

# Use of Hidden Markov Models for Partial Discharge Pattern Classification

L. Satish and B. I. Gururaj

Department of High Voltage Engineering,  
Indian Institute of Science, Bangalore, India

## ABSTRACT

The importance of partial discharge (PD) measurements for diagnosis of defects in insulation systems is well known. The image patterns obtained in these measurements contain features whose analysis leads to identification of the PD cause. These features are the phase position and amplitudes of PD pulses appearing on the image pattern (usually displayed on elliptic time base on conventional detectors). There is a close similarity between PD signals and speech. Both are time-varying and similar in behavior. Hidden Markov models (HMM) have been very successful in modeling and recognizing speech. Hence, an attempt was made to employ them to classify PD image patterns. Basis for selection of model and training parameters and the obtained results are discussed. It is shown that successful recognition of PD image patterns using HMM is possible. The ability of HMM to classify some actual PD image patterns has also been ascertained.

## 1. INTRODUCTION

PARTIAL discharge (PD) measurement has gained wide acceptance as a diagnostic tool to assess the adequacy of insulation systems in power apparatus during and after manufacture. In this measurement, the PD pulses are displayed in relation to the electrical phase angles at which they occur; an elliptic time base is often used for this purpose. Useful features available from these PD image patterns are the number of pulse groups, relative phase positions and spreads, behavior with change in voltage, duration of voltage application, etc. These features are used by experts for PD interpretation. Over the years, systematic collections of various PD image patterns have been made. These and annotations on them are available in CIGRE WG 21.03 [1] and in [2]. However, specific clas-

sification of any image pattern obtained in a PD test to one from the listed types still needs human expertise because of inherent randomness. Codifying these randomly varying feature information into rules is very difficult, if not impossible. Therefore, a need arises to automate the interpretation by an approach possessing abilities for learning from examples.

Consider, the time domain waveform Figure 1(a) of one cycle of a PD signal. PD is a stochastic process in which phase position, number and amplitude of pulses vary randomly with time, duration of applied voltage and insulation aging. These variations are within certain lower and upper bounds, which primarily depend on the PD cause. These signal variations translate on to the image pattern as changes in phase position, phase spread of pulse groups and amplitudes. Thus, the PD process can be described

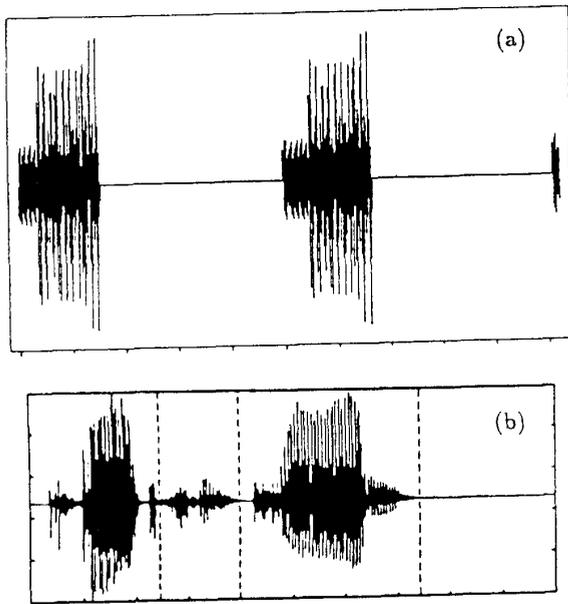


Figure 1.

(a) One cycle of a PD signal. (b) Speech signal due to utterance, "Collect from Tom (silence)". Adapted from [3], Figure 5(a).

mathematically as a time-varying phenomenon. Figure 1(b) gives the time domain speech signal due to an utterance, "Collect from Tom (silence)" [3].

It is well recognized in speech understanding that it is the temporal order of speech signals (analogous to phase information in PD) which conveys information. It varies with individuals, accent, context, emotion, and environment (similar to variations with time, duration of applied voltage, aging etc. in PD). Thus these two processes are similar in this respect.

In the speech domain, hidden Markov models (HMM) have been exceptionally successful for tasks such as recognition of isolated and connected words (speaker dependent and independent), and, to some extent, continuous speech. Some of the important advantages of using HMM are:

1. The ability to learn from examples. The model parameters are automatically tuned to represent the inputs.
2. The temporal characteristic of the input signal (left-right model) is modeled inherently.
3. The tolerance to randomness observed in the input signal due to probabilistic formulation is implicit.
4. It does not need a *a priori* distribution of inputs for estimating the parameters, which is usually not the case in other statistical approaches.

Close similarity between speech and PD signals suggested exploration of HMM use for PD pattern classification.

This paper presents results of this approach to model and automatically classify PD image patterns using HMM. After an introduction to HMM, the methodology and algorithms for evolving them are explained. Results obtained are then discussed and the utility of this approach has been evaluated. A brief account of the approach's ability when fed with five types of actual PD image patterns is given. This approach has been compared with that of neural networks which has gained prominence in recent years.

## 2. BRIEF INTRODUCTION TO HMM

### 2.1 DESCRIPTION OF HMM

ALTHOUGH the basic theory of Markov chains is well known from the beginning of this century, it is only in the last decade that it has been explicitly and successfully applied for different speech recognition tasks. Lack of an optimization algorithm for estimating the model parameters from example data delayed its implementation. These advances were reported in a series of classic papers by Baum and his colleagues in the late 1960's (a principal reference being [4]). Thereafter, it was applied for speech recognition by Baker [5] and, from then on, numerous attempts have been made and are still continuing on various tasks/versions of automatic speech recognition. The results obtained so far are very encouraging.

An HMM is defined by Rabiner [6, 7] as a doubly stochastic process comprising an underlying stochastic process that is not directly observable but can only be visualized through another set of stochastic processes that produce the sequence of observations. In simpler terms, it is a collection of states connected by transitions and emitting an output during each transition. The model is named 'hidden' since the state of the model at a time instant  $t$  is not observable directly.

There are basically two types of HMM. The ergodic, or fully connected HMM, is one in which every state can be reached (in a single or finite number of steps) from every other state of the model. The other type is the non-ergodic or left-right model in which, as time increases, the state index also may increase i.e., the states proceed from left to right (Figure 2(a)). All other HMM are merely a combination of these two types.

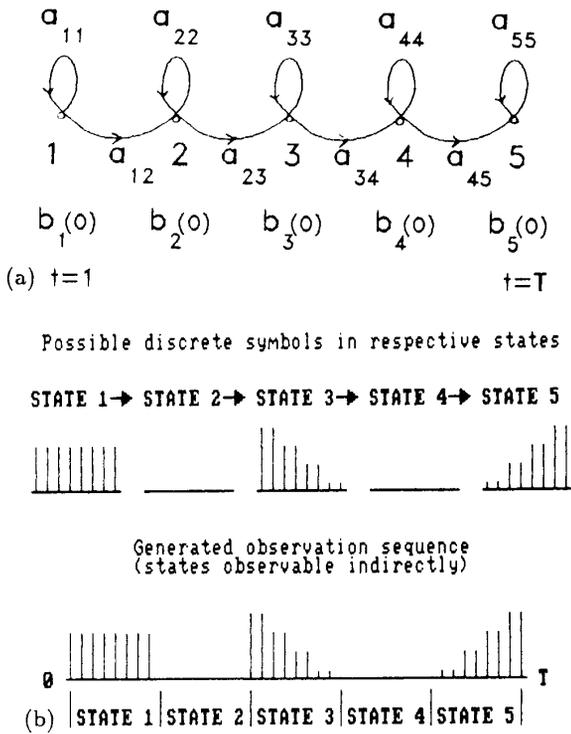


Figure 2. (a) A five state left-right HMM. (b) Schematic of generation of observation sequence given the model.

The left-right model best describes the speech and PD processes because it inherently incorporates temporal ordering, a must for modeling time-varying signals. Figure 2(a) shows a left-right model which starts from the left in state  $q_1$  at time  $t$  and sequentially proceeds to the right and stops at time  $T$  (probably in state  $q_N$ , if  $T$  is large).

### 2.2 ELEMENTS OF HMM

The main elements of an HMM are [6, 7] the following.  $N$  is the number of states in the model. Individual states are denoted as  $(q_1, q_2, \dots, q_N)$ . Usually some physical significance is attached to these states.

$M$  is the number of distinct observable symbols per state. It corresponds to the actual output of the system being modeled. Individual symbols are  $V = (V_1, V_2, \dots, V_M)$ .

The initial state distribution is given by the matrix  $\Pi = \pi_i$ ,  $\pi_i = Pr(q_i \text{ at } t = 1)$ ,  $i = 1 \dots N$ .  $\pi_i$  is the probability of starting the process in state  $q_i$  at  $t = 1$ .

Obviously for the left-right model,  $\pi_1 = 1.0$ ,  $\pi_i = 0.0$ ,  $2 \leq i \leq N$ .

The state transition probability distribution is the matrix  $A = a_{ij}$ ,  $a_{ij} = Pr(q_j \text{ at } t + 1 | q_i \text{ at } t)$ ,  $1 \leq i, j \leq N$ .  $a_{ij}$  is the probability of making a transition from state  $q_i$  at time  $t$  to state  $q_j$  at time  $t + 1$ , and this depends on the state  $q_i$  (Markovian property). The transition may be such that the process remains in state  $q_i$  at  $t + 1$  or moves to state  $q_j$ .

For the left-right model of Figure 2(a),  $a_{ij} = 0.0$ ,  $j < i$  and for  $j > i + 1$ . Choice of values of the other matrix elements will be discussed in Section 3.3. The symbol probability distribution is the matrix  $B = b_j(k)$ ,  $b_j(k) = Pr(V_k \text{ at } t | q_j \text{ at } t)$ ,  $j = 1$  to  $N$ ;  $k = 1$  to  $M$ .  $b_j(k)$  is the probability of emitting the symbol  $V_k$  from state  $q_j$  at time  $t$ , from among the discrete set of symbol observations.

In short, the model is denoted by  $\theta = (A, B, \Pi)$ . The model starts in a particular state  $q_i$  at  $t = 1$  depending on the initial state distribution, and produces an output symbol  $O_t = V_k$ , according to  $b_i(k)$ . It moves to state  $q_j$  or remains in state  $q_i$  according to  $a_{ij}$ . This process of outputting a symbol and transition to the next state repeats until a goal (such as completing a fixed number of time steps) is reached. In the left-right model, it starts in state  $q_1$  at  $t = 1$  and stops after  $T$  steps i.e. at  $t = T$  (probably in state  $q_N$ , if  $T$  is large). Thus, one can even generate an observation sequence  $O = O_1, O_2, O_3, \dots, O_T$ , given the model, as depicted pictorially in Figure 2(b).

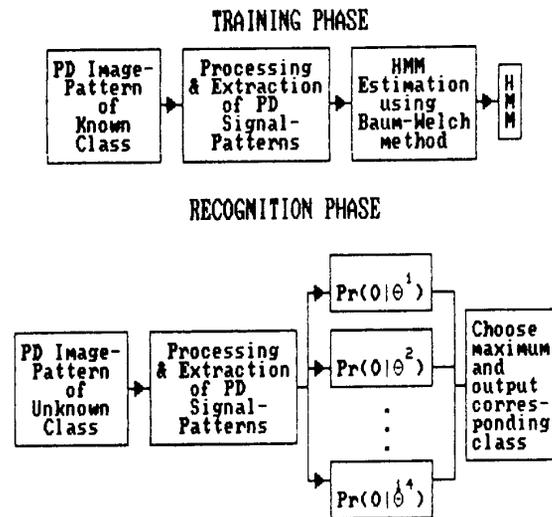


Figure 3. Block diagrams of training and recognition phases.

### 2.3 ISSUES IN HMM IMPLEMENTATION

The actual application of HMM essentially involves first a training phase during which the model is built. This attempts to optimize the model parameters  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{\Pi}$ , so as to best describe the observation sequences. Then comes the recognition phase which involves classification of new sequences (also referred to as the 'evaluation' or 'scoring' problem by speech analysts), where computation of  $Pr(O|\theta)$ ; the probability that the observed sequence  $O = O_1, O_2, O_3, \dots, O_T$ , was produced by the model  $\theta$ , is performed. Figure 3 gives block diagrams of these two phases. A computationally efficient solution to the recognition problem has been reported in [6, 7], and is named 'forward-backward' procedure. With the 'Forward' variable  $\alpha_t(i)$  and 'Backward' variable  $\beta_t(i)$  defined as:

$$\begin{aligned}\alpha_t(i) &= Pr(O_1, O_2, \dots, O_t, \quad q_i \text{ at } t|\theta) \\ \beta_t(i) &= Pr(O_{t+1}, O_{t+2}, \dots, \quad O_T|q_i \text{ at } t, \theta)\end{aligned}\quad (1)$$

the equations involving its computation are initialization Equation (2) and (4) and induction Equations (3) and (5)

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N \quad (2)$$

for  $t = 1 \dots T - 1, 1 \leq j \leq N$

$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(O_{t+1}) \quad (3)$$

$$\begin{aligned}Pr(O|\theta) &= \sum_{i=1}^N \alpha_T(i) \\ \beta_T(i) &= 1.0 \quad 1 \leq i \leq N\end{aligned}\quad (4)$$

for  $t = T - 1 \dots 1, 1 \leq i \leq N$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad (5)$$

There is as yet no analytical procedure to re-estimate the model parameters  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{\Pi}$  which maximizes  $Pr(O|\theta)$ , given any finite observation sequence as training data. Therefore an iterative method (due to Baum and Welch [6]) is employed to choose  $\theta$  such that  $Pr(O|\theta)$  is only locally maximized. Baum and his colleagues [4] showed that the re-estimated model  $\bar{\theta} = (\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{\Pi}})$  is better than or equal to the previous model estimate  $\theta$ , in the sense that  $Pr(O|\bar{\theta}) \geq Pr(O|\theta)$ . Thus, if iteratively  $\bar{\theta}$  can be used in place of  $\theta$  to repeat the re-estimation process, the probability of  $O$  being observed from the model improves until some limiting point is reached. The final result or estimate is called maximum likelihood estimate of HMM.

The re-estimation formulae are evaluated as follows (details can be found in [6-8]). The value  $\bar{a}_{ij}$  is the ratio of the expected number of transitions from  $q_i$  to  $q_j$  to the expected number of transitions from  $q_i$ .

The value  $\bar{b}_j(k)$  is the ratio of the expected number of transitions from  $q_j$  and observing symbol  $V_k$  to the expected number of times in state  $q_j$ , and  $\bar{\pi}_i$  is not estimated in the left-right model since any HMM parameter set to zero initially will remain zero throughout the re-estimation procedure.

A major problem with left-right models is that a single observation sequence cannot be used to train the model using the above principle. This is because the transient nature of the states within the model only allow a small number of observations for any state. Thus reliable estimates are possible only with multiple observation sequences. It is assumed that each observation sequence is independent of the other and individual frequencies of occurrence for each of the sequences are added [7, 8].

Computation of  $\alpha_t(i)$  and  $\beta_t(i)$  as per Equations (2) to (5) involve probabilities ( $< 1.0$ ) and quickly lead to mathematical underflow on any machine. A reasonable solution is to perform scaling of  $\alpha_t(i)$  and  $\beta_t(i)$  when necessary, or at every time step  $t$ . These scaled values are used in the re-estimation formulae and yield unscaled estimates [7, 8].

## 3. CHOICE OF HMM PARAMETERS FOR PD PATTERNS

TRAINING an HMM involves firstly the choice of a number of quantities  $N$ ,  $M$ , initializations of  $\pi$ ,  $a_{ij}$  and  $b_j(k)$  etc.. It has been observed by many speech researchers that a proper choice of initial estimates is very essential and leads to better results [7-9]. But, unfortunately, no simple rules exist to aid such a choice and moreover the choice depends on the type of signal being modeled. Therefore a justification for any chosen model parameter becomes very essential.

### 3.1 INPUT SIGNAL PATTERNS

The image pattern displayed on conventional PD detector screen can be transferred to a computer with the help of a CCD camera, from which the PD signal pattern is extracted. The details of extraction from the displayed image pattern are given in the Appendix. If the digitized

PD signal is available (as is possible with modern instrumentation), the signal pattern can be obtained directly (Figure 10). Here, these signal patterns are 36 points long and scaled in the range 0 to 1.0. Thirty such signal patterns form the training set. Only the positive excursions of the PD pulses in the image pattern have been considered in line with conventional PD measurements.

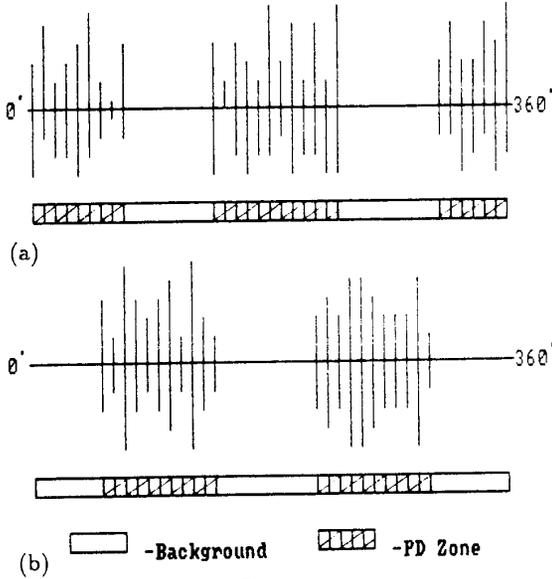


Figure 4.

Basis for choosing  $N$  by segmentation of signal record. Two possible types (a) and (b) are shown.

### 3.2 NUMBER OF STATES $N$

All the PD image patterns can, as a first approximation, be considered to consist essentially of two zones; 'PD' and 'Background', analogous to 'Signal' and 'Silence' in speech. The relative durations and positions of these zones vary depending on the type of PD. In Figure 4, the two variations shown cover all these possibilities and the PD input patterns can be accounted for by this approach. As can be seen, there are in all 5 zones and hence,  $N$  was chosen to be 5. In speech domain, values of  $N$  from 5 to 7 have been considered and found to give low word recognition error rates [7-9]. Therefore, the choice of  $N = 5$  can be considered as reasonable.

### 3.3 INITIALIZATION OF $a_{ij}$

The state transition probability distribution basically models the duration for which a process resides in a particular state ( $a_{ii}$ ), before making a transition to the next

state ( $a_{ij}$ ). In practice, uniform estimates of  $a_{ij}$  have been considered quite adequate [7, 8]. This leads to the selection of  $a_{ii}$  and  $a_{ii+1} = 0.5$ , for  $1 \leq i \leq N - 1$  and  $a_{NN} = 1.0$ , for all the models chosen.

### 3.4 INITIALIZATION OF $b_j(k)$

Initial estimates of  $b_j(k)$  have a very high influence on the final estimates. Various methods such as hand tuning of the initial estimates, maximum likelihood segmentation of observations with averaging, segmental  $K$ -means segmentation with clustering, etc. are used to obtain better initial estimates in speech recognition [7, 9]. All these involve much preprocessing. Therefore, all the symbols in the states are assumed to be 'equally probable' and  $b_j(k)$  is initialized to  $1/M$  for all  $j, k$ , for simplicity.

### 3.5 OBSERVATION SEQUENCE AND SYMBOLS

The PD signals are 'pulsive' in nature and do not contain any predominant sinusoidal component, as is contained in speech. In the present case, only relative time information is available. Therefore, linear predictive coding analysis (used in speech domain to obtain observation vectors), cannot be adopted here. So, as a first step, only quantized amplitudes were considered. The signal pattern (scaled from 0 to 1.0) containing the averaged amplitudes are quantized in steps of 0.05 to yield the observation sequence. Therefore, the corresponding symbols are  $V = 0.0, 0.05, 0.10, \dots, 1.0$  and  $M = 21$ .

### 3.6 CONSTRAINTS ON $b_j(k)$ DURING TRAINING

Serious recognition errors will result [7-9] if any element of  $b_j(k)$  is allowed to attain a zero value during the training phase. This is because, the recognition phase involves the computation of  $Pr(O|\theta)$  from  $\alpha_t(i)$  (Equation (2) and (3)) for each of the models and assigning the input to the most probable model (i.e. the model with the maximum  $Pr(O|\theta)$ ). This problem is avoided by clamping the defaulting elements to a small value,  $\epsilon$ . Thereafter,  $b_j(k)$  is rescaled such that  $\sum_k b_j(k) = 1.0$ . This means that the probability of none of the symbols in any state is allowed a zero value, however remote its chance of occurrence may be. Such a constraint check is done each time the model parameters are updated. Values of  $\epsilon$  between  $10^{-3}$  to  $10^{-7}$ , used in speech have resulted in low word recognition errors. Hence, a value  $10^{-4}$  was used throughout this work unless otherwise stated.

## 4. RESULTS AND DISCUSSIONS

### 4.1 STUDY OF SIGNAL PATTERNS

A study of the patterns in [1,2] resulted in the identification of 21 distinct patterns; of them, 7 are due to interferences. As the latter patterns are not related to the insulation system under study, present with 'zero' applied voltage, and often found to rotate due to relative asynchronism with the time base, no definite phase position can be assigned to these patterns. Hence, pattern stationarity or otherwise cannot be conveyed simply through a single image and they have to be considered separately. Image patterns due to interferences have a finer structure when compared to the rest. Hence, caution is to be exercised while forming the signal patterns so as to retain as many discriminating features as possible. This is achieved with longer observation sequences, and hence results in increased training time. These results will be reported later.

A study of the phase angle characteristics of the first set of patterns (14, herein after called classes) led to the following subgroupings:

Group A (3) - 2 pulse groups in first and third quadrants.

Group B (6) - 1 or 2 pulse groups near voltage peaks.

Group C (5) - 2 pulse groups near voltage zero.

After preliminary processing of each of the image patterns 6 signal patterns were extracted by choosing its starting positions randomly, within a phase interval of  $\sim 20^\circ$ . This simulates the random variation of phase angle at which the pulses appear on the image pattern. 30 signal patterns are obtained by repeating the process 5 times for a good statistical representation in the training set. Figure 5 shows this process.

### 4.2 TRAINING OF HMM

HMM are built for each of the 14 classes by the training procedure outlined in the flow chart shown in Figure 6. This method is a supervised learning approach and, as only a local minimum is reachable, no absolute error value can be fixed a priori for indicating convergence. So, the training is stopped when the change in probability of observing a sequence conditioned on the model,  $Pr(O|\theta)$ ,

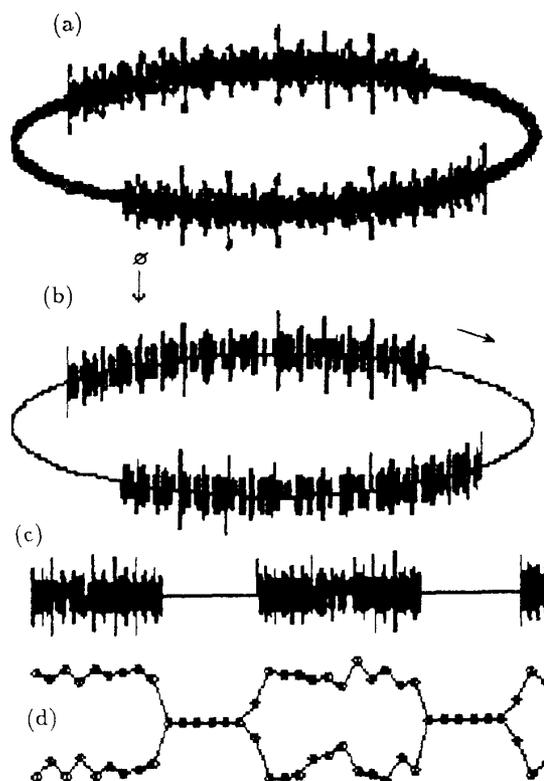


Figure 5.

Procedure for extraction of signal patterns. (a) Processed camera output. (b) Extracted image pattern. (c) Signal pattern before averaging. (d) Scaled signal pattern 36 points long.

is reasonably small or when the number of iterations exceeds a preset limit [8]. To prevent underflow in computation of  $Pr(O|\theta)$ , the mean logarithmic probability  $MP$  is computed as

$$MP_j = \sum_{i=1}^p \frac{\log[Pr(O^i|\theta^j)]}{p} \quad (6)$$

where  $O^i$  is the  $i$ th observation sequence from training set,  $\theta^j$  the updated HMM after  $j$ th iteration, and  $p$  is the number of training set patterns.

In this work, the latter convergence condition was chosen for simplicity. Results showed that  $\sim 25$  to 50 iterations are sufficient to reach these local minima for all the models.

### 4.3 HMM PERFORMANCE

HMM corresponding to each of the 14 classes were trained independently assuming them to be distinct. As

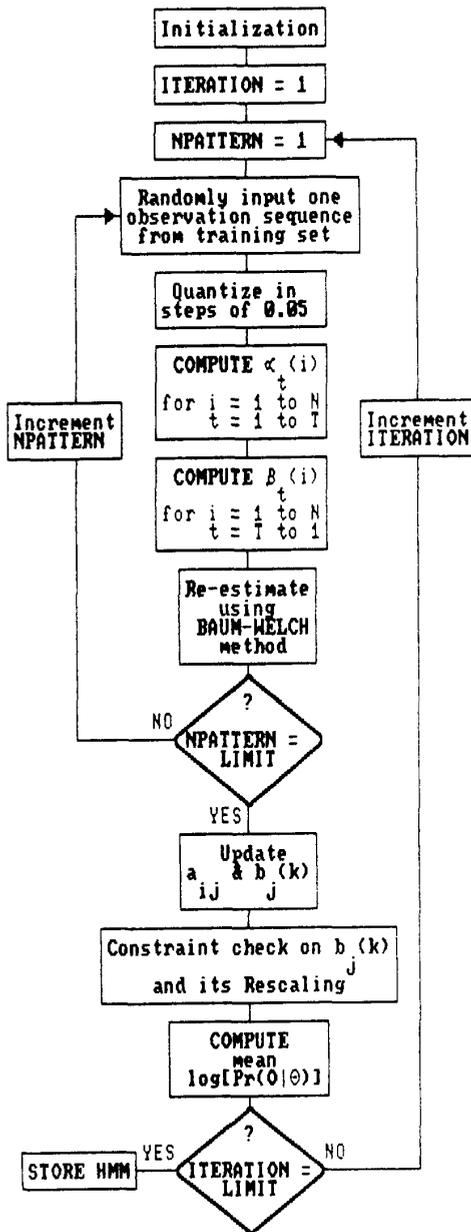


Figure 6.

Flow chart for the training procedure.

an HMM is required to recognize any given input pattern correctly, tests were run to evaluate this capability by using inputs drawn from the training set itself.

Initially, models were updated after 10 randomly chosen observation sequences were drawn from the training

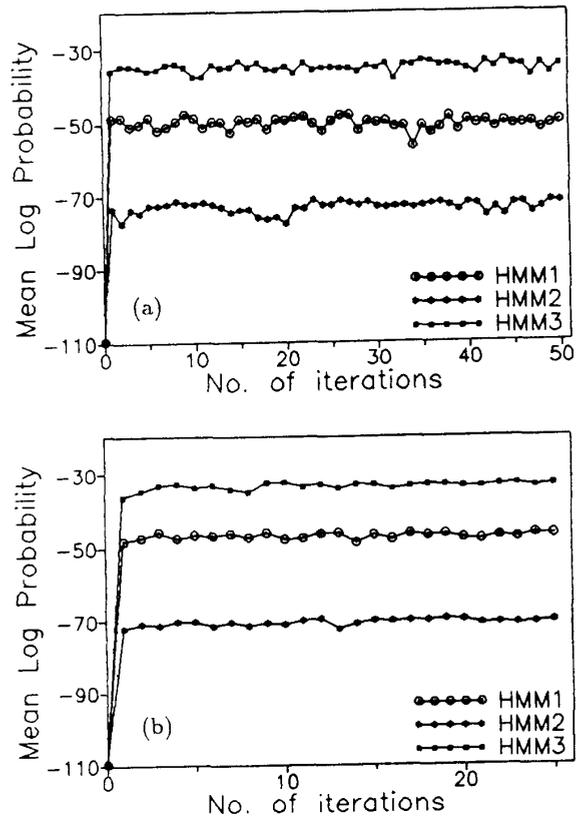


Figure 7.

Sample HMM learning curves for Cases 1 and 2. (a) Case 1, (b) Case 2.

Table 1.

Value of  $\epsilon$  and number of inputs  $N$  before updating the HMM.

Case	$\epsilon$	Number of inputs $N$
1	$10^{-4}$	10
2	$10^{-4}$	20
3	$10^{-3}$	20

set. Initializations were done as already explained. The resulting learning curves were observed to be oscillatory for most of the models. Figure 7(a) gives sample learning curves. Tests for recognition capability were run using inputs drawn from the training set and results indicated need for further improvement. It was conjectured that these oscillations are due to premature model updating. Hence, the number of inputs fed in before updating was increased from 10 to 20 (the other parameters being unchanged) and a new set of models were trained. These resulted in smoother learning curves as shown in Figure

7(b) and better recognition ability. To evaluate the effect of  $\epsilon$  on recognition errors, a new set of models were also trained using  $\epsilon = 10^{-3}$ . Table 1 gives all the parameters used during training. Table 2 shows these results. It can be seen that only marginal improvement was obtained for Case 3 when compared to Case 2.

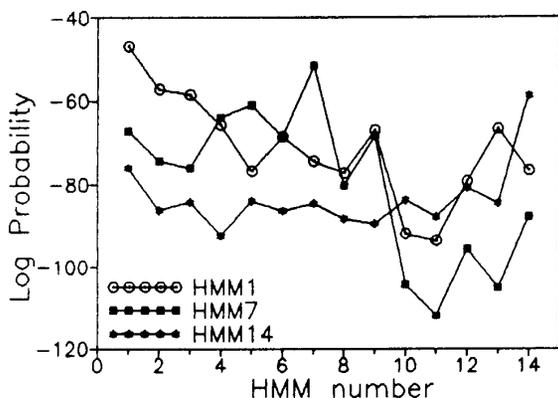


Figure 8. Discrimination ability of HMM (Case 3).

Table 2. Number of correct classifications when tested with training set itself (30 patterns each).

Group	HMM (No.)	Case 1	Case 2	Case 3
A	1	22	28	29
	2	28	29	28
	3	26	26	27
B	4	20	24	29
	5	27	30	27
	6	27	30	28
	7	29	25	30
	8	17	16	21
	9	25	30	30
C	10	22	29	26
	11	21	30	30
	12	18	30	30
	13	24	23	24
	14	26	29	30

Generalization capability of HMM also was tested by using additionally extracted signal patterns. These results are summarized in Table 3. Recognition rates obtained for these new inputs were equally good for both  $\epsilon$  values of  $10^{-4}$  and  $10^{-3}$ . The ability of HMM to classify correctly any single input if it belongs to its class and discriminate it from others is shown in Figure 8.

#### 4.4 DISSIMILARITY MEASURE FOR HMM

All along, the models have been trained on the assumption that the individual classes they represent are distinct. But it may so happen that two classes assumed to be distinct may in fact be very similar in a statistical sense even though the estimated model parameters may appear different. Such circumstances lead to random misclassification within the two classes (assumed distinct) during the recognition phase. This necessitates a dissimilarity check to be performed on the trained models.

In speech domain, such a model distance or statistical equivalence check is generally done [7, 10]. Following this approach, a dissimilarity measure in terms of model distance has been calculated. The model distance indicates how well HMM  $\theta_1$  matches observations generated by HMM  $\theta_2$ , relative to how well HMM  $\theta_2$  matches observations generated by itself [7, 10]. The model distance between HMM  $\theta_1$  and HMM  $\theta_2$  is defined as

$$D(\theta_1, \theta_2) = \frac{\log Pr(O^{(2)}|\theta_2) - \log Pr(O^{(2)}|\theta_1)}{T} \quad (7)$$

where  $O^{(2)} = O_1, O_2, O_3, \dots, O_T$  is a  $T$  length sequence generated by model  $\theta_2$ .

This distance measure being asymmetrical, a symmetrized version is defined as:

$$D_s(\theta_1, \theta_2) = \frac{1}{2}[D(\theta_1, \theta_2) + D(\theta_2, \theta_1)] \quad (8)$$

The larger the distance, the greater the dissimilarity. The distances aid in locating those models which are statistically similar, but very different in appearance.

These distances were calculated for models within a group. Separate sequences were generated to compute the distance each time. The model distances for HMM belonging to Group C are given in Table 4. below. These were obtained as averages of 100 such distances.

### 5. EXPERIMENTS WITH ACTUAL PD PATTERNS

The HMM which were trained using synthetic patterns had performed well in generalization tests. Even then, their performance when fed with actual PD patterns had to be ascertained because, in practice, noise and interference levels are usually higher. So, some five typical

Table 3.  
Recognition results with new inputs (Check of generalization ability).

Group	HMM No.	No. Inputs	% correct	
			Case 2	Case 3
A	1	97	82	78
	2	73	93	92
	3	75	85	84
B	4	77	78	91
	5	75	88	83
	6	70	90	94
	7	70	76	100
	8	69	81	72
	9	71	92	97
C	10	76	91	87
	11	80	93	91
	12	71	100	98
	13	55	82	91
	14	60	98	98

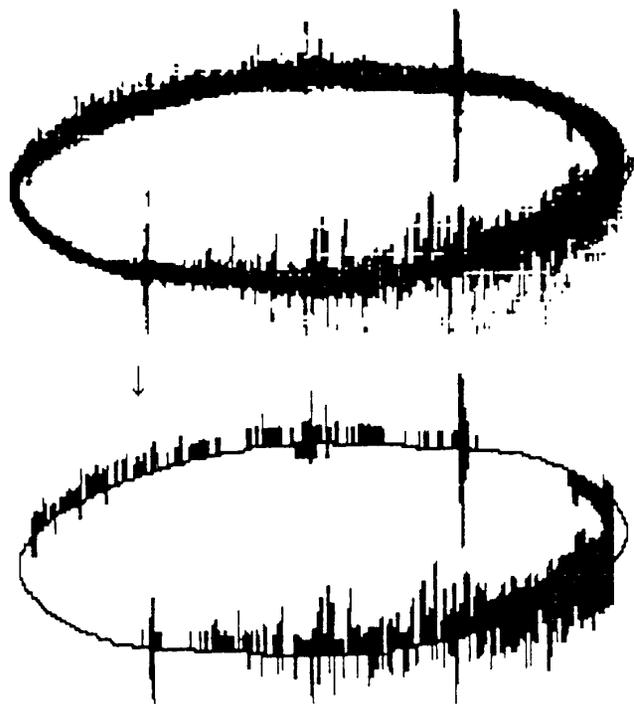


Figure 9.

Sample of processed camera output and extracted pattern for actual PD due to surface discharge on a cable. Two large and symmetrical pulsive interferences present have also been captured.

(both complex and simple patterns) PD from among the

Table 4.  
Model distances for HMM classes 10...14

	10	11	12	13	14
10	0.0	0.456	0.721	0.789	0.433
11		0.000	0.930	0.860	0.813
12			0.000	0.922	0.461
13				0.000	0.507
14					0.000

14 sets considered, were generated and displayed on the PD detector screen using an HV test system. An overall recognition rate of  $\sim 84\%$  was obtained. Further details can be found in [13]. An example of such a PD (due to surface discharge on a cable specimen) along with the extracted pattern is shown in Figure 9. The small reduction in overall performance of the HMM when fed with actual PD patterns may be attributable to higher levels of noise and pulsive interference present, as is evident in Figure 9. Details regarding training HMM to recognize both synthetic and actual interference patterns have also been reported in [13].

Table 5. Comparison of HMM and NN.

Criteria	HMM	NN
Type of classifier	Statistical	Non-parametric (implements input-output mapping.)
Training	Trained independently assuming each is distinct, verified later.	Exact implementation of training set ensured: 1 at output node, 0 at other output nodes.
Convergence	Only local maxima attainable. Hence convergence limit = iteration limit or when change in $Pr(O \theta)$ is small.	Global maxima reachable, almost always. So convergence limit can be fixed easily.
Programming	Relatively difficult and involves periodic scaling and constraint checks on $b_j(k)$ .	Easier to program.

## 5.1 COMPARATIVE PERFORMANCE OF NEURAL NETWORKS AND HMM

Neural networks (NN) offer another approach incorporating ability to learn from examples, thereby matching the needs of this problem domain. Its suitability has been examined by the authors in a recent paper [11]. A comparison of basic features of HMM and NN is given in Table 5. NN's inherent drawback is its inability to handle time-varying sequences such as PD, speech, etc. [12]. As HMM is known to model them better, it has an edge over NN. In this work, it was also observed that HMM demonstrated greater tolerance to realistic changes in phase and amplitudes during the generalization tests compared to NN for the same inputs. In these tests, the time-varying property of PD signals has been adequately simulated by choosing different starting positions (within 20 to 30°) for generating additional test inputs. This typically corresponds to variations in PD signals due to changes in voltage, duration of voltage application, aging etc. A more recent trend in continuous speech recognition has been to combine NN and HMM; NN being used for its discriminating and interpolative capabilities and HMM to capture the dynamics of speech and to segment continuous speech to successive words [12]. Perhaps, this approach may be useful in the PD domain as well.

## 6. CONCLUSIONS

PD and speech signals are time varying processes which are behaviorally very similar. This observation and the success of HMM in many speech recognition tasks led to studies on HMM for modeling and classification of PD signals. The PD image patterns were synthetically generated and the data for the training set was extracted from them.

Before the training phase is started, structural assumptions of the model have to be made; these are similar to those in the speech domain but with appropriate modifications to suit this problem. The training phase iteratively optimizes the models. The recognition phase relates the input to that model which yields the highest likelihood of reproducing it. The HMM are trained assuming them to be distinct which should be ascertained by computing the model distances.

Tests carried out on synthetic PD image patterns have yielded overall recognition rates of > 88% with inputs different from those used for training. This generalization can be improved further by using larger training sets, continuous densities for  $b_j(k)$ , etc. Therefore the capability

of HMM for PD recognition is established. Tests with actual PD image patterns (5 types) yielded an overall recognition rate of 84%, thus demonstrating its practical utility. With easy availability of fast and cheap A/D converter boards and a variety of computer platforms, on-line training and recognition of PD signals, using HMM, can be visualized.

## ACKNOWLEDGMENT

The first author acknowledges help offered in preparation of the manuscript by Mr. B. S. Rajanikanth and Mr. V. Nagesh.

## 7. APPENDIX

### 7.1 GENERATION AND DISPLAY OF PD IMAGE PATTERNS

A simple and effective, though synthetic, method to display any complicated PD image pattern on the oscilloscope has been developed. This avoids use of HV transformers and components and the need for specimens with specific defects for generating required patterns. The idea is to use a real-time digitizer in the reverse way i.e., the required signals are computed and transferred to its memory; these are then used to create the pattern on an oscilloscope. To generate any pattern, the  $X$  and  $Y$  axis signals corresponding to one 50 Hz cycle (20 ms) have to be computed. The  $X$  signal (sine wave) is computed and stored while the  $Y$  signal (cosine wave plus PD pulses) is computed each time depending on the type of pattern to be displayed. The two signals are transferred via a GPIB bus to the digitizer as 2048 points at 10  $\mu$ s sampling rate and 10 bit resolution.

### 7.2 EXTRACTION OF INPUT SIGNAL

A digitizing camera is used to capture the displayed image. The compressed raw video file is then unpacked and displayed on the PC screen after normalization and thresholding. A spot on the oscilloscope as seen by the eye does not correspond to a single pixel on the CCD array, but to as many as 8 to 10 pixels, depending on its intensity. Therefore, all displayed traces have to be corrected for this dispersion. First, a reference ellipse is captured and displayed on the PC screen from which a single pixel wide reference ellipse is extracted in the following manner. A boundary tracing algorithm was developed and used to trace the inner and outer contours, from which

the reference ellipse is computed and stored. Movable zero voltage markers on the elliptic trace are provided on the PD detectors for clarity in viewing patterns and hence made software selectable. The reference ellipse is then moved, if necessary, to coincide with the center of the displayed pattern. This is to correct for any misalignment due to camera shake, shift of oscilloscope trace, etc. Then the pattern is scanned vertically on either side along the reference ellipse, starting from the selected zero marker, in a clockwise direction to extract the PD pulses. The positive and negative portions are stored separately and averaged every  $10^\circ$  to obtain 36 sample points. The averaged version of the positive portion forms the training set. Figure 5 illustrates these steps.

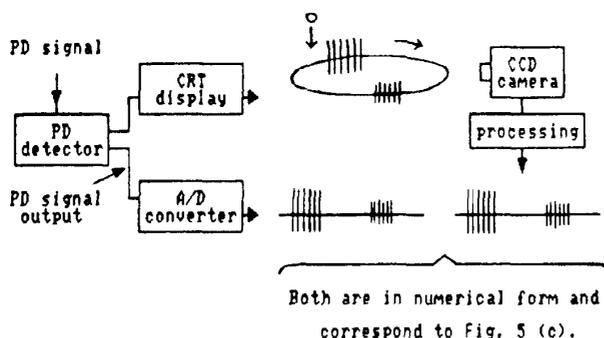


Figure 10.

Schematic illustrating the two possible approaches for obtaining signal patterns.

In this context of obtaining signal patterns, as mentioned in Section 3.1, the approach using the A/D converter can also be taken to obtain signal patterns. Figure 10 illustrates the two approaches schematically. As is evident, both of them lead to the same information.

## REFERENCES

- [1] CIGRE Working Group 21.03, "Recognition of Discharges", *Electra*, no. 11, pp. 61-98, 1969.
- [2] D. A. Natrass, "Partial Discharge Measurement and Interpretation", *IEEE Electrical Insulation Magazine*, Vol. 4, no. 3, pp. 10-23, May/June, 1988.
- [3] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech using Hidden Markov Models", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, no. 11, pp. 1870-1878, Nov. 1990.
- [4] L. E. Baum and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology", *Bulletin of the American Mathematical Society*, Vol. 73, pp. 360-363, 1967.
- [5] J. K. Baker, "The Dragon System - An Overview", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 23, no. 1, pp. 24-29, Feb. 1975.
- [6] L. R. Rabiner, and B. H. Juang, "An Introduction to Hidden Markov Models", *IEEE Acoustics, Speech, and Signal Processing Magazine*, Vol. 3, no. 1, pp. 4-16, Jan. 1986.
- [7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of the IEEE*, Vol. 77, no. 2, pp. 257-286, Feb. 1989.
- [8] S. E. Levinson, L. R. Rabiner and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Model Process to Automatic Speech Recognition", *The Bell Systems Technical Journal*, Vol. 62, no. 4, pp. 1035-1074, April 1983.
- [9] L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition", *The Bell Systems Technical Journal*, Vol. 62, no. 4, pp. 1075-1105, April 1983.
- [10] B. H. Juang, and L. R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models", *AT & T Technical Journal*, Vol. 64, no. 2, pp. 391-408, Feb. 1985.
- [11] L. Satish and B. I. Gururaj, "Partial Discharge Pattern Classification using Multilayer Neural Networks", *Proc. of IEE-A*, acceptance communicated.
- [12] H. Bourlard and N. Morgan, "Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition", in *Neural Networks: Advances and Applications*, E. Gelenbe (Editor), Elsevier Science Publishers, pp. 215-239, 1991.
- [13] L. Satish, *Study of some Artificial Intelligence Paradigms for Partial Discharge Pattern Recognition*, Ph.D Thesis, Indian Institute of Science, Bangalore, Sept. 1992.

Manuscript was received on 30 March 1992, in revised form 23 Dec. 1992.