Abstract - Hidden Markov Models (HMMs) have been successfully applied to problems from different areas over the last years. Independent of the application it is well known that an important step considering the use of HMMs is the initialization of the parameters of the model. This initialization should take into account the knowledge about the addressed problem and also optimization techniques to estimate the best initial parameters given a cost function. Currently, there exist distinct techniques to initialize and evaluate HMMs, however, there is no a common consensus concerning the choice of these tools. In this paper we illustrate, through examples, the effects of an inadequate initialization of HMM parameters, and also discuss important issues on the selection or estimate of these parameters. The presented results and examples highlight the relevance of studying the initial model parameters prior to signal modeling.

Keywords – Hidden markov models, mathematical models, statistical models.

I. INTRODUCTION

Mathematical models are very useful tools for processing and understanding random signals [1, 2]. Hidden Markov Models (HMMs) are statistical models, which have been successfully applied to areas, such as speech analysis and biomedical signal processing [1]. An HMM is considered to be an extension of Markov models from which an observed state is the result of a stochastic process in one of several unobservable states [3].

Although HMMs have been applied in practice, a frequent problem that researchers have to face is that of selecting or estimating initial values for the parameters of the model [4]. It is well known that improperly initialized models lead to inaccurate estimates of the HMM parameters and in a convergence to a local optimum, which may be significantly worse than the global optimum [5].

A number of strategies have been developed to initialize the main parameters of HMMs, such as the application of clustering algorithms (e.g., k-means clustering) [5-11], Gaussian Mixture Models [12-16], and the use of random values [4, 5, 17-19]. However, there is not a common consensus concerning the use of any criterion to select the technique to be used. So, how is it possible to know if an initialization is really good? How to know which are the best parameters to be used in the training and in the test of an HMM?

In this work we illustrate, through examples, the effects of an inadequate initialization of HMM parameters, and also discuss important issues on the selection or estimate of these parameters. The presented results and examples highlight the relevance of studying the initial model parameters prior to signal modeling. These results may be relevant to those investigations that seek to apply HMMs in a practical context.

II. DEFINITION OF HMMS

An HMM is a statistical method that employs probability measures to model sequential data represented by a sequence of observations [20]. Rabiner [17] also defines an HMM as "a doubly embedded stochastic process with an underlying process that is not observable (i.e., it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations".

A standard HMM has the following parameters [4]:

- \( N \) = number of states;
- \( S = (S_1, S_2, ..., S_N) \) – set of states;
- \( M \) = number of symbols;
- \( V = \{v_1, v_2, ..., v_M\} \) - set of observations, which may be scalars or vectors;
- \( \pi = [\pi_1, \pi_2, \pi_3, ..., \pi_N] \) – set of initial probabilities of starting in state \( S_i \) at \( t = 1 \);
- \( O = \{O_1, O_2, ..., O_T\} \); \( O_i \in V \); \( \forall i \);
- \( T \) = length of the observation sequence (total number of clock times);
- \( t = \) clock time; \( t \in (1, 2, 3, ..., T) \);
- \( A = \{a_{ij}\} \) transition probabilities \( a_{ij} \) for moving from state \( S_i \) to state \( S_j \);
- \( B = \{b_j(k)\} \) observation probabilities \( b_j(k) \) of observing symbol \( V_k \) while being in state \( S_j \).

HMM parameters are often referred to as \( \lambda = (\pi, A, B) \). With this set of parameters, it is possible to generate a series of observations, and given the above parameters one can compute the likelihood, \( L(O|\lambda) \), of such series by taking the product of the individual probabilities of moving from one state to the next and producing the observations \( O_t, t = 1, ..., T \in V \) in those states [3]. However, it is a challenge task to initialize HMM parameters, and a proper initialization is essential [4]. The best initialization must be the one that have the highest likelihood [21].
III. STRATEGY FOR DATA ANALYSIS

As mentioned above, one of the aims of this study is to highlight one of the central issues when applying HMMs in practice, which is the estimate of the initial parameters of the model. In order to illustrate this problem, we present, in the following sections, examples that show how inappropriate initial conditions of HMMs may influence on the optimization of the parameters of the model. Although this is a well-known problem in the theory of HMMs, after a literature survey, regarding the study of 108 papers applying HMMs to solve practical problems, we concluded that most of them did not take into account a detailed study of the model initialization before its use. This was the main motivation of our work.

In our study, we first generated synthetic data (a Hidden Markovian Process), and then modeled them by means of HMMs taking into consideration distinct initial conditions for the model, which were strategically defined as a function of parameters, \( \lambda_{\text{actual}} \) used to generate artificial data.

IV. GENERATION OF SYNTHETIC DATA

In order to generate the observations \( (O) \), the HMM Matlab Toolbox was used as:

\[
[O, \text{states}] = \text{hmmgenerate}(T, A_{\text{actual}}, B_{\text{actual}});
\]

where \( O \) is a vector of observations generated, \( \text{states} \) is a vector that contains the state of each observation, \( T \) is the total number of observations \( (O_1, O_2, ..., O_T) \), \( A_{\text{actual}} \) is the transition probability matrix and \( B_{\text{actual}} \) is the observation probability matrix. Note that when applying HMMs one is interested in obtaining the actual parameters \( (A_{\text{actual}}, B_{\text{actual}}) \) that define the rules for data generation. These parameters could have a physical meaning, for instance, they could represent an underlying process responsible for the generation of biological signals.

V. OPTIMIZATION OF THE MODEL PARAMETERS

In order to optimize the parameters of the model, the Matlab function \text{hmmtrain} was used as follows:

\[
[\text{new}_A, \text{new}_B] = \text{hmmtrain}(O, \text{Trial}_A, \text{Trial}_B);
\]

where \( O \) is the observation vector obtained from the function \text{hmmgenerate} (as explained above), \( \text{Trial}_A \) is an initial guess for the matrix \( A_{\text{actual}} \), \( \text{Trial}_B \) is an initial guess for the matrix \( B_{\text{actual}} \), and \( \text{new}_A \) and \( \text{new}_B \) are estimated (optimized) matrices resulting from the Matlab function \text{hmmtrain}, which optimizes the model parameters based on the Baum-Welch algorithm [4].

Note that the main aim of our study is to illustrate the sensitivity of this optimization algorithm with regard to the choice of the initial parameters of the model \( (\text{Trial}_A, \text{Trial}_B) \).

Simulations considering three distinct conditions were made:

1. Initialize \( \text{Trial}_A \) as a function of \( A_{\text{actual}} \) and maintain \( \text{Trial}_B \) equal to \( B_{\text{actual}} \);
2. Initialize \( \text{Trial}_B \) as a function of \( B_{\text{actual}} \) and maintain \( \text{Trial}_A \) equal to \( A_{\text{actual}} \);
3. Initialize \( \text{Trial}_A \) and \( \text{Trial}_B \) as a function of \( A_{\text{actual}} \) and \( B_{\text{actual}} \), respectively.

VI. DEFINITION OF A FUNCTION TO GENERATE INITIAL PARAMETERS OF HMMs

In this study, a linear projection was employed to define the initial parameters, \( \text{Trial}_A \) and \( \text{Trial}_B \), of the model. The idea was to estimate these parameters by varying the distance from their actual values \( (A_{\text{actual}}, B_{\text{actual}}) \), which were employed for data generation.

In order to generate the initial parameters and evaluate them according to the distance from their actual values, it was considered the Euclidian Distance in \( R^2 \) for visualization purposes. Figure 1 illustrates a set of possible parameters positioned around a central point defined at the coordinate \((500, 500)\). These points are separated into bins (as indicated by the different colours) and each bin is composed of points that have a Euclidian distance in the interval \([d_{\text{min}}, d_{\text{max}}]\) from the central point, where \( b \) is one of the eight available bins (see distinct colors in Figure 1), and \( d_{\text{min}} \) and \( d_{\text{max}} \) are the minimum and maximum Euclidian distances that define the bin (see Figure 2).

Our aim is to study the behavior of the likelihood function \( (L) \) and the mean squared error (a measure of the difference between estimated and actual parameters) as a function of the central Euclidian Distance, \((d_{\text{max}} - d_{\text{min}})/2\), in a bin.

Figure 1: Points in \( R^2 \) used to generate the estimated parameters to test the HMM model. These points are separated into bins according with their Euclidian Distance from the central point \((x=500, y=500)\).

The representations in \( R^2 \) in Figure 1 are only for visualization purposes. In practice both the \( A \) and \( B \) matrices may have a dimension larger than 2, for instance, a matrix \( A \) with 3 states would need a vector in \( R^2 \) in order to represent it. Note that, although the matrices \( A \) and \( B \) of an HMM are bi-dimensional arrays these may be seen, for practical purposes, as one-dimensional arrays. This observation is
important because it is always possible to obtain a linear projection from a low-dimensional space vector onto a high dimensional space guaranteeing the same Euclidian distance in both spaces. This transformation will allow for the generation of A and B matrices that follows the same Euclidian distances like those for the points in Figures 1 and 2.

This linear projection can be arbitrarily defined as in Equation 1 [22], where \( p(r^* ) \) is a vector in \( R^n \) (i.e., \( p(r^* ) = [w_1, w_2, ..., w_n] \)), \( P(r^* ) \) is a vector in \( R^2 \) (used only for visualization purposes) and \( O_r \) is an orthogonal matrix.

\[
P(r^n) = O_r \ast P(r^2)
\]  

(1)

Note that the Euclidian Distance from points estimated by (1) will be preserved in the higher dimensional space (\( R^n \)), where \( n > 2 \), and vice-versa. In (2) we have a generic example of a representation of an A matrix using the multi-dimensional vector \( P(r^n) \). The same reasoning is true for the representation of the matrix B.

\[
A = \begin{bmatrix}
    w_1 & w_2 & w_N \\
    w_{N+1} & w_{N+2} & ... \\
    ... & ... & w_n
\end{bmatrix}
\]  

(2)

In our simulations, the vector of initial probabilities (\( \Pi \)) was defined as in (4), where \( N \) is the number of states.

\[
\Pi = \begin{bmatrix}
    \frac{1}{N} & \frac{1}{N} & \frac{1}{N}
\end{bmatrix}
\]  

(3)

VII. RESULTS

As described in the previous section, the synthetic Markovian process was generated through parameters (\( A_{\text{actual}} \), \( B_{\text{actual}} \)), obtained by means of the projection of the central point (500,500) in \( R^2 \) onto a higher dimensional space. In order to assess the sensitivity of the HMM with regard to the initialization parameters, different initial conditions, which are represented by the matrices \( \text{Trial}_A \), \( \text{Trial}_B \), were used. Note that these matrices are variations of the matrices \( A_{\text{actual}} \) and \( B_{\text{actual}} \) that consider distinct Euclidian distances from the central point (the actual point used for generation of the Markovian process) to other points. This is illustrated in Figures 1 and 2.

After optimizing the parameters of an HMM two new matrices are estimated \( \text{New}_A \), \( \text{New}_B \). Ideally these matrices should be equal to \( A_{\text{actual}} \) and \( B_{\text{actual}} \), however, due to limitations of the optimization algorithm, what one can obtain in practice is only an estimate, which is dependent of the initialization parameters. Therefore, it is possible to calculate the Mean Square Error between the estimated and actual parameters of an HMM.

In addition, it is possible to investigate the behavior of the probability of a model given a data set, that is, it is possible to investigate the log-likelihood of the data given the model. These two measures were employed to investigate the sensitivity of an HMM with regard to its initialization parameter.

For the first simulation, which considers \( \text{Trial}_A \) as a function of \( A_{\text{actual}} \) and \( \text{Trial}_B \) equals to \( B_{\text{actual}} \), we obtained the result depicted in Figure 2. Each mark in this graph is a mean error (computed as the average of the Mean Square Error measure) together with its variance (indicated by the vertical bar over the mark). From this result it is possible to conclude that as the mean Euclidian Distance, indicated by the bins in the x-axis increases the overall error (in the y-axis) also increases.

For the second simulation, which considers \( \text{Trial}_B \) as a function of \( B_{\text{actual}} \) and \( \text{Trial}_A \) equals to \( A_{\text{actual}} \), we obtained the result depicted in Figure 4. From this result it is possible to conclude that as the mean Euclidian Distance between \( \text{Trial}_B \) and \( A_{\text{actual}} \) (in the x-axis) increases the overall error (in the y-axis) also increases.

Figure 2: The average error between the matrices \( A_{\text{actual}} \) and \( \text{New}_A \) according with the distance between the matrices \( A_{\text{actual}} \) and \( \text{Trial}_A \).

In this simulation, it was also possible to observe that as the distance between \( \text{Trial}_A \) and \( A_{\text{actual}} \) (in the x-axis) increases, the likelihood (in the y-axis) decreases (Figure 3).

Figure 3: The average likelihood according with the distance between the matrices \( A_{\text{actual}} \) and \( \text{Trial}_A \).
Figure 4: The average error between the matrices $B_{\text{actual}}$ and $\text{New}_B$ according with the distance between the matrices $B_{\text{actual}}$ and $\text{Trial}_B$.

It was also possible to observe that as the distance between $\text{Trial}_B$ and $B_{\text{actual}}$ (in the x-axis) increases, the likelihood (in the y-axis) decreases (Figure 5).

Figure 5: The average likelihood according with the distance between the matrices $B_{\text{actual}}$ and $\text{Trial}_B$.

For the third simulation, which considers $\text{Trial}_A$ as a function of $A_{\text{actual}}$ and $\text{Trial}_B$ as a function of $B_{\text{actual}}$, we obtained the result depicted in Figure 6. From this result it is possible to conclude that as the mean Euclidian Distance between $\text{Trial}_A$ and $A_{\text{actual}}$ (in the x-axis) increases, the overall error (in the y-axis) also increases.

Figure 6: The average error between the matrices $A_{\text{actual}}$ and $\text{New}_A$ according with the distance between the matrices $A_{\text{actual}}$ and $\text{Trial}_A$.

It was also possible to conclude that as the mean Euclidian Distance between $\text{Trial}_A$ and $A_{\text{actual}}$ (in the x-axis) increases, the overall error (in the y-axis) also increases (Figure 7).

Figure 7: The average error between the matrices $B_{\text{actual}}$ and $\text{New}_B$ according with the distance between the matrices $B_{\text{actual}}$ and $\text{Trial}_B$.

Finally, it was also possible to observe that as the mean Euclidian Distance between $\text{Trial}_B$ and $B_{\text{actual}}$ (in the x-axis) increases, the overall error (in the y-axis) also increases.

Figure 8: The average likelihood according with the distance between the matrices $A_{\text{actual}}$, $\text{Trial}_A$, and $B_{\text{actual}}$, $\text{Trial}_B$.

VIII. DISCUSSION AND CONCLUSION

The results obtained from the simulations above show that the optimization algorithm employed in the study (Baum-Welch) is sensitive to the initial conditions of the HMM. Although this may be a well-known fact in the theory of HMMs a number of practical studies do not consider this issue. As shown by our results a poor initialization of the HMM parameters may lead to unexpected results which may be far from the actual ones.

In general we noted that there exists an increase in the Mean Square Error as the Euclidian distance between actual and initial parameters increases. We also noted that there is a decrease in the log-likelihood as this distance increases.

From our results, it is possible to conclude that independent of the technique employed to initialize the parameters of an HMM, this technique should provide initial parameters that are somehow close to the true solution of the
problem. However, in practice this true solution is not often known, what makes the knowledge of the practical problem a bonus that should be used together with other initialization methods.

REFERENCES


