Processing of magnetic diagnostics data using Hidden Markov Models

A.A. Lukianitsa[1], F.S. Zaitsev, S.V. Nosov

Moscow State University, Faculty of Computational Mathematics and Cybernetics, RF
[1]*e-mail: luk@ic.msu.su*

1. Introduction. Development of mathematical methods and convenient software tools for processing diagnostics data is an important direction of research in controlled fusion. Such methods and tools help the automation of huge data analyses and the prompt extraction of valuable information about plasma behavior. One approach to the solution of many complicated problems in plasma physics and controlled fusion involves modern adaptive data mining tools. The efficiency and productivity of this kind of tools are based on their ability to extract very valuable information from seemingly chaotic data without deep knowledge of underlying physical laws, which in fact can be too complicated or still unformulated. Indeed, the extracted information can help the formulation of these laws and the making of correct decisions.

Here we consider the problem of analyses of magnetic diagnostics data. Many gigabytes of such data are collected for plasma pulses on different tokamaks. However, at present magnetic data is not used as fully as it might be. In this contribution, a technique new to fusion is presented for diagnostics data compression, the extraction and analyses of additional information from measurements together with appropriate software. As a result the researcher can avoid many routine operations and concentrate on physics and navigation between pulses. The technique is based on hidden Markov models (HMM) [1,2]. HMM are well suited to magnetic diagnostics analyses, since magnetic frequency characteristics have much in common with speech ones, for which HMM are successfully used in speech recognition problems (http://leader.cs.msu.su/~luk/ContinuousSpeech.html).

HMM allow indexing of time dependent data of different length without substantial loss of the main information, i.e. compression, and bringing it to a standardized form, which can be used then for many different purposes.

2. Processing of magnetic diagnostics data using HMM. Each HMM is characterized by the following parameters: 1) probability of initial stage; 2) matrix of transition from one stage to another; and 3) matrix of observation probability of some given set of features at each stage.

In order to find these parameters one should replace the original signal with the sequence of vectors from some given set, which is called the code book. The code book is constructed once as the result of processing of all or at least the major part of the diagnostics data. The processing consists of determining the set of characteristic features of the magnetic data and dividing them into classes (clustering) using the K-means algorithm. The set of vectors consisting of centroid (centre of mass) vectors [3] forms the code book.

In this paper we consider the signal from one Mirnov coil. Generalization of the method for all components of magnetic diagnostics is a subject of further research.

For determining the characteristic features the duration of each signal is divided into a set of overlapping segments. The influence of the boundary effects of segments is reduced by application of the Hamming function [4]:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N}\right).$$

Here n is the index of the time point in a segment $1 \le n \le N$, N is the number of points in each segment. Then the so-called cepstral coefficients

$$C_i(n) = F^{-1}[\log|F[s_i(n)w(n)]|],$$

see e.g. ref. [1], are calculated for each segment i, where $s(n)$ is the magnetic diagnostic data on the segment, F and F^{-1} are the direct and inverse Fourier transform. Cepstral coefficients allow one to compress data preserving the essential information of the signal, since usually $C_i(n)$ decays quickly with n. Typically the first $n \approx$20-30 components of $C_i(n)$ are enough.

Cepstral coefficients vectors are clustered by the K-means algorithm [1]. The result of clustering forms the code book, which is used for construction of the HMM. Each characteristic stage of the code book can be supplemented with appropriate data from other (non-magnetic) diagnostics for the purpose of integrated analyses. The HMM for each plasma pulse is constructed using the Bauman-Welsh procedure [2]. The code book renewal can be done only when the content of the magnetic diagnostics database substantially increases or new pulses are added.

3. Software. The HMM technique is implemented in the code MAGDI (Magnetic Diagnostics), which gives a convenient tool for visualization, clusterising, statistical analyses of magnetic data, study of the correlation of different diagnostics and integral analyses of data.

Development of user-friendly graphical software requires application of automation tools for the many routine operations required. In the problem under consideration this is access to data bases of different types and structure. A convenient system for diagnostics data processing and analyses can be built with the help of Kepler – a relatively new tool for modeling scientific workflows [5]. Kepler helps in the construction of complicated models, visualization of data flows in a model and integration of separate numerical and analytical applications.

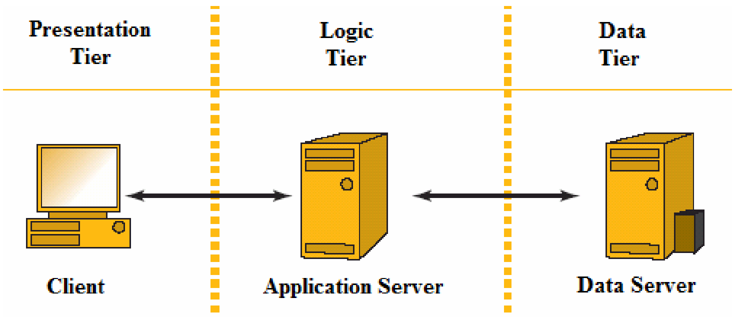

Fig. 1. General structure of the diagnostics data analyses system.

The code MAGDI is planned to be built in Kepler. The general concept of this integration is presented here. It can be useful for other systems of diagnostics data analyses. The three-tier model is applied "client - application server – data server (repository)", fig. 1. Application and data servers are responsible for the whole system functioning, data storage and security. Client is a stand-alone application for accessing servers. In the simplest case Client is a standard Internet browser running on a remote user computer, and application server and data server is a computer running Kepler and software for accessing databases on the network. The system allows a user to select a pulse from a database, build appropriate HMM, navigate in the database, i.e. find the nearest, the most distant, etc. pulse, find and visualize for each stage of HMM different (i.e. non-magnetic) diagnostic data. An example of this is given in the next section.

The structure of the Kepler-based system is presented in fig. 2. Kepler notation is used. Execution of this structure consists of the following steps.

1. Client sets the path to required data ("Path to client's data") and the scratch directory on the server. Then a ssh-session is opened with authentication, if downloading of the user files is required (Actor "Start Session").
2. The data together with a description is copied from the application server ("Transfer data") to the data-server ("Put to data base").
3. Client runs the data analyses program on the server, for example the code MAGDI, ("Execute cmd") by setting the path to an application on the server if necessary ("Path to application"), where the parameter "target" is the machine for job submission.
4. In the case of normal execution ("Is execution correct"), the path to video ("Path to video file") or other data to be plotted together with magnetic HMM are set for "Video

Diagnostic" actor and the result is displayed ("Show HMM"). The visual presentation of HMM consists of preset data of the code book and the newly calculated parameters of HMM for the pulse, see the example in the next section.

5. The parameters of HMM appropriate for data from different diagnostics (such as video) for the pulse are saved on the server and, if necessary, are copied to the client ("Copy HMM to client") to the client's scratch directory ("Result target").
6. Actors "Display the result of transfer", "Display the result of execution", and "Display the result copying" allow controlling the system at each stage.

This sequence of operations can be implemented by tuning standard Kepler actors. These take upon themselves the major routine work, which is usually reiterated manually and requires a lot of time.

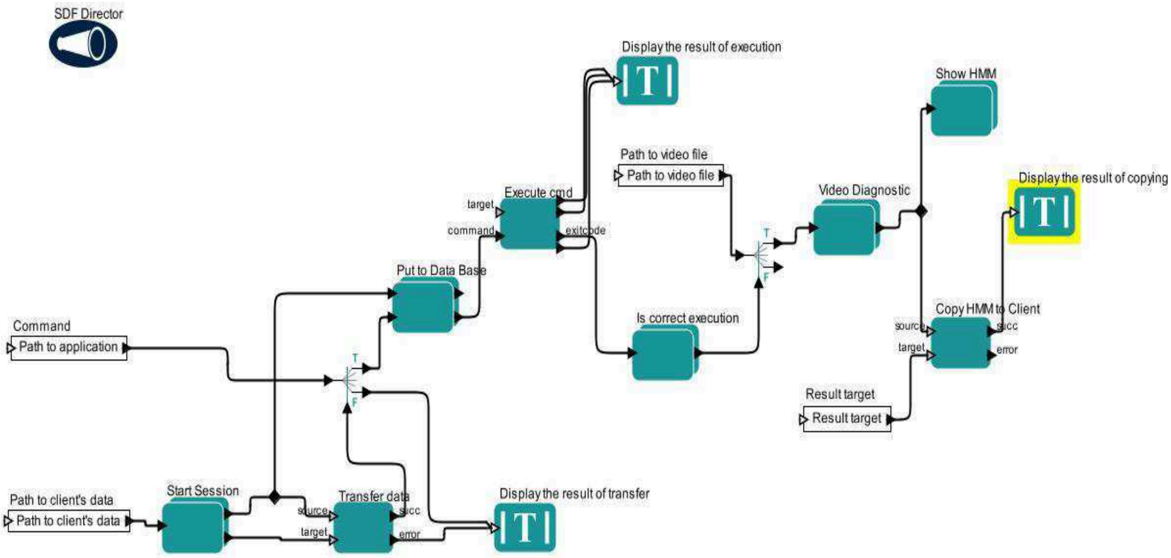

Fig. 2. General structure of HMM diagnostics data analyses in Kepler.

4. Results of modeling. By processing magnetic data from various Mirnov coils we found that practically each plasma pulse corresponds to a unique Markov model with just a few stages. Information can therefore be compressed from Gigabytes to kilobytes. Each stage of the HMM was characterized by the probability of initial status, the matrix of probabilities of transition to other stages and the matrix of probabilities of observing certain given plasma attributes, e.g. those obtained from other diagnostics (X-ray, optical, neutral, neutron, currents in poloidal field coils and solenoid, etc.) or from numerical modeling (safety factor, current density, type of instability, plasma shape, temperature and density profiles, etc.). In order to facilitate magnetic interpretation of diagnostics one can assign to each stage the characteristic sound de-

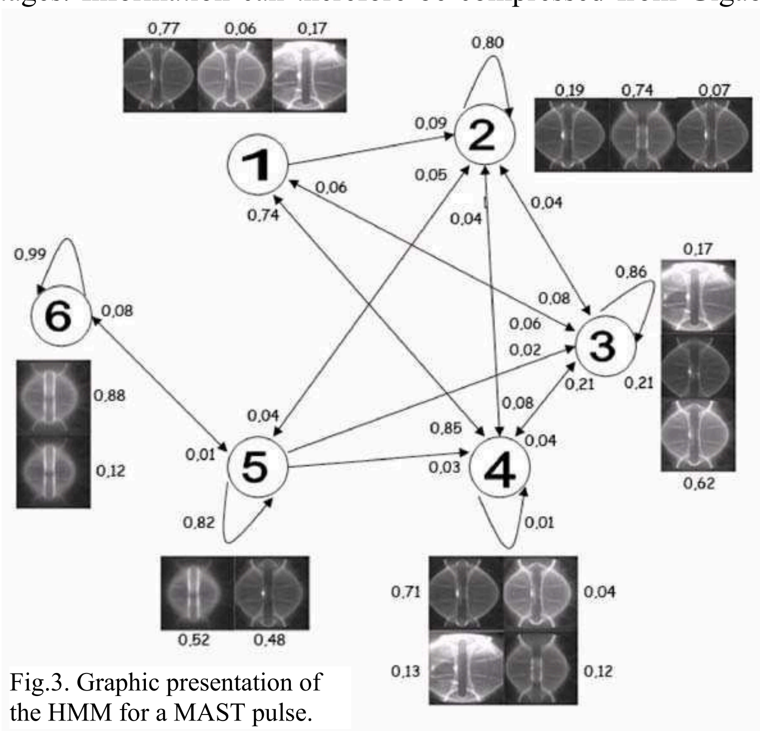

Fig.3. Graphic presentation of the HMM for a MAST pulse.

rived from magnetic oscillations and/or a video image. The HMM processed database allows quick navigation, i.e. finding the closest or the most distant pulse, or a pulse with preset properties, and comparison of different diagnostics.

An example of the HMM for a MAST pulse is given in fig. 3. The data from one Mirnov coil located in the equatorial plane was analyzed, though generalization for accounting for all coils is possible. Circles present plasma stages, arrows show possible transitions between stages, and digits near arrows give transition probabilities. Each circle is supplied with characteristic plasma images appropriate to the code book elements present in the stage. The probability of observation of the element of the code book in the stage is indicated near each video. A development feature useful to aid intuition matches each stage with an appropriate sound derived from magnetic oscillations.

Fig. 4 illustrates the elements of the code book. Graphs show the values of cepstral coefficients $C_i(n)$ as functions of n for some time interval i and video gives the appropriate plasma image. One can see a noticeable difference in the coefficients even in the case of similar videos, which indicates that cepstral coefficients catch substantial information about the plasma pulse.

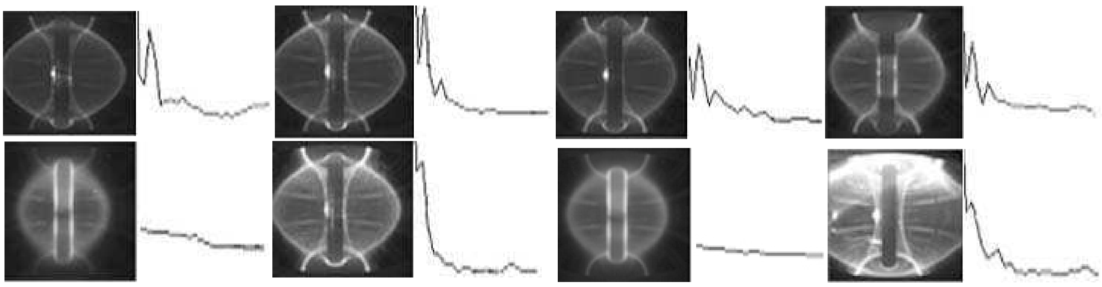

Fig.4. Some elements from the cepstral coefficients code book and the equivalent video image.

5. Conclusions. An example of a successful application of hidden Markov models (HMM) is given, which shows the efficiency of this new approach and motivates further applications.

A user-friendly software implementation is also presented. This can be helpful for standardization of software development for systems of diagnostics data analyses.

The HMM treatment of magnetic data with automatic superposition with other, non-magnetic diagnostics can bring the research to a higher integrated level of understanding of plasma behaviour allowing a non-specialist in diagnostics and data storage to access, visualize and study the data.

Acknowledgement. The work was supported by the Russian Foundation for Basic Research, grant N 07-07-00064 and FARG Ltd. (http://leader.ic.msu.su/~farg). UKAEA is gratefully acknowledged for permission to use data from the MAST spherical tokamak fusion experiment www.fusion.org.uk/mast.

References.

[1] Rabiner L.R. A Tutorial on Hidden Markov Models and Selected applications in speech recognition. Proceeding of the IEEE, Feb 1989. Vol. 77. N 2. [2] Hidden Markov Models. http://leader.cs.msu.su/~luk/HMM. [3] Moore A. K-means and Hierarchical Clustering - Tutorial Slides. http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html. [4] Hamming R.W. Digital Filters, Prentice Hall, third edition, 1989, 296 p. [5] http://www.kepler-project.org.