(x_i), (x_i)\}}$$

(7)

where $a(x_i)$ is the average dissimilarity of the i^{th} object to all other objects in the same cluster and $b(x_i)$ is the minimum of the average dissimilarity of the i^{th} object to all objects in other clusters (in the closest cluster).

The silhouette width ranges from -1 to +1. A value of +1 indicates a very good clustering, that is the data point has been assigned to the best possible cluster. A silhouette width of 0 means that the data point could have been assigned to either of two clusters, while a value of -1 indicates a bad clustering. The largest overall average silhouette width indicates a good clustering.

6 Experimental Results

Recorded ELF signals transmitted through the subsurface were used in our experiments. Clustering was performed on a few datasets, and the cluster centroids of the "best" algorithm were used to compute a lookup table which was then used to classify the segments of test files to check if the clustering algorithms yielded valid clusters. As stated in [19], though the use of cluster centroids to represent clusters is a popular scheme, it works well only when the clusters are compact and well separated. We do not analyze the different methods of cluster representation in this work, though other methods such as the use of boundary points could have been used. Feature Extraction as detailed in Section 3 was performed on the datasets, and the transformed datasets were then normalized before being subjected to the different clustering algorithms.

Since the k-means and the SOM clustering algorithms are known to be sensitive to the initial choice of cluster centroids and the weights of the neural network respectively, both these methods were run multiple times and the clustering with the lowest mean square error was chosen in both cases. The cluster ensemble algorithm was implemented using the toolbox provided in [31] which makes use of three different consensus functions and evaluates all three approaches against an objective function based on the Average Normalized Mutual Information to pick the best solution. To establish a baseline performance, a random clustering of the data points was also done in which each input sample was assigned a cluster label drawn from a uniform distribution of labels from 1 to k.

The number of clusters was varied from five through twenty and the validity measures for the different clustering algorithms across varying cluster numbers was computed. It has to be kept in mind that higher values of Dunn indices and Silhouette widths and lower values of Davies-Bouldin Indices correspond to good clusterings. Figure 4 shows the variations of the different indices with respect to the number of clusters. It can be seen that k-Means performs the best across all three measures followed by SOMs, Agglomerative Clustering, Ensemble clustering, and then Graph Partitioning. There seems to be no strong consensus for the ideal number of clusters across all the three indices with Dunn index showing five as the optimum number while Silhouette Width chooses fifteen. There seems to be no marked variation in the Davies-Bouldin index across the different cluster numbers as far as the best performing k-Means and SOM algorithms are concerned.

While these indices reveal how compact and separable the clusters are, it is also important that the clusters map into human-interpretable and useful partitions. Figure 5a is an instance of one of the few files in which the different signal segments have been identified manually while Figure 5b shows the cluster memberships superimposed on the signal file for the case of SOM clustering. It can be seen that Cluster 5 clearly maps into the downlinks while Cluster 4 maps into the uplinks. Clusters 1, 2 and 3 correspond to the noisy signal segments. At higher values of cluster numbers the uplinks and downlinks get further divided into yet smaller clusters. The level of granularity required depends on the human user. The different noise clusters can then be studied and whenever a similar signal segment is observed we can choose one of the predetermined filters to extract whatever information that segment may contain or choose to ignore that particular signal segment.

We then measured the performance of the look-up table computed from the cluster centroids on independent test data, and as can be seen from Figure 6a and 6b the uplinks, downlinks and noise segments have been identified correctly-which proves the utility of the clustering algorithms in identifying the signal segments of interest thereby saving much human effort.

We did not study the performance of these algorithms with respect to the computational speed and memory required. SOMs are computationally expensive depending on the rate of convergence of the learning process. Since our objective was to construct a look-up table that could be used offline to classify signals, questions about the relative speed and complexity of the different methods were not addressed.

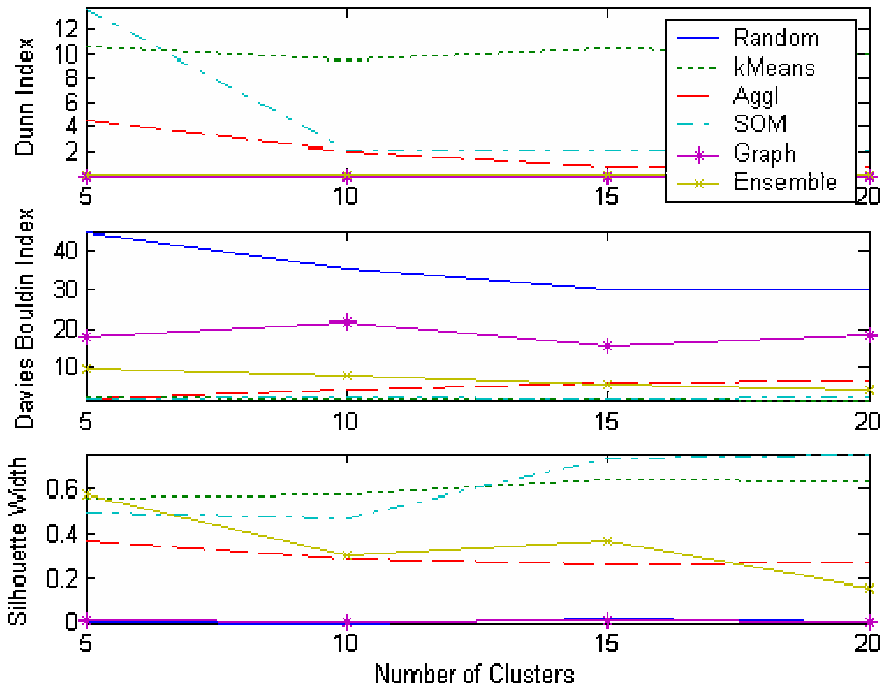
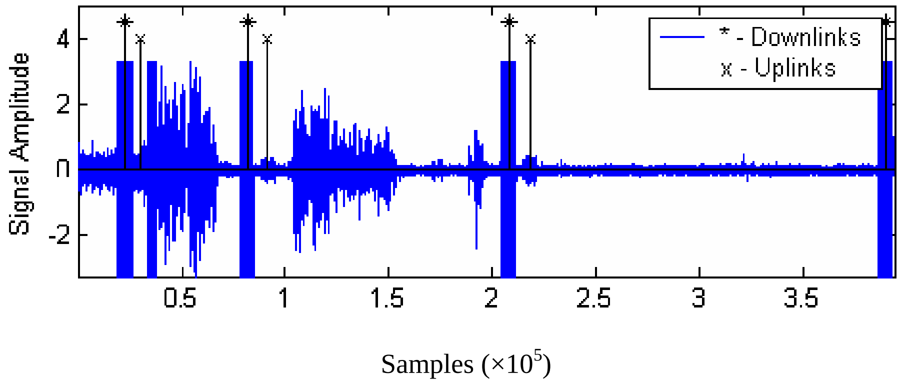


Figure 4: Performance of the clustering algorithms with respect to the validity indices



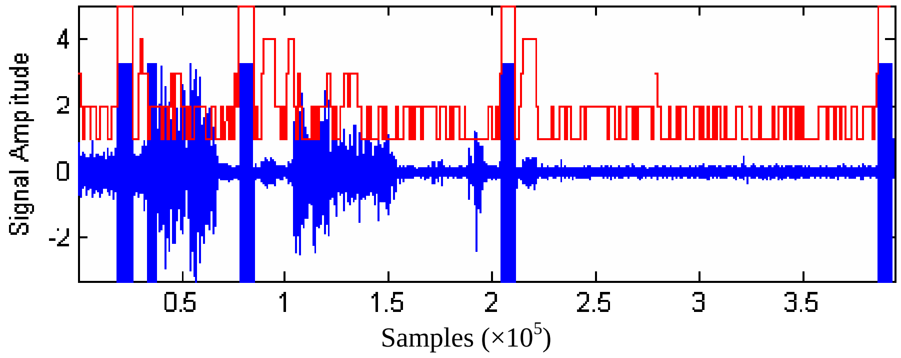


Figure 5a & 5b: Results of clustering using SOMs

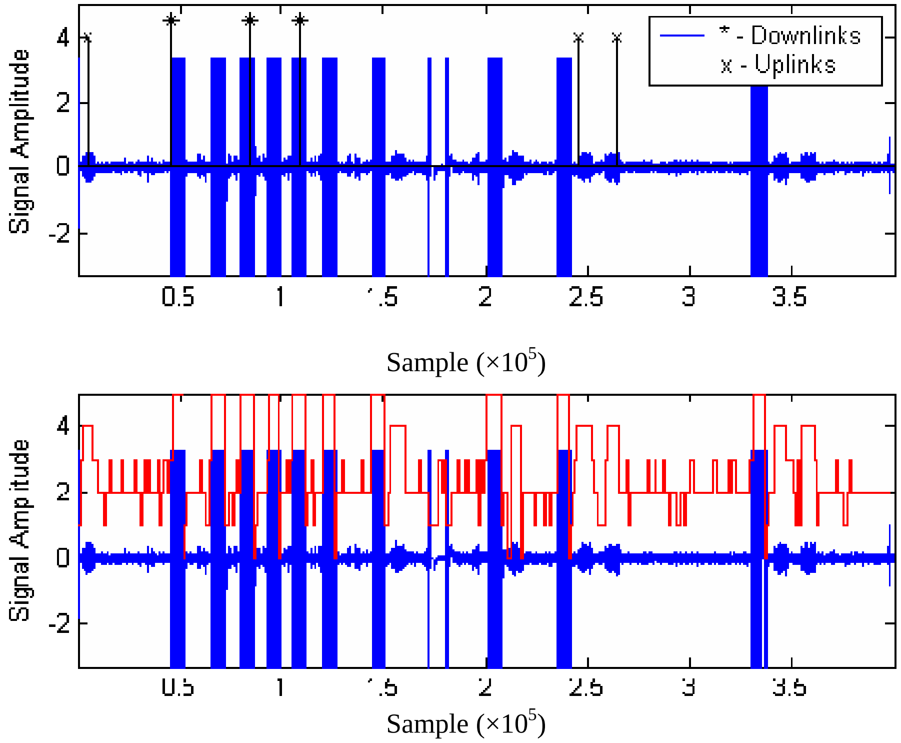


Figure 6a & 6b: Results of pattern matching using the lookup table

Conclusions

We evaluated the performance of five different algorithms for the purpose of being able to identify those signal segments that contain useful information in the presence of severe interference at Extremely Low Frequencies. A feature space defined in terms of the Short Time Fourier Transform was found to be the best suited. The clusters obtained were evaluated in terms of several cluster validity indices. From our studies it was found that the k-Means and SOMs performed best at being able to identify good clusters in terms of the validity indices.

While Graph Partitioning suffers from the drawbacks of the balancing constraint, we chose to investigate this method because one generally does not have *a priori* information about the sizes of different clusters, unlike the case presented here. So while Graph Partitioning proved to be inappropriate for this dataset it might yield good results in those cases where there are roughly equal sized clusters. The Cluster Ensemble techniques did not perform best, as was expected since we are working with the entire feature space [31], but it may be of interest to note that it does much better than the worst method.

Even in the nominal case (see Figure 3), signal segments corresponding to uplinks, noise and corrupted uplinks can show some overlap. In Agglomerative Clustering, since merges are done based on the individual signal segments, it is possible that the above types of segments were merged early on resulting in a huge cluster of these signal segments. Unlike k-Means and SOMs, these clusters once formed are not revisited. Figure 7 implies that our intuition about this method is correct. A similar result was reported in [39] when clustering was performed on text documents.

K-Means and SOMs are similar as they both work with the entire dataset instead of individual data samples and since they are free from any assumptions about the relative sizes of the clusters. Unlike the hierarchical and graph partitioning methods, clusters are revisited and the data samples are relocated until the best partitioning is obtained. As stated in [39] this approach of relocation of the data samples to better clusters is reliable when there is significant overlap between the different clusters. It has to be also borne in mind that these two methods were run several times and the best clustering was chosen, a single-run of the same could, in general, yield clusters similar to that of Agglomerative clustering.

The intent of this work was not to formally establish the optimal algorithm for our particular application as the search space for the solution is very large. Instead we identified which algorithms did well and which did not. It is hoped that this study will serve as a guide to subsequent analysis of non-stationary ELF communication signals.

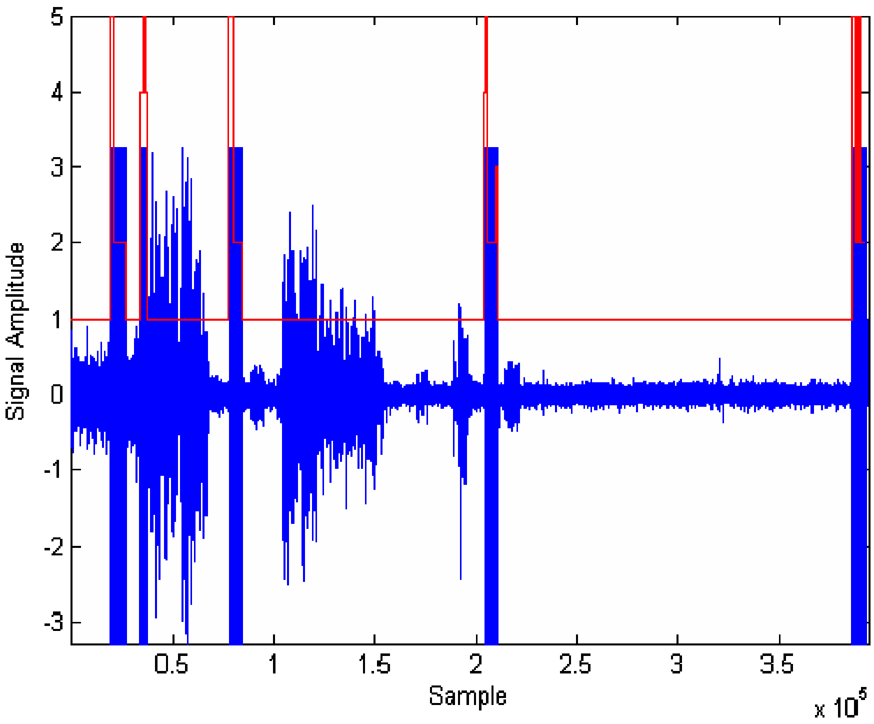


Figure 7: Clustering using the Hierarchical Agglomerative Method

Acknowledgements

This work was supported by the US Department of Energy (grant DE-FC26-02NT41656).

References:

[1] R. Gabillard, P. Degauque, and J. Wait, “Subsurface Electromagnetic Telecommunication--A Review,” *IEEE Trans. on Comm.*, vol. 19, no. 6, Dec. 1971, pp. 1217-1228.

[2] A. D. Chave, A. H. Flosadottir, and C. S. Cox, “Some Comments on the Seabed Propagation of VLF/ULF Electromagnetic Fields,” *Radio Science*, vol. 25, 1990, pp. 825 – 836.

[3] D. A. Chrissan and A. C. Fraser-Smith, “A Clustering Poisson Model for Characterizing the Interarrival Times of Sferics,” *Radio Science*, vol. 38, no. 4, 2003, pp. 17-3 – 17-14.

[4] L.Hasselgren and J.Luomi, “Geometrical Aspects of Magnetic Shielding at Extremely Low Frequencies,” *IEEE Trans. Electromag. Compat*., vol. 37, pp. 409-420, Aug. 1995.

[5] Rusch, L.A.; Poor, H.V., “Narrowband Interference Suppression in Spread Spectrum Communications via Multiuser Detection Techniques,” *Seventh IEE European Conference on Mobile and Personal Communications*, vol., no., 13-15, Dec 1993, pp. 84-89.

[6] E. J. Wegman, S. C. Schwartz, and J. B. Thomas (Editors), *Topics in Non-Gaussian Signal Processing*, Springer Verlag, 1989.

[7] Duda, O.R.and Hart, P.E., *Pattern Classification and Scene Analysis*, 1973, John Wiley & Sons

[8] Emamian, Vahid., Kaveh, Mostafa., Twefik, Ahmed H., “Robust Clustering of Acoustic Emission Signals Using the Kohonen Network,” *Proceedings of ICASSP 2000,* Istanbul, Turkey, 2000

[9] Owsley, L., L. Atlas, and G. D. Bernard, “Automatic Clustering of Vector Time-Series for Manufacturing Machine Monitoring,” *Proceedings of ICASSP 97,* Munich, Germany, April, 1997

[10] Plicker, Shai., and Geva, Amir, B., “Nonstationary Time Series Analysis by Temporal Clustering,” *IEEE Transactions on Systems, Man and Cybernetics*, Vol.30, No.2 , 2000

[11] Guedalia, Issac D., London, Mickey and Werman, Michael., “An On-line Agglomerative Clustering Method for Nonstationary Data,” *Neural Computation*, Vol.11, No. 2, 1999

[12] Karahmeh, Fadi N. and Dahleh, Munther A., “Automated Classification of EEG Signals in Brain Tumor Diagnostics,” *Proceedings of the American Control Conference 2000,* 2000.

[13] Eickeler, Stefan., Muller, Stefan. And Rigoll Gerhard, “Video Indexing Using Face Detection and Face Recognition Methods,” *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Salt Lake City , Utah, May 2001

[14] Sim, A. W. K., Jin, C., Chan, L. W., and Leong, P. H. W., “A comparison of methods for clustering of electrophysiological multineuron recordings,” *IEEE Engineering in Medicine and Biology Society,* 1381-1384, 1998

[15] Millecchia, R., McIntyre, T., “Automatic Nerve Impulse Identification and Separation,” *Computers and Biomedical Research*, Vol.11, pp.459-468,1978

[16] Mundra, P.S.; Singal, T.L.; Kamal, T.S., “Radio frequency interference-an aspect for designing a mobile radio communication system,” *Vehicular Technology Conference*, 1992 IEEE 42nd, Vol., Iss., 10-13 May 1992, pp. 860-865 vol.2.

[17] http://www.dallaspaleo.org/paleo/surf_geol.htm, Dallas Paleontological Society, 2003.

[18] Oppenhiem, A., Scahfer, R., *Discrete-Time Signal Processing*, Englewood Cliffs, Nj: Prentice-Hall: New Jersey, 1989

[19] Jain, A.K., Murthy, M.N., Flynn, P.J., “Data Clustering: A review,” *ACM Computing Surveys*, Vol 31, No.3, pp 264-323, 1999

[20] Berkhin, P., *Survey of Clustering Data Mining Techniques*, Accrue Software, San Jose, CA, 2002

[21] Fasulo, D., **“**An analysis of recent work on clustering algorithms**,”** *Technical Report 01-03-02*, Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, April 1999

[22] Hartigan, J.A., *Clustering Algorithms*, John Wiley and Sons Inc., New York, NY, 1975

[23] Mac Queen, J. “Some methods for classification and analysis of multivariate observations,” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297, 1967

[24] Han, J., Kamber, .M, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, LA, 2001

[25] Kamvar, S.D., Klein, D., Manning, C.D., “Interpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach,” *Proc. of the Nineteenth International Conference on Machine Learning*, July 2002

[26] Kohonen, T., “The Self Organizing Map,” *Proc. of the IEEE*, Vol. 78, No.9,1464-1480, 1990

[27] Haykin, S., *Neural Networks, A Comprehensive Foundation*, McMillan Publishing Company, Englewood Cliffs, NJ, 1994

[28] Garey. M.R., Johnson, D.S., *Computers and Intractability : A Guide to the Theory of NP-completeness*, W.H. Freeman, San Francisco, CA 1979

[29] Karypis, G., Kumar, V., “A fast and high quality multilevel scheme for partitioning irregular graphs,” *SIAM Journal of Scientific Computing*, 20(I);359-392,1998

[30] Hendrickson, B., Leland, R., “An improved spectral graph partitioning algorithm for mapping parallel computations,” *SIAM Journal on Scientific Computing*, 16(2):452-469, 1995

[31] Strehl, A., Ghosh, J., “Cluster Ensembles - A Knowledge Reuse Framework for Combining Partitionings,” *Proc. of 18th National Conference on Artificial Intelligence (AAAI 2002)*, July 2002, Edmonton, Canada

[32] Strehl, A., Ghosh, J., “A scalable approach to balanced, high-dimensional clustering of market-baskets,” *In Proc. High Performance Computing (HiPC 2000)*, Bangalore, Vol. 1970 of *LNCS*, pp 525-536. Springer, December 2000

[33] Pal, N.R.; Bezdek, J.C., “On Cluster Validity for the Fuzzy C-means Model,” *IEEE Transactions on Fuzzy Systems*, Vol. 3, No. 3, pp. 370-379, Aug. 1995

[34] Halkidi, Maria., Batisakis, Yannis., Vazirgiannis, Michalis., “On Clustering Validation Techniques,” *Intelligent Information Systems Journal*, Kluwer Pulishers, 2001

[35] Dunn, J.C., “A Fuzzy Relative of the Isodata Process and its Use in Detecting Compact Well-Separated Clusters,” *J. Cybernetics*, 3(3): 32-57,1973

[36] Bezdek, J.C., and Pal, N.R., “Some New Measure of Cluster Validity,” *IEEE Transactions on Systems, Man and Cybernetics*, Vol.28, No.3, 1998

[37] Davies, D.L., and Bouldin, D.W., “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(4): 2224-227, 1979

[38] Rousseeuw, P.J., “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” 1987. *Journal of Computational and Applied Mathematics*. 20, pp. 53-65

[39] Steinbach, M., Karypis, G., Kumar, V., “A Comparison of Document Clustering Techniques,” *KDD Workshop on Text Mining*, 2000.