Clustering Strategies applied to Binder Identification using Phantom Measurements

Reginaldo Santos and Claudomiro Sales Jr. and Manoel Lima and Caio Rodrigues
Alessandra Araujo and Marx Freitas and Antoni Fertner and João C. W. A. Costa

***Abstract*— This paper proposes a method for the identification of TPs sharing the same binder, based on the analysis of phantom circuit measurements. Herein, phantoming is used to reveal if a 4-wire loop composed by two TPs are close enough in order to be considered in the same binder. K-means and Gaussian Mixture Model are evaluated on S_{11} parameter features obtained from the phantom-mode measurement of two TPs. Also, an automatic method to labelling the clusters and a method to estimate the length which two TPs share the same binder are briefly presented. Laboratory results confirm the accuracy of the methods.**

***Keywords*— Binder Identification, DSL, Phantom-mode measurement, K-means, Gaussian Mixture Model.**

I. Introduction

Loop qualification is a relevant topic in Digital Subscriber Line (DSL) research, since most of the actual telephone lines are inherited from the Plain Old Telephone Service (POTS). Knowledge about telephone loop plants is crucial for operators to have a full exploitation of DSL service, managing the lines and discovering which kind of service can be provided for a specific customer.

Undoubtedly, multimedia services has increased over the last years. Also, users are demanding more and more high data transmission rates to use the bundle of services offered by the cable companies. Hence, lots of technologies for improving data transmission rates in copper systems has been deployed such as G.fast [1] and XG.fast [2]. The usage of high frequencies on DSL systems also offers new crosstalk challenges. Then, studies about mitigation or removal of crosstalk, e.g., on phantom circuits (more details on Sec. II), are also well investigated [3], [4]. While Fiber-To-The-Home (FTTH) is not completely deployed, regarding the optical fiber wire connecting the central office to every customers premises, there will still be an important role to play for DSL systems.

Most Loop Topology Identification (LTI) techniques in DSL focus on a single loop and no information is given about which other customer's loops are near to it. This paper expands the loop qualification principle from the single loop identification approach to binder – structure that groups twisted pairs (TPs) of telephone lines – and cable identification. Knowledge about how the TPs are distributed along the cable contributes for a better management of the telephone transmission lines network, such as prediction of crosstalk noise, since most of the noise comes from TPs in the same binder.

To the best of our knowledge, no machine learning algorithms have been studied for solving the binder identification issue. Nevertheless, there are some important investigations in the same field of study. In [6], a binder identification method is presented, based on quiet line noise measurements. This method takes considerable time to produce results (a few days are required to identify closer TPs) and no information is given about the binder length shared by the TPs. In [7], a study is presented revealing the possibility of binder identification via interpretation of line parameters such as the input impedance and reflection of the phantom circuit composed by two TPs which are intended to be identified. It is pointed out that the phantom circuit composed by TPs in different binders presents high impedance, which is an evidence for the identification. Although the study was done, no direct application was proposed. Also, the definition of binder is not clear. They seem to be working more likely on cable identification than binder identification. It is also pointed out that the length of the phantom circuit can be found by using LTI technique as if it were a common TP loop, but no further details are given, as the condition in which this length can be discovered.

The main focus of this work is to define, via one-port phantom measurements from the transmission device, when two TPs are sharing the same binder and which length they go together into the binder. In this paper, we present an automatic method for binder identification using four features extracted from scattering parameter measures in phantom-mode between two TPs, S_{11}^{PM}, where PM stands for phantom-mode. Two clustering strategies are used to reach the proposed goal. K-means and Gaussian Mixture Model (GMM) are machine learning algorithms used as unsupervised learning method, characterizing the appropriate technique for binder identification issue.

The paper is organized as follows: Sec. II presents the theoretical explanation of phantom-mode signal; Sec. III presents the features for binder identification; Sec. IV presents the methodology; Sec. V presents the analysis and results; Sec. VI presents the concluding comments.

II. Phantom-mode signal

Signal transmission using phantom-mode is a technique primarily used for creating a new virtual communication channel. The so called phantom circuit (Fig. 1) is obtained using one TP (two wires) for transmitting common mode

Applied Electromagnetism Laboratory, Federal University of Pará (UFPa), 66075-900, Belém, Pará, Brazil, E-mails: {reginaldo.filho, manoel.lima, caio.rodrigues, alessandra.araujo}@icen.ufpa.br, {cssj, jweyl}@ufpa.br, marx.freitas@itec.ufpa.br, antoni.fertner@ericsson.com. This work was partially supported by Innovation Center, Ericsson Telecomunicações S.A., CNPq, CAPES and FAPESPA.

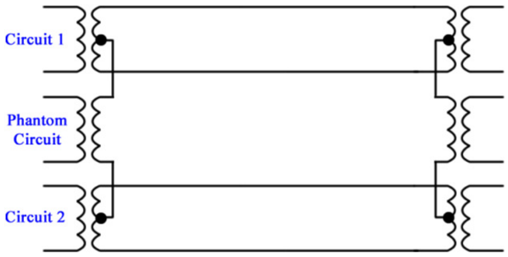

Fig. 1. phantom circuit configuration.

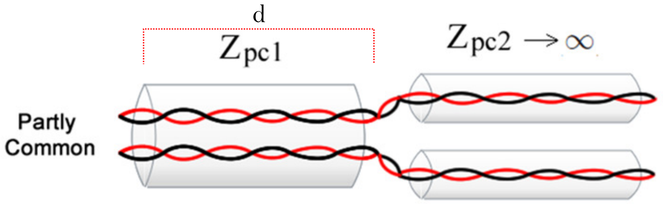

Fig. 3. Two TPs sharing a first common binder and subsequently splitting into different binders. The position where TPs going into different binders behaves like an open-circuit.

signals traveling in the positive direction and other TP (two wires) for common mode signals traveling in the negative direction. To achieve this, two center tapped baluns are used in order to transmit and receive the phantom-mode signal.

The side circuits are considered as a single conductor for the phantom circuit. For a proper functioning of the phantom signal, there should be a proper balancing between the pairs which are composing it, because a difference in the impedance series or in the parallel admittance may generate common mode currents that will cause a coupling between the phantom circuit and the side circuits [8]. Some factors may limit the transmission using the phantom-mode: a) Distance between the pairs will decrease the capacitive and inductive coupling of the phantom circuit; b) TPs in different binders will be more unbalanced than TPs in the same binder, since the TPs in the same binder are twisted together with different twist rate.

The medium between two TPs is inhomogeneous, composed by air, metallic wires and the dielectric between the TPs, which will strongly influence the characteristic impedance of the phantom circuit. This inhomogeneity, including the distance variation between the TPs (resulting in non-uniformity), makes general physical modeling of phantom circuits a difficult task. When both TPs of the phantom circuit are in the same binder, this variation of the medium and the greater proximity between them cause a lower mismatch between the measurement device and the phantom circuit, allowing the signal to propagate along the line, creating standing waves (Fig. 2).

Phantom circuits composed by TPs in different cables have strong Near-End Reflection (NER) caused by distance and unbalance between pairs. Thus, most part of the signal is reflected before propagating along the line, reducing the amplitude of the standing waves created in the frequency domain (Fig. 2). For TPs inside a cable but in different binders, the behavior of S_{11}^{PM} tends to be the midterm between same binder and different cables cases.

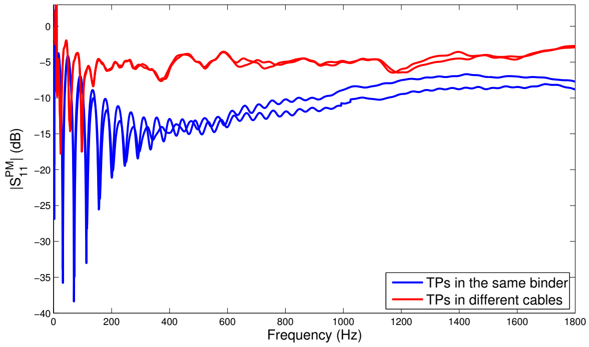

Fig. 2. Phantom measurements in the same binder and in different cables.

Another relevant characteristic of the phantom circuit is that when the distance between close TPs suddenly increases, i.e., when they go into different binders or cables as shown in Fig. 3, the characteristic impedance of the phantom circuit (Z_{pc}) will also increase. In effect, it can be considered as the impedance of an open-circuit in the phantom transmission line. These different behaviors of S_{11}^{PM} will be exploited in this paper for identifying if TPs are in the same binder and estimating the length that TPs share inside the cable.

III. Features for binder identification

The presence of periodicities on the measured S_{11}^{PM} parameter was revealed as a relevant evidence for identifying TPs in the same binder. The following notations demonimate the features: f_1) variance of the period of $|S_{11}^{PM}|$: σ_p^2; f_2) variance of the magnitude of $|S_{11}^{PM}|$: σ_M^2; f_3) number of equivalent spectral lines applying the PSD equation in S_{11}^{PM}, $\mathcal{F}(S_{11}^{PM}(f))$: n_Φ; f_4) the first point of the reflective time domain response calculated from the inverse Fourier transform of S_{11}^{PM}, $\mathcal{T}(S_{11}^{PM}(f))$: $\mathcal{A}_1$.

The variance of the period (σ_p^2) is estimated by calculating the mean square error of the distance between peaks from the magnitude of $|S_{11}^{PM}|$. The signal is divided into K segments and for each segment is applied Eq. 1.

$$p_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k - 1} \| t_{k,i+1} - t_{k,i} \|^2, \qquad (1)$$

where p_k is the variance of the segment k, $t_{k,i+1}$ and $t_{k,i}$ are the position of two consecutive peaks within a segment k and n_k is the total number of peaks in k. Finally, the mean value of all calculated variances is the σ_p^2 feature. As revealed before, S_{11}^{PM} from TPs in the same binder has a well-defined periodicity, then σ_p^2 tends to be smaller. As the frequency increases, the periodic behavior becomes less clear, so it is recommended not to use very high frequencies.

The variance of the magnitude (σ_M^2) is more related to the influence caused by attenuation of the phantom circuit on $|S_{11}^{PM}|$. Two TPs in different cables have a very high characteristic impedance, causing strong NER. In contrast, when the two TPs are situated in the same cable and even in the same binder, the characteristic impedance is small and most energy propagates along the line. This effect is intermediary for TPs in the same cable but in different binders. The variance of all peaks on the $\left|S_{11}^{PM}\right|$ is σ_M^2 feature.

The third feature (n_Φ) applies the PSD, which reveals the harmonic components (spectral lines) of a signal [9], in S_{11}^{PM} as it were a time-domain signal. This operation is indicated by expression $\mathcal{F}(S_{11}^{PM}(f))$, which means the number of equivalent spectral lines, named n_Φ. If n_Φ is small, meaning that $\mathcal{F}(S_{11}^{PM}(f))$ has few peaks, this indicates a well-defined periodicity of the signal and the measurement probably corresponds to TPs in the same binder.

The fourth feature ($\mathcal{A}_1$) attempts to capture the level of mismatch between the input impedance of the phantom circuit and the impedance of the measurement equipment, via the NER level. Throughout experimental tests, a clear difference among the NER levels of TPs in different cables, different binders and same binder of phantom circuits has been verified. The method uses an operator $\mathcal{T}$ that is the reflective time domain response calculated from the inverse Fourier transform of S_{11}^{PM}: $\mathcal{T}(S_{11}^{PM}(f))$. The fourth feature is the first point of $\mathcal{T}(S_{11}^{PM}(f))$, called $\mathcal{A}_1$.

Since every phantom measurement with different pairs will exhibit different features, it is possible to characterize a phantom circuit measurement as a sample by a vector $f = (\sigma_p^2,\, \sigma_M^2,\, n_\Phi,\, \mathcal{A}_1) = (f_1, f_2, f_3, f_4)$, forming a space $\varphi \rightarrow \Re \times \Re \times Z \times \Re$. As will be seen, the pattern of a sample will be identified based on its location on φ, with the estimate accuracy depending on its position and choice of features.

IV. Methodology

The dataset was filled by 267 samples of S_{11}^{PM} measurements performed using the setup shown in the Fig. 4 through 164 different TPs from a cable farm present in our DSL lab. The cables used in this work are TEL 313 000 ELQXBE and TEL 481 02 ELAFQBU/120. The former is an insulated cable composed by 3 binders holding 10 pairs each and the latter is a shielded cable (aluminum foil) with only one binder, holding 16 pairs. Each sample is constructed by the feature extraction $(\sigma_p^2,\, \sigma_M^2,\, n_\Phi,\, \mathcal{A}_1)$, formming the space $\varphi \rightarrow \Re \times \Re \times Z \times \Re$. Thus, the dataset is a vector $\mathbf{X} \in \varphi^m$, with $m = 267$ samples. Each class (same binder, different binders and different cables) has 89 S_{11}^{PM} measurements.

A. Stratified cross-validation

Cross-validation is called stratified when the original dataset is randomly particioned into k equal size subsamples (also called k-folds) and there is an effort to maintain the same original proportion of each class in each fold. The stratified cross-validation runs k times, where each time a single fold is retained as the validation set and the remainig $k-1$ folds are used as the training set. Each fold is used exactly once as the validation set. The goal of the cross-validation is to generalize the results by statistical analysis that represents the insights of how a model will behave with an unknown dataset, and also reduces the so called overfitting. The output is the average accuracy of correctly classified samples from each round.

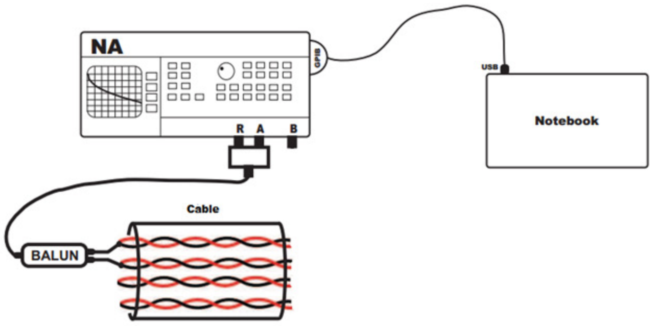

Fig. 4. Setup for one port phantom measurements. The Network Analyser is controlled by a computer and generates a common mode signal that is converted in two differential mode signals, connected in parallel to both TPs.

Herein, the dataset $\mathbf{X}$ is stratified into 10-folds. There will be a combination of feature extraction for each technique to find the best one for the proposed binder identification issue. Later, that best feature combination will be explored in the same dataset $\mathbf{X}$ under specific experiments.

B. Machine learning algorithms

There are several methods for cluster analysis in the literature. In this paper, two consecrated clustering strategies are evaluated for the binder identification issue: K-means and Gaussian Mixture Model (GMM). The former is well known as hard clustering, where each sample from the dataset is classified into exactly one clusters. Samples from the same cluster are as similar as possible and samples from different clusters are as dissimilar as possible, more details can be found in [10]. The latter is a soft clustering, where sample can be classified into more than one cluster with some membership levels associated to each cluster. GMM uses the expectation-maximization (EM) algorithm to optimize its parameters, more details in [11].

C. Labelling clusters

When working with unsupervised clustering strategies, an important issue appears: the clusters must be labelled in order to calculate the accuracy of the model. That task is not trivial and it is usually done by a specialist who has the specific knowledge to discern the clusters and to give labels to them.

In this paper, the embedded knowledge is responsible for labelling the clusters found in the clustering techniques. The clusters can have their labels automatically defined through the fourth feature ($\mathcal{A}_1$) in a decreasing order of value: different cables, different binders and same binder.

D. Sharing length identification through S_{11}^{PM}

The LTI technique performed in this paper is used to determine the length d that is shared by TPs in the binder (Fig. 3). The length is estimated through an automated technique to find the peak related to the point that TPs split out in different cables, which is equivalent to an open-end. First, the method identifies all real peaks in the time-domain response. Second, two thresholds related to the horizontal distance between the rising and falling edges of the time-domain pulse and the vertical discrepancy between these edges are assumed. If both conditions related to these thresholds are fulfilled for the first peak found, the length shared by the TPs is calculated from the location of this peak.

V. Results

The results were generated by combining all four features extracted from S_{11}^{PM} measurements. The goal was to identify what is the best combination of features for the binder classification using clustering techniques. After the identification of the best feature combination, there will be an exploration of this information in the dataset.

K-means results are presented in the Table I (only combinations with average accuracy above 65% are presented). The results were sorted by the average accuracy rate (last column in the table). The percentages mean the correctly classified samples regarding the three classes. The best combination was using feature $[\mathcal{A}_1]$ only, reaching 100% of hit rate for different cables (DC) and different binders (DB) classes, and 67.42% of hit rate for same binder (SB) class. There is a perfect match in differentiating different cables and different binders classes, but the best combination of feature has difficulties with the samples from same binder class.

Knowing the best combination of features, the goal now is to explore that best combination in the dataset $\mathbf{X}$ in order to discover how accurate is the best combination after applying it in the whole dataset. Thus, three tests were performed:

1) Runs the clustering technique with stratified 10-fold cross-validation. Computes the average accuracy among all classifiers and the overall average.
2) Accomplished during the first step, retain the best classifier found in the cross-validation and highlights its average accuracies in the validation set.
3) Runs the best classifier in all dataset in order to verify its overall accuracy.

Table II resumes the accuracies found in the tests. As it can be seen, K-means technique produces notable results with overall accuracy around 90%, hitting 100% of hit rate in all different binders and cables samples and 67.42% for the same binder samples. That classifier had a clear difficulty in identifying the same binder S_{11}^{PM} measurements, which was expected due to the overlap of feature $\mathcal{A}_1$ values for samples of same binder and different binders. Fig. 5 shows the clusters formed by K-means where is possible to notice the nearest samples between same binder and different binders classes. However, despite of that percentage error of 10%, these lines will still be classified in same cable (considering the leftmost two classes as one) as result of the clear separation of the feature $\mathcal{A}_1$ for lines in different cables and same cable.

TABLE I

Average accuracy for each combination and class by K-means.

#	Features	SB (%)	DB (%)	DC (%)	Avg. (%)
1	[4]	67.42	100	100	89.14
2	[2,4]	78.65	69.66	95.51	81.27
3	[2]	73.03	73.03	89.89	78.65
4	[2,3]	83.15	40.45	78.65	67.42

TABLE II

K-means accuracy for each class using feature $\mathcal{A}_1$.

Type	SB (%)	DB (%)	DC (%)	Avg. (%)
Mean	66.29	100	97.75	88.01
Best	88.89	100	100	96.3
Best in all Dataset	67.42	100	100	89.14

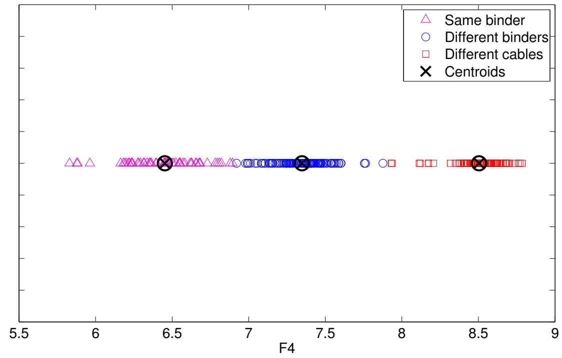

Fig. 5. K-means clusterization result for the best classifier in all dataset.

The Table III presents all possible combinations of feature for GMM technique. The winning combination with the best average accuracy was the same of K-means technique, which was feature $\mathcal{A}_1$ only, reaching 91.3% in the average accuracy.

Applying the same three tests as in K-means technique regarding the winning combination, GMM produces the results in the Table IV. The table confirms that classifying correctly a same binder S_{11}^{PM} measurement, which means two TPs in the same cable, is much difficult than classifying a different cables S_{11}^{PM} measurement. When the best classifier is applied in the whole dataset, GMM reaches 90.64% of overall accuracy.

GMM also produces the Posterior Probabilities (PPs) of each sample regarding the three clusters found in the dataset. Fig. 6 (top) depicts the PPs regarding the different cables

TABLE III

Average accuracy for each combination and class by GMM.

#	Features	SB (%)	DB (%)	DC (%)	Avg. (%)
1	[4]	76.25	98.75	98.9	91,3
2	[2]	77.53	61.8	92.13	77.15
3	[2,4]	65.17	53.93	95.51	71.54
4	[2,3]	83.15	38.2	78.65	66.67

TABLE IV

GMM accuracy for each class using feature $\mathcal{A}_1$.

Type	SB (%)	DB (%)	DC (%)	Avg. (%)
Mean	78.65	94.38	97.75	90.26
Best	100	100	100	100
Best in all Dataset	77.53	96.63	97.75	90.64

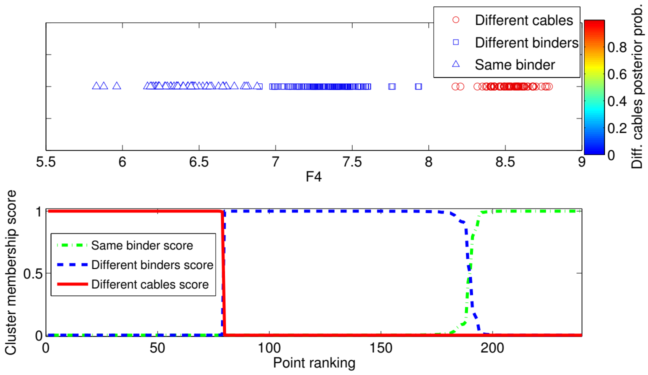

Fig. 6. GMM results for the best classifier: (top-figure) the posterior probabilities considering the component Different cables; (bottom-figure) Presents the cluster membership score of each sample from the training set.

cluster in the training set which make possible to confirm the clear separation between different cables class with high PPs and the other ones with low PPs. Fig. 6 (bottom) shows the membership score of each sample regarding the clusters. There are just two line intersections among the three clusters with few samples with compromised membership scores. The lower the number of compromised membership scores, the better the differentiation between classes.

A. Sharing length identification results

The S_{11}^{PM} dataset used to evaluate the sharing length identification is formed by 78 samples of same binder and 65 samples of different binders, which is 71% composed by lines of 500 m long and around 29% composed by lines of 150 m, 200 m and 250 m long. The following results compiles the average error distribution with its standard deviation: a) same binder obtained 6.4 ± 4.52; b) different binders obtained 12.94 ± 4.79. The general average error is less than 10%, considering the whole range of tested TPs length (150-500 m). Table V gives another view of the errors separated by tracks. For almost 56% lines, the error was below 10%. Around 43% with error between 10-20%. Besides the proximity, TPs in the same binder are twisted all together but with different twist-rate than other binders. Thus, the TPs of a binder have larger coupling than TPs in different binders. Hence, the reflective time domain response and one-port scattering parameter in phantom-mode are very similar to measurement in differential operation mode of individual TPs, which favors the LTI techniques used with only 21.8% between 10-20%. On the other hand, the lower coupling between TPs in different binders decreases the quality of estimation with only 29.2% of the lines in different binders with errors below 10%.

VI. Conclusion

This paper developed a general method for binder identification, i.e, a method that can work with any phantom measurement. Analysing the results, we can afirm that cable identification is an easier task relative to binder identification. It probably happens because on cable identification, distance and material (shield) between the conductors strongly affects the characteristic impedance, creating a strong signature on the data with a great influence in the classification task.

TABLE V

Error per track (%)

Track	Error	SB	DB
[0-10]	55.94	78.21	29.23
[10-20]	43.36	21.79	69.23
[20-30]	0.7	0	1.54
[30-100]	0	0	0

GMM outperformed K-means in the automatic classification of a S_{11}^{PM} measurement. Also, GMM is a soft clustering technique, which produces a membership score for each sample to belong to a cluster. On the other hand, K-means is a hard clustering technique, assigning always crisp values for each sample belonging to a cluster. But it is still a simple, robust and fast clustering technique.

The algorithm for sharing length identification maintained the total error rate below 10%, showing that from different binders S_{11}^{PM} measurements is more difficult to estimate the length of the line than same binder S_{11}^{PM} measurements. Physical modeling is being studied and developed to improve the physical understanding of the problem and the methods developed.

References

[1] M. Timmers, M. Guenach, C. Nuzman, and J. Maes, "G.fast: evolving the copper access network," *Communications Magazine, IEEE*, vol. 51, no. 8, pp. –, 2013.

[2] W. Coomans, R. Moraes, K. Hooghe, A. Duque, J. Galaro, M. Timmers, A. van Wijngaarden, M. Guenach, and J. Maes, "Xg-fast: Towards 10 gb/s copper access," in *Globecom Workshops (GC Wkshps), 2014*, Dec 2014, pp. 630–635.

[3] C. Leung, S. Huberman, K. Ho-Van, and T. Le-Ngoc, "Vectored dsl: Potential, implementation issues and challenges," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 4, pp. 1907–1923, 2013.

[4] A. Forouzan, M. Moonen, M. Timmers, M. Guenach, and J. Maes, "On the achievable bit rates of dsl vectoring techniques in the presence of alien crosstalkers," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, Dec 2012, pp. 3086–3091.

[5] J. Gambini and U. Spagnolini, "Wireless over cable for femtocell systems," *Communications Magazine, IEEE*, vol. 51, no. 5, pp. 178–185, May 2013.

[6] W. Rhee, B. Lee, I. Almandoz, J. Cioffi, and G. Ginis, "Binder identification," U.S. Patent US 8 073 135 B2, 2011.

[7] C. Neus, W. Foubert, L. Van Biesen, Y. Rolain, P. Boets, and J. Maes, "Binder identification by means of phantom measurements," *Instrumentation and Measurement, IEEE Transactions on*, vol. 60, no. 6, pp. 1967–1975, 2011.

[8] C. F. MYERS and L. S. CROSBY, *Principles of Electricity applied to Telephone and Telegraph Work.* AMERICAN TELEPHONE & TELEGRAPH CO., 1953.

[9] K. Kim, I. Akbar, K. Bae, J.-S. Um, C. Spooner, and J. Reed, "Cyclostationary approaches to signal detection and classification in cognitive radio," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, 2007, pp. 212–215.

[10] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.