System Identification of Electronic Nose Data From Cyanobacteria Experiments

Graham E. Searle, Julian W. Gardner, *Member, IEEE*, Michael J. Chappell, Keith R. Godfrey, and Michael J. Chapman

***Abstract*—Linear black-box modeling techniques are applied to both steady state and dynamic data gathered from two electronic nose experiments involving cyanobacteria cultures. Analysis of the data from a strain identification experiment shows that very simple low order MISO black box model structures are able to produce very high success rates (up to 100%) when identifying the toxic strain of cyanobacteria. This is comparable with the best success rates for steady state data reported elsewhere using artificial neural networks. Analysis of data from a growth phase identification experiment using MIMO black-box models produces success rates of 82.3% for steady state data and 76.6% for dynamic data. This compares poorly with the best performing nonlinear artificial neural networks, which obtained a 95.1% success rate on the same data. This demonstrates the limitations of these linear techniques when applied to more difficult problems.**

***Index Terms*—Biological systems, identification, modeling, sensors.**

I. Introduction

In RECENT years, electronic nose instruments have been applied to numerous tasks using various different data processing techniques. They have shown promise for a wide range of applications, including the classification of the quality of tomatoes [1], determination of banana ripeness [2], and discrimination between different blends of coffee [3]. Commercial electronic noses are now routinely employed in the food, beverage and cosmetics industries.

In this paper, we consider the application of an electronic nose system to the classification of odors emitted by biological samples. The objective is to determine whether the odor discrimination abilities of electronic nose systems are sufficient so that in the future they can be used for such applications as bioprocess monitoring and medical diagnosis.

Most of the well-established techniques for analyzing data gathered from electronic nose experiments only involve the use of the steady state (static) responses of the sensors. The data are preprocessed to extract the steady state information from the dynamic data produced by the system. This steady state information is then used by a pattern recognition system to classify the odor. Popular techniques for analyzing this steady state information include principal components analysis, cluster analysis, discriminant function analysis, and artificial neural networks [1], [4], [5]. It has been suggested that there could be significant discriminatory information contained within the traditionally discarded dynamic data [6]–[8]. In this paper, linear time invariant black-box models are used to analyze electronic nose data. Such models have been shown to be effective when applied to steady state data [9], but here, their application to dynamic data is also investigated.

Two biological experiments are considered. The first was designed to test the ability of the electronic nose system to discriminate between two similar strains of bacteria, one of which was toxic, the other not. The identification of different strains of bacteria could clearly be beneficial for many applications; not only in environmental monitoring, but also in food industries (i.e., bacterial food spoilage) and medical diagnosis.

The second experiment was designed to investigate the ability of the electronic nose system to classify the growth phase of a bacterial colony. The levels of biological activity of the bacteria vary as the culture progresses through its life cycle, which consists of four distinct growth phases, known as the lag, log, stationary, and late stationary (or death) phases [10]. The experiment considered here was designed to test the ability of the electronic nose system to discriminate between the different growth phases of a single strain of bacteria. This presents a challenging problem for the system since the odors given off by the bacteria change only slightly over the course of their life cycle. Thus the classes between which we wish to discriminate are perhaps closer together (and less sharply defined) than those for the simpler first experiment. The knowledge of which growth phase a sample of bacteria is currently in provides a useful indication of the likely future progress (viability) of the bacteria. The phase defines the rates at which the cells take in substances from around them. Thus, information about the current growth phase could be used to accurately predict the dosage levels of antibiotic required to challenge the bacteria. If utilized in similar medical applications, such predictions would yield substantial benefits to the healthcare industries.

II. Strain Identification Experiment

A. Black-Box Modeling

The models considered here are, in effect, inverse black-box models of the electronic nose system. A forward model of the system would have inputs corresponding to the odor input $y(t)$ to the system, and outputs corresponding to the electrical resistances of the sensors responding to those odors $u(t)$. Our models

Manuscript received February 14, 2001; revised March 24, 2002. This work was supported by the EPSRC under Grant GR/M 11943. The associate editor coordinating the review of this paper and approving it for publication was Dr. H. Troy Nagle.

G. E. Searle, J. W. Gardner, M. J. Chappell, and K. R. Godfrey are with the School of Engineering, University of Warwick, Coventry, U.K.

M. J. Chapman is with the School of MIS-Mathematics, Coventry University, Coventry, U.K.

Publisher Item Identifier 10.1109/JSEN.2002.800286.

work in the opposite direction, taking the sensor resistances as inputs, and producing a classification of the odor as an output. See Fig. 1 for a pictorial representation of this process.

For the experiment considered here, the output required is a simple classification as toxic or nontoxic, rather than a numerical value. In other applications, there may need to be more output classes (e.g., the similar problem considered in Section III). These simple classifications can be obtained from black-box models outputs in a number of ways, requiring the use of either single or multi-output black-box models.

The *System Identification Toolbox* [11] within MATLAB (Version 5) provides the user with many functions for creating, evaluating and using black-box models. The general form for the discrete time multi-input, single-output (MISO) models considered is given as follows[12]:

$$\begin{aligned} A(q^{-1})y(t) = &\frac{B_1(q^{-1})}{F_1(q^{-1})}u_1(t-k_1)+\cdots \\ &+\frac{B_i(q^{-1})}{F_i(q^{-1})}u_i(t-k_i)+\cdots \\ &+\frac{B_n(q^{-1})}{F_n(q^{-1})}u_n(t-k_n)+\frac{C(q^{-1})}{D(q^{-1})}e(t) \quad (1)\end{aligned}$$

where y is the discretised model output, u_i is the i-th input (for $i = 1 \ldots n$), e is white noise, and k_i is the delay from input u_i to the system. The functions $A(q^{-1}), B_i(q^{-1}), C(q^{-1}), D(q^{-1})$, and $F_i(q^{-1})$ are polynomials in the backward shift operator q^{-1} that is defined by

$$q^{-1}x(t) = x(t-1). \quad (2)$$

Not all of the polynomials A, B_i, C, D, and F_i are used in a particular model. The models considered in this paper are listed in Table I, together with details of the polynomials used. The model structures were selected because of their simplicity and previously reported success [9]. Finite impulse response (FIR) models only use the present and past values of the inputs (the u_is) in order to produce an output. Auto-regressive with exogenous inputs (ARX) models also use past values of the (simulated) outputs. The addition of a moving average term to FIR and ARX models to produce MAX and ARMAX models, respectively, corresponds to the inclusion of a C polynomial to the model structure [see (1)]. This allows more effective modeling of the noise characteristics of the system. Note that the polynomials $D(q^{-1})$ and $F_i(q^{-1})$ do not appear in Table I. They were not used in any of the model structures considered, but are included in (1) for generality.

B. Experimental Details

The strain experiment was designed to evaluate the ability of an electronic nose system to discriminate between two strains of cyanobacteria (blue-green algae), one toxic and the other nontoxic. The ability of such a system to classify quickly and accurately the strain of bacteria present in an algal bloom could clearly be useful to environmental agencies, monitoring reservoirs and lakes.

The headspaces of separate cultures of the two strains of cyanobacteria, grown in a nutrient medium (BG11), were sampled periodically by an electronic nose system over 40 days [10]. The nose system used consisted of six commercial metal oxide resistive odor sensors (Alpha MOS, France), and two other sensors to monitor ambient temperature (LM35CZ, National Instruments) and humidity (MiniCap 2, Panametrics) [10]. The repeated exposure cycle was as follows.

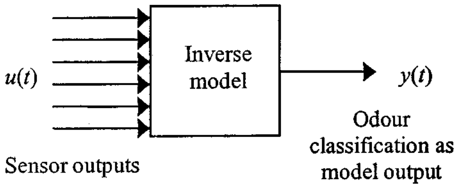

Fig. 1. Diagrammatic representation of the electronic nose modeling problem.

TABLE I
LINEAR BLACK-BOX MODEL STRUCTURES USED TO ANALYSE THE ELECTRONIC NOSE DATA

Model structure:	Polynomials used:
FIR	B_i
ARX	A, B_i
MAX	B_i, C
ARMAX	A, B_i, C

- 23 min 20 s—medium only
- 2 min—medium with toxic *microcystis aeruginosa* PCC 7806 strain
- 23 min 20 s—medium only
- 2 min—medium with nontoxic *microcystis aeruginosa* PCC 7941 strain.

The outputs from the sensors were sampled every 10 s, producing 350 358 data vectors corresponding to 1 150 exposure cycles. A plot of a section of the data showing the response of a single odor sensor is given in Fig. 2. The fluctuations in the baseline signals are attributed to variations in the ambient air.

The data obtained from the experiment have previously been preprocessed to extract static parameters and subsequently analyzed using artificial neural networks with considerable success [10]. Here, we use system identification techniques to analyze the same data in order to compare the efficacy of linear time-invariant black-box models, for both static and dynamic data, with the static data based nonlinear neural network methods.

Two distinct classes of models were analyzed: in Section II-C, models for the steady state responses of the sensors (static models) and in Section II–D, models for the full temporal data (dynamic models).

C. Static Models

Previous work on the analysis of data from electronic nose experiments has mostly concentrated on using preprocessing algorithms to extract the steady state features from the sensor signals. Accurate results have been achieved using only this steady state information [10]. There are also other benefits to using this steady state data, some of which are listed as follows.

- Information compression—the full (dynamic information intact) data-sets can be very large, especially when the data sampling rate is high. The preprocessing compresses this information to a single value per sensor per odor exposure cycle.

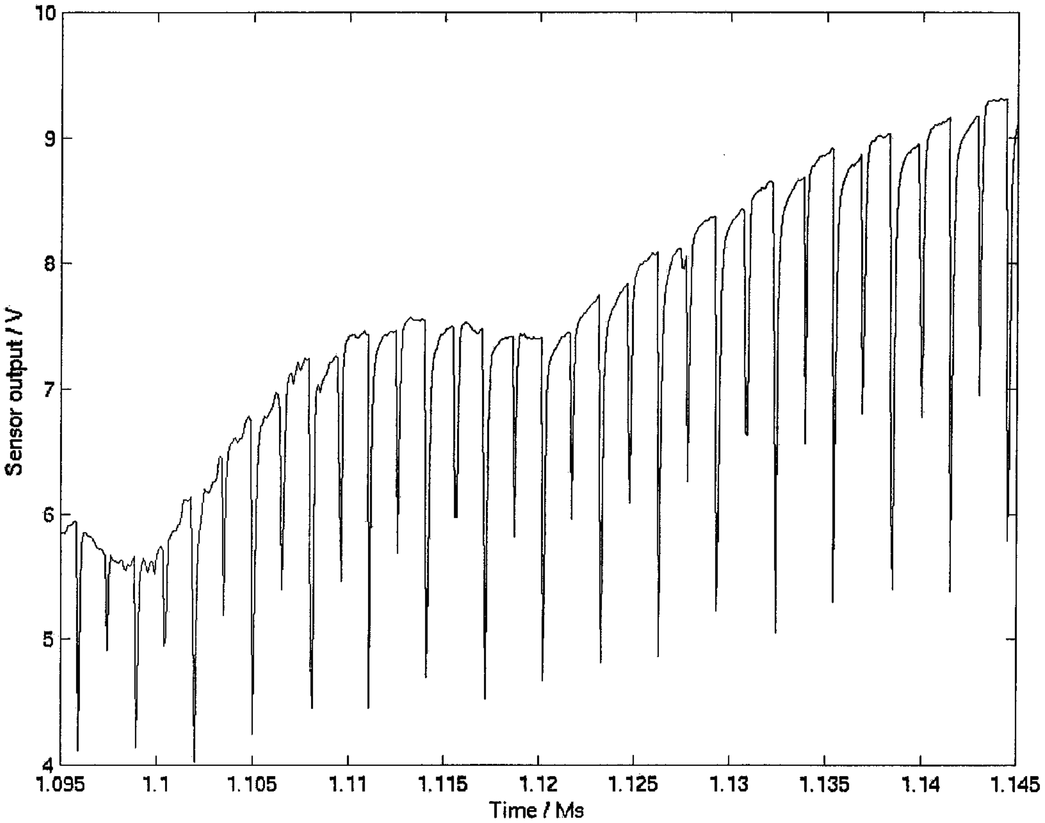

Fig. 2. A 14 h section of the raw data, showing considerable variation in the output level of a single odor sensor during part of the experiment. The voltage is directly proportional to sensor resistance.

- Baseline drift removal—the gas sensors employed are susceptible to poisoning effects that produce long term systematic drift in the baseline (sensor resistance values in air). Some of the preprocessing algorithms help to counteract these effects, an example is the difference algorithm below.
- Reduction of temperature dependence—the resistances of the odor sensors are highly dependent on the temperatures of the various components of the nose system. Some physical models for the effects of temperature variations on the sensor responses suggest that an appropriate choice of preprocessing algorithm should reduce the effects of the temperature variations on the data being analyzed [13].

These preprocessed data are traditionally then passed on to pattern recognition algorithms or artificial neural networks for classification [5], [14].

In this section, black-box models for these preprocessed (or static) data are analyzed. The four preprocessing algorithms tested are as follows.

- Absolute response: V_b;
- Difference: $(V_b - V_m)$;
- Relative difference: (V_b/V_m);
- Fractional difference: $((V_b - V_m)/V_m)$

where V_b is the sensor voltage when exposed to the headspace of the bacteria sample (final value at end of exposure period) and V_m is the sensor voltage when exposed to the headspace of the medium (immediately before the sensor is exposed to bacteria). The preprocessed sensor outputs for each of the six gas sensors are considered to be the inputs to MISO black-box models.

The preprocessed values are heavily influenced by the intensity of the odor to which the sensors are exposed. For certain quantitative applications, this can be useful, since one might wish to obtain an estimate of the concentration of a gas in some sample. However, for this experiment, the aim is only to discriminate between the two classes of odor (the toxic bacteria and the nontoxic bacteria). For this reason, the effects of normalizing the data vectors in order to try and remove the concentration information contained within the data are investigated. The normalization process used for this experiment is simply to take each six-dimensional data vector and to divide it by its Euclidean length, so that each data point is projected onto the surface of the six-dimensional unit hypersphere.

The evaluation of each model structure and preprocessing algorithm was carried out using the following procedure.

- The chosen preprocessing algorithm was applied to the data to produce an array (2300 × 6) of input data for the inverse models, normalized if required.
- An output vector was formed using knowledge of the experimental details. A classification output of toxic was encoded as $+1$, and nontoxic as -1.
- The data set was split at the half-way point. The first half to be used for training of the models, and the second to be used for testing.

- A MISO black-box model was trained, i.e., MATLAB was used to estimate the polynomial coefficients in the selected model structure.
- The model was used to produce a simulated output vector from the second half of the data set.
- The elements of the simulated output vector were converted to classifications simply by taking their sign. Finally, this classification vector was compared with the known classification vector in order to evaluate the success of the model. Note that this method forces the classification system to always produce a classification and does not allow for a "not known" result.

1) FIR Models: Finite impulse response (FIR) models are the simplest of the black-box model structures considered here. A zero-delay FIR model forms a prediction of the system output at time t by simply taking a linear combination of the system inputs at time t and at times $t-1, t-2, \dots t-b_i$, where b_i is the order of the polynomial B_i in the model (see Section II.A). Depending on the nature of the system being modeled, it may be appropriate to include delay terms [the k_i parameters in (1)]. These could be different for each of the model inputs; thus, for a 6-input 1-output FIR model, the structure of the model is determined by the orders of the six B polynomials, and the six delays k_i. However, due to the fundamental similarity between our six inputs (i.e., same type of semiconducting oxide gas sensor), our investigation was restricted to models using the same order for all six B polynomials, and the same delay for all six inputs.

Early investigations showed that the most appropriate delay was zero. This was expected with this static data since a single data point corresponds to a whole exposure cycle. Hence, to introduce a nonzero delay would be to expect the model to classify an odor using none of the data recorded during the sampling of that odor.

The most effective orders for the B polynomials were either two or three for each of the eight preprocessed data sets (for all four different algorithms and normalized/raw data). The results obtained using models with these optimal structures are given in Table II.

The normalization process carried out on the preprocessed data had the effect of reducing the success rates slightly. As mentioned before, the aim of the normalization procedure is to remove some of the intensity of odor information to leave mainly type of odor information. The fact that this process reduced the ability of the models to classify correctly the odors in the test data set would seem to indicate that the models for preprocessed data (raw) were using a small amount of this intensity information to produce the classifications. In a field-based application, it would be impossible to control the intensities of the signals, so for such purposes, one might expect normalized data processes to yield greater success rates.

The fact that the absolute response algorithm produced a noticeably poorer success rate than the other three algorithms could be attributed to the fact that this algorithm is the most affected by the long term baseline drift evident in the data. The normalization process reduces these effects somewhat, enabling the normalized absolute response models to perform comparably with the other models using normalized data.

TABLE II
CLASSIFICATION SUCCESS RATES OBTAINED USING ZERO-DELAY FIR MODELS OF ORDERS TWO OR THREE FOR STATIC (PREPROCESSED) DATA

	Percentage success rate using	
Pre-processing (static) algorithm used:	Pre-processed data:	Normalised pre-processed data:
Absolute response	98.8	97.6
Difference	99.7	97.4
Fractional difference	99.5	97.0
Relative difference	99.5	98.4

2) More Complex Model Forms: The FIR models in the previous section produced very high classification success rates, however the inclusion of either A or C polynomials to the model structure to form auto-regressive with extra inputs (ARX) or moving average with extra inputs (MAX) models increased the success rates obtained, perhaps because the more complex model structures are better able to handle the drift evident in the data.

By selecting appropriate orders for the B and A polynomials in an ARX model, a successful classification rate of 100% was obtained using each of the eight preprocessed data-sets. The orders required were mostly one for the B polynomial and between one and three for the A polynomial.

The extension of the FIR model structure to a MAX model structure produced similar success rate increases. Success rates of 100% were achieved for seven of the eight data sets, with 99.9% achieved on the remaining data set. The orders required for these models were mostly two for the B polynomials and between zero and three for the C polynomial. The MAX models generally required more parameters (polynomial coefficients) than the ARX models, making the ARX models preferable in terms of simplicity and computational efficiency. Further extension of the model structures to form ARMAX models was not found to produce any significant improvement over the ARX and MAX models.

It should be noted that, for the training and testing of these static models, the temporal order of the data was maintained. It might be considered more realistic to randomly reorder the data vectors to avoid the possibility of the models learning the characteristics of the particular experiment rather than the characteristics of the different odors. In order to enable fair comparison with the dynamic models in the next section (for which random reordering would not be as straightforward), the data sets used in this section were not reordered. However, it was noted that, when the static data sets were randomly reordered (100 times, average results taken), the success rates were only reduced by a few percent (typically around 4%).

D. Dynamic Models

The models for the static data used in the previous section achieved considerable success, but it has been suggested [6]–[8] that there is also useful information contained within the dynamic data which the preprocessing algorithms discard. In this section, we form linear time-invariant black-box models for the full dynamic data in order to investigate this possibility. We also investigate the effects of normalizing the sample vectors (as in Section II.C).

The model structures tested are the same as those for the static data in the previous section. The crucial difference is that, for the static models, a model of order d, say, takes into account the last d exposure cycles, whilst for the dynamic models, a model of order d takes into account the last d sampled data points, i.e., the last $10 \times d$ s (since our data was sampled at 0.1 Hz).

The evaluation of each model structure was carried out using the following procedure.

- If required, the data-set was normalized to form an array ($350\,358 \times 6$) of input data.
- As before, an output of toxic was encoded as $+1$, and nontoxic as -1. Here, however, an output of medium was required for the periods where no bacteria were sampled. This was encoded as 0. An output vector was accordingly formed.
- The data-set was split into halves. Data from the first 20 days of the experiment were used for training the models, and data from the remaining 20 days were used for testing.
- A MISO black-box model of the chosen structure was trained on the first half of the data using MATLAB.
- The model obtained was used to produce a simulated output vector from the second half of the data.
- The simulated output was sampled at the time point just before the system switched from bacterial odor input to medium input, for each exposure cycle, thus producing a single numerical output for each exposure cycle.
- The vector of numerical outputs was converted to a vector of encoded classifications by taking the sign of each element. This vector was then compared with the known sequence of classifications in order to evaluate the success of the model.

1) FIR Models: The first model structures considered were FIR models. As with the static models, the set of structures considered was restricted to those having the same order polynomials for each of the inputs, and similarly the same delay on each input. FIR models for both the raw and normalized data-sets with B polynomials of orders from one to 25 and having delays from zero to 5 data points (corresponding to zero to 50 s) were formed and evaluated. Note that the orders of each of the six B polynomials were the same for any given model. The optimal structures for each data-set were chosen as a compromise between a high successful classification rate, and a simple model structure, i.e., one having few parameters requiring estimation.

With the static models in the previous section, the optimum delay was found to be zero, as expected. However, for the dynamic models in this section, a nonzero optimum delay would not be unexpected. The delay required in the model reflects the physical characteristics of the system in question. The sample vessels were connected to the sensor chamber via a system of pipes and valves, thus one might expect there to be a noticeable time lapse between the switching of the valves and the arrival of the new odor at the sensor array. However, for our data-set, no delay was observed. This could be explained by the relatively low dead volume in the pipework (V_d) compared with the volumetric flow-rate of the pump $(\dot{Q}_p)$ producing a physical delay that was short enough to be undetectable at the 0.1 Hz sampling frequency used [i.e., $(V_d/\dot{Q}_p)$ much less than 10 s].

Using the raw data-set, the optimum model order was found to be 12. This corresponds to an input information window of 120 s being used by the model in order to calculate an estimated output at a given time. The successful classification rate achieved by this model on the test data was 91.3%. A slightly higher success rate (91.8%) could be achieved using a model of order 22, but the lower order model was deemed more suitable due to the insignificant (and unlikely to be reproducible) difference and the greatly reduced computing costs associated with using a model with 72 polynomial coefficients rather than 132.

The models for the normalized data-set achieved greater success than those for the raw data. The optimum model order was found to be 11, producing a success rate of 99.3%. The fact that the normalization process improved the success rates for the dynamic models contrasts with the situation for the static models, where the normalization process reduced slightly the classification success.

As discussed in Sections II.C and II.C.1, the normalization process reduces the intensity information within the data and increases the significance of the odor type information. For this dynamic data-set, it would appear that the models for raw data were partially classifying the odors based on correlations in the training set between the intensity of the signal and the correct classification. Thus, when the models were applied to the testing set, where these correlations were no longer valid due to long term drifts in sensor responses, incorrect classifications were made. With the intensity information removed by the normalization process, the models were able to make their classifications based on the type of odor and thus generalize more successfully from the training to testing sets.

The sensor chamber contained a temperature sensor and a humidity sensor in addition to the six odor sensors. The activity levels of the bacteria and the responses of the odor sensors are both known to be sensitive to changes in ambient temperature and humidity. For this reason, the inclusion of the outputs of these sensors as extra inputs to the FIR models was investigated, but found to produce no significant improvement using either the raw or normalized data set.

2) ARX Models: The inclusion of an A polynomial to the black-box structure to form ARX models allows the model to consider past values of the model output as well as past and present values of the inputs. The delays on the inputs were again set to zero. Models with orders from zero to 15 for the A polynomial and from one to 15 for the B polynomials were formed and evaluated for both the raw and normalized data sets. The percentage successful classification rates obtained on the test data set by each of the 240 models for the raw data are plotted in Fig. 3.

It is evident from Fig. 3 that the order of the B polynomials is a more significant factor in the success of the model than the order of the A polynomial. For low order B polynomials (around 3–7), the inclusion of a low order A polynomial increased the success rate slightly. However, the greatest success rate overall for this raw data set was obtained using the order 12 FIR model of the previous section (91.3%).

The results for the normalized data set are similar in that the greatest success was observed when no A polynomial was in-

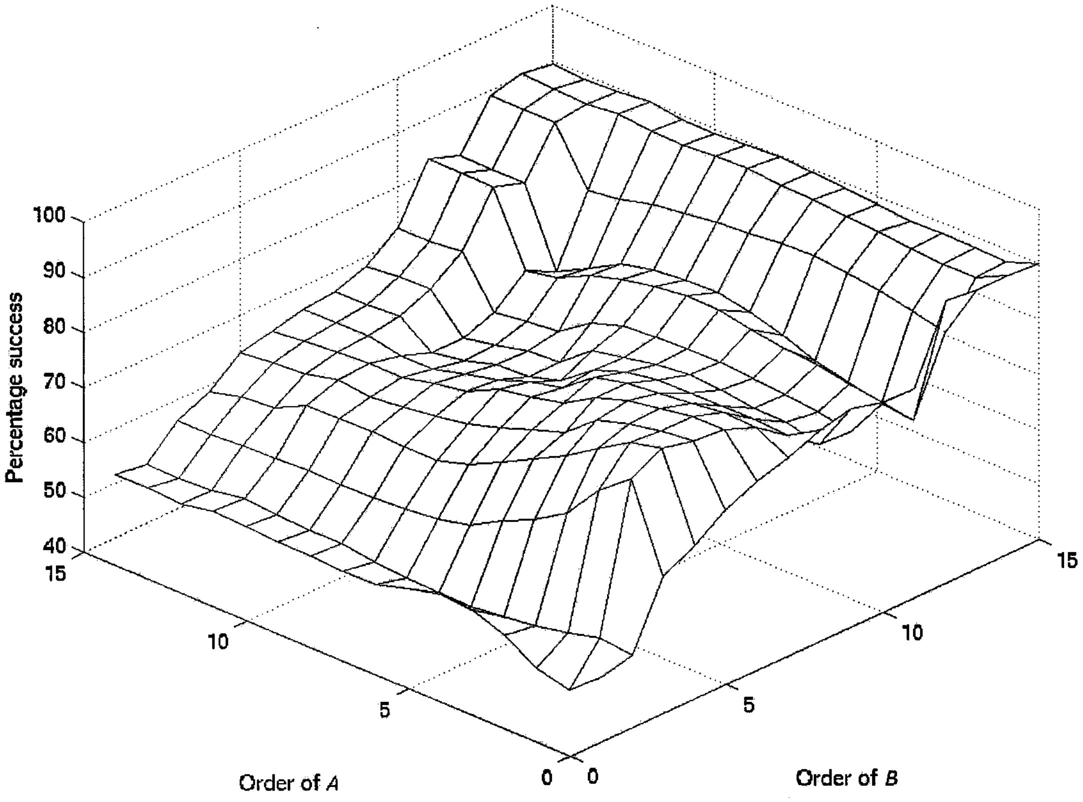

Fig. 3. Plot of successful classification rates against orders of A and B polynomials when ARX model structures (with zero input delays) were tested on the raw data set.

cluded in the model structure, so that the maximum success rate (99.3%) was obtained with an 11th order FIR model.

3) MAX and ARMAX Models: The addition of a C polynomial (a moving average term) to the FIR and ARX model structures to form MAX and ARMAX model structures was also investigated.

MAX models with orders from one to 15 for the B polynomials and from zero to 15 for the C polynomial (and again, zero delays) were investigated. Again, it was found that the optimum models only involve the B polynomials (using both the raw and normalized data sets). This indicates that the extension from an FIR model to an MAX model is not beneficial.

Similarly, the addition of an A polynomial (an auto-regressive term) to the MAX models to form ARMAX models did not increase the success of the models, This suggests that (recent-10 to 150 s previous) past output values do not provide any useful information that can be used to produce a more accurate value for the current output. This makes sense when the observed sensor drift seems to vary diurnally rather than by the minute—probably associated with changes in the ambient temperature of the bacterial samples.

III. Growth Phase Identification Experiment

A. Black-Box Modeling

For this experiment, the bacteria must be classified according to which of four possible growth phases the bacteria are currently in. Quantitative information about the strength or concentration of the odor inputs is unnecessary. These simple classifications can be obtained from black-box model outputs in a number of ways, requiring the use of either single or multi-output black-box models.

In this section, the model structures used were, without exception, multi-input, multi-output (MIMO) finite impulse response (FIR) models. This is in contrast with the simpler strain identification experiment considered in Section II, where FIR, ARX, MAX, and ARMAX models were considered. The model structure considered was selected due to its simplicity (and thus computational efficiency) and previous success (in Section II and [9]). FIR models effectively just take a linear combination of the present and past values of the inputs (the u_is) in order to produce an output.

For example, MIMO FIR models with n inputs and m outputs have the general form [12]

$$\mathbf{y}(t) = \mathbf{B}(q^{-1})\mathbf{u}(t) + \mathbf{e}(t) \tag{3}$$

where $\mathbf{y}$ and $\mathbf{e}$ are m-dimensional vectors, $\mathbf{u}$ is an n-dimensional vector and $\mathbf{B}$ is an $m \times n$ matrix. The elements of $\mathbf{B}$ are polynomials in the backward shift operator q^{-1}, which is defined by

$$q^{-1}x(t) = x(t-1). \tag{4}$$

Thus, $\mathbf{B}$ can be written

$$\mathbf{B}(q^{-1}) = \begin{pmatrix} b_{11}(q^{-1}) & b_{12}(q^{-1}) & \cdots & b_{1n}(q^{-1}) \\ b_{21}(q^{-1}) & b_{22}(q^{-1}) & \cdots & b_{2n}(q^{-1}) \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1}(q^{-1}) & b_{m2}(q^{-1}) & \cdots & b_{mn}(q^{-1}) \end{pmatrix} \tag{5}$$

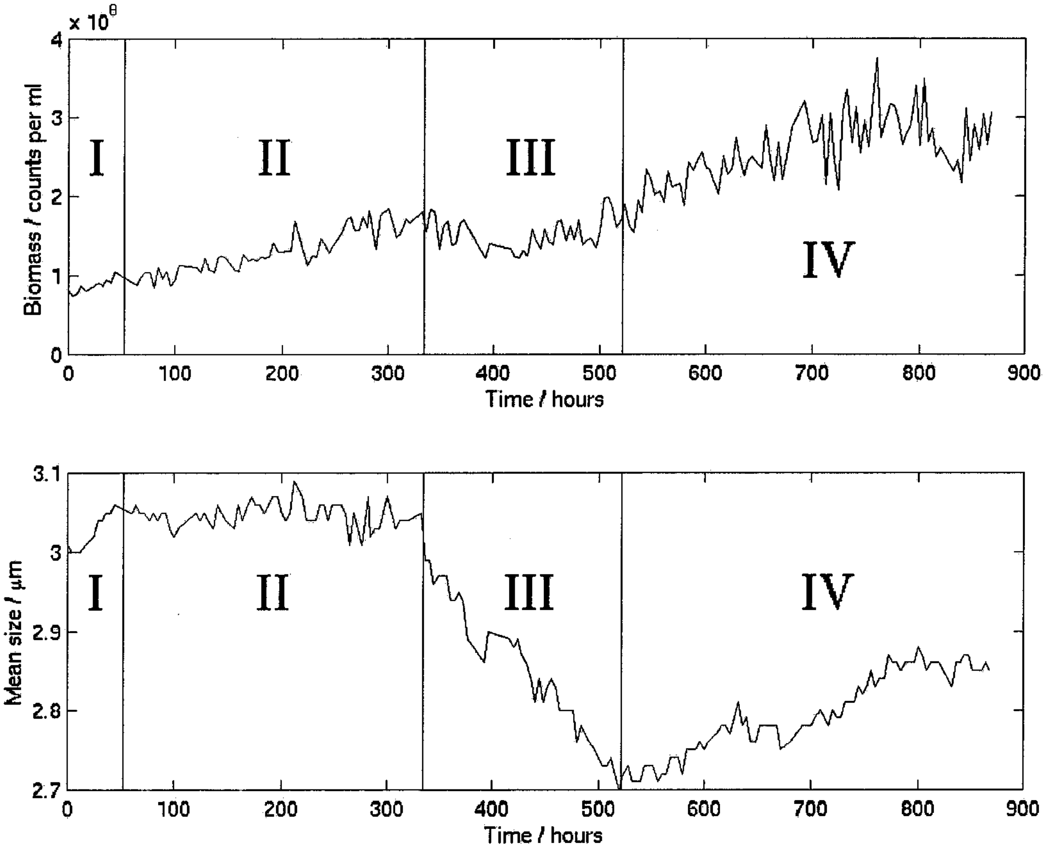

Fig. 4. Plot of the data from the CellFacts instrument for the growth phase identification experiment. The upper plot shows the general increase in biomass (cell counts) with time. The lower plot shows the variation in mean size of the bacteria cells with time. The four growth phases (lag, log, stationary and late stationary), are labeled I to IV in each plot. This plot shows the subtle changes in biomass between phases.

where for $i = 1 \ldots m$ and for $j = 1 \ldots n$:

$$b_{ij}(q^{-1}) = b_{ij}^{1} q^{-k_{ij}} + \cdots + b_{ij}^{d_{ij}} q^{-(d_{ij}+k_{ij})} \tag{6}$$

where d_{ij} is the order of the polynomial b_{ij}, and k_{ij} is the delay from input j to output i.

B. Experimental Details

The experiment considered here involved a single (toxic) strain of cyanobacteria, monitored over a 40 day period. As well as electronic nose data, information concerning the mean size of the bacterial cells and the biomass present in the cultures was recorded using a CellFacts instrument (Microbial Systems, Ltd.). This enabled the identification of the four distinct growth phases through which the bacteria pass during their life cycle. Neural networks have been successfully used by Shin *et al.* [10] with the electronic nose data to classify the bacteria into each of the four growth phases, obtaining classification success rates of up to 95.1%. In this section, we apply MIMO linear black-box models to the same data in order to investigate the ability of such techniques to tackle this more challenging problem.

The experiment in question was intended not only to test the ability of an electronic nose to discriminate between the different growth phases of a cyanobacteria strain, but also to investigate the reproducibility of the measurements and success rates. For this reason, the experimental system consisted of three vessels, two containing nominally identical cultures of toxic *microcystis aeruginosa* PCC 7806 in nutrient medium (BG11) and one reference vessel containing only the nutrient medium. The headspaces of these vessels were connected via a system of pipes and computer-operated valves to an electronic nose system. The nose system used consisted of six commercial metal oxide resistive odor sensors (Alpha MOS, France), and two other sensors to monitor ambient temperature (LM35CZ, National Instruments) and humidity (MiniCap 2, Panametrics) [10]. The repeated exposure cycle was

- 50 min—medium only;
- 5 min—medium and toxic *microcystis aeruginosa* PCC 7806 strain sample 1;
- 50 min—medium only;
- 5 min—medium and toxic *microcystis aeruginosa* PCC 7806 strain sample 2.

The sensor outputs were again sampled every 10 s, producing 361 698 data vectors corresponding to 548 full exposure cycles.

The information collected using the CellFacts instrument was used to produce a correct classification vector for the data, identifying the four growth phases of the bacteria using understanding of the biological processes involved. It should be noted that the boundaries between the growth phases were by no means sharp, thus making the identification problem highly nontrivial. Some of the data from the CellFacts instrument are plotted in Fig. 4.

TABLE III
SUCCESS RATES FOR FIR MIMO MODELS OF ORDERS BETWEEN ONE AND FOUR FOR STATIC (PREPROCESSED) DATA FROM THE GROWTH PHASE EXPERIMENT. THE PERCENTAGES ARE AVERAGES OVER TEN DIFFERENT RANDOM REORDERINGS USING TWO-FOLD (50%) CROSS-VALIDATION.

	Percentage successful classification using pre-processing algorithm:			
Data set:	Absolute response	Difference	Fractional difference	Relative difference
Toxic 1	66.1	75.7	78.1	78.6
Toxic 1 (normalised)	70.2	68.5	73.9	77.5
Toxic 2	67.3	82.3	81.4	78.9
Toxic 2 (normalised)	67.5	80.3	75.2	77.7

C. Static Models

Both static and dynamic linear black-box models are considered. This section deals with static models for the system. The preprocessing algorithms used were the same as those in Section II, namely absolute response, difference, relative difference, and fractional difference (for definitions of these, see Section II–C). However, the details of how the black-box techniques should be applied to the data were less clear-cut. In Section II, the basic aim of the modeling was to produce an algorithm for classifying an odor into one of two classes (toxic or nontoxic), so the classification output was encoded simply as $+1$ for a classification of toxic, and -1 for nontoxic. In this experiment, the odor must be classified into one of four classes corresponding to the current growth phase of the bacteria. Hence, the simple numerical encoding of the output used in Section II is no longer appropriate.

Two methods of encoding the classification output of the model into a numerical output were considered. The first was an analogue of the method used in Section II, whereby the four growth phases were each allocated a numerical label [a (possibly zero) integer] between -2 and $+2$ and a multi-input, single-output model was formulated using the given phase labeling system. All of the different labeling combinations possible were tested to find the optimum labeling system for each model structure. This method was time consuming, though moderately successful, producing a maximum success rate of approximately 66% when trained on half of the static data, and tested on the second half.

The second method considered was the use of multi-output models to encode the different classification. An inverse model with four outputs (and six inputs, as before) was chosen, corresponding to the four different possible classifications of the odor. Thus a perfect classification of growth phase 1 would be a four-dimensional vector with a $+1$ in the first coordinate and zeros elsewhere. This method was found to be more successful than the single-output model method. Though it should be mentioned that analogous two-output models were tested for the data in Section II but were out-performed by the single-output models previously used.

The effects of normalizing the sample data vectors were also investigated. The evaluation of each model structure and preprocessing algorithm was carried out as follows.

- The chosen preprocessing algorithm was applied to the data to produce an array (1096×6) of input data for the inverse models, normalized if required.
- An output array (1096×4) was formed using the data from the CellFacts instrument. A classification output of growth phase i was encoded as a 4-dimensional vector with $+1$ in the i-th position and zeros elsewhere.
- The data vectors corresponding to the Toxic 1 bacteria culture were separated from those corresponding to the Toxic 2 culture, thus forming two separate data-sets (each with 548 input and output vectors).
- Each data-set was randomly reordered and then split into two halves. The first to be used for training and the second for testing of the models.
- A MIMO black-box model was trained (see Section II for details of the MIMO model structure used).
- The model was used to produce a simulated output array from the second half of the data-set.
- Each output vector (row in the output array) was converted to a growth phase classification by choosing the output with the largest (positive) value.
- The resulting classification vector was compared with the actual growth phase data to evaluate the success of the model.

Note that, for this growth phase data set, the data were randomly reordered (unlike those in Section II). This was done for two main reasons:

- to give a more realistic indication of how well the models could perform on real world data;
- to enable cross validation testing; unlike the experiment in Section II, the different classes are not distributed evenly (with time) through the data-set. Thus, training the model on the first half of the data and testing on the second half would be nonsensical, since only growth phases 1 and 2 would be seen by the model training algorithm, so it would be impossible to get useful results when testing the model on data corresponding to phases 3 and 4.

1) MIMO Static Models: As with the strain identification experiment detailed in Section II, fairly low order models were found to produce the best compromise between model simplicity and success rate. The exact order used varied across the 16 preprocessed data sets (choice of two bacteria cultures, four algorithms, each one subsequently normalized or left unchanged) varied between one and four.

For the Toxic 1 bacteria, the most successful of the preprocessing algorithms was the relative difference algorithm, not normalized, producing a success rate of 78.6% using a model of order four. For the Toxic 2 bacteria, the success rate was highest (82.3%) using the static difference algorithm, again not normalized, this time using a model of order three. For the results using other preprocessing algorithms, see Table III.

Notice that, as with the strain identification experiment of Section II, the normalization process reduced the success rate slightly in most cases. This can be attributed again to the odor

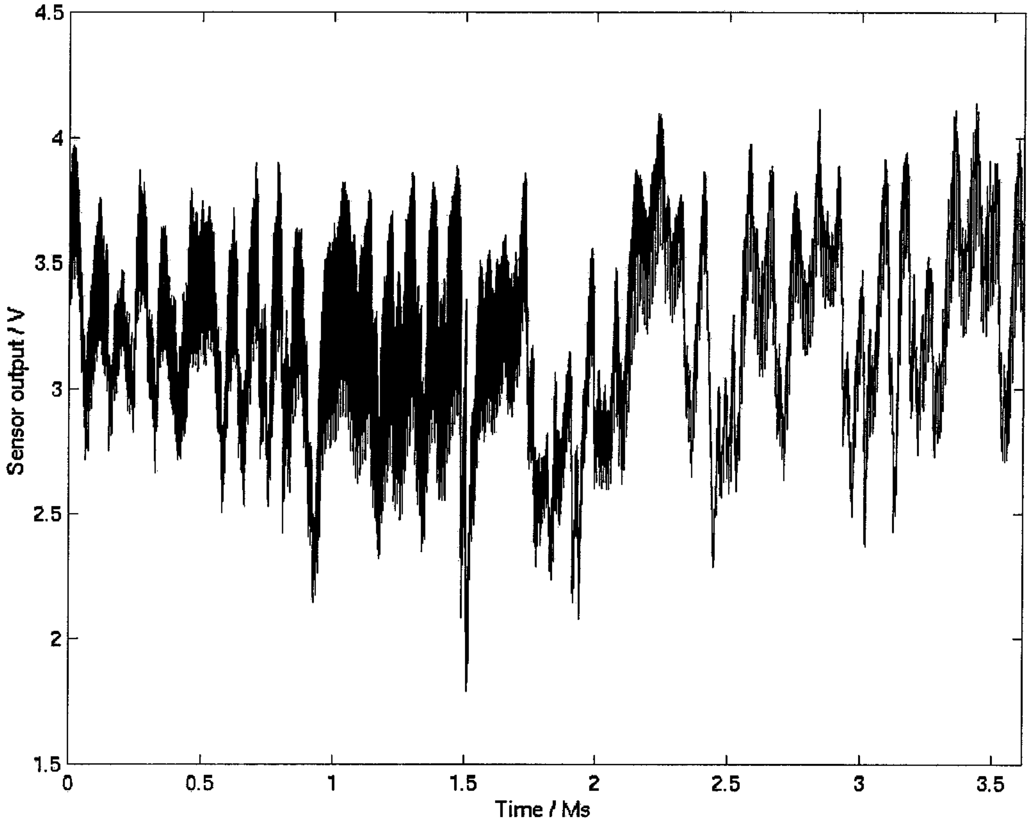

Fig. 5. Plot showing the considerable long term variations in the output voltage from a single sensor over the whole (40 day) period of the growth phase experiment. There is clear evidence of diurnal peaks probably associated with the temperature of the biomass. The temperature of the room fluctuated daily.

intensity being related to the cell count and hence growth phase. Also note that the absolute response preprocessing algorithm was markedly less successful than the other three algorithms.

D. Dynamic Models

The MISO dynamic models for the bacterial strain identification experiment considered in Section II achieved considerable success. In this section we consider similar MIMO models applied to the bacterial growth phase data.

As with the static models in Section III–C, the survey of MIMO model structures is restricted by the available computing algorithms and time constraints to FIR MIMO models of various orders. However, unlike the case for the strain identification experiment in Section II, the appropriate method for preparing the data for the models is not obvious. As mentioned in Sections III–B and III–C, the experimental procedure was such that the data-set obtained was effectively two data-sets interleaved. For the preprocessed (static) data, this posed no real problem; however, for the dynamic data, things are less straightforward. As with the static data, it was desirable to treat the data from the two cultures separately, but this necessitated splitting and reforming the data-set into two halves. The sensor values drifted significantly over the course of the experiment (see Fig. 5) so, whilst splitting up the data-set into two, it was decided to shift each response segment to remove baseline drift in an analogous manner to the difference preprocessing algorithm for the static data. The effects of this dynamic preprocessing can be seen in Fig. 6.

Since the data were already being split up into individual response cycles, it was decided to also randomly reorder these response cycles to better simulate real-world application and facilitate fair comparison between the performances of the static and dynamic models.

MIMO FIR models of various orders for these dynamically preprocessed data-sets were formulated, and the effects of two different normalization methods were investigated. The evaluation of each model was carried out as follows.

- The dynamic preprocessing method was applied to the original growth phase data-set to produce two separate dynamic input data-sets (each being 180 840 × 6) corresponding to the Toxic 1 and Toxic 2 cultures.
- An output array (180 840 × 4) was formed for each of the two cultures using the classifications gained from the CellFacts information.
- The response cycles of each data-set were randomly reordered.
- If required, a normalization process was applied to the input data.
- A MIMO FIR model was trained on the first half of the set in question using MATLAB.
- The model obtained was used to produce a simulated output from the second half of the data-set.
- The simulated output was sampled at the appropriate points, and the output vectors obtained were converted to a sequence of (274) classifications.

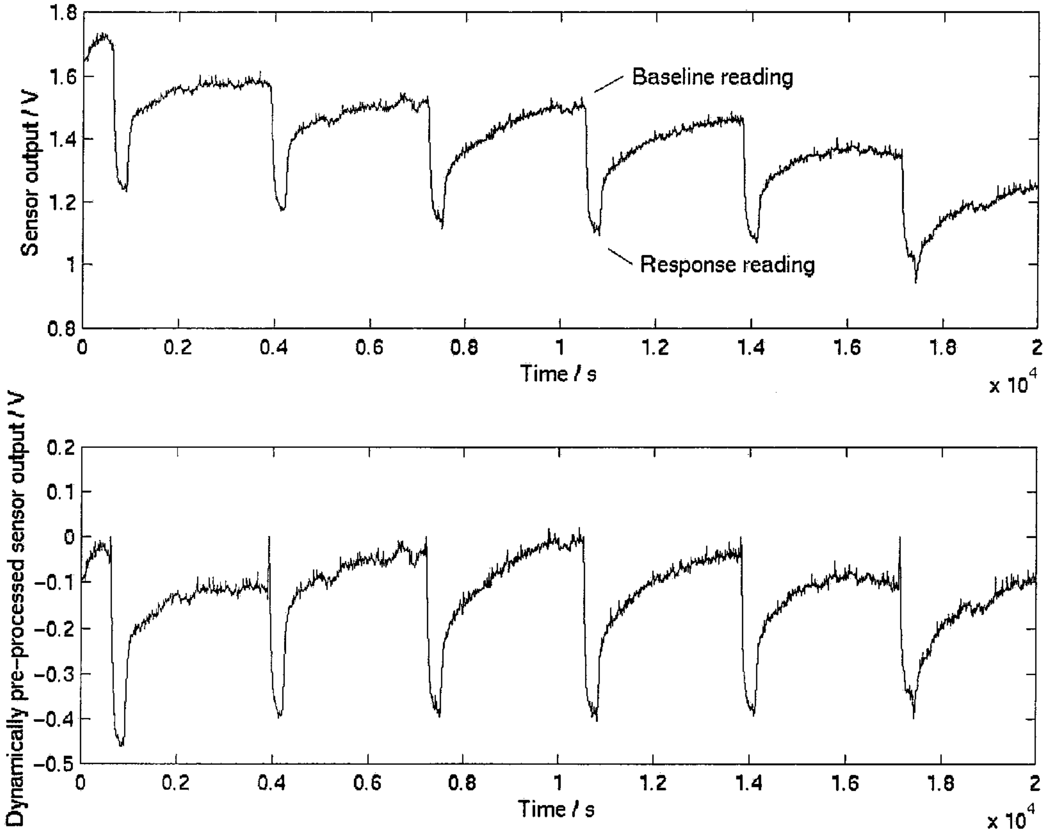

Fig. 6. Upper plot shows a section of the output data from a single sensor for the growth phase experiment. The lower plot shows the same section after dynamic preprocessing to shift each response cycle and thus remove some of the effects of the baseline drift evident in the upper plot.

- The sequence of classifications obtained was compared with the correct classifications to evaluate the success of the model.

The above process was repeated ten times with different random reorderings, and the results averaged.

One of the two normalization processes investigated was the same as used in Section II, where each data vector was scaled so as to have unit length. This constrains the position of the input vector in sensor space to move about on the surface of the six dimensional hypersphere (there were six sensors used), removing to some extent information regarding the intensity (or strength) of the odor and forcing the system to utilize instead information regarding the type (or note) of odor. The other normalization process considered was to normalize (i.e., scale to unit length) only the input data vectors corresponding to the times when the bacterial headspace was sampled, leaving the data vectors corresponding to the headspace samples from the (nonbacterial) control vessel (containing only the nutrient medium) in their original form. The latter normalization process consistently produced superior success rates to those obtained using the former method.

FIR model orders between one and 20 were investigated and generally a model of order around 10 was deemed an appropriate compromise between model complexity and success rate. The success rates obtained with order 10 models are given in Table IV.

IV. Conclusion

The models for the static strain identification data were able to produce very high success rates: 98.4% for normalized, 99.7% for nonnormalised data, using 50% cross validation (using MISO FIR models of order two or three). These success rates increased to 100% on the extension of the FIR models to ARX or MAX model structures.

The models for static data work best with a 'memory' of the last three or so response cycles in order to make a classification. In practical applications it may not be acceptable to wait for three complete exposure cycles to obtain a classification. In contrast, the dynamic models, even though they use higher order polynomials, require only a single exposure cycle to produce a classification (since for the dynamic data, 10 data points corresponds to 100 s, rather than 10 exposure cycles).

The models for the dynamic data produced maximum success rates of 91.3% for the raw data (using an FIR model of order 12) and 99.3% for the normalized data. This showed that the normalization process can be an effective tool for improving the classification success rates when long term drift might otherwise adversely affect the system performance. This might be attributable to the fact that the normalization process reduces the strength of smell information in the data, and thus forces the model to learn (and subsequently make classifications based on) the type of smell encountered.

The success rates of the models for static data, and those for normalized dynamic data, compare well with the success rates achieved using the best artificial neural networks [10] where a successful classification rate of 100% was obtained using a Fuzzy ARTMAP applied to static data from the same experiment.

TABLE IV
SUCCESS RATES OBTAINED FOR 10TH ORDER FIR MIMO MODELS FOR DYNAMICALLY PRE-PROCESSED AND RANDOMLY REORDERED GROWTH PHASE DATA. THE MODELS WERE TRAINED ON HALF OF THE DATA AND TESTED ON THE REMAINING HALF, I.E., TWO-FOLD (OR 50%) CROSS-VALIDATION WAS EMPLOYED.

	Percentage success rate for dynamically pre-processed data-set with		
Bacteria culture:	No normalisation:	Normalisation of whole input data-set:	Normalisation only during bacterial sampling:
Toxic 1	40.4	45.3	61.5
Toxic 2	52.2	61.9	76.6

The models for the growth phase identification experiment data were significantly less successful than those for the simpler strain experiment. The static data from the experiment were randomly reordered to simulate real-world applications (and also to avoid the fact that if the data had not been reordered then the model would have encountered a training set consisting of all the phase 1 data points, followed by all the phase 2 data points etc.). Successful classification rates of up to 82.3% were obtained with the static data. This compares with 95.1% obtained elsewhere [10] using an LVQ artificial neural network.

The models for dynamic growth phase data were considerably less successful than the static models. Although a maximum success rate of 76.6% was achieved using one particular data-set with one of the normalization techniques considered, this success was not repeated using the alternative data-set (where a maximum of 61.5% success was achieved).

The lack of success of linear black box techniques (both static and dynamic) to identify growth phase in comparison to nonlinear neural network techniques [10] may be attributed to the fact that the models used here are linear in nature, whilst the processes being modeled are clearly not. Only FIR models were investigated for the growth phase data due to the computational demands of more complex model structures when dealing with very large data-sets. It is possible that more complex model structures (such as ARX, MAX, ARMAX etc.) may be capable of producing better results.

The success of the black-box models (for these and other applications) might be improved by the use of nonlinear system identification techniques. It should also be noted that the criteria used for parameter estimation were not precisely the same as those used for the evaluation of the resulting models. Thus, further work to produce a more appropriate parameter estimation algorithm may improve the success rates obtained.

In conclusion, it has been shown that simple linear black-box (inverse) models for an electronic nose system can be successfully employed for strain classification of cyanobacteria. The models performed as well as the previously employed artificial neural network techniques, with the advantage that they require less computing power to implement. However, for the more complex problem of growth phase classification, the technique was only moderately successful, failing to compete with the results obtained elsewhere [10] using nonlinear neural network techniques. Thus such modeling techniques could be appropriate for use in relatively simple applications where available computing power is limited, such as in a handheld instrument. Future refinements of the techniques could make them suitable for more challenging classification problems, where currently artificial neural networks are most suitable.

ACKNOWLEDGMENT

The authors would like to thank Dr. C. S. Dow (Warwick University) and Dr. H. W. Shin (LG Electronics, Korea) for their work carrying out the biological experiments. We also thank Alpha-MOS (France) for the odor sensors.

REFERENCES

[1] C. Di Natale, A. Macagnano, R. Paolesse, A. Mantini, E. Tarizzo, A. D'amico, F. Sinesio, F. M. Bucarelli, E. Moneta, and G. B. Quaglia, "Electronic nose and sensorial analysis: Comparison of performances in selected cases," *Sens. Actuators B*, vol. 50, pp. 246–252, 1998.

[2] E. Llobet, E. L. Hines, J. W. Gardner, and S. Franco, "Non-destructive banana ripeness determination using a neural network-based electronic nose," *Proc. Inst. Elect. Eng., Meas. Sci. Technol.*, vol. 10, pp. 538–548, 1999.

[3] J. W. Gardner, H. V. Shurmer, and T. T. Tan, "Application of an electronic nose to the discrimination of coffees," *Sens. Actuators B*, vol. 6, pp. 71–75, 1992.

[4] A. Hierlemann, U. Weimar, G. Kraus, M. Schweizer-Berberich, and W. Göpel, "Polymer-based sensor arrays and multicomponent analysis for the detection of hazardous organic vapors in the environment," *Sens. Actuators B*, vol. 26–27, pp. 126–134, 1995.

[5] J. W. Gardner and P. N. Bartlett, "A brief history of electronic noses," *Sens. Actuators B*, vol. 18–19, pp. 211–220, 1994.

[6] E. L. Hines, E. Llobet, and J. W. Gardner, "Electronic noses: A review of signal processing techniques," in *Proc. Inst. Elect. Eng., Circuits Devices Syst.*, vol. 146, 1999, pp. 297–310.

[7] D. M. Wilson and S. P. Deweerth, "Odor discrimination using steady-state and transient characteristics of tin oxide sensors," *Sens. Actuators B*, vol. 28, pp. 123–128, 1995.

[8] X. Vilanova, E. Llobet, R. Alcubilla, J. E. Sueiras, and X. Correig, "Analysis of the conductance transient in thick-film tin oxide gas sensors," *Sens. Actuators B*, vol. 31, pp. 175–180, 1996.

[9] S. Marco, A. Pardo, F. A. M. Davide, C. Di Natale, A. D'amico, A. Hierlemann, J. Mitrovics, M. Schweizer, U. Weimar, and W. Gopel, "Different strategies for the identification of gas sensing systems," *Sens. Actuators B*, vol. 34, pp. 213–223, 1996.

[10] H. W. Shin, E. Llobet, J. W. Gardner, E. L. Hines, and C. S. Dow, "Classification of the strain and growth phase of cyanobacteria in potable water using an electronic nose system," in *Proc. Inst. Elect. Eng., Sci. Meas. Technol.*, vol. 147, 2000, pp. 158–164.

[11] L. Ljung, *System Identification Toolbox User's Guide*. Natick, MA: The MathWorks, Inc., 1995.

[12] ——, *System Identification—Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[13] J. W. Gardner, "Detection of vapors and odors from a multisensor array using pattern recognition. Part 1: Principal component and cluster analysis," *Sens. Actuators B*, vol. 4, pp. 109–115, 1991.

[14] J. W. Gardner and P. N. Bartlett, *Electronic Noses: Principles and Applications*. Oxford, U.K.: Oxford Univ. Press, 1999.

Graham E. Searle received the M.Math. degree in mathematics from the University of Warwick, U.K., in 1998. He is currently pursuing the Ph.D. degree in electrical and electronic engineering at the same university.

Julian W. Gardner (M'92) joined the School of Engineering at Warwick University, Coventry, U.K., in 1987 and is now Professor of electronic engineering, heads the Electrical and Electronics Engineering Division, and heads the Sensors Research Laboratory. He is author or co-author of over 250 technical papers and patents, as well as six technical books. He serves on several advisory panels on sensors, e.g., for EPSRC, DTI, and IEE Professional Network. His research interests include the modeling of silicon microsensors, chemical sensor array devices, MEMS, and electronic noses.

Michael J. Chappell was born in 1960. He received the B.Sc., M.Sc., and Ph.D. degrees in mathematics, all from from the University of Warwick, U.K.

He is currently Senior Lecturer in the Electrical and Electronics Division of the School of Engineering, University of Warwick. He has research interests in the mathematical modeling and simulation of biomedical, biological and pharmacokinetic processes, structural identifiability, and system identification/parameter estimation. He is presently the Deputy Director of the University of Warwick's Mathematics in Medicine Initiative (MiMI). He is the author or co-author of over 90 research papers.

Keith R. Godfrey received the Dr.Sci. degree from the University of Warwick in 1990 for publications with the collective title "Applications of Modeling, Identification and Parameter Estimation in Engineering and Biomedicine."

He is Head of the Systems Modeling and Simulation Research Group in the School of Engineering at the University of Warwick, U.K., and he is author of a book on compartmental modeling published by Academic Press in 1983, and is author, or co-author, of more than 170 papers.

Dr. Godfrey is a member of the IFAC Technical Committees on Biomedical Engineering and Control, and on Modeling, Identification and Signal Processing. He was one of the recent recipients of the IEE Snell Premium for a paper on wavelet analysis of heart rate variability and its application in detection of sleep apnea.

Michael J. Chapman was born in 1956. He received the B.Sc. degree in applied mathematics in 1977 and the Ph.D. degree in realization theory for infinite dimensional linear systems in 1981 from the University of Warwick, U.K.

During 1980–1984, he held a position as Research Fellow in the Control Theory Centre, University of Warwick. In 1984, he became a Lecturer in Mathematics at Coventry (Lanchester) Polytechnic, now Coventry University, U.K. His areas of interest are modern control theory and compartmental models. He is a Member of the IMA.